Prevention Science (2019) 20:598–608
https://doi.org/10.1007/s11121-019-0980-3

Evaluating Community-Based Translational Interventions Using Historical Controls: Propensity Score vs. Disease Risk Score Approach

Luohua Jiang[1] · Shuai Chen[2] · Janette Beals[3] · Juned Siddique[4] · Richard F. Hamman[5] · Ann Bullock[6] · Spero M. Manson[3] · the Special Diabetes Program for Indians Diabetes Prevention Demonstration Project

Published online: 12 February 2019

Abstract

Many community-based translations of evidence-based interventions are designed as one-arm studies due to ethical and other considerations. Evaluating the impacts of such programs is challenging. Here, we examine the effectiveness of the lifestyle intervention implemented by the Special Diabetes Program for Indians Diabetes Prevention (SDPI-DP) demonstration project, a translational lifestyle intervention among American Indian and Alaska Native communities. Data from the landmark Diabetes Prevention Program placebo group was used as a historical control. We compared the use of propensity score (PS) and disease risk score (DRS) matching to adjust for potential confounder imbalance between groups. The unadjusted hazard ratio (HR) for diabetes risk was 0.35 for SDPI-DP lifestyle intervention vs. control. However, when relevant diabetes risk factors were considered, the adjusted HR estimates were attenuated toward 1, ranging from 0.56 (95% CI 0.44–0.71) to 0.69 (95% CI 0.56–0.96). The differences in estimated HRs using the PS and DRS approaches were relatively small but DRS matching resulted in more participants being matched and smaller standard errors of effect estimates. Carefully employed, publicly available randomized clinical trial data can be used as a historical control to evaluate the intervention effectiveness of one-arm community translational initiatives. It is critical to use a proper statistical method to balance the distributions of potential confounders between comparison groups in this kind of evaluations.

Keywords Comparative effectiveness evaluation · Disease risk score · Prognostic score · Propensity score · Single-arm intervention · Historical controls

Introduction

The growing, global pandemic of type 2 diabetes (T2DM) is a major public health concern. Randomized clinical trials (RCTs) have convincingly shown that lifestyle interventions consisting of exercise and diet behavioral modifications are highly efficacious in preventing or delaying the onset of T2DM for those at risk (Knowler et al. 2002; Norris et al. 2005; Pan et al. 1997; Tuomilehto et al. 2001). A critical next step in stemming this epidemic is to translate interventions developed in rigorously controlled clinical trials into everyday settings. A number of translational diabetes prevention initiatives have yielded promising results, albeit considerable variability in outcomes exists and the magnitude of risk reduction is generally less than that achieved in landmark clinical trials (Cefalu et al. 2016; Dunkley et al. 2014; Wareham 2015).

While implementing interventions of proven efficacy, these translational projects often forego inclusion of a randomized control group given ethical concerns about denying or delaying treatment. Thus, quasi-experimental designs are

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11121-019-0980-3) contains supplementary material, which is available to authorized users.

✉ Luohua Jiang
lhjiang@uci.edu

[1] Department of Epidemiology, School of Medicine, University of California Irvine, Irvine, CA 92697-7550, USA

[2] Division of Biostatistics, Department of Public Health Sciences, University of California Davis, Davis, CA, USA

[3] Centers for American Indian and Alaska Native Health, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[4] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

[5] Department of Epidemiology, Colorado School of Public Health, LEAD Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[6] Division of Diabetes Treatment and Prevention, Indian Health Service, Rockville, MD, USA

common in community-based translational research (Henry et al. 2017). Evaluations of such designs must be conducted with great caution. They are prone to mis-estimation of intervention effects due to potential biases resulting from selective enrollment and/or lack of control for placebo, historical, and other effects (Buntin et al. 2009; Flamm et al. 2012). Hence, appropriate analytical methods that can account for the potential biases but are also practical for program evaluation in routine settings are highly desirable.

Data from the National Institutes of Health funded Diabetes Prevention Program (DPP), a large-scale randomized trial that provided evidence for the efficacy of lifestyle intervention in a diverse sample of US adults (Knowler et al. 2002), has become publicly available through the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Data Repository. These data provide the potential opportunity to serve as a historical control to evaluate the effectiveness of translated versions of the DPP lifestyle intervention, provided the translational projects have similar eligibility criteria and outcome measures. The Special Diabetes Program for Indians Diabetes Prevention (SDPI-DP) demonstration project (Jiang et al. 2013) meets these criteria. Based on DPP and using very similar eligibility criteria and outcome measures, this one-armed intervention project implemented the DPP lifestyle intervention among over 2500 American Indian and Alaska Native (AI/AN) participants across 36 diverse grantee sites.

Simply combining intervention data with historical control data and analyzing the combined data as if they were from a single randomized trial can result in a biased treatment effect estimate due to imbalance in observed confounders between treatment conditions. A statistical method that has often been used when considering causal effects in observational studies without randomization is the propensity score (PS) approach whereby the distributions of potential confounders between comparison groups are statistically balanced (Guo and Fraser 2009; Rosenbaum and Rubin 1983). Another type of summary score, the disease risk score (DRS), has been suggested as an alternative method to control for confounding (Sturmer et al. 2005). The DRS method has been applied widely to predict the occurrence of chronic diseases, such as cardiovascular disease and diabetes (D'Agostino Sr. et al. 2008; Kahn et al. 2009; Lee et al. 2006; Lindstrom and Tuomilehto 2003; Noble et al. 2011; Wilson et al. 1998). Recently, simulation studies have shown when the PS distributions do not overlap well between the comparison groups, the DRS approach might allow researchers to assess treatment effects in a larger proportion of the treated population and yield effect estimates with improved precision. In this study, we explored and compared the use of PS and DRS approaches in evaluating the effectiveness of SDPI-DP, using the DPP data as historical controls.

Research Design and Methods

Data Sources

SDPI-DP The SDPI-DP program is a demonstration project designed to reduce diabetes incidence among American Indians and Alaska Natives (AI/ANs) with prediabetes through local translation of the DPP lifestyle intervention. The details of this project are described elsewhere (Jiang et al. 2013). Briefly, 36 AI/AN health care programs implemented the 16-session *Lifestyle Balance Curriculum* drawn from the DPP (Knowler et al. 2002). The primary goal of the intervention was to achieve and maintain a weight reduction of at least 7% of initial body weight through a healthy diet and increased physical activity. Grantees used the DPP curriculum covering diet, exercise, and behavior modification to help participants achieve this goal. Adaptation to local culture and conditions was allowed, provided the same basic information was presented and adaptations were well documented.

The eligibility criteria of SDPI-DP included being AI/AN (based on eligibility to receive Indian Health Service [IHS] services), being 18+ years of age, no previous diagnosis of diabetes, and having either impaired fasting glucose (i.e., a fasting blood glucose (FBG) level of 100 to 125 mg/dL and an oral glucose tolerance test (OGTT) result < 200 mg/dL) or impaired glucose tolerance (i.e., an OGTT of 140 to 199 mg/dL 2 h after a 75-g oral glucose load and a FBG level < 126 mg/dL). Enrollment began January 2006. The analyses here included baseline and annual data for up to 3 years for 2553 participants who completed the baseline assessment and started intervention by July 31, 2008.

