Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions

Romain Neugebauer[a]*, Julie A. Schmittdiel[a], Zheng Zhu[a], Jeremy A. Rassen[b], John D. Seeger[b], Sebastian Schneeweiss[b]

The high-dimensional propensity score (hdPS) algorithm was proposed for automation of confounding adjustment in problems involving large healthcare databases. It has been evaluated in comparative effectiveness research (CER) with point treatments to handle baseline confounding through matching or covariance adjustment on the hdPS. In observational studies with time-varying interventions, such hdPS approaches are often inadequate to handle time-dependent confounding and selection bias. Inverse Probability Weighting (IPW) estimation to fit Marginal Structural Models (MSMs) can adequately handle these biases under the fundamental assumption of no unmeasured confounders (NUC). Upholding of this assumption relies on the selection of an adequate set of covariates for bias adjustment. We describe the application and performance of the hdPS algorithm to improve covariate selection in CER with time-varying interventions based on IPW estimation and explore stabilization of the resulting estimates using Super Learning. The evaluation is based on the analysis of electronic health records (EHR) data in a CER study of adults with type 2 diabetes. Results from randomized experiments provide a surrogate gold standard to evaluate inferences. This report 1) establishes the feasibility of IPW estimation with the hdPS algorithm based on large EHR databases, 2) demonstrates little impact on inferences when supplementing the set of expert-selected covariates using the hdPS algorithm in a setting with extensive background knowledge, 3) supports the application of the hdPS algorithm in discovery settings with little background knowledge or limited data availability, and 4) motivates the application of Super Learning to stabilize effect estimates based on the hdPS algorithm.

Keywords: high-dimensional propensity score; marginal structural model; inverse probability weighting; super learning; comparative effectiveness; diabetes

1. Introduction

The high-dimensional propensity score (hdPS) algorithm was proposed [1] for automation of confounding adjustment in problems involving large healthcare databases where expert-selection of confounders “by hand” is not practical due to the high-dimensionality of these databases and suspected to fall short of preventing confounding bias. This algorithm has been evaluated in comparative effectiveness research (CER) with point treatments to handle baseline confounding through matching or covariance adjustment on the hdPS [2, 3, 4]. In observational studies with time-varying interventions, such

[a] *Division of Research, Kaiser Permanente Northern California, Oakland, CA*
[b] *Department of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA*
* *Correspondence to: Romain.S.Neugebauer@kp.org*

Contract/grant sponsor: This project was funded under Contract No. HHSA29020050016I from the Agency for Healthcare Research and Quality, US Department of Health and Human Services as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program. The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the US Department of Health and Human Services.

hdPS approaches are inadequate when covariates that need to be adjusted for to handle time-dependent i) confounding and ii) selection bias are also mediators of treatment effects on the outcome [5, 6, 7]. Inverse Probability Weighting (IPW) estimation to fit Marginal Structural Models (MSMs) can then adequately handle these biases under the fundamental assumption of no unmeasured confounders (NUC) [8, 6].

In practice, upholding of this assumption relies on the selection of a sufficient set of covariates for bias adjustment. Substantive subject-matter knowledge encoded in causal diagrams may be used to rigorously motivate covariate selection (e.g., sequential back-door criterion) [9, 10, 11, 12] but the adequacy of the selection then rests on the correct specification of the corresponding nonparametric causal model and in particular on the ability to assume that all common causes of any two nodes in a diagram are known [13, 14]. In many, if not all, real-world observational studies, drawing a reliable causal diagram based on subject-matter expertise is not possible due to incomplete knowledge. Expert-selection of a set of covariates using the causal diagram principle is then suspected to fall short of upholding the NUC assumption and thus preventing bias. Given these challenges, one practical approach which underlies the hdPS algorithm is to supplement the covariates that are known risk factors for the outcome and that are also suspected to affect treatment or censoring with additional large databases in hope that the NUC assumption can be better approximated with the new information. Machine learning algorithms such as the hdPS algorithm are then used to filter through the supplemental data to identify new relevant covariates for bias adjustment. We discuss concerns that have been raised over such approaches at the end of this report.

First, using a real-world CER study, we describe the application, computation burden and performance of the hdPS algorithm to improve and automate time-dependent confounding and selection bias in CER with time-varying interventions and right-censoring based on IPW estimation to fit a MSM. Second, we explore stabilization of the resulting IPW estimates using Super Learning to data-adaptively reduce the dimensionality of hdPS covariate adjustments based on cross-validation. The evaluation is based on a secondary analysis of electronic health records (EHR) data in a cohort study of adults with type 2 diabetes to contrast cumulative risks of albuminuria development/progression under four more or less aggressive strategies for treatment intensification in patients already on 2+ oral agents or basal insulin. Results from previous randomized experiments provide a surrogate gold standard to validate confounding and selection bias adjustment.

2. Evaluation with a real-world CER study

In this section, we describe the CER question, answers from previous randomized studies, and the observational study on which is based the evaluation in this report. We also introduce formal notation for representing the data structure and the parameter of interest in this analysis.

2.1. Research question and previous trial results

It has long been hypothesized that aggressive glycemic control is an effective strategy to reduce the occurrence of common and devastating microvascular and macrovascular complications of type 2 diabetes (T2DM). A major goal of clinical care of T2DM is minimization of such complications through a variety of pharmacological treatments and interventions to achieve recommended levels of glucose control. The progressive nature of T2DM results in frequent revisiting of treatment decisions for many patients as glycemic control deteriorates. Widely accepted stepwise guidelines start treatment with metformin, then add a secretagogue if control is not reached or deteriorates. Insulin or (less frequently) a third oral agent is the next step. Thus, it is common for T2DM patients to be on multiple glucose-lowering medications.

Current recommendations specify target hemoglobin A1c of $<7\%$ for most patients [15, 16]. However, evidence supporting the effectiveness of a blanket recommendation is inconsistent across several outcomes [17, 18, 19, 20, 21, 22, 23, 24], especially when intensive anti-diabetic therapy is required. In this report, we aim to evaluate the impact of progressively more aggressive glucose-lowering strategies on the development or progression of albuminuria, a microvascular complication in T2DM.

In the ACCORD and ADVANCE clinical trials published from 2008 to 2010 [25, 26, 27], intensive glucose-lowering strategies using multiple classes of glucose-lowering agents succeeded in reducing A1c levels substantially. In the ADVANCE trial, the more intensive therapy arm aimed to reach an A1c level $<6.5\%$ and achieved a mean A1c level of 6.5%, compared to a mean level of 7.3% in the control arm. In the ACCORD trial, the more intensive arm aimed for an A1c of $<6\%$, and achieved a mean A1c of 6.4% (vs. 7.5% in controls). There is substantial data from both trials [28, 29] to support the hypothesis [30, 31] that, in general, those with T2DM who are treated to lower A1c levels may have lower rates of onset and progression of albuminuria (e.g., HR: 0.79,0.66-0.93 in ADVANCE).

2.2. An observational, multi-center, retrospective, cohort study

The effects of intensive treatment remain uncertain, and the optimal target levels of A1c for balancing benefits and risks of therapy are not clearly defined. In addition, no additional major trials addressing these questions are underway.

For these reasons, using expert-selected data from the electronic health records (EHR) from patients of seven sites of the HMO Research Network [32], a large retrospective cohort study of adults with T2DM was assembled to evaluate the impact of progressively more aggressive glucose-lowering strategies on several clinical outcomes. To properly account for time-dependent confounding and informative selection bias, a dynamic MSM [8, 33, 34, 35, 36, 37] was fitted using IPW estimation [8, 38, 39] for the purpose of contrasting cumulative risks under the following four treatment intensification (TI) strategies denoted by d_θ: 'patient initiates TI at the first time (no grace periods allowed [37]) her A1c level reaches or drifts above θ% and patient remains on the intensified therapy thereafter' with θ =7, 7.5, 8, or 8.5.

Details of the study design, analytic approach, and results are described elsewhere [40, 41]. In brief, results were consistent with that of ACCORD and ADVANCE and imply that the pattern of results in these trials are applicable to a large population of adults with T2DM treated in routine clinical settings. In particular, findings from the observational study confirmed the benefit of tight glycemic control with respect to the development or progression of albuminuria.

Here, we report on results from a secondary analysis of the same expert-selected observational data but supplemented with additional claims data. This secondary analysis aims to contrast the same four counterfactual survival curves indexed by the TI strategies described above for the purpose of evaluating the performance of the hdPS algorithm for improving time-dependent confounding and selection bias adjustment by IPW estimation. We now formally describe the expert-selected observational data, supplemental claims data and the parameter of interest before describing the use of the hdPS algorithm within IPW estimation.

2.3. Data, parameter of interest and assumptions

The observed data on each patient in the cohort consist of measurements on exposure, outcome, and confounding variables made at 90-day intervals between study entry and until each patient's end of follow-up. The time (expressed in units of 90 days) when the patient's follow-up ends is denoted by $\tilde{T}$ and is defined as the earliest of the time to failure, i.e., albuminuria development or progression, denoted by T or the time to a right-censoring event denoted by C. When a patient is right-censored, i.e., $\tilde{T} = C$, the type of right-censoring event experienced by the patient is recorded and denoted by Γ with possible values 1, 2, or 3 to represent end of follow-up by administrative end of study, disenrollment from the health plan, or death respectively. For patients with normoalbuminuria at study entry, i.e., microalbumin-to-creatinine ratio (ACR) <30, we defined failure as an ACR measurement indicating either microalbuminuria (ACR 30 to 300) or macroalbuminuria (ACR>300). For patients with microalbuminuria at study entry, we defined failure as an ACR measurement indicating macroalbuminuria. We thus excluded patients with a baseline ACR measurement missing (5884) or indicating macroalbuminuria (1608), which yielded the sample size $n = 51,179$. The indicator that the follow-up time $\tilde{T}$ is equal to the failure time T is denoted by $\Delta = I(\tilde{T} = T)$. At each time point $t = 0, \ldots, \tilde{T}$, the patient's exposure to an intensified diabetes treatment is represented by the binary variable $A_1(t)$, and the patient's right-censoring status is denoted by the indicator variable $A_2(t) = I(C \leq t)$. The combination $A(t) = (A_1(t), A_2(t))$ is referred to as the action at time t. At each time point $t = 0, \ldots, \tilde{T}$, expert-selected covariates (listed in Table 1) are denoted by the multi-dimensional variable $L(t)$ and defined from measurements that occur before the action at time t, $A(t)$, or are otherwise assumed not to be affected by the actions at time t or thereafter, $(A(t), A(t+1), \ldots)$. In addition, the covariates at time t include an outcome measurement denoted by $Y(t)$, i.e., $Y(t) \in L(t)$ for $t = 0, \ldots, \tilde{T}$. For each time point $t = 0, \ldots, \tilde{T}+1$, the outcome is the indicator of past failure, i.e., $Y(t) = I(T \leq t-1)$. By definition, the outcome is thus 0 for $t = 0, \ldots, \tilde{T}$, missing at $t = \tilde{T}+1$ if $\Delta = 0$ and, 1 at $t = \tilde{T}+1$ if $\Delta = 1$. To simplify notation, we use overbars to denote covariate and exposure histories, e.g., a patient's exposure history through time t is denoted by $\bar{A}(t) = (A(0), \ldots, A(t))$. Following the MSM framework [8], we approach the observed data in this study as realizations of n independent and identically distributed copies of $O = (\tilde{T}, \Delta, (1-\Delta)\Gamma, \bar{L}(\tilde{T}), \bar{A}(\tilde{T}), \Delta Y(\tilde{T}+1))$ denoted by O_i for $i = 1, \ldots, n$. The longest observed follow-up time is $\max_{i=1,\ldots,n} \tilde{T}_i = 36$ (9 years). Details about the approach implemented for mapping EHR data into these coarsened exposure, covariate and outcome data for each patient was described elsewhere [40, Appendix E].

