Harm Reduction of Switching from Metformin Plus Sulfonylureas to Metformin Plus DPP4s in Older Adults: A Target Trial Emulation using German Routine Claims Data

Paula Starke
Chair of Statistics, University of Göttingen

December 18, 2023

Abstract

Design, setting and participants: The average risk for adverse events in patients over 65 years old who currently take metformin as first-line treatment is compared between those who initiated dipeptidyl peptidase-4 inhibitors (DPP4) and those who initiated sulfonylurea (SU) as add-on treatment. Initiations between 2011 and 2018 were analysed using routine claims data from a German health insurance provider (Barmer). Generalized linear models with overlap weighting were used to estimate the average treatment effects in the overlap population.
Treatment: metformin + DPP4 versus metformin + SU
Outcomes and measures: Rates of combined all-cause hospitalisations and outpatient visits compared via rate ratio as primary outcome and the odds for at least one event of severe hypoglycemia within one year, death within one year and at least one all-cause hospitalisation within 30 days as secondary outcomes.
Subgroups: new users, age>80, with severe hypoglycemia, with heart failure, with severe renal insufficiency
Results: Among the 171,318 eligible patients, 111.865 received DPP4 and 59,453 were in the control group receiving SU. Patients treated with DPP4 had a higher prevalence of all selected comorbidities and were more often naive to both treatments of interest (53% vs 21%). After applying overlap weighting, baseline characteristics including confounding variables for treatment history were well balanced between groups. In the main analysis, patients treated with DPP4 had a higher rate of combined all-cause hospitalisations and outpatient visits compared with those not treated with DPP4 (rate ration, 1.03; 95% CI, 1.02-1.03).

Contents

List of Tables

List of Figures

Acronyms

ATE average treatment effect

ATT average treatment effect on the treated

ATU average treatment effect on the untreated

DDD daily defined doses

DMP disease management program

DPP4 dipeptidyl peptidase-4 inhibitors

eGFR estimated glomerular filtration rate

ICD-10-GM German modification of the international classification of diseases 10th revision

OW overlap weights

PIMs potentially inappropriate medications

RCT randomized controlled trial

SU sulfonylurea

SUTVA stable unit treatment value assumption

Acknowledgements

I would like to express a huge thank you to Prof. Dr. Tim Mathes for the excellent and close supervision, as well as for the numerous engaging discussions. Special thanks also go to Dr. Thomas Grobe for the extensive guidance in handling routine data and to Prof. Dr. Petra Thürmann for the immensely important professional input.

I deeply appreciate the generosity of Barmer health insurance for giving me the opportunity to utilize their data for this analysis. I would also like to thank the entire team at the aQua-Institute. It was a pleasure to work with you during this last year.

Last but not least, I want to acknowledge Prof. Dr. Tim Friede for the valuable feedback in the final stages, and big thanks to Leon for the countless hours dedicated to proofreading.

1 Introduction

Elderly adults often suffer from multiple chronic illnesses which often require them to take a high amount of different medications but also lead to general frailty. This is problematic as elderly patients are more susceptible to the adverse effects of pharmacotherapy than younger patients due to the aging process. Efforts are taken in different countries to assemble evidence and develop recommendations for drug therapy of elderly patients tailored to the current practice and available options in each country. The *Priscus* list is the German version and has just received an update in 2022. The publication lists medications that may be potentially inappropriate for use in patients over the age of 65 and should therefore be avoided [25]. Findings from RCTs and expert knowledge were assembled to select substances that are harmful to elderly patients. Alternative medication options are presented to promote the conversion of affected patients. There is evidence that suggests that elderly patients who take any of the potentially inappropriate medications (PIMs) instead of alternative substances have a higher risk for hospitalisations connected to adverse events [15]. However, efforts to reduce inappropriate prescription behaviour could be bolstered up by more specific knowledge about the real-world effect of deprescribing specific substances to specific groups of patients and further research is warranted to identify those PIMs that are particulary harmful in practice.

This thesis aims to explore the feasibility of producing such real-world evidence by conducting observational studies using routine claims data. In the form of a proof-of-concept study I evaluate whether the findings from the *Priscus* list on the superiority of dipeptidyl peptidase-4 inhibitors (DPP4) compared to sulfonylurea (SU) can be reproduced using an observational study. I use German routine claims data from Barmer health insurance to estimate the average effect of taking DPP4 compared to SU as a second-line therapy in addition to metformin on outcomes connected to adverse events. Furthermore, I aim to identify heterogenous treatment effects in subgroups within the population of elderly patients that might particularly benefit from taking DPP4 instead of SU.

Diabetes was chosen for this application study both because of the high prevalence of the disease in particular in the elderly population in Germany and a relatively high level of existing evidence on medication options. This makes it suitable for a proof-of-concept of the target trial framework. The choice of the contrasted treatments was then primarily guided by the *Priscus* list 2.0. With glimepirid, glibenclamid, gliquidon and gliclazid, four SU are listed [25]. Gliquidon and gliclazid already had a negligible amount of prescriptions in Germany in recent years but glimepirid and glibenclamid are still prescribed on a larger scale although use decreased in recent years (from a prevalence of about 4.3% of the total population of over 65 year olds in Germany in 2011 down to 1.11% in 2021 according to the standardized prevalence in the Barmer population). DPP4 are listed as a viable alternative medication in the *Priscus* list besides metformin, insulin and SGLT2 inhibitors. They are currently one of the most commonly prescribed substance classes in Germany, in particular among the elderly. About 3% of the over 65 year olds in Germany have received at least one prescription of sitagliptin in 2021 (according to standardized prevalence in Barmer population).

A recent systematic review on DPP4 conducted in the context of the *Priscus* list update [13] included five randomized controlled trials on the safety of DPP4 compared with SU,

both as add-on to standard care. The meta-analysis of these studies suggests that older patients might benefit from taking DPP4 compared to SU as it reduces mortality and the risk for hypoglycemia. The evidence for other outcomes like overall adverse events, risk for hospitalization, falls or pancreatitis is insufficient [13]. The expert survey concluded that DPP4 should be prefered over SU in elderly patients due to their better benefit-risk ratio [13] but the magnitude of the effect is unknown because existing evidence is too heterogeneous.

Contrary to these randomised controlled trials this analysis of claims data in principle allows an evaluation of a real-world effect within the population of interest under real-world prescribing and treatment practices. The absence of a study context in principle increases the external validity of the study. However, only a carefully developed design ensures the applicability of results on the desired target population and several sources of bias have to be eliminated to achieve internal validity. I use the concept of „target trial emulation“ that was first introduced by Hernán and Robins [18] to support the design of well-defined observational studies with high internal and external validity. Besides using the concept as a structure to follow, I also aim to gain insights into the methodological demands of answering similar questions on the basis of insurance claims data.

My hypothesis was that the average risk for adverse events in patients over 65 years old, who currently take metformin as first-line treatment, but who are in need for additional diabetes medication, can be reduced under real-world conditions, if only DPP4 are prescribed as secondline treatment instead of SU. I expected the magnitude of the benefit to be most pronounced in four high-risk subgroups: patients older then 80 years, patients with a severe renal insufficiency, patients with heart failure and patients with a recent history of severe hypoglycemia. As a primary outcome I chose the rate of combined all-cause hospitalisations and outpatient visits within one year as a proxy for the overall amount of adverse events and general state of heath. As secondary outcomes I considered the odds for death within one year, at least one event of severe hypoglycemia within one year and at least one all-cause hospitalisation within 30 days.

In all medical aspects I received comprehensive guidance from Dr. Thürmann. Apart from selecting the pair of substances to be compared, she also contributed her expert knowledge to the selection of relevant confounders and provided valuable insights on the interpretation of results. During the process of data preparation I was supported by Dr. Grobe who introduced me to the particularities of claims data as well as the Barmer data warehouse and provided assistance when I encountered coding problems.

In the following section I begin by outlining the theoretical foundations of comparative treatment analysis in observational studies and the „target trial emulation“ concept (chapter 2), followed by the practical application of the method on the DPP4-SU-study with a focus on the decisions taken choose and apply a suitable method for confounding adjustment (chapter 3). The implementation of the study in the Barmer dataset and methods used for statistical analysis are described in chapter 4 and 5. The results are presented in chapter 6. Chapter 7 discusses the strengths and limitations of the study before I sum up the gained insights in a brief conclusion.

2 Theory

2.1 Target Trial Emulation

The design of the study follows the approach of a „target trial emulation“. The concept was first introduced under that name by Hernán and Robins (2016)[18] and it offers a structured approach to estimating causal effects using observational data. The central idea is to design studies and define effects as if they were observed in a hypothetical randomized controlled trial (RCT). The hypothetical RCT is then emulated as closely as possible. For both the target trial and the emulation the following items have to be defined:

- Treatment strategies
- Eligibility criteria
- Assignment procedures
- Follow-up period
- Outcome
- Causal contrasts of interest
- Analysis plan

The final trial should be as close as possible to the „ideal“ trial that was designed first but should be reasonably supported by the available data.

Hernán and Robins (2016) map out some of the most important methodological challenges of such an emulation. Their most important insights concern the temporal order of events that define a patients inclusion or exclusion from the trial or treatment allocation. I summarize their findings as follows:

- Attention should be paid to both validity and interpretation of the data. This includes verifying the reliability of information on diagnoses and the availability of confounders in the dataset.
- Included patients have to be in database long enough to apply all inclusion criteria prior to baseline.
- Eligibility criteria must not include post-baseline information. Events that occur after the initiation of treatment are possibly affected by the treatment strategy itself, which can introduce selection bias.
- On the other hand, immortal time bias can result from a delay between cohort entry and treatment initiation, as it induces a period of follow-up during which no outcome events or deaths can occur. Ideally, treatment initiation and inclusion should align and occur at the same time.
- Eligible patients who do not start any of the specified therapies should be excluded from the analysis. The resulting estimated effect is then not valid for this excluded subpopulation. While this seems obvious, applying such an explicit procedure can help to clarify the temporal structure if inclusion occurs prior to treatment initiation.

- Loss-to-follow-up can be reduced by including patients who have been in regular contact with the health care system prior to *time zero*. This is mainly important in study contexts where the data stems from non-compulsory or employer-dependant health insurance systems where individuals often have incomplete coverage for financial reasons or frequently switch between healthcare providers.
- Only target trials without blind assignment can be emulated. Again, this is obvious, but in some cases this can be important to consider, for instance to clarify differences to extisting RCTs.

I try to implement these guidelines in my application study as comprehensively as possible.

2.2 Estimands

On several occasions I noticed that it is very helpful to complement the items of the target trial framework by explicitly defining an estimand for the question of interest, in order to clarify the argumentation when making decisions about both the design of the target trial and the emulation. Considering the influence of methodological decisions on the estimand definition can help to choose an approach that best fits the estimand of interest. For instance, different methods to adjust for confounding, e.g. different matching or weighting schemes, are connected to different target populations and population-level summary statistics. Method and estimand mutually define each other and an exact definition is important to allow an exact interpretation of the estimated effect as I will briefly outline in the next chapter 2.3. I consider this specific aspect throughout the process of the target trial definition and emulation.

Another important element of the estimand framework, besides the target population and the definition of a population-level summary, that is not explicitly addressed in the target trial framework, is the handling of intercurrent events. The „ICH-E9 R1 addendum on estimands and sensitivity analysis in clinical trials“ [8], which was published by the European Medical Agency as part of the guideline on statistical principles for clinical trials, describes four possible strategies for handling intercurrent events. Under the *treatment policy* strategy, the occurrence of the intercurrent event is considered not to be relevant for the treatment effect. Instead it is considered as part of the treatment regimen and the values for the outcome of interest are used, irrespectively of whether the intercurrent event occurs or not. Therefore, the strategy is not applicable to terminal events like death, where post-event values are unavailable. The *hypothetical* strategy envisions a scenario where the intercurrent event does not occur and defines the value for the outcome of interest under those hypothetical conditions. The goal is to understand the potential impact of various hypothetical situations, such as the absence of additional medication or different outcomes for subjects experiencing adverse events. Under the *composite* strategy, an intercurrent event is considered to be informative about the patient's outcome and is therefore added to the definition of the outcome to form a composite outcome variable. When all events or the value of the outcome prior to the intercurrent event is of interest, a *while-on-treatment* strategy is chosen.

Luijken et al. (2023) [24] recently applied the „ICH-E9 R1 addendum on estimands“ to observational pharmacoepidemiologic comparative effectiveness and safety studies. The publication describes three case studies for different study scenarios and outlines possible

estimands. The first scenario of the second case study corresponds to an intention-to-treat effect of a sustained treatment and is similar to the case of the DPP4-SU study that I want to design. The intercurrent events *discontinuation of treatment*, *switching to an alternative treatment* and *switch to intermediate-acting-insulin* are handled under the *treatment policy* strategy and a *while-on-treatment* strategy is applied to *death*. I proceed similarily in chapter 3.1.

2.3 Causal effect estimation in observational studies

Just as the definition of the estimand has to be adjusted to the observational setting, the subsequent estimation of a causal effect is not as straightforward in an observational study as it is in an RCT. In this chapter, I briefly summarize the general procedure of estimating causal effects and introduce possible approaches for selecting and controlling for confounders.

Causal effect estimation generally aims to estimate an average treatment effect (ATE) as the average individual effect in a specific population. Knowing the exact individual effect of a treatment $Y_i(1) - Y_i(0)$ for an individual would be ideal, but as only the outcome of the treatment that was actually received is known, such an estimation is impossible in practice. Instead, the ATE is defined as the expectation $E(Y_1 - Y_0)$ across all members of the population, where Y_0 and Y_1 denote the potential outcomes in the absence and presence of treatment respectively. In an RCT under random treatment allocation, $E(Y_1 - Y_0) = Y_1 - Y_0$ is an unbiased estimator of the ATE.

In an observational setting however, where the characteristics of patients in the treatment group might differ from subjects in the control group, an ATE can only be estimated in a population if the treatment assignment is strongly ignorable. According to Greifer et al. (2023) [17], this assumption entails three conditions. The first condition, known as „conditional exchangeability“, demands the inclusion of all relevant confounders in the analysis, ensuring that no unmeasured confounding influences the treatment-outcome relationship. Confounding here references to factors that are correlated with both the independent variable (treatment) and the dependent variable (outcome), potentially leading to bias. The second condition, „positivity“, dictates that all patients should have a non-zero probability of receiving either treatment. The third condition, the stable unit treatment value assumption (SUTVA), demands that the treatment status of one patient does not impact another patient's outcomes and no unmeasured versions of treatment exist. The SUTVA assumption has to be accounted for in the study design. The assumption of „positivity“ does not necessarily hold for the total population of interest. If there is limited overlap between the compared groups, the target population has to be restricted to a population in which the assumption holds. The first assumption of „conditional exchangeability“ requires the identification of possible confounders and their incorporation into the treatment effect estimation by using suitable statistical methods.

There exist several approaches for selecting the set of confounding variables. In practice, it is usually not possible to observe or even unambiguously identify all confounders, but the validity of the observational analysis depends on how well confounding bias can be minimized. Different approaches on how restrictive the set of confounders should be are discussed in the scientific community. Including a large set of possible confounders bears the risk of controlling for colliders, which introduces additional bias to the analysis [16].

On the other hand, omitting important confounders also introduces bias. Among others, three different sets of covariates are frequently used as possible confounders in the literature [32]: firstly, all covariates that can be determined prior to exposure, secondly, all common causes of the outcome and the exposure, or thirdly, all pre-exposure covariates that are a cause of the exposure, the outcome or both. The last approach has several advantages compared to the other two sets, as it in principle includes all relevant covariates and thus always sufficiently controls for confounding if the true set of confounders is in fact observed [32]. However, depending on the availability of variables in the data and existing knowledge of the causal mechanisms, there often remains some unobserved confounding. In such cases, unnecessarily controlling for causes of the exposure that are not related to the outcome other than through the exposure, so-called instruments, can reinforce the bias introduced by the unmeasured confounders [32]. Taking this additional aspect into account, Vanderweele et al. (2019)[32] proposed the following, practically applicable strategy for confounder selection:

1. Control for each covariate that is a cause of the exposure (choice of treatment), of the outcome, or of both

2. Exclude from this set any variable known to be an instrumental variable (effect on exposure but not outcome)

Adjustment for the selected confounders can be conducted using three general approaches: outcome regression, instrumental variable analysis or propensity score methods. Outcome regression is widely used, but has disadvantages with respect to interpretation of the estimated effect. Targeting specific estimands is difficult and can only be validly achieved by using additional methods, notably g-computation [17]. Instrumental variable analysis targets the average effect in the population of patients who comply with the treatment recommended to them, which is not often the population of interest in medical studies. In the following chapter I will therefore only consider propensity score methods in more detail.

2.4 Propensity score methods for confounding adjustment

Propensity scores are estimated by modeling the probability of treatment assignment based on observed covariates. The covariate distributions are then balanced between the groups by using either one of two general approaches: matching and weighting. Both approaches aim to achieve covariate balance between the treatment and control group. Matching on the propensity score involves pairing treated and control units with similar propensity scores. The resulting population often has well-balanced covariates. However, the high precision comes at the cost of potential sample size reduction, as not all treated units may find suitable matches.

Weighting on the other hand assigns different weights to units based on functions of their propensity scores. Zhou et al. (2020)[37] define different weighting methods in terms of a tilting function $h(x)$ that transforms the distribution of covariates in the sampled population to a target distribution of interest: If $f(x)$ is the sampled distribution, then the target distribution is defined as $f(x)h(x)$. With $f_z(x) = Pr(X = x|Z = z) = e(x)$ as the density of covariates in group z, balancing weights $w(x)$ can create a balanced covariate distribution between group 1 and 0 when $f_1(x)w_1(x) = f_0(x)w_0(x) = f(x)h(x)$. As the propensity score $e(x)$ is defined as $Pr(Z = 1|X)$, a general definition of a weight given a tilting function $h(x)$

and a treatment indicator Z_i for observation i can be written as:

$$w_i(x) = Z_i \frac{h(x)}{e_i(x)} + (1 - Z_i) \frac{h(x)}{(1 - e_i)} \tag{1}$$

The weighted average treatment effect over all considered observations i is then equal to

$$\Delta_h = \frac{E[h(X)(Y(1) - Y(0))]}{E(h(X))} \tag{2}$$

When the treatment effect is constant across the entire population, the weighted average treatment effect is the same for all tilting functions $h(x)$. However, under heterogenous treatment effects the choice of tilting function has to match the target population which is either of interest or statistically optimal.

The simplest case is a tilting function $h(x) = 1$ which corresponds to the ATE in the sampled population without any re-distribution of the population. The corresponding very established weighting scheme is called „Inverse Probability of Treatment weighting“ (IPW) where the weights correspond to the inverse probability of treatment:

$$w_i(x) = Z_i \frac{1}{e_i(x)} + (1 - Z_i) \frac{1}{(1 - e_i)} \tag{3}$$

Alternative, well established tilting functions are $h(x) = e(x)$ and $h(x) = 1 - e(x)$ which correspond to the average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATU) respectively. However, all three functions have limitations when the positivity assumption is violated. When there is a large proportion of extreme propensity scores equal or close to 0 or 1, these individuals get assigned large weights as the weights are calculated by dividing by the propensity score. Extremely large weights overemphasize the influence of these observations on the estimation of the treatment effect. As Zhou et al. (2020)[37] deduced in detail, the asymptotic variance of the IPW estimator gets very large in these cases. Furthermore, if extreme weights occur in combination with heterogenous treatment effects and a misspecified propensity score model, the bias caused by the misspecification is reinforced. A common solution is to trim the PS-distribution at the edges to exclude extreme propensity scores. This can however substantially reduce the sample size [21].