During the design phase of SDPI-DP, the inclusion of a control group was deemed an unethical delay of treatment due to strong evidence supporting the efficacy of the lifestyle intervention (Knowler et al. 2002; Norris et al. 2005; Pan et al. 1997; Tuomilehto et al. 2001). Rather, the goal of SDPI-DP was to pursue a comprehensive public health evaluation of the translation of a proven intervention in diverse AI/AN communities. Therefore, all SDPI-DP participants received the intervention. Previous analyses of SDPI-DP data based on one-arm design revealed significant improvements in weight and a number of secondary outcomes (Jiang et al. 2013). However, without a control group, a causal interpretation of intervention effectiveness is not straightforward.

DPP We used de-identified DPP data obtained from the NIDDK Data Repository as historical controls (Cuticchia et al. 2006). The DPP was a RCT conducted at 27 US sites enrolling individuals at high risk for diabetes. Its methods are published elsewhere (The Diabetes Prevention Program 1999). Briefly, eligible participants were randomly assigned to one of the three groups: (1) placebo medication twice daily and standard lifestyle recommendations; (2) metformin twice

daily and standard lifestyle recommendations; or (3) intensive lifestyle modification. The first group (placebo) from DPP served as the historical control to evaluate the effectiveness of SDPI-DP.

The eligibility criteria for most DPP participants were ≥ 25 years old, BMI ≥ 24, FBG level of 95–125 mg/dL, and OGTT 2-h result of 140–199 mg/dL. Compared to SDPI-DP, DPP had more stringent eligibility criteria because its participants needed to have both impaired fasting glucose and impaired glucose tolerance; additionally, BMI defined eligibility in DPP.

We obtained the DPP data following established policies of the NIDDK Data Repository. The University of Colorado Anschutz Medical Center and University of California Irvine IRBs approved this supplementary analysis.

Measures

Both studies have the same primary outcome: incidence of diabetes diagnosed by an annual OGTT or a semiannual FBG test, according to the American Diabetes Association 2004 criteria: a FBG ≥ 126 mg/dL or a 2-h result ≥ 200 mg/dL after a 75-g oral glucose load. In addition to the semiannual tests, FBG was measured if symptoms suggestive of diabetes developed. The diagnosis required confirmation by a second test, usually within 6 weeks of the first test.

Basic demographic characteristics and key diabetes risk factors comprised the common baseline measurements of SDPI-DP and DPP. Age, gender, and race were available in both data sets. Given the importance of de-identification in the public available DPP data, race/ethnic groups were simply coded as Caucasian, African American, Hispanic, and All Other. Similarly, age at baseline was collapsed into 5-year age groups, with truncation of those < 40 and those ≥ 65.

Well-known diabetes risk factors measured at baseline in both data sources are BMI, family history of T2DM, FBG, OGTT, systolic blood pressure (SBP), diastolic blood pressure (DBP), triglycerides, high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C). In both studies, baseline physical examination included measurements of height, weight, and sitting SBP and DBP. BMI was calculated from height and weight (kg/m^2). Blood was drawn from DPP and SDPI-DP participants after a 9–12-h fast to measure blood glucose level, triglycerides, HDL-C, and LDL-C.

Statistical Analysis

To quantify the effectiveness of the lifestyle intervention among SDPI-DP participants, Cox proportional hazards regression models were constructed after merging the SDPI-DP data with those from the DPP placebo group. We investigated four approaches to estimate the hazard ratio (HR) of developing T2DM among SDPI-DP vs. the DPP placebo group: (1) no adjustment for confounders; (2) multivariable regression adjustment; (3) propensity score matching; (4) disease risk score matching. We describe each of these methods in turn.

Unadjusted Estimate After merging the two data sources, the Cox regression models with a dummy variable indicating SDPI-DP (intervention group) or DPP cohort (control group) as the only independent variable produced the estimate for unadjusted HR of diabetes between SDPI-DP and DPP placebo participants.

Multivariable Regression Adjustment The Cox regression models were then adjusted by previously reported diabetes risk factors identified in three published diabetes prediction models validated by Mann et al. in a multiethnic population of US adults (Mann et al. 2010). These were based on the Framingham Offspring (FO) Study (Wilson et al. 2007), Atherosclerosis Risk in Communities (ARIC) Study (Schmidt et al. 2005), and the San Antonio Heart (SAH) Study (Stern et al. 2002). The risk factors included in the FO risk prediction model were overweight and obesity, impaired fasting glucose, low HDL-C, elevated triglycerides, high blood pressure, and parental history of diabetes. The ARIC model included age, height, waist circumference, black race/ethnicity, SBP, FBG, HDL-C, triglycerides, and parental history of diabetes. Finally, the San Antonio diabetes risk prediction model included age, sex, Mexican-American ethnicity, FBG, SBP, HDL-C, BMI, and family history of diabetes. In addition to the risk factors included in these three models, we added another variable, OGTT 2-h test result, to each model to account for the potential unbalance in it between the SDPI-DP and DPP participants. Thus, for each risk prediction model, we estimated adjusted HR with and without OGTT 2-h result included in the model.

Propensity Score Method Some drawbacks of using regression adjustment include the strong assumptions in model specification of the outcome, computational complexity when many potential confounders exist, and the danger of extrapolation in situations with insufficient common support between groups of comparison. The propensity score (PS) approach was proposed to overcome these challenges. The PS, as defined by Rosenbaum and Rubin (Rosenbaum and Rubin 1983), is the predicted probability of exposure/intervention for a given vector of observed covariates. It is considered a balancing score and is usually estimated using a logistic regression model with the exposure variable as the response and all variables related to the outcome of interest as the covariates (D'Agostino Jr. 1998; Guo and Fraser 2009; Rosenbaum and Rubin 1983). Here, a logistic regression model was used to estimate the PS of belonging to the SDPI-DP cohort. The covariates included for estimating the PS were the same as the adjustment variables

used in the multivariable regression models described above in order to compare the methods. After the PS has been estimated, it can then be used in various ways to obtain adjusted estimates of intervention effectiveness.

This study focused on the PS matching approach. We used a nearest-neighbor methodology with a caliper set to 0.1 of the standard deviation of the PS (Rosenbaum and Rubin 1985) to match SDPI-DP subjects to DPP placebo subjects with a maximum of one control matched per appropriate case. After matching on the PS, estimates of intervention effectiveness were calculated using the Cox regression models. The models include the exposure as the only independent variable as the comparison samples have already been matched. Covariate balance before and after matching was checked by calculating the absolute standardized difference (ASD) of each variable between the two treatment groups.