In addition, these expert-selected data O are supplemented by the following time-stamped claims data collected on each patient in the cohort between 90 days prior to study entry and the end of follow-up: 1) All diagnoses from inpatient, outpatient, and emergency department visits. Both primary and secondary diagnoses from inpatient visits are included; 2) All procedures from inpatient, outpatient, and emergency department visits; 3) All types of laboratory visits (not the test measurements but only the names of the tests undergone by the patient, e.g., hemoglobin A1c versus HDL cholesterol); 4) All filled prescriptions; 5) All types of clinical encounters (i.e., ambulatory, emergency department, email, telephone, acute inpatient hospital stay, non-acute institutional stay, laboratory only, radiology only, or other). Note that the granularity of all claims data based on the ICD-9 coding system (including V and E codes) was restricted to all digits before the dot. As described in the next section, these claims data are mined with the hdPS algorithm to identify new summary measures

which are referred to as hdPS covariates from now on. These hdPS covariates are then appended to the covariate vector $L(t)$ of the data structure O for the purpose of improving bias adjustment by IPW estimation.

In this study, we aim to evaluate the effect of dynamic treatment interventions on the cumulative risk of failure at a pre-specified time point t_0, e.g., $t_0 = 11$ to investigate cumulative risks of failure over three years. The dynamic treatment interventions of interest correspond to treatment decisions made according to given clinical policies for initiation of an intensified therapy based on the patient's evolving A1c level. These policies denoted by d_θ were described above. Formally, these policies are individualized action rules [34] defined as a vector function $d_\theta = (d_\theta(0), \ldots, d_\theta(t_0))$ where each function, $d_\theta(t)$ for $t = 0, \ldots, t_0$, is a decision rule for determining the action regimen (i.e., a treatment and right-censoring intervention) to be experienced by a patient at time t. A decision rule $d_\theta(t)$ maps the action and covariate history measured up to a given time t to an action regimen at time t: $d_\theta(t) : (\bar{L}(t), \bar{A}(t-1)) \mapsto (a_1(t), a_2(t))$. In this study, the decision rules of interest are defined such that $d_\theta(t)((\bar{L}(t), \bar{A}(t-1))$ is:

- $(a_1(t), a_2(t)) = (0, 0)$ (i.e., no use of an intensified treatment and no right-censoring) if and only if the patient was not previously treated with an intensified therapy (i.e., $\bar{A}(t-1) = 0$) and the A1c level at time t (an element of $L(t)$) was lower than or equal to the threshold θ.
- $(a_1(t), a_2(t)) = (1, 0)$ (i.e., use of an intensified treatment and no right-censoring) otherwise.

To simplify notation below, the action regimen $(a(0) = d_\theta(0)(L(0)), a(1) = d_\theta(1)(\bar{L}(1), a(0)), \ldots, a(t) = d_\theta(t)(\bar{L}(t), \bar{a}(t-1)))$ through time t is denoted by $d_\theta(\bar{L}(t))$ for any given observed covariate history through time t, $\bar{L}(t)$. The parameter of interest denoted by ψ^{θ_1,θ_2} is the differences between the cumulative risks at time t_0 associated with any two distinct treatment strategies d_{θ_1} and d_{θ_2}:

$$\psi^{\theta_1,\theta_2} = P(Y_{d_{\theta_1}}(t_0+1) = 1) - P(Y_{d_{\theta_2}}(t_0+1) = 1).$$

For conciseness, we refer the reader to earlier work [40, Appendices B and D] for a description of the concepts and the counterfactual statistical framework on which relies the definition of this parameter of interest.

Identifiability of this parameter with the observational data above relies on at least three assumptions detailed elsewhere [40, Appendices C]: NUC, positivity, and consistent estimation of the action mechanism (defined in the next section).

If the MSM framework above (missing data framework) is not explicitly resting on the more general structural framework through additional explicit assumptions encoded by a causal diagram [42], then an additional assumption referred to as consistency assumption is made [43, 44].

In addition, a more or less flexible non-saturated MSM may be assumed [45, 46, 47, 36, 48]. The assumption encoded by such an MSM typically imposes constraints on the survival curves that underlie the definition of the parameter of interest ψ^{θ_1,θ_2}. In practice, specification of a non-saturated MSM is essentially an arbitrary choice that does not encode real knowledge about the true survival curves of interest. The previous CER analysis of these observational data was based on such a MSM although minimal constraints were actually imposed since the MSM chosen and shown below is relatively close to saturation:

$$logit(P(Y_d(t+1) = 1 \mid Y_d(t) = 0)) = \sum_{\theta \in \{7, 7.5, 8, 8.5\}} \Big(\sum_{j=0}^{7} \beta_j^\theta I(t = j, d = d_\theta) + \beta_8^\theta I(8 \le t \le 11, d = d_\theta) + \beta_{12}^\theta I(12 \le t \le 15, d = d_\theta) + \beta_{16}^\theta I(16 \le t \le 23, d = d_\theta) + \beta_{24}^\theta I(t \ge 24, d = d_\theta) \Big). \quad (1)$$

This MSM is also posited for all analyses in this report.

3. Inverse Probability Weighting estimation and the hdPS algorithm

3.1. IPW estimation with expert-selected covariates only

In previous work [40], IPW estimation was implemented to fit the parametric dynamic MSM (1) based on expert-selected covariates only. Estimates of the hazards were subsequently mapped into estimates of each of the four corresponding counterfactual survival curves from which inferences about risk differences (RDs) ψ^{θ_1,θ_2} were derived.

Here, we focus on IPW estimation of 6 distinct RDs at $t_0 = 11$: $\psi^{8.5,8}$, $\psi^{8.5,7.5}$, $\psi^{8.5,7}$, $\psi^{8,7.5}$, $\psi^{8,7}$, and $\psi^{7.5,7}$. IPW estimation of these RDs relies on consistent estimation of the nuisance parameter $g^\theta = (g^\theta_{A(0)}, \ldots, g^\theta_{A(t_0)})$ for $\theta \in \{7, 7.5, 8, 8.5\}$ where $g^\theta_{A(t)} = P(A(t) = d_\theta(\bar{L}(t)) \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}(t-1) = d_\theta(\bar{L}(t-1)))$ represents the conditional probability that a patient's exposure and right-censoring status at time t remain concordant with the action implied

by the decision rule d_θ given i) that the patient did not fail before t, ii) that her past actions are concordant with action decisions according to rule d_θ, and iii) her past observed covariates $\bar{L}(t)$. The approach implemented to estimate $g^\theta_{A(t)}$ is based on separate estimation of each element of the factorization of the action mechanism at time t, i.e., $P(A(t) \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}_1(t-1), \bar{A}_2(t-1) = 0)$. Specifically, the following probabilities were estimated separately as detailed later:

- Propensity score (PS) for TI initiation denoted by μ_1:

$$P(A_1(t) = 1 \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}_1(t-1) = 0, \bar{A}_2(t) = 0)$$

- PS for TI continuation denoted by μ_2:

$$P(A_1(t) = 1 \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}_1(t-2), A_1(t-1) = 1, \bar{A}_2(t) = 0)$$

- PS for right-censoring by administrative end of study denoted by μ_3:

$$P(I(A_2(t) = 1, \Gamma = 1) = 1 \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}_1(t-1), \bar{A}_2(t-1) = 0),$$

  where $I(\cdot)$ denotes an indicator variable
- PS for right-censoring by disenrollment from the health plan denoted by μ_4:

$$P(I(A_2(t) = 1, \Gamma = 2) = 1 \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}_1(t-1), \bar{A}_2(t-1) = 0, I(A_2(t) = 1, \Gamma = 1) = 0)$$

- PS for right-censoring by death denoted by μ_5:

$$P(I(A_2(t) = 1, \Gamma = 3) = 1 \mid \bar{L}(t), \bar{Y}(t) = 0, \bar{A}_1(t-1), \bar{A}_2(t-1) = 0, I(A_2(t) = 1, \Gamma = 1) = 0, I(A_2(t) = 1, \Gamma = 2) = 0)$$

For patients who followed rule d_θ through t (i.e., for whom $\bar{A}(t) = d_\theta(\bar{L}(t))$), an estimate of the nuisance parameter $g^\theta_{A(t)}$ can be derived from estimates of these 5 PSs based on the following equality implied by factorizing the action mechanism at time t using the chain rule:

$$g^\theta_{A(t)} = \big(I(\bar{A}_1(t-1) = 0)\mu_1^{A_1(t)}(1-\mu_1)^{1-A_1(t)} + I(A_1(t-1) = 1)\mu_2^{A_1(t)}(1-\mu_2)^{1-A_1(t)}\big)(1-\mu_3)(1-\mu_4)(1-\mu_5). \quad (2)$$

Similar to previous work [40], right-censoring due to administrative end of study is assumed uninformative in this report. Thus, implementation of IPW estimation can be simplified by ignoring the PSs μ_3 in the calculation of stabilized IPW weights because weight stabilization results in cancellation of such probabilities in both the numerator and denominator of the weights. In this secondary analysis, the approach implemented to estimate the numerator of the weights is identical to that implemented in previous work whereas estimation of the denominator of the weights was altered to permit a direct comparison of IPW results derived with and without the hdPS algorithm. Indeed, data had been pooled across time in previous work to fit 4 main-term logistic models for the PSs μ_1, μ_2, μ_4, and μ_5, respectively. Because the current implementation of the hdPS algorithm does not permit such simultaneous estimation of PSs by pooling data across time, we adopted a stratified estimation approach for each PS in this report. In other words, for each time point t separately, 4 main-term logistic models were fitted here to estimate each of the 4 PSs μ_1, μ_2, μ_4, and μ_5. By 'main-term logistic model', we mean a logistic model with only main terms for each explanatory variable considered (i.e., no interaction terms between explanatory variables). The explanatory variables considered were all time-independent covariates and the last measurement of time-varying covariates. In addition, exposure to TI in the last period was included as an explanatory variable for the 2 PSs μ_4 and μ_5 for right-censoring and the latest change in A1c was included as an explanatory variable for estimating all 4 PSs.

In this section, all logistic models for estimating the 4 relevant PSs over time involve main terms for 48 expert-selected covariates only (Table 1). The resulting estimate of g^θ is denoted by $g^\theta_{n,expert}$. The distribution of the stabilized IPW weights estimated based on $g^\theta_{n,expert}$ is described in Table 2. Weight truncation [49, 50, 51] is often implemented in practice to improve the precision of IPW estimation. In this report, IPW weights were truncated at 20, i.e. all stabilized weights above 20 were replaced with value 20. Table 4 and Figure 1 display the estimates resulting from the IPW estimation approach just described. In addition, Figure 2 displays the results from a crude analysis that consists in fitting the logistic model for the discrete-time hazards without weights, i.e., without adjustment for time-dependent confounding and selection bias. A comparison of the crude and IPW estimates of the survival curves on Figures 1 and 2 clearly demonstrates successful adjustment for time-dependent confounding and selection bias with the IPW approach. The IPW estimates indicate an early separation and consistent ordering of the four survival curves suggesting an increasing beneficial effect of more aggressive therapy initiation rules (i.e., of rules indexed by decreasing A1c thresholds). These results are consistent with that of the ACCORD and ADVANCE randomized trials.

