Siyuan Dong

A time dependent adaptive learning process for estimating drug exposure from register data - applied to insulin and its analogues

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 1.6.2013

Thesis supervisor:

Prof. Erik Aurell

Prof. Torbjörn Gräslund

Thesis advisor:

D.Sc. Fabian J. Hoti

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Siyuan Dong

Title: A time dependent adaptive learning process for estimating drug exposure from register data - applied to insulin and its analogues

Date: 1.6.2013 Language: English Number of pages:5+54

Professorship: Code:

Supervisors: Prof. Erik Aurell, Prof. Torbjörn Gräslund

Advisor: D.Sc. Fabian J. Hoti

In register based drug research time dependent drug exposure is evaluated based on information available from the prescription register. In Finland the exact daily dosage information is not available and thus has to be estimated from the total amount purchased and the purchase patterns. The aim of this master thesis project is to develop a new algorithm, named Time Dependent Adaptive Learning Process (TDALP), which is a prospective method and applies the historical records to estimate the current insulin daily dosage. Through testing under different situations including both constant daily dosage cases and alterable daily dosage cases, it is demonstrated that the performance of the new algorithm is superior over the more traditional methods.

Keywords: Insulin daily dosage, Time dependent adaptive learning process, Prospective method

Preface

I would like to express my sincere gratitude to all the persons and organizations supporting me during the two years master program.

My gratitude goes first to the European Commission who gave me a precious opportunity to join in Erasmus Mundus programme and supports my study with scholarship generously.

At the same time, My gratitude goes specifically to D.Sc. Fabian J. Hoti who directly guided my thesis work and EPID Research Oy providing this lovely internship position for me. And I also own my great gratitude to my two supervisors: Prof. Erik Aurell and Prof. Torbjörn Gräslund who guided me from macroscopic level and provided me their treasured comments.

Finally, I would like to thank D.Sc. Tuire Tirkkonen, M.Sc. Pia Vattulainen and M.Sc. Solomon Christopher for their cozy cares from academic area and daily life during this thesis work.

Otaniemi, 1.6.2013

Siyuan Dong

Contents

Symbols and abbreviations

Abbreviations

ATC	Anatomical Therapeutic Chemical
SSN	Social Security Number
FPR	The Finnish Prescription Register
FHCR	The Finnish Hospital Care Register
FCDR	The Finnish Causes of Death Register
THL	The National Institution for Health and Welfare
SF	The Statistics Finland
DDD	Defined Daily Dosage
ICD-10	The 10th revision of the International Statistical Classification of Diseases and Related Health Problems
WHO	World Health Organization
TDALP	Time Dependent Adaptive Learning Process
CIR	Carbohydrate-to-insulin ratio
ISF	Insulin sensitivity factor
ERH	Expected relative hazard
ERR	Expected relative ratio
HRT	Hormone replacement therapy

1 Introduction

Since the 1950s and especially after gene recombination technology was successfully applied into pharmaceutical industry, insulin therapy has become the most important medical method to resist severe diabetes. With the revealment of the pathogenesis of diabetes and pharmacology of insulin, different insulin analogues have been designed with different properties and emerged in the pharmacy market during recent 10 years. Insulins treatment has been associated with the risk of hypoglycaemic coma which dues to too low blood glucose levels. A current research problem is to evaluate the risk of the hospitalization due to hypoglycaemic coma in relation to use of insulin and its analogues.

A key problem in evaluating drug exposure is that the individual's daily dosage is unknown. The simplest daily dosage estimation algorithm is simply setting the daily dosage as the universal daily dosage defined by WHO, though it may be far away from the truth. Another approach is to assume each individual has a fixed daily dosage and estimate it by dividing the total amount prescribed by the corresponding follow-up time. Nevertheless the method cannot adapt to real cases in which the daily dosage may vary as the disease progresses and depend on other diabetes medication. Also, the approach violates a fundamental assumption of survival analysis by using future events to make inference on present. Thus there is a need for a new approach, which should be prospective and be able to overcome the drawbacks of the previous approaches by adapting to changes in individuals daily dosage.

Due to the restrictions in the data permit agreement and the daily dosage unknown, this project does not include any real data collected from Finish medical system. In order to demonstrate and evaluate the performance of the new algorithm in comparison to the simpler methods, simulated data were used. Data were simulated to mimic real data by using parameters originating from real data.

2 Background

Diabetes

Diabetes is a chronic disease defined as high sugar (glucose) levels, hyperglycemia, in the blood, in which there are two forms: type 1 and type 2 [41]. Type 1 diabetes is mostly diagnosed in children and adolescence. Patients with type 1 diabetes produce little or no insulin [31]. Type 2 diabetes occurs typically in adults and contributes most diabetes cases. In Type 2 diabetes, patients become less responsive to insulin [41]. Insulin therapy is used to treat hyperglycemia which results in increased risk on the incidence and progression of microvascular and macrovascular complications [26].

Insulin is a peptide hormone generated by β cells of the pancreas. Its main biological function is to cause hepatocytes, muscle cells and fat tissues to store glucose, thus it plays an important role to keep the carbohydrate and fat metabolism balance [28]. Insulin analogues are peptides that are modified from natural human insulin to improve medical features. Insulin and its analogues can be roughly categorized into long acting insulins and short acting insulins. The long acting insulin or insulin analogue is injected subcutaneously once or twice a day to maintain the basal level of blood glucose. The short acting insulin analogue is injected before each meal in order to offset the glucose peak after diet. In this master thesis project, the emphasize is on the long acting insulin and insulin analogues in Finland, including insulin NPH (neutral protamine Hagedorn) , insulin glargine and insulin determir.

- Insulin NPH (ATC code: A10AC01), abbreviated as **NPH** in the follow-up, is a long acting insulin by prepared a suspension of human insulin with neutral protamine and zinc [14]. Because of zinc ion, the complex constructed by insulin and protamine stays equilibrium and release insulin slowly in the body fluid environment.

- Insulin glargine (ATC code: A10AE04), abbreviated as **Glargine** in the follow-up, is a long acting insulin analogue developed by Sanofi-Aventis Corp. It is created by adding two arginine residues (ArgB31, ArgB32) onto the tail of the B chain of human insulin and replacing the asparagine residue (AspA21) into glycine residue (GlyA21) [8], which make Glargine precipitate in the subcutaneous tissue and be gradually absorbed into the blood [18].

- Insulin determir (ATC code: A10AE05), abbreviated as **Determir** in the follow-up, is a long insulin analogue created by Novo Nordisk. Compared with the natural human insulin, a myristic acid is covalently bound to the lysine residue at the 29th position of human insulin's B chain, Determir thus can form a crystal structure with albumins, a family of globular proteins in blood [40]. And as a long acting insulin, Determir can be also slowly released from the complex.

After the first successful diabetes treatment in 1922 [16], insulin therapy has been proved to be a solid method against diabetes. However, exogenous insulin or

its analogues can cause the over attenuation of the blood glucose level that results in hypoglycemia symptoms. Severe hypoglycemia can provoke several health problems, including neuronal death [29], cardiovascular disease and the disorder of haematological values [42]. Elder individuals are more vulnerable against repeated episodes of hypoglycemia than younger ones [7], and have an increased risk of dementia if they have a history of the episodes of severe hypoglycemia symptoms [39]. Therefore it is important to evaluate the hypoglycemia risk related to the use of insulin and its analogues in current pharmaceutical market by epidemiology methods.

Epidemiology

The epidemiology is a set of methodology about the study of the distribution and determinants of disease frequency in human populations [30]. The three core components of epidemiology: distribution, determinants and frequency are the compass of all epidemiologic principles and methods. The measure of disease frequency is the first component, which quantifies the existence or occurrence of disease; the second component to be considered, distribution, is the pattern of disease describing the people getting the disease in the population, the time and location of disease occurrence; the third, determinants, derives from the first two for testing the necessary epidemiologic hypothesis [24].

In epidemiology, the methodologies contain two main approaches: cohort study and case-control study. In a cohort study a group of people in different exposure groups are followed through a period of time to analyse the relationship between the exposures and disease incidence. In a case-control study the subjects occurring disease, named case group, are registered by other methods rather than the follow-up, and health people are used to present control group. If no relationship exists between exposure levels and the occurrence of disease, the distribution of exposure in the case group should be the same as the one of the control group [11].

Although the case-control study eliminating the need of the tracking in the follow-up got thrived in developed countries since the twentieth century and especially after the major public health problems shifting from the acute to chronic diseases [12], it may obtain wrong answer if selection bias happened. The selection bias generated by incorrectly sampling control or case members from the study base is one source of error in the case-control study [11]. Thus the cohort study is still being widely used in many sorts of epidemiology topics. Cohort studies can be classified into two categories: prospective and retrospective. The difference between these two categories is the temporal relationship between the initiation of study and the disease incidence [24]. In the prospective cohort study the events of interest such as the occurrence of disease have certainly not occurred before the study though the subjects may or may not be under the exposure. However in the retrospective cohort study all the events have occurred at the beginning time of the study, thus the retrospective study is quicker and cheaper than the prospective one. Due to the logistics benefit the retrospective design is particular efficient for the study with long latency period [24]. Since the severe diabetes patients need lifelong insulin therapy, the retrospective cohort study was selected for this project.

Registry-based Pharmacoepidemiological Research

Pharmacoepdemiology is the research of the effect of drugs in large numbers of people by epidemiological methods [38]. Besides the randomized clinical trials, the effectiveness and safety of approved drugs in the real world still need to be assessed by pharmacoepidemiological approaches [34]. Since the 1970s, the Nordic countries have begun to collect drug dispensation data to study drug utilization trends and make comparisons between nations [9], however these data cannot guarantee the accurate measure on drug exposure in the population since they are wholesale records rather than individual ones [17]. During the late 1980s, the Nordic countries started to computerize their drug prescription records that made it possible to record pharmacy dispensation for individuals efficiently [21]. In 1994, Finland established a computerized central register on all reimbursed medications that is one of the core components in registry-based epidemiological research[25]. The prescription database is connected with other registries by an unique personal identifier - the social security number (SSN) (fig.1). The other registries including health data, census data, immigration and etc can be selected based on the research question.

In Finland a real epidemiologic research usually needs the Finnish Prescription Register, the Finnish Hospital Care Register and the Finnish Causes of Death Register.

- The Finnish Prescription Register (**FPR**): the prescription database of Finland, contains the prescription records for reimbursed medicine. The prescription data in FPR include the ATC code, the package size, the amount prescribed, the date of prescription, the indication for use and the place of patients' residence.

- The Finnish Hospital Care Register (**FHCR**):the hospital discharge registry of Finland is administrated by the National Institute for Health and Welfare (THL). FHCR contains the detailed records for each hospitalization event including diagnosis information (ICD-10 codes), start date of hospitalization, end date of hospitalization, operation information, the name and district of hospital [33].

- The Finnish Causes of Death Register (**FCDR**): the cause of death registry of Finland is administrated by the Statistics Finland (SF). FCDR records the deceased person's sex, age, place of residence and the cause of death [33].

In this work the simulated data were used to replace the data from real registers for two reasons. The first reason is the restrictions in the data permit agreement, therefore the data from FPR, FHCR and FCDR cannot be directly used in this thesis work. The second reason is the lack of accurate insulin daily dosage information in FPR prevents the comparison of different algorithms. Thus simulated data containing accurate daily dosage were used. For the reason of simplification, only prescription data (FPR) and hospitalization data (FHCR) were simulated. In order to mimic patients' age and gender information, the personal information data were also imitated. Then the estimated daily dosage, cumulative dosage generated

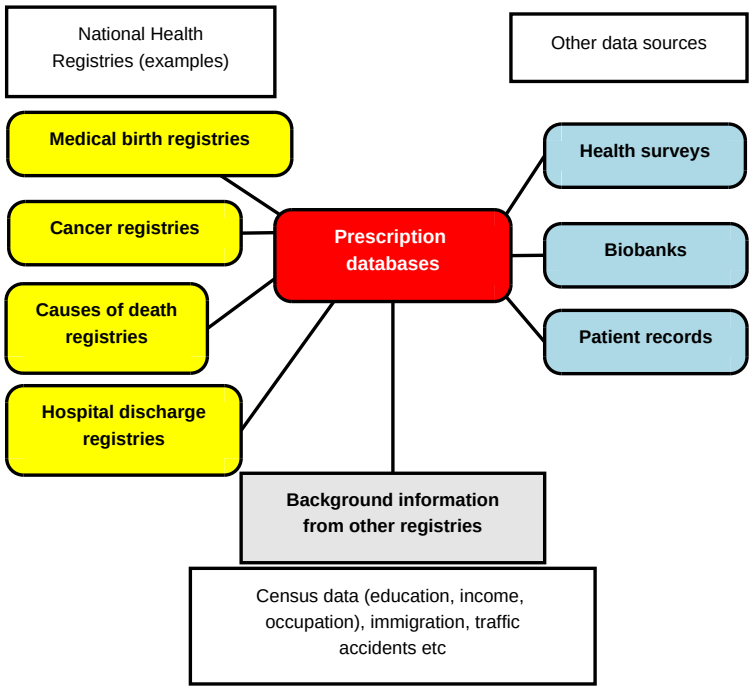


Figure 1: Potential linkages of the prescription databases to other databases and sources. Adapted from "The Nordic Countries as a Cohort for Pharmacoepidemiological Research" by Kari Furu et al, 2009 [17].

from algorithms and the results from Cox model were used to compare with the real values in order to make the correct evaluation (fig.2).

Time Dependent Adapative Learning Process

In order to accurately evaluate the risk of hypoglycemia related to insulin or its analogues, exposure information on individuals is needed. In exposure information, the two most important contents used in the risk evaluation are "current daily dosage" and "cumulative dosage". The cumulative dosage can be calculated from the prescription dosage and current daily dosage, such as defined daily dosage (DDD). According to the definition of WHO, the DDD is the assumed average maintenance dosage per day for a drug used for its main indication in adults [1].

Although for some drugs (pills and capsules) the daily dosage can be reliably derived from the DDD, the approach is not suitable for insulin or its analogues since the DDD does not take into account the fact that daily dosage for different patients can vary significantly (more than 100 time differences) in real cases. Therefore, it is not accurate and reasonable to estimate the length of drug exposure based on the DDD. Another approach that estimates the daily dosage from the average during the follow-up time is a retrospective approach, and cannot deal with time dependent daily dosage. In order to make an improvement, a new approach to estimate the daily dosage, named the time dependent adaptive learning process (TDALP) is

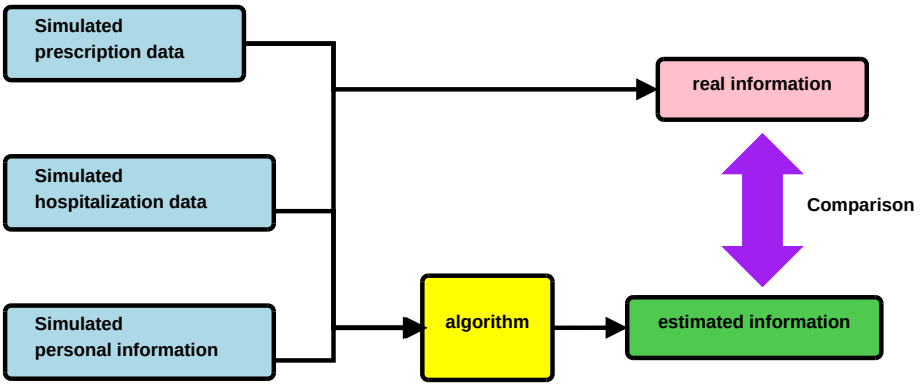


Figure 2: The application of simulated data to evaluate the peformances of different algorithms. In the scheme, the prescription data, hospitalization data and personal information are simulated, then they are used by different algorithms to generate estimated information such as daily dosage, cumulative dosage and so on. Finally the estimated information are compared with the real values embed in these simulated data.

developed. The idea of TDALP is to apply the patient' previous usage pattern to predict his/her current daily dosage. Briefly, for the first prescription the daily dosage is set as 1 DDD, for the remaining records the daily dosage is equal to the dosage of the previous one or several prescriptions divided by the related exposure time.

With the assumption that patients keep the same usage pattern during the period of one prescription, the daily dosage estimated by TDALP can approach to the real one closely. However the estimation may be bigger than the true value that causes the occurrence of gaps (fig.3). Such gaps are only technical results rather than true interruption of treatment. The common sense is that the usage of insulin or its analogues must be uninterrupted if diabetes is so severe that the patient has to receive insulin therapy. In order to eliminate the bias, a prospective method is applied to fill the gaps. According to *Lars Hougaard Nielsen. et al*, the prospective method is better to fill gaps than the retrospective one which violates one fundamental predictability assumption of survival analysis [32].

The Cox Proportional-Hazards Regression Model

The Cox proportional-hazards regression model, abbreviated as Cox model in the follow-up, is one of most commonly used method in survival analysis [22]. Survival analysis is a set of methods for analysing data where the outcome variable is the time until the occurrence of an event of interest. The typical event is death, but the scope of the survival analysis' application is much border at present, such as occurrence of a disease, marriage, divorce, etc [2].

The Cox model, developed by *Cox*, is a proportional-hazards model [13]. For example, a model followed the exponential distribution can be written as [15]

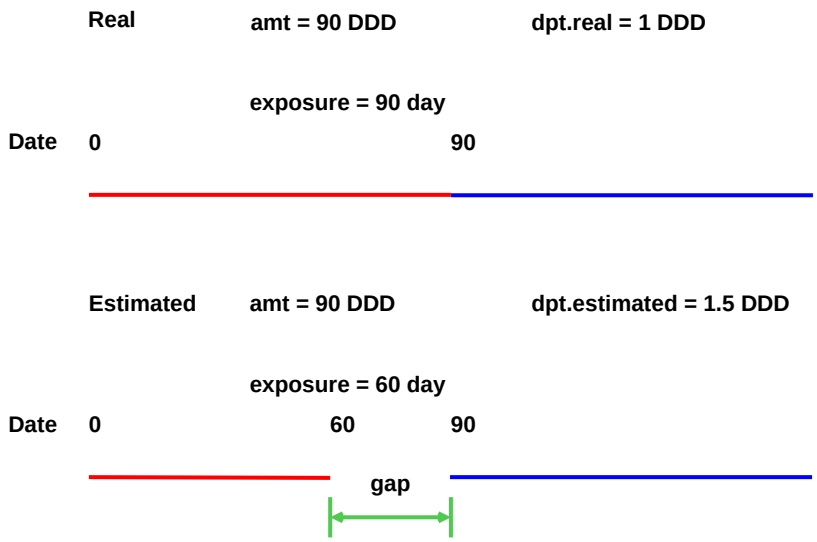


Figure 3: Gap occurrence. The upper is the real case, and the lower is the one estimated by TDALP. The red bonds indicate the first insulin exposure, and the blue bonds indicate the second insulin exposure. The 'amt' is the abbrevation of the amount of insulin, and the 'dpt' indicates dosage per time (day).

$$h_i(t) = \exp\left(\alpha(t) + \beta_1 x_{i1}^t + \beta_2 x_{i2}^t + \cdots + \beta_k x_{ik}^t\right) \tag{1}$$
$$= h_0(t) \exp\left(\beta_1 x_{i1}^t + \beta_2 x_{i2}^t + \cdots + \beta_k x_{ik}^t\right), \tag{2}$$

or equivalently,

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1}^t + \beta_2 x_{i2}^t + \cdots + \beta_k x_{ik}^t \tag{3}$$
$$= \log h_0(t) + \beta_1 x_{i1}^t + \beta_2 x_{i2}^t + \cdots + \beta_k x_{ik}^t, \tag{4}$$

where i is an index for subject, $h_i(t)$ is the hazard of ith subject at time t, $h_0(t)$ is the baseline hazard at time t, and x_{ij}^t is the ith subject's jth covariate at time t. Additionally, a new argument is input to simplify above formula, here

$$\eta_i^t = \beta_1 x_{i1}^t + \beta_2 x_{i2}^t + \cdots + \beta_k x_{ik}^t, \tag{5}$$
$$\theta_i^t = \exp\left(\eta_i^t\right). \tag{6}$$

Therefore, the Cox model is semi-parametric because the form of the baseline hazard $h_0(t)$ would not influence the linear features of covariates. For instance, the relative hazard between two subjects: m and n is:

$$\frac{h_m(t)}{h_n(t)} = \frac{h_0(t)\exp(\eta_m^t)}{h_0(t)\exp(\eta_n^t)} \tag{7}$$

$$= \frac{\exp(\eta_m^t)}{\exp(\eta_n^t)} \tag{8}$$

$$= \frac{\theta_m^t}{\theta_n^t}. \tag{9}$$

The result indicate that the relative hazard is independent from the baseline hazard. And the feature allows to more easily reveal the relationship between the survival time and covariates including age, gender, medication usage without heavy calculation on baseline hazards[15].