A promising method that has not yet been studied and applied as extensively are balancing weights that target the population of patients with clinical equipoise, namely overlap weights (OW), matching weights and entropy weights. The targeted population of patients with clinical equipoise relates to the population of patients for whom both treatment options are currently considered in practice but where the treatment decision is less certain. Each patient's overlap weight is the probability of that patient being assigned to the opposite group [21]:

$$w_i = Z_i(1 - e_i) + (1 - Z_i)e_i \tag{4}$$

Unlike inverse probability of treatment (IPT) weights, these weights do not involve the inversion of probabilities and can only take values between 0 and 1. They are designed to tilt $f(x)$ towards a propensity score of 0.5 [37]. The three methods differ only in how sharply the tails of $f(x)$ are weighted down. As was deduced mathematically by Zhou

et al. (2020)[37], OW, matching weights and entropy weights clearly outperform IPW weighting asymptotically in settings with propensity scores $e(X) \approx 0$ or 1. Such extreme observations get assigned weights near 0 and therefore do not increase the variance or add much additional bias in case of propensity score misspecification. Compared to IPW with trimming, OW lead to lower bias, as extreme propensity scores are not omitted and the need for an arbitrary data-driven choice of cutoffs and mode of trimming is avoided. Simulations conducted by Zhou et al. (2020)[37] confirmed these asymptotic properties. While all three analysed balancing weights outperformed IPW, overlap weighting provided the most efficient estimate. OW were also more accurate and efficient in simulation studies under propensity score misspecification, unless a very important confounder was omitted which none of the methods can handle [37],[28]. Furthermore, as the target population is not defined a priori, methods targeting the overlap population can also be useful for estimating a precise and robust estimand in settings where no specific hypothesis on the characteristics of the population of interest exists. An example might be a study that aims to discover whether at least some patients benefit from a treatment while the treatment is clearly not appropriate for other subgroups [17].

Once sufficient covariate balance between the groups has been achieved through weighting, the weighted observations are used to estimate the average treatment effect. A straightforward approach is to calculate the difference in the weighted means of the outcome variable between the treatment and control groups. For effect measures other than the difference in means, a weighted generalized regression model $g(\mu_i) = \beta_0 + \beta_1 T_i$ with the treatment indicator T_i as the independent variable and a link function $g()$ that is suitable to the outcome type is chosen as suggested by [29]. In principle, any parametric or non-parametric model can be used. In the presence of non-linear relations and complex interactions, non-parametric models can offer more flexibility. To account for the violation of the homoskedasticity assumption of the maximum likelihood estimation caused by the use of weights, robust standard errors should be used. Another more complicated but potentially superior method is to use bootstrapping if sufficient computational resources and time are available.

If there is a risk for a severely misspecified propensity score model, substantial unobserved confounding or if the covariate balance after propensity score weighting is still insufficient, a doubly robust method can improve the validity of the causal analysis. The idea behind doubly robust estimates is to specify both a propensity score model and an outcome model. As long as one of the two models is specified correctly, the causal effect estimation is valid. One doubly robust approach that is often recommended [12] is to use targeted maximum likelihood estimation and conduct a g-estimation as an additional step after fitting the outcome model. The outcome model is defined as the 'Q-model' which is used to predict the counterfactual outcomes under each of the two treatments alternatives. The resulting „full“ dataset is then used to calculate the causal estimate as the contrast of the two average estimated potential outcomes. G-estimation also the advantages that it offers flexibility to analyse and compare different types of outcomes and outcome models using an identical workflow where only the outcome model has to be adjusted each time.

3 Application of the concept of target trial emulation

In the following chapter I apply the comprehensive process of a „target trial emulation“ to the DPP4-SU study case. The five essential elements of the target trial defined by Hernán and Robins [18] are specified in chapter 3.1, followed by a description of the decisions taken to emulate the target trial (3.2). As I conduct an observational analysis and, more specifically, use routine claims data that were not created for research purposes, I can not replicate the target trial exactly as it was specified. The last subchapter outlines the process of confounder selection (3.3. The target trial that I designed in the very first step was the basis for a study protocol that I got approved by the responsible ethics commission in July 2023 (Appendix 9).

3.1 Definition of the target trial and estimand

I defined a target trial by following the recommendations given by Hernán and Robins (2016)[18]. They name seven essential elements that need to be defined: the definition of eligibility criteria, treatment strategies, assignment procedures, the follow-up period, the outcome, the causal contrast of interest and an analysis plan. Table 1 defines these elements for our study. Table 2 supplements four additional elements that describe the estimand in more detail. The structure and definitions correspond to considerations outlined by Luijken et al. (2023) [24], who applied the „ICH-E9 (Statistical Principles for Clinical Trials) R1 addendum on estimands“ to observational pharmacoepidemiologic comparative effectiveness and safety studies.

Prevalent users of both DPP4 and SU are included, as the study targets elderly patients who have usually already been suffering from diabetes type 2 for many years and thus rarely start an entirely new treatment. Existing literature on „target trial emulation“ encourages the exclusion of prevalent users to reduce potential bias, but in order to answer my research question it does not make sense to generally exclude these long-term users as the study population would then be limited to a unrepresentative subgroup of patients. Thus, I cannot avoid difficulties by switching to a new-user design and instead need to try to reduce bias as well as possible and be mindful of possible remaining bias. I also decided to control for covariates that might be influenced by prior treatment, more specifically I included the variable *severe hypoglycemia* as a confounder in all analyses except for the outcome *risk for severe hypoglycemia within one year*. The latter is likely to be more severely impacted by potential collider bias. I briefly take up this question in chapter 7, but the topic warrants further consideration and discussions.

As a primary outcome I selected the composite outcome *all-cause hospitalisations and physician visits within one year*. The secondary outcomes *1-year risk of severe hypoglycemia while alive* and *1-year risk of all-cause mortality* are more specific endpoints. The risk for severe hypoglycemia is the main event that is supposed to be reduced by any diabetes medication, so this endpoint reflects the efficacy of the treatments. The mortality risk on the other hand is a more rigorous endpoint that should reflect any difference in severe adverse events between the two treatments. The occurrence of all-cause hospitalisations within 30 days is supposed to represent a potential serious short-term effect, especially in new initiators. All three secondary outcome are binary and odds ratios are used as a measure of effect.

The definition of an appropriate estimand is another consideration that is also affected by the inclusion of prevalent users. The treatment decision between DPP4 and SU is currently primarily decided due to personal preference of the physician rather than other criteria like comorbidities or patient risk factors. Thus, the ATT and ATU are not of interest here, as they strongly depend on current prescribing practice. The hypothesis is that DPP4 are superior to SU in all patients. I designed the hypothetical intervention as a system-wide *Priscus*-informed medication review that is supposed to address all eligible patients, irrespective of their prior treatment. Thus, I am in principle interested in the ATE, the effect in the total population. However, the ATE requires the strictest adherence to the assumptions of „conditional exchangeability“, „positivity“, and „SUTVA“ as no restriction of the population is allowed if the effect is to be valid for the entire population. If assumptions are violated adequate covariate balance can not be achieved. The „positivity“ assumption in particular might be a critical issue in our case. As I decided to include prevalent users of the treatments, the treatment history will be an influential confounder. Patients and physicians have a high preference to continue the current treatment. It also makes sense from a medical perspective not to reevaluate a patients treatment at every visit as long as medical preconditions did not change drastically. However, the inclusion of the treatment history as a supposedly very influential confounder might lead to limited overlap which is a violation of the positivity assumption. I therefore chose the average treatment effect in the overlap population as a population-level summary.
Overlap weighting as a method that defines a population of clinical equipoise without requiring a prior active restriction from our part is also useful to define the real-world population of interest. The variables we chose as confounders are supposed to influence the treatment decision in the sense that they are risk factors for adverse events confirmed by expert judgment and often also randomized controlled trials and other studies. However, the extent to which each factor is considered in practice is not clear.

Most intercurrent events are handled under a *treatment policy* strategy. The only exception is *death before end of follow-up* for which I use a *while-on-treatment* strategy. For the primary outcome this means that the number of hospitalisations and physician visits is counted up to the date of death, resulting in a relatively low count for patients who die early. This handling is not optimal, as the treatment effect has to be interpreted more from a healthcare system perspective than from the patient's perspective as I discuss in more detail in section 7.2. A patient who dies before the end of follow-up will not use any more resources. While these saved resources may lower the overall costs on the system level this is clearly a negative outcome from a patients' perspective. This has to be taken into account when interpreting the estimated effects while the importance of this aspect depends on the mortality rate and effect of treatment on mortality present in the dataset. An extension of the approach is to introduce an offset to the outcome model that accounts for the varying lengths of follow-up. I estimate such a model for the primary outcome as a sensitivity analysis. If the follow-up lenghts differ a lot between the two groups, this approach would be more appropriate to ensure the effect still reflects the question of interest. On the other hand, the additional parameter makes the model more complex which might complicate the interpretation of the resulting estimand.

Table 1: Definition of the target trial

Eligibility criteria	Eligibility of clusters: German physicians who treat diabetes patients (both general practitioners and specialists) Physicians identify eligible patients and conduct one medication review at the next appointment. Patient eligibility: 1) at least 65 years old 2) currently takes metformin 3) needs SU or DPP4 as additional treatment according to assessment of the physician (no contraindication against glibenclamid, Glimepirid or DPP4 and no clear indication to prescribe a different additional treatment Current users of the intervention or control treatment are included. Their individual risk profile is reassessed and they might be assigned a different medication than before
Treatment Strategies	Intervention: *Priscus*-informed prescription of medications that will usually result in intake of DPP4 Control: Standard of care that will usually result in prescription of SU to some patients
Treatment Assignment	Physicians are randomly assigned to one strategy and are aware of their strategy. Physicians start identifying all their eligible patients at the start of the trial. For the duration of a recruitment period, physicians also include any new patients as soon as they fulfill the criteria. Each patient is included only once when he/she is first eligible.
Outcomes	• 1-year number of all-cause hospitalisations and all-cause outpatient visits (as composite outcome) while alive • 30-day risk of all-cause hospitalisation while alive • 1-year risk of hypoglycemia while alive • 1-year risk of all-cause mortality
Follow-Up	From treatment assignment until death, loss-to-follow-up, or end of follow-up (1 year after scheduled chart review), whichever occurs first.
Causal contrast of interest	Intention-to-treat effect (of prescribed treatment)
Analysis Plan	Intention-to-treat effect estimated as rate ratio (primary outcome) and odds ratio (secondary outcomes) of outcomes.

Table 2: Definition of the estimand

Population of interest	Whole population of patients over 65 years old, who currently initiate SU or DPP4 as add-on medication to metformin or who need to initiate SU or DPP4 as add-on treatment to metformin, and who have a similar distribution of prior duration of intake, treatment adherence and dosing to the study sample.
Intercurrent events	...handled under *treatment policy* strategy and considered as part of the treatment regimen: • switch to alternative treatment • early treatment discontinuation • hospital stay • new initiation of other medication that can also cause adverse events • change of disease status that requires additional or different treatment (treatment escalation) ...handled under the *while-on-treatment* strategy: • death during follow-up
Population-level summary measure	ATE in patients with clinical equipoise (overlap weights)
Estimand stated as a research question	What would be the difference in adverse outcomes if all elderly patients who require additional medication to metformin had participated in a *Priscus*-informed medication review which would result in initiation of treatment with an DPP4 inhibitor, compared to if they had not received such a medication review and continued treatment with SU?

3.2 Reducing bias caused by emulation

In the following I outline the study design of the emulated trial. I discuss the most relevant of the aspects outlined in chapter 2.1 that were identified by Hernán and Robins (2016)[18] as being especially important for bias reduction when emulating a trial with routinely collected data.

By first defining a hypothetical RCT I aimed to avoid as much bias as possible without following any predefined "good practice" design. Rather than reproducing existing designs for observational studies I primarily tried to emulate the target RCT as closely as possible. Figure 1 visualizes my final study design. The structure of the diagram is modeled on the study diagram that was suggested in the Structured Template and Reporting Tool for RWE (STaRT-RWE) studies on the safety and effectiveness of treatments [34]. A public-private consortium created the template on behalf of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). It provides a guiding tool to enhance the reproducibility of RWE studies, support the transparent communication of methods and reduce misinterpretation. Besides the detailed study design it also includes information on the exact definition of all inclusion criteria, confounders and analysis plans. I completed the template for this study and appended it in the zip-file that also contains all skripts. The start and end of the trial was defined following

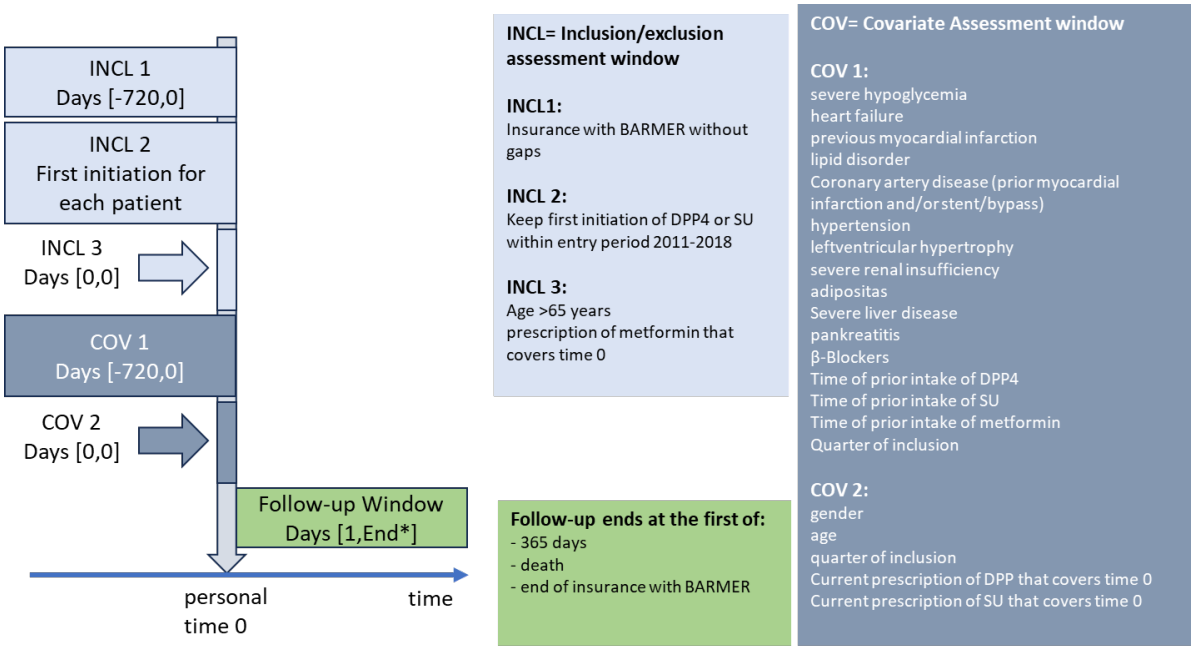


Figure 1: Study design of the emulated trial

primarily practical considerations. In principal, data from 2005 until 2022 were at disposal. The earliest meaningful year to start is 2007, as this was the year the first DPP4 inhibitor was marketed in Germany. Other DPP4 products followed shortly after (Sitagliptin in 2007 [4], Vildagliptin 2008 [5] and Saxagliptin in 2009 [3]. Also, to ensure comparability of the data, e.g. to avoid influences of health-political decisions about data management and other practical reasons, I restrict the study period further to 2009-2018. Especially our main outcome, the number of healthcare contacts, is majorly affected by changing documentation rules. In 2008, a new billing concept („pauschalisierte Versorgung") was

introduced. This resulted in a lower number of contacts per patient and quarter than in the previous years as only one contact per patient can now be invoiced at the health insurance provider by each physician [1]. Thus, it is no longer possible to deduct the number of days with contacts from the documented data. Additionally, starting from the fourth quarter of 2008, direct settlement for joint laboratories („Direktabrechnung der Laborgemeinschaften“) was introduced. Laboratory services now have to be billed separately from other services [1]. This led to a higher number of total documented cases. To ensure that all baseline covariates are measured in a comparable way, I set the start of the trial to 2011 with 2009-2010 as pre-baseline period. The end of the follow-up was set to 2019 with an end of recruitment in 2018, as this avoids all periods influenced by the coronavirus epidemic. Again, especially our main outcome but also all areas of the health system were severely impacted by the pandemic and data were not comparable to prior periods.

Immortal time bias and selection bias are avoided in this study as each participant gets assigned a personal *time zero*. The patient is assigned to the treatment strategy that is initiated the first time this participant meets the inclusion criteria. Only data prior to each index initiation are used to determine eligibility. The baseline assessment is carried out up-to and on the day of treatment initiation and participant follow-up starts on the day of initiation.

One of the inclusion criteria is a simultanous treatment with metformin. In the hypothetical RCT, a treatment plan is devised by the physician and fixed before treatment allocation, so he or she can actively select those patients who are in need for add-on treatment and who are supposed to continue taking metformin in the future. As the treatment plan can not be controlled in an observational study, I will check whether the initiation of the add-on treatment falls within the active days supply of the last metformin prescription. This procedure follows the approach from a “refill pattern method” developed by Liu et al. (2016) [23]. The implications of this implementation are discussed in chapter 7.2.

3.3 Confounder selection

In comparative analysis of treatments in general, but particularly in this case, it is complex to identify all causes of the outcomes but more straightforward to identify the causes of the exposure. Therefore, we used the general method for confounder selection described in chapter 2.3 but focus on identifying all covariates that cause the exposure instead of trying to also select additional covariates that primarily cause the outcome. Causes of the exposure here are factors that influence a physicians decision for one treatment or the other. A physician who prescribes in compliance with the *Priscus* list 2.0 [25] should always prefer DPP4 over SU in patients over 65 years independently from dosing, treatment plan or prior treatment with the exception of contraindications. However, in real-world practice SU were and still are prescribed frequently. Thus, the task of choosing an appropriate set of confounders aims at identifying those decision criteria that were used by physicians in everyday-practice during the study period. To get an exposure set that is as comprehensive as possible, expert knowledge from three different sources was combined:

1. Factors that are recommended as decision criteria in the current guidelines for treatment of diabetes [9] & [7]. The current S3 guideline defines risk factors that should be taken into account when choosing the treatment.

2. Factors that were considered in Table I of relevant RCTs. To identify RCTs I did not conduct a literature review but used an existing systematic review on the safety of DPP4 compared with SU that was conducted in the development process of the *Priscus* list 2.0 [13].

3. An expert (Dr. Thürmann) checked and extended the list with her expert knowledge.

In the next step we determined which variables can be found in our data, how comprehensive the representation of each variable is and I drew a causal directed acyclic graph to identify potential instrumental variables using the browser-based program DAGitty v3.1 [2]. Dr. Thomas Grobe (aQua) provided advice on the reliability of the representation of each of the considered factors in claims data. A complete list of factors considered in the expert review and the comments on reliability can be found in Table 6 in the Appendix.