3.2. hdPS algorithm to supplement expert-guided bias adjustment

In this section, we append hdPS covariates to the set of expert-selected covariates $L(t)$ for the purpose of improving bias adjustment which may be insufficient due to violations of the NUC assumption when relying on expert-selected covariates only. The hdPS covariates were derived from application of the hdPS algorithm to claims data grouped into so-called dimension data sets. The following seven dimension data sets were considered in this analysis: all outpatient and emergency department diagnoses combined (referred to as 'oped.dx'), all primary and secondary inpatient diagnoses combined (referred to as 'ip12.dx'), all outpatient and emergency department procedures combined (referred to as 'oped.px'), all inpatient procedures (referred to as 'ip.px'), all filled prescriptions (referred to as 'rx'), all types of laboratory tests performed (referred to as 'lab'), and all types of clinical encounters (referred to as 'utilization'). Each data set contains a separate row for each distinct claim encounter with three columns that identify the patient, date, and claim code associated with the encounter.

For each time point t and each of the four PS at time t (i.e., μ_1, μ_2, μ_4, and μ_5), a distinct set of hdPS covariates was derived based on all seven dimension data sets. Specifically, for each time point t, we first identified all $n_{t,j}$ patients who contributed to the logistic fit for estimating a given PS μ_j at time t in the previous section. Second, we filtered each dimension data set to only retain the claim encounters from these patients recorded in the 90-day interval $t-1$ to ensure anteriority of the resulting hdPS covariates with respect to the exposure and censoring status at time t. Third, the resulting 7 reduced dimension data sets were considered by the hdPS algorithm to identify $k_{t,j}$ explanatory covariates for predicting the independent variable associated with the PS μ_j at time t. Finally, the logistic model for estimating μ_j at time t in the previous section was modified here to include main terms for the $k_{t,j}$ hdPS covariates. The resulting estimate of the nuisance parameter g^θ is denoted by $g^\theta_{n,expert,hdPS}$ (Table 1).

The derivation of hdPS covariates was implemented based on publicly available Java programs for which interfaces are available in both the SAS and R software. R version 2.13.0 was used for implementing all analyses in this report based on version 2.4.10 of the hdPS algorithm software [52]. The hdPS algorithm requires user-specification of a set of options that determine the number of hdPS covariates to be derived but also the method for deriving them. The following Java arguments were used:

- `k` is the maximum number of hdPS covariates derived for each time t and PS type μ_j. It was determined based on the number $n_{t,j}$ of patients who can contribute to the estimation of the PS μ_j at time t to ensure that a minimum of 50 observations were available per main term in the logistic model for the PS. The value of the argument `k` was capped at 500 and we thus chose $k = \min(500, \max(0, \text{floor}(\frac{n}{50} - 48)))$. Note that $k_{t,j} \leq k$.
- `frequencyMin` restricts the set of claim codes that are used to define hdPS covariates. If a code appears in the reduced dimension data for less than `frequencyMin` patients then this code is ignored in the derivation of the hdPS covariates. The value of the argument `frequencyMin` was set to 100.
- `topN` restricts the number of claim codes that are used to define hdPS covariates. Each code passing the `frequencyMin` criterion is ranked based on the magnitude of its prevalence and only the first `topN` codes of the resulting ranked list are used to define hdPS covariates. The value of the argument `topN` was set to 200.
- `inferServiceIntensityVars` is the indicator that variables that describe the intensity of service usage in each reduced dimension data set should be used to define hdPS covariates. This indicator was set to 1.
- `variableRankingMethod` indicates whether hdPS covariates should be ranked based on their association with either 1) the PS independent variable (i.e., the exposure for PSs μ_1 and μ_2, the indicator of censoring by death for PS μ_4, and the indicator of censoring by health plan disenrollment for PS μ_5), 2) the outcome (i.e., albuminuria development/progression) or, 3) both. This argument was set to `BIAS` for a ranking based on both the PS independent variable and the outcome. More specifically, for deriving the hdPS covariates associated with a PS μ_j at time t, we defined the outcome as the indicator of observed failure between and at time $t+1$ and $t+8$ (i.e., the indicator of failure in the next 2 years of follow-up).

Tables 5 through 8 summarize the numbers k and $k_{t,j}$ of hdPS covariates considered and selected for estimating each PS μ_j at each time point t through $t_0 = 11$. These tables also display the number of patients $n_{t,j}$ contributing to the estimation of each PS and the total number of claim encounters in each reduced dimension data set. Given the relative large number of claim encounters processed by the hdPS algorithm for each time point and each PS μ_j, it is worth noting that computing time to identify and define the set of $k_{t,j}$ covariates is only measured in seconds (most often less than one minute) when providing 12GB of memory to the Java programs[†]. Thus, it is the fitting of the resulting main-term logistic models that dominates overall computing time when implementing IPW estimation with the hdPS algorithm.

[†] All programs ran on a compute server operating under RedHat RHEL 6.1 with 4 Physical CPU Sockets (Intel (R) Xeon(R) CPU X7560 @ 2.27 GHz with 16 Cores) and 128 GB of physical memory.

The distribution of the stabilized IPW weights estimated based on $g^{\theta}_{n,expert,hdPS}$ is described in Table 2. Table 4 and Figure 1 display the corresponding IPW estimates of the 6 RDs and 4 survival curves after weight truncation at 20.

A comparison of the distributions of the stabilized weights derived based on the estimates $g^{\theta}_{n,expert}$ and $g^{\theta}_{n,expert,hdPS}$ in Table 2 indicates that the application of the hdPS algorithm results in more extreme weighting of some person-time observations contributing to the MSM fit. While the 990^{th} and 999^{th} permiles of the stabilized weights derived without the hdPS algorithm were 7.6 and 15.67, respectively, they increased to 8.98 and 1,571,154,565,775,085 for the stabilized weights based on the hdPS algorithm. As shown in Table 4, even after weight truncation at 20, the overall increase in the number of large weights results in a clear increase (of about 10% to 110%) of the variability $\Gamma^{\theta_1,\theta_2}_n$ of the IPW estimates for all 6 RDs. A comparison of the crude and IPW estimates based on the hdPS algorithm on Figures 2 and 1 with results from the ACCORD and ADVANCE trials demonstrates however successful adjustment for time-dependent confounding and selection bias with IPW estimation based on expert-selection of covariates supplemented by automated covariate selection with the hdPS algorithm. Table 4 also shows that the IPW point estimates $\psi^{\theta_1,\theta_2}_n$ of the RDs at $t_0 = 11$ based on the hdPS algorithm are mostly larger than that that derived based on only expert-selection of covariates. Figure 1 confirms this trend for RDs defined between study entry and t_0: there is a clearer separation of the four survival curves starting at study entry when using the hdPS algorithm. In particular, the survival curve associated with the TI decision rule d_7 is now consitently above all other curves when using the hdPS algorithm. These results do not substantiate concerns over potential bias from over-adjustment with the hdPS algorithm (e.g., so-called M-bias) [9, 2]. As shown in Table 4, the increase in the IPW standard errors with the hdPS algorithm is counterbalanced by the corresponding increase in the magnitude of the point estimates such that the IPW inference based on both expert-selected and hdPS covariates remains essentially unchanged compared to that derived without hdPS covariates. We note, however, that almost all p-values increased when using the hdPS algorithm.

3.3. hdPS to semi-automate bias adjustment with claims data only

To evaluate semi-automation of confounding and selection bias adjustment with the hdPS algorithm based on claims data only, we evaluated the approach described in the previous section with the difference that only the subset of expert-selected covariates that would be available in typical claims databases are considered as expert-selected covariates for IPW estimation in this section. As a result, all logistic models for estimating the PSs now include main terms for only 24 expert-selected covariates available in claims data and main terms for all hdPS covariates selected with the hdPS algorithm applied as described in the previous section with $k = \min(500, \max(0, \text{floor}(\frac{n}{50} - 24)))$. The resulting estimate of the nuisance parameter g^{θ} is denoted by $g^{\theta}_{n,claims,hdPS}$ (Table 1)

Note that the evaluation in this section is artificial since we aim to evaluate the extent to which the information from claims data only and their mining with the hdPS algorithm are sufficient for proper bias and confounding adjustment while the definitions of the exposure groups and outcomes in this analysis do require availability of two time-varying EHR variables, i.e., hemoglobin A1c and ACR. The analysis in this section could thus not be implemented in practice if only claims data were available.

The distribution of the stabilized IPW weights estimated based on $g^{\theta}_{n,claims,hdPS}$ is described in Table 2. Table 4 and Figure 1 display the corresponding IPW estimates of the 6 RDs and 4 survival curves after weight truncation at 20.

A comparison of the distributions of the stabilized weights derived based on the estimates $g^{\theta}_{n,expert,hdPS}$ and $g^{\theta}_{n,claims,hdPS}$ in Table 2 indicates that ignoring EHR data in the estimation of the PSs results in a greater number of larger weights even though extreme weighting (weights$\geq$150) becomes slightly less frequent. While the 990^{th} and 999^{th} permiles of the stabilized weights derived with $g^{\theta}_{n,expert,hdPS}$ were 8.98 and 1,571,154,565,775,085, respectively, they increased to 22.1 and 1,667,206,539,714,011 for the stabilized weights based on $g^{\theta}_{n,claims,hdPS}$. As shown in Table 4, after weight truncation at 20 the variability $\Gamma^{\theta_1,\theta_2}_n$ of the IPW estimates based $g^{\theta}_{n,claims,hdPS}$ is nevertheless mostly lower than that of the IPW estimates based on $g^{\theta}_{n,expert,hdPS}$. A comparison of the crude and IPW estimates based on $g^{\theta}_{n,claims,hdPS}$ on Figures 2 and 1 with results from the ACCORD and ADVANCE trials demonstrates some, albeit very limited, adjustment for time-dependent confounding and selection bias in the expected directions when using IPW estimation with $g^{\theta}_{n,claims,hdPS}$. Table 4 indeed shows that the IPW point estimates $\psi^{\theta_1,\theta_2}_n$ of the RDs at $t_0 = 11$ based on $g^{\theta}_{n,claims,hdPS}$ are much smaller than that derived based on $g^{\theta}_{n,expert}$ and $g^{\theta}_{n,expert,hdPS}$. Figure 1 confirms this trend for RDs defined between study entry and t_0: there is only a weak separation of the four survival curves starting at study entry when using $g^{\theta}_{n,claims,hdPS}$. As shown in Table 4, the combination of large standard errors and small point estimates from IPW estimation with $g^{\theta}_{n,claims,hdPS}$ does not permit to rule out the null hypotheses since all p-values for all 6 RDs are great than 0.4. These results underscore the importance of EHR data in CER.

3.4. *hdPS to automate bias adjustment with claims data only*

To evaluate automation of confounding and selection bias adjustment with the hdPS algorithm without subject-matter expertise, we evaluated the approach described in the previous section with the difference that all expert-selected covariates were ignored in the derivation of IPW estimates. As a result, all logistic models for estimating the PSs now only include main terms for all hdPS covariates selected with the hdPS algorithm applied as described earlier with $k = \min(500, \max(0, \text{floor}(\frac{n}{50})))$. The resulting estimate of the nuisance parameter g^θ is denoted by $g^\theta_{n,only\ hdPS}$ (Table 1).