In real cases, the baseline hazards can be time-variates rather than a constant value

$$h_0(t) = \lim_{\Delta t \to 0} \frac{Pr[(t < T \leqslant t + \Delta t) | T > t]}{\Delta t} \tag{10}$$

$$= \lim_{\Delta t \to 0} \frac{Pr[(T > t, \leqslant t + \Delta t)]}{Pr(T > t)\Delta t} \tag{11}$$

$$= \lim_{\Delta t \to 0} \frac{p(t)\Delta t)}{S(t)\Delta t} \tag{12}$$

$$= \frac{p(t)}{S(t)}, \tag{13}$$

where T represents the survival time, $Pr(T \leqslant t)$ is the probability of an interesting event happened in the time t, $p(t) = dPr(T \leqslant t)/dt$ which is the probability density function and survival function $S(t)$ is the complement of the $Pr(T \leqslant t)$, $S(t) = Pr(T > t)$.

As mentioned above, the baseline hazards are not necessary to access the relationship analysis, thus they are nuisance parameters. In the profile likelihood method the nuisance parameter is replaced by its most likely value expressed by a set of interesting parameters [3]. For example, in survival analysis the log likelihood of occurrence of one subject's interesting events usually have the Poisson form [11]. Thus the sum of log likelihood over all subjects and all the time in the exposure can be written as

$$\Sigma_{i,t}[d_i^t \log(h_i(t)) - y_i^t h_i(t)] \tag{14}$$

$$= \Sigma_{i,t}[d_i^t \log(h_0(t)\theta_i^t) - y_i^t h_0(t)\theta_i^t], \tag{15}$$

where i is the index for subject, t is the time band, y_i^t is the observation time for ith subject in the time band t, and d_i^t is the indicator for the occurrence of event. If one of ith subject's events happened in the time band t, $d_i^t = 1$, otherwise $d_i^t = 0$.

It is easy to see that given θ_i^t, the most likely value for $h_0(t)$ is

$$\frac{d^t}{\Sigma_i y_i^t \theta_i^t}, \tag{16}$$

where $d^t = \Sigma_i d_i^t$. Substituting the most likely values of the baseline hazards in the formulation (15), it can be rewritten as

$$\Sigma_{i,t}[d_i^t \log(\frac{\theta_i}{\Sigma_j y_j^t \theta_j^t})] + \Sigma_{i,t} d_i^t \log(d^t) - d^t \frac{\Sigma_i y_i^t \theta_i^t}{\Sigma_j y_j^t \theta_j^t} \tag{17}$$

$$= \Sigma_{i,t}[d_i^t \log(\frac{\theta_i^t}{\Sigma_j y_j^t \theta_j^t})] + \Sigma_{i,t} d_i^t \log(d^t) - d^t. \tag{18}$$

In the formula (18), only the one term is related to θ_i^t, thus it can be simplified further as

$$\Sigma_{i,t}[d_i^t \log(\frac{\theta_i^t}{\Sigma_j y_j^t \theta_j^t})]. \tag{19}$$

Finally, the sum of log likelihood expressed as the formula (19) is applied to search the optimal parameters for θ_i^t which maximize the log likelihood.

3 Materials and methods

Simulate data

Due to the restrictions in the data permit agreement and the ineffectiveness of real data in algorithm evaluation, simulated data were used in this project. The simulated data include three components: prescription data, hospitalization data and personal information (fig.2), which are imitated by functions `"simulation.purchase"`, `"simulation.hospital"` and `"generate.personal.information"` respectively. The assumptions for simulation are summarized as:

- Once a diabetes patient initiated insulin therapy, this treatment cannot be interrupted in the follow-up;
- The prescription for each diabetes patient should be prescribed for a full month (30 days) or full months. In each prescription only one insulin or its analogues is assigned to the patient. Patients will go to purchase insulin immediately when they get prescriptions;
- It is impossible for one patient to obtain two or more prescriptions in the same day, though it may occur in real cases;
- The number of hospitalization events obeys a Poisson process, and each event lasts full days (One patient can be sent to hospital and leave hospital by himself or herself on the same date.);
- The occurrence of hospitalization event was influenced by insulin, gender and age which were assumed to be independent from each other.

The function `"simulation.purchase"` was built to simulate the insulin prescriptions given by doctors. Following the assumptions made above, each prescription was given with the dosage of full months, and one month was set as 30 days. Since the number of day during the full months is the duration of the doctors' expected exposure, they were denoted as `'expected.diff'`. The number of month of `'expected.diff'` obeys a normal distribution whose mean was set as 3, and truncated between 1 and 6. The daily dosages for each prescription, denoted as `'dpt'`, were the positive integral multiples of 0.5 DDD. Considering the patients needing high insulin daily dosage are rare, the integral multiples were generated by Poisson distribution whose expected value was set as 1. The amount of insulin or its analogues dosage of each prescription, denoted as `'amt'`, was the product of `'expected.diff'` and `'dpt'`. In real treatments patients may consume insulin or its analogues slower or faster than doctors' expectations. To simulate the fluctuation of medicine consumption, Gaussian noise, whose mean and SD value were set as 0 and 3 respectively, was added into the `'expected.diff'` for each prescription. The original expected exposure modified by Gaussian noise were the real exposures, denoted as `'diff'`. Since one patient may or may not change his/her insulin type, a Markov chain model was implemented to simulate the insulin type transition events.

The transition hazards of the model were given by the previous research result in EPID Research Oy [23]. The maximal number of insulin transitions for each patient was set as 1, and if transition occurs on one patient, the date of transition will be selected from the purchase dates of the relevant patient with equal probability. Besides the properties described above, the `"simulation.purchase"` are also able to simulate the cases with alterable daily dosage in which the daily dosage of one insulin was multiplied with a ratio factor in regular interval, depending on which value was set to parameter `'increase.mode'`.

The function `"generate.personal.information"` was applied to simulate the personal information for each individual. The personal information includes each patient's gender and the age when they begun to accept insulin therapy. The gender of each patient was assigned to male, denoted as `'M'`, or female, denoted as `'F'`, with the equal probability. And the individual age was sampled from 30 to 90 years old with the equal probability.

The function `"simulation.hospital"` was used to simulate the hospitalization events. Since the hazard of hospitalization event occurrence depends on insulin (NPH, Determir, Glargine), gender (male, female) and age ($<$40, 40$\sim$49, 50$\sim$59, 60$\sim$69, 70$\sim$79, $\geqslant$ 80 years old) factors, the function will be called after prescription data and personal information have been simulated. In the simulation of hospitalization event, the three factors were independent, thus the hazard of hospitalization occurrence of one specific category is equal to the product of the reference category group's and the effect of the three factors:

$$h_i = h_{ref} \times Insulin \times Gender \times Age,$$

where h_i is the ith category group's hazard, h_{ref} is the baseline hazard for a male younger than 40 years old who uses NPH, $Insulin$, $Gender$ and Age indicate the insulin, gender and age factors respectively. In order to simply the simulation process, each category's hazard was assumed to keep constant during the period of study. Because it is impossible for a new hospitalization event to occur before the previous event has ended, the beginning date of new event should be equal to or after the end date of the previous event. The interval between the new start and the previous end obeyed the exponential distribution whose rate was equal to h_i. If one hospitalization event crosses the transition boundary between two different insulins, the part that crosses the border will be truncated. If the start date of one new event is bigger than the transition date, the event will be discarded and another simulation will automatically start from the transition date on. The simulation will not stop until all the observation time in the study has been run out.

For the purpose of convenience, the three functions have been packed into `"simulation.package"` (fig.4) in which prescription data, personal information and hospitalization events were simulated and saved as `"mimic_purchase.csv"`, `"mimic_personal_information.csv"` and `"mimic_hospital.csv"` respectively. To evaluate the quality of simulation, function `"simulation.test"` was built which returned the relative hazard of each level in insulin, gender and age factors respectively. The relative hazard were calculated as follow:

$$h_i = \frac{N_i}{T_i} = \frac{N_i}{T_{i,follow-up} - T_{i,hospital}},$$

$$h_{i,relative} = \frac{h_i}{h_{ref}},$$

where h_i, h_{ref}, $h_{i,relative}$ indicate the hazard of ith level, the hazard of reference level and the relative hazard of ith level, N_i indicates the number of hospitalization events in ith level no matter what are the other two factors, $T_{i,follow-up}$ and $T_{i,hospital}$ indicate the related sum of follow-up and the related total duration of events. For example NPH is the reference level for the calculation of the relative hazard of insulin factor. Therefore $h_{Determir,relative}$ was equal to $h_{Determir}$ divided by h_{NPH}. And $N_{Determir}$ and $T_{Determir}$ were the sum of event number and hospitalization time related to Determir no matter patients' gender or age.

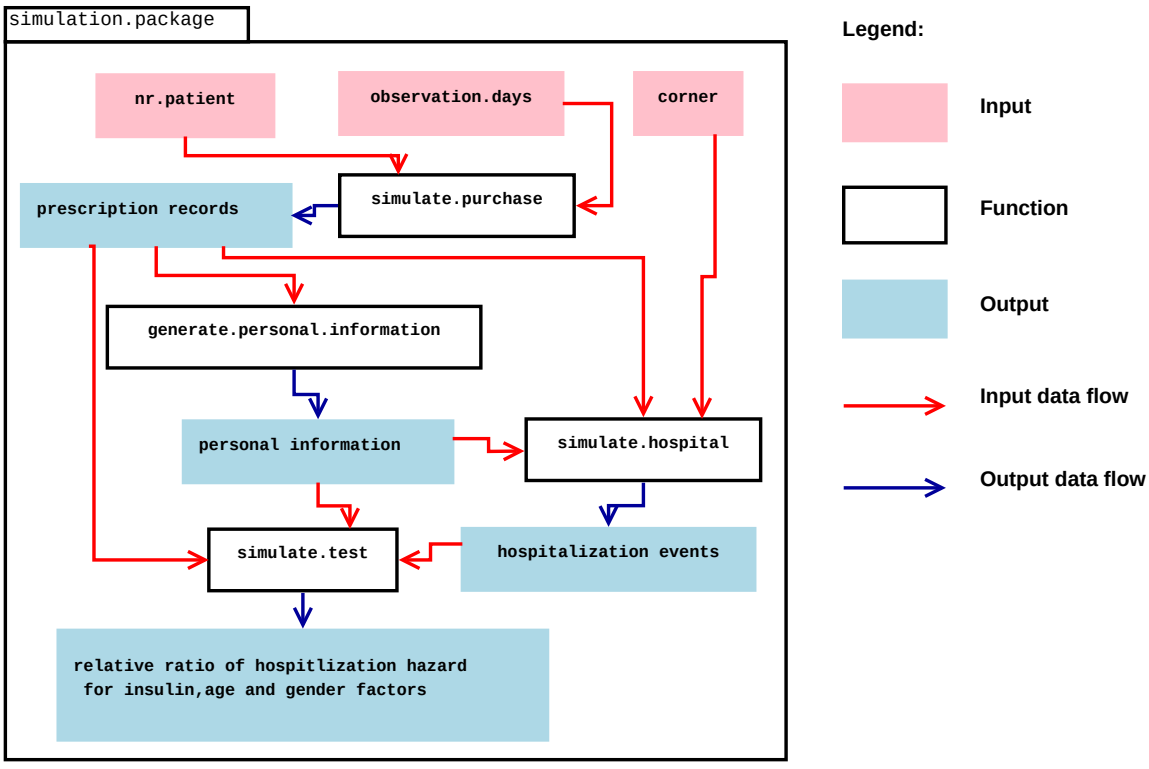


Figure 4: The flow chart of simulation package. Briefly, after inputting three parameters: the number of patient, the length of observation time and the hospitalization event hazard of reference group, named `'corner'`, `"simulation.package"` returns simulated data frames and a quality report.

TDALP

Before periodizing the follow-up time, both the real and the simulated prescription records need to be pre-processed by function `"TDALP"` that was built to estimate the insulin or its analogues' daily dosage for each prescription. Dividing prescription records related to the same type of insulin into clusters is the first step of procedure in TDALP. The cluster is a concept of prescription record group in which the interval of purchase date between two adjacent prescription records is less than or equal to a fixed number, named as `'max.interval'`. Otherwise a splitting point will be set at the purchase date of the latter prescription, since the prescriptions given before are supposed to be so remote that they cannot be applied for the learning process of the new one (fig.5).

In each cluster the estimated daily dosage of the first prescription was set as either 1 DDD or the daily dosage of the last prescription in the previous cluster. The remaining prescriptions' daily dosages were learned from the historical records in the same cluster by this equation (fig.6):

$$dpt_i = \frac{Amt_i}{T_i},$$

where dpt_i indicates the estimated daily dosage of ith prescription in one cluster, Amt_i indicates the total dosage of insulin or its analogues in the historical records before ith in the same cluster, and T_i indicates the total exposure time of related historical records before ith prescription in the same cluster. The number of considered historical records was determined by a parameter, named as `'prev.time'`.

In order to compare with the simpler algorithm used before which assumes the daily dosages of each individual are equal to his/her average daily dosage during the follow-up, it was also built in this project. To conveniently state, this algorithm was called as `average-dpt`. If there is only one prescription record for one patient, the daily dosage of this record will be set as 1 DDD. If there are more than one record, all daily dosages with the same insulin will be equal to the sum dosage of all related prescriptions except the last record divided by the sum of all the exposure time related to this insulin except the last record. The reason for excluding the last record results from the censoring.

Periodization

The periodization is to divide the whole of observation time into small consecutive time bands in which the insulin type, daily dosage and health status of individuals keep homogeneous. To implement the intention, three functions were built: `"use.amt.dpt"`, `"purchase.hospital.fusion"` and `"personal.information.pur.hosp.fusion"`.

For each prescription, the function `"use.amt.dpt"` firstly takes advantage of the amount of insulin dosage and estimated daily dosage generated from algorithms to calculate the length of each estimated exposure:

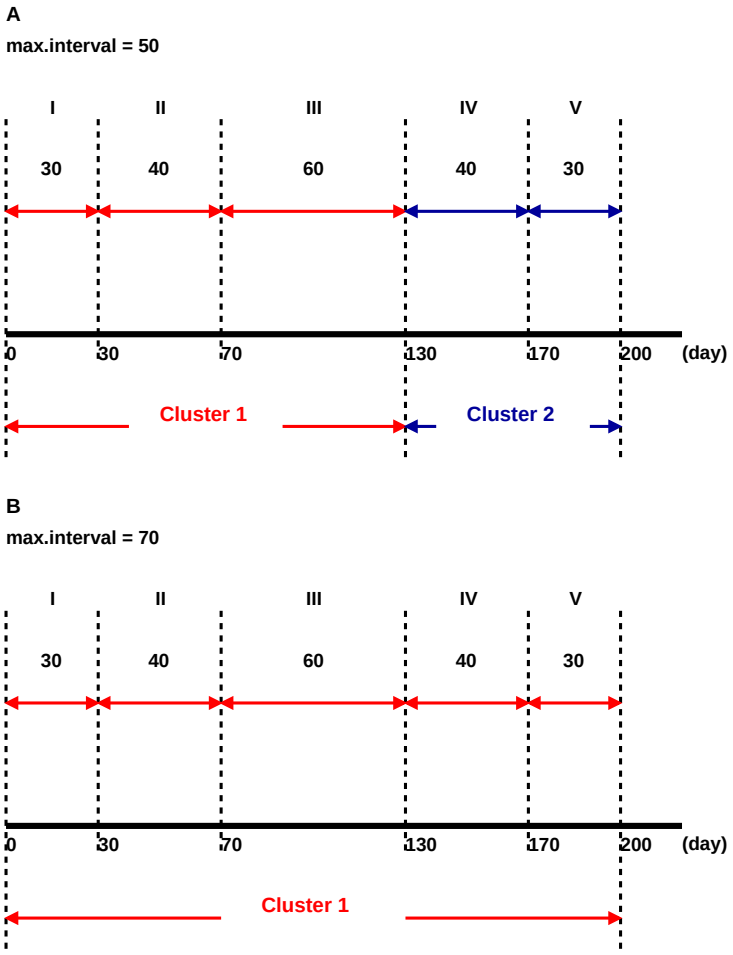


Figure 5: Splitting prescription records into clusters. In the example, there are five prescription records whose purchase dates are 0, 30, 70, 130 and 170 d respectively. If the parameter `'max.interval'` is set as 50 days shown in panel A, the splitting point will be set at 130th day and the five prescription records will be divided into two clusters: the first three grouped into one cluster, and the last two grouped into another cluster; if `'max.interval'` is set as 70 days that is bigger than any interval shown in panel B, no splitting will occur and the five records will be kept into one cluster.

$$exposure_i = \frac{amt_i}{dpt_i},$$
$$end.date_i = start.date_i + exposure_i,$$

where amt_i, dpt_i, $exposure_i$, $start.date_i$ and $end.date_i$ indicate the ith prescription's dosage, daily dosage, the exposure time, the start date and the end date of exposure respectively. If a time band is not covered by any insulin or its analogues, it is named as gap (fig.3).