Only three comorbidities are specified as contraindications for one of the two treatment options according to the summaries of product characteristics. The diagnoses *severe renal insufficiency* and *severe liver insufficiency* are contraindications for *glibenclamid*. The *Priscus* list 2.0 also states these as comorbidities that should be avoided and both are also mentioned as critical factors in the current S3-guidelines. Pancreatitis is a severe side effect of DPP4, after which the treatment should be discontinued or switched. The extent to which physicians take any other factors into account is more ambigous and certainly also varies a lot between different physicians. The summary of product characteristics of glibenclamid [6] and also the *Priscus* list 2.0 advise particular caution and monitoring in the presence of several factors that increase the risk for hypoglycemia, among others an age over 65 years. The summary of product characteristics also mentions that beta-blockers can disguise the symptoms of hypoglycemia.

The existing guidelines on diabetes treatment [9] & [7] provide general factors that should influence the treatment decision but no specific criteria to choose between different options for second-line diabetes treatment. The S3 guideline for type 2 diabetes is currently under revision but the relevant chapter on medication therapy was published in 2021 [9]. The guideline recommends a specified treatment algorithm. Both DPP4 and SU should only be prescribed as second-line treatment in addition to metformin if the patient's individual therapy target is not achieved. Metformin is recommended as first-line treatment for all patients except for contraindications. Patients with a diagnosis of cardiovascular diseases should receive a combination therapy with metformin and SGLT2 inhibitors or GLP1-analogues. Otherwise, the choice of additional treatment depends on personal risk factors that are individually assessed. The risk should be assessed on these 14 criteria (translated from the German S3-guideline):

- biological age
- sex
- diabetes duration
- lifestyle/diet/lack of physical activity
- family/genetic predisposition
- hypertension
- dyslipidemia
- adipositas
- kidney insufficiency

- albuminuria
- smoking status
- strong metabolic instability and severe hypoglycemia
- left ventricular hypertrophy
- subclinical arteriosclerosis or subclinical cardiovascular disease

There also exists a S2k guideline on diabetes in elderly patients from 2018 [7]. The level of evidence of this guideline is generally lower than for the general S3 guideline, which underlines the relevance of focusing research directed at elderly patients. Nevertheless, the guideline provides further arguments for choosing DPP4 as the treatment of interest instead of other options. DPP4 are generally recommended also for elderly patients. In contrast, the guidelines state that SGLT2 inhibitors are not indicated in patients with impaired renal function, which is a common condition among elderly patients. In addition, several side effects of SGLT2 inhibitors appear to be more prevalent among elderly patients [7]. When risk factors are respected they have advantages, e.g. no increased risk for hypoglycemia, and thus are very commonly prescribed also among elderly ($\approx$ 3% of over 65 year olds in Germany according to standardized Barmer population). GLP1-analogues are also not generally recommended for elderly patients due to the necessity of injection and side effects like nausea and weight loss, but they can be used in particular cases [7]. This is consistent with the recommendations of the *Priscus* list 2.0, which does not mention GLP1-analogues as an alternative medication for elderly patients. We used these criteria as a first proposal in the selection of possible confounders as these are factors that physicians take into account when deciding which treatment they prescribe. However, the criteria in the guidelines are very vague and thus needed to be complemented by the exact information on contraindications from the summary of product characteristics and expert knowledge of the current prescribing practices.

Table 3 provides the final list of 16 expert-selected confounders and their representation in the data. All diagnoses are coded according to the German modification of the international classification of diseases 10th revision (ICD-10-GM). The ICD-10-GM was valid in Germany during the entire study period. The translation of some of the diagnoses to ICD10 Codes follows the coding algorithms used in Wilke et al.(2014)[36] [11]. The causal directed acyclic graph (DAG) shown in Figure 2 was designed based on the final list of confounders. All comorbidities and the co-medication *betablockers* clearly influence both the treatment and the outcome and should be included as confounders (bottom left corner of the graph). We also controlled for the most recent prescription and the time under each of the two treatments by including the total amount of prescribed daily defined doses (DDD) in the past two years. *Ethnicity*, the *estimated glomerular filtration rate (eGFR)* and *diabetes duration* were identified as confounders, but are not represented adequately in the data. As claims data generally do not include laboratory values and medical test results the eGFR is only indirectly represented in the data in the form of ICD-10-GM diagnosis code N18.x. The subcategory (3rd digit) denotes the stage of a chronic kidney disease where the eGFR is the distinguishing criterion. Codes N18.4 and N18.5 correspond to severe cases of renal insufficiency with an eGFR below 30 ml/min, which are taken into account by the included confounder *severe renal insufficiency*. Diabetes diagnoses and therefore in principle also the *prior duration of diabetes* are present in the data. However, as only a pre-baseline period of two years is considered, the duration will be severely left censored for most patients. Diabetes type 2 as a chronic disease develops and remains present in a patients medical

history over long periods of time, often decades or even the entire lifespan. Therefore, a period of two years is much too short to identify the true onset of the disease. Also, a lack of diagnoses codes at the beginning of the observation period does not necessarily imply that the disease first started later, as chronic diseases are often not consistently encoded by physicians at every visit.

Table 3: Final list of confounding variables and their representation in the data

variable	**in data**	**ICD-10 code/ comment**
age	yes	
sex	yes	no evidence, but generally important
ethnicity	no	
diabetes duration	no	too severely left censored (max. 2 years pre-baseline)
severe hypoglycemia	yes	E11.61, E12.61, E13.61, E14.61, E16.0, E16.1, E16.2
heart failure	yes	I50.0, I50.1, I50.9, I11.0, I13.0, I13.2
myocardial infarction	yes	I21, I22
lipid disorder	yes	E78
Coronary Artery Disease:		
old myocardial infarction	yes	I25.2
stent or bypass	yes	Z95, I25.1
hypertension	yes	I10, I11, I12, I13, I15
leftventricular hypertrophy	yes	I51.7
severe renal insufficiency	yes	N18.4, N18.5
eGFR	no	
adipositas	yes	E66, potentially underreported
severe liver insufficiency	yes	I85.0, I86.4, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7
pancreatitis	yes	K85
Beta- blockers	yes	C07
treatment history	yes	5 variables: *current DPP*, *current SU*, *time DPP*, *time SU*, *time metformin*

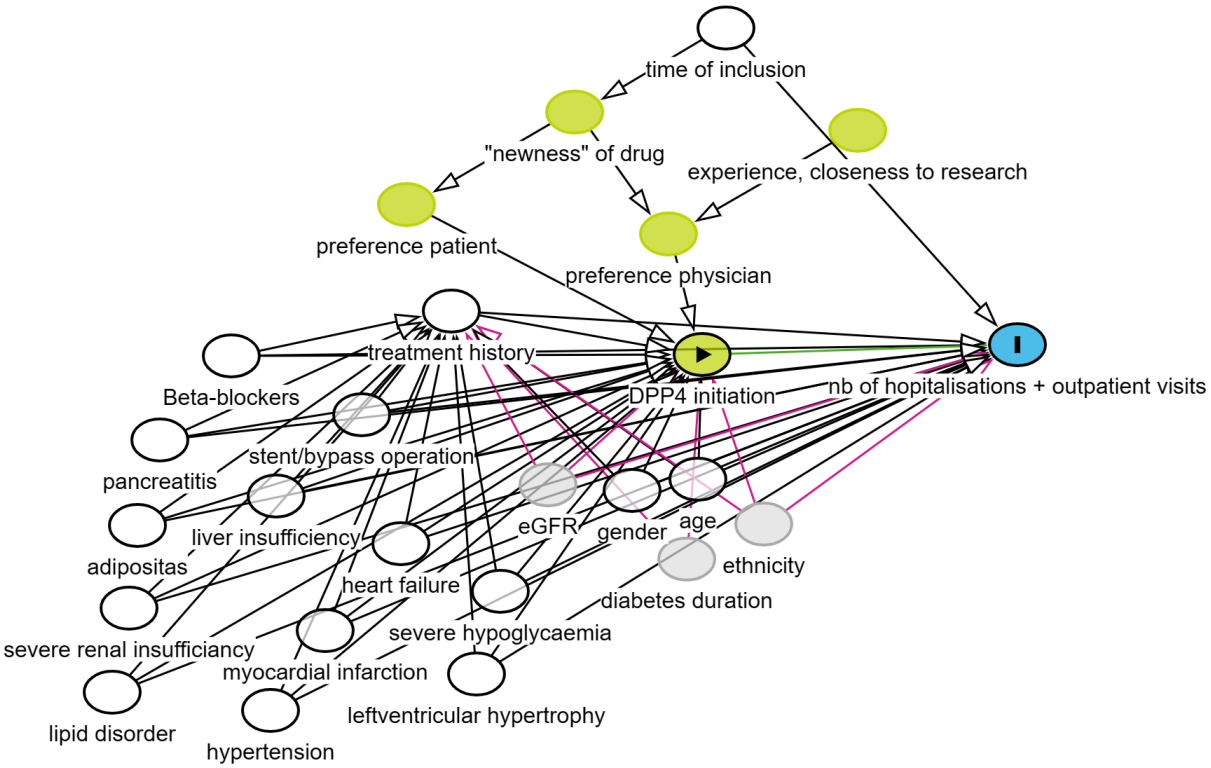


Figure 2: Directed acyclic graph (DAG) with Treatment (green with triangle), outcome (blue), confounders (white) and other causes of treatment (green). Open causal pathways are red.

4 Data source and implementation

I used data available in the Barmer scientific data warehouse (W-DWH). All datasets were pseudonymized by Barmer. Re-identification of an insured person can only be conducted by authorized Barmer employees. All data processing and analysis was conducted via secured remote access inside the data warehouse so that a linkage of data with other sources was not possible. The source files of the database could only be viewed but not accessed directly. Any direct download of datasets was prohibited and technically not possible. Results tables and aggregated data without personal references that are in the narrow sense connected to the research project could be downloaded on a personal computer on request by an authorised person. The data usage approval is part of the research partnership between Barmer and aQua GmbH as I used the same personal access to the data warehouse also for another project within the scope of my employment at aQua GmbH. The usage approval was extended to this master thesis by Barmer.

The preparation of the data and calculation of covariates was conducted in SAS (Version 9.4.4). For the statistical analysis I used R (Version 4.2.0). A *zip*-folder is appended that contains all skripts.

Available datasets include pseudonymised information on all persons insured by Barmer health insurance between 2005 and 2022 e.g. 1) demographics (longitudinal), 2) medical prescription data, 3) ambulatory data, 4) inpatient and outpatient hospital data, 5) therapeutic remedies and aids, 6) care data, 7) incapacity to work data, 8) dental data. Relevant for the implementation of the planned study were information from 1) to 4).
The insurance history is captured in the database in the form of single insurance episodes that sometimes overlap or continue on seamlessly. To better capture the practically relevant periods with or without valid insurance, it was necessary to create an aggregated table that includes completed episodes with aggregated information. I was able to use an existing table prepared earlier within other projects.

The basic identification level is an ID-Variable that each insured person gets assigned when a new insurance contract is concluded. In principle, each person should be identifiable with the same number for life. However, new temporary numbers often get assigned when persons for instance switch their insurance provider and later come back to Barmer. The temporary identifiers are regularly consolidated with the lifetime identifiers. Nevertheless, prescriptions or cases are sometimes identified only over the temporary ID. This makes it necessary to manually create a masterdata file that connects the relevant lifelong ids with all associated temporary ids.
Prescription data are directly identified by the person-ID. Diagnoses on the other hand belong to ambulatory or hospital cases that are identified via a combination of person-ID and a case number. Case numbers alone are not sufficient to identify all diagnoses of a specific patient as the case-IDs sometimes get reassigned.

As specified in the study design in chapter 3.2, I applied the inclusion criteria to distinct initiations of either DPP4 inhibitor or SU. Thus, the first data extraction step was to create a table with all such initiations within the time period of interest 2009-2019. Subsequently, each initiation gets assigned the corresponding lifelong ID number from the self-created masterdata, if possible. A few initiations (about 0.4%) could not be matched and were

excluded already at this point. By means of the lifelong ID other information associated with each initiation could now be added subsequently to check the eligibility of inclusion and to add covariates. Some relevant variables like *age*, *gender*, *time-to-death* or the *number of days with valid insurance* can be calculated using simple operations after merging the corresponding masterdata and insurance table to the table of initiations. Some factors that serve as inclusion criteria or confounders require more extensive operations. I prepared pseudo-code before writing the actual skripts to ensure all operations are carried out in appropriate order and achieve the intended goals. In the following I briefly depict some of the most important and complicated tasks.

One inclusion criterion is that the patient takes metformin as a firstline treatment. I chose an approach that focuses on checking the overlap between the end of the metformin prescription and the date of the index initiation as is justified in more detail in chapter 7.2. Two prescriptions overlap if the coverage period of the prior prescription of metformin extends beyond *time zero* with a tolerance of 20%. This means that at least 80% of the time in between the two prescriptions is covered. Coverage is calculated as the medication possession ratio: the number of days' supply of medication that was prescribed in the time period, divided by the number of days in the time period. The number of days' supply is calculated using the prescribed DDD. As the study is conducted on an elderly population with a high prevalence of reduced renal and liver function, I adjusted the DDD and assumed only half of the defined dose to better reflect the actual average dose of metformin prescribed to this population. A similar algorithm was useful to determine the current treatment status of each patient immediately prior to personal *time zero*. Any prior prescription of DPP4 or SU that has overlap with the date of the index initiation (with 20% tolerance) is designated as a current treatment. The full DDD is used here as renal and liver insufficiency influence the dosis of DPP4 and SU to a much smaller extent as for metformin.
The duration of prior treatment with DPP4, SU and metformin is simplified to the time since first initiation of the respective treatment within the observation period. Accordingly, potential gaps in treatment are ignored. I also considered more complicated solutions. For instance, the time could be calculated by determining the maximum time that is covered by a medication possession ratio of at least 80%. However, while this approach would contain more information, it would have a less straightforward interpretation, as it does not capture the whole extent of different treatment histories.
The identification of diagnoses associated to each included initiation required some considerations with regard to the efficiency of the skript. The datasets containing ambulatory diagnoses are extremely large as each of the included patients had approximately 15 physician visits per year and many diagnoses are recorded at each visits. Iterating through the entire tables to merge diagnoses to specific patients can easily result in calculation times of several days to weeks. Therefore, it was extremely important to reduce the size of the datasets as much as possible prior to any merging. I could substantially reduce the necessary calculation capacity in in the very first step by applying filters that only kept lines with those diagnosis codes I was interested in. Creating a dataset that included all relevant prior diagnoses from the 2-year pre-baseline period for each of the included patients then only took about 14 hours of calculation over night.
Finally, in one of last steps, information from the ambulatory and inpatient sector had to be merged to generate the final covariates and outcomes. Ambulatory diagnoses in particular are not always reliable in the sense that single incorrect diagnoses are common.

A common practice is to check that diagnoses from the ambulatory sector occur in at least two subsequent quarters. As a slight simplification I require diagnoses to occur at least twice within the relevant period to be valid, as incorporating the timing of diagnoses just for this issue would require considerable additional effort and no major gain in precision is to be expected.

5 Statistical Analysis

5.1 Confounding adjustment

Confounding adjustment through propensity score weighting was conducted. The log-odds of the propensity scores $e(x)$ were estimated using a generalized linear model with logit link function:

$$\log\left(\frac{e(x)}{1-e(x)}\right) = \beta' x$$

with β as the vector of j coefficients including an intercept and the identified confounders as covariates x_j

$$\begin{aligned}\text{confounders} = \{&\text{age, sex, heart failure, adipositas, coronary artery disease, leftventricular}\\ &\text{hypertrophy,lipid disorder, myocardial infarction, pancreatitis, severe liver}\\ &\text{disease, severe renal insufficiency, betablockers, current SU, current DPP}\\ &\text{+time SU, time dpp, time\_metformin}\}\end{aligned} \tag{5}$$

To decide which weighting method is best suited to balance the covariates and to determine the extent of overlap between the two groups I plotted and analysed the propensity score distribution. I used the R package *PSweight* as it includes diagnostic tools to analyse covariate balance and visualize overlap. As described in chapter 2.3 the weighting method that corresponds to the ATE in the combined total population is *inverse probability weighting* [21]. However, in the presence of a large proportion of extreme propensity scores, OW outperform inverse probability weights. As recommended by Zhou et al. (2020)[37] I therefore considered OW as the method of choice if extreme propensity scores are encountered. In any case, I applied both IPW and overlap weighting to compare the differences in achieved covariate balance. The balance of covariates between the treatment and control group that is achieved through weighting was assessed using standardized absolute differences. Most of the confounders are binary variables for which the balance between groups was measured as the standardized difference in proportion. For continuous variables standardized mean differences were calculated and the covariate balance was additionally assessed using the variance ratio and the p-value from a two-sample Kolmogorov-Smirnov test.

5.2 Effect estimation

Average treatment effects in the population with clinical equipoise were estimated by regressing the treatment on the outcome variable in the weighted population. In cases like this one where I do not control for any additional covariates in the outcome model, the coefficient for the treatment variable estimated by the outcome model is identical to an estimate that would be produced by a g-estimation [29].
Before defining an appropriate effect measure for the primary outcome I considered the distribution of the outcome as it is a count variable. Count outcomes often violate assumptions of ordinary least squares regression, namely conditional normality and homoskedasticity so that standard errors and tests of significance will be biased. Even so count data with a relatively high mean less often violate these assumptions, it is important to consider the outcome distribution in order to ensure an appropriate effect estimate is used. The

standard estimate for a continuous outcome is the absolute difference or ratio of means. However, if the outcome variable is skewed or displays excess positive kurtosis, a poisson or negative-binomial model is more appropriate. A negative binomial model with a log-link function estimates the rate ratio $exp(\beta_1)$:

$$\log(\mu) = \beta_0 + \beta_1 T \tag{6}$$

where μ is the expected count and T is the treatment indicator variable with the coefficient β_1 and intercept β_0.
The secondary outcomes are all binary and odds ratios $\frac{\pi}{1-\pi}$ were estimated using regression:

$$logit(\pi) = \beta_0 + \beta_1 T \tag{7}$$

where π is the probability of success (event coded as 1).

The weights were incorporated into the likelihood estimation in a similar way as balancing weights in a survey design. I used the R package „survey“ with the function „svyglm()“. The function estimates generalized linear models with robust standard errors, which are necessary to account for the additional error produced by the estimation of the weights. The survey package produces conservative standard errors by using Horvitz-Thomson-type standard errors, which are a generalization of the 'sandwich' estimator.

5.3 Sensitivity Analyses

An e-value was calculated for all effects to assess the robustness of the estimates to potential unobserved confounding [33]. To determine how the handling of the intercurrent event of death before end of follow-up influences the primary outcome, I estimated an alternative model for the rate ratio that includes an offset for the time-to-death:

$$\log(\mu) = \beta_0 + \beta_1 T + \log(\text{time-to-death+1}) \tag{8}$$

In addition, I conducted a subgroup analysis in the population of those patients that are naive to both treatments prior to their respective entry into the trial. As the treatment history was very difficult to model in sufficient detail, this subgroup analysis in new users helps to narrow down the extent of possible unobserved confounding. However, new users form a distinct group of patients, so the comparability of the estimated effects depends on how greatly the group of new users differs from the overall population with respect to age and the presence of comorbidities.