Note that the evaluation in this section remains artificial. Indeed, we aim to evaluate the extent to which information from claims data may be successfully mined with the hdPS algorithm only to properly account for confounding and selection bias while the definitions of the exposure groups and outcomes in this analysis do require availability of two time-varying EHR variables, i.e., hemoglobin A1c and ACR. The analysis in this section could thus not be implemented in practice if only claims data were available.

The distribution of the stabilized IPW weights estimated based on $g^\theta_{n,only\ hdPS}$ is described in Table 2. Table 4 and Figure 1 display the corresponding IPW estimates of the 6 RDs and 4 survival curves after weight truncation at 20.

A comparison of the distributions of the stabilized weights derived based on the estimates $g^\theta_{n,claims,hdPS}$ and $g^\theta_{n,only\ hdPS}$ in Table 2 indicates that ignoring expert-selected claims covariates in the estimation of the PSs results in little impact on the weight distribution. While the 990^{th} and 999^{th} permiles of the stabilized weights derived with $g^\theta_{n,claims,hdPS}$ were 22.1 and 1,667,206,539,714,011, respectively, they slightly decreased to 21.91 and 1,629,739,264,204,867 for the stabilized weights based on $g^\theta_{n,only\ hdPS}$. As shown in Table 4, after weight truncation at 20 the variability $\Gamma^{\theta_1,\theta_2}_n$ of the IPW estimates based $g^\theta_{n,only\ hdPS}$ is mostly lower than that of the IPW estimates based on $g^\theta_{n,claims,hdPS}$. A comparison of the crude and IPW estimates based on $g^\theta_{n,claims,hdPS}$ on Figures 2 and 1 demonstrates improved, albeit still limited, adjustment for time-dependent confounding and selection bias in the expected directions when using $g^\theta_{n,only\ hdPS}$ instead of $g^\theta_{n,claims,hdPS}$. The IPW estimate of the survival curve associated with the TI decision rule d_7 is however no longer consistently above all other estimated curves when using $g^\theta_{n,only\ hdPS}$. Table 4 also shows that the IPW point estimates $\psi^{\theta_1,\theta_2}_n$ of all RDs not involving d_7 and based on $g^\theta_{n,only\ hdPS}$ are larger than that derived based on $g^\theta_{n,claims,hdPS}$. Figure 1 confirms this trend between study entry and t_0. As shown in Table 4, the combination of large standard errors and small point estimates from IPW estimation with $g^\theta_{n,only\ hdPS}$ does not permit to rule out the null hypotheses since the p-values for almost all 6 RDs remain dismal. These results further underscore the importance of EHR data in CER.

4. Super Learning to stabilize IPW estimation with the hdPS algorithm

Applications of the hdPS algorithm in the previous section led to a large increase in estimation variability. This loss in efficiency results from estimates of the denominator of the weights approaching 0 more frequently when main terms for hdPS covariates are incorporated in the logistic models for the PSs (Table 2). Such a change in the weight estimates may be legitimate if it is caused by some hdPS covariates truly being strong predictors of the exposure or censoring events (near violation of the positivity assumption) or it may be an artifact of the aggressive estimation approach for the PSs which results in overfitting (excessive number of main terms in the logistic models compared to the available number of observations).

In this section, we evaluate the use of Super Learning for protection against overfitting when estimating PSs with the hdPS algorithm. In our previous applications of the hdPS algorithm, we arbitrarily selected the number k of hdPS covariates that could enter the logistic models for the PSs. This number was chosen as a function of the sample size available at each time point as a feeble attempt to manually guard against overfitting. Super Learning (SL) has been proposed and implemented in CER [53] to hedge against erroneous causal inference due to such arbitrary estimation decisions. SL [54] is a data-adaptive estimation algorithm that combines predicted values from a library of various candidate estimators (a.k.a. learners) through a weighted average. The selection of the optimal combination of the candidate learners is based on cross-validation [55, 56, 57, 58] to protect against overfitting such that the resulting learner (called 'super learner') performs asymptotically as well (in terms of mean error) or better than any of the candidate learners considered.

Here, instead of arbitrarily fixing the number of relevant hdPS covariates that should enter a logistic model for estimating each PS, we define a sequence of candidate estimators for each PS based on nested logistic models that incorporate an increasing number of hdPS covariates up to a maximum number k and SL is then used to derive an estimate of each PS based on these candidate estimators. More specifically, the hdPS algorithm is first used to define and rank up to k hdPS covariates relevant to a given PS. Second, a candidate logistic model for estimating each PS is defined by combining main terms for expert-selected covariates with main terms for the first 0, 10, 20, 30, 40, 50, 100, 200, 300, 400, or 500

covariates from the ranked list of hdPS covariates. In addition to the resulting 11 candidate estimators for each PS, we also considered the candidate estimator defined by the intercept logistic model (i.e., prediction with the average value).

This SL approach to PS estimation with the hdPS algorithm was applied to 1) supplement expert-guided bias adjustment, i.e., we combined SL and the general approach described in section 3.2 to define a new estimator for g^θ denoted by $g^\theta_{n,expert,hdPS,SL}$, and 2) semi-automate bias adjustment with claims data only, i.e., we combined SL and the general approach described in section 3.3 to define a new estimator for g^θ denoted by $g^\theta_{n,claims,hdPS,SL}$ (Table 1). In a third application, we combined SL and the general approach described in section 3.3 with the difference that two EHR expert-selected variables (current A1c level and difference between the last two A1c levels) were also added to the list of explanatory variables in each main-term logistic models. The resulting new estimator for g^θ is denoted by $g^\theta_{n,claims+A1c,hdPS,SL}$ (Table 1).

Tables 9 through 12 describe the composition of the super learners for each PS μ_j at each time point t through $t_0 = 11$ when the hdPS algorithm is used to supplement 48 expert-selected covariates for confounding and selection bias adjustment. It is worth noting that the computation burden to derive such super learners increases significantly: about 10, 1, 18, and 23 hours of computation time were needed for estimating the PSs μ_1, μ_2, μ_4, and μ_5, respectively. Table 3 describes the distribution of the stabilized IPW weights based on the resulting estimates $g^\theta_{n,expert,hdPS,SL}$ but also the estimators $g^\theta_{n,claims,hdPS,SL}$ and $g^\theta_{n,claims+A1c,hdPS,SL}$ derived from the other two SL approaches described above. Table 4 and Figure 2 display the corresponding IPW estimates of the 6 RDs and 4 survival curves after weight truncation at 20.

A comparison of the distributions of the stabilized weights derived based on the estimates $g^\theta_{n,expert,hdPS,SL}$ and $g^\theta_{n,expert,hdPS}$ in Table 3 reveal a clear reduction in the number of large weights. While the 990^{th} and 999^{th} permiles of the stabilized weights derived without SL ($g^\theta_{n,expert,hdPS}$) were 8.98 and 1,571,154,565,775,085, respectively, they decrease to 7.65 and 15.31 for the stabilized weights based on SL ($g^\theta_{n,expert,hdPS,SL}$). A similar shift in the distribution of the stabilized weights due to SL was observed when comparing the weights derived based on the estimates $g^\theta_{n,claims,hdPS}$ and $g^\theta_{n,claims,hdPS,SL}$. The 990^{th} and 999^{th} permiles of the stabilized weights derived based on $g^\theta_{n,claims,hdPS,SL}$ were 16.79 and 29.45, respectively. As shown in Table 4, after weight truncation at 20, the overall decrease in the number of large weights with SL results in a clear stabilization of the IPW estimates: the variability $\Gamma^{\theta_1,\theta_2}_n$ of the IPW estimates based on $g^\theta_{n,expert,hdPS,SL}$ is much smaller than that of the IPW estimates based on $g^\theta_{n,expert,hdPS}$ for all 6 RDs (a decrease of about 12% to 53%). The same trend in variability is observed when comparing the IPW estimators based on $g^\theta_{n,claims,hdPS}$ to that based on $g^\theta_{n,claims,hdPS,SL}$ (a decrease of about 3% to 56%).

A comparison of the IPW estimates based on $g^\theta_{n,expert,hdPS,SL}$ and $g^\theta_{n,expert,hdPS}$ on Table 4 and Figures 2 and 1 demonstrates however a slight decrease in the separation of the four survival curves. The estimated survival curves and RDs based on expert-selected covariates, the hdPS algorithm and SL ($g^\theta_{n,expert,hdPS,SL}$) are closer to that based on expert-selected covariates only ($g^\theta_{n,expert}$) than that based on expert-selected covariates and the hdPS algorithm without SL ($g^\theta_{n,expert,hdPS}$). In particular, the survival curve associated with the TI decision rule d_7 is no longer consistently above all other curves when using SL. This observation suggests that the use of SL to data-adaptively incorporate the information from hdPS covariates into estimates of the PSs results not only in a major increase in the precision of effect estimates but also in minor bias from incomplete confounding and selection bias adjustment. A similar observation can be noted when comparing the IPW estimates based on $g^\theta_{n,claims,hdPS}$ and $g^\theta_{n,claims,hdPS,SL}$.

These results motivate the application of SL as a protection against overfitting from arbitrary decision for estimating the PSs with the hdPS algorithm. They do not substantiate concerns over finite-sample bias from near violation of the positivity assumption [59] due to the identification of instruments with the hdPS algorithm (so-called Z-bias) [60, 2]. The decrease in estimation variability from the application of SL seems however to come at a cost of a small increase in finite-sample bias that is particularly apparent when comparing the survival curves based on $g^\theta_{n,claims,hdPS}$ and $g^\theta_{n,claims,hdPS,SL}$. It is interesting to note however that the lack of confounding and selection bias adjustment with $g^\theta_{n,claims,hdPS,SL}$ is largely overcome when including only two main terms for A1c measurements in the logistic models for the PSs as shown by the proximity of the IPW estimates based on $g^\theta_{n,expert}$ and $g^\theta_{n,claims+A1c,hdPS,SL}$ on Figure 2 and Table 4. This last result highlights the importance of A1c measurements as major confounding factors in the subject-matter problem studied. While A1c is known to be a determinant of treatment decisions in diabetes and thus to be associated with TI initiation, it is worth noting that its inclusion in the PS models results in a more compact distribution of IPW weights as shown by the distributions of the stabilized weights based on $g^\theta_{n,claims,hdPS,SL}$ and $g^\theta_{n,claims+A1c,hdPS,SL}$ in Table 4.

5. Discussion

To our knowledge, this report describes the first evaluation of the hdPS algorithm in CER with time-varying interventions based on IPW estimation to adjust for the time-dependent confounding and selection bias expected in observational studies. It establishes the feasibility of IPW estimation to fit MSM with the hdPS algorithm in real-world CER based on large healthcare databases [61]. It demonstrates little impact on inferences when supplementing the set of expert-selected covariates using the hdPS algorithm in studies where the no unmeasured confounders assumption is likely tenable solely based on expert-selection of covariates, i.e., in settings with extensive background knowledge and rich data [4]. It demonstrates sufficient, albeit incomplete, confounding and selection bias adjustment with the hdPS algorithm to differentiate point estimates of survival curves with little or no expert-selection of covariates based on relatively limited databases. It thus provides an additional data point [62] for supporting the application of the hdPS algorithm in discovery studies based on claims data and characterized by little background knowledge. Finally, this work also motivates the application of SL to stabilize effect estimates that are based on the hdPS algorithm.