Disease Risk Score Method The disease risk score (DRS), sometimes referred to as prognostic score, is a score summarizing the associations between a set of observed covariates and a disease outcome, such as diabetes incidence. It was originally proposed by Miettinen and was called a "multivariate confounder score" to overcome the difficulty of multiple cross-classification in stratified analysis based on a number of confounding factors (Miettinen 1976). Specifically, if the association between observed covariates and the outcome follows a generalized linear model, the DRS score is usually calculated as the conditional expectation of the outcome given the values of the observed covariates using the estimated parameters of the generalized linear model. In most previous studies using DRS, it was developed for prediction purposes and has been mainly used to estimate the probability of developing a disease among individuals not exposed to an intervention.

Recently, it also has been suggested that the DRS could be used as a balancing score similar to the PS, because the DRS can be used to account for not only unbalanced propensities of exposure but also different disease risks between comparison groups (Hansen 2008). Recent simulation studies suggested that the DRS could be a reasonable alternative to the PS approach when the association between covariates and exposure is, at most, moderate (Arbogast et al. 2008; Arbogast and Ray 2009, 2011). Additionally, when the PS distributions are severely separated, matching on DRS is often able to match a larger proportion of the treated population and yield effect estimates with improved precision (Wyss et al. 2015). In this study, DRSs were calculated based on the three diabetes risk prediction models described above: FO, ARIC, and SAH. DRS matching was then performed using an approach similar to the PS methods. Because none of the original version of these diabetes risk models included the OGTT 2-h test result, we fitted the three DRS models with and without OGTT 2-h result within the DPP placebo group and calculated the DRSs for all participants afterwards.

The confounding control ability of DRS matching cannot be evaluated through balance checks commonly used for the PS approach. Here, we used a newly proposed alternative, the "dry-run" analysis (Wyss et al. 2017), to assess the ability of DRS matching in controlling potential confounding effects. Briefly, in the dry-run analysis, we split the unexposed population (i.e., the DPP placebo group) into "pseudo-exposed" and "pseudo-unexposed" groups so that the differences on observed covariates between the two "pseudo" groups are similar to those between the DPP placebo and the SDPI-DP participants. We then evaluated the ability of each DRS model in confounding control by calculating the pseudo-bias, defined as the difference between the pseudo-effect estimate and the true null effect. A pseudo-bias close to 0 indicates adequate ability of the DRS matching approach in retrieving the unconfounded null estimate.

Results

Table 1 compares baseline characteristics of the DPP placebo group and the SDPI-DP participants. Overall, compared to the DPP, SDPI-DP participants were younger (46.8 vs. 50.3 years old), included more females (75% vs. 68%) and more obese participants (80% vs. 68%). They also had significantly lower FBG, OGTT 2-h result, and LDL-C level, but significantly higher weight, waist circumference, and systolic BP. The SDPI-DP subgroup ($n = 648$) who met the DPP eligibility criteria had similar differences when compared with the DPP placebo group. Here, however, FBG and OGTT 2-h result were not significantly different between the SDPI-DP subgroup and the DPP placebo group.

Unadjusted and adjusted HRs estimated by various statistical models are presented in Table 2. When we compare all SDPI-DP participants ($N = 2553$) to the DPP placebo group ($N = 1030$), the unadjusted HR for diabetes risk is 0.35, suggesting a 65% risk reduction by the lifestyle intervention among SDPI-DP participants. When OGTT 2-h test result was not included in the adjustment methods, all the statistical models produced similar HR estimates, ranging from 0.29 to 0.41, with a very small P value (< 0.0001). However, when OGTT 2-h test result was included, all the adjusted HR estimates were larger than the unadjusted HR, ranging from 0.56 to 0.69 (Table 2), indicating a weaker effectiveness of the SDPI-DP lifestyle intervention. These are close to the unadjusted HR comparing the SDPI-DP participants who met the DPP eligibility criteria to the DPP placebo group, which is 0.64 (95% CI 0.49–0.84). Regardless of including OGTT 2-h test results or not, the DRS matching resulted in more pairs of SDPI-DP and DPP participants to be matched. All the estimated HRs are significantly different from 1 ($P < 0.05$), suggesting lifestyle intervention was significantly effective at reducing diabetes risk among SDPI-DP participants.

Table 1 Baseline characteristics of DPP placebo group and SDPI-DP participants

Characteristics	DPP placebo ($N = 1030$)	All SDPI-DP participants ($N = 2553$)		SDPI-DP participants met DPP eligibility criteria ($N = 648$)	
	N (%)	*N* (%)	*P* value[a]	*N* (%)	*P* value[b]
Gender			< 0.001		< 0.001
Female	699 (67.9%)	1901 (74.5%)		520 (80.2%)	
Male	331 (32.1%)	652 (25.5%)		128 (19.8%)	
Age group			< 0.001		< 0.001
< 40	151 (14.7%)	731 (28.6%)		148 (22.8%)	
40 to < 50	378 (36.7%)	774 (30.3%)		186 (28.7%)	
50 to < 60	301 (29.2%)	645 (25.3%)		166 (25.6%)	
≥ 60	200 (19.4%)	403 (15.8%)		148 (22.8%)	
Family history of type 2 diabetes			< 0.001		< 0.001
No	335 (32.5%)	512 (20.2%)		125 (19.3%)	
Yes	695 (67.5%)	2026 (79.8%)		522 (80.7%)	
BMI			< 0.001		< 0.001
< 30	326 (31.6%)	512 (20.1%)		100 (15.4%)	
30 to < 40	519 (50.3%)	1427 (56.0%)		380 (58.6%)	
≥ 40	185 (17.9%)	612 (24.0%)		168 (25.9%)	
	Mean (SD)	Mean (SD)	*P* value[a]	Mean (SD)	*P* value[b]
FBG (mg/dL)	107.4 (7.8)	104.6 (9.2)	< 0.001	108.0 (8.2)	0.15
OGTT 2-h glucose (mg/dL)	164.6 (17.2)	122.9 (35.2)	< 0.001	164.7 (18.5)	0.86
Weight (lbs)	208.9 (44.6)	217.5 (51.2)	< 0.001	217.7 (48.7)	< 0.001
Waist circumference (cm)	105.1 (14.4)	111.7 (15.9)	< 0.001	111.8 (15.6)	< 0.001
Systolic BP (mmHg)	123.8 (14.4)	126.6 (15.0)	< 0.001	128.2 (15.5)	< 0.001
Diastolic BP (mmHg)	78.2 (9.2)	78.8 (10.1)	0.08	78.9 (10.2)	0.13
HDL-C (mg/dL)	44.7 (11.4)	45.0 (12.1)	0.47	44.8 (12.1)	0.96
LDL-C (mg/dL)	125.2 (33.3)	111.7 (31.3)	< 0.001	111.3 (31.0)	< 0.001
Triglycerides (mg/dL)	167.2 (92.6)	163.3 (98.1)	0.26	175.8 (105.7)	0.09

Abbreviations: BP, blood pressure; DPP, Diabetes Prevention Program; FBG, fasting blood glucose; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; OGTT, oral glucose tolerance test; SDPI-DP, Special Diabetes Program for Indians Diabetes Prevention Program; SD, standard deviation

[a] *P* value for chi-square test or two-sample *t* test comparing the DPP placebo group and the SDPI-DP participants

[b] *P* value for chi-square test or two-sample *t* test comparing the DPP placebo group and the SDPI-DP participants who met the DPP eligibility criteria

When comparing the DPP placebo participants and the SDPI-DP participants who met the DPP eligibility criteria, the unadjusted HR for diabetes risk is 0.64, indicating a 36% risk reduction by the SDPI-DP lifestyle intervention. This HR estimate is closer to the HRs estimated using the adjustment models with OGTT 2-h result included in Table 2, but is much larger than the unadjusted HR when comparing all SDPI-DP participants with the DPP placebo group. The adjusted HRs for SDPI-DP subgroup vs. the DPP placebo group are all slightly smaller than the unadjusted HR, but fairly close to it in general, regardless of including OGTT 2-h result in the model or not (Supplementary Table 1).

