For diseases assessed by disease symptoms it is difficult to distinguish whether interventions slow, stabilize, or reverse the disease, or only reduce symptoms. For example, when testing glucose-lowering drugs for delaying or preventing type 2 diabetes, reduced rates of diabetes diagnoses based on glycemic values do not directly answer this question, because glucose lowering reduces this surrogate without necessarily benefiting prediabetic individuals. A washout could evaluate whether effects persist after eliminating any direct glucose lowering. Several trials analyzed cumulative diabetes diagnoses, from a treatment period including a washout in not-yet-diagnosed subjects. This approach is severely biased, as demonstrated by simulations, because different misclassification errors occur unequally on drug and placebo as a result of the glucose-lowering effect during active treatment and the variability of glycemic values. An alternative is to analyze continuous end-of-washout values for all patients. This requires an imputation of glycemic values after diabetes diagnosis, which can no longer be observed without distortion by physician intervention. Valid imputation is possible, because the known diagnostic criteria lead to a missing at random situation. Trials with a washout in all subjects or delayed start designs are more efficient than current trial designs, and also minimize reliance on data imputation.

Key Words: Delayed start design; Deterministic missing at random; Disease prevention trial; Disease progression model; Multiple imputation; Washout for glucose-lowering drugs.

1. Introduction

The presence and worsening of many chronic progressive diseases, such as type 2 diabetes mellitus (T2DM), Parkinson’s disease, and Alzheimer’s disease, are primarily diagnosed and tracked by disease symptoms. The nature of the symptoms used for this purpose can vary widely from disease to disease. For example, in T2DM this would initially be elevated blood glucose values (American Diabetes Association 2010). In Parkinson’s disease, it would be reduced mental function, motor function, and ability to perform activities of daily living. In Alzheimer’s disease, the symptoms tracked are worsened memory and language skills, attention, and other cognitive and practical abilities (Ploeger and Holford 2009).

Commonly, interventions are initially developed to achieve short-term symptom improvements. If such an intervention additionally prevents or slows down further worsening of the disease, it would be preferable to other interventions, or it might make sense to use it at a disease stage when treatment of symptoms is not yet justified. However, it can be hard to distinguish short-term symptom improvements, which cease when treatment is stopped, from long-term symptom improvements that are a result of a disease-modifying effect and last beyond the end of treatment.

This becomes even harder when it is unethical to leave patients without additional symptomatic treatment in case of worsening symptoms. This article discusses the
issues arising in one such situation, namely, for trials that test whether a glucose-lowering drug delays or prevents T2DM.

1.1 Type 2 Diabetes Mellitus

During the progression toward T2DM, blood glucose levels rise increasingly above those seen in nondiabetics, because insulin becomes less effective at lowering blood glucose (insulin resistance) and because the body no longer releases enough insulin. Currently, the recommended method for the diagnosis of T2DM is based directly on biochemical measures of blood glucose levels (American Diabetes Association 2010). T2DM is associated with an increase in mortality as well as vascular, kidney, retinal, and neuropathic complications (Amos, McCarty, and Zimmet 1997). Randomized controlled clinical trials have shown that some of these consequences can be reduced or prevented by lowering blood glucose levels using drug therapy in addition to lifestyle measures, whether abnormal glucose levels are a consequence of T2DM (UK Prospective Diabetes Study Group 1998) or type I diabetes mellitus (The Diabetes Control and Complications Trial Research Group 1993).

The total number of people with diabetes has been predicted to more than double between 2000 and 2030 (Wild et al. 2004). Thus, it is intuitively attractive to avoid or delay the consequences of T2DM by preventing or delaying the progression to T2DM in high-risk individuals. A number of interventions for this purpose have been studied in clinical trials—primarily diet and exercise, as well as drugs normally used to lower blood glucose in patients already diagnosed with T2DM. Diet and exercise are by now widely accepted as the first-line intervention for individuals at risk of developing T2DM and have been shown to reduce the number of T2DM diagnoses in multiple randomized controlled trials (US Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008; American Diabetes Association 2010; Committee for Medicinal Products for Human Use 2012). Any new intervention is therefore expected to be evaluated with all trial participants receiving effective lifestyle counseling (US Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008; Committee for Medicinal Products for Human Use 2012).

1.2 Outcomes for T2DM Prevention Trials

If starting an intervention prior to the diagnosis of T2DM (as opposed to only starting the intervention after diagnosis) were to reduce diabetic complications, this would constitute a clear benefit to prediabetic individuals, whether the underlying disease mechanism is truly delayed or not. Thus, aiming to demonstrate a reduction in complications is one way to evaluate the value of interventions in prediabetic subjects. However, trials with diabetic complications as the primary outcome have to be of a very long duration for two reasons. First, most diabetic complications develop slowly. Second, it may well take a long time for the positive effects of any intervention to materialize to a meaningful extent. For example, the currently ongoing long-term extension of the Diabetes Prevention Program (DPP) aims for approximately 15 years of total follow-up to assess micro- or macrovascular diabetes complications. Results are expected in 2014, 18 years after the start of recruitment, making patient retention difficult: 22% of the original study participants were no longer seen 5 years into the long-term extension (The Diabetes Prevention Program Research Group 2009).