There has been and is extensive methodological research devoted to the problem of identifiability of causal effects based on graphical modeling [11, 63]. Based on this theoretical framework, formal results were developed and led to algorithms for drawing causal directed acyclic graphs (DAGs) from data (i.e., based on statistical criteria) and for the identification of a sufficient set of covariates for bias adjustment from such causal DAGs. Although the hdPS algorithm aims to address the problem of covariate identification for drawing causal inferences from observational data, it is not based on such formal causal results and it can thus be criticized for its simplicity and the potential for over-adjustment (e.g., M or Z-bias) [62, 2]. This report does not substantiate these concerns but does not provide evidence of strong performance from the hdPS algorithm to draw reliable causal inferences from observational data without subject-matter expertise either [64]. It does suggest however that the hdPS algorithm can be used to elicit effect signals in a discovery analysis based on claims data where background expertise or data are restricted.

The applications in this report underscore two limitations of the current hdPS algorithm: 1) The hdPS algorithm cannot be used for simultaneous identification of hdPS covariates relevant to multiple PSs over time through data pooling; 2) The hdPS algorithm does not automate the selection of possibly relevant continuous covariates such as laboratory measurements in EHR data.

Finally, the application of SL in this report was implemented for automating data-adaptive dimension reduction in settings with a large number of binary explanatory variables (hdPS covariates) and pre-specified nested subsets of these variables (arbitrarily derived here based on the default ranking of hdPS covariates provided by the hdPS algorithm). The application of SL with the hdPS algorithm could also be extended to 1) consider different 'screening' algorithms that define alternate subsets of hdPS covariates (e.g., based on a ranking of hdPS covariates using p values and a subsetting using various p value thresholds) [65], and 2) hedge against incorrect inference from arbitrary parametric assumptions [53], e.g., SL can be used to consider interactions between covariates and alternate functional forms for continuous covariates (e.g., A1c) when estimating the PSs.

Acknowledgement

The authors thank the following investigators from the HMO research network for making data from their sites available to this study: Denise M. Boudreau (Group Health), Cynthia Nakasato (Kaiser Permanente Hawaii), Gregory A. Nichols (Kaiser Permanente Northwest), Marsha A. Raebel (Kaiser Permanente Colorado), Kristi Reynolds (Kaiser Permanente Southern California), and Patrick J. O'Connor (HealthPartners).

Table 1. Covariates involved in each of the 7 estimators† of the action mechanism considered for time-dependent confounding and selection-bias adjustment. Covariates that are time-varying are listed in bold text.

Covariates‡ ($L(t)$)	$g^{\theta}_{n,expert}$	$g^{\theta}_{n,expert,hdPS}$ $g^{\theta}_{n,expert,hdPS,SL}$	$g^{\theta}_{n,claims,hdPS}$ $g^{\theta}_{n,claims,hdPS,SL}$	$g^{\theta}_{n,only,hdPS}$	$g^{\theta}_{n,claims+A1c,hdPS}$
Group health member	✓	✓	✓		✓
Kaiser Permanente (KP) Northwest member	✓	✓	✓		✓
KP Southern California member	✓	✓	✓		✓
KP Hawaii member	✓	✓	✓		✓
KP Colorado member	✓	✓	✓		✓
HealthPartners member	✓	✓	✓		✓
Female	✓	✓	✓		✓
Black	✓	✓	✓		✓
Hispanic	✓	✓	✓		✓
Hawaiian/Pacific Islander	✓	✓	✓		✓
Asian	✓	✓	✓		✓
Native American	✓	✓	✓		✓
Baseline (Bn) age	✓	✓	✓		✓
Bn median household income (census block)	✓	✓	✓		✓
Bn use of alpha-glucosidase inhibator	✓	✓	✓		✓
Bn use of long-acting insulin	✓	✓	✓		✓
Bn use of metformin	✓	✓	✓		✓
Bn use of sulfonylurea	✓	✓	✓		✓
Bn use of thiazolidinenedione	✓	✓	✓		✓
Use of 2 T2DM medications at bn	✓	✓	✓		✓
Use of 3 T2DM medications at bn	✓	✓	✓		✓
Use of 4+ T2DM medications a bn	✓	✓	✓		✓
Bn prescription-based risk score [66]	✓	✓	✓		✓
Bn diagnosis-based risk score [66]	✓	✓	✓		✓
Bn normoalbuminuria (ACR<30)	✓	✓			
HDL cholesterol	✓	✓			
LDL cholesterol	✓	✓			
Triglyceride	✓	✓			
Diastolic blood pressure	✓	✓			
Systolic blood pressure	✓	✓			
A1c	✓	✓			✓
A1c change since last quarter	✓	✓			✓
History (Hr) of arrhythmia	✓	✓			
Hr of coronary heart disease	✓	✓			
Hr of chronic heart failure	✓	✓			
Hr of cerebrovascular disease	✓	✓			
Hr of diabetic macular edema	✓	✓			
Hr of peripheral artery disease	✓	✓			
Body mass index	✓	✓			
60≤ eGFR≤89	✓	✓			
45≤ eGFR≤59	✓	✓			
30≤ eGFR≤44	✓	✓			
15≤ eGFR≤29	✓	✓			
eGFR<15	✓	✓			
Background retinopathy	✓	✓			
Mild retinopathy	✓	✓			
Moderate to severe retinopathy	✓	✓			
Proliferative retinopathy	✓	✓			
hdPS covariates		✓	✓	✓	✓

† $g^{\theta}_{n,expert}$ denotes the estimator defined by logistic models with main terms for 48 expert-selected covariates only; $g^{\theta}_{n,expert,hdPS}$ denotes the estimator defined by logistic models with main terms for 48 expert-selected covariates supplemented by up to k covariates selected by the hdPS algorithm; $g^{\theta}_{n,expert,hdPS,SL}$ denotes the estimator defined by Super Learning based on nested candidate logistic models with main terms for 48 expert-selected covariates and an increasing number of hdPS covariates; $g^{\theta}_{n,claims,hdPS}$ denotes the estimator defined by logistic models with main terms for 24 expert-selected covariates typically available in claims databases supplemented by up to k hdPS covariates; $g^{\theta}_{n,claims,hdPS,SL}$ denotes the estimator defined by Super Learning based on nested candidate logistic models with main terms for 24 expert-selected claims covariates and an increasing number of hdPS covariates; $g^{\theta}_{n,only,hdPS}$ denotes the estimator defined by logistic models with main terms for up to k hdPS covariates only; $g^{\theta}_{n,claims+A1c,hdPS}$ denotes the estimator defined by logistic models with main terms for 24 expert-selected claims covariates supplemented by 2 A1c covariates (typically available only in EHR databases) and up to k hdPS covariates.

‡ Linear combinations of the covariates encode the following information: 'White', 'Use of 1 T2DM medication at Bn', 'Bn use of either DPP4 inhibitor, exenatide, insulin combo, meglitinide, pramlintide, or short-acting insulin' , 'KP Northern California', '**eGFR≥90**', '**No retinopathy**'.

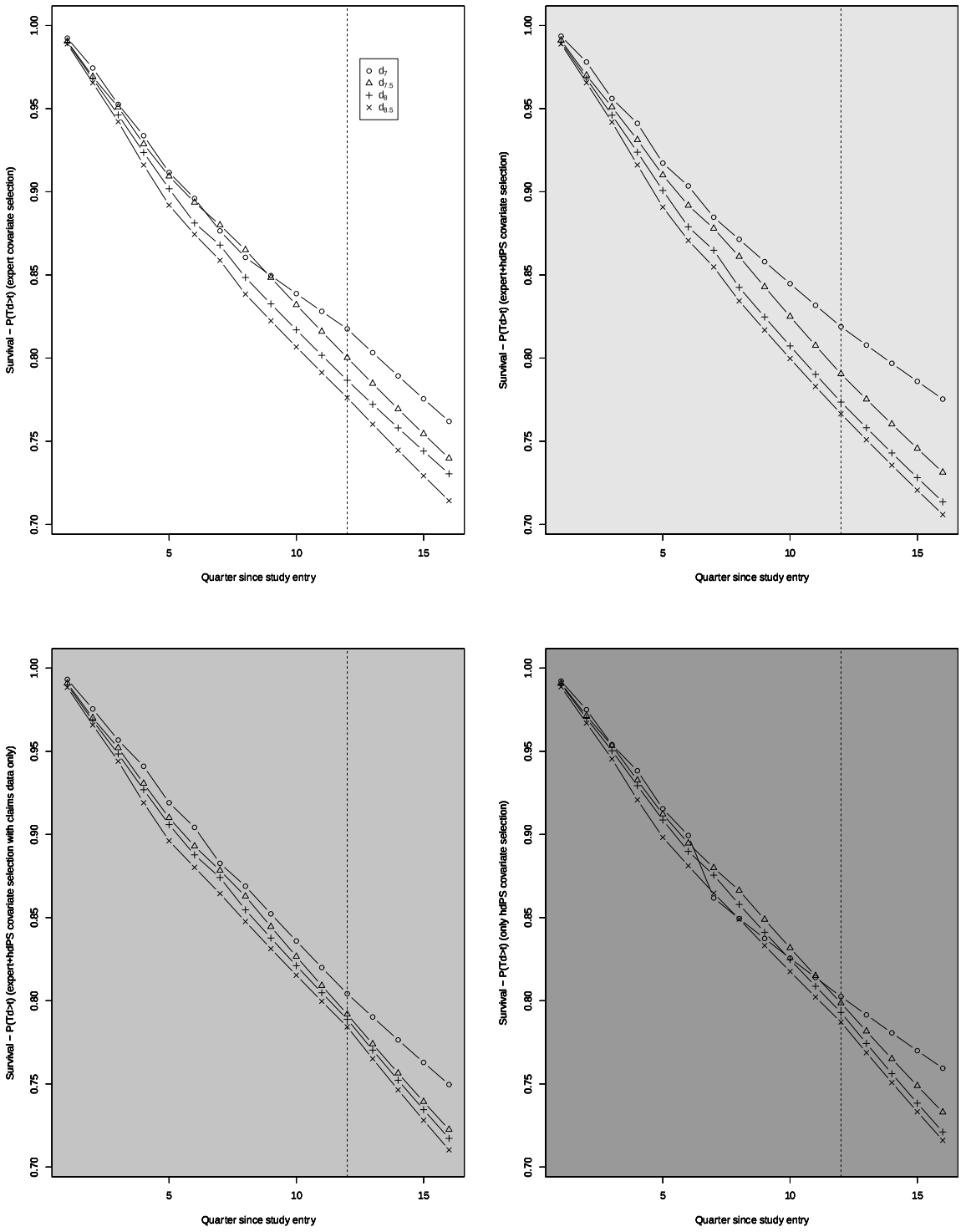

Figure 1. Each plot represents IPW estimates over 16 quarters of the four counterfactual survival curves corresponding with the four TI initiation strategies d_θ with $\theta = 7, 7.5, 8, 8.5$. The plots located at the top left, top right, bottom left, and bottom right are obtained based on the estimates $g^\theta_{n,expert}$, $g^\theta_{n,expert,hdPS}$, $g^\theta_{n,claims,hdPS}$, and $g^\theta_{n,only\ hdPS}$ of the nuisance parameter g^θ, respectively.