Figure 1 illustrates the PS and DRS distributions by intervention group before matching. The PS distributions of the SDPI-DP and the DPP placebo group do not overlap very well with each other (overlapping coefficient [a measure of the agreement between two probability distributions (Inman and Bradley 1989)] ranges from 0.34 to 0.40) with many SDPI-DP participants having very high probability of belonging to the intervention group. Meanwhile, for DRS, the overlapping coefficients are much larger (0.49–0.66).

As shown in Fig. 2, before PS matching, the absolute standardized differences (ASDs) between the two treatment groups were larger than 0.1 for most of the diabetes risk factors included in our regression models. However, after matching on PS scores without OGTT 2-h test results included in the PS model, the ASDs were smaller or close to 0.1 for all risk factors except OGTT 2-h. Furthermore, after matching on PS scores with OGTT 2-h test results in the model, the ASD was smaller than 0.1 even for OGTT 2-h. Particularly, upon matching on the PS scores calculated based on the ARIC model, the ASDs are smaller than 0.1 for all risk factors except

Table 2 Intervention effectiveness of SDPI-DP based on different estimation methods using data from SDPI-DP participants and DPP placebo group

Method	HR	SE	95% CI	*P* value
Unadjusted ($N_T = 2553$, $N_C = 1030$) (All SDPI-DP vs. DPP)	0.35	0.10	(0.29, 0.43)	< 0.0001
Unadjusted ($N_T = 648$, $N_C = 1030$) (SDPI-DP subgroup* vs. DPP)	0.64	0.14	(0.49, 0.84)	0.001
Without OGTT 2-h result in the models				
Regression adjustment				
FO ($N_T = 2553$, $N_C = 1030$)	0.29	0.11	(0.24, 0.36)	< 0.0001
ARIC ($N_T = 2553$, $N_C = 1030$)	0.35	0.11	(0.28, 0.43)	< 0.0001
SAH ($N_T = 2553$, $N_C = 1030$)	0.38	0.11	(0.31, 0.47)	< 0.0001
Propensity score matching				
FO ($N_T = 716$, $N_C = 716$)	0.32	0.15	(0.24, 0.43)	< 0.0001
ARIC ($N_T = 880$, $N_C = 880$)	0.35	0.14	(0.27, 0.46)	< 0.0001
SAH ($N_T = 932$, $N_C = 932$)	0.41	0.13	(0.32, 0.53)	< 0.0001
Disease risk score matching				
FO ($N_T = 857$, $N_C = 857$)	0.32	0.14	(0.24, 0.42)	< 0.0001
ARIC ($N_T = 1011$, $N_C = 1011$)	0.33	0.13	(0.26, 0.43)	< 0.0001
SAH ($N_T = 1001$, $N_C = 1001$)	0.39	0.13	(0.31, 0.50)	< 0.0001
With OGTT 2-h result in the models				
Regression adjustment				
FO ($N_T = 2553$, $N_C = 1030$)	0.56	0.12	(0.44, 0.71)	< 0.001
ARIC ($N_T = 2553$, $N_C = 1030$)	0.58	0.12	(0.45, 0.73)	< 0.001
SAH ($N_T = 2553$, $N_C = 1030$)	0.63	0.12	(0.50, 0.79)	< 0.001
Propensity score matching				
FO ($N_T = 535$, $N_C = 535$)	0.66	0.14	(0.50, 0.87)	0.004
ARIC ($N_T = 556$, $N_C = 556$)	0.69	0.15	(0.56, 0.96)	0.013
SAH ($N_T = 603$, $N_C = 603$)	0.64	0.14	(0.48, 0.84)	0.002
Disease risk score matching				
FO ($N_T = 788$, $N_C = 788$)	0.58	0.12	(0.45, 0.74)	< 0.001
ARIC ($N_T = 920$, $N_C = 920$)	0.63	0.12	(0.49, 0.77)	< 0.001
SAH ($N_T = 903$, $N_C = 903$)	0.66	0.12	(0.52, 0.82)	< 0.001

Abbreviations: ARIC, Atherosclerosis Risk in Communities Study; CI, confidence interval; DPP, Diabetes Prevention Program; FO, Framingham Offspring Study; HR, hazard ratio; N_C, sample size of the control group, i.e., the DPP placebo group; N_T, sample size of the treatment group, i.e., the SDPI-DP group; OGTT, oral glucose tolerance test; SAH, San Antonio Heart Study; SDPI-DP, Special Diabetes Program for Indians Diabetes Prevention Program; SE, standard error

*SDPI-DP subgroup who met the DPP eligibility criteria

gender. The fitted PS model based on the ARIC covariates has excellent predictive performance, with a *C* statistic of 0.885.

Table 3 exhibits the estimated dry-run pseudo-bias before and after DRS matching. Before matching, the pseudo-bias of the unadjusted HR was about − 0.30 (95% CI − 0.47, − 0.12), indicating the differences on observed covariates between the DPP placebo and the SDPI-DP participants would result in an estimated preventive effectiveness of − 0.30 even when the true intervention effect was 0. After matching based on DRS without OGTT 2-h test included in the models, the mean pseudo-biases ranged from − 0.14 to − 0.36. After matching based on DRS with OGTT 2-h results included, the estimated pseudo-biases were much closer to 0. The smallest mean pseudo-bias was 0.01 (95% CI − 0.25, 0.22), which was matched on the ARIC score.

Discussion

Using historical control data, this study attempted to formally evaluate the translational effectiveness of a one-arm project focusing on translating an evidence-based intervention into a community setting. Many translational or programmatic diabetes prevention programs have sought to translate the well-established evidence of lifestyle interventions into "real-world" settings. Given the existing strong evidence supporting the intervention being implemented, such programs often adopt a pre-post study design without including a concurrent control group. The evaluation of the translational effectiveness for such programs is thus, challenging. While methods have been developed for continuous outcome variables in one-arm intervention designs (Chevreul et al. 2014), approaches are needed that allow one to assess intervention effects on a time-to-event outcome, such as diabetes incidence.