It is of interest to find other more practical means to demonstrate a delay or stopping of the underlying disease mechanism, which would be expected to reduce diabetic complications. However, there are currently no accepted surrogate outcomes that are not also potentially influenced by the intervention under study—particularly if it is a blood glucose-lowering drug. T2DM diagnosis and assessments of disease progression are currently based on surrogate endpoints such as fasting plasma glucose (FPG), 2 hr post-oral glucose tolerance test glucose (OGTT), or glycated hemoglobin (HbA1c) (de Winter et al. 2006; American Diabetes Association 2010), all of which respond to short-term blood glucose lowering. The time to, or the incidence of, new diabetes diagnoses based on these glycemic values exceeding diagnostic thresholds—often with a retest to confirm the result—has been used as an endpoint in T2DM prevention trials. This is an appealing approach, because this is the way T2DM is diagnosed in clinical practice, and these diagnostic thresholds were defined by their correlation with the risk of occurrence of diabetic complications (American Diabetes Association 2010). However, the correlation between having blood glucose values above the diagnostic thresholds and diabetic complications is not perfect (Wong et al. 2008). Subjects at a high risk for developing T2DM, based on glycemic values between normal and diabetic levels, are already at some risk of complications and thus might derive some benefit from glucose lowering. Consequently, the diagnostic thresholds for diabetes is simply the approximate point beyond which the benefits of reducing a patient’s blood glucose values are currently considered to outweigh the risks and costs of doing so by the medical community. Thus, to justify why individuals should receive any glucose-lowering intervention prior to this point, other effects of treatment beyond just lowered blood glucose values should be demonstrated (U.S. Department of Health and Human
1.3 What do We Mean by a Delay or Prevention of Diabetes?

Most drugs tested in T2DM prevention trials are known to lower glucose in diabetics and expected to also do so in subjects at high risk of developing T2DM. As a result, glucose-lowering drugs may simply delay the diagnosis (rather than the onset) of diabetes, by keeping glycemic parameters below the diagnostic thresholds for a time. Therefore, they mask the expected progressive worsening of glycemic values (U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008; Committee for Medicinal Products for Human Use 2012).

For this reason, it is important to clarify what is meant by a delay or prevention of T2DM. For this purpose, hypothetical scenarios similar to those described by Buchanan (2007) for the disease progression on treatment and after drug discontinuation are useful.

Figure 1 contrasts a range of possible treatment outcomes to the untreated disease progression. In scenario A, the drug has no additional benefit on top of its glucose lowering, while in scenario B, it has a negative impact on disease progression. In contrast, it is clear that scenarios C to G all leave treated individuals with a better glycemic state after the discontinuation of treatment, compared with if they had never been treated.

If one understands a prevention of diabetes as a modification of the underlying disease process to totally avoid the progressive worsening of glycemic values, then scenarios E, F, and G constitute a prevention of diabetes during treatment. In contrast, the drugs in scenarios C and D only slow down the worsening of glycemic values during treatment. This could mean that diabetes has either been delayed or prevented for an individual, depending on their initial glycemic status and remaining life expectancy. Since even substantially slowing the worsening of glycemic control delays the development of diabetes and subsequent diabetic complications, either outcome would benefit treated individuals.

Thus, a logical first step for the evaluation of a glucose-lowering drug for the delay or prevention of diabetes is to establish whether it at least provides a benefit in this sense. This will be a key focus of this article, while further distinctions based on the extent to which diabetes is delayed or totally prevented will only be briefly touched upon. These are nevertheless of interest for risk-benefit and cost considerations, even though they are challenging to make due to regression to the mean following the inclusion into a trial based on glycemic values, as well as the initial effects of the mandated lifestyle intervention. For example, scenario D is more desirable than scenario C, because diabetes would be delayed for longer.

However, if lifelong treatment is foreseen, a distinction between scenarios F and E is of limited value. If no further disease progression occurs despite treatment having been stopped, as in scenario G, then this would be particularly attractive, because only a temporary course of treatment would be needed rather than a possibly lifelong treatment.

1.4 Regulatory Standards for Obtaining an Indication for Preventing or Delaying Diabetes

No drugs have been approved in the United States or Europe for an indication of preventing or delaying diabetes. However, we understand from the current regulatory guidance documents (U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008; Committee for Medicinal Products for Human Use 2012) that the FDA and the European Medicines Agency (EMA) foresee the possibility of such an indication, if scenarios A and B have been excluded in favor of scenarios C to G. Both the draft FDA guidance for industry on developing drugs and therapeutic biologics for treatment and prevention of diabetes mellitus (U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008) and the European guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus (Committee for Medicinal Products for Human Use 2012) suggest that for glucose-lowering drugs, there should still be a difference compared to placebo in the proportion of patients meeting diabetes criteria after treatment is stopped during
a washout. This would then be accepted by the FDA and EMA as a surrogate for a modification of the underlying disease progression (U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008; Committee for Medicinal Products for Human Use 2012), while the extent of the delay of diabetes would presumably play a role in risk benefit considerations. Due to the difficulty of judging the clinical relevance of a delay of diabetes, the recent European guideline additionally seeks evidence showing benefits in terms of diabetic complications to support the clinical relevance of a reduced rate of T2DM diagnoses (Committee for Medicinal Products for Human Use 2012). However, it is not clear what level of evidence would be required by the European regulators in this respect. The option to use a glycemic data-based surrogate outcome would presumably be meaningless, if the evidence about diabetes complications would have to be sufficiently compelling for an approval on that basis alone. No experience yet exists about how European regulators apply this recently approved guideline in practice.

1.5 Current Study Designs and Analysis Approaches for Diabetes Prevention Trials

Appropriate analysis methods for surrogate data from diabetes prevention trials with a washout have not yet been systematically investigated and are not discussed in the existing health authority guidance documents. This may be because the FDA and EMA guidance documents (U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research 2008; Committee for Medicinal Products for Human Use 2012) possibly envision a trial where at the end of the treatment period, all initial trial participants, irrespective of whether they have already been diagnosed as diabetic or not, stop all trial medications and all other concomitant glucose-lowering drugs for long enough to eliminate all direct glucose-lowering effects of these drugs. In this case, glycemic values could quite simply be compared between treatment groups, either in terms of groups’ means or some dichotomization.