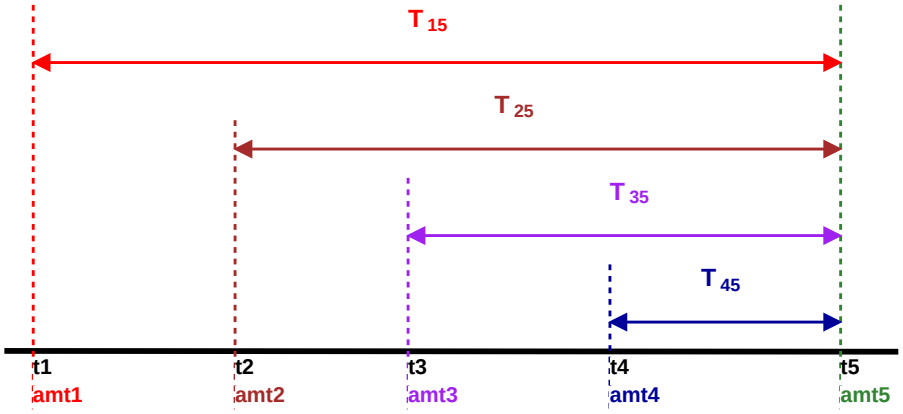


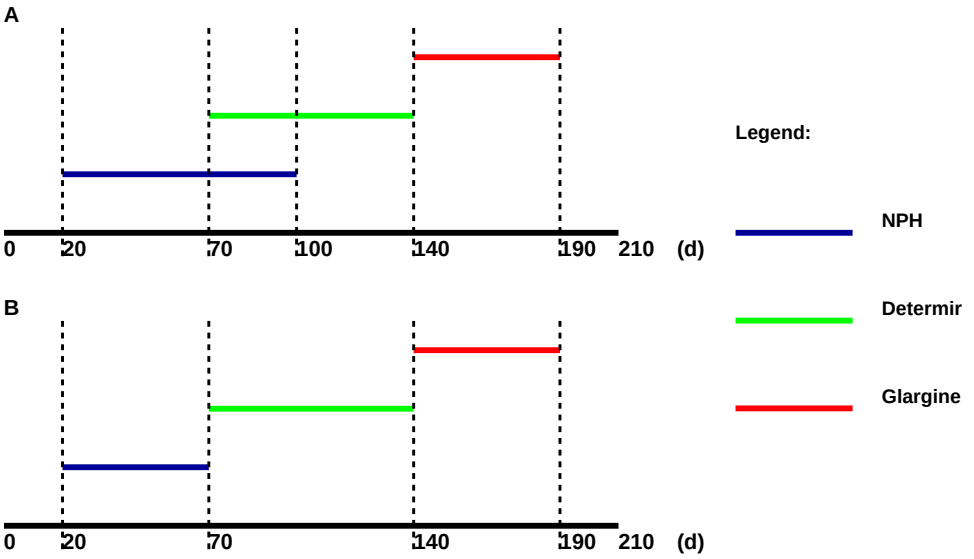
Figure 6: TDALP estimates the daily dosage. One cluster has 5 prescription with the same `ATC` code, which occurred at time points: t_1, t_2, t_3, t_4 and t_5. amt_i indicates the amount of insulin or its analogues in ith prescription, and T_{ij} indicates the interval between the dates of ith and jth purchase. For example, if `'prev.time'` is set as 2, the 3rd and 4th purchase are used to estimate the daily dosage of 5th one. In this case, Amt_5 is the sum of amt_3 and amt_4, and T_5 is equal to T_{35}.

Different long acting insulins are not permitted to be used during the same time band. The simplest case is that one type insulin was prescribed and before running out the insulin in this batch, a new type of insulin was prescribed to the patient. The strategy to treat the overlapping time band is to delete the insulin record from the old batch and to keep the new one (fig.7), since it is assumed that the reason of insulin transition comes from urgent medical requirements, for example, Determir has a better performance for some diabetes patients.

In real cases, the more complicated situation is possible. If the exposures of two prescriptions with the same `ATC` code have overlap, two options are available: *(i)* discard the insulin record from the old batch during the overlapping time band and preserve the new one; *(ii)* keep using the old batch during a permitted time that is determined by parameter `'push.max'` or until using out of all of them, then start to use the new batch (fig.8). Although prescribing different types of insulins to one patient in the same date is rare, it was once recorded in the FPR. In order to deal with the special case, only one type of long acting insulins, NPH, Determir or Glargine, was randomly selected as the really used one and the rest were supposed to be discarded by the patient (fig.9). Due to the independence assumption between long acting and short acting insulins, all the records about short acting insulins were remained.

Since the gaps in the periodized result are technical rather than real, the function `"use.amt.dpt"` was implanted the ability to fill the gaps with virtual extensions whose lengths were determined by the previous estimated exposure of the gap and the extension ratio, named as `'extension'` (fig.10):

$$extension_i = exposure_i \times ratio_{ext},$$



C

start	end	dur	ATC	dpt (DDD)
0	20	20	/	/
20	70	50	NPH	1
70	140	70	Determir	1
140	190	50	Glargine	1
190	210	20	/	/

Figure 7: A simple example of periodization. In the example, one patient got prescriptions of NPH, Determir and Glargine sequentially and all of the prescriptions are given 1 DDD per day. In the time band from the 70d to the 100d there is an overlapping between NPH and Determir as shown in the panel A. In the panel B, the record of NPH, the older batch, is deleted in the time band and the new one, Determir, is retained. The panel C provides the periodized data in which `'start'`, `'end'`, `'dur'`, `'ATC'` and `'dpt'` indicated each time band's start date, end date, duration time, ATC code and daily dosage.

where $extension_i$ is the length of extension for the ith gap, $exposure_i$ indicates the estimated length of previous exposure just before the ith gap and $ratio_{ext}$ indicates the extension ratio parameter, `'extension'`. If the extension is not long enough to cover all the gap, a new cutting point will be set at the ending time point of the extension. If the extension is so long that it overlays with the next estimated exposure, the overlapping part of the extension will be truncated. Although the time bands covered by extension are equivalent to the real exposure, their insulin daily dosages were set as 0.

Periodizing the follow-up time (observation time) based on the prescription information is the first step of periodization. Then the function `"purchase.hospital.fusion"`

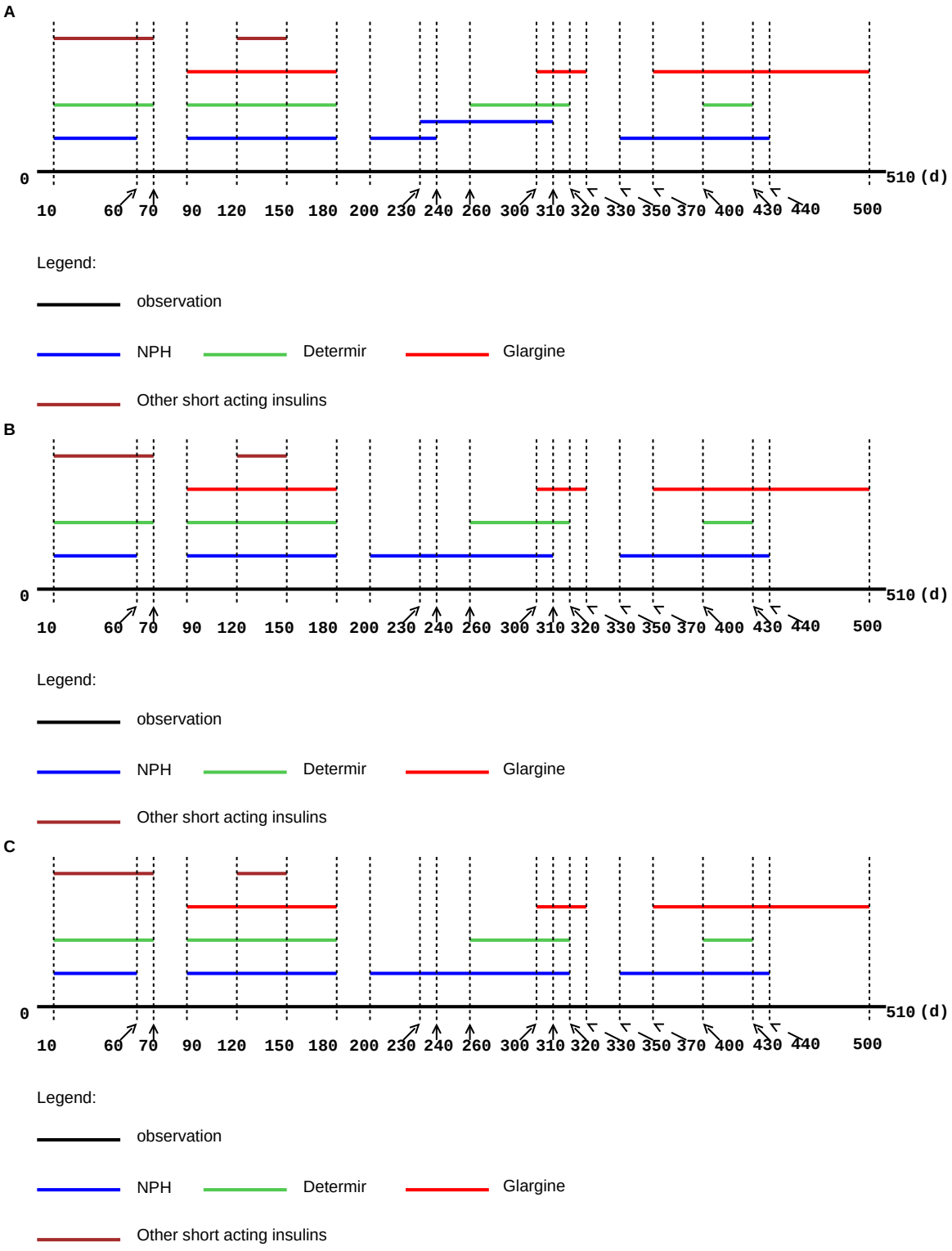


Figure 8: Dealing with the expsoure overlaps occurred in the same insulin. The colorful bands indicate the estimated insulin or its analogues' exposure. The panel A indicates the situation before dealing with overlaps. In the panel A, the 3rd and 4th NPH prescriptions have an overlap in the time band from the 230 d to the 240 d. The panel B and C indicate the results after the application of option *i* and *ii* respectively.

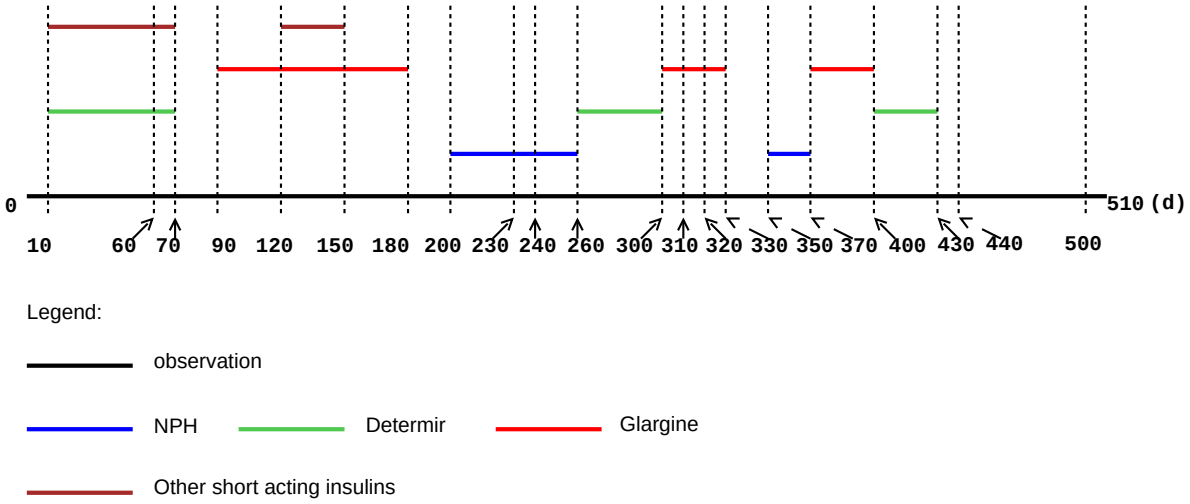


Figure 9: Dealing with the expsoure overlaps among different long acting insulins. After correcting the overlap between the same insulin exposures with option *ii* in the figure 8 C, the new type insulin is prescribed to the patient immediately when the insulin transition occurred. If several different long acting insulins are prescribed on the same date, only one of them will be randomly selected as the one used. For example, on the 10 d, one patient bought NPH, Determir and Other short acting insulins, however only NPH and short acting insulins were supposed to be used.

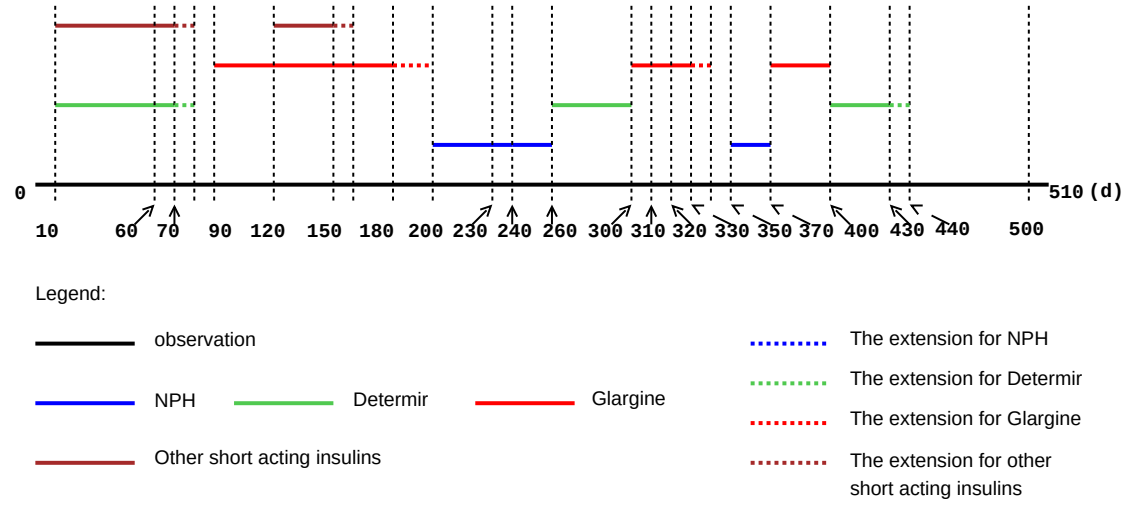
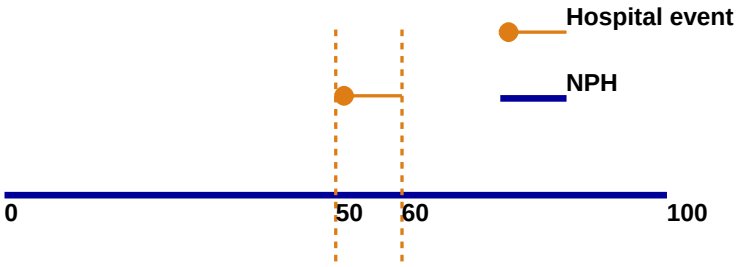


Figure 10: Gap filling. The example in the figure is based on the result in the fig.9. In the case, some gaps were quite long thus they are only partly filled, such as the gap between the 70 d and the 90 d. Others are so short that they are totally covered, such as the gap between the 180 d and the 200 d.

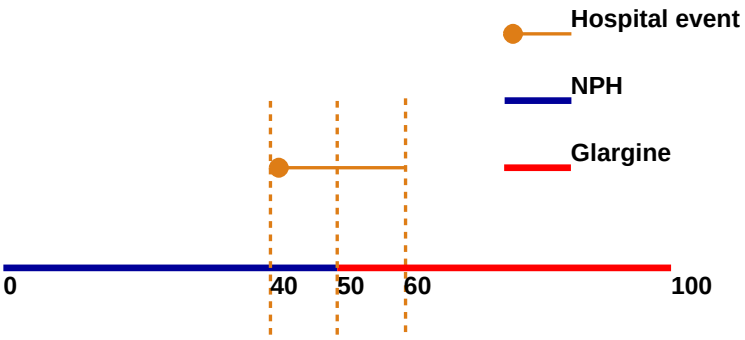
was applied to periodize the follow-up further based on the hospitalization event records. Hospitalization events can be split into two sub-events: `'to.hospital'` and `'in.hospital'`. The former sub-event indicates the moment of hospitalization occurrence, and the latter one indicates patients staying in hospital and receiving medical treatments. Since the `'to.hospital'` events were one moment, their duration time were set as 0. When one hospitalization event is that one patient was sent to hospital and left the hospital by himself/herself on the same date, the duration time of the related `'in.hospital'` was recorded as 0.

For one event that completely locates in one time band, the strategy of the periodization is similar to the one about prescription record (fig.11). If one event spans the transition boundary between two different types of insulins, the event will be divided into three parts: one `'to.hospital'` and two `'in.hospital'` and each insulin exposure will have its own contribution on the related part(fig.12). Besides the simple events above, the function `"purchase.hospital.fusion"` is also capable to treat more complicated one that may be recorded in the real register, such as two hospitalization events occurring on the same date, hospitalization events exactly occurring on the insulin transition boundary (fig.13). To rationalize the temporal and causal relationships between the insulin therapy and the occurrence of hospitalization event, an assumption was made: If one hospitalization event occurs on the transition boundary, the new type insulin rather than the old one will take responsibility for the hospitalization event.



start	end	dur	ATC	to.hos	in.hos
0	50	50	NPH	1	0
50	60	10	NPH	0	1
60	100	40	NPH	0	0

Figure 11: One hospitalization event completely located in one insulin exposure. In the example, the exposure of NPH is from the 0 d to the 100 d, and the hospital event occurs on the 50 d then lasted for 10 days. In the table under the scheme, `'start'`, `'end'`, `'dur'` indicate the start date, end date and duration of each time band respectively. `'ATC'` indicate the abbreviation of insulin's name. `'to.hos'` and `'in.hos'` are binary indicators recording the occurrence and the duration of the hospital event.



start	end	dur	ATC	to.hos	in.hos
0	40	40	NPH	1	0
40	50	10	NPH	0	1
50	60	10	Glargine	0	1
60	100	40	Glargine	0	0

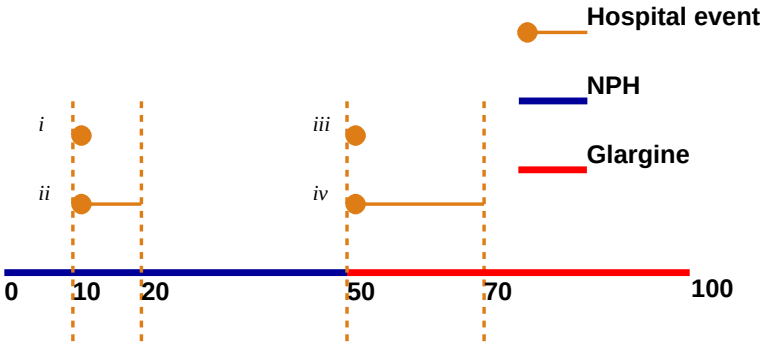
Figure 12: One hospitalization event spans the boundary between two different insulins' exposures. The event has been splitted into three parts: one `'to.hospital'`, the occurrence of hospitalization, and two `'in.hospital'`, locating in NPH exposure and Glargine exposure.