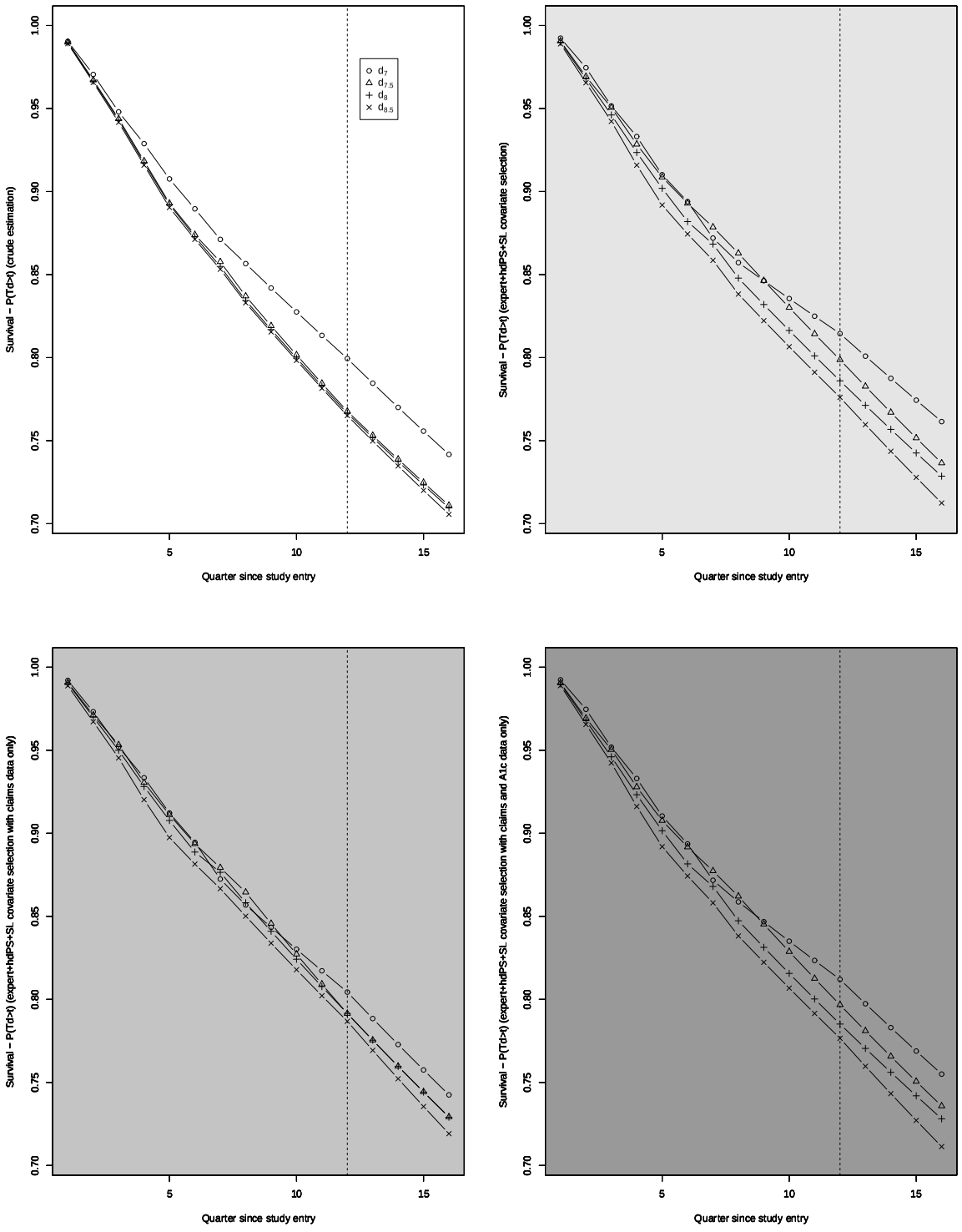

Figure 2. Each plot represents crude or IPW estimates over 16 quarters of the four counterfactual survival curves corresponding with the four TI initiation strategies d_θ with $\theta = 7, 7.5, 8, 8.5$. The plot located at the top left represents the crude estimates (equivalent to IPW estimates with equal weights for all person-time observations). The plots located at the top right, bottom left, and bottom right represent IPW estimates based on the estimates $g^\theta_{n,expert,hdPS,SL}$, $g^\theta_{n,claims,hdPS,SL}$, and $g^\theta_{n,n,claims+A1c,hdPS,SL}$ of the nuisance parameter g^θ, respectively.

Table 2. Counts of the number of estimated stabilized IPW weights within specific intervals. For each of four candidate estimators for the weights, these counts describe the distribution of the weight estimates associated with all rule-person-time observations (930956) consistent with a patient following any of the 4 TI decision rules d_θ with $\theta = 7, 7.5, 8, 8.5$. Note that if a patient follows more than one rule at a given time point, her corresponding person-time observation is replicated as many times as the number of rules followed and each replicate is assigned a separate stabilized IPW weight.

Weight range	$g^{\theta}_{n,expert}$	$g^{\theta}_{n,expert,hdPS}$	$g^{\theta}_{n,claims,hdPS}$	$g^{\theta}_{n,only\ hdPS}$
<0	0	0	0	0
[0, 0.5[	312288	326515	280150	275223
[0.5, 1[	520669	508363	533526	544085
[1, 10[	93589	88173	88405	81029
[10, 20[	4086	4592	17911	19619
[20, 30[	282	819	5253	5782
[30, 40[	18	309	1866	1828
[40, 50[	4	113	796	786
[50, 100[	3	116	1036	673
[100, 150[	4	28	140	44
≥ 150	13	1928	1873	1887

Table 3. Counts of the number of estimated stabilized IPW weights within specific intervals. For each of three Super Learning approaches for estimating the weights, these counts describe the distribution of the weight estimates associated with all rule-person-time observations (930956) consistent with a patient following any of the 4 TI decision rules d_θ with $\theta = 7, 7.5, 8, 8.5$. Note that if a patient follows more than one rule at a given time point, her corresponding person-time observation is replicated as many times as the number of rules followed and each replicate is assigned a separate stabilized IPW weight.

Weight range	$g^{\theta}_{n,expert,hdPS,SL}$	$g^{\theta}_{n,claims,hdPS,SL}$	$g^{\theta}_{n,claims+A1c,hdPS,SL}$
<0	0	0	0
[0, 0.5[	305132	268060	304173
[0.5, 1[	529237	556448	530050
[1, 10[	92173	75991	92773
[10, 20[	4152	25500	3754
[20, 30[	217	4115	177
[30, 40[	27	658	18
[40, 50[	13	135	6
[50, 100[	0	44	2
[100, 150[	4	2	1
≥ 150	1	3	2

Table 4. Comparison of inferences from IPW estimation of the 6 RDs at 3 years (12 quarters) based on 7 approaches to estimate the nuisance parameter g^{θ}. For each of these 7 approaches, the corresponding IPW point estimate, estimate of the standard error, estimate of the lower and upper bound of the 95% confidence interval, and the associated p-value are denoted by $\psi_n^{\theta_1,\theta_2}$, $\Gamma_n^{\theta_1,\theta_2}$, $\psi_n^{\theta_1,\theta_2,-}$, $\psi_n^{\theta_1,\theta_2,+}$, and p, respectively.

θ_1	θ_2	g^{θ}	$\psi_n^{\theta_1,\theta_2}$	$\Gamma_n^{\theta_1,\theta_2}$	$\psi_n^{\theta_1,\theta_2,-}$	$\psi_n^{\theta_1,\theta_2,+}$	p
8.5	8	$g^{\theta}_{n,expert}$	0.0105	5.4e-03	-1e-04	0.0211	0.053
8.5	7.5	$g^{\theta}_{n,expert}$	0.024	7.9e-03	8.5e-03	0.0395	2e-03
8.5	7	$g^{\theta}_{n,expert}$	0.0414	0.0128	0.0163	0.0664	1e-03
8	7.5	$g^{\theta}_{n,expert}$	0.0136	7.3e-03	-7e-04	0.0278	0.062
8	7	$g^{\theta}_{n,expert}$	0.0309	0.013	5.4e-03	0.0564	0.017
7.5	7	$g^{\theta}_{n,expert}$	0.0173	0.0121	-6.4e-03	0.0411	0.153
8.5	8	$g^{\theta}_{n,expert,hdPS}$	7e-03	6.4e-03	-5.5e-03	0.0196	0.271
8.5	7.5	$g^{\theta}_{n,expert,hdPS}$	0.0239	9.4e-03	5.5e-03	0.0423	0.011
8.5	7	$g^{\theta}_{n,expert,hdPS}$	0.0524	0.027	-6e-04	0.1054	0.052
8	7.5	$g^{\theta}_{n,expert,hdPS}$	0.0169	8.1e-03	1e-03	0.0328	0.037
8	7	$g^{\theta}_{n,expert,hdPS}$	0.0454	0.0258	-5.2e-03	0.0959	0.079
7.5	7	$g^{\theta}_{n,expert,hdPS}$	0.0285	0.0247	-2e-02	0.077	0.25
8.5	8	$g^{\theta}_{n,claims,hdPS}$	4.5e-03	5.9e-03	-7.1e-03	0.0161	0.444
8.5	7.5	$g^{\theta}_{n,claims,hdPS}$	7.6e-03	9.1e-03	-0.0102	0.0254	0.402
8.5	7	$g^{\theta}_{n,claims,hdPS}$	0.0199	0.0269	-0.0327	0.0726	0.458
8	7.5	$g^{\theta}_{n,claims,hdPS}$	3.1e-03	7.4e-03	-0.0115	0.0176	0.68
8	7	$g^{\theta}_{n,claims,hdPS}$	0.0154	0.0265	-0.0366	0.0674	0.562
7.5	7	$g^{\theta}_{n,claims,hdPS}$	0.0123	0.0246	-0.0359	0.0606	0.616
8.5	8	$g^{\theta}_{n,only\ hdPS}$	5.8e-03	5.8e-03	-5.6e-03	0.0172	0.319
8.5	7.5	$g^{\theta}_{n,only\ hdPS}$	0.0114	8.6e-03	-5.5e-03	0.0284	0.185
8.5	7	$g^{\theta}_{n,only\ hdPS}$	0.0153	0.021	-0.0257	0.0564	0.464
8	7.5	$g^{\theta}_{n,only\ hdPS}$	5.7e-03	7.3e-03	-8.6e-03	0.0199	0.437
8	7	$g^{\theta}_{n,only\ hdPS}$	9.5e-03	0.0206	-0.0308	0.0499	0.643
7.5	7	$g^{\theta}_{n,only\ hdPS}$	3.9e-03	0.0194	-0.0342	0.042	0.841
8.5	8	$g^{\theta}_{n,expert,hdPS,SL}$	9.9e-03	5.4e-03	-7e-04	0.0206	0.067
8.5	7.5	$g^{\theta}_{n,expert,hdPS,SL}$	0.0226	7.8e-03	7.3e-03	0.038	4e-03
8.5	7	$g^{\theta}_{n,expert,hdPS,SL}$	0.0384	0.0127	0.0134	0.0634	3e-03
8	7.5	$g^{\theta}_{n,expert,hdPS,SL}$	0.0127	7.1e-03	-1.2e-03	0.0265	0.073
8	7	$g^{\theta}_{n,expert,hdPS,SL}$	0.0284	0.0128	3.4e-03	0.0535	0.026
7.5	7	$g^{\theta}_{n,expert,hdPS,SL}$	0.0157	0.0117	-7.2e-03	0.0386	0.178
8.5	8	$g^{\theta}_{n,claims,hdPS,SL}$	4.7e-03	5.7e-03	-6.4e-03	0.0159	0.408
8.5	7.5	$g^{\theta}_{n,claims,hdPS,SL}$	4.7e-03	8.2e-03	-0.0115	0.0208	0.571
8.5	7	$g^{\theta}_{n,claims,hdPS,SL}$	0.0176	0.0128	-7.4e-03	0.0426	0.168
8	7.5	$g^{\theta}_{n,claims,hdPS,SL}$	0	6.6e-03	-0.013	0.0129	0.994
8	7	$g^{\theta}_{n,claims,hdPS,SL}$	0.0129	0.0121	-0.0108	0.0365	0.287
7.5	7	$g^{\theta}_{n,claims,hdPS,SL}$	0.0129	0.0109	-8.4e-03	0.0342	0.235
8.5	8	$g^{\theta}_{n,claims+A1c,hdPS,SL}$	8.6e-03	5.3e-03	-1.9e-03	0.019	0.108
8.5	7.5	$g^{\theta}_{n,claims+A1c,hdPS,SL}$	0.0202	7.8e-03	4.9e-03	0.0354	1e-02
8.5	7	$g^{\theta}_{n,claims+A1c,hdPS,SL}$	0.0354	0.0127	0.0105	0.0604	5e-03
8	7.5	$g^{\theta}_{n,claims+A1c,hdPS,SL}$	0.0116	7.3e-03	-2.6e-03	0.0258	0.11
8	7	$g^{\theta}_{n,claims+A1c,hdPS,SL}$	0.0269	0.0128	1.7e-03	0.052	0.036
7.5	7	$g^{\theta}_{n,claims+A1c,hdPS,SL}$	0.0153	0.0117	-7.7e-03	0.0383	0.192