However, all trials will have missing data, and to date T2DM prevention trials have only entered subjects not yet diagnosed as diabetic into a washout. This was presumably due to ethical concerns about taking diabetic patients off their glucose-lowering medications, and a lack of appreciation of the serious interpretational problems that arise when drawing inferences from a washout in a subset of subjects that is selected based on on-treatment data—likely with a different extent of losses to follow-up in each treatment group. The only analysis approach that has been used in these trials is a comparison of the total cumulative incidence of diabetes diagnoses across both the treatment and the washout period between test and control group.

In Section 2 of the article, this analysis approach is appraised using the completed DPP as an example (Knowler et al. 2002; The Diabetes Prevention Program Research Group 2003) and further evaluated using simulations in Section 3.

Buchanan has retroactively applied another analysis approach to several T2DM prevention trials (Buchanan 2007). He suggests that a delay or prevention of diabetes has occurred, if the risk difference for diabetes diagnoses between treatment and control group continues to diverge in favor of treatment for each year of active treatment. However, this effectively assumes that differences in glycemic values between groups translate proportionally into risk differences between the groups. This is not a reasonable assumption. Diverging risk differences could just as easily be the result of, for example, the expected increase of glycemic values over time pushing increasingly many placebo patients over the threshold for diagnoses, while the glucose-lowering effect in the drug group results in a lesser number of additional diagnoses per year in the treatment group. Nevertheless, his concept when applied to continuous glycemic values underlies the discussion of what we mean by a delay or prevention of diabetes.

An alternative analysis method that uses the continuous end-of-washout glycemic values is proposed in Section 4 and evaluated using simulations.

2. The Diabetes Prevention Program Example

2.1 Overview of DPP Results, Analyses, and Conclusions

The DPP will be used as an example of a randomized double-blind placebo-controlled trial of a glucose-lowering drug (metformin) for the prevention of diabetes in subjects at high risk of developing T2DM defined as having a body mass index of at least 24, an FPG of 5.3–6.9 mmol/l (95–125 mg/dl), and a plasma glucose of 7.8–11.0 mmol/l (140–199 mg/dl) 2 hr post-OGTT. It is the largest such study in which a washout for the not yet diabetic subjects was conducted (Knowler et al. 2002; The Diabetes Prevention Program Research Group 2003) and fully published.

Over an average follow-up of 2.8 years, 225 out of 1073 (21%) metformin subjects and 304 out of 1082 (28%) placebo subjects (odds ratio 0.66; 95% confidence interval 0.54–0.82) were diagnosed as diabetic based on either an FPG ≥ 7.0 mmol/l (126 mg/dl) or a plasma glucose ≥ 11.1 mmol/l (200 mg/dl) 2 hr post-OGTT.
In either case, the diagnosis had to be confirmed by a retest, usually within 6 weeks (Knowler et al. 2002). However, after a 1–2 week washout in those not yet diagnosed as diabetic, the glycemic values increased more on metformin than on placebo relative to just before the washout. The difference between metformin and placebo was 0.23 mmol/l (4.1 mg/dl) for FPG and 0.08 mmol/l (1.4 mg/dl) in terms of the OGTT. Consequently, 48 out of 688 (7%) metformin subjects that entered the washout were diagnosed as diabetic during the washout (The Diabetes Prevention Program Research Group 2003). As to be expected with a glucose-lowering drug, this suggests that at least some of the reduction in diabetes diagnoses seen on treatment was due to a direct glucose-lowering effect.

The DPP research group synthesized these results by recalculating the odds ratio for the main trial considering washout diabetes diagnoses as events. Ignoring the adjustment for year of randomization employed by the DPP research group, which can only be taken into account with access to the trial data, the original core trial odds ratio

\[
\frac{D_{\text{met}} \times (N_{\text{pbo}} - D_{\text{pbo}})}{D_{\text{pbo}} \times (N_{\text{pbo}} - D_{\text{pbo}})} = \frac{[225]/(1073 - 225)]/[304]/(1082 - 304)] = 0.68 (95\% CI 0.55 - 0.83)
\]

was updated taking into account the 48 new washout diabetes diagnoses in the metformin group and the 30 new diagnoses in the placebo group in the following way:

\[
\frac{(D_{\text{met}} + DW_{\text{met}}) \times (N_{\text{pbo}} - D_{\text{pbo}} - DW_{\text{pbo}})}{(D_{\text{pbo}} + DW_{\text{pbo}}) \times (N_{\text{pbo}} - D_{\text{pbo}} - DW_{\text{pbo}})} = \frac{[225 + 48]/(1073 - 225 - 48)]/[304 + 30]/(1082 - 304 - 30)] = [273/800]/[334/748] = 0.76 (95\% CI 0.63 - 0.93).
\]

Based on this analysis and because the difference between 0.76 and 0.68 is 25% of 1.0–0.68, the DPP research group asserted that “it appears that approximately one-quarter of the beneficial effect of metformin to prevent type 2 diabetes in the DPP was attributable to a pharmacological effect that did not persist when the drug was withdrawn” and concluded that the overall relative reduction in the odds of developing diabetes with metformin “remained a substantial 25% after withdrawal” (The Diabetes Prevention Program Research Group 2003, p. 979).

### 2.2 Potential Issues With the DPP Analysis and Its Interpretation

The way in which the DPP was analyzed is biased in favor of metformin in a number of ways. First, a considerable proportion of subjects that were not yet diagnosed as diabetic died or failed to complete the trial, enter the washout, or complete the washout (22%). While such subjects were reasonably well balanced between the metformin and the placebo group, the problem is that this prevents the washout from revealing such subjects as diabetic, if they were undiagnosed due to the glucose-lowering effects of metformin. No correction for this—such as scaling up the number of cases in the washout anti-proportionally to the proportion of eligible subjects that entered the washout—was made.

Second, the natural uncertainty in the diagnosis of diabetes not only results in missed diabetes diagnoses during the core study, but also means that the washout could not identify all the missed diagnoses. One would expect more of these on metformin as compared to placebo, due to the glucose-lowering effect of metformin during the treatment period. Thus, it is likely that more diabetics remained undiagnosed in the metformin group despite the washout period.