Finally the function `"personal.information.pur.hosp.fusion"` was built to merge the personal information into the periodized data frame. For the purpose of convenience, the three functions have been packed into function `"fusion.package"` (fig.14). In the function, the `"use.amt.dpt"`, `"purchase.hospital.fusion"` and `"personal.information.pur.hosp.fusion"` were called sequentially, and their outputs were save as `"periodized_purchase.csv"`, `"purchase_hospital.csv"` and `"pur_hosp_person_inf.csv"` respectively.

Cox model

The Cox proportional-hazards regression model has been implemented by function `"coxph"` located in the `survival` library in R. The detailed manual of `"coxph"` can be read from the booklet written by *John Fox* in 2002 [15].

Before calling function `"coxph"`, some pre-processes for the periodized data frame are necessary. First, the rows whose `'in.hospital'` is equal to 1 were deleted from the data frame, since the time of patients staying in hospital has no contribution to the occurrence of hospitalization events. After that, function `"recode"` located in the `car` library in R was applied to stratify diabetes patients into categories by their ages: younger than 40, 40 to 49, 50 to 59, 60 to 69, 70 to 79 and elder than 80. A new column named `'age.group'` was added into the periodized data frame to record the label of each age category. Besides that, the `'dur'` with 0 value were added 1e-8



start	end	dur	ATC	to.hos	in.hos
0	10	10	NPH	1	0
10	10	0	NPH	0	1
10	10	0	NPH	1	0
10	20	10	NPH	0	1
20	50	30	NPH	0	0
50	50	0	Glargine	1	0
50	50	0	Glargine	0	1
50	50	0	Glargine	1	0
50	70	20	Glargine	0	1
70	100	30	Glargine	0	0

Figure 13: The complicated example of hospitalization events. In the example, one patient is exposed under two different insulins: NPH (0 - 50 d) and Glargine (50 - 100 d) (The 50 d is the insulin transition boundary). During the 100 days, 4 hospitalization events occur. The *i* and *ii* occur on the 10 d, and lasted for 0 and 10 days respectively. The *iii* and *iv* occur on the 50 d when the insulin transition occurs, and lasted for 0 and 20 days respectively. Due to their zero duration, events *i* and *iii* are supposed to occur before the events *ii* and *iv* respectively. Since the events *iii* and *iv* occur on the boundary, their occurrences are supposed to be related to Glargine.

automatically for technical reason, otherwise the function `"coxph"` cannot work.

Then the `"coxph"` was invoked. Brief, `"coxph"` has lots of arguments including `'formula'`, `'data'`, `'weights'`, `'subset'` and etc, which supply `"coxph"` with great flexibility. However only two arguments among them were frequently used in

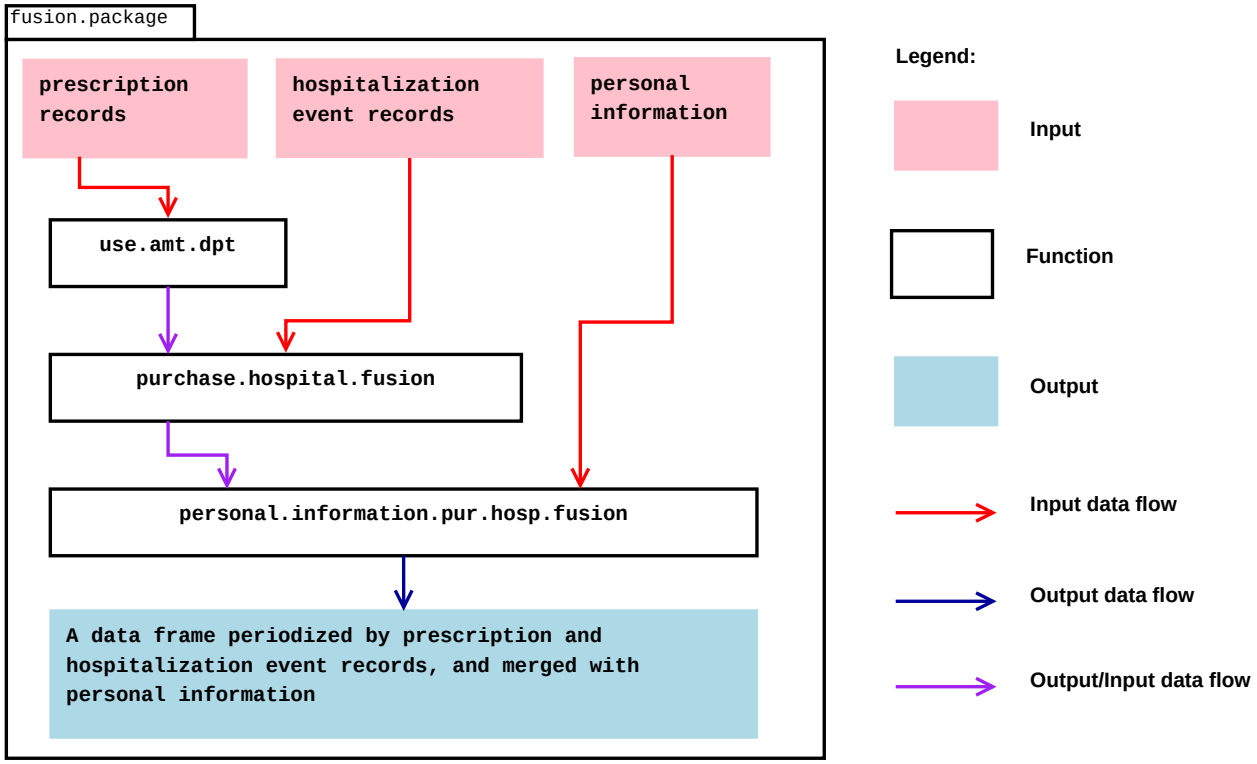


Figure 14: The flow chart of "fusion.package". Briefly, the prescription and hospitalization event records are used to periodize the follow-up time. And the personal information is also merged into the periodized data frame. Then the output, a data frame, is applied for the Cox model analysis.

this project. One is 'data', which is assigned to the periodized data frame after pre-processes. The other is 'formula' whose expression can be written as:

$$\texttt{Left side} \sim \texttt{Right side},$$

where *Left side* is a survival object created by function "Surv". The call to "Surv" takes the form Surv(start.date, end.date, to.hospital) in which 'start.date' and 'end.date' indicate the beginning date and the end date of each time band respectively, and 'to.hospital' is a binary variable indicating whether hospitalization events occurred in related time bands. Since the reference group's hazard was allowed to change with time, the event duration, 'dur', cannot replace the censor time: 'start.date' and 'end.date', otherwise it may make bias. The right side of formula's expression was the sum of covariates. For example, if the factor age and the factor insulin are the potential determinants of the hospitalization occurrence, the command line will be written as follow:

$$\begin{aligned}&\texttt{coxph}(\texttt{Surv}(\texttt{start.date}, \texttt{end.date}, \texttt{to.hospital})\\ &\qquad \sim \texttt{age.group} + \texttt{insulin}.*,\ \texttt{data} = \texttt{periodized.data}),\end{aligned}$$

where `'age.group'` is the label of stratified age covariate, and `'insulin.*'` indicate the binary covariates for sorts of long acting and short acting insulins.

Finally, `"simulation.package"`, `"TDALP"`, `"fusion.package"` and `"coxph"` were assembled and ran along the pipeline sequentially (fig.15). And the results were output by chart or table.

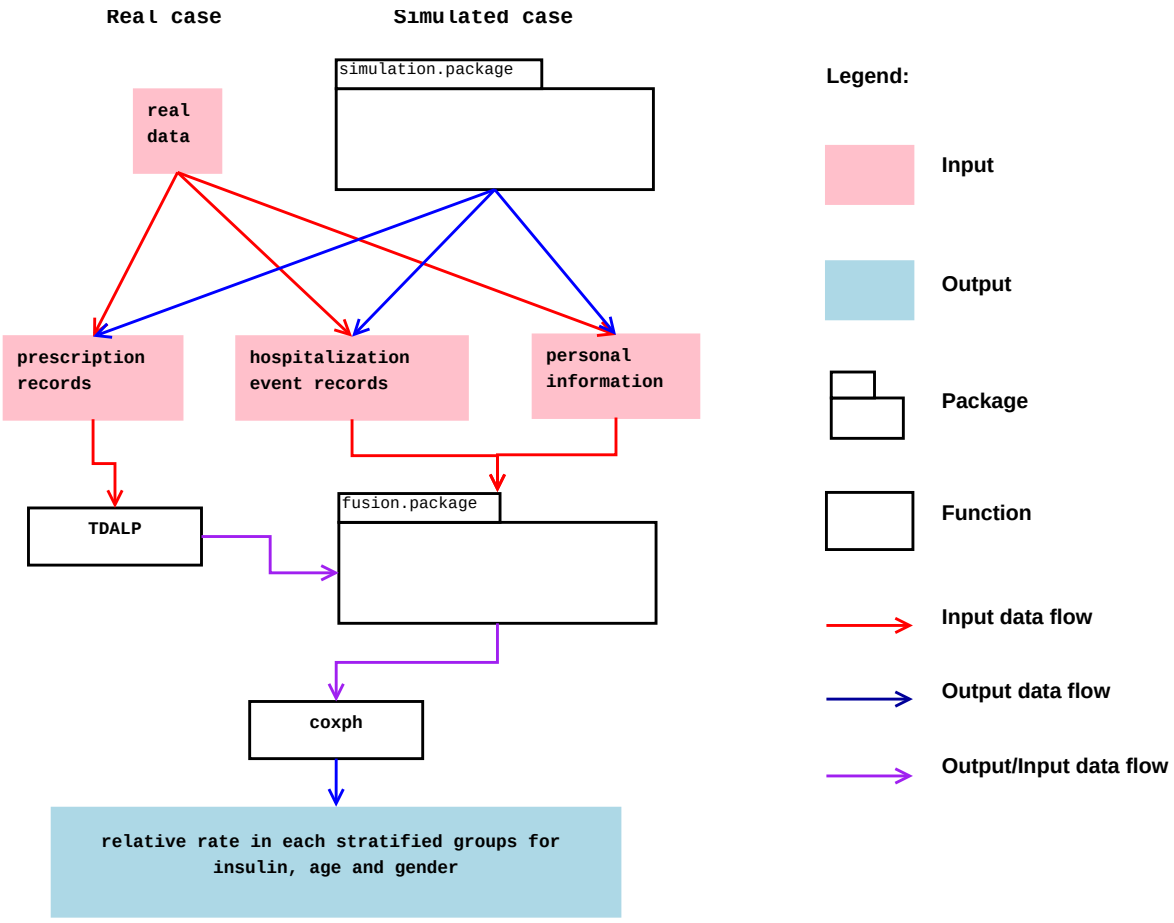


Figure 15: The pipeline for hospitalization relative hazard analysing. Although the real data are not available temperally, the simulated data can be used to optimize and validate the analysis protocal.

4 Results

Test the quality of simulated data

Before moving forward further, the quality of simulated data has to be tested. Briefly, using different baseline hazards for the reference group including 0.01, 0.001 and 0.0001, simulated data were repeatedly generated for 500 times. In each simulation, the simulated data included prescription records, hospitalization event records and personal information of 5000 patients for a follow-up time of 365 days. Within the simulations the baseline hazards were adjusted insulin, gender and age factors. The kernel density estimations of the distributions of relative hazard estimates from each of the 500 simulations were drawn to control the sampling bias in the comparisons (fig.16 and fig.17). The effect of different baseline hazard values is clear: the simulated data with bigger baseline hazard value has higher quality. For example, the distributions of data simulated with 0.01 baseline hazard were bell shape and sharp. The mean and mode values with bigger baseline hazard were more close to expected relative hazard (ERH) (table.1). When the baseline hazard got smaller, the distributions tended to be more stretched, irregular and skewed (fig.16 B and fig.17 F). The mode values of distributions with 0.0001 baseline hazard were far from related ERHs though their mean values were still close to the ERHs (table.1). Therefore it is rational to set the baseline hazard to be the order of magnitude at 10^{-2} in order to take account of quality and computation burden.

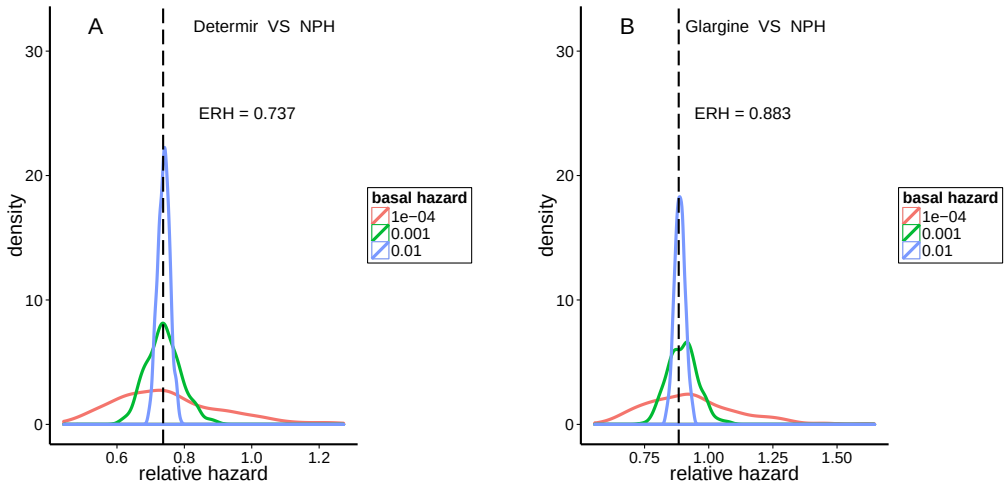


Figure 16: The kernel density estimations of the estimated relative hazard between insulin categories and their reference group. The blue , green and red curves indicate the data simulated with 0.01, 0.001 and 0.0001 baseline hazards respectively. The black broken lines indicate the expected relative hazard (ERH) in each category used for simulation. The panel A and B are the distributions of insulin categories: Determir and Glargine respectively. Their reference group is constructed by all the time bands related to NPH no matter patients' age and gender. The statistic results of the distributions are shown in table.1.

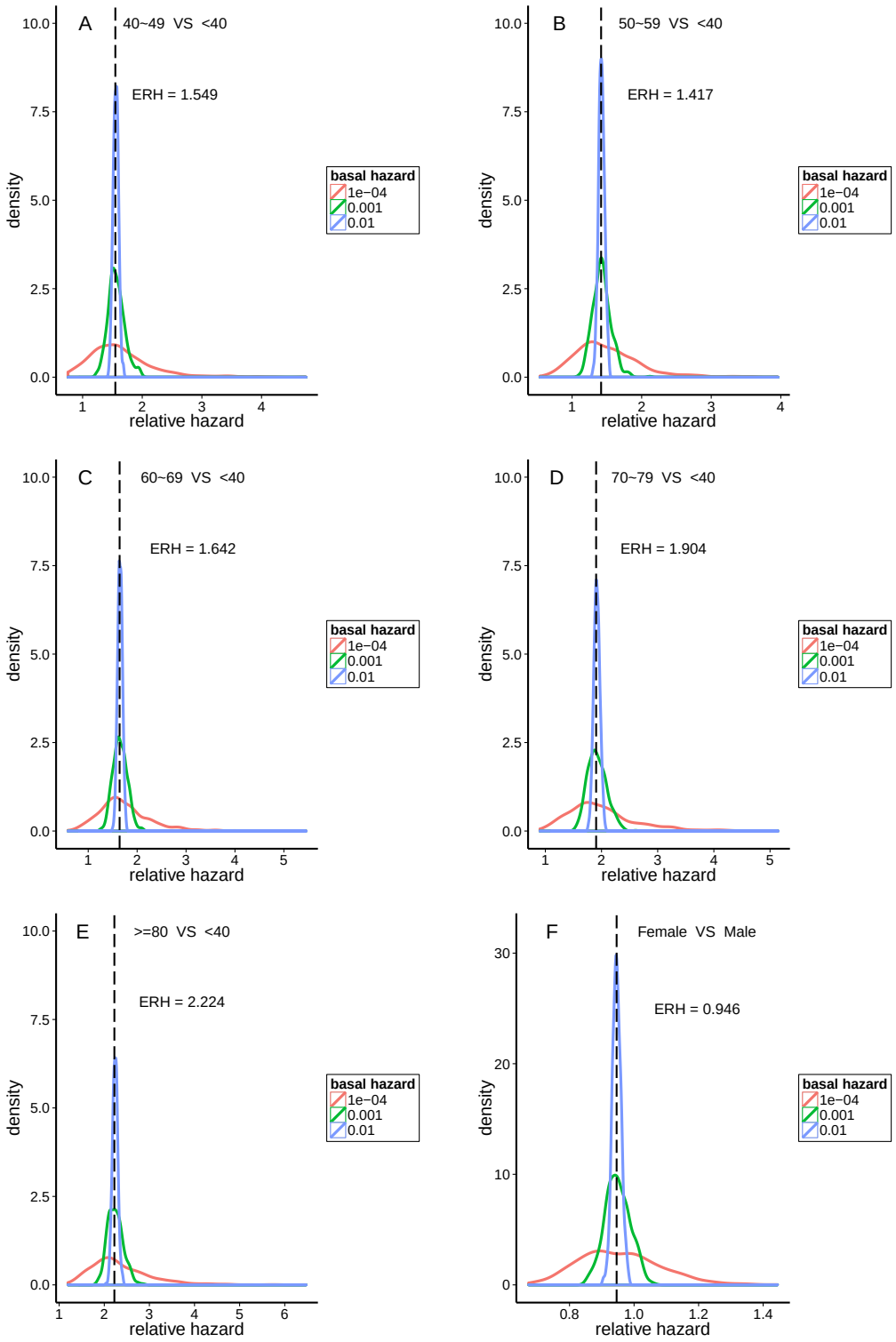


Figure 17: The kernel density estimations of the estimated relative hazard between the age and gender categories and the their reference groups. The panel A, B, C, D and E indicate the age group $40 \sim 49$, $50 \sim 59$, $60 \sim 69$, $70 \sim 79$ and ≥ 80 respectively. Their reference group is constructed by the diabetes patients younger than 40 years old. The panel F is the distribution of the gender category: female, whose reference group is constructed by male patients. The legend of the figure 17 is the same as the figure 16's.

Table 1: The statistical results of data simulation testing

		RH[1] baseline hazard = 0.01			RH[1] baseline hazard = 0.001			RH[1] baseline hazard = 0.0001		
Category	ERH[2]	Mean	SD	Mode	Mean	SD	Mode	Mean	SD	Mode
Insulin[3]:										
Determir	0.737	0.739	0.018	0.741	0.739	0.052	0.736	0.750	0.158	0.727
Glargine	0.883	0.887	0.021	0.886	0.899	0.058	0.915	0.933	0.180	0.920
Gender[4]:										
Female	0.946	0.946	0.013	0.945	0.949	0.039	0.941	0.956	0.126	0.892
Age[5]:										
$40 \sim 49$	1.549	1.557	0.044	1.572	1.566	0.138	1.515	1.637	0.550	1.510
$50 \sim 59$	1.417	1.423	0.043	1.420	1.428	0.132	1.420	1.508	0.484	1.288
$60 \sim 69$	1.642	1.650	0.048	1.636	1.651	0.148	1.618	1.756	0.588	1.550
$70 \sim 79$	1.904	1.916	0.054	1.908	1.922	0.170	1.876	2.011	0.631	1.758
≥ 80	2.224	2.240	0.059	2.259	2.246	0.174	2.228	2.364	0.710	2.088

[1] estimated relative hazard between categories and related reference group
[2] expected relative hazard
[3] the reference group is constructed by NPH exposures
[4] the reference group is constructed by Male diabetes patients
[5] the reference group is constructed by diabetes patients who began to accept insulin therapy younger than 40 years old

Comparison of the performance of different algorithms on the constant daily dosage estimation

The proposed algorithm, `TDALP`, was compared with two older ones: `"default-1-DDD"` and `average-dpt`. As described above, the `default-1-DDD` algorithm sets the daily dosage of each prescription as 1 DDD. In the `average-dpt` algorithm the daily dosage is assumed to be equal to the average daily dosage calculated from the individuals prescription records. In the `TDALP` algorithm the daily dosage was learnt from the historical prescription records. In order to evaluate the difference among the performance on the daily dosage estimation, simulated data were used.