The current study illustrates a potential solution to this challenge, based on publicly available data from the original RCT that generated the evidence for the intervention being translated. While our results provide initial evidence for the usefulness of such an approach, they also underscore that great care is essential. As shown in Table 2, after simply merging the DPP and all SDPI-DP data, the unadjusted HR greatly overestimated the effectiveness of the SDPI-DP lifestyle intervention. Yet, when we restricted the analysis to the SDPI-DP subgroup or adjusted for all baseline diabetes risk factors, the magnitude of the estimated risk reduction was smaller. These observations highlight the importance of considering the eligibility criteria and baseline characteristics of two studies involved when conducting program evaluations using historical control data (Baker and Lindeman 2001).

Further, we found the omission of an unbalanced confounder, the OGTT 2-h result, produced biased estimates of the intervention effectiveness no matter which statistical method was employed. As noted in the PS literature (Drake 1993), even such sophisticated methods cannot correct the bias introduced by omitting important confounders. Many DPP translational projects did not conduct OGTT due to cost and feasibility considerations. Yet, our results suggest that, for accurate estimation of diabetes prevention effect of a translational intervention, this variable may be too important to ignore. Furthermore, emerging evidence highlighted the importance of OGTT in detecting prediabetes and T2DM (NCD Risk Factor Collaboration 2015).

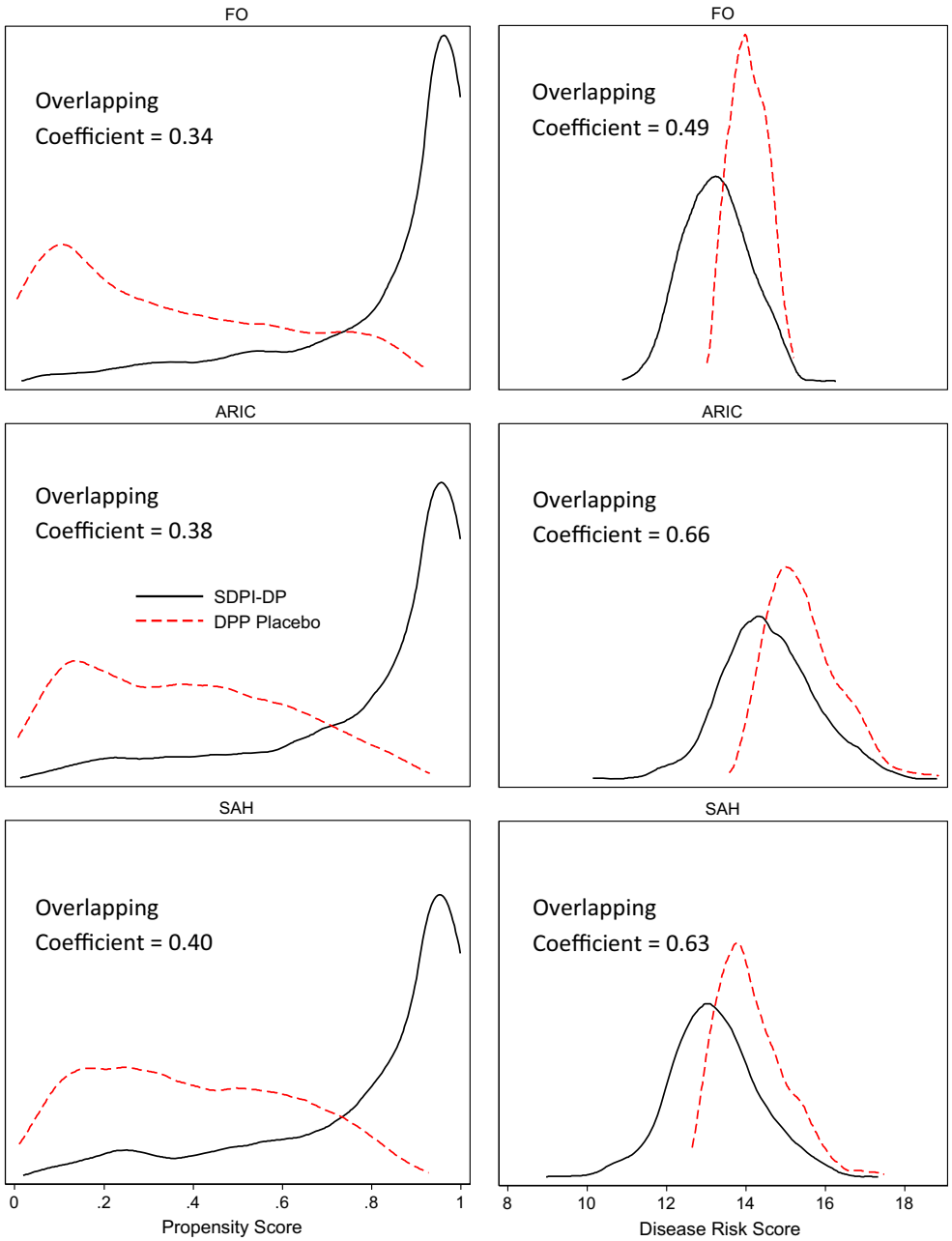

Fig. 1 Propensity score and disease risk score distributions across treatment groups. Abbreviations: ARIC, Atherosclerosis Risk in Communities Study; DPP, Diabetes Prevention Program; FO, Framingham Offspring Study; SAH, San Antonio Heart Study; SDPI-DP, Special Diabetes Program for Indians Diabetes Prevention Program

Indeed, a recent study in a high-risk population found that 47.3% of newly diagnosed patients with T2DM would have been missed if OGTTs were not performed (Meijnikman et al. 2017).

The regression adjustment method is the standard method to control for potentially unbalanced confounders, but suffers from computational complexity and model selection issues. It has been shown to produce biased estimates for regression coefficients when the number of events per covariate is less than 10 (Harrell Jr. et al. 1985; Peduzzi et al. 1996). Hence, dimension reduction methods such as the PS or DRS are preferred in the presence of a large number of potential confounders. When comparing PS matching with DRS matching in this study, we found the DRS approach resulted in more matched pairs than PS. This is consistent with a recent simulation study demonstrating that DRS can match a larger proportion of the treated population when the PS distributions across comparison groups are strongly separated (Wyss et al. 2015). Consequently, DRS matching can improve the precision and potential generalizability of the effect estimate due to larger sample size. The DRS approach has been shown to require a weaker condition than the positivity assumption of the PS approach (Hansen 2008). It only assumes no levels of disease risk at which each intervention or control is received with certainty, which means DRS matching can allow researchers to include individuals who would otherwise be