Table 5. Selection of hdPS covariates for estimating the PS for TI initiation μ_1 at each time point t through $t_0 = 11$. The columns oped.dx, oped.px, ip.12.dx, ip.px, rx, lab, and utilization contain the number of encounters in each of the following 7 reduced dimension data sets relevant to the estimation of μ_1 at time t.

t	$n_{t,1}$	k	$k_{t,1}$	oped.dx	oped.px	ip.12.dx	ip.px	rx	lab	utilization
0	50377	500	500	272389	102252	10783	6539	402312	127690	144994
1	46262	500	500	377663	144743	15483	9580	395759	547923	197760
2	41255	500	500	255045	98870	14312	8722	334654	244257	139138
3	36371	500	500	225015	88476	11651	7066	292020	220341	122452
4	32552	500	500	205121	78874	11368	7307	260969	200627	110129
5	29117	500	500	191895	74427	10383	6006	233894	191042	102131
6	25901	470	470	168010	64640	10273	6464	207188	168642	89191
7	22768	407	407	149522	55647	8861	4746	180898	147321	79012
8	20202	356	356	133071	49920	8776	4695	161317	134584	70216
9	17901	310	310	121909	45574	6920	3722	143100	123003	63354
10	15591	263	263	106484	39297	6493	3339	125325	105213	55741
11	13410	220	220	93884	34906	5870	3070	107779	93480	49251

Table 6. Selection of hdPS covariates for estimating the PS for TI continuation μ_2 at each time point t through $t_0 = 11$. The columns oped.dx, oped.px, ip.12.dx, ip.px, rx, lab, and utilization contain the number of encounters in each of the following 7 reduced dimension data sets relevant to the estimation of μ_2 at time t.

t	$n_{t,2}$	k	$k_{t,2}$	oped.dx	oped.px	ip.12.dx	ip.px	rx	lab	utilization
1	2762	7	0	34525	12002	2351	1385	33352	42731	17815
2	4108	34	34	37762	13835	2458	1166	44122	34168	20612
3	4782	47	47	43765	16616	3053	1789	51991	38788	24327
4	5423	60	60	50111	18639	3694	1735	57853	47260	26939
5	5838	68	68	54777	19913	4003	1835	62629	49279	29895
6	6147	74	74	56513	21457	4385	2383	65740	52728	30502
7	6318	78	78	56565	20556	3811	1990	67093	51463	30274
8	6310	78	78	57860	20966	4511	2307	67143	56270	30726
9	6378	79	79	56702	20978	3684	2273	67085	51510	29825
10	6251	77	77	55169	20321	3457	1998	65522	49869	29405
11	5982	71	71	53532	19655	4065	1828	62594	51943	28637

Table 7. Selection of hdPS covariates for estimating the PS for right-censoring by disenrollment from the health plan μ_4 at each time point t through $t_0 = 11$. The columns oped.dx, oped.px, ip.12.dx, ip.px, rx, lab, and utilization contain the number of encounters in each of the following 7 reduced dimension data sets relevant to the estimation of μ_4 at time t.

t	$n_{t,4}$	k	$k_{t,4}$	oped.dx	oped.px	ip.12.dx	ip.px	rx	lab	utilization
0	51179	500	500	276810	103917	11066	6670	409115	129913	147421
1	49832	500	500	419581	159688	18640	11343	436568	602719	219470
2	46123	500	500	298031	115010	17369	10260	385256	284241	162709
3	42308	500	500	277224	108354	16069	9569	354298	270023	151344
4	39438	500	500	266655	101897	16788	9888	332186	260914	143570
5	36630	500	500	260206	100034	15743	8973	311776	253204	139443
6	33901	500	500	239358	91915	16189	9655	289980	236915	127754
7	31175	500	500	222981	82615	14701	7443	267105	217076	118557
8	28814	500	500	210114	77901	15369	7768	249608	211328	111310
9	26736	486	486	199916	73836	12705	6823	233163	196905	104361
10	24362	439	439	182988	66813	12441	6314	214190	178893	96688
11	21938	390	390	170110	62240	12615	6417	194306	169783	89811

Table 8. Selection of hdPS covariates for estimating the PS for right-censoring by death μ_5 at each time point t through $t_0 = 11$. The columns oped.dx, oped.px, ip.12.dx, ip.px, rx, lab, and utilization contain the number of encounters in each of the following 7 reduced dimension data sets relevant to the estimation of μ_5 at time t.

t	$n_{t,5}$	k	$k_{t,5}$	oped.dx	oped.px	ip.12.dx	ip.px	rx	lab	utilization
0	50439	500	451	272876	102428	10904	6554	402976	128238	145315
1	49113	500	500	413676	157373	18356	11178	430202	594485	216469
2	45437	500	500	293814	113111	17173	10124	379660	280778	160394
3	41697	500	500	274030	107011	15842	9457	349480	266384	149577
4	38832	500	500	263081	100321	16439	9616	327122	256697	141443
5	36123	500	500	256968	98597	15616	8885	307560	250248	137727
6	33460	500	500	236763	90781	16070	9507	286516	234659	126324
7	30735	500	500	220538	81670	14559	7326	263661	214631	117183
8	28457	500	500	207912	77070	15174	7646	246796	209018	110182
9	26388	479	479	197671	73010	12526	6752	230131	194532	103139
10	24075	433	433	181525	66236	12256	6219	211879	176919	95837
11	21679	385	385	168311	61421	12369	6198	192109	168256	88909

Table 9. Weighted combination (%) of the 12 candidate learners that define the super learners for estimating the PS for TI initiation μ_1 at each time point t through $t_0 = 11$. Eleven candidate learners are defined by nested logistic models with main terms for 48 expert-selected covariates and an increasing number of hdPS covariates. They are denoted by 'glm.x' where 'x' denotes the total number of main terms in the underlying logistic model. An additional candidate learner is defined by the intercept logistic model and denoted by 'mean'.

t	glm.48	glm.58	glm.68	glm.78	glm.88	glm.98	glm.148	glm.248	glm.348	glm.448	glm.548	mean
0	62.32	0	0	0	0	0	18.46	5.99	0	3.13	1.82	8.28
1	48.2	0	0	0	11.73	0	16.76	3.8	0	14.81	3.19	1.51
2	13.99	0	12.73	16.55	0	0	13.35	12.68	0	0	11.72	18.98
3	59.62	0	0	0	0	0	0	0	0	0	19.26	21.12
4	56.51	0	0	0	0	0	14.85	4.3	0	0	4.66	19.69
5	44.91	0	0	0	0	9.25	0	0	0	13.76	6.34	25.73
6	20.64	0	20.61	0	16.22	0	0	0	19.82	0	0	22.72
7	37.88	0	0	0	25.68	0	0	0	11.18	0	0	25.25
8	32.35	0	22.36	0	5.64	0	0	0.4	1.73	10.16		27.37
9	29.92	0	16.61	0	0	0	0	0	0	27.5		25.96
10	36.26	0	0	0	0	0	14.04	0	25.26			24.43
11	52.93	0	0	0	0	0	4	0	17.61			25.46

Table 10. Weighted combination (%) of the 12 candidate learners that define the super learners for estimating the PS for TI continuation μ_2 at each time point t through $t_0 = 11$. Eleven candidate learners are defined by nested logistic models with main terms for 48 expert-selected covariates and an increasing number of hdPS covariates. They are denoted by 'glm.x' where 'x' denotes the total number of main terms in the underlying logistic model. Note that four of these 11 candidate learners are ignored because they are identical to the learner 'glm.148' due to the limited number ($k_{t,2} < 100$) of hdPS covariates identified as relevant for PS μ_2 (Table 6). An additional candidate learner is defined by the intercept logistic model and denoted by 'mean'.

t	glm.48	glm.58	glm.68	glm.78	glm.88	glm.98	glm.148	mean
1	0							100
2	0	0	0	23.63	48.95			27.42
3	0	0	0	15.13	13.14	45.26		26.47
4	0	0	0	0	47.58	0	23.82	28.6
5	0	0	0	0	0	25.04	43.04	31.92
6	0	0	0	0	0	52.63	17.63	29.73
7	0	0	0	0	0	8.57	46.6	44.83
8	0	0	0	0	5.1	67.88	0	27.03
9	0	0	0	0	0	49.12	11.05	39.84
10	0	0	0	0	0	43.42	19.94	36.64
11	0	0	0	0	0	36.68	17.63	45.69

Table 11. Weighted combination (%) of the 12 candidate learners that define the super learners for estimating the PS for right-censoring by disenrollment from the health plan μ_4 at each time point t through $t_0 = 11$. Eleven candidate learners are defined by nested logistic models with main terms for 48 expert-selected covariates, previous TI exposure, and an increasing number of hdPS covariates. They are denoted by 'glm.x' where 'x' denotes the total number of main terms in the underlying logistic model. An additional candidate learner is defined by the intercept logistic model and denoted by 'mean'.

t	glm.49	glm.59	glm.69	glm.79	glm.89	glm.99	glm.149	glm.249	glm.349	glm.449	glm.549	mean
0	21.18	0	5.54	40.7	0	0	1.52	0	0.69	0	1.32	29.05
1	6.84	0	43.65	0	1.35	0	0	12.01	0	0	0	36.16
2	34.75	0	24.32	0	0	0	1.98	0	0	0	1.52	37.43
3	27.96	19.37	0	0	0	1.96	12.39	0	0	0	1.01	37.32
4	22.11	21.32	3.01	0	10.32	0	0	2.08	0	0	1.51	39.64
5	0	8.77	10.24	0	0	0	8.69	10.57	0	0	0.78	60.94
6	18.12	0	0	10.64	0	3.41	18.54	0	0	0	0	49.29
7	39.29	0	0	0	0	4.02	0	7.94	0	0	0	48.76
8	38.17	0	0	0	0	0	0	0	0	0	0	61.83
9	22.85	0	0	0	0	8.63	5.37	0	0	0	0.18	62.98
10	32.95	8.44	0	0	0	0	0	5.29	0	0	0.24	53.07
11	4.4	0	24.97	0	0	0	0.55	3.83	0	0		66.25