5.4 Subgroup analysis

In addition, analyses for four more subgroups were conducted. Besides the subgroup of new users, I also consider a particularly old population of patients over 80 years old and three groups with specific comorbidities: patients with severe hypoglycemia, severe renal diseases and heart failure respectively. I hypothesize that DPP4 should be superior to SU in the the entire population of over 65 year old patients. However, no further evidence exists on both the question whether the age limit of 65 is meaningful and whether there are other factors that modify the effect. It is of interest to determine more specific subgroups that are at higher risk for developing adverse events when taking SU and therefore would profit most from switching to DPP4.

5.5 Missing data

A valid ID was required for each for each initiation to connect administrative information with claims and diagnosis. If an initiation of DPP4 or SU could not be matched with administrative information because there was no ID at provided or because the ID1 (temporary ID) could not be connected to a valid ID these lines were excluded. For some claims the information on the daily defined dose contained in the issued drug package was missing. In these cases, I imputed the overall mean of the DDD for the respective substance.
For most covariates only positive values are recorded. In a strict sense, the presence of only one comorbidity does not necessarily imply the absence of others. Missing diagnosis can result from either a lack of collection when patients are not asked about all possible conditions at each physician visit or from a lack of documentation by the physician [35]. However, any imputation would be very complex and practically impossible to implement for the entirety or even a large number of diagnoses at once as the overall health status always needs to be taken into account but depends on all other present conditions. Also, the level of the patients healthcare utilisation has a large influence. For instance, people who visit physicians frequently are more likely to have data also on minor conditions not directly connected to the cause of the visit than patients with fewer contact. I assume that no claims are missing and impute a 0 for any missing values in any column except for ID, *age* and *gender*. No missing values were encountered for *age* and *gender*.
I excluded patients that had gaps in their insurance history prior to baseline. I conducted a complete case analysis if cases with incomplete insurance during follow-up were encountered. As insurance is mandatory in Germany and people do not frequently switch between insurance companies, especially not elderly people, these missing cases should not be systematically correlated with the outcome.

6 Results

In the upcoming section, I describe the characteristics of the enrolled population, assess the covariate balance that could be achieved through propensity score weighting, and present effect estimates for the primary and secondary outcomes, along with the sensitivity and subgroup analyses.

6.1 Population characteristics

I identified a total of 7.1 Mio initiations of DPP4 or SU between 2009 and 2019 in the Barmer population. Of these, 31,803 observations had to be excluded because they could not be matched to an insured person from the master data of the database. The exclusion criteria were subsequently applied to the resulting population as specified in the flowchart in Figure 5. Only the first eligible initiation of any insured person was included. 171,318 initiations were included, of which 111,865 initiated the treatment (DPP4) and 59,453 the control (SU).
As prevalent users are included and this aspect is very relevant to the analysis and interpretation, Figure 3 and Figure 4 provide a breakdown of the included population into groups with a specific treatment history. Detailed counts are displayed at the bottom of the flowchart in Figure 5. In the treatment group, the largest subgroup (52%) is formed by *new users* who are naive to both treatment alternatives. The second largest group consists of *continued users* (also referred to as current users) who already received DPP4 immediately before the index initiation (23%). 18% are *re-initiators* who also previously received DPP4 in the 2 year assessment period prior to their personal *time zero* but had a longer gap since their last prescription. Only a minority of about 7% of patients in the DPP4 group have switched from SU or both to only DPP4 since their previous prescription. The control group on the other hand contains a larger proportion of *continued users* (44%) and *re-initiators* (30%). Only 21% of the control group are *new users* and 5% are *switchers*.

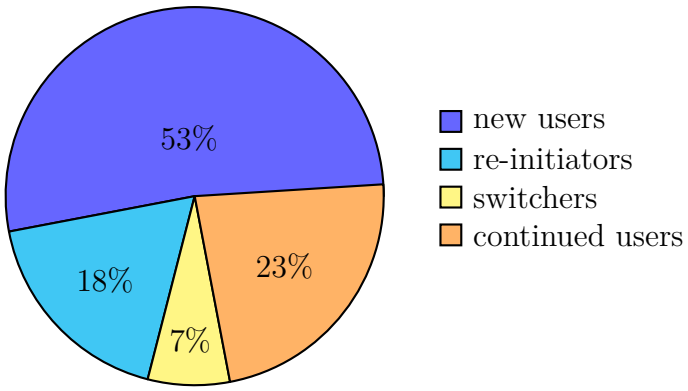


Figure 3: DPP4 group

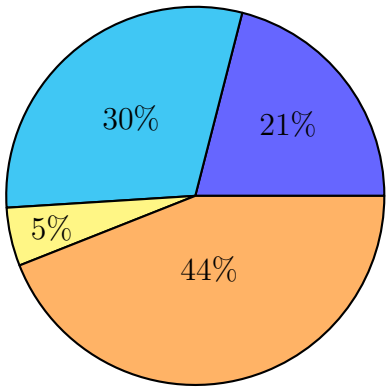


Figure 4: SU group

Figure 5: Flowchart of inclusion of observations in the study

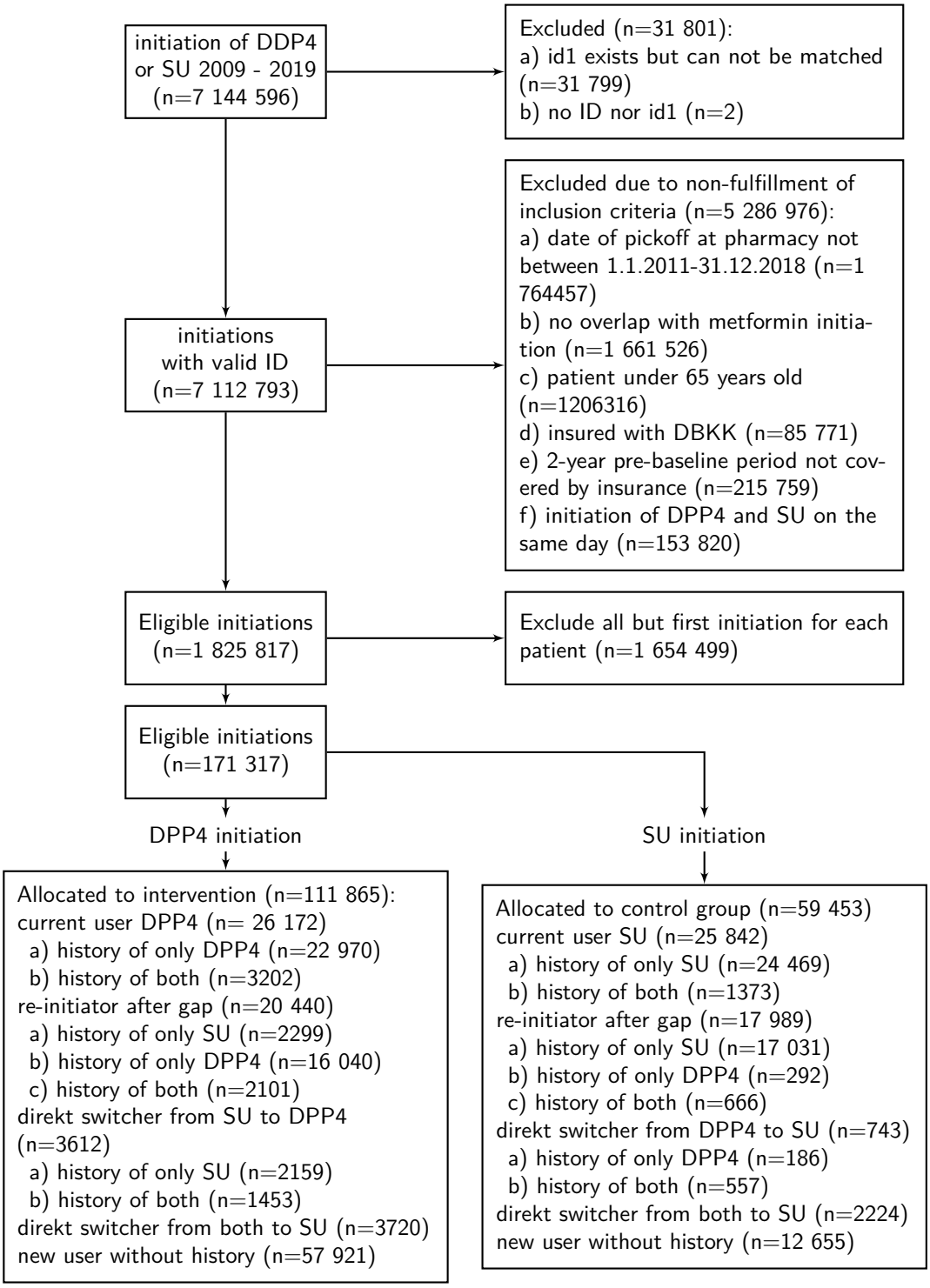


Table 4: Baseline characteristics and absolute standardized difference (ASD) between treatment and control group

Characteristic	**DPP4** (n = 111865)	**SU** (n = 59453)	**ASD**
heart failure = 1 (%)	22942 (20.5)	10426 (17.5)	0.076
adipositas = 1 (%)	39818 (35.6)	17464 (29.4)	0.133
betablocker = 1 (%)	67873 (60.7)	34771 (58.5)	0.045
hypertension = 1 (%)	100517 (89.9)	52821 (88.8)	0.033
CAD = 1 (%)	18662 (16.7)	8120 (13.7)	0.084
leftventricular hypertrophy = 1 (%)	3254 (2.9)	1263 (2.1)	0.050
lipid disorder = 1 (%)	69032 (61.7)	34352 (57.8)	0.080
myocardial infarction = 1 (%)	4550 (4.1)	1769 (3.0)	0.059
pancreatitis = 1 (%)	2816 (2.5)	1291 (2.2)	0.023
severe hypoglycemia = 1 (%)	271 (0.2)	90 (0.2)	0.021
severe liver disease = 1 (%)	1654 (1.5)	778 (1.3)	0.014
severe renal insufficiency = 1 (%)	1064 (1.0)	196 (0.3)	0.078
age (median [IQR])	72.00 [67.00, 77.00]	73.00 [69.00, 78.00]	0.229
sex = 1 (%)	56687 (50.7)	29974 (50.4)	0.005
cohort n (median [IQR])	14.00 [6.00, 23.00]	4.00 [1.00, 11.00]	0.806
current DPP = 1 (%)	29892 (26.7)	2967 (5.0)	0.623
time DPP (median [IQR])	0.00 [0.00, 482.00]	0.00 [0.00, 0.00]	0.783
current SU = 1 (%)	7332 (6.6)	28066 (47.2)	1.032
time SU (median [IQR])	0.00 [0.00, 0.00]	600.00 [134.00, 682.00]	1.644
time metformin (median [IQR])	659.00 [375.00, 700.00]	676.00 [537.00, 706.00]	0.143

Table 4 displays the baseline characteristics of the included population and the balance of the identified confounders between the treatment and control group before applying propensity score weighting. The last column shows the baseline absolute standardized difference between the treatment and control group.

In the total population the average age was 72 years (IQR, 67-77) and 50.6% of the participants were males. The most prevalent of the comorbidities observed in the total population are hypertension (89.5%), lipid disorder (60.3%) and adipositas (33.4%). DPP4 initiators were slightly younger (median 72 vs 73) and were more frequently suffering from all the observed comorbidities compared with those initiating SU. However, the standardized difference in proportion for the comorbidities is generally small (<0.1), except for adipositas, which is present in 35.6% the DPP4 initiators but only in 29.4% of the initiators in the SU group. The majority of patients in both groups received beta-blockers in the two years prior to cohort entry, 60.7% in the treatment group and 58.5% in the control group.

Differences between the groups are larger with respect to their median entry into the cohort, their most current prior treatment and their treatment history. The median cohort entry is much later for DPP4 initiators (14 [6,23]) than SU initiators (4 [1,11]). The majority of SU initiators entered the cohort at the beginning of the study with less than 25% entering after quarter 11 which corresponds to the second half of the year 2013. In the DPP4 group the median entry is in quarter 14. Figure 3 and 4 already visualized the proportion of *new users*, *continued users*, *re-initiators* and *switchers* in both groups. As there are few *switchers* in both groups, there is little overlap in the sense that few patients have a similar

treatment history prior to baseline. While the majority of patients in the SU group has taken SU during almost the entire pre-baseline period of 720 days (median 600 [134,682] days), the majority of DPP4 initiators has not taken any SU prior to baseline (median 0 [0,536] days). Most patients from both groups are naive to treatment with DPP4 (median 0 [0,482] and 0 [0,0]), but while 44% of DPP4-initiators have at least some prior history of DPP4, this proportion is only 12% in the SU group.

6.2 Confounder adjustment via propensity score weighting

Figure 6(a) shows the unadjusted propensity score distribution in treatment (blue) and control (red) group. The treatment group displays a narrow peak at approximately 0.8 and a second peak very close to 1. The propensity scores of the control group on the other hand are concentrated in one high peak close to zero with a lower second peak at around 0.8. The overlap between the two groups is best in the right part of the distribution where the value of the propensity score is about 0.7 or higher but not yet close to 1. Both extremes, propensity score very close to 1 or 0 have very little to no overlap which is to be expected.

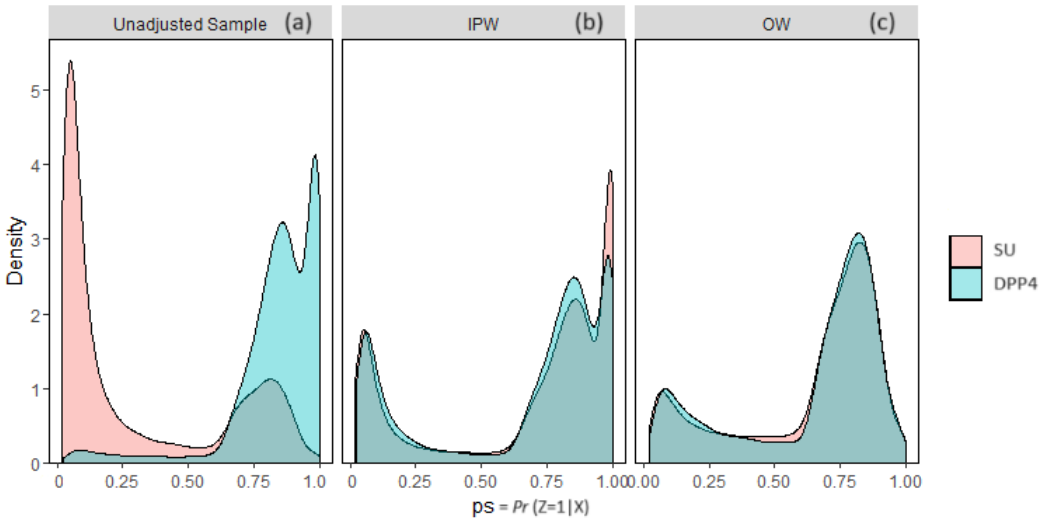


Figure 6: Propensity score distribution in treatment and control group. (a) Unadjusted and adjusted using IPW (b) and overlap weighting (c)

A comparison of the weighted propensity score distributions achieved by IPW and overlap weighting is displayed for our data in the middle (b) and right (c) plots in Figure 6. Both IPW and overlap weighting perform well. As our sample is relatively large (171,318 patients) the number of patients in the areas with little overlap still seems to be high enough to achieve adequate balance also with IPW. However, the highest propensity score is larger than 0.99 and the 75% quantile lies at 0.91 (6). As inverse-probability weights are calculated by taking the inverse of the propensity score, these extreme propensity scores produce a high amount of large weights with the highest weight in IPW taking a value as high as 1,441 (6). A detailed analysis of balance for each covariate clearly favors overlap weighting as it achieves nearly perfect balance in means for all confounders (Figure 7). With IPW on the other hand, the achieved balance is worse, in particular with respect to two variables that represent the prior treatment with DPP4, *currentDPP* and *timeDPP*. Both variables

also displayed very large baseline differences in proportion respectively median (4). IPW seems to overcompensate the difference as weighting switches the sign of the difference in mean from positive to negative.
The Kolmogorov-Smirnov statistics and variance ratios for the continuous variables (Figure 8) show that there remains some imbalance also when using overlap weighting even so the standardized mean difference is zero. Nevertheless, the overall resulting balance achieved with overlap weights is clearly better than with IPW, so I will use these for analysing the average treatment effect.