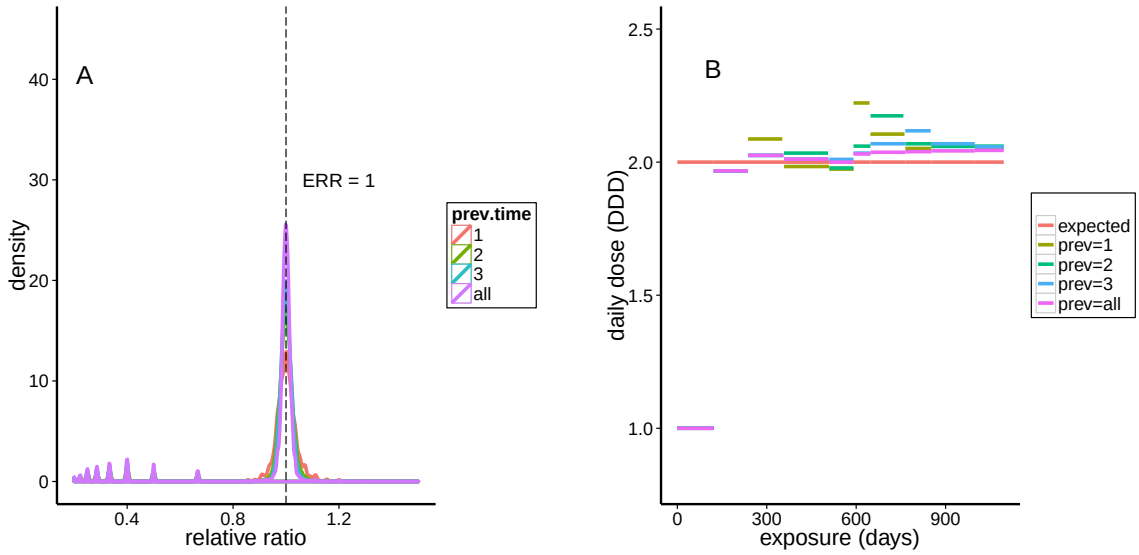


Figure 18: The performance of TDALP with dfifferent `'prev.time'` on the constant daily dosage estimation. The `'prev.time'` is a parameter of function `"TDALP"` determining the size of historical records that will be applied for the current daily dosage learning. For example, if `'prev.time'` is equal to 1, the current daily dosage will be learnt from the previous record; if `'prev.time'` is equal to Inf, the current daily dosage will be learnt from all the historical records in the same cluster. The panel A is the kernel density estimation of relative ratio between estimated daily dosage and expected one in the 500 repeated simulations. Each repetition is set as 5000 patients lasting for 1095 days and the daily dosages kept contant during the whole period. `ERR` indicates expected relative ratio, which is 1 in the case. The more concentrated distribution means the better performance on the learning process. The panel B indicates one example of learning results by `TDALP` with diffferent `'prev.time'` in the 500 repetitions. In the example, the expected daily dosage is 2 DDD indicated by a red long solid line. The other colorful segments indicate the daily dosages estimated by `TDALP` with different `'prev.time'` parameters. The line closer to the expected one means its performance better.

The simulated prescription records can be categorized into two classes. In one

class the daily dosage is constant during the whole exposure period, and in the other one where reduced insulin sensitivity syndrome is imitated the individual insulin daily dosage increases at regular interval. Before comparing the performances of different algorithms, the parameter `'prev.time'` needs to be fixed, due to its importance for `TDALP`. To search the optimal `'prev.time'` for the constant daily dosage cases, 500 times repetitions were made to evaluate the performances of four different `'prev.time'`: 1, 2, 3, Inf. Each simulation contained the prescription records of 5000 patients during 1095 days (3 years). The relative ratios between the daily dosages estimated by `TDALP` with different `'prev.time'` and expected one were drawn into the kernel density estimations (fig.18). In the figure 18 A, the purple curve indicating the `'prev.time'` set as Inf was sharper and more concentrated around the expected relative ratio. Thus the optimal `'prev.time'` was Inf for `TDALP` in the case with constant daily dosage. At the same time, a series of subsidiary peaks from around 0.6 to 0.75 were clear and detectable, which were generated by the `TDALP` mechanically setting the daily dosage of the first prescription record as 1 DDD. The mechanical treatment can also be found in the figure 18 B in which the first daily dosage has been set as 1 DDD though the expected one was 2 DDD.

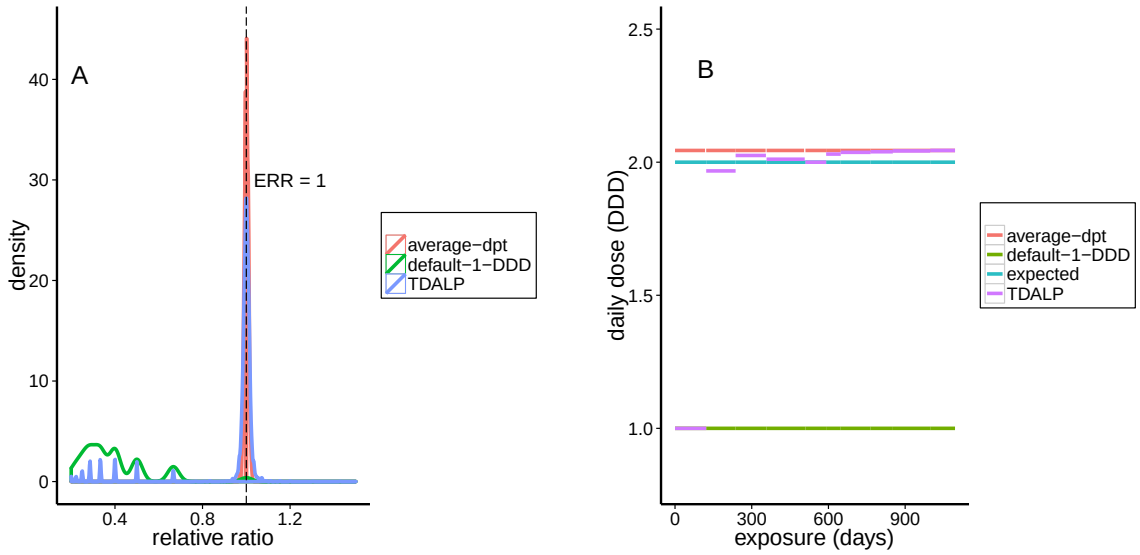


Figure 19: The performance of different algorithms on the constant daily dosage estimation. The legend and description of the figure are similar to the figure 18. Special Note: in the figure, the `'prev.time'` of `TDALP` was set as Inf.

Due to the better performance of `TDALP` with *Inf* `'prev.time'`, the configuration was applied for the comparison with other estimation algorithms including `default-1-DDD` and `average-dpt` on the constant daily dosage estimation. To make the comparison, 500 times repetitions were made and each repetition also contained 5000 patients' prescription records during 3 years. The kernel density estimations of relative ratio of daily dosages estimated by different algorithms were drawn (fig.19). Among the three methods, the performance of `default-1-DDD` was worst,

whose estimation was far from the ERR value in the figure 19 A. Compared with average-dpt, the performance of TDALP was a little worse, since the average-dpt's estimation was sharper and clean without the series of subsidiary peaks. As explained above, the subsidiary peaks in TDALP were formed by its default treatment for the first prescription record.

Comparison of the performance of different algorithms on the alterable daily dosage estimation

In the comparison the daily dosages during the whole follow-up time were allowed to increase at regular interval, which was called as the alterable daily dosage case. The conclusions about optimal configuration and algorithm for the alterable daily dosage estimation are different from the ones for the constant daily dosage.

The alterable daily dosage simulations can mimic the behavior pattern of diabetes patients who were using larger daily dosage than usual to overcome the reduced insulin sensitivity syndrome. Through 500 times repetitions, the optimal 'prev.time' was found to be equal to 1 (fig.20). In the figure 20 A, the kernel density estimation of relative ratio between the daily dosages estimated by TDALP when the 'prev.time' was equal to 1 was sharper and more concentrated on ERR than other configurations, and the subsidiary peak at around 0.7 was smaller. As the same as the previous explanation, the subsidiary peak was formed by the default treatment of TDALP on the first prescription (fig.20.B).

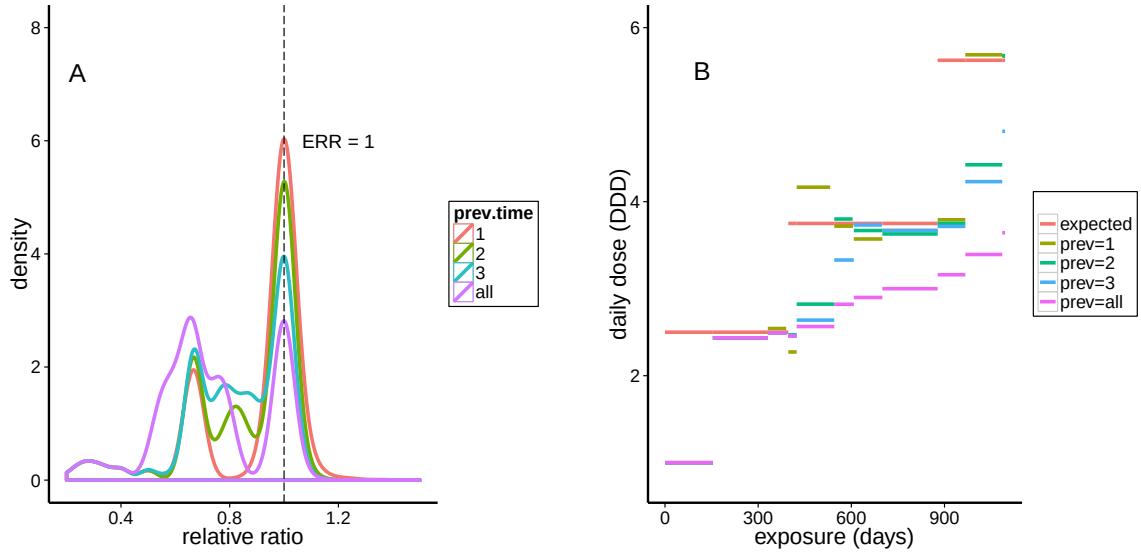


Figure 20: The performances of TDALP with dfifferent 'prev.time' on the alterable daily dosage estimation. The daily dosages in individuals prescription records are allowed to increase 1.5 times at regular interval that is set as 365 days in the comparison. The legend and description of the figure are similar to the ones of figure 18.

Then the performances of different algorithms on the alterable daily dosage estimation were compared (fig.21). The performance of the algorithm `default-1-DDD` was as bad as its behavior in the constant daily dosage case, since the distribution of relative ratio of the estimated daily dosage was irregular and far from the ERR value. For `average-dpt` and `TDALP`, their performances on the alterable daily dosage estimation swapped. In the case with constant daily dosage, the difference between these two algorithms was slight though `average-dpt` performed a little better. However in the case with alterable daily dosage the performance of `TDALP` surpasses the one of `average-dpt` largely. Although the subsidiary peak of `TDALP`'s distribution was significantly detected at around 0.65, its main peak was sharper and more concentrated on the ERR value. In contrast, the distribution of `average-dpt` had three peaks with roughly equal height. In the figure 21 B an example of the daily dosage estimation result among the 500 repetitions was shown. The expected daily dosage was stepwise due to the increment taking account of the reduced insulin sensitivity syndrome in the long-term therapy. The daily dosage estimated by `aveage-dpt` was close to the expected one in the second step segment, whereas it was greatly far from the expected daily dosage in the first and third segments.

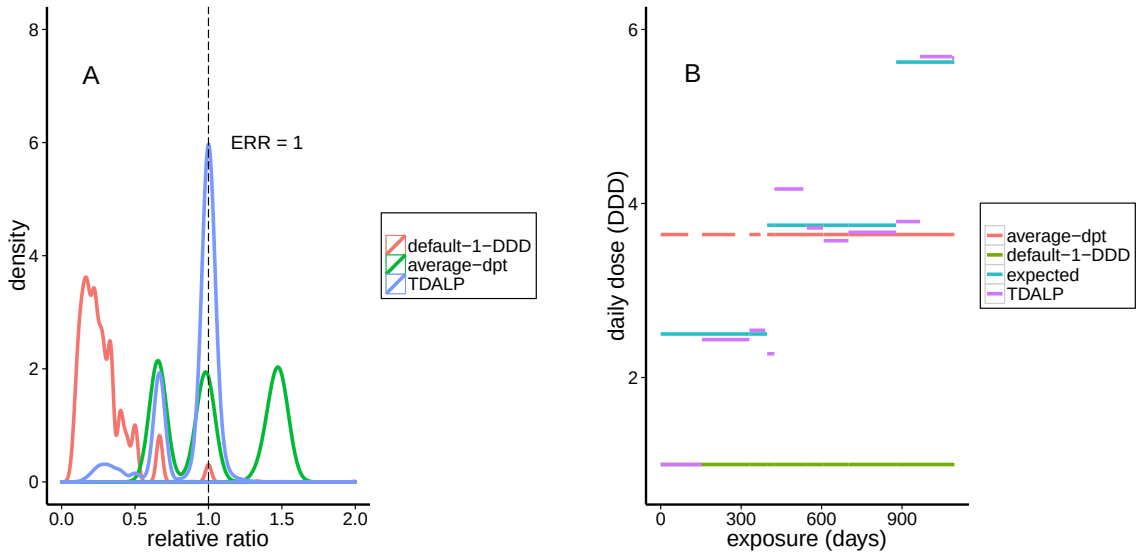


Figure 21: The performance of different algorithms on the alterable daily dosage estimation. The legend and description of the figure are similar to the figure 18. Special Note: in the figure the `'prev.time'` of `TDALP` was set as 1.

Therefore the conclusion can be drawn: In the case whose prescription records were simulated with the constant daily dosage, the algorithm `average-apt` wined a razor-thin victory, and the optimal `'prev.time'` for `TDALP` was Inf; In the case whose prescription records were simulated with the alterable daily dosage, the algorithm `TDALP` had a great advantage compared with `default-1-DDD` and `average-dpt`, and the optimal `'prev.time'` for `TDALP` was 1.

Comparison of the performance of different algorithms on the cumulative dosage estimation

The accuracy of the cumulative dosage estimation is another important factor that needs to be included in the consideration, since the cumulative insulin dosage is a highly potential covariate influencing the hazard of hospitalization occurrences.

In the case with constant daily dosage, the kernel density estimations of the relative ratio between the cumulative dosages estimated by `default-1-DDD` and the expect one was multi-modal and far away from the expected value (fig.22 A). Compared with the `average-dpt`, the `TDALP` was insufficient though its distribution has been quite close to the expected value. The conclusion can also be read from the figure 19 A. In the figure 22 B, the cumulative dosage - time curve of `average-dpt` almost coincides with the expected one, but the curves of `TDALP` and the expected one kept parallel and the difference between them was significant which dues to the `TDALP`'s default treatment in the daily dosage estimation on the first prescription record. Besides that, the figure 22 B also displays the huge deviation of the `default-1-DDD`'s curve from the expected one.

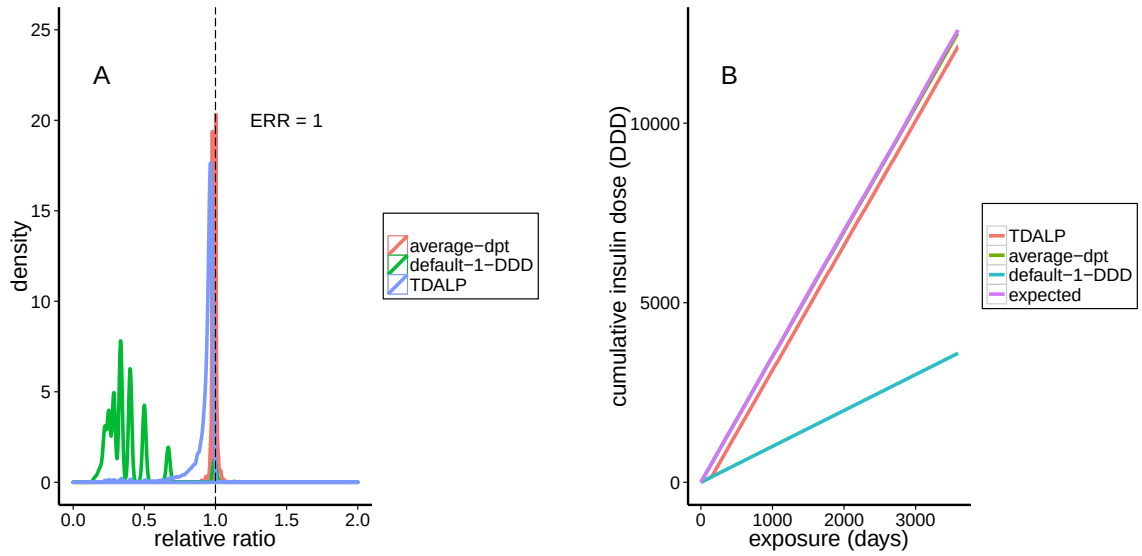


Figure 22: The performances of `TDALP`, `average-dpt` and `default-1-DDD` on the cumulative dosage estimation when daily dosages is kept constant during 10 years. The panel A is the kernel density estimations of the relative ratios between the estimated cumulative dosage and the expected one in the 500 times simulations. The `ERR` indicates the expected relative ratio. The estimation of one algorithm more concentrated on `ERR` indicates that its performance is better. The panel B is one example in the 500 times repetitions which shows the estimated and expected cumulative dosage increments with the follow-up time during the 10 years.

In the case with the alterable daily dosage, the performance of `default-1-DDD` was as bad as its behavior in the constant daily dosage case (fig.23 A). Although

the performance of TDALP on the daily dosage estimation was much better than the one of average-dpt, the average-dpt seems to win the competition in the cumulative dosage estimation. In the figure 23 A, the kernel density estimation of average-dpt was constructed by one main peak and two wings. The main peak was sharp and concentrated on the expected value which indicates the estimated results of average-dpt in the mid-piece of the 10 years exposure. The right wing, which is bigger than the expected value and shows step-wise shape in the figure 23 B, corresponds the initial segment of the follow-up time. The left wing, smaller than the expected value, corresponds the terminal segment in which it shows a huge deviation from the expected curve.