Finally, uncertainty in the diagnosis of diabetes leads to some subjects being falsely or prematurely diagnosed as diabetic. During the treatment period, this would have occurred more frequently in the placebo group than in the metformin group due to the glucose-lowering effects of metformin. Subjects misdiagnosed as diabetic never entered the washout and thus this bias is not addressed by the presence of the washout period. It is not immediately clear whether the requirement of a retest for the diagnosis of diabetes alleviates this problem to some extent. On the one hand, it reduces misdiagnoses in the placebo group, but on the other hand, it also increases the number of metformin subjects not diagnosed as diabetic due to the glucose-lowering effect of the drug.

### 3. Simulation Study to Quantify Bias in DPP Style Analysis

A DPP style analysis is subject to several biases, but their severity is not immediately obvious. Simulations were used to evaluate the extent of the resulting Type I error rate inflation—in the sense of wrongly claiming a disease-modifying effect—when the test intervention is a glucose-lowering drug that does not otherwise delay or prevent T2DM, which is being compared versus a placebo.

#### 3.1 Simulation Setup

For the main simulation scenario I that attempts to imitate the DPP as closely as possible, it was assumed that

- FPG increases on average linearly by 0.1 mmol/l per year in placebo subjects based on the FPG
trajectory in the placebo group of the DPP (Knowler et al. 2002).

- The test drug does not change the slope of this disease progression during treatment nor during the washout.
- The test drug shifts FPG downward by a constant 0.23 mmol/l based on the FPG increase observed on metformin compared to placebo after the DPP washout (The Diabetes Prevention Program Research Group 2003).

Based on up to 5.5 years of data from the placebo group of another T2DM prevention study (The NAVIGATOR Study Group 2010a, 2010b) including data post-diabetes diagnosis, the correlation in FPG between the 6-month visit appears to decrease the larger \( |i−j| \), but remains nonzero. Over the limited study duration, a correlation of the form \( a = b \times |i−j| \) for \( i \neq j \) with \( a = 0.48 \) and \( b = 0.037 \) appeared to approximate the correlation matrix reasonably well. In contrast, a correlation structure of the form \( c + d^2 |i−j| \) or \( d^2 |i−j| \) did not fit the observed correlation matrix. During the first 3 years of the study, the variance appeared to be relatively constant around 0.68, but there was some indication of an increasing variability in years 3–6. For simulation scenarios, in which a lower variability was assumed, a random subject effect was added to maintain the level of variability at baseline to ensure a distribution of baseline values across the whole inclusion range for FPG values from 5.3 to < 7 mmol/l.

For each scenario, 1000 trials were simulated, each with 1000 subjects per arm with observed baseline FPG values between 5.3 and <7 mmol/l matching the FPG inclusion criteria of the DPP. Trial duration was assumed to be 3 years, with FPG assessed at baseline and every 6 months thereafter, as well as at the end of the trial after a washout of sufficient length to eliminate all direct glucose-lowering effects.

Diabetes was considered to have been diagnosed when FPG \( \geq 7.0 \) mmol/l, confirmed by a retest performed within a month to mimic the DPP criteria. Additional diagnostic criteria have come into use since the design of DPP (American Diabetes Association 2010), but the general issues investigated in this simulation should hold equally for these. Analysis was done in the same fashion as for the primary analysis of the DPP washout using the one-sided 2.5% significance level. The main reference scenario 1 is intended to assess the biases in the DPP style analysis introduced solely by the combination of direct glucose lowering and variability. Thus, it assumes no losses to follow-up during the main study and that all subjects complete the washout.

### 3.2 Simulation Results for a DPP Style Analysis

As can be seen from the simulation results for scenario 1 in Table 1, using a DPP style analysis in the

| Scenario | No. of visits \( y \) | Slope \( /\text{year}^a \) | FPG effect \( b \) | \( \sigma^2 \), \( /\text{visit}^c \) | Lost \( /\text{washout}^d \) | Not in FPG retest \( /\text{washout}^e \) | One-sided Type I errors (%) \( (95\% \text{ confidence interval}) \) | Median odds ratio |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|------------------|------------------|
| 1        | 6              | 0.1            | 0.23           | 0.68           | None           | None           | Yes              | 99.3 (98.6 to 99.7) | 0.63             |
| 1a       | 6              | 0.1            | 0.23           | 0.68           | None           | None           | No               | 99.8 (99.3 to 100%) | 0.65             |
| 2        | 6              | 0.1            | 0              | 0.68           | None           | None           | Yes              | 3.2 (2.1 to 4.4)   | 0.99             |
| 2a       | 6              | 0.1            | 0.23           | 0              | None           | None           | Yes              | 2.3 (1.4 to 3.2)   | 1.00             |
| 3        | 6              | 0.1            | 0.23           | 0.68           | 2%             | None           | Yes              | 99.6 (99.0 to 99.9) | 0.63             |
| 4        | 6              | 0.1            | 0.23           | 0.68           | 10%            | None           | Yes              | 99.6 (99.0 to 99.9) | 0.63             |
| 5        | 6              | 0.1            | 0.23           | 0.68           | 20%            | None           | Yes              | 99.6 (99.0 to 99.9) | 0.63             |
| 6        | 6              | 0.1            | 0.23           | 0.68           | 2%             | 20%            | Yes              | 99.3 (98.6 to 99.7) | 0.63             |
| 7        | 6              | 0.1            | 0.23           | 0.68           | 2%             | 20%            | Yes              | 40.4 (37.3 to 43.5) | 0.83             |
| 8        | 6              | 0.1            | 0.23           | 0.68           | 2%             | 20%            | Yes              | 63.1 (60.0 to 66.1) | 0.81             |
| 9        | 6              | 0.02           | 0.68           | None           | None           | None           | Yes              | 99.2 (98.4 to 99.7) | 0.59             |
| 10       | 6              | 0.25           | 0.23           | 0.68           | None           | None           | Yes              | 98.4 (97.4 to 99.1) | 0.67             |

**NOTES:** Based on 1000 simulated trials with 1000 subjects in each treatment group. Simulations were performed using SAS/STAT® software, version 9.3 of the SAS System for Linux on a cluster of Linux nodes. Copyright © 2002–2010 SAS Institute Inc., Cary, NC, USA.