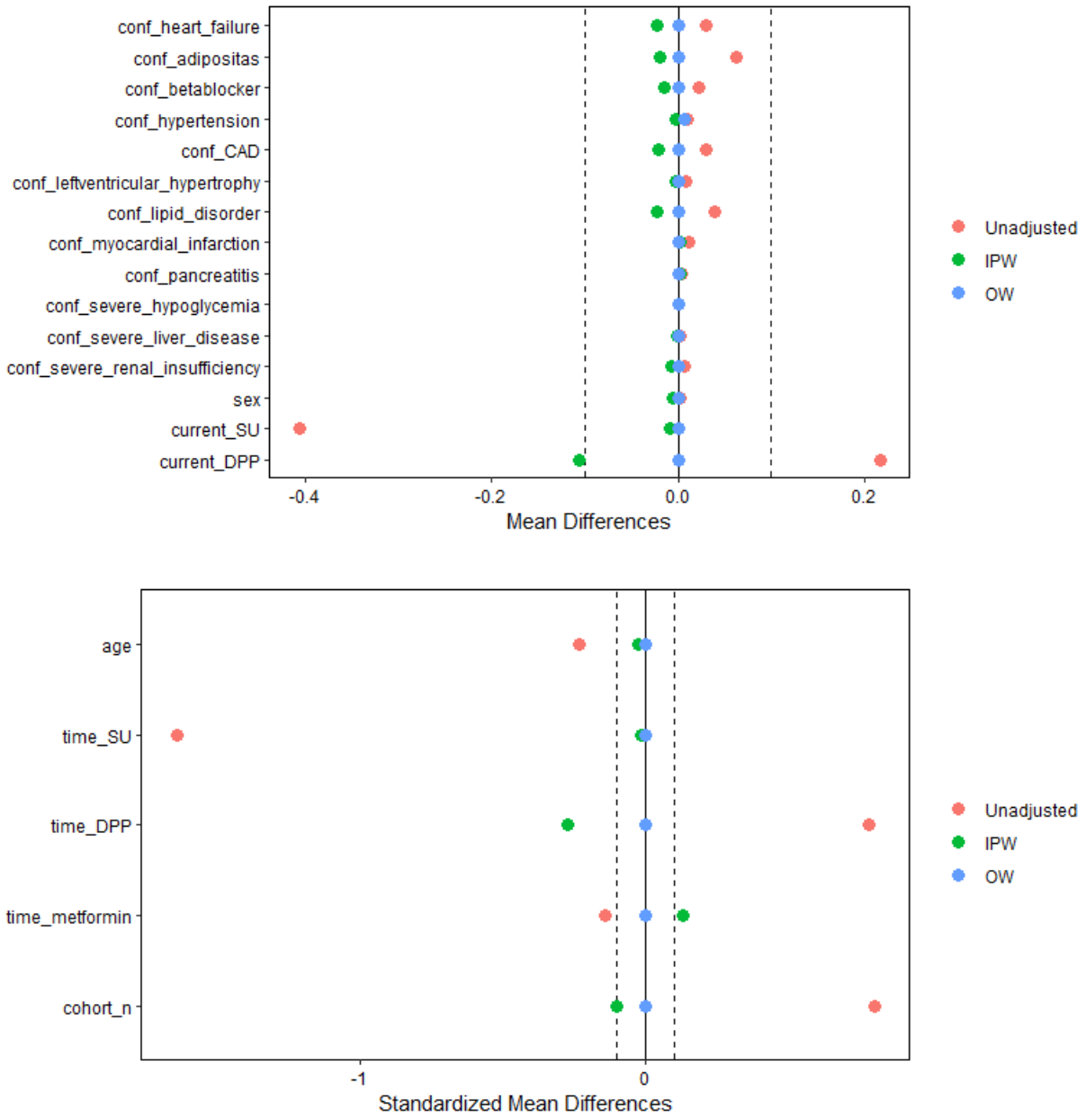


Figure 7: Covariate balance

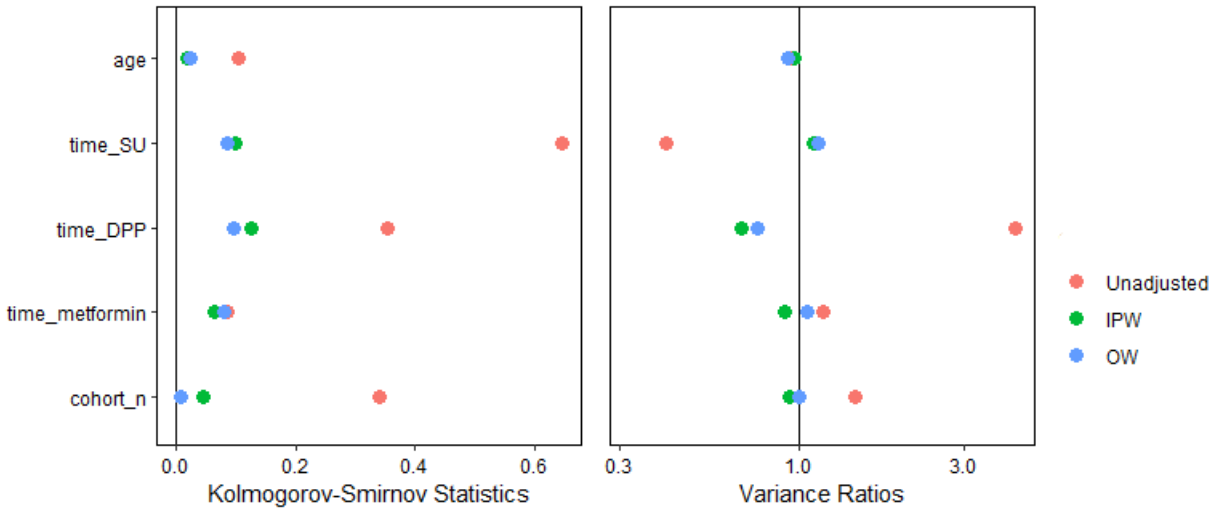


Figure 8: Kolmogorov-Smirnov Statistics and Variance Ratios for continous convariates

6.3 Average treatment effect in the overlap population

The distribution of the weighted primary outcome shown in Figure 9 is not normally distributed. It is slightly skewed with an inflation below the mean. Therefore, I modeled it as a count outcome using a negative-binomial model. Using a negative-binomial model instead of a standard poisson model is appropriate as there is some overdispersion, which is visible in the deviation from the line in the QQ-plot in Figure 10.

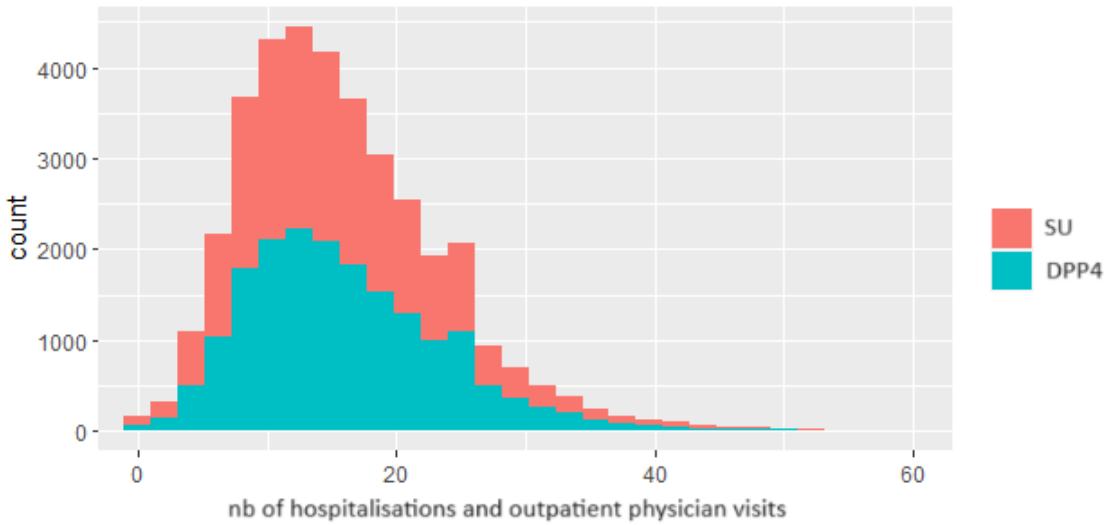


Figure 9: Weighted distribution of the primary outcome

Figure 11 shows the effect estimates for the average treatment effect in the overlap population for the primary outcome. The estimated effects for the secondary outcomes are shown in Figure 12. Both figures highlight the main analysis in blue and display the subgroup analyses underneath.

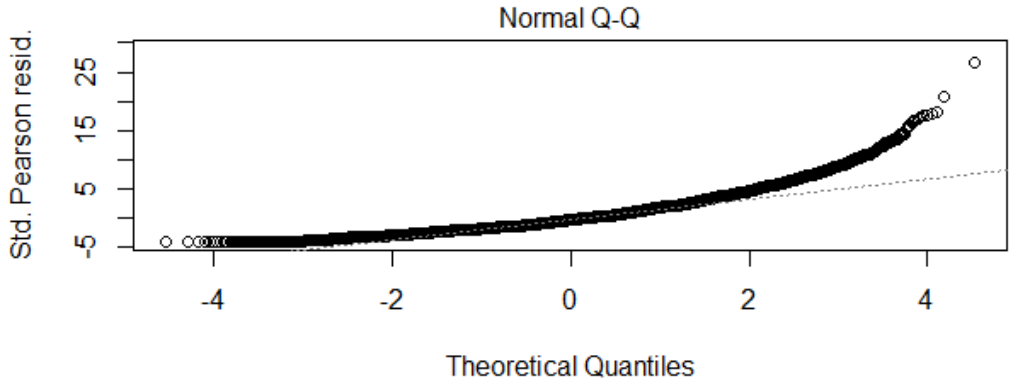


Figure 10: QQ-plot of the standardized pearson residuals for a GLM for the outcome regressed on treatment

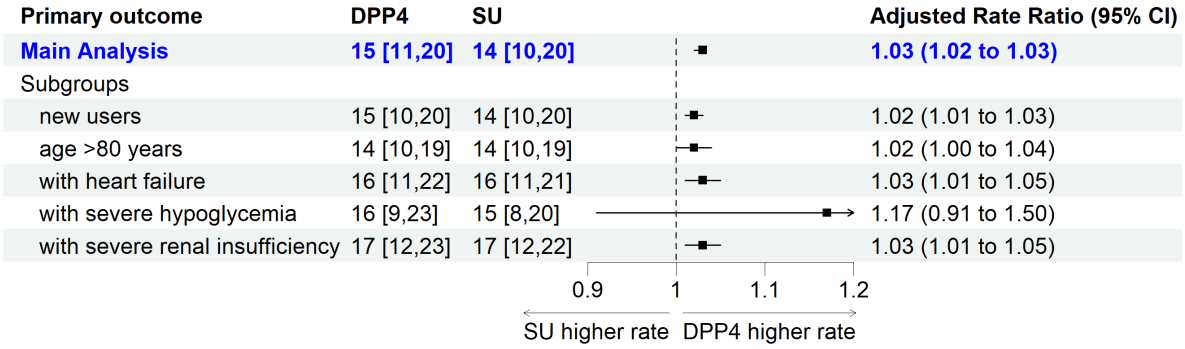
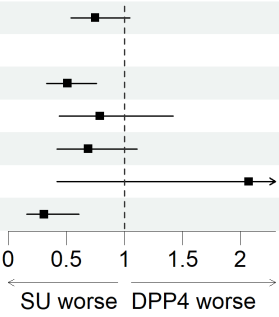


Figure 11: ATO of the primary outcome in total population and subgroups

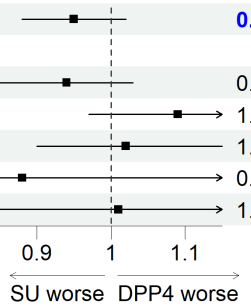
In the total population, the mean rate of hospitalisations and outpatient visits within one year was higher in the DPP4 group (15 [11,20]) than in the SU group (14 [10,20]) with a rate ratio of 1.03 (95% CI 1.02-1.03) (Figure11). According to the e-value, the observed effect could be nullified by an unmeasured confounder that was associated with both the treatment and the outcome by a rate ratio of 1.03 each, taking into account the already measured confounders. The confidence interval could be moved to include the null if an unmeasured confounder had a rate ratio of at least 1.02 each. The rate ratio was identical in sensitivity analyses with an offset for time-to-death included in the outcome model (Table 8 in the Appendix). In all subgroup analysis, the direction of the effect of the primary outcome was similar. Except for the very small subgroup of patients with a prior diagnosis of severe hypoglycemia the effect was also significant (12) with similar e-values as in the main analysis (Table 8 in the Appendix).

The odds for the secondary outcomes *1-year all-cause mortality*, *1-year severe hypoglycemia* and *all-cause hospitalisation within 30 days* were higher in the SU group than in the DPP4 group of the total population, but the differences were not significant (Figure 12).
However, the effects were more distinct in some subgroups. The odds for *severe hypoglycemia* were significantly higher in the SU group than in the DPP4 group among the subpopulations of *new users* and those patients with a severe renal insufficiency. The e-values of 5.51 and 54.78 suggest, that the harmful effects of SU in these subgroups are robust to unobserved

Outcome Severe Hypoglycemia	DPP4	SU	Adj. Odds Ratio (95% CI)
Main Analysis	**35.4 (0.2%)**	**46.9 (0.3%)**	**0.75 (0.54 to 1.05)**
Subgroups			
new users	13.2 (0.1%)	27.6 (0.3%)	0.51 (0.33 to 0.76)
age >80 years	10.9 (0.4%)	14.4 (0.5%)	0.79 (0.44 to 1.42)
with heart failure	16.1 (0.4%)	24.4 (0.7%)	0.69 (0.42 to 1.11)
with severe hypoglycemia	1.0 (3.2%)	1.5 (2.4%)	2.07 (0.42 to 10.18)
with severe renal insufficiency	3.9 (0.2%)	13.7 (0.8%)	0.31 (0.16 to 0.61)



Outcome Mortality	DPP4	SU	Adj. Odds Ratio (95% CI)
Main Analysis	**680.2 (3.7%)**	**717.3 (3.9%)**	**0.95 (0.88 to 1.02)**
Subgroups			
new users	420.5 (4.2%)	448.3 (4.5%)	0.94 (0.85 to 1.03)
age >80 years	318.2 (10.9%)	294.6 (10.0%)	1.09 (0.97 to 1.23)
with heart failure	300.0 (8.2%)	295.4 (8.1%)	1.02 (0.90 to 1.15)
with severe hypoglycemia	6.4 (20.4%)	7.1 (22.6%)	0.88 (0.35 to 2.20)
with severe renal insufficiency	110.4 (6.3%)	109.0 (6.2%	1.01 (0.83 to 1.23)



Outcome Hospitalisations	DPP4	SU	Adj. Odds Ratio (95% CI)
Main Analysis	**1028.9 (5.6%)**	**1074.4 (5.8%)**	**0.95 (0.90 to 1.02)**
Subgroups			
new users	622.3 (6.2%)	693.2 (7.0%)	0.89 (0.82 to 0.96)
age >80 years	257.4 (8.8%)	223.3 (7.6%)	1.17 (1.02 to 1.34)
with heart failure	359.2 (9.9%)	318.7 (8.8%)	1.14 (1.02 to 1.28)
with severe hypoglycemia	6.5 (20.6%)	3.7 (11.6%)	1.98 (0.74 to 5.29)
with severe renal insufficiency	155.3 (8.8%)	119.5 (6.8%)	1.33 (1.11 to 1.59)

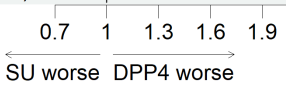


Figure 12: ATO of the secondary outcomes in total population and subgroups

confounding. For the outcome *hospitalisations* the direction of the effect differed between the subgroup of *new users*, where SU is harmful and the other subgroups, where the odds for at least one hospitalisation within 30 days are higher in the DPP4 group. Accept for the very small subgroup of patients with prior hypoglycemia, all these effects were significant. The effects on the outcome *mortality* were inconclusive between the subgroups and no significant effects were observed.

7 Discussion

The conducted study has several strengths. The „target trial emulation“ concept was applied to detect and control important sources of bias by design. The overlap weighting used for confounder adjustment lead to adequate balance of the observed confounders. Also, the use of the e-value allows to assess the impact of unobserved confounding. Two important sources of bias were addressed in sensitivity analyses. The subgroup analysis in new users offered a possibility to narrow-down the influence of the partially incomplete modeling and data representation of the treatment history. The inclusion of an offset for the time-to-death showed that the intercurrent event *death before end of follow-up* had no substantial influence on the estimated effect. One of the most important advantages of the analysis is its high external validity. The results are only fully representative for patients insured with Barmer. However, as Barmer is the second- largest public health insurer in Germany and has more than eight million members, the included population is also be considered as highly representative for the overall German population.
One the other hand, observational analysis and particularly the usage of claims data also has important limitations. I discuss the most important of the encountered issues in the following sections.

7.1 Definition of the target trial and estimand

One particularly complicated task with regard to the general study design was to decide whether to include and how to handle prevalent users of the treatments of interest or patients who switched between the two substances prior to the start of the trial. Existing literature on target trial emulation encourages the use of new-user designs as this generally leads to lower risk of bias [18] [27]. Prevalent users „survived“ the early period of treatment, which can introduce selection bias if risk is not constant over time [27]. For instance, an extreme case would be to include patients only after a surgery has taken place. As the surgery itself and the immediate postoperative period have very high risk, the treatment effect will be overestimated as patients who died during or shortly after surgery are excluded [27]. However, the study targets older patients who rarely start a new medication. It is of special interest to research how long-term users could profit from a treatment switch. In order to answer our research question it does not make sense to generally exclude these long-term users as the study population would then be limited to patients who have a shorter diabetes duration, less severe disease status and who are younger than the total population. The diabetes medication I am looking at should have a relatively constant risk over time as the effect of the medication is short-term and does not accumulate. However, the occurrence of adverse events certainly depends on how well the treatment plan and dosing is adjusted to the patient, which does in turn depend on the experiences that patient and physician have made. Also, the occurrence of adverse events certainly causes patients to switch or discontinue treatment which leads to a relative depletion of patients with a disposition or high risk for adverse events in the group that receives the inferior drug.

I estimated the effect in first-time users separately as a sensitivity analysis to see how important these problems are. However, *new users* form a distinct group of patients, so any comparability of the estimated effects depends on how greatly the group of *new users* differs from the overall population with respect to important effect modifiers like age and the presence of comorbidities. Table 5 provides a comparison of characteristics of the weighted

total population and the weighted population of *new users*. The most important difference lies in the proportion of patients with a prior event of hypoglycemia, which is only half as high in the population of *new users* compared to the total population. *New users* are slightly older, more often male and entered the cohort later. Overall however, the observed covariate distributions are very similar. An important unobserved effect modifier is most certainly diabetes duration. Assuming though, that such unobserved characteristics do not modify the treatment effect substantially, the difference in effect sizes between the new-user group and the overall population might indicate the extent and direction of bias that was introduced by unobserved confounding related to the treatment history. For the primary outcome, the effect in *new users* is slightly closer to null and for the secondary outcome, the harmful effect of SU is considerably more pronounced. In both cases, bias seems to most likely conceal harmful effects of SU, as the results in the potentially less biased analysis in the *newuser*-population are closer to the expected harmful effect of SU. A

Table 5: Characteristics of weighted total population and weighted population of new users

Characteristic	**total population** (n=36,915.9)	**new users** (n=19,927.6)
heart_failure = 1 (%)	7215.6 (19.5)	4055.8 (20.4)
adipositas = 1 (%)	11746.0 (31.8)	6166.5 (30.9)
betablocker = 1 (%)	22199.5 (60.1)	12105.8 (60.7)
hypertension = 1 (%)	32977.4 (89.3)	17784.1 (89.2)
CAD = 1 (%)	5658.5 (15.3)	3158.7 (15.9)
leftventricular_hypertrophy = 1 (%)	967.4 (2.6)	547.8 (2.7)
lipid_disorder = 1 (%)	21969.7 (59.5)	11676.7 (58.6)
myocardial_infarction = 1 (%)	1325.9 (3.6)	744.6 (3.7)
pancreatitis = 1 (%)	928.4 (2.5)	515.9 (2.6)
severe_hypoglycemia = 1 (%)	69.6 (0.2)	27.9 (0.1)
severe_liver_disease = 1 (%)	547.7 (1.5)	301.2 (1.5)
severe_renal_insufficiency = 1 (%)	221.3 (0.6)	140.1 (0.7)
age (median [IQR])	73.00 [68.00, 78.00]	74.00 [70.00, 78.00]
sex = 1 (%)	19074.7 (51.7)	10724.0 (53.8)
cohort_n (median [IQR])	9.00 [3.00, 18.00]	12.00 [6.00, 20.00]
time_metformin (median [IQR])	659.00 [276.00, 700.00]	655.00 [252.34, 699.00]

topic related to the estimand definition that I considered in more detail is how to handle the intercurrent event *death before end of follow-up*. The effect of the treatment on mortality and the exact mortality rate was not clear beforehand which made it difficult to choose the best estimand and to assess the importance of this decision. Kahan et al. (2020)[20] discuss the interpretation of possible estimands for outcomes truncated by death. According to the authors, a strategy that should be avoided whenever treatment might affect mortality, is a complete case analysis with the exclusion of patients who die during follow-up, as this can introduce substantial bias depending on the magnitude of difference in mortality rates [20]. One recommended approach is a composite strategy which would assign the worst possible outcome to individuals who die before the end of follow-up [20]. This ensures that death is

not counted as a positive outcome. The disadvantage of this approach, in particular for count outcomes like our primary outcome, is that the meaning of the estimated effect size will no longer be clear, as a difference in the number of visits between the groups can be solely due to a difference in mortality rates between the groups and the exact choice of the values assigned to the dying patients has a large influence on interpretation. Therefore, I chose the second proposed estimand, a *while-on-treatment* strategy. This handling is not optimal either, as the treatment effect then has to be interpreted more from a healthcare system perspective than from the patient's perspective. A brief analysis of the weighted distribution of the primary outcome suggests that the treatment does not have a substantial effect on mortality. Figure 13 in the Appendix shows the distribution of time-to-death in the two groups in the included population. About 3.4% of the included cases die between *time zero* and end of follow-up. While small differences are visible, the overall distribution is very similar in both groups. The sensitivity analysis with an offset for the varying lengths of follow-up and the secondary analysis of the outcome *mortality* both confirm this finding. Thus, the handling of this intercurrent event does not seem to have an important impact on the estimated effect.
Another terminal intercurrent event that is present in the included population is the end of a patients insurance before the end of follow-up. Similar to death such cases cannot be observed after the intercurrent event occured, which for instance excludes the possibility to handle these events under a *treatment policy* strategy. Contrary to death however, the end of insurance should be completely independent from the intervention, so I consider these values to be missing completely at random and conduct a complete case analysis. In total, only 962 (0.4%) of the included cases have incomplete insurance coverage between *time zero* and end of follow-up which indicates that again no substantial impact of the intercurrent event is to be expected.