Table 12. Weighted combination (%) of the 12 candidate learners that define the super learners for estimating the PS for right-censoring by death μ_5 at each time point t through $t_0 = 11$. Eleven candidate learners are defined by nested logistic models with main terms for 48 expert-selected covariates, previous TI exposure, and an increasing number of hdPS covariates. They are denoted by 'glm.x' where 'x' denotes the total number of main terms in the underlying logistic model. An additional candidate learner is defined by the intercept logistic model and denoted by 'mean'.

t	glm.49	glm.59	glm.69	glm.79	glm.89	glm.99	glm.149	glm.249	glm.349	glm.449	glm.549	mean
0	15.68	0	0	0	0	0	0.09	0	0.88	1.51	0	81.84
1	0	50.8	5.27	0	0	0	0	2.96	0	0	0.84	40.14
2	3.7	5.23	0	0	0	5.38	0	0	0.29	2.4	0	83.01
3	2.44	0	37.06	0	0	2.58	0.84	5.1	2.86	1.01	0.1	48
4	11.73	43.4	0	0	0	0	0	0	0	3.62	1.46	39.8
5	32.71	0	0	0	0	0	11.14	0.17	1.8	0	1.2	52.98
6	5.83	23.19	0	7.76	0	0	0	0	0	1.62	0.41	61.2
7	19.89	0	21.6	0	0	0	5.7	0	0	0.62	1.01	51.18
8	0	35.04	0	0	0	0	4.48	3.29	1.27	2.42	0.75	52.76
9	27.75	0	0	0	18.16	0	0	1.22	0.16	0	0.35	52.36
10	9.4	0	34.6	4.34	0	0	0	2.28	1.03	1.25	1.83	45.28
11	2.76	28.86	0	0	0	0	7.36	1.76	0.33	0.23		58.7

References

1. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* Jul 2009; **20**(4):512–522.
2. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiol Drug Saf* Jan 2012; **21 Suppl 1**:41–49.
3. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am. J. Epidemiol.* Jun 2011; **173**(12):1404–1413.
4. Toh S, Garcia Rodriguez LA, Hernan MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol Drug Saf* Aug 2011; **20**(8):849–857.
5. Rosenbaum PR. The consequence of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, General* 1984; **147**:656–66.
6. Robins JM. Association, causation and marginal structural models. *Synthese* 1999; **121**:151–179.
7. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**(5):615–625.
8. Robins J. Marginal Structural Models. *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, 1998; 1–10.
9. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**(4):669–710.
10. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* Jan 1999; **10**(1):37–48.
11. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd edn., Cambridge University Press, Cambridge, 2009.
12. Pearl J. An introduction to causal inference. *Int J Biostat* 2010; **6**(2):Article 7.
13. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* May 2001; **12**(3):313–320.
14. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am. J. Epidemiol.* Jan 2002; **155**(2):176–184.
15. Nathan DM, Buse JB, Davidson MB, Heine RJ, Holman RR, , Sherwin R, Zinman B. Management of Hyperglycemia in Type 2 Diabetes: A Consensus Algorithm for the Initiation and Adjustment of Therapy: A consensus statement from the American Diabetes Association and the European Association for the Study of Diabetes. *Diab Care* 2006; **29**:1963–72.
16. Skyler JS, Bergenstal R, Bonow RO, Buse J, Deedwania P, Gale EA, Howard BV, Kirkman MS, Kosiborod M, Reaven P, *et al.*. Intensive Glycemic Control and the Prevention of Cardiovascular Events: Implications of the ACCORD, ADVANCE, and VA Diabetes Trials: A position statement of the American Diabetes Association and a scientific statement of the American College of Cardiology Foundation and the American Heart Association. *Diab Care* 2009; **32**:187–92.
17. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med.* 1993; **329**:977–86.
18. Nathan DM, Cleary PA, Backlund JY, Genuth SM, Lachin JM, Orchard TJ, Raskin P, Zinman B. Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Study Research Group. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. *N Engl J Med* 2005; **22**(353):2643–53.
19. UK Prospective Diabetes Study (UKPDS) Group. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). *Lancet* 1998; **352**:854–865.
20. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HA. 10-year follow-up of intensive glucose control in type 2 diabetes. *N Engl J Med* 2008; **359**:1577–89.
21. Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med* 2008; **358**:2545–9.
22. ADVANCE Collaborative Group. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2008; **358**:2560–2.
23. Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, Zieve FJ, Marks J, Davis SN, Hayward R, *et al.*. Glucose Control and Vascular Complications in Veterans with Type 2 Diabetes. *N Engl J Med* 2009; **360**:129–39.
24. Ray KK, Seshasai SR, Wijesuriya S, Sivakumaran R, Nethercott S, Preiss D, Erqou S, Sattar N. Effect of intensive control of glucose on cardiovascular outcomes and death in patients with diabetes mellitus: a meta-analysis of randomised controlled trials. *Lancet* 2009; **373**:1765–72.
25. Gerstein HC, Miller ME, Byington RP, Goff DC, Bigger JT, Buse JB, Cushman WC, Genuth S, Ismail-Beigi F, Grimm RH, *et al.*. Effects of intensive glucose lowering in type 2 diabetes. *N. Engl. J. Med.* Jun 2008; **358**(24):2545–2559.
26. Patel A, MacMahon S, Chalmers J, Neal B, Billot L, Woodward M, Marre M, Cooper M, Glasziou P, Grobbee D, *et al.*. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.* 2008; **358**(24):2560–2572.
27. Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, Zieve FJ, Marks J, Davis SN, Hayward R, *et al.*. Glucose control and vascular complications in veterans with type 2 diabetes. *N. Engl. J. Med.* Jan 2009; **360**(2):129–139.
28. Ismail-Beigi F, Craven T, Banerji MA, Basile J, Calles J, Cohen RM, Cuddihy R, Cushman WC, Genuth S, Grimm RH, *et al.*. Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes: an analysis of the ACCORD randomised trial. *Lancet* Aug 2010; **376**:419–430.
29. O'Connor PJ, Ismail-Beigi F. Near-normalization of glucose and microvascular diabetes complications: data from ACCORD and ADVANCE. *Therapeutic Advances in Endocrinology and Metabolism* 2011; **2**(1):17–26.
30. *Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection* 2002.
31. Nathan DM, Buse JB, Davidson MB, Ferrannini E, Holman RR, Sherwin R, Zinman B. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* Jan 2009; **32**(1):193–203.
32. Vogt TM, Elston-Lafata J, Tolsma D, Greene SM. The role of research in integrated healthcare systems: the HMO Research Network. *Am J Manag Care* 2004; **10**(9):643–8.
33. Murphy S, van der Laan M, Robins J. Marginal mean models for dynamic treatment regimens. *Journal of the American Statistical Association* 2001; **96**:1410–1424.
34. van der Laan M, Petersen ML. Causal Effect Models for Realistic Individualized Treatment and Intention to Treat Rules. *Int J Biostat* 2007; **3**(1):Article 3.
35. Robins J RA Orellana L. Estimation and extrapolation of optimal treatment and testing strategies. *Stat Med* 2008; **27**(23):4678–4721.

36. Hernan MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin. Pharmacol. Toxicol.* Mar 2006; **98**(3):237–242.
37. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernan MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *The international journal of biostatistics* 2010; **6**(18).
38. Hernan MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine* 2002; **21**:1689–1709.
39. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
40. Neugebauer R, Fireman B, Roy JA, O'Connor PJ, Selby JV. Dynamic marginal structural modeling to evaluate the comparative effectiveness of more or less aggressive treatment intensification strategies in adults with type 2 diabetes. *Pharmacoepidemiol Drug Saf* May 2012; **21 Suppl 2**:99–113.
41. Neugebauer R, Fireman B, Roy JA, O'Connor PJ. Dynamic marginal structural modeling to assess impact of more versus less intensive glycemic control strategies on microvascular and macrovascular outcomes in 58,000 adults with type 2 diabetes mellitus. *Submitted to Diabetes care.* 2013; .
42. Pearl J. Causal inference in statistics: An overview. *Statistics Surveys* 2009; **3**:96–146.
43. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* Nov 2009; **20**(6):880–883.
44. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology* Nov 2010; **21**(6):872–875.
45. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* Sep 2000; **11**(5):561–570.
46. Cole SR, Hernan MA, Robins JM, Anastos K, Chmiel J, Detels R, Ervin C, Feldman J, Greenblatt R, Kingsley L, *et al.*. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am. J. Epidemiol.* Oct 2003; **158**(7):687–694.
47. Bodnar LM, Davidian M, Siega-Riz AM, Tsiatis AA. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *Am. J. Epidemiol.* May 2004; **159**(10):926–934.
48. The HIV-CAUSAL Collaboration. When to initiate combined antiretroviral therapy to reduce mortality and aids-defining illness in HIV-infected persons in developed countries: an observational study. *Ann Intern Med* 2011; **154**(8):509–515.
49. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* Sep 2008; **168**:656–664.
50. Bembom O, van der Laan MJ. Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators. *Technical Report 230*, Division of Biostatistics, UC Berkeley 2008.
51. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* Feb 2012; **21**(1):31–54.
52. Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital. Pharmacoepidemiology Toolbox including hd-PS. http://www.hdpharmacoepi.org/ 2012. Version 2.4.10.
53. Neugebauer R, Fireman B, Roy JA, Raebel MA, Nichols GA, O'Connor PJ. Super learning to hedge against incorrect inference from arbitrary parametric assumptions in marginal structural modeling. *Journal of Clinical Epidemiology* Accepted, 2013; .
54. van der Laan M, Polley E, Hubbard A. Super learner. *Statistical Applications in Genetics and Molecular Biology* 2007; **6**(1).
55. Dudoit S, van der Laan M. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology* 2005; **2**:131–154.
56. van der Laan M, Dudoit S, Keles S. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1).
57. van der Vaart A, Dudoit S, van der Laan M. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions* 2006; **24**(3):351–371.
58. van der Laan M, Dudoit S, van der Vaart A. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions* 2006; **24**(3):373–395.
59. Neugebauer R, van der Laan, MJ. Why prefer double robust estimates. *Journal of Statistical Planning and Inference* 2005; **129**(1-2):405–26.
60. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am. J. Epidemiol.* Jun 2006; **163**(12):1149–1156.
61. Schneeweiss S, Rassen JA. Letter to the editor. *Pharmacoepidemiology and drug safety* 2011; **20**.
62. Joffe MM. Exhaustion, automation, theory, and confounding. *Epidemiology* Jul 2009; **20**(4):523–524.
63. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd edn., Springer: New York, 2009.
64. Robins J, Wasserman L. On the impossibility of inferring causation from association without background knowledge. *Computation, Causation, and Discovery*, Glymour C, Cooper G (eds.). AAAI Press/The MIT Press: Menlo Park, CA, Cambridge, MA, 1999; 305–321.
65. Neugebauer R, Chandra M, Paredes A, Graham D, McCloskey C, Go A. A Marginal Structural Modeling Approach with Super Learning for a Study on Oral Bisphosphonate Therapy and Atrial Fibrillation. *Journal of Causal Inference* 2012; **Accepted**.
66. DxCG Inc., Boston, MA. *Risk Smart®Models and Methodologies Guide* November 2002.