- \(^a\)With visits every half-year a 3-year trial has six post-baseline visits, 1.5 \( y = 3 \) visits and 0.5 \( y = 1 \) visit.
- \(^b\)Amount by which FPG values increase per year.
- \(^c\)Assumed direct glucose-lowering effect on FPG in the active treatment group.
- \(^d\)Multivariate normal distribution with covariance matrix \( \text{cov}(\text{FPG}, \text{FPG}) = \sigma^2 (\sigma^2 = 0.48 − 0.037 \times |i−j|) \) for \( i \neq j \); \( N(0, 0.68 − \sigma^2) \) random subject effect added to mean baseline of 6.1 mmol/l when \( \sigma^2 < 0.68 \).
- \(^e\)Percentage of subjects lost at each post-baseline visit during active treatment.
- \(^f\)Percentage of subjects that complete active treatment, but do not enter or complete the washout.
- \(^g\)Whether a retest for FPG values \( \geq 7 \) mmol/l is conducted or not.
presence of both a direct glucose lowering and variability in FPG values inflates the Type I error rate of wrongly claiming a disease-modifying effect close to 100%, whether a retest for confirming the diagnosis is required or not (scenario 1a).

If there were either no direct glucose-lowering effect of the test drug (scenario 2) or no variability in the data other than different baseline levels for each patient (scenario 2a), as well as no losses to follow-up during the treatment period and 100% washout completion, then as expected the Type I error rate was approximately at the nominal level.

Losses to follow-up that occur completely at random as simulated in scenarios 3 to 6 do not change that the Type I error rate is close to 100%. Among the scenarios that explored the effect of the assumed level of variability, scenario 7 with the lowest assumed variance had the lowest, but still substantially inflated, Type I error rate. However, the relationship between variability in glycemic values and type error rate appears to be nonmonotonic, because scenario 8 with the highest tested variance had the second lowest Type I error rate. That is, increased variability in glycemic values increases the Type I error rate only up to a certain point, but the Type I error rate remained substantially inflated above the significance level in all scenarios.

The effect of the background disease progression rate on the Type I error rate appears to be small. Simulations under a first-order autoregressive correlation structure also resulted in substantial Type I error rate inflations and are not shown.

4. Alternative Analysis Approaches

4.1 Comparing End-of-Washout Values

One alternative analysis approach is to perform a washout for all trial and nontrial glucose-lowering medications in all patients, both in those not yet diagnosed as diabetic, as well as in those already diagnosed as diabetic. If this is feasible and data are otherwise missing completely at random, then a direct comparison of either the continuous glycemic values at the end of the washout or of the proportion of patients with glycemic values meeting diabetes diagnosis criteria would be possible.

A hypothetical trial outcome might be that after 3 years of treatment FPG on average increased from a baseline of 5.90 mmol/l to 6.20 mmol/l in the placebo group and decreased to 5.82 mmol/l in the treatment group, while after a 3-month washout in all patients FPG was 6.23 mmol/l in the placebo group and 6.08 mmol/l in the treatment group with a treatment difference of 0.15 mmol/l. If these results were statistically sufficiently compelling, it would demonstrate a treatment effect beyond just direct glucose lowering. Assuming that no regression to the mean or initial improvement due to lifestyle advice occurred, one would conclude that the treatment prevented approximately half of the worsening of glycemic values on placebo over a 3-year period.

4.2 Considerations About Data Post-Diabetes Diagnosis

Problems with the analysis of end-of-washout values arise, if a washout cannot be performed in those already diagnosed as diabetic, or if the trial has already been performed with a washout only in those not yet diagnosed as diabetic—as in the case of the DPP.

In that case, glycemic values undistorted by medical intervention will not be available after patients are diagnosed as diabetic. Even if they are observed, they would likely be confounded by the use of concomitant anti-diabetic medications or any other physician intervention in reaction to the diabetes diagnosis. For this reason, glycemic values after the initiation of a glucose-lowering drug—and perhaps already after the diagnosis of diabetes—need to be considered effectively missing as far as the availability of a nonconfounded assessment of disease progression is concerned.

Even when conducting a washout in all patients, it could be that patients would be more (or less) likely to withdraw consent or be lost to follow-up after being informed of a diabetes diagnosis. Thus, both data censored at the time of diabetes diagnosis and unobserved data cannot be assumed to be missing completely at random, because we expect patients diagnosed as diabetic to have higher future off-treatment glycemic values than those not diagnosed as diabetic.

Murray and Findlay (1988) gave a similar hypertension example, which is quoted in the National Academy of Sciences report on the prevention and treatment of missing data in clinical trials (National Research Council 2010). They pointed out that considering data after exceeding a threshold as missing leads to a deterministic missing at random (MAR) situation (Rubin 1976), because the observed preceding glycemic values that lead to the diabetes diagnosis determine that the data are considered missing.

Thus, if diabetes diagnoses are based on only a single glycemic variable, appropriate likelihood methods can be used to analyze this single glycemic variable on its own (Kenward and Molenberghs 1998). However, this is not necessarily the case, because diabetes diagnoses may be based on multiple glycemic variables (e.g., HbA1c, FPG, glucose 2 hr after an oral glucose tolerance test). In that case, all preceding data that could enter the diabetes diagnosis must be taken into account to justify an MAR assumption. This follows the recommendations of Liu and Gould (2002) to use any post-baseline information in the
imputation of missing data that is potentially informative about why data are missing, to come as close as possible to a true MAR situation.