One aspect of the target trial that I was not able to emulate is the clustering of patients over different physicians. In the target cluster-RCT, randomisation would be conducted on the physician level as the intervention of a *Priscus*-informed medication review can not be implemented on an individual level. Single physicians would be randomised to either intervention or control group and would then prescribe their patients a medication according to their best knowledge. Apart from patient-characteristics, the decision criteria can also depend on characteristics of the physician. The used claims data do not contain any physician-level variables apart from the geographical region and the medical specialty the physician works in. Thus, I did not have enough information to directly include cluster-level covariates as confounders. Other possibilities are to do a very strict emulation of a cluster RCT by estimating the propensity scores within each cluster [10], or to include the cluster membership as either a fixed or a random effect. In an extensive simulation study conducted by Arpino et al. (2011), directly including all relevant cluster information performed best in terms of bias and MSE, but a fixed or random effect for the cluster membership also reduces the imbalance of the unobserved covariates considerably and lead to a better model compared to a method that ignores all cluster information [10]. However, the power of all three analyses strongly despends on both a sufficient cluster size and that the positivity assumption holds within the clusters as there needs to be enough overlap between the groups for all relevant covariates within each cluster. I originally planned modeling the physician as a random factor in a generalized linear mixed model to estimate the propensity scores. Unfortunately, both the cluster sizes and the overlap between the two groups within

the clusters are insufficient to allow any modeling of the cluster design in our study. Even so the population and the number of clusters is very large, many of the physician only treat few patients. Many physicians have a preference to predominantely or even exclusively prescribe either DPP4 or SU (see Appendix Table 14). Despite these apparent problems I tried to estimate the linear mixed model but as expected, convergence could not be reached. I consider this unsuccessful attempt to incorporate the cluster aspect into the study design to be a significant weakness of the analysis.

7.2 Reducing bias caused by emulation

One of the inclusion criteria defined in the target trial is that patients need to take metformin as first-line therapy at *time zero*. This is a good example to underline how helpful it is to both follow the „target trial emulation“ framework and to define a clear estimand. In the emulation with claims data, I have no information on the treatment plan and therefore cannot distinguish between patients who are prescribed SU or DPP4 as an add-on treatment to metformin and those who are switching away from metformin. Polypharmacy studies often distinguish between polypharmacy and switching by using methods that check the overlap of days supply [23]. If both medications are taken simultanously for more than an arbitrary number of days the medications are classified as polypharmacy. Since in our emulation study the *time zero* for each patient is set to the initiation of the add-on treatment, counting the overlap days would necessarily require information about the period that follows *time zero*. Under the „target trial emulation“ framework such a setting is critical. Patients should never be included or excluded based on an event that occurs after treatment assignment as this would be an impossible procedure in a real trial and introduces survivorship bias. Patients who for example die or experience serious side effects soon after the start of the trial will not be able to refill their metformin prescription and would consequently be excluded from the study. The study sample would then be artificially healthier than the total population of interest. Instead, in this case we need to accept that there will be a lack of information and some risk that patients are wrongfully included. I did not find a possibility to ensure that metformin and DPP4/SU are taken simultanously and not successively without incorporating information from after *time zero*. To at least take into account as much information as possible I checked whether the initiation of the add-on treatment falls within the active days supply of the last metformin prescription. This procedure follows the approach from a “refill pattern method” developed by Liu et al. (2016) [23]. Instead of defining a fixed overlap, the “refill pattern method” considers medications as simultanous if each prescription is refilled within the active days supply of the other. Following Liu et al. (2016)[23], at least two prescriptions for each of the two medications are required. In my case, I only have one prescription each that occurs before or at *time zero*. I do not know whether the patient continued to take the remaining doses of metformin and whether he/she received a refill soon enough before the coverage with the last prescription was over, but I can at least exclude patients without any overlap between the two medications. The lack of information changes the target population to which the estimate applies from patients who receive metformin and an add-on treatment to patients who still had an active metformin prescription at the time when they initiated DPP4 or SU. As I chose to model the adverse event *treatment switching or discontinuation* under the *treatment policy* strategy, I did not check for adherence to the treatment regimen.

7.3 Confounder Selection

The process of confounder selection and definition, especially the expert input in the last step turned out to be time-consuming. For future projects it could be reasonable to develop an even more structured approach and involve more than just one pharmacological expert. Also, the summaries of product characteristics of each of the compared treatment alternatives might be a better starting point for choosing confounders than guidelines and RCTs, as they allow to more reliably identify contraindications and interactions with simultaneous medication. Building on these, less unambiguous criteria can then be supplemented from relevant RCTs and guidelines. The process should ideally also include a defined procedure for selecting the appropriate ICD-10 codes for each condition that is to be included as a confounder. Both, the agreement of codes with the medically appropriate diagnosis and the extent to which coding adheres to guidelines, varies greatly and is not seldom influenced by financial considerations. The selection of the practically most relevant codes and an assessment of the extent of under- or overrepresentation of a diagnosis in the data therefore requires substantial expert knowledge. Ideally, comprehensive validation of codes should be considered for important covariates and the outcomes, which was out of the scope of this project.

Particularily relevant in this respect is the incomplete representation of hypoglycemia in the claims data. As the occurence of hypoglycemia in the patient history is a confounder for treatment choice, any systematic missclassification of such events will also have an impact on the estimated effects of the primary outcome. A comprehensive analysis of hospital stays due to adverse drug reactions based on ICU admission records, internal reports and discharge letters conducted by Schmiedl et al. (2017)[30] has revealed that coding of hypoglycemic events in German hospitals is very unreliable. Only about half of the occuring hypoglycemia were coded as such during discharge. Accompanying illnesses like renal insufficiency or heart insufficiency are often coded as main diagnosis instead of the hypoglycemia. There are no indications, that this misleading coding process is dependant of the exact drug that was taken by the patient so there should be no systematic differences between the two groups. Nevertheless, the magnitude of the missing data asks for a cautious interpretation of the estimated effects.

The role of the factor *treatment history* was the topic of particularly thorough discussions. As the study includes prevalent users of both DPP4 and SU as well as patients that have switched between the two in the past, both patients and physicians are certainly influenced by prior experiences when making the decision of continuing the current treatment or switching to an alternative. The question whether the factor should be controlled as a confounder depends on whether it also influences the outcome independently from the treatment. If many patients did not tolerate one medication in the past they might be over-represented in the other group while the first group appears artificially healthier, as patients with severe side effects already switched to the alternative. Also, while I expect both benefits and harms to occur mostly while or shortly after administering the respective treatment, effects can still spill-over to the study period, especially if the patient recently switched between treatments. Events like severe hypoglycemia immediately lead to contacts with the healthcare system in the form of physician visits or even hospitalisations. Indirectly, this influences the primary outcome as I can not assume that the future number of contacts is independent of the intensity of prior visits. An aspect that might have biased our analysis

is the incomplete representation of the exact treatment history. I controlled for the most recent prescription and the time under each of the two treatments by including the total amount of prescribed DDD in the past two years. However, the time of a switch between treatments is not captured. Spill-over effects might occur in patients who switched very shortly before their inclusion in the study. Also, a new treatment is sometimes introduced gradually, so in the beginning of a new treatment, dosing is adjusted frequently and some patients receive prescriptions for both options simultaneously so that it is not possible to deduct from the prescription data, which medication was in fact taken. I excluded patients who received both medications on the same day, but as one prescription usually supplies the patient for several months, it is impossible to determine an exact switching time. In the light of problems like these it makes sense that the inclusion of prevalent users in observational analysis is often discouraged.

Another question related to this is whether it is appropriate to adjust for variables that are influenced by prior treatment. The prior occurrence of severe hypoglycemia is such a critical variable. An inclusion of these variables would be problematic if they are colliders and induce a spurious association between exposure and outcome. Colliders are defined as variables that are independently caused by both the exposure and the outcome [19]. I argue that in principle, including the variable *severe hypoglycemia* as a confounder should not block the causal pathway as the events that are considered for determining the occurrence of hypoglycemia occur prior to baseline while the outcome value is assessed strictly after *time zero*. Not controlling for an important confounder might also introduce bias. I therefore decided to include *severe hypoglycemia* as a confounder in all analyses except for the outcome *risk for severe hypoglycemia within one year* which is most likely to be more severely impacted by the decision than the other outcomes. However, the topic warrants further discussions, which were outside the scope of this project.

I assume that there might be some even stronger unobserved confounding that is related to the health care conditions a patient is treated in. The medical covariates that we included as confounders were balanced unexpectedly well between the two groups already at baseline. As we carefully selected the medical factors that might be considered by physicians based on existing evidence, it is unlikely that we missed very influential comorbidities or comedications. One exception might be treatment with insulin. We did not consider the factor as a potential confounder, as it has no explicit and defined influence on treatment allocation. However, insulin can cause severe hypoglycemia and thus might have a very strong influence on all considered outcomes. Further analyses should determine whether the factor is really not associated with the treatment assignment, else it should be included as a confounder. The treating physician definitely also has a strong influence on the choice of treatment. Many physicians have a preference to exclusively prescribe either DPP4 or SU that seems to be independent of any patient characteristics. Thus, it would be appropriate to control for confounding factors at the physician level but that was not possible within this analysis as was justified above in chapter 7.2. Another possible unobserved confounding factor that Dr. Thürmann came up with when we discussed the results, is the participation of a patient in a disease management program (DMP). Existing evidence suggests that the DMP for diabetes type 2 has improved the quality of pharmacotheraphy in the participating patients [22][26][31]. Several observational analyses on claims data also suggest that DMP-participation influences the overall use of the health care system. The closer and more thorough monitoring of patients within the program increases the overall number of claims,

prescriptions and contacts with resident physicians [22][31], the frequency and duration of hospitalisations, and reduces diabetes complications as well as overall mortality [14]. This suggests that medication plans of patients who take part in a DMP follow existing evidence more closely; and adverse events and interaction effects with other drugs are better monitored by physicians for these patients. Thus, these patients might less often receive SU than other patients. In any future analysis similar to this one the factor should certainly be taken into account.

7.4 Data source and implementation

A problem that complicates the interpretation of the primary outcome relates to current coding practice. German claims data do not allow to deduce the total number of physicians visits from the recorded cases. According to the reimbursement regulations of the social health insurance scheme, physician can only invoice one visit for each patient within one quarter. Additional visits are thus not financially relevant. As elderly patients generally have a high contact rate, many patients will visit each of their physicians more than once every quarter. Thus, the recorded number of visits might be much lower than the true total number of visits. As we use an active comparator design and no differences in recording practice are to be expected between the two treatments, no bias is expected. Nevertheless, it is highly questionable whether the number of all-cause doctors visits within one year really is suitable to give an impression of the overall amount of adverse events or the overall health status of the patient.

8 Conclusion

The target trial emulation approach proved to be very helpful for developing a study design that effectively addresses the question of interest. I profited most from the structured approach when deciding how to handle prevalent users, how to account for the cluster structure and when determining which initiations of DPP4/SU should be included (the first for each patient). However, several of the most decisive decisions that I had to make during the „emulation step“ were not explicitely addressed by the elements included in the „target trial table“ that was suggested by Hernán and Robins (2016). Therefore, the additional elements from the estimand framework that I also defined prior to the practical implementation were a useful addition. First and foremost, the consideration of different possible population-level summary measures allowed me to straightforwardly compare the implications of choosing different methods for confounding adjustment. Besides that, the handling of intercurrent events and an exact definition of the outcomes can be very complex, and without such a structure, reducing bias and providing transparent reporting would have been even more difficult.

Overlap weighting achieved a very good balance of all observed confounders. It outperformed inverse probability weighting with respect to balancing those variables of treatment history that had limited overlap between the DPP4 and SU group. However, a high risk for bias due to unobserved confounding remains. In particular, strong unobserved confounding related to the personal preferences of the treating physician and the health care conditions a patient is treated in is possible.

The effect estimated for the primary outcome contradicts my initial hypothesis of a harmful effect of SU on all considered outcomes and in all subgroups. Among the elderly patients included in this cohort study, the rate of hospitalisations and physicians visits within one year was significantly higher in patients who received DPP4 than in those who were treated with SU. However, this primary outcome, as I defined it, might not be an adequate surrogate for overall adverse events or the general state of health of patients, as increased contact with the healthcare system can also be caused by preventative measures like DMPs. In addition, the effect on the primary outcome seems to be very susceptible to unmeasured confounding, as the e-value is very low in the total population and even lower in the subgroup analyses. The already low effect size might be additionally inflated by remaining confounding related to the treatment history. In light of of the numerous possible sources of unmeasured confounding that I identified, the estimated effect in favour of SU is not reliable. This very result is in line with the existing evidence from RCTs concerning the difference between DPP4 and SU in overall risk for adverse events, which is also inconclusive. With hindsight, it might have been better to focus on a primary outcome with a higher level of existing evidence, as this thesis was constructed as a proof-of-concept study. Clearer evidence on the harmful effect of SU exists for *mortality* and the very specific outcome *risk for hypoglycemia*. None of the effects for any of the secondary outcomes was significant in the overall population, but a clearly harmful effect of SU could be shown for the risk of severe hypoglycemia in some subgroups. The direction and approximate effect sizes of the two secondary outcomes correspond to the findings from RCTs. The effect is most pronounced and clearly significant among new users and patients with a pre-existing renal insufficiency. The e-values for these two subgroup-effects also suggest that they are relatively robust to unmeasured confounding.

References

[1] BARMER GEK Arztreport 2011. https://www.barmer.de/resource/blob/1026932/d5630a0f349e388b65fd28ad616b7257/barmer-gek-arztreport-2011-data.pdf. Accessed: 2023-12-04.

[2] Dagitty v3.1. https://dagitty.net/dags.html. Accessed: 2023-10-18.

[3] KBV Information Saxagliptin. https://www.kbv.de/media/sp/Wirkstoff_AKTUELL_Saxagliptin.pdf. Accessed: 2023-12-04.

[4] KBV Information Sitagliptin. https://www.kbv.de/media/sp/Wirkstoff_AKTUELL_Sitagliptin.pdf. Accessed: 2023-12-04.

[5] KBV information vildagliptin. https://www.akdae.de/fileadmin/user_upload/akdae/Arzneimitteltherapie/WA/Archiv/Vildagliptin.pdf. Accessed: 2023-12-04.

[6] Fachinformation glibenclamid abz tabletten. https://www.abz.de/assets/products/de/label/Glibenclamid%20AbZ%20Tabletten%20-%202.pdf?pzn=1015995, Jul 2018. Accessed: 2023-12-04.

[7] S2k Diabetes mellitus im Alter. 2018. Accessed: 2023-12-04.

[8] ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf, 2020. Accessed: 2023-12-10.

[9] S3 guideline typ 2 diabetes. https://register.awmf.org/assets/guidelines/nvl-001l_S3_Typ_2_Diabetes_2021-03.pdf, 2021. Accessed: 2023-12-04.

[10] Bruno Arpino and Fabrizia Mealli. The specification of the propensity score in multilevel observational studies. *Computational Statistics amp; Data Analysis*, 55(4):1770–1780, April 2011.

[11] Je-Wook Chae, Chang Seok Song, Hyang Kim, Kyu-Beck Lee, Byeong-Sung Seo, and Dong-Il Kim. Prediction of Mortality in Patients Undergoing Maintenance Hemodialysis by Charlson Comorbidity Index Using ICD-10 Database. *Nephron Clinical Practice*, 117(4):379–384, 11 2010.

[12] Arthur Chatton, Florent Le Borgne, Clémence Leyrat, Florence Gillaizeau, Chloé Rousseau, Laetitia Barbin, David Laplaud, Maxime Léger, Bruno Giraudeau, and Yohann Foucher. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports*, 10(1), June 2020.

[13] Katharina Doni, Stefanie Bühn, Alina Weise, Nina-Kristin Mann, Simone Hess, Andreas Sönnichsen, Dawid Pieper, Petra Thürmann, and Tim Mathes. Safety of dipeptidyl peptidase-4 inhibitors in older adults with type 2 diabetes: a systematic review and meta-analysis of randomized controlled trials. *Ther. Adv. Drug Saf.*, 13:20420986211072383, January 2022.

[14] Anna Drabik, Christian Graf, Guido Büscher, and Stephanie Stock. Evaluation der effektivität eines disease management programms diabetes mellitus in der gkv - erste ergebnisse und methodische Überlegungen. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 106(9):649–655, 2012. Aktuelle Probleme der Gesundheitsversorgung ? Neue Erkenntnisse und Lösungsvorschläge.

[15] Heinz G. Endres, Petra Kaufmann-Kolle, Valerie Steeb, Erik Bauer, Caroline Böttner, and Petra Thürmann. Association between potentially inappropriate medication (PIM) use and risk of hospitalization in older adults: An observational study based on routine data comparing PIM use with use of PIM alternatives. *PLOS ONE*, 11(2):e0146811, February 2016.

[16] Sander Greenland and Hal Morgenstern. Confounding in health research. *Annual Review of Public Health*, 22(1):189–212, 2001. PMID: 11274518.

[17] Noah Greifer and Elizabeth A. Stuart. Choosing the causal estimand for propensity score analysis of observational studies. 2023.

[18] Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available: Table 1. *American Journal of Epidemiology*, 183(8):758–764, March 2016.

[19] Mathias J. Holmberg and Lars W. Andersen. Collider bias. *JAMA*, 327(13):1282, April 2022.

[20] Brennan C. Kahan, Tim P. Morris, Ian R. White, Conor D. Tweed, Suzie Cro, Darren Dahly, Tra My Pham, Hanif Esmail, Abdel Babiker, and James R. Carpenter. Treatment estimands in clinical trials of patients hospitalised for covid-19: ensuring trials ask the right questions. *BMC Medicine*, 18(1), September 2020.

[21] Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.*, 188(1):250–257, January 2019.