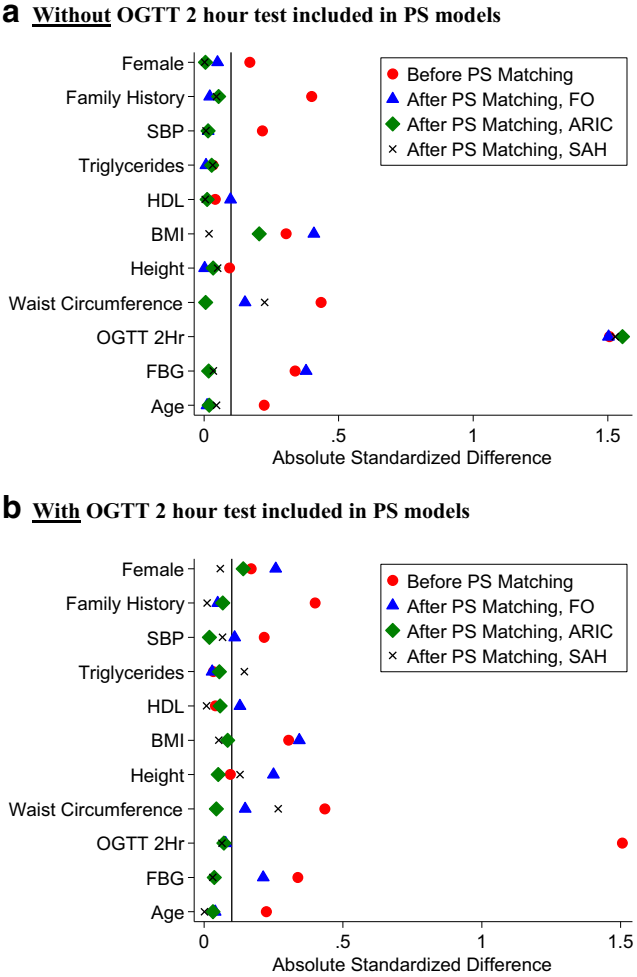

Fig. 2 Absolute standardized differences before and after propensity score matching in covariate values for all SDPI-DP vs. DPP placebo participants. **a** Without OGTT 2 h test included in PS models. **b** With OGTT 2 h test included in PS models. Abbreviations: ARIC, Atherosclerosis Risk in Communities Study; DPP, Diabetes Prevention Program; FO, Framingham Offspring Study; OGTT, oral glucose tolerance test; SAH, San Antonio Heart Study; SDPI-DP, Special Diabetes Program for Indians Diabetes Prevention Program

excluded with PS matching, especially in regression discontinuity designs such as the current study where participants with an OGTT 2-h result < 140 mg/dL were excluded from the DPP study.

Regarding the three different DRS models we explored, all demonstrated adequate capability to correct the bias in the effectiveness estimates, as long as all important confounders were considered. Although due to lack of a true control group, it is difficult to assess which method produced the least biased estimate for the effectiveness, the PS matched samples based on the ARIC model exhibited the best covariate balance with ASD < 0.1 for almost all diabetes risk factors. Similarly, among the three DRS models we compared, matching on the ARIC score (with OGTT 2 h included in the score) produced the least pseudo-bias which is very close to 0. The ARIC score included waist circumference instead of BMI in its model, which has been demonstrated to be more predictive of diabetes risk than BMI (Klein et al. 2007). This might explain the better performance of ARIC model in confounding control shown here.

Several limitations exist in this study. First, the race/ethnic compositions of the two data sources were substantially different. The SDPI-DP only recruited AI/ANs while the DPP was a multiethnic cohort with only 49 AI/ANs in the placebo group (3). Since the publicly available DPP data coded AIs in the "Other" category along with other race/ethnicity groups, adjusting for AI/AN status was impossible. However, the DPP findings showed no significant racial differences in the efficacy of lifestyle intervention, including the AI/AN subgroup (Knowler et al. 2002). Second, except for the OGTT 2-h result, we only adjusted for diabetes risk factors that were included in one of the three diabetes risk models, which may not capture all the potentially unbalanced confounders. Third, we could only find a match for < 40% of the SDPI-DP participants. This means valid inference can only be made for a proportion of the SDPI-DP participants. Yet, a previous study reported no treatment heterogeneity in lifestyle intervention

Table 3 "Dry-run" analysis evaluating disease risk scores for confounding control

	HR	95% CI	Pseudo-bias	95% CI
Unadjusted	0.35	(0.29, 0.43)	− 0.30	(− 0.47, − 0.12)
DRS matching (without OGTT 2-h results in the models)				
FO ($N_T = 857$, $N_C = 857$)	0.32	(0.24, 0.42)	− 0.36	(− 0.12, − 0.53)
ARIC ($N_T = 1011$, $N_C = 1011$)	0.33	(0.26, 0.43)	− 0.18	(− 0.39, 0.04)
SAH ($N_T = 1001$, $N_C = 1001$)	0.39	(0.31, 0.50)	− 0.14	(− 0.38, 0.09)
DRS matching (with OGTT 2-h results in the models)				
FO ($N_T = 788$, $N_C = 788$)	0.58	(0.45, 0.74)	− 0.11	(− 0.33, 0.14)
ARIC ($N_T = 920$, $N_C = 920$)	0.61	(0.49, 0.77)	0.01	(− 0.25, 0.22)
SAH ($N_T = 903$, $N_C = 903$)	0.66	(0.52, 0.82)	0.06	(− 0.17, 0.32)

Abbreviations: ARIC, Atherosclerosis Risk in Communities Study; CI, confidence interval; FO, Framingham Offspring Study; HR, hazard ratio; N_C, sample size of the control group, i.e., the DPP placebo group; N_T, sample size of the treatment group, i.e., the SDPI-DP group; OGTT, oral glucose tolerance test; SAH, San Antonio Heart Study

effects based on baseline diabetes risk of the DPP participants, suggesting potential generalizability of our results (Sussman et al. 2015).

Last, although the PS and DRS approaches appear to be useful statistical tools for evaluating intervention effectiveness in studies with observational data or quasi-experimental design, they cannot substitute RCTs in assessing the efficacy of a new intervention. The Women's Health Initiative (WHI) study reported a well-known example where the conclusions from observational studies were different from those based on RCT: although several large observational studies with sound statistical design and analyses suggested postmenopausal hormone use reduced CHD risk (Grodstein et al. 1996) (Varas-Lorenzo et al. 2000), the WHI randomized trial reported those in the hormone therapy arm had a higher incidence of CHD than the women in the placebo group (Manson et al. 2003). The discrepancies between the results of the WHI RCT and observation studies could largely be explained by the time-varying HRs of the treatment effects (Prentice et al. 2005), which cannot be detected and solved by the statistical methods used here. Furthermore, several threats to internal validity as listed by Campbell and Stanley (Campbell and Stanley 1963) might exist in our study. The potential applicability of those threats to our study is listed and discussed in Supplementary Table 2.

In summary, this study illustrates how one can use publicly available RCT data as historical controls to evaluate the intervention effectiveness of community translational projects without a concurrent control. Carefully employed, this approach shows promise in obtaining relatively accurate estimates for the translational effectiveness of projects wherein the eligibility criteria and outcome measures are similar. Indeed, future translational initiatives without a control group may consider using similar eligibility criteria and outcomes as the original clinical trial(s), at least for a proportion of the participants, in order to allow for formal evaluation of translational effectiveness using historical control data. To overcome potentially severe selection bias while using historical controls, it is critical to employ a proper statistical method to balance the distributions of potential confounders between comparison groups. Both PS matching and DRS matching are good choices when the number of confounders that needs to be adjusted is large (Cepeda et al. 2003; Harrell Jr. et al. 1985; Peduzzi et al. 1996). Further, the DRS approach may be particularly suitable in circumstances when the PS distributions of the comparison groups do not overlap well with each other.