### 4.3 Implementation of Imputation Approaches to be Tested

In our simulations, we imputed data separately in each treatment group using multiple imputation (Rubin 1976; Little and Rubin 1987) taking all available post-baseline data (in the case of these simulations only FPG data) into account, followed by a subsequent analysis of the FPG values alone. Multiple imputation was performed using a Markov chain Monte Carlo (MCMC) approach assuming a joint normal distribution of glycemic values and implemented using the SAS procedure PROC MI. Parameter estimates obtained using the expectation–maximization algorithm were used as the initial values for the imputation-posterior method (Tanner and Wong 1987) to create 50 independent Markov chains that converge in distribution to the conditional distribution given the observed data of the unobserved or censored values, and distribution parameters (Schafer 1997, 72ff; SAS Institute Inc. 2011, pp. 4595–4599). For each of the 50 imputation samples, the end-of-washout values were then compared between the two treatment groups and aggregate estimates obtained. Program code is provided as online supplementary material.

Additionally, a last prediabetes diagnosis value carried forward analysis was implemented, because researchers might be tempted to use such an easy-to-implement approach.

### 4.4 Simulations

Table 2 shows the results of simulations conducted to evaluate the performance of the proposed alternative analysis approaches. The alternative analysis methods were applied to the same simulated 1000 trials for each of the scenarios, on which the DPP-style analysis was evaluated. Two additional modified versions of scenarios 1 and 6, in which the treatment actually reduced the slope of glycemic values by 0.05 mmol/l/year during active treatment, but not during the washout, were added and are designated scenarios 1b and 6b.

All methods performed well in scenario 2, where no glucose-lowering effect is present and thus, no bias is expected to arise.

The simplistic last prediabetes diagnosis value carried forward analysis resulted in a clear inflation of the Type I error rate in all scenarios with a glucose-lowering effect, but was particularly strongly biased when end-of-washout assessments were missing for many patients as in scenarios 3 to 6.

Simulated Type I error rates when using multiple imputation followed by comparing end-of-washout values between treatment groups did not exceed the nominal significance level by more than what could be explained by the expected variability in simulation outcomes.

This was the case for both simulated studies where all patients, or only those not diagnosed as diabetic, entered a washout. However, when conducting a washout in all patients, power was substantially higher than when only not yet diabetic patients entered the washout (99% vs. 60% for 1b and 96% vs. 51% for 6b). Analyzing end-of-washout values dichotomized as diabetic or nondiabetic instead of continuous data substantially reduced power.

### 5. Discussion

The most conclusive approach for demonstrating a benefit to prediabetic individuals is to conduct a large long-term trial to show that treatment with a drug prior to...
the diagnosis of diabetes reduces future diabetic complications. However, such a trial poses enormous practical challenges, as illustrated by the duration of and extent of losses to follow-up in the DPP long-term extension. As a result, researchers have been drawn to shorter, but harder to analyze and interpret, trials using surrogate outcomes based on glycemic values.

The previously used analysis methods for claiming at least a delay of T2DM for a glucose-lowering drug, based on the cumulative incidence of diabetes diagnoses, are a drastic example of between-group comparisons becoming biased when misclassifications of a binary outcome based on continuous data occur unequally between groups—something that Cochran already warned about over 40 years ago, when he wrote that in such a setting there is an “increased risk that an entirely spurious relationship is found to be statistically significant, as several writers have warned” (Cochran 1968, p. 647). Some issues related to the DPP—such as the length of the washout and the laboratory tests measured—have been previously critiqued (Buchanan 2003a; Buchanan 2003b; Scheen 2003), but the biases discussed in this article appear to have been ignored so far. As our simulations indicate a DPP style analysis for a drug that only lowers glucose and does not affect disease progression is likely to produce results similar to those seen for metformin in the DPP.

Comparing continuous end-of-washout values, as proposed in this article, is a more suitable alternative. Appropriately dealing with the glycemic data post-diabetes diagnosis that will be distorted by medical intervention is the key challenge in doing so. The observation that this constitutes an MAR situation motivates the proposed multiple imputation using an MCMC method. Our simulations indicate that, unlike the investigated alternatives, this approach controls the Type I error rate for claiming at least a delay in T2DM, if the glycemic data follow a joint normal distribution as in our simulations.

An equivalent approach could be to jointly model the different post-baseline variables separately in each treatment group while assuming neither a specific correlation structure between variables or visits nor a specific time trend. Both approaches are easy to implement with commonly available statistical software packages and have the same assumptions, but the multiple imputation approach offers more scope for sensitivity analyses, for example, by modifying the imputed data.

One could also consider applying Buchanan’s approach (Buchanan 2007) mentioned at the end of the introduction and illustrated in Figure 1 to the glycemic values over time after imputing data as proposed in this article. If a drug only lowers glucose, but does not otherwise affect the underlying disease process, then one might expect to see two parallel curves that are separated by a constant distance reflecting the size of the direct glucose-lowering effect. If on the other hand, the drug affects the underlying disease process, then the curves should diverge. Disease progression models aiming to detect such slope differences have already been used in other disease areas (Chan and Holford 2001; Tashkin et al. 2008) and as a sensitivity analysis in the ADOPT trial, which studied disease progression in already diabetic patients on different glucose-lowering drugs (Kahn et al. 2006). Such models were described for the Parkinson’s disease setting by Bhattacharam et al. (2009). One attraction is that these may also offer a way of analyzing previous trials that were conducted without a washout. Additionally, if there is more than one assessment during the washout, slope changes after the end of treatment can be investigated in this fashion to distinguish scenarios such as E, F, or G in Figure 1.