Despite of the illusion of good performance in average-dpt algorithm appearing in the distribution of the relative ratio between estimated cumulative dosage and the expected one during the whole 10 years, it is easy to notice the severe bias occurred in the terminal segment of the follow-up time. In the terminal exposure, the difference between the cumulative dosage estimated by average-dpt and the expected value was getting bigger and bigger with the increment of exposure time. When the kernel density estimations of the data only in the terminal exposure were drawn, the statement got strong support (fig.23 C and D). The figure 23 C was the kernel density estimation for the data during the last 5 years in which the main peak of average-dpt was diminished and its left wing became more apparent. When the observation was narrowed into the last 3 years shown in figure 23 D, the average-dpt's main peak disappeared completely and only left wing remained. Opposite from the average-dpt, the performance of TDALP was not interfered by the size of the observation window and kept the relative ratio stably around 0.85.

Comparison of the performances of different daily dosage estimation algorithms on the cooperation with Cox model

Since the computational load of the periodization before Cox model is heavy, only 50 simulations were repeated for both the case with the constant daily dosage and the one with alterable daily dosage. Each repetition contained 5000 diabetes patients' simulated data including prescription records, hospitalization event records and personal information during 1095 days (3 years). When the prescription records were simulated with the alterable daily dosage, the regular interval for the daily dosage increment was set as 365 days (1 year), and the increasing rate of daily dosage was set as 1.5. Due to the reduced size of repetitions, boxplot was applied to describe the distributions of relative hazards between each category and the related reference group.

In insulin categories related to Determir and Glargine, all of the combinations of algorithms and extension rates had almost the same relative hazard distribution pattern, when the daily dosages kept constant (fig.24). And all the distributions of the relative hazard generated from the Cox model was concentrated on their expected values that was denoted as red long lash horizontal lines, no matter which algorithm and how big the extension rate were used. The similar results were also observed in insulin categories when daily dosages were allowed to increase during

the follow-up time (fig.25). In the age and gender categories, all the combinations of algorithm and extension rate had the same and good performance on the Cox model's results no matter when insulin daily dosages kept constant or were alterable (fig.26 and fig.27). Therefore the Cox model is not a sensitive approach to screen the optimal algorithm and parameter in this project.

However different performances on gap filling among the combinations were detected. In order to measure residual gaps after the filling treatment, residual gap rate was calculated. The rate is defined as the sum of the residual gaps divided by the follow-up time. When the daily dosages was kept constant, the residual gap rate of the `default-1-DDD` was smaller than the ones of the `average-dpt` and the `TDALP`, and reduced linearly rather than exponentially in other two algorithms with the increment of the extension rate (fig.28 A). After the extension rates exceed 0.10, none of the `average-dpt` and the `TDALP` has significant improvements on the gap filling. Therefore the golden mean value of extension rate was 0.10. The performance of `average-dpt` was slightly better than the one of `TDALP`, which was consistent with the conclusion in the comparison of the daily dosage estimation.

When the daily dosages were alterable, the performance of `default-1-DDD` was similar to itself in the case with the constant daily dosage (fig.28 C). Although the residual gap rate of `average-dpt` switched to reduce linearly with the increment of the extension rate (fig.28 B), the pattern of `TDALP` did not change (fig.28 C). The performance of `TDALP` was much better than the one of `average-dpt` (fig.28 B). For example, when the extension rate was set as 0.10, the residual gap rate of `average-dpt` was around 0.09, which was almost 45 times of the one of `TDALP` in the same condition. The low residual gap rate of the `default-1-DDD` was an illusion that is explained in the summary section.

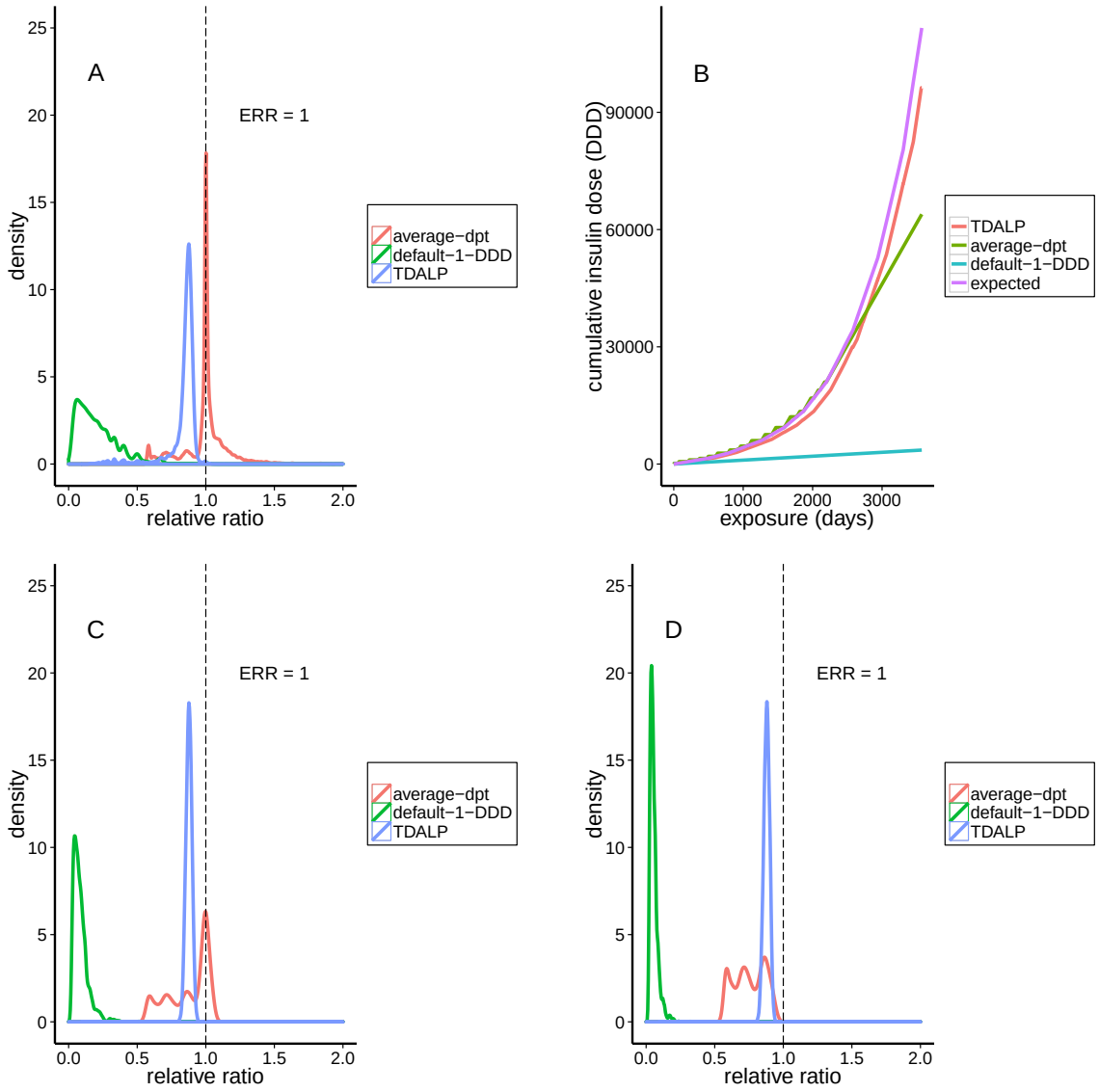


Figure 23: The performances of `TDALP`, `average-dpt` and `default-1-DDD` on the cumulative dosage estimation when daily dosages are allowed to increase during the 10 years. The panel A is the kernel density estimations of the relative ratios between the estimated cumulative dosage and expected one during the whole 10 years in the 500 times repetitions. The panel B is one example among the 500 times repetitions which shows the estimated and expected cumulative dosage increments with the follow-up time during the 10 years. The panel C and D are kernel density estimations of the relative ratios between the estimated cumulative dosage and expected one during the last 5 years and 3 years respectively.

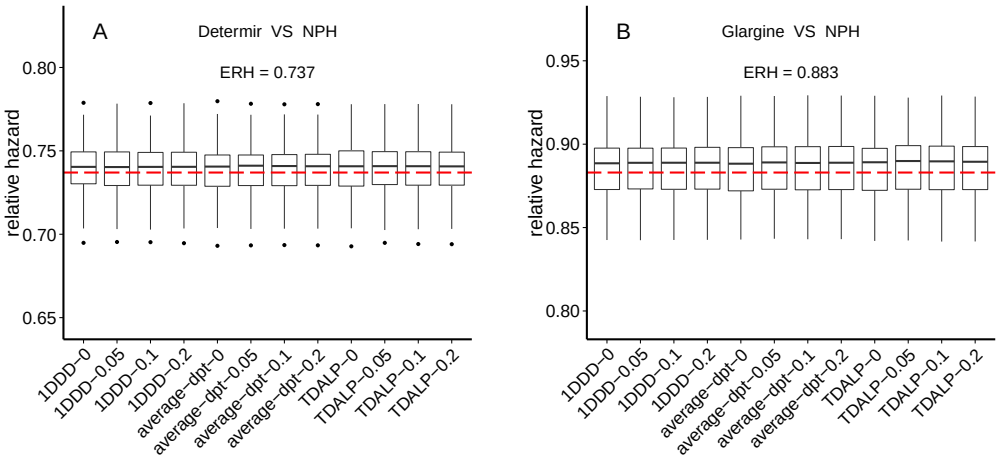


Figure 24: The performance of all the combinations of algorithms and extension rates on the Cox model in insulin categories when daily dosages kept constant. There are three different algorithms: `default-1DDD`, `average-dpt`, `TDALP`, and four different extension rates: 0, 0.05, 0.10, 0.20, thus twelve different combinations are tested. In each combination, relative hazards calculated by the Cox model are drawn with boxplot. Besides that the expected relative hazard (ERH) are drawn with a red longlash horizontal line. The panel A is the result of Determir category, and the panel B is the one of Glargine category.

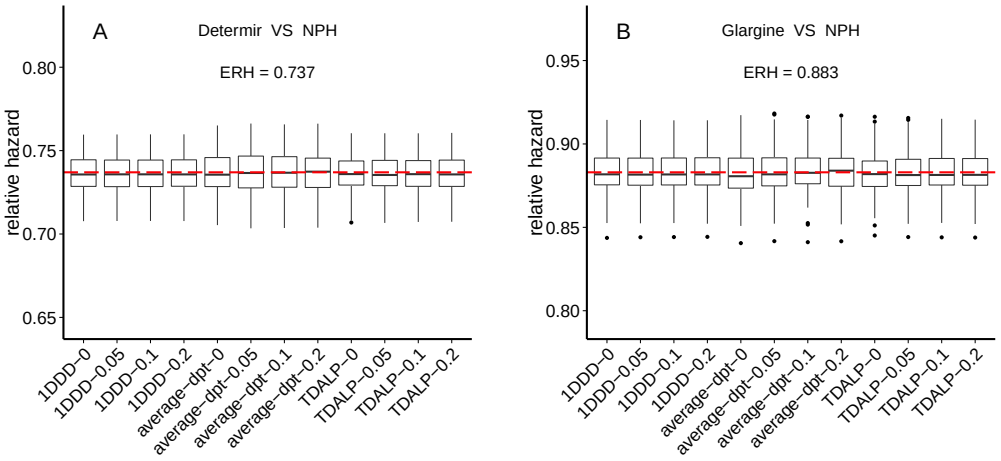


Figure 25: The performance of all the combinations of algorithms and extension rates on the Cox model in insulin categories when daily dosages are allowed to increase. The panel A is the result of Determir category, and the panel B is the one of Glargine category. The legend and annotation of the figure are the same as the ones of figure 24.

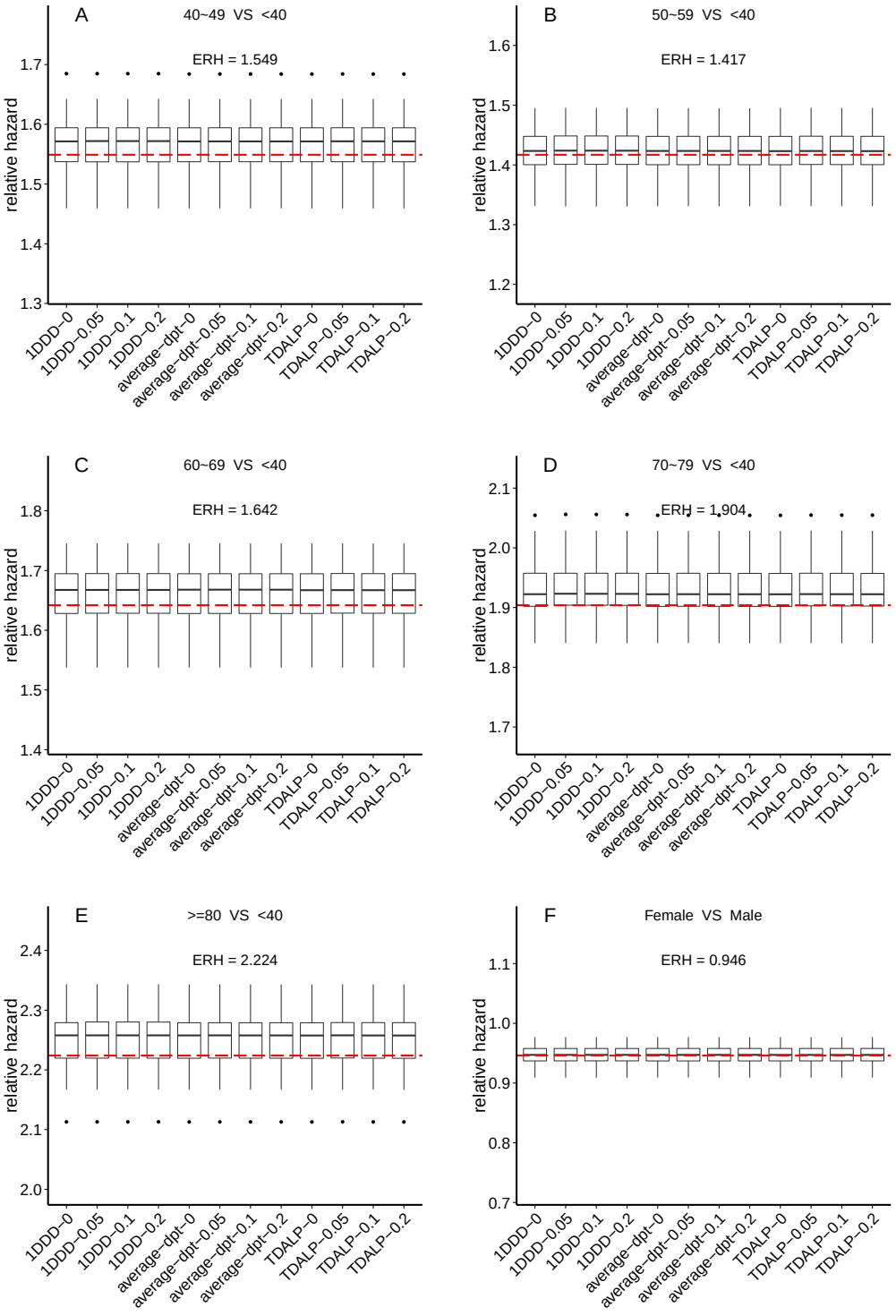


Figure 26: The performance of all the combinations of algorithms and extension rates on the Cox model in age and gender categories when daily dosages are kept constant. The panel A, B, C, D and E indicate the age group $40 \sim 49$, $50 \sim 59$, $60 \sim 69$, $70 \sim 79$ and ≥ 80 respectively. Their reference group is constructed by the diabetes patients who begun to accept insulin therapy younger than 40 years old. The panel F is the result of gender category: female, whose reference group is constructed by male patients. The legend and annotations of the figure is the same as the one of the figure 24.

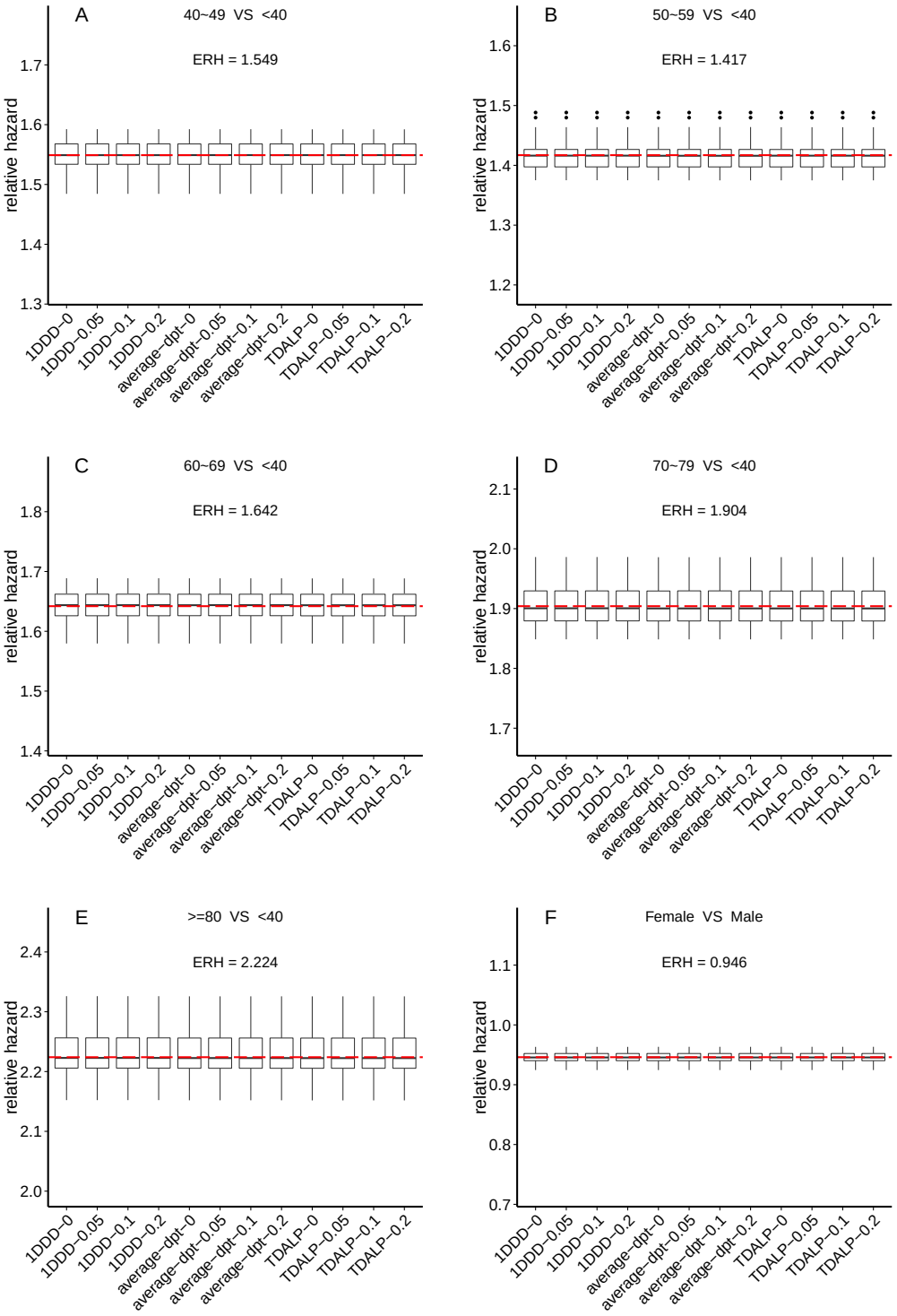


Figure 27: The performance of all the combinations of algorithms and extension rates on the Cox model in age and gender categories when daily dosages are allowed to increase. The legend and annotations of the figure are the same as the one of figure 24.