Acknowledgments The authors would like to express our gratitude to the Indian Health Service (IHS) as well as tribal and urban Indian health programs and participants involved in the SDPI-DP. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the IHS. The DPP was conducted by the DPP Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This manuscript was not prepared in collaboration with the Investigators of the DPP study and does not necessarily reflect the opinions or views of the DPP study or the NIDDK. The authors would also like to thank Dr. James Hill for his valuable scientific suggestions and comments.

Grant programs participating in the Special Diabetes Program for Indians Diabetes Prevention Program are as follows: Confederated Tribes of the Chehalis Reservation, Cherokee Nation, Cheyenne River Sioux Tribe, the Chickasaw Nation, Coeur d'Alene Tribe, Colorado River Indian Tribes, Colville Confederated Tribes, Cow Creek Band of Umpqua Tribe, Klamath Tribes, and Coquille Tribe, Fond du Lac Reservation, Gila River Health Care, Haskell Health Center, Ho-Chunk Nation, Indian Health Board of Minneapolis, Indian Health Center of Santa Clara Valley, Native American Rehabilitation Association of the NW, Hunter Health, Kenaitze Indian Tribe IRA, Lawton IHS Service Unit, Menominee Indian Tribe of Wisconsin, Mississippi Band of Choctaw Indians, Norton Sound Health Corporation, Pine Ridge IHS Service Unit, Pueblo of San Felipe, Quinault Indian Nation, Rapid City IHS Diabetes Program, Red Lake Comprehensive Health Services, Rocky Boy Health Board, Seneca Nation of Indians, Sonoma County Indian Health Project, South East Alaska Regional Health Consortium, Southcentral Foundation, Trenton Indian Service Area, Tuba City Regional Health Care Corporation, United American Indian Involvement, Inc., United Indian Health Services, Inc., Warm Springs Health & Wellness Center, Winnebago Tribe of Nebraska, Zuni Pueblo.

Funding Funding for the SDPI-DP project was provided by the Indian Health Service (HHSI242200400049C, S.M. Manson). Manuscript preparation was supported in part by American Diabetes Association (ADA #7-12-CT-36, L. Jiang), and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (1P30DK092923, S.M. Manson, and R21DK108187, L. Jiang).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The SDPI-DP protocol was approved by the institutional review board (IRB) of the University of Colorado and the National IHS IRB. The use of the DPP data from the NIDDK Central Repositories was approved by the University of California Irvine IRB.

Informed Consent All participants provided written informed consent and Health Insurance Portability and Accountability Act authorization.

Transparency and Openness Promotion Guidelines

Availability of Data De-identified DPP data can be obtained from the NIDDK Data Repository following the data request instructions posted on the Repository's website: https://repository.niddk.nih.gov/pages/overall_instructions/. Due to confidentiality concerns and previous tribal agreements, the SDPI-DP data cannot be made publicly available. Access to the SDPI-DP data can only be requested by contacting the Division of Diabetes Treatment and Prevention of the Indian Health Service.

Availability of SAS Code The SAS code for DRS matching and dry-run analysis used in the statistical analysis section of this study is included in Appendix 1 of the supplementary materials

References

Arbogast, P. G., Kaltenbach, L., Ding, H., & Ray, W. A. (2008). Adjustment for multiple cardiovascular risk factors using a summary risk score. *Epidemiology, 19*, 30–37. https://doi.org/10.1097/EDE.0b013e31815be000.

Arbogast, P. G., & Ray, W. A. (2009). Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research, 18*, 67–80. https://doi.org/10.1177/0962280208092347.

Arbogast, P. G., & Ray, W. A. (2011). Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology, 174*, 613–620. https://doi.org/10.1093/aje/kwr143.

Baker, S. G., & Lindeman, K. S. (2001). Rethinking historical controls. *Biostatistics, 2*, 383–396. https://doi.org/10.1093/biostatistics/2.4.383.

Buntin, M. B., Jain, A. K., Mattke, S., & Lurie, N. (2009). Who gets disease management? *Journal of General Internal Medicine, 24*, 649–655. https://doi.org/10.1007/s11606-009-0950-8.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research on teaching*. Chicago: Rand McNally.

Cefalu, W. T., Buse, J. B., Tuomilehto, J., Fleming, G. A., Ferrannini, E., Gerstein, H. C., et al. (2016). Update and next steps for real-world translation of interventions for type 2 diabetes prevention: Reflections from a diabetes care editors' expert forum. *Diabetes Care, 39*, 1186–1201. https://doi.org/10.2337/dc16-0873.

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology, 158*, 280–287.

Chevreul, K., Brunn, M., Cadier, B., Nolte, E., & Durand-Zaleski, I. (2014). Evaluating structured care for diabetes: Can calibration on margins help to avoid overestimation of the benefits? An illustration from French diabetes provider networks using data from the ENTRED survey. *Diabetes Care, 37*, 1892–1899. https://doi.org/10.2337/dc13-2141.

Collaboration, N. C. D. R. F. (2015). Effects of diabetes definition on global surveillance of diabetes prevalence and diagnosis: A pooled analysis of 96 population-based studies with 331,288 participants. *The Lancet Diabetes and Endocrinology, 3*, 624–637. https://doi.org/10.1016/S2213-8587(15)00129-1.

Cuticchia, A. J., Cooley, P. C., Hall, R. D., & Qin, Y. (2006). NIDDK data repository: A central collection of clinical trial data. *BMC Medical Informatics and Decision Making, 6*, 19. https://doi.org/10.1186/1472-6947-6-19.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265–2281.

D'Agostino, R. B., Sr., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation, 117*, 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49*, 1231–1236. https://doi.org/10.2307/2532266.

Dunkley, A. J., Bodicoat, D. H., Greaves, C. J., Russell, C., Yates, T., Davies, M. J., & Khunti, K. (2014). Diabetes prevention in the real world: Effectiveness of pragmatic lifestyle interventions for the prevention of type 2 diabetes and of the impact of adherence to guideline recommendations: A systematic review and meta-analysis. *Diabetes Care, 37*, 922–933. https://doi.org/10.2337/dc13-2195.

Flamm, M., Panisch, S., Winkler, H., & Sonnichsen, A. C. (2012). Impact of a randomized control group on perceived effectiveness of a disease management programme for diabetes type 2. *European Journal of Public Health, 22*, 625–629. https://doi.org/10.1093/eurpub/ckr147.

Grodstein, F., Stampfer, M. J., Manson, J. E., Colditz, G. A., Willett, W. C., Rosner, B., . . . Hennekens, C. H. (1996). Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. The New England Journal of Medicine, 335, 453–461. https://doi.org/10.1056/NEJM199608153350701.

Guo, S. Y., & Fraser, M. W. (2009). *Propensity score analysis: Statistical methods and applications*: SAGE Publications, Inc.

Hansen, B. (2008). The prognostic analogue of the propensity score. *Biometrika, 95*, 481–488.