However, there are some difficulties when interpreting disease progression models, which become even harder to resolve in the absence of a washout. Additional evidence would be needed to distinguish a change in the rate of disease progression from a direct pharmacological glucose-lowering effect that becomes larger as the disease progresses. Without a washout, the glucose-lowering effect is estimated solely across subjects and one is unable to investigate nonslope-changing treatment effects that would have persisted after a washout.

Our simulations also indicate that in a trial with the same sample size, duration and design as the DPP, the proposed analysis of partially imputed end-of-washout data does not achieve an acceptable power. While trials with substantially larger sample sizes than the DPP have already been conducted, it would be more efficient to simply reduce the number of patients not entering the washout.

Our simulations show that, if end-of-washout assessments are available for all trial subjects, a trial with the DPP sample size would be adequately powered to detect a halving of the rate of worsening of glycemic values. Additionally, the reduced extent of imputation minimizes the reliance of the analysis result on the assumptions underlying the proposed imputation, such as that of a joint normal distribution of glycemic values including for the unobserved values after the diagnosis of diabetes. This would also reduce concerns around the proposed analysis not being an intention-to-treat approach, in the sense that it does not directly analyze post-diabetes diagnosis values to the extent that they are available. However, in line with Lewis and Machin (1993) we feel that such a narrow definition of intention-to-treat does not address the question of interest, because post-diabetes-diagnosis glucose values will be distorted by physician intervention.

One idea to obtain an end-of-washout assessment in as many patients as possible would be to attempt to overcome the ethical issues around conducting a washout.
in all participants—even if they are already diabetic—for example, by making rescue medication available during the washout. In such a trial, one would either compare end-of-washout values with imputation for only those requiring rescue medication during the washout, consider rescue medication as a worse outcome than the worst observed end-of-washout value akin to the approach of Gould (1980), or use a composite outcome that counts both those meeting diagnostic criteria at the end of the washout and those requiring rescue medication during the washout as events.

Another trial design with a similar aim would be to stop all study and nonstudy glucose-lowering drugs and start all subjects in all treatment groups on an identical glucose-lowering therapy. If the identical therapy given to all patients is the test drug, then this is a delayed start design (D’Agostino 2009). Delayed start designs have so far primarily been used to study disease progression in other disease areas such as Parkinson’s disease (Olanow et al. 2009), but were also proposed by Garg (2007) to clarify the effect of rosiglitazone on disease progression of already diabetic patients in the ADOPT trial (Kahn et al. 2006). This approach seems promising, because compared to a washout it appears less likely that subjects would start rescue medication or need to continue rescue medication once the identical glucose-lowering therapy has been started in all patients. Thus, it would become possible to directly observe data in a greater proportion of patients. As a result, Ploeger and Holford’s (2009) conclusion that when fitting a disease progression model a washout design leads to a higher power than a delayed start design may not apply for the T2DM prevention setting, even if that conclusion were otherwise applicable to the analysis method proposed in this article.

6. Conclusions

If a glucose-lowering drug is compared to placebo to show that it prevents or delays progression to T2DM, then it is invalid to draw inferences about such disease-modifying effects based only on cumulative diabetes diagnosis rates. This is irrespective of whether data from a washout period are included or whether retests were required to confirm the diagnosis.

Imputing continuous glycemic values post-diabetes diagnosis under an MAR assumption should be preferred and allows a comparison of end-of-washout values. When conducting such an analysis, study designs with a washout that includes all trial participants or delayed start designs provide a substantially higher power and at the same time rely less heavily on imputation. Practical experience with these approaches is currently lacking in the T2DM prevention setting.

Acknowledgments

The author thanks Sheila Dickinson for her suggestion of a washout that allows rescue medications, Georgina Bermann for a number of useful suggestions regarding disease progression models, and the anonymous reviewers for helping to make this a more focused manuscript.

Supplementary Materials

Online supplementary material contains the SAS program code to perform the proposed analysis approach with imputation of end-of-washout values using multiple imputation.

References

American Diabetes Association (2010), “Standards of Medical Care in Diabetes—2010,” Diabetes Care, 33, S11–S61. [104,105,109]
Amos, A. F., McCarty, D. J., and Zimmern, P. (1997), “The Rising Global Burden of Diabetes and its Complications: Estimates and Projections to the Year 2010,” Diabetic Medicine, 14, 57–585. [105]
Bhattaram, V. A., Siddiqui, O., Kapcala, L. P., and Gobburu, J. V. S. (2009), “Endpoints and Analyses to Discern Disease-Modifying Drug Effects in Early Parkinson’s Disease,” The AAPS Journal, 11, 456–464. [112]
Buchanan, T. A. (2003a), “Prevention of Type 2 Diabetes—What is it Really?” Diabetes Care, 26, 1306–1308. [112]
——— (2003b), “Preventing, Delaying, or Masking Type 2 Diabetes With Metformin in the Diabetes Prevention Program?” Diabetes Care, 26, 2702–2703. [112]
——— (2007), “(How) Can We Prevent Type 2 Diabetes?” Diabetes, 56, 1502–1507. [106,107,112]
Chan, P. L. S., and Holford, N. H. G. (2001), “Drug Treatment Effects on Disease Progression,” Annual Review of Pharmacology and Toxicology, 41, 625–659. [112]
Cochran, W. G. (1968), “Errors in Measurement in Statistics,” Technometrics, 10, 637–666. [112]
Committee for Medicinal Products for Human Use (2012), “Guideline on Clinical Investigation of Medicinal Products in the Treatment or Prevention of Diabetes Mellitus,” CPMP/EWP/1080/00 Rev. 1, Committee for Medicinal Products for Human Use. [105,106,107]
D’Agostino, R. B. (2009), “The Delayed-Start Study Design,” New England Journal of Medicine, 361, 1304–1306. [113]
de Winter, W., DeJongh, J., Post, T., Ploeger, B., Uruhart, R., Moules, I., Eckland, D., and Danhof, M. (2006), “A Mechanism-Based Disease Progression Model for Comparison of Long-term Effects of Pioglitazone, Metformin and Glitazide on Disease Processes Underlying Type 2 Diabetes Mellitus,” Journal of Pharmacokinetics and Pharmacodynamics, 33, 313–343. [105]
Garg, R. (2007), “Letter to the Editor,” New England Journal of Medicine, 356, 1379. [113]
Gould, A. L. (1980), “A New Approach to the Analysis of Clinical Drug Trial with Withdrawals,” Biometrics, 36, 721–727. [113]
Kahn, S. E., Haffner, S. M., Heise, M. A., Herman, W. H., Holman, R. R., Jones, N. P., Kravitz, B. G., Lachin, J. M., O’Neill, M. C., Zinman, B., and Viberti, G. (2006), “Glycemic Durability of
Rosiglitazone, Metformin, or Glyburide Monotherapy,” *New England Journal of Medicine*, 355, 2427–2443. [112,113]