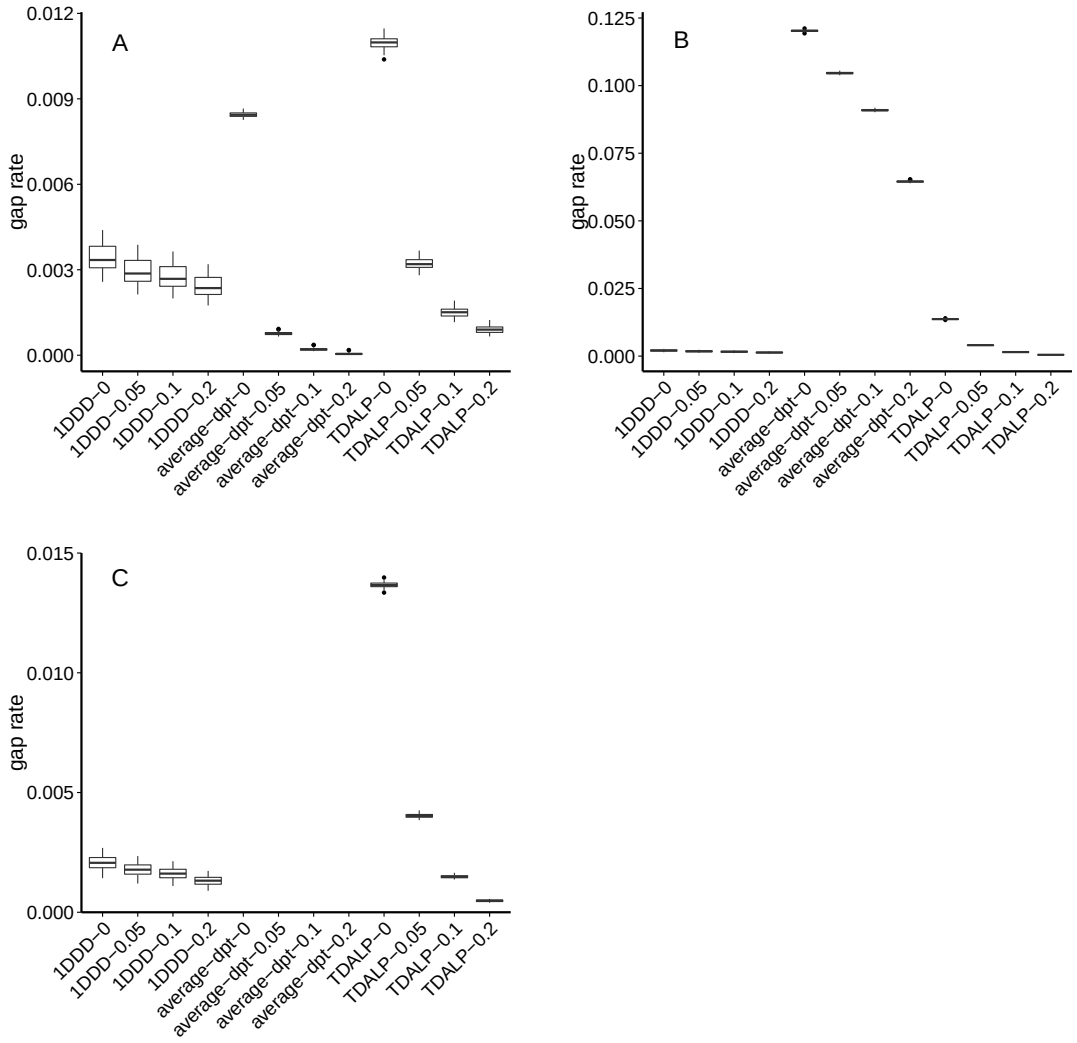


Figure 28: The performance of the gap filling of different combinations of algorithms and extension rates when daily dosages are kept constant and were alterable. The panel A is the result when daily dosages kept constant, the panel B is the result when daily dosages were alterable, and the panel C is the zoom-in of the panel B.

5 Summary

Due to the restriction of the data permit agreement and the lack of insulin's daily dosage in the FPR, the real data cannot be used to test the sensitivity of TDALP. To solve the problem the simulated data were generated. In the simulated prescriptions, each record was assumed to be prescribed by the doctor as the number of full months. Nevertheless, the real exposure can be longer or shorter for many reasons, such as the fluctuation of carbohydrate intake and the occasional early purchase in which patients will go to get prescriptions several days earlier before they run out all their insulin. In order to simulate the moderate fluctuation happened in insulin exposure, Gaussian noise was added to the length of each expected exposure. In this project work only long acting insulins' prescriptions including NPH, Determir and Glargine were simulated.

Besides the exposure time, another key information in simulated prescriptions is the insulin's daily dosage. In this project, each patient' health condition and diet habit was assumed fixed during a single prescription, thus their insulin daily dosages stay constant in each prescription as well. To match the real situation better, the function `"simulation.purchase"` has considered the insulin insensitivity that can be induced by the immunocreativity for DNA recombinant human insulin [27] [37], the reduced affinity between insulin and its receptor after long term therapy [36], and the impact from other medicine [6] [19]. Therefore, if `insulin.resistance` mode is open, the daily dosage will be allowed to increase at regular interval. In addition, the value of each patient's daily dosage may be controversial, since the `"simulation.purchase"` randomly and uninterpretably set them as the integral multiple of 0.5 DDD. In the real therapy, insulin's daily dosage is calculated according to the amount of carbohydrate in meals, carbohydrate-to-insulin ratio (CIR) and insulin sensitivity factor (ISF). CIR is the amount of intaken carbohydrate that covered by 1 unit of insulin [4], and can be quite different in individuals. For example, one adult' CIR is around 15, but a young child's CIR might be 150 [5]. ISF is the amount of blood glucose lowered by the injection of 1 unit of insulin [4]. However none of the three factors has been recorded and analyzed. If the three factors are randomly generated, the action will be equivalent to randomly generate daily dosage. Thus in the case the simplification on daily dosage is still acceptable.

As described above, the simulated prescription records are mainly determined by the two parameters: the exposure time and the daily dosage. The simulated hospitalization event records also has its own determinants. The hospitalization events are dependent on the baseline hazard, the hazard of the reference group, and the other group's relative hazard. The bigger baseline hazard makes hospitalization events occur more frequently in the simulation process. The simulated hospitalization event records with bigger baseline hazard have higher quality, thus in this project the baseline hazard was final set as 0.01 though a more realistic value would be 0.0005 as estimated in the previous result of EPID Research Oy [23]. The relative hazards used in the simulation of the hospitalization events were chosen according to values obtain from ENCePP. It is already known that the long acting basal insulin analogues, whose sustained release makes them more similar to the endogenous

insulin secretion, have better performance to control blood glucose level and reduce the risk of hypoglycemia [20][35]. However, the solution of NPH in blood is rapid. If NPH is injected subcutaneously at night, its drug effect peak can coincide with the lower serum glucose levels associated with nocturnal metabolism which potentially aggravates nocturnal hypoglycaemia. Therefore, the relative hazards of Determir and Glargine in the result of ENCePP are 0.737 and 0.883 respectively. Besides that, the covariates, gender, age and insulin, are simply assumed to be independent from each other in the simulations. The assumption is relatively weak, since the three factor may be correlated in real cases. In the future, some improvements may be possible by considering the interaction terms among gender, age and insulin factors.

When dealing with gaps between the drug exposures, exposure extensions were applied to fill them. As described above, during such extensions the binary indicators of sorts of insulins named `'insulin.*'` are set as 1, but their daily dosages named `'*.dpt'` are assigned to 0. Thus plateaus will be formed in the cumulative dosage time curve (fig.29). In the figure 29, the blue solid line indicates the real cumulative dosage time curve, and the red solid lines indicate the estimated cumulative dosage time curve from TDALP. Compared with real exposure, the estimated daily dosages were bigger which cause the curve to be steeper than the truth, and exposure lengths showing up shorter resulted in gaps. Then the exposure extensions (red broken lines) were supplied to fill the gaps. The deviation of the estimated curve from the real one is obvious. For example, at time points T_a and T_b, the estimated cumulative dosages (the values of A_2 and B_2 on *cum.amt* axis) are bigger than the real cumulative dosages (the values of A_1 and B_1 on *cum.amt* axis). Although the relative error between the real cumulative dosage and the estimated one is acceptable when the follow-up time is long enough such as T_b, the relative error at T_a, when is the beginning of the follow-up, can be significant and is actually inevitable in the prospective approach. If the estimated exposure is counted as the interval between two adjacent prescriptions, it takes the retrospective approach and violates the assumption of survival analysis. Another simplified method for approximating the cumulative dosage assumes that each exposure in realistic is quite short and all the insulin is run out in its beginning. Compared with the approximation of TDALP, the method, indicated by green lines, is even worse. For example, at time T_a, the error between the real case and the 'run out in the head' is much bigger than the one estimated by TDALP. Thus, in this project, the cumulative dosage estimated from TDALP is adopted as the best we currently have, even through it is not a prefect solution yet.

The behavior of `default-1-DDD` on the daily dosage estimation is weak. The result is not surprised, since the action to assign daily dosage as 1 DDD is not flexible, and is only able to correctly match the cases with 1 DDD insulin daily. However in real cases different patients have different CIR, ISF and carbohydrate intake. In some extreme cases, the difference of insulin daily dosage between different individuals can exceed 100 times. The parameter `prev.time` of the `TDALP`, which defines the size of historical records for learning process, is important to the estimation performance. When dealing with the constant daily dosage case, the op-

timal `prev.time` is equal to Inf. That is reasonable because taking more historical information is helpful to diminish the noise in the exposure time. In contrast, the optimal `prev.time` is 1 in the case with the alterable daily dosage, and the reason is that too much consideration on historical information would weaken the daily dosage estimation accuracy after daily dosages increased or reduced. For example, if the daily dosages of the first three prescriptions have been assigned with 1 DDD, then in order to overcome the reduced insulin sensitivity syndrome, the daily dosage of last two ones were assigned with 2 DDD (fig.30). The daily dosage estimation for the $4th$ prescription always approximates to 1 DDD no matter 1 or Inf was assigned to `prev.time`, however different `prev.time` parameters will impact the accuracy of the $5th$ prescription's daily dosage estimation. When the `prev.time` is equal to 1, the daily dosage of the $5th$ prescription will be calculated by amt_4 divided by T_4 which approximates to 2 DDD; but when `prev.time` is equal to Inf, it will be calculated by the sum of amt_1, amt_2, amt_3, amt_4 divided by the sum of T_1, T_2, T_3, T_4, which deviates from 2 DDD but approaches to 1 DDD. Although another parameter, `max.interval`, was not considered in this project, it will be useful to deal with long prescription interval cases in the insulin therapy. For instance, one diabetes patients moved aboard for some reasons and moved back 1 year later. In the case, if TDALP does not make a breaking point during the period, the daily dosage of the new prescription after homecoming will be much smaller than the expected one.

Although the `average-dpt` is a retrospective method and violates one fundamental assumption in survival analysis, it has a good behavior and even wins the `TDALP` narrowly on the daily dosage estimation when the daily dosage is kept constant during the follow-up time. But the insulin daily dosage is highly influenced by patients' health conditions, carbohydrate intake and other factors, thus it seldomly keeps constant in real cases. When the daily dosage is alterable, the `average-dpt`'s performance deteriorates and cannot compete with the `TDALP` (fig.21 A).

Considering the importance of the cumulative dosage in pharmacoepidemiology, the estimation accuracies of different algorithms on the cumulative dosage were evaluated. To achieve the purpose, the relative ratios between the estimated cumulative dosage from different algorithms and the expected one were drawn into kernel density estimation. As described above, the `default-1-DDD` was out of the competition first. Although the `average-dpt` is a retrospective approach, it showed an excellent performance in the case with the constant daily dosage (fig.22 B). However the behavior of the `average-dpt` crashed down in the case with the alterable daily dosage: the estimated cumulative dosage severely deviated from the expected one in the terminal part of the follow-up time(fig.23 B). Even through the bias is not so fatal for the study of insulin, since the main disease related to the insulin therapy, coma, is weakly related to the cumulative dosage, it will be a disaster for other studies, such as the relationship between post-menopausal hormone replacement therapy (HRT) and breast cancer. In 2002, *Chi-Ling Chen, et.al* have indicated that the long-term use of HRT was associated with the increased risk of lobular and nonlobular cancer [10]. For the `TDALP`, it still has some distances away from "perfect" daily dosage estimation algorithm in fact. The most significant problem of the `TDALP` is its default treatment on the first prescription record which is the main source of biases

in the daily dosage and cumulative dosage estimations. At present, two approaches are selectable to diminish the bias of TDALP. One is to simulate physicians' action. In real cases, the common sense is that physician will give some recommendations about the insulin daily dosage to diabetes patients who were the first time to receive the insulin therapy. The recommended daily dosage should be calculated by a formula whose parameters may include patient's baseline blood glucose level, carbohydrate intake, weight, gender, age and etc. However the most tough obstacle of the solution is patients' blood glucose level, carbohydrate intake and weight were not recorded in the prescription register. Therefore the approach is hard to reach though sound beautiful. The other approach is a speculative strategy ingratiating the result. In the speculative strategy, the daily dosage of the first prescription record will be set as the average daily dosage of the initial part, such as one third, of the follow-up time or be set as the amount of insulin of the first prescription divided by the interval between the first and the second prescription. The flaw of the latter approach is that it is a retrospective method and is still possible to make bias when there is only one prescription record for one patient in which, for example, the patient was prescribed one and half year insulin, but the censoring happened and the observation ended in one year.

Cox proportional-hazards regression model is a powerful statistical tool to analyze the relative hazard between different categorized groups. Although it is possible to easily calculate each group's hazard by the division of events number and the related exposure time, the simple method cannot replace the Cox model to work out the coefficient of one continuous covariate such as the impacts of cumulative time and cumulative dosage on the occurrence of hospitalization events. However, the Cox model is not appropriate for the algorithm screening in this project. The first reason is the heavy computation. In order to keep the high quality of the simulated data, each simulation included the data of 5000 patients during at least one year (365 days) and the baseline hazard was set as 0.01. Then a heavy job is needed to merge and reshape the prescription records, hospitalization event records and personal information into a new data frame which has a readable structure for the Cox model. The process for each combination of algorithms and extension rates will spend approximately 5 minutes in the server of Aalto University (HP BL460c 2 $\times$ X5650 2.67GHz 96 GB RAM). Secondly, due to the lack of the cumulative time and dosage's expected coefficients, there is no benefit to spend time to run the process specifically designed for the Cox model. The third reason is that one of the Cox model's advantages is to deal with the case with the baseline hazard changing with time or age. But in this project, the simulation of hospitalization events has been simplified to have a constant baseline hazard from the beginning to the end.

Albeit listing disadvantages of Cox model in the algorithm evaluation, some attempts have still been made in this project. After applying the Cox model for different combinations of daily dosage estimation algorithms and extension rates in the gap filling, none of difference has been detected. No matter which algorithm, how large the extension rate, or whether daily dosages are alterable, the relative hazards generated from the Cox model are close to the expected value and share the similar distribution patterns, thus it is impossible to distinguish the combinations.

The inability is due to the large baseline hazard chosen in this project. The baseline hazard in real cases is around 0.0005, but it was set as 0.01 which causes the hospitalization records overstocked. The solution is to reduce the baseline hazard to the real level, 0.0005, and increase the patients' number or the exposure time. Due to the heavy computation, the solution will not be tested in this project. Besides that, the residual gap rate of `default-1-DDD` was found to be lower than the ones of `average-dpt` and `TDALP`. The illusion was caused by the method of simulating daily dosage in each prescription. When simulating the prescription records, the daily dosages were generated from a Poisson distribution and most of them were bigger than 1 DDD. Therefore, the estimated exposures time will be longer and have better behavior in the coverage.

Through testing the behaviors of different algorithms on the daily dosage and cumulative dosage estimation, the `TDALP` has been proved as an accurate one in both cases with constant and with alterable daily dosage. Compared with the `average-dpt`, the `TDALP` has better performance when the daily dosage is alterable. In addition the `TDALP` is a safer algorithm adopting prospective approach which avoids to violate the fundamental predictability assumption of survival analysis. Therefore it is a rational decision to replace the `average-dpt` with the new method in the future work.

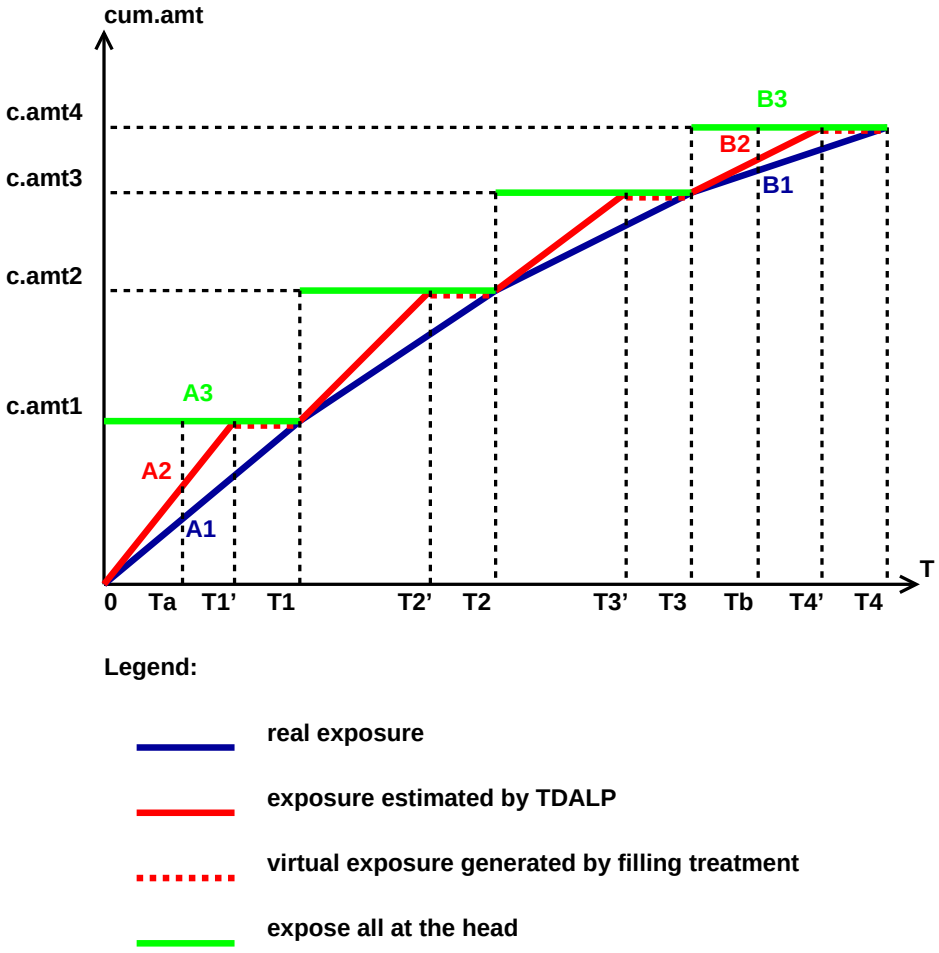


Figure 29: The calculation on cumulative dosage. In the figure, the horiontal axis is time, and the vertical axis is cumulative dosage. The blue, red and green solid lines indicate the real cumulative dosage, the estimated ones from TDALP and the 'run out in the head' solution. Additionally, the red broken lines indicate the virtual exposures for the gap filling. Four prescriptions are recorded at 0, T_1, T_2 and T_3, whose cumulative dosages are $c.amt_1$, $c.amt_2$, $c.amt_3$ and $c.amt_4$ if all of them are supposed to be run out. In the real case, there is no gap among exposures and the last exposure ended at T_4. In the estimation from TDALP, four exposures are estimated to end at T_1', T_2', T_3' and T_4' respectively. Since all the insulin is supposed to be run out at the beginning moment in the 'run out in the head' approach, its cumulative dosage time curve in each exposure keeps plateau. Finally, the $\{A_1, A_2, A_3\}$ and $\{B_1, B_2, B_3\}$ are related points of real case, estimation based on TDALP, 'run out in the head' method at time points T_a and T_b individually.