[22] Roland Linder, Susanne Ahrens, Dagmar Köppel, Thomas Heilmann, and Frank Verheyen. The benefit and efficiency of the disease management program for type 2 diabetes. *Deutsches Ärzteblatt international*, March 2011.

[23] Xinyue Liu, Paul Kubilis, Regina Bussing, and Almut G Winterstein. Development of a refill pattern method to measure polypharmacy in administrative claims databases. *Pharmacoepidemiology and Drug Safety*, 25(12):1407–1413, August 2016.

[24] Kim Luijken, Rik van Eekelen, Helga Gardarsdottir, Rolf H. H. Groenwold, and Nan van Geloven. Tell me what you want, what you really really want: Estimands in observational pharmacoepidemiologic comparative effectiveness and safety studies. *Pharmacoepidemiology and Drug Safety*, 32(8):863–872, March 2023.

[25] Nina-Kristin Mann, Tim Mathes, Andreas Sönnichsen, Dawid Pieper, Elisabeth Klager, Mahmoud Moussa, and Petra A. Thürmann. Potentially inadequate medications in the elderly: PRISCUS 2.0—first update of the PRISCUS list. *Deutsches Ärzteblatt international*, January 2023.

[26] Michael Mehring, Ewan Donnachie, Florian Cornelius Bonke, Christoph Werner, and Antonius Schneider. Disease management programs for patients with type 2 diabetes mellitus in germany: a longitudinal population-based descriptive study. *Diabetology & Metabolic Syndrome*, 9(1), May 2017.

[27] W. A. Ray. Evaluating medication effects outside of clinical trials: New-user designs. *American Journal of Epidemiology*, 158(9):915–920, November 2003.

[28] Jinma Ren, Paul Cislo, Joseph C. Cappelleri, Patrick Hlavacek, and Marco DiBonaventura. Comparing g-computation, propensity score-based weighting, and targeted maximum likelihood estimation for analyzing externally controlled trials with both measured and unmeasured confounders: a simulation study. *BMC Medical Research Methodology*, 23(1), January 2023.

[29] Ryan D. Ross, Xu Shi, Megan E. V. Caram, Phoebe A. Tsao, Paul Lin, Amy Bohnert, Min Zhang, and Bhramar Mukherjee. Veridical causal inference using propensity score methods for comparative effectiveness research with medical claims. *Health Services and Outcomes Research Methodology*, 21(2):206–228, October 2020.

[30] S Schmiedl, M Rottenkolber, J Szymanski, B Drewelow, W Siegmund, M Hippius, K Farker, I R Guenther, J Hasford, and P A Thuermann. Preventable adrs leading to hospitalization — results of a long-term prospective safety study with 6, 427 adr cases focusing on elderly patients. *Expert Opinion on Drug Safety*, 17(2):125–137, December 2017.

[31] Stephanie Stock, Anna Drabik, Guido Büscher, Christian Graf, Walter Ullrich, Andreas Gerber, Karl W. Lauterbach, and Markus Lüngen. German diabetes management programs improve quality of care and curb costs. *Health Affairs*, 29(12):2197–2205, December 2010.

[32] Tyler J. VanderWeele. Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219, March 2019.

[33] Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*, 167(4):268, July 2017.

[34] Shirley V Wang, Simone Pinheiro, Wei Hua, Peter Arlett, Yoshiaki Uyama, Jesse A Berlin, Dorothee B Bartels, Kristijan H Kahler, Lily G Bessette, and Sebastian Schneeweiss. Start-rwe: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ*, 372, 2021.

[35] Brian J. Wells, Amy S. Nowacki, Kevin Chagin, and Michael W. Kattan. Strategies for handling missing data in electronic health record derived data. *eGEMs (Generating Evidence amp; Methods to improve patient outcomes)*, 1(3):7, December 2013.

[36] Thomas Wilke, Antje Groth, Andreas Fuchs, Lisa Seitz, Joachim Kienhöfer, Rainer Lundershausen, and Ulf Maywald. Real life treatment of diabetes mellitus type 2 patients: an analysis based on a large sample of 394,828 german patients. *Diabetes Res. Clin. Pract.*, 106(2):275–285, November 2014.

[37] Yunji Zhou, Roland A Matsouaka, and Laine Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29(12):3721–3756, July 2020.

Appendices

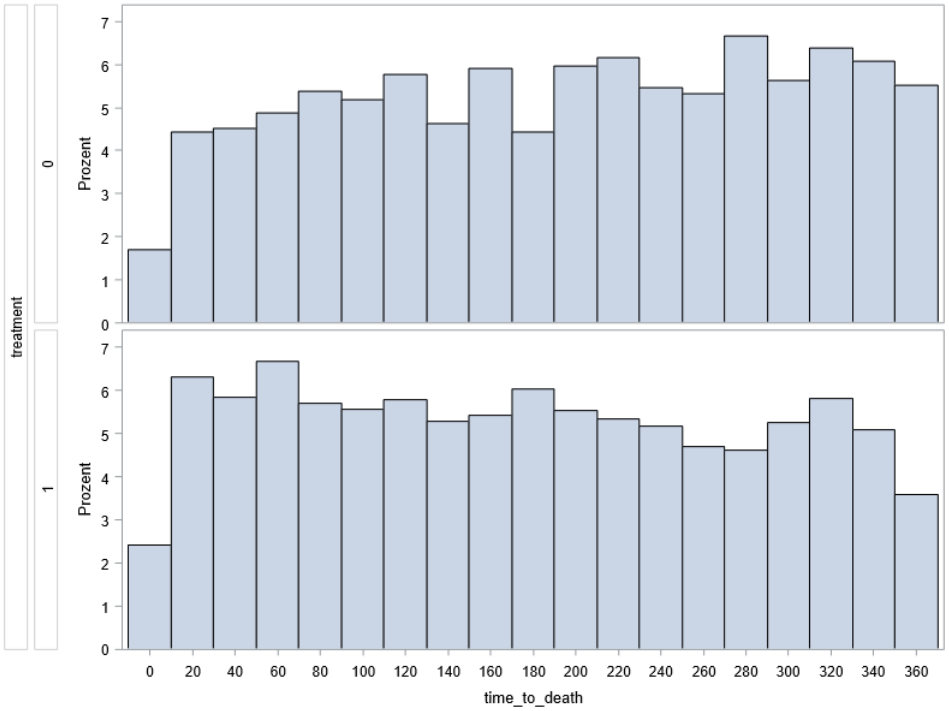


Figure 13: Distribution of time to death in treatment compared to control group

Table 6: All variables considered as confounders. The blue section was completed by Dr. Thürmann, the grey section by Dr. Grobe. Green lines correspond to the variables that were included.

Variable	in mind. einer RCT als Baseline-Kovariable berück-sichtigt	Entscheidungsfaktor in aktueller S3 Leitlinie	Variable sollte aufgenommen werden	ggf. Entscheidungs-kriterium	Kommentar	in Routinedaten zuverlässig darstellbar
Age	ja	ja	ja	65 years	gemäß unseres systemat Reviews	ja
Sex	ja	ja	nein		es gibt keine Hinweise in der Literatur. Man könnte höchstens sagen, dass sex immer berücksichtigt werden sollte - dann ja.	ja
Ethnicity	ja		ja		es gibt schon Unterschiede bei Diabetes allgemein	nein
Medical history:						
Smoking Status (Never smoker, Ex-smoker, current smoker)	ja	ja	nein		sollte keinen Unterschied auf die Therapieauswahl haben	geringe Sensitivität, Dokumentation könnte von bestimmten Erkrankungen/Therapien abhängen
Diabetes duration, years	ja	ja	ja			in dokumentierter Vorbeobachtungszeit ja
schwere Hypoglykämien		ja	ja			in dokumentierter Vorbeobachtungszeit ja
Microvascular complications	ja		nein		kein Entscheidungsfaktor	in dokumentierter Vorbeobachtungszeit ja
Heart failure (narrow SMQ 'cardiac failure')	ja		ja		ganz wichtiger Faktor!	in dokumentierter Vorbeobachtungszeit ja
Previous myocardial infarction	ja		ja			in dokumentierter Vorbeobachtungszeit ja
Cardiovascular accident	ja					in dokumentierter Vorbeobachtungszeit ja ????
History of lipid disorder	ja	ja	ja			in dokumentierter Vorbeobachtungszeit ja
comorbidities:						
subklinische kardiovaskuläre Erkrankung		ja	nein		ist nicht in Studien als relevanter Faktor belegt	ob "subklinisch" ist in Routinedaten i.d.R. schwer abgrenzbar und setzt bei ICD-10-Kodes entsprechende Differenzierungen voraus
manifeste kardiovaskuläre Erkrankung	ja	ja	ja		previous MI, heart failure	
Atherosclerotic CV disease	ja					in dokumentierter Vorbeobachtungszeit ja
Coronary artery disease	ja		ja		CAD als Diagnose liegt wahrscheinlich bei Z.n. Infarkt vor, den würde ich auf alle Fälle nehmen und wenn Z.n. Stent/Bypass Op.	in dokumentierter Vorbeobachtungszeit ja
Cerebrovascular disease	ja		nein		kein Entscheidungskriterium	in dokumentierter Vorbeobachtungszeit ja
Peripheral artery occlusive disease	ja		nein		kein Entscheidungskriterium	in dokumentierter Vorbeobachtungszeit ja
vascular disease	ja		nein		zu unscharf	
Microvascular disease	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Diabetic neuropathy	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Diabetic nephropathy	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Diabetic retinopathy	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Hypertension	ja	ja	ja			in dokumentierter Vorbeobachtungszeit ja
linksventrikuläre Hypertrophie		ja	ja			in dokumentierter Vorbeobachtungszeit ja
Systolic blood pressure	ja		nein		es wurde bisher kein überzeugender Vorteil für Diabetike mit/ohne kardiovaskuläre Erkrankungen gezeigt. Daher würden mir auch keine Grenzwerte einfallen.	nein (nur in DMP-Daten)
Diastolic blood pressure	ja		nein			nein (nur in DMP-Daten)
Stable angina	ja		nein		nur wenn previous History of myocardial infarction	in dokumentierter Vorbeobachtungszeit ja(?)
Niereninsuffizienz		ja	ja		bei schwerer NI ist die Ausschreidung von Glibenclamid (und den meisten SH) reduziert und daher Kontarindikation; bei z.B. Sitagliptin konkrete Dosisanpassungen möglich	in dokumentierter Vorbeobachtungszeit ja
Mild renal impairment	ja		nein			in dokumentierter Vorbeobachtungszeit ja
starke Stoffwechselinstabilität		ja	nein			???
Musculoskeletal and connective tissue disorders	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Gastrointestinal disorders	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Reproductive system and breast disorders	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Neoplasms	ja		nein			in dokumentierter Vorbeobachtungszeit ja
eGFR (estimated glomerular filtration rate)	ja		ja	< 30 ml/min		nein (nur in DMP-Daten)
UACR (urinary albumin-to-creatinine ratio)	ja					nein
Albuminurie		ja	nein		zu ungenau	in dokumentierter Vorbeobachtungszeit ja
subklinische Arteriosklerose		ja	nein			
Adipositas		ja	ja			fragliche Sensitivität, Dokumentation dürfte von bestimmten Erkrankungen/Therapien abhängen
schwere Leberinsuffizienz			ja		SH werden eher über die Leber abgebaut und sollten dann nicht gegebn werden, bei Gliptinen spielt das nicht so eine Rolle	in dokumentierter Vorbeobachtungszeit ja

Pankreatitis			ja		Pankreatitis ist eine schwere, aber seltene Nebenwirkung der DPP4-ijhibitoren und danach sollte es nicht mehr gegeben werden.	in dokumentierter Vorbeobachtungszeit ja
(Labor-)werte						gibt es nur bei DMP, nicht im W-DWH
Body weight	ja		nein		diese Werte sind kein Entscheidungskriterium für oder gegen einen der beiden Wirkstoffklassen	nein
BMI	ja		nein			nein (ggf. Adipositas-Grade, s.o.)
HbA1c (glycated haemoglobin)	ja		nein			nein
FPG (Fasting plasma glucose)	ja		nein			nein
2-h PPG (postprandial glucose)	ja		nein			nein
Triglycerides	ja		nein			nein
Co-Medikation	ja					
Glucose-lowering therapy	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Alpha-glucosidase inhibitor	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Glinide (meglitinide)	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Metformin total daily dose/mean dose	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Number of background glucose-lowering therapies (0,1,2 or 3)	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Antihypertensives	ja					in dokumentierter Vorbeobachtungszeit ja
ACE inhibitors	ja		nein			in dokumentierter Vorbeobachtungszeit ja
β-Blockers	ja		ja		Betablocker verschleiern die Symptome der Hypoglykämie v.a. bei SH, bei Gliptinen Hypoglykämien seltener, daher ist die Kombination weniger gefährlich	in dokumentierter Vorbeobachtungszeit ja
Diuretics	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Angiotensin receptor blockers	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Calcium antagonists	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Acetylsalicylic acid (aspirin)	ja		nein			in dokumentierter Vorbeobachtungszeit ja - ohne OTC
Statins	ja		nein			in dokumentierter Vorbeobachtungszeit ja
Lebensstil/Ernährung/Bewegungsmangel		ja				nein
familiäre/genetische Disposition		ja				nein
weitere von uns als wichtig erachtete Variablen:						
bisherige Einnahmedauer der aktuellen Add-on-Diabetesmedikation (DPP4 bzw. SU)			ja			in dokumentierter Vorbeobachtungszeit ja
bisherige Einnahmedauer von Metformin			ja	z.B. 1,2,…,>24 Monate	6 Monate, 12 Monate und länger	in dokumentierter Vorbeobachtungszeit ja
Anzahl der Arztbesuche und Krankenhausaufenthalte im letzten Jahr					pro Jahr macht Sinn, da meist Hxpoglykämien vorkommen. Auch Hospitalisierungen wegen kardiovaskulärer Ereignisse und Thrombosen/Embolien sind häufig. Interessant sind auch die Anzahl der Hospitalisierungen mit Diagnose Herzinsuffizienz.	in dokumentierter Vorbeobachtungszeit ja
Zeitpunkt des Einschlusses in die Studie (Quartal)						vermutlich SEHR WESENTLICH - ja (Daten vor Einführung der pauschalierten Versorgung nicht vergleichbar),Rückgang Sulfonylharnstoff, Anstieg DPP4

Table 7: Characteristics after overlap weighting and absolute standardized mean difference (ASD) between treatment and control group

Characteristic	**DPP4** (n = 18461.5)	**SU** (n = 18454.4)	**ASD**
heart failure = 1 (%)	3608.1 (19.5)	3607.5 (19.5)	<0.001
adipositas = 1 (%)	5873.7 (31.8)	5872.2 (31.8)	<0.001
betablocker = 1 (%)	11102.5 (60.1)	11097.0 (60.1)	<0.001
hypertension = 1 (%)	16548.6 (89.6)	16428.8 (89.0)	0.020
CAD = 1 (%)	2831.1 (15.3)	2827.4 (15.3)	<0.001
leftventricular hypertrophy = 1 (%)	483.4 (2.6)	484.0 (2.6)	<0.001
lipid disorder = 1 (%)	10989.3 (59.5)	10980.4 (59.5)	0.001
myocardial infarction = 1 (%)	663.0 (3.6)	662.9 (3.6)	<0.001
pancreatitis = 1 (%)	463.4 (2.5)	465.0 (2.5)	0.001
severe hypoglycemia = 1 (%)	34.7 (0.2)	34.9 (0.2)	<0.001
severe liver disease = 1 (%)	274.4 (1.5)	273.3 (1.5)	<0.001
severe renal insufficiency = 1 (%)	110.5 (0.6)	110.9 (0.6)	<0.001
age (median [IQR])	73.00 [69.00, 78.00]	73.00 [68.00, 78.00]	<0.001
sex = 1 (%)	9542.8 (51.7)	9531.9 (51.7)	0.001
cohort n (median [IQR])	10.00 [3.00, 18.00]	9.00 [3.00, 18.00]	<0.001
current DPP = 1 (%)	2033.5 (11.0)	2029.6 (11.0)	0.001
time DPP (median [IQR])	0.00 [0.00, 35.00]	0.00 [0.00, 0.00]	<0.001
current SU = 1 (%)	3862.1 (20.9)	3856.4 (20.9)	0.001
time SU (median [IQR])	0.00 [0.00, 532.00]	0.00 [0.00, 457.00]	<0.001
time metformin (median [IQR])	660.00 [288.00, 701.00]	657.00 [265.00, 700.00]	<0.001

Table 8: Unadjusted and propensity score weighted effect estimates and e-values for all outcomes and subgroup analysis