Kenward, M. G., and Molenberghs, G. (1998), “Likelihood Based Frequentist Inference When Data are Missing at Random,” *Statistical Science*, 13, 236–247. [110]

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., and Nathan, D. M. (2002), “Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin,” *New England Journal of Medicine*, 346, 393–403. [107,109]

Lewis, J. A., and Machin, D. (1993), “Intention to Treat—Who Should use ITT?,” *British Journal of Cancer*, 68, 647–650. [112]

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley. [111]

Liu, G., and Gould, A. L. (2002), “Comparison of Alternative Strategies for Analysis of Longitudinal Trials With Dropouts,” *Journal of Biopharmaceutical Statistics*, 12, 207–226. [110]

Murray, G. D., and Findlay, J. G. (1988), “Correcting for the Bias Caused by Dropouts in Hypertension trials,” *Statistics in Medicine*, 7, 941–946. [110]

National Research Council (2010), *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials*, Washington, DC: The National Academies Press. [110]

Olanow, C. W., Rascol, O., Hauser, R., Feigin, P. D., Jankovic, J., Lang, A., Langston, W., Melamed, E., Poewe, W., Stocchi, F., and Tolosa, E. (2009), “A Double-Blind, Delayed-Start Trial of Rasagiline in Parkinson’s Disease,” *New England Journal of Medicine*, 361, 1268–1278. [113]

Ploeger, B. A., and Holford, N. H. G. (2009), “Washout and Delayed Start Designs for Identifying Disease Modifying Effects in Slowly Progressive Diseases Using Disease Progression Analysis,” *Pharmaceutical Statistics*, 8, 225–238. [104,113]

Rubin, D. B. (1976), “Inference and Missing Data,” *Biometrika*, 63, 581–592. [110,111]

SAS Institute Inc. (2011), SAS®/STAT® 9.3 User’s Guide, Cary, NC: SAS Institute Inc. [111]

Schefer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall. [111]

Scheen, A. J. (2003), “Preventing, Delaying or Masking Type 2 Diabetes with Metformin in the Diabetes Prevention Program?,” *Diabetes Care*, 26, 2701. [112]

Tanner, M. A., and Wong, W. H. (1987), “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82, 528–550. [111]

Tashkin, D. P., Celli, B., Senn, S., Burkhardt, D., Kesten, S., Menjoge, S., and Decramer, M. (2008), “A 4-Year Trial of Tiotropium in Chronic Obstructive Pulmonary Disease,” *New England Journal of Medicine*, 359, 1543–1554. [112]

The Diabetes Control and Complications Trial Research Group (1993), “The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus,” *New England Journal Medicine*, 329, 977–986. [105]

——— (2003), “Effects of Withdrawal From Metformin on the Development of Diabetes in the Diabetes Prevention Program,” *Diabetes Care*, 26, 977–980. [107,108,109]

——— (2009), “10-Year Follow-up of Diabetes Incidence and Weight Loss in the Diabetes Prevention Program Outcomes Study,” *Lancet*, 374, 1677–1686. [105]

The NAVIGATOR Study Group (2010a), “Effect of Nateglinide on the Incidence of Diabetes and Cardiovascular Events,” *New England Journal Medicine*, 362, 1463–1476. [109]

——— (2010b), “Effect of Valsartan on the Incidence of Diabetes and Cardiovascular Events,” *New England Journal Medicine*, 362, 1477–1490. [109]

UK Prospective Diabetes Study Group (1998), “Intensive Blood- Glucose Control With Sulphonylureas or Insulin Compared With Conventional Treatment and Risk of Complications in Patients With Type 2 Diabetes (UKPDS 33),” *Lancet*, 352, 837–853. [105]

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (2008), “Guidance for Industry Diabetes Mellitus: Developing Drugs and Therapeutic Biologics for Treatment and Prevention (Draft Guidance),” available at http://www.fda.gov/downloads/Drugs/.../Guidances/ucm071624.pdf (accessed 28 June 2012). [105,106,107]

Wild, S., Sicree, R., Roglic, G., King, H., and Green, A. (2004), “Global Prevalence of Diabetes—Estimates for the Year 2000 and Projections for 2030,” *Diabetes Care*, 27, 1047–1053. [105]

Wong, T. Y., Liew, G., Tapp, R. J., Schmidt, M. I., Wang, J. J., Mitchell, P., Klein, R., Klein, B. E. K., Zimmet, P., Shaw, J. (2008), “Relation Between Fasting Glucose and Retinopathy for Diagnosis of Diabetes: Three Population-Based Cross-Sectional Studies,” *Lancet*, 371, 736–743. [105]

**About the Author**

Björn Holzhauer is Associate Director Statistical Scientist, Novartis Pharma AG, Basel, Switzerland (E-mail: bjoern.holzhauer@novartis.com).