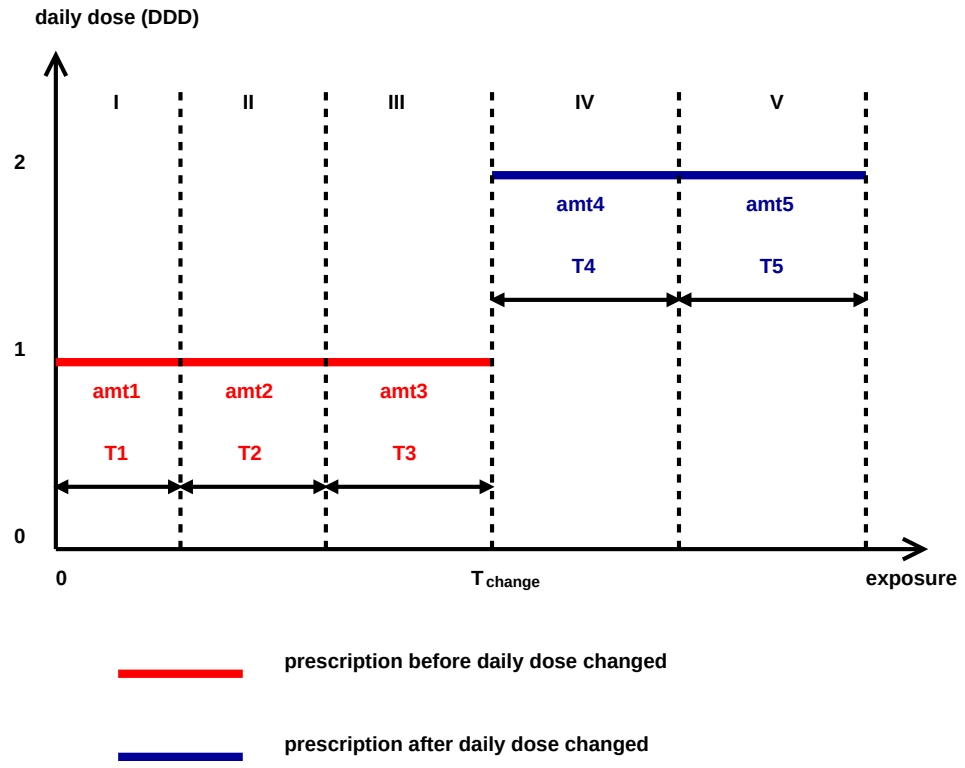


Figure 30: Increaing the insulin daily dosage when reduced insulin sensitivity syndrome happened on one patient. In the example, there are five purchase records: the first three purchase prescriptions are given 1 DDD as daily dosage. Then reduced insulin sensitivity syndrome occurs at T_{change}, thus the doctor doubles the daily dosage to 2 DDD in the new prescriptions: the last two purchases. In the figure, the amt_i indicates the ith purchase's insulin amount, and T_i indicates the exposure time for ith purchase.

References

[1] http://www.whocc.no/ddd/definition_and_general_considera/.

[2] http://www.cscu.cornell.edu/news/statnews/stnews78.pdf.

[3] http://www.stat.tamu.edu/~suhasini/teaching613/profile_likelihood.pdf.

[4] http://www.bd.com/us/diabetes/download/insulin_adjustment_workbook_complete.pdf.

[5] http://www.bd.com/us/diabetes/page.aspx?cat=7001&id=7303.

[6] Ulf Adamson and Erol Cerasi. Diminished sensitivity of the insulin response to glucose following growth hormone infusion in man. *Acta endocrinologica*, 81(4):735–742, 1976.

[7] A. Artola, A. Kamal, GMJ Ramakers, F. Gardoni, M. Di Luca, G.J. Biessels, F. Cattabeni, and WH Gispen. Synaptic plasticity in the diabetic brain: advanced aging? *Progress in brain research*, 138:305–314, 2002.

[8] Monika Bähr, Thomas Kolter, Gerhard Seipke, Jürgen Eckel, et al. Growth promoting and metabolic activity of the human insulin analogue [glya21, argb31, argb32] insulin (hoe 901) in muscle cells. *European journal of pharmacology*, 320(2-3):259, 1997.

[9] U Bergman, P Elmes, M Halse, T Halvorsen, H HoOD, PKM Lunde, and F SJoQVIsr. Wade, 0. l. and westerholm, b.(1975). the measurement of drug consumption. drugs for diabetes in northernireland, norway and sweden. *European J. clin. Pharmacol*, 8:83.

[10] Chi-Ling Chen, Noel S Weiss, Polly Newcomb, William Barlow, and Emily White. Hormone replacement therapy in relation to breast cancer. *JAMA: the journal of the American Medical Association*, 287(6):734–741, 2002.

[11] David Clayton, Michael Hills, and A Pickles. *Statistical models in epidemiology*, volume 41. IEA, 1993.

[12] Philip Cole. The evolving case-control study. *Journal of Chronic Diseases*, 32(1-2):15, 1979.

[13] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[14] Torsten Deckert. Intermediate-acting insulin preparations: Nph and lente. *Diabetes Care*, 3(5):623–626, 1980.

[15] John Fox. Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, pages 1–18, 2002.

[16] A Fry. Insulin delivery device technology 2012: where are we after 90 years? *Journal of diabetes science and technology*, 6(4):947, 2012.

[17] Kari Furu, Björn Wettermark, Morten Andersen, Jaana E Martikainen, Anna Birna Almarsdottir, and Henrik Toft Sørensen. The nordic countries as a cohort for pharmacoepidemiological research. *Basic & clinical pharmacology & toxicology*, 106(2):86–94, 2010.

[18] Peter S Gillies, David P Figgitt, Harriet M Lamb, et al. Insulin glargine. *Drugs*, 59(2):253–260, 2000.

[19] John R Guyton. Niacin in cardiovascular prevention: mechanisms, efficacy, and safety. *Current opinion in lipidology*, 18(4):415–420, 2007.

[20] T. Haak, A. Tiengo, E. Draeger, M. Suntum, and W. Waldhäusl. Lower within-subject variability of fasting blood glucose and reduced weight gain with insulin detemir compared to nph insulin in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism*, 7(1):56–64, 2004.

[21] J Hallas and A Nissen. Individualized drug utilization statistics. *European journal of clinical pharmacology*, 47(4):367–372, 1994.

[22] Frank E Harre, Kerry L Lee, and Barbara G Pollock. Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80(15):1198–1202, 1988.

[23] Jari Haukka. Assessment of association between severe hypoglycaemia and use of detemir, glargine and nph insulins (er11-9417/ u1111-1120-7164). nationwide register-based study, EPID Research Oy, April 2013. `http://www.encepp.eu/encepp/viewResource.htm?id=3815`.

[24] Charles H Hennekens, Julie E Buring, and Sherry L Mayrent. *Epidemiology in medicine*, volume 515. Little Brown & Company, 1987.

[25] Timo Klaukka. The finnish database on drug utilisation. *Norsk epidemiologi*, 11(1), 2001.

[26] R. Klein et al. Hyperglycemia and microvascular and macrovascular disease in diabetes. *Diabetes care*, 18(2):258, 1995.

[27] Dinesh Kumar, Charles M Alexander, Adina Zeidler, James J Rhodes, Christopher C Boarman, and Michael T Hoopes. Immunoreactivity of human insulin of recombinant dna origin. *Diabetes*, 32(6):516–519, 1983.

[28] RD Lawrence. Interactions of fat and carbohydrate metabolism¡ªnew aspects and therapies:(section of therapeutics and pharmacology). *Proceedings of the Royal Society of Medicine*, 35(1):1, 1941.

[29] X. Luo, G.N. Lambrou, J.A. Sahel, and D. Hicks. Hypoglycemia induces general neuronal death, whereas hypoxia and glutamate transport blockade lead to selective retinal ganglion cell death in vitro. *Investigative ophthalmology & visual science*, 42(11):2695–2705, 2001.

[30] Brian MacMahon, Thomas F Pugh, et al. Epidemiology: principles and methods. *Epidemiology: principles and methods.*, 1970.

[31] M.L. Misso, K.J. Egberts, M. Page, D. O Connor, and J. Shaw. Continuous subcutaneous insulin infusion (csii) versus multiple insulin injections for type 1 diabetes mellitus. *Cochrane Database Syst Rev*, 1, 2010.

[32] L.H. Nielsen, E. Løkkegaard, A.H. Andreasen, and N. Keiding. Using prescription registries to define continuous drug use: how to fill gaps between prescriptions. *Pharmacoepidemiology and drug safety*, 17(4):384–388, 2008.

[33] Pia Pajunen, Heli Koukkunen, Matti Ketonen, Tapani Jerkkola, Pirjo Immonen-Räihä, Päivi Kärjä-Koskenkari, Markku Mähönen, Matti Niemelä, Kari Kuulasmaa, Pertti Palomäki, et al. The validity of the finnish hospital discharge register and causes of death register data on coronary heart disease. *European Journal of Cardiovascular Prevention & Rehabilitation*, 12(2):132–137, 2005.

[34] RJ Prescott, CE Counsell, William J Gillespie, Adrian M Grant, Ian T Russell, S Kiauka, IR Colthart, Susan Ross, SM Shepherd, Daphne Russell, et al. Factors that limit the quality, number and progress of randomised controlled trials, 1999.

[35] M.C. Riddle, J. Rosenstock, and J. Gerich. The treat-to-target trial randomized addition of glargine or human nph insulin to oral therapy of type 2 diabetic patients. *Diabetes care*, 26(11):3080–3086, 2003.

[36] C Ronald Kahn. Insulin resistance, insulin insensitivity, and insulin unresponsiveness: a necessary distinction. *Metabolism*, 27(12):1893–1902, 1978.

[37] Guntram Schernthaner. Affinity of igg-insulin antibodies to human (recombinant dna) insulin and porcine insulin in insulin-treated diabetic individuals with and without insulin resistance. *Diabetes Care*, 5(Supplement 2):114–118, 1982.

[38] Brian L Strom. *Pharmacoepidemiology*. Wiley, 2006.

[39] R.A. Whitmer, A.J. Karter, K. Yaffe, C.P. Quesenberry Jr, and J.V. Selby. Hypoglycemic episodes and risk of dementia in older patients with type 2 diabetes mellitus. *JAMA: the journal of the American Medical Association*, 301(15):1565–1572, 2009.

[40] Jean L Whittingham, Svend Havelund, and Ib Jonassen. Crystal structure of a prolonged-acting insulin with albumin-binding properties. *Biochemistry*, 36(10):2826–2831, 1997.

[41] P.J. Wiffen. Enhanced glucose control for preventing and treating diabetic neuropathy. *Journal of pain & palliative care pharmacotherapy*, 26(4):380, 2012.

[42] R.J. Wright and B.M. Frier. Vascular disease and diabetes: is hypoglycaemia an aggravating factor? *Diabetes/metabolism research and reviews*, 24(5):353–363, 2008.

Appendix

The input arguments and the output variables of functions programmed especially for this thesis project are listed in this appendix.

- `simulation.purchase`

  Input:

  - `nr.purchaser`: the number of purchaser.
  - `observation.days`: the observation duration of this study.
  - `increase.mode`: the parameter to determine whether daily dosages are alterable.
    If `increase.mode` is equal to `TRUE`, the daily dosages can be multiplied in each regular intervals.
    If `increase.mode` is equal to `FALSE`, the daily dosages keep constant.
  - `regular.interval`: the regular interval for daily dosage multiplication. When the cumulative exposure time of one kind of insulin with one daily dosage exceeds this value, a multiplication action will happen.
  - `increase.ratio`: the rate of the multiplication of daily dosage.
  - `trans.mode`: the parameter to determine whether the insulin transition happened.
    If `trans.mode` is equal to `TRUE`, the insulin transition is allowed to happen.
    If `trans.mode` is equal to `FALSE`, the insulin transition cannot happen therefore only one insulin can be prescribed to patients in this study.

  Output: A data frame for purchase records

  - `id`: personal id.
  - `ATC`: the ATC code of insulin or its analogues.
  - `dop`: the date of purchase.
  - `doe`: the date of end of insulin exposure.
  - `amt`: the amount of dosage in one purchase.
  - `dpt`: the daily dosage.
  - `expected.diff`: the duration of expected exposure in the prescription.
  - `diff`: the duration of real happened insulin exposure.

- `generate.personal.information`

  Input:

  - `purchase`: the data frame of purchase records.

  Output: A data frame of personal information

  - `id`: personal id.
  - `gender`: the gender information for individuals.
  - `age`: the age when patients begun to use the insulin or its analogues.

- `simulation.hospital`

  Input:

  - `purchase`: the data frame for purchase records.
  - `personal.information`: the data frame for personal information.
  - `corner`: the hospital event hazard of the reference group.

  Output: A data frame of hospital events.

  - `id`: personal id.
  - `start.date`: the dates of the occurrence of hospitalization event.
  - `end.date`: the dates of the end of hospitalization event.

- `TDALP`

  Input:

  - `purchase`: the data frame of prescription records, which contains `'id'`, `'ATC'`, `'amt'` and `'dop'` information.
  - `prev.time`: how many previous purchases are needed for TDALP in one cluster.
  - `max.interval`: the maximum interval allowed between two adjacent purchase dates in one cluster.
  - `initial.mode`: the parameter to control which method is used for the estimation of the first purchase's daily dosage in the non-first cluster.
    If initial.mode = 1, this daily dosage is equal to the one of the last purchase in the previous cluster;
    If initial.mode = 2, it is equal to 1 DDD.

  Output: the data frame of prescription records with estimated daily dosage.

  - `id`: personal id (the same as the one in `"purchase"`).

- `ATC`: ATC code (the same as the one in `"purchase"`).
- `dop`: the date of purchase (the same as the one in `"purchase"`).
- `amt`: the amount of insulin dosage in one prescription (the same as the one in `"purchase"`).
- `dpt`: the estimated daily dosage for each prescription (different from the one in `"purchase"`).

- `use.amt.dpt`

  Input:

  - `purchase`: the data frame processed by the function `"TDALP"`.
  - `push.max`: the maximal allowed delay of one insulin from old prescription after new one given.
  - `start.mode`: how to deal with the overlap of exposure between two prescription with the same insulin.
    If start.mode = 1, discard the insulin from the old prescription and start to use the new one immediately;
    If start.mode = 2, the insulin from the old prescription is kept to use within `'push.max'`.
  - `observ.start`: the beginning date of observation (study).
  - `observ.dur`: the duration of observation (study).
  - `history.mode`: how to treat the prescriptions given before the observation (study).
    If history.mode = 1, discard these records.
    If history.mode = 2, keep the parts of these records overlapping with the observation, and discard non-overlapping parts.
  - `remove.mode`: how to treat the overlaps among different long acting insulins.
    If remove.mode = 1, retain these overlaps.
    If remove.mode = 2, discard the old type insulin and start to use the new one immediately.
    If remove.mode = 3, keep using the old type insulin and discard the new one in the overlapping time bands.
  - `extension`: the ratio of extension for gap filling.
  - `fill.gap.mode`: whether the gaps need filling treatment.
    If fill.gap.mode = T, fill them;
    If fill.gap.mode = F, do not do that.

  Output: the data frame of periodized result based on prescription records.

- `id`: personal id.
- `start.date`: the beginning date of time band.
- `end.date`: the end date of time band.
- `dur`: the duration of time band.
- `insulin.*`: binary indicator recording insulin usage in each time band (`'*'` indicates the abbreviation of insulin's name).
- `insulin.*.ori`: the original records of `'insulin.*'` before eliminating the overlaps among different long acting insulins.
- `*.dpt`: the daily dosage of `'insulin.*'` in each time band.
- `*.dpt.ori`: the original records of `'*.dpt'` before eliminating the overlaps among different long acting insulins.
- `*.cum.time`: the cumulative time of `'insulin.*'` until the end of each time band.
- `*.cum.amt`: the cumulative dosage of `'insulin.*'` until the end of each time band.

- `purchase.hospital.fusion`

  Input:

  - `purchase`: the data frame of periodized information processed by the function `"use.amt.dpt"`.
  - `hospital`: the data frame of hospitalization records.
  - `study.start`: the beginning date of observation (study).
  - `study.end`: the end date of observation (study).
  - `history.mode`: how to deal with hospitalization events partially overlaying with the observation.
    If history.mode = T, discard the non-overlapping parts and keep the overlapping ones;
    If history.mode = F, discard these events directly.

  Output: the data frame periodized based on the prescription and hospitalization event records. Compared with the output of function `"use.amt.dpt"`, this data frame has two new columns:

  - `to.hospital`: binary indicator of the occurrence of hospitalization event.
    If to.hospital = 1, hospitalization event occurred in the beginning of corresponding time band;
    If to.hospital = 0, do not occur.

  - `in.hospital`: binary indicator of the status of staying in hospital.
    If in.hospital = 1, the patient stayed in hospital during corresponding time band;
    If in.hospital = 0, negative.

- `personal.information.pur.hosp.fusion`

  Input:

  - `personal.information`: the personal information including gender and age information of each individual.
  - `pur.hosp.result`: the periodized data frame processed by the function `"purchase.hospital.fusion"`.

  Output: the periodized data frame which has been integrated with personal information. Compared with the output of function `"purchase.hospital.fusion"`, this data frame was added two new columns:

  - `gender`: `'M'` indicates male, and `'F'` indicates female.
  - `age`: the age when each individual started to accept insulin therapy.