			Outcome Model (without any			unadjusted					adjusted for confounding via overlap weighting									
	total sample size	complete cases	family	link	Effect measure	estimate	std error	p value	exp(estimate)	95% CI of ratio	effective sample size	DPP4 median[IQR]/proportion(%)	SU median[IQR]/proportion(%)	estimate	std error	p value	exp(estimate)	95% CI of ratio	E-value	e value for limit of confidence interval closest to ratio=1
main analysis	171318	170626									36915.9	18461.5	18454.4							
number of outpatient visits+hospitalisations within 1 year			**negative binomial**	**logit**	**rate ratio**	**0.06**	**0.0024**	**<0.0001**	**1.06**	**[1.05,1.06]**		**15.00 [11.00, 20.00]**	**14.00 [10.00, 20.00]**	**0.03**	**0.0036**	**<0.0001**	**1.03**		**1.03**	**1.02**
1 year all-cause mortality risk			binomial	logit	OR	0.02	0.028	0.4546	1.02	[0.97,1.08]		680.2 (3.7)	717.3 (3.9)	-0.06	0.0389	0.1532	0.95	[0.88,1.02]	1.06	1.00
1 year risk of severe hypoglycemia (at least one diagnosis) while alive			binomial	logit	OR	-0.73	0.1126	<0.0001	0.48	[0.39,0.6]		35.4 (0.2)	46.9 (0.3)	-0.28	0.1677	0.0925	0.75	[0.54,1.05]	1.53	1.00
30 day risk of all-cause hospitalisation while alive			binomial	logit	OR	0.18	0.0235	<0.0001	1.19	[1.14,1.25]		1028.9 (5.6)	1074.4 (5.8)	-0.05	0.0321	0.1489	0.95	[0.9,1.02]	1.06	1.00
Sensitivity																				
with offset for time to death																				
number of outpatient visits+hospitalisations within 1 year			negative binomial	logit	rate ratio									0.03	0.0036	<0.0001	1.03	[1.02,1.03]	1.03	1.02
new users	70576	70283									19927.6	9964.2	9963.4							
number of outpatient visits+hospitalisations within 1 year			negative binomial	logit	rate ratio	0.04	0.0047	<0.0001	1.04	[1.03,1.05]		15.00 [10.00, 20.00]	14.00 [10.00, 20.00]	0.02	0.0049	0.0001	1.02	[1.01,1.03]	1.02	1.01
1 year all-cause mortality risk			binomial	logit	OR	-0.03	0.0474	0.4624	0.97	[0.88,1.06]		420.5 (4.2)	448.3 (4.5)	0.07	0.0483	0.1651	0.94	[0.85,1.03]	1.07	1.00
1 year risk of severe hypoglycemia (at least one diagnosis) while alive			binomial	logit	OR	-0.76	0.2049	0.0002	0.47	[0.31,0.7]		13.2 (0.1)	27.6 (0.3)	-0.68	0.2106	0.0012	0.51	[0.33,0.76]	5.51	1.49
30 day risk of all-cause hospitalisation while alive			binomial	logit	OR	-0.04	0.0387	0.2576	0.96	[0.89,1.03]		622.3 (6.2)	693.2 (7.0)	-0.12	0.0395	0.0034	0.89	[0.82,0.96]	1.14	1.04
Subgroups																				
age>80	25160	25126									5866.8	2932.4	2934.4							
number of outpatient visits+hospitalisations within 1 year			negative binomial	logit	rate ratio	0.07	0.0064	0.0001	1.07	[1.06,1.08]		14.00 [10.00, 19.00]	14.00 [10.00, 19.00]	0.02	0.0094	0.0678	1.02	[1,1.04]	1.02	1.00
1 year all-cause mortality risk			binomial	logit	OR	0.26	0.0435	<0.0001	1.30	[1.2,1.42]		318.2 (10.9)	294.6 (10.0)	0.09	0.0609	0.1532	1.09	[0.97,1.23]	1.10	1.00
1 year risk of severe hypoglycemia (at least one diagnosis) while alive			binomial	logit	OR	-0.7	0.2017	0.0005	0.49	[0.33,0.73]		10.9 (0.4)	14.4 (0.5)	-0.23	0.2975	0.4337	0.79	[0.44,1.42]	1.38	1.00
30 day risk of all-cause hospitalisation while alive			binomial	logit	OR	0.46	0.0504	<0.0001	1.58	[1.43,1.74]		257.4 (8.8)	223.3 (7.6)	0.16	0.0687	0.0239	1.17	[1.02,1.34]	1.21	1.00
heart failure	33368	33273									7282.1	3641.4	3640.6							
number of outpatient visits+hospitalisations within 1 year			negative binomial	logit	rate ratio	0.07	0.0058	0.0001	1.07	[1.06,1.09]		16.00 [11.00, 22.00]	16.00 [11.00, 21.00]	0.03	0.0082	<0.0001	1.03	[1.01,1.05]	1.03	1.01
1 year all-cause mortality risk			binomial	logit	OR	0.11	0.0459	0.0135	1.12	[1.02,1.23]		300.0 (8.2)	295.4 (8.1)	0.02	0.0613	0.7856	1.02	[0.9,1.15]	1.02	1.00
1 year risk of severe hypoglycemia (at least one diagnosis) while alive			binomial	logit	OR	-0.82	0.1758	<0.0001	0.44	[0.31,0.62]		16.1 (0.4)	24.4 (0.7)	-0.38	0.2465	0.1261	0.69	[0.42,1.11]	1.87	1.00
30 day risk of all-cause hospitalisation while alive			binomial	logit	OR	0.36	0.0435	<0.0001	1.44	[1.32,1.57]		359.2 (9.9)	318.7 (8.8)	0.13	0.0579	0.0231	1.14	[1.02,1.28]	1.17	1.02
severe hypoglycemia	361	359									63.0	31.4	31.6							
number of outpatient visits+hospitalisations within 1 year			negative binomial	logit	rate ratio	0.08	0.0734	0.2731	1.08	[0.94,1.25]		16.00 [9.00, 23.00]	15.00 [8.00, 20.00]	0.16	0.128	0.2246	1.17	[0.91,1.5]	1.21	1.00
1 year all-cause mortality risk			binomial	logit	OR	0.39	0.3603	0.2753	1.48	[0.73,3]		6.4 (20.4)	7.1 (22.6)	-0.13	0.4709	0.7788	0.88	[0.35,2.2]	1.16	1.00
1 year risk of severe hypoglycemia (at least one diagnosis) while alive			binomial	logit	OR	0.3	0.6573	0.6447	1.35	[0.37,4.91]		1.0 (3.2)	0.5 (1.6)	0.73	0.8118	0.3698	2.07	[0.42,10.18]	6.98	1.00
30 day risk of all-cause hospitalisation while alive			binomial	logit	OR	0.53	0.3719	0.1565	1.69	[0.82,3.51]		6.5 (20.6)	3.7 (11.6)	0.68	0.5025	0.176	1.98	[0.74,5.29]	5.75	1.00
renal	17718	17633									3515.0	1759.2	1755.8							
number of outpatient visits+hospitalisations within 1 year			negative binomial	logit	rate ratio	0.05	0.0079	<0.0001	1.05	[1.04,1.07]		17.00 [12.00, 23.00]	17.00 [12.00, 22.00]	0.03	0.0112	0.0087	1.03	[1.01,1.05]	1.03	1.01
1 year all-cause mortality risk			binomial	logit	OR	0.14	0.0755	0.0666	1.15	[0.99,1.33]		110.4 (6.3)	109.0 (6.2)	0.01	0.1006	0.9146	1.01	[0.83,1.23]	1.01	1.00
1 year risk of severe hypoglycemia (at least one diagnosis) while alive			binomial	logit	OR	-0.69	0.2657	0.0093	0.50	[0.3,0.84]		3.9 (0.2)	13.7 (0.8)	-1.17	0.3477	0.0008	0.31	[0.16,0.61]	54.78	2.74
30 day risk of all-cause hospitalisation while alive			binomial	logit	OR	0.45	0.0684	<0.0001	1.57	[1.37,1.8]		155.3 (8.8)	119.5 (6.8)	0.28	0.0924	0.0023	1.33	[1.11,1.59]	1.52	1.12

Häufigkeit Prozent Prozent Zeile Prozent Spalte	Tabelle von n_s nach prop_SU_s				
	prop_SU_s				
n_s	0	<=0.5	>0.5	1	Summe
1	7726 17.91 51.97 44.76	0 0.00 0.00 0.00	0 0.00 0.00 0.00	7141 16.56 48.03 48.91	14867 34.47
1-3	7122 16.51 46.06 41.26	1112 2.58 7.19 19.57	1182 2.74 7.65 21.15	6045 14.01 39.10 41.40	15461 35.84
4-5	1639 3.80 23.81 9.49	1892 4.39 27.49 33.30	2270 5.26 32.98 40.62	1082 2.51 15.72 7.41	6883 15.96
6-10	568 1.32 12.28 3.29	1974 4.58 42.69 34.74	1779 4.12 38.47 31.84	303 0.70 6.55 2.08	4624 10.72
11-20	93 0.22 9.35 0.54	569 1.32 57.19 10.01	303 0.70 30.45 5.42	30 0.07 3.02 0.21	995 2.31
21-50	22 0.05 11.00 0.13	127 0.29 63.50 2.24	51 0.12 25.50 0.91	0 0.00 0.00 0.00	200 0.46
51-100	18 0.04 66.67 0.10	6 0.01 22.22 0.11	3 0.01 11.11 0.05	0 0.00 0.00 0.00	27 0.06
>100	74 0.17 97.37 0.43	2 0.00 2.63 0.04	0 0.00 0.00 0.00	0 0.00 0.00 0.00	76 0.18
Summe	17262 40.02	5682 13.17	5588 12.96	14601 33.85	43133 100.00

Figure 14: Number of patients per physician (n_s) vs. probability to get prescribed SU (prop_SU_s)

Table 9: Protocol submitted for approval of ethics commission

Study protocol

Version: 1.1

Date: 07.06.2023

1. Study title

Harm reduction of switching from metformin plus sulfonylureas to metformin plus DPP4s in older adults: A target trial emulation using German routine claims data

2. Abstract

This study will evaluate the real-world benefit of switching elderly patients from potentially inappropriate medications (PIMs) to alternative medications using an example from diabetic treatment. More concretely, we apply the concept of target trial emulation to analyse the hypothetical intervention of prescribing only DPP4 inhibitors in addition to metformin in patients who are currently prescribed sulfonylurea as add-on diabectic medication. The hypothesis is that the intervention could reduce adverse events. In addition, we aim to identify heterogenous treatment effects regarding these PIMs. Thus, we aim to identify subgroups within the population of elderly patients that have an especially high risk for adverse events when taking the selected sulfonylurea. The study will use routine claims data from a German health insurance provider.

3. Responsibilities

Principal researcher:

Name	Institution	Contact
Paula Starke	Institut für Medizinische Statistik, Universitätsmedizin Göttingen, Göttingen; aQua Institut GmbH, Göttingen	paula.starke@stud.uni-goettingen.de

Participating researchers:

Name	Institution	Contact
Prof. Tim Mathes	Institut für medizinische Statistik, Universitätsmedizin Göttingen, Göttingen (principal investigator)	tim.mathes@med.uni-goettingen.de
Prof. Dr. Petra Thürmann	Lehrstuhl für Klinische Pharmakologie, Fakultät für Gesundheit, Universität Witten/Herdecke;	petra.thuermann@helios-gesundheit.de

	Helios Universitätsklinikum Wuppertal, Witten	
Dr. Thomas Grobe	aQua Institut GmbH, Göttingen	thomas.grobe@aqua-institut.de

4. Scientific background

The PRISCUS list 2.0 (updated in 2022) identifies medications that may be potentially inappropriate for use in patients over the age of 65 and should therefore be avoided. There is evidence that suggests that elderly patients that take a PIM instead of alternative substances have a higher risk for hospitalisations connected to adverse events[1]. Efforts to reduce inappropriate prescription behaviour could be bolstered up by more specific knowledge about the real-world effect of prescribing specific substances to specific groups of patients but further research is warranted to identify those PIMs that are especially responsible for the effect.

5. Project goals

We will prove the superiority of DPP4 inhibiors compared to Sulfonylurea as add-on treatment in older adults receiving metformin with regard to harms such as hospitalisations, the occurrence of severe hypoglycemia and mortality.

Apart from the question whether the entire population would profit from a treatment switch (average treatment effect), we also aim to determine subgroups that might be at higher risk for developing adverse events when taking PIMs.

6. Endpoints

Primary:

- 1 year number of all-cause hospitalisations and all-cause doctors visits (as composite outcome) while alive

Secondary:

- 30 day risk of all-cause hospitalisation while alive
- 1 year risk of severe hypoglycemia (at least one diagnosis) while alive
- 1 year all-cause mortality risk

[1] Endres HG, Kaufmann-Kolle P, Steeb V, Bauer E, Böttner C, Thürmann P (2016) Association between Potentially Inappropriate Medication (PIM) Use and Risk of Hospitalization in Older Adults: An Observational Study Based on Routine Data Comparing PIM Use with Use of PIM Alternatives. PLoS ONE 11(2): e0146811. doi:10.1371/journal.pone.0146811

7. Study population

Inclusion criteria: adults insured with BARMER health insurance

- that are older than 65 years
- are currently prescribed metformin

Time period: 2010-2018

Expected number of patients: 250 000 - 500 000

8. Methodology and Implementation

A retrospective cohort-type study will be conducted using routine claims data. The data analysis will follow the approach of "target trial emulation"[2] which offers a structured approach to use observational data to estimate the causal effect of an exposure on an outcome as if it were observed in a hypothetical randomized trial. A hypothetical RCT will first be designed and then emulated with the routine data as closely as possible.

9. Biometry

Given this is a retrospective, real-world analysis of all patients insured with BARMER who meet the inclusion criteria, a sample size calculation is not warranted[3].

A statistical analysis plan will specify the methods of causal inference that will be used. We will use generalized linear models for analysing the differences in harms between groups. For confounding adjustment,[4][5]. The heterogenous average treatment effects will be estimated using Baysian[6] , a nonparametric ensemble learning method that additively

[2] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758-64. doi: 10.1093/aje/kwv254. Epub 2016 Mar 18. PMID: 26994063; PMCID: PMC4832051.

[3] Hernán MA. Causal analyses of existing databases: no power calculations required. J Clin Epidemiol. 2022 Apr;144:203-205. doi: 10.1016/j.jclinepi.2021.08.028. Epub 2021 Aug 27. PMID: 34461211; PMCID: PMC8882204.

[4] Chatton, A., Le Borgne, F., Leyrat, C. *et al.* G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Sci Rep* **10**, 9219 (2020). https://doi.org/10.1038/s41598-020-65917-x

[5] Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. Statistical Methods in Medical Research. 2020;29(12):3721-3756. doi:10.1177/0962280220940334

[6] Wendling, T.; Jung, K.; Callahan, A.; Schuler, A.; Shah, N. H.; Gallego, B. (2018): Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. In: Statistics in medicine 37 (23), S. 3309–3324. DOI: 10.1002/sim.7820.

combines multiple decision trees. Each new tree is fitted to the residuals of the previous[7]

Data preprocessing will be done using SAS while we will use R to conduct the statistical analysis.

10. Data management and security

We will use data available in the BARMER scientific data warehouse (W-DWH). The detailed data protection concepts of the BARMER (**Appendix A**) and the BARMER W-DWH (**Appendix B**) including information on the technical, organisational and legal basis of the data collection, storage and scientific usage are appended (in German).

All datasets were pseudonymized by BARMER. Re-identification of an insured person can only by conducted by authorized BARMER employees. All data processing and analysis can only be conducted via secured remote access inside the W-DWH so that a linkage of data with other sources is not possible. The source files of the database can only be viewed but not accessed directly. Any download of datasets is prohibited and technically not possible. Results tables and aggregated data without personal references that are in the narrow sense connected to the research project can be downloaded from the W-DWH on a personal computer on request by an authorised person but only if the number of cases in each cell of the results table is large enough that results cannot be reconstructed for individual insured persons.

Available datasets include pseudonymised information on all persons insured by BARMER health insurance between 2005 and 2022 e.g. 1) demographics (longitudinal), 2) medical prescription data, 3) ambulatory data, 4) inpatient and outpatient hospital data, 5) therapeutic remedies and aids, 6) care data, 7) incapacity to work data, 8) dental data. I can access the database through a personal secured terminal. The data usage approval is part of the research partnership between BARMER and aQua GmbH as I use the same personal access to the W-DWH also for another project within the scope of my employment at aQua GmbH. The usage approval was extended to this master thesis by BARMER.

[7] Jennifer L. Hill (2011) Bayesian Nonparametric Modeling for Causal Inference, Journal of Computational and Graphical Statistics, 20:1, 217-240, DOI: 10.1198/jcgs.2010.08162 p. 223

Table 10: Approval letter of ethics comission

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

Ethik-Kommission der Universitätsmedizin Göttingen, Von-Siebold-Straße 3, 37075 Göttingen

Herrn
Prof. Dr. Tim Mathes
Institut für Medizinische Statistik
Humboldtallee 32 + 34
37073 Göttingen

Ethik-Kommission der
Universitätsmedizin Göttingen
Vorsitzender: Prof. Dr. Jürgen Brockmöller
Referentin
Regierungsrätin Doris Wettschereck
0551 / 39-68644 **Telefon**

Von-Siebold-Straße 3, 37075 Göttingen
Adresse
0551 / 39-61261 **Telefon**
0551 / 39-69536 **Fax**
ethik@med.uni-goettingen.de **E-Mail**
www.ethikkommission.med.uni-goettingen.de

per E-Mail: tim.mathes@med.uni-goettingen.de
Nachrichtlich an: paula.starke@stud.uni-goettingen.de

10.07.2023 br – fr - gö Datum

Antragsnummer:	**23/723 (bitte stets angeben)**
Studientitel:	Harm reduction of switching from metformin plus sulfonylureas to metformin plus DPP4s in older adults: A target trial emulation using German routine claims data
Antragsteller:	Prof. Dr. Tim Mathes, Institut für Medizinische Statistik, UMG
	Doktorand*innen: Paula Starke

Zur Begutachtung lagen vor:
E-Mail vom 22.06.2023
Anschreiben vom 21.06.2023
Zusammenfassung
Studienprotokoll gemäß der Checkliste für retrospektive Datenauswertung (Englisch) Version 1.1 vom 07.06.2023
Appendix A: Datenschutzkonzept der BARMER
Appendix B: Datenschutzkonzept des BARMER Wissenschafts-Datawarehouse (W-DWH)
Lebenslauf der Studienleiterin

Sehr geehrter Herr Prof. Dr. Mathes, sehr geehrte Damen und Herren,

die Ethik-Kommission der Universitätsmedizin Göttingen hat den oben genannten Antrag in der Sitzung vom 05.07.2023 beraten.

Die Ethik-Kommission hat keine ethischen oder rechtlichen Bedenken gegen das vorgelegte Studienvorhaben.

Bitte beachten Sie noch folgende Hinweise:
Bitte reichen Sie die Genehmigung zur Studiendurchführung von Prof. Dr. Mathes nach.

Es liegt in der Verantwortung der Antragsteller, die oben angeführten Hinweise vollständig umzusetzen. Ein weiteres Votum ergeht nicht.

Wir wünschen Ihnen viel Erfolg bei der Durchführung Ihres Projektes.

Unabhängig vom Beratungsergebnis macht die Ethik-Kommission darauf aufmerksam, dass die ethische und rechtliche Verantwortung für die Durchführung einer wissenschaftlichen Studie beim verantwortlichen Studienarzt und aller an der Studie beteiligten Ärzte liegt.

Alle Änderungen im Studienprotokoll müssen der Ethik-Kommission vorgelegt werden und dürfen erst nach der zustimmenden Bewertung umgesetzt werden.

Über alle schwerwiegenden unerwarteten unerwünschten Ereignisse, die während der Studie auftreten und die Sicherheit der Studienteilnehmer oder die Durchführung der Studie beeinträchtigen könnten, muss die Ethik-Kommission unterrichtet werden.

Der Abschluss/Abbruch der Studie ist mitzuteilen und ein Abschlussbericht vorzulegen.

Verification of examination registration in FlexNow

Name: Ms Paula Starke
Matriculation No.: 21662886

Semester: SoSe23
Degree Course: Angewandte Statistik (Master of Science)
Module: Masterarbeit
Exam: Masterarbeit
Lecturer: Jun.-Prof. Dr. Tim Mathes

Declaration

I hereby declare that I have produced this work independently and without outside assistance, and have used only the sources and tools stated.

I have clearly identified the sources of any sections from other works that I have quoted or given in essence.

I have complied with the guidelines on good academic practice at the University of Göttingen.

If a digital version has been submitted, it is identical to the written one.

I am aware that failure to comply with these principles will result in the examination being graded “nicht bestanden”, i.e. failed.

Göttingen, 13th December 2023 Paula Starke

Erklärung zur Nutzung von ChatGPT und vergleichbaren Werkzeugen im Rahmen von Prüfungen:

In der hier vorliegenden Arbeit habe ich ChatGPT wie folgt genutzt:

- [ ] gar nicht
- [ ] bei der Ideenfindung
- [ ] bei der Erstellung der Gliederung
- [ ] zum Erstellen einzelner Passagen, insgesamt im Umfang von ...
- [ ] zur Entwicklung von Software-Quelltexten
- [x] zur Optimierung oder Umstrukturierung von Software-Quelltexten (ausschließlich Korrektur-Tipps bei Fehlermeldungen in SAS, R und Latex)
- [ ] zum Korrekturlesen oder Optimieren
- [x] Weiteres, nämlich: Suche nach synonymen Formulierungen um häufige Wiederholungen von ähnlicher Satzstruktur und Worten zu vermeiden

Ich versichere, alle Nutzungen vollständig angegeben zu haben. Fehlende oder fehlerhafte Angaben werden als Täuschungsversuch gewertet.