Harrell, F. E., Jr., Lee, K. L., Matchar, D. B., & Reichert, T. A. (1985). Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treat Rep, 69*, 1071–1077.

Henry, D., Tolan, P., Gorman-Smith, D., & Schoeny, M. (2017). Alternatives to randomized control trial designs for community-based prevention evaluation. *Prevention Science, 18*, 671–680. https://doi.org/10.1007/s11121-016-0706-8.

Inman, H. F., & Bradley, E. L. (1989). The overlapping coefficient as a measure of agreement between probability-distributions and point estimation of the overlap of 2 normal densities. *Communications in Statistics-Theory and Methods, 18*, 3851–3874.

Jiang, L., Manson, S. M., Beals, J., Henderson, W. G., Huang, H., Acton, K. J., et al. (2013). Translating the diabetes prevention program into American Indian and Alaska native communities: Results from the special diabetes program for Indians diabetes prevention demonstration project. *Diabetes Care, 36*, 2027–2034. https://doi.org/10.2337/dc12-1250.

Kahn, H. S., Cheng, Y. J., Thompson, T. J., Imperatore, G., & Gregg, E. W. (2009). Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years. *Annals of Internal Medicine, 150*, 741–751.

Klein, S., Allison, D. B., Heymsfield, S. B., Kelley, D. E., Leibel, R. L., Nonas, C., et al. (2007). Waist circumference and cardiometabolic risk: A consensus statement from shaping America's health: Association for Weight Management and Obesity Prevention; NAASO, the Obesity Society; the American Society for Nutrition; and the American Diabetes Association. *Diabetes Care, 30*, 1647–1652. https://doi.org/10.2337/dc07-9921.

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England Journal of Medicine, 346*, 393–403.

Lee, E. T., Howard, B. V., Wang, W., Welty, T. K., Galloway, J. M., Best, L. G., et al. (2006). Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: The strong heart study. *Circulation, 113*, 2897–2905. https://doi.org/10.1161/CIRCULATIONAHA.105.593178.

Lindstrom, J., & Tuomilehto, J. (2003). The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care, 26*, 725–731.

Mann, D. M., Bertoni, A. G., Shimbo, D., Carnethon, M. R., Chen, H., Jenny, N. S., & Muntner, P. (2010). Comparative validity of 3 diabetes mellitus risk prediction scoring models in a multiethnic US cohort: The multi-ethnic study of atherosclerosis. *American Journal of Epidemiology, 171*, 980–988. https://doi.org/10.1093/aje/kwq030.

Manson, J. E., Hsia, J., Johnson, K. C., Rossouw, J. E., Assaf, A. R., Lasser, N. L., et al. (2003). Estrogen plus progestin and the risk of coronary heart disease. *The New England Journal of Medicine, 349*, 523–534. https://doi.org/10.1056/NEJMoa030808.

Meijnikman, A. S., De Block, C. E. M., Dirinck, E., Verrijken, A., Mertens, I., Corthouts, B., & Van Gaal, L. F. (2017). Not performing an OGTT results in significant underdiagnosis of (pre)diabetes in a high risk adult Caucasian population. *International Journal of Obesity, 41*, 1615–1620. https://doi.org/10.1038/ijo.2017.165.

Miettinen, O. S. (1976). Stratification by a multivariate confounder score. *American Journal of Epidemiology, 104*, 609–620.

Noble, D., Mathur, R., Dent, T., Meads, C., & Greenhalgh, T. (2011). Risk models and scores for type 2 diabetes: Systematic review. *BMJ, 343*, d7163. https://doi.org/10.1136/bmj.d7163.

Norris, S. L., Zhang, X., Avenell, A., Gregg, E., Bowman, B., Schmid, C. H., & Lau, J. (2005). Long-term effectiveness of weight-loss interventions in adults with pre-diabetes: A review. *American Journal of Preventive Medicine, 28*, 126–139. https://doi.org/10.1016/j.amepre.2004.08.006.

Pan, X. R., Li, G. W., Hu, Y. H., Wang, J. X., Yang, W. Y., An, Z. X., et al. (1997). Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. *Diabetes Care, 20*, 537–544.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*, 1373–1379.

Prentice, R. L., Pettinger, M., & Anderson, G. L. (2005). Statistical issues arising in the Women's Health Initiative. Biometrics, 61, 899–911; discussion 911-841. https://doi.org/10.1111/j.0006-341X.2005.454_1.x.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*, 33–38.

Schmidt, M. I., Duncan, B. B., Bang, H., Pankow, J. S., Ballantyne, C. M., Golden, S. H., et al. (2005). Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. *Diabetes Care, 28*, 2013–2018.

Stern, M. P., Williams, K., & Haffner, S. M. (2002). Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? *Annals of Internal Medicine, 136*, 575–581.

Sturmer, T., Schneeweiss, S., Brookhart, M. A., Rothman, K. J., Avorn, J., & Glynn, R. J. (2005). Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology, 161*, 891–898. https://doi.org/10.1093/aje/kwi106.

Sussman, J. B., Kent, D. M., Nelson, J. P., & Hayward, R. A. (2015). Improving diabetes prevention with benefit based tailored treatment: Risk based reanalysis of Diabetes Prevention Program. *BMJ, 350*, h454. https://doi.org/10.1136/bmj.h454.

The Diabetes Prevention Program. (1999). The Diabetes Prevention Program. Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care, 22*, 623–634.

Tuomilehto, J., Lindstrom, J., Eriksson, J. G., Valle, T. T., Hamalainen, H., Ilanne-Parikka, P., et al. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *The New England Journal of Medicine, 344*, 1343–1350.

Varas-Lorenzo, C., Garcia-Rodriguez, L. A., Perez-Gutthann, S., & Duque-Oliart, A. (2000). Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation, 101*, 2572–2578.

Wareham, N. J. (2015). Mind the gap: Efficacy versus effectiveness of lifestyle interventions to prevent diabetes. *The Lancet Diabetes and Endocrinology, 3*, 160–161. https://doi.org/10.1016/S2213-8587(15)70015-X.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation, 97*, 1837–1847.

Wilson, P. W., Meigs, J. B., Sullivan, L., Fox, C. S., Nathan, D. M., & D'Agostino, R. B., Sr. (2007). Prediction of incident diabetes mellitus in middle-aged adults: The Framingham Offspring Study. *Archives of Internal Medicine, 167*, 1068–1074. https://doi.org/10.1001/archinte.167.10.1068.

Wyss, R., Ellis, A. R., Brookhart, M. A., Jonsson Funk, M., Girman, C. J., Simpson, R. J., Jr., & Sturmer, T. (2015). Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety, 24*, 951–961. https://doi.org/10.1002/pds.3810.

Wyss, R., Hansen, B. B., Ellis, A. R., Gagne, J. J., Desai, R. J., Glynn, R. J., & Sturmer, T. (2017). The "dry-run" analysis: A method for evaluating risk scores for confounding control. *American Journal of Epidemiology, 185*, 842–852. https://doi.org/10.1093/aje/kwx032.