STATISTICS,OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **2**, June 2014, pp 93–104.
Published online in International Academic Press (www.IAPress.org)

Accuracy of hemoglobin A1c imputation using fasting plasma glucose in diabetes research using electronic health records data †

Stanley Xu [a, *], Emily B. Schroeder[a], Susan Shetterly[a],Glenn K. Goodrich[a], Patrick J. O'Connor[b], John F. Steiner[a], Julie A. Schmittdiel[c], Jay Desai[b], Ram D Pathak[d], Romain Neugebauer[c], Melissa G. Butler[e], Lester Kirchner[f] and Marsha A. Raebel[a]

[a]*Kaiser Permanente Colorado Institute for Health Research, Denver, Colorado, USA.*
[b] *HealthPartners Institute for Education and Research, Minneapolis, Minnesota, USA.*
[c]*Division of Research, Kaiser Permanente Northern California, Oakland, California, USA.*
[d]*Marshfield Clinic Research Foundation, 1000 N Oak Ave (ML2), Marshfield, WI 54449, USA.*
[e] *Kaiser Permanente Center for Health Research, Atlanta, Georgia, USA.*
[f]*Geisinger Health System, Danville, Pennsylvania, USA.*

Received 30 January 2014; Accepted 5 April 2014
Editor: David G. Yu

Abstract In studies that use electronic health record data, imputation of important data elements such as Glycated hemoglobin (A1c) has become common. However, few studies have systematically examined the validity of various imputation strategies for missing A1c values. We derived a *complete dataset* using an incident diabetes population that has no missing values in A1c, fasting and random plasma glucose (FPG and RPG), age, and gender. We then created missing A1c values under two assumptions: missing completely at random (MCAR) and missing at random (MAR). We then imputed A1c values, compared the imputed values to the true A1c values, and used these data to assess the impact of A1c on initiation of antihyperglycemic therapy. Under MCAR, imputation of A1c based on FPG 1) estimated a continuous A1c within ± 1.88% of the true A1c 68.3% of the time; 2) estimated a categorical A1c within ± one category from the true A1c about 50% of the time. Including RPG in imputation slightly improved the precision but did not improve the accuracy. Under MAR, including gender and age in addition to FPG improved the accuracy of imputed continuous A1c but not categorical A1c. Moreover, imputation of up to 33% of missing A1c values did not change the accuracy and precision and did not alter the impact of A1c on initiation of antihyperglycemic therapy. When using A1c values as a predictor variable, a simple imputation algorithm based only on age, sex, and fasting plasma glucose gave acceptable results.

Keywords Glycated hemoglobin; Diabetes; Multiple imputation; Missing completely at random; Missing at random.

DOI: 10.19139/soic.v2i2.68

[†]This project was supported by grant number R01HS019859 from the Agency for Healthcare Research and Quality.

[*]Correspondence to: Kaiser Permanente Colorado Institute for Health Research, Denver Highlands, 10065 E. Harvard Ave., Suite 300, Denver CO 80231, Phone: (303)614-1200, Email: stan.xu@kp.org.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1. Introduction

Large electronic health records (EHR) data now enable health care researchers to conduct comparative effectiveness studies, monitor post-market safety of medical products, and carry out pharmacoepidemiological studies. However, as these data are collected in the process of delivering health care, and not collected specially for research purposes, data relating to important risk factors may be missing. One such large EHR database is SUrveillance PREvention and ManagEment of Diabetes Mellitus (SUPREME-DM), a consortium of 11 U.S. health care systems, including approximately 1.1 million diabetes patients. Within SUPREME-DM, we collect patient demographic, health care utilization, diagnosis, procedure, medication, and laboratory results data from EHR and other clinical and administrative databases [1].

In a recent study using SUPREME-DM data, the association between initial antihyperglycemic therapy and patient-level baseline characteristics was examined among 241,327 adults with newly identified diabetes between 2005 and 2010 [2]. Initial results suggested that glycated hemoglobin A1c (A1c), a measure of a patient's average blood glucose level during the past 2-3 months [3], was a key predictor of antihyperglycemic initiation. However about 33% of patients had no A1c available at the time of or within the two years preceding diabetes identification.

The importance of A1c as an objective and reliable measure of long term glucose control and its utility in diabetes diagnosis and care is well established [4, 5]. Clinical trials, such as the Diabetes Control and Complications Trial (DCCT) and the United Kingdom Prospective Diabetes Study (UKPDS) have shown that improving A1c measures decreased the development and progression of eye, kidney and nerve complications in both type 1 and type 2 diabetes [6, 7]. Whereas A1c reflects average glucose over a 2-3 month period, measures for fasting plasma glucose (FPG) or random plasma glucose (RPG) reflect glucose values at a single point in time. Because glucose values vary with eating patterns, exercise, stress, and other factors within a single day or even hour, A1c provides a better estimate of glucose control that FPG or RPG does. In the SUPREME-DM antihyperglycemic initiation study, the large proportion of missing A1c values could be problematic. Because A1c is an important risk factor, it is crucial to handle missing A1c values appropriately in diabetic research using EHR data. Simply excluding a third of the patients with missing A1c values could result in biased results. This simulation study focuses on the situation where A1c is an independent variable used for predicting clinical outcomes.

Imputation of important risk factors has become a common practice in clinical trials studies [8, 9] and observational studies using EHR data [10, 11]. However, few studies have systematically examined the validity of various imputation strategies for A1c. Using the diabetic cohort compiled for the antihyperglycemic drug initiation study, we explored the implications of using auxiliary variables (FPG and RPG) and other covariates (age and gender) for imputation of missing A1c values for use as a predictor (independent) variable under two assumptions: missing completely at random (MCAR) and missing at random (MAR) [12].

Key contributions of this papaer: 1) Imputed categorical A1c is within ± one category 50% of the time. This is not very precise when considered from a clinical standpoint, but it may still be useful in research where the alternative is to exclude those individuals with missing A1c and possibly introduce substantial bias. 2) For categorical A1c, the accuracy did not improve by including RPG, age and gender in imputation. 3) The accuracy of imputation did not vary much as a function of the proportion of A1c values that were imputed. 4) An analysis of A1c as a predictor of medication initiation in these newly diagnosed diabetes patients showed that results were not altered by up to 33% of A1c values being imputed using this simple method.

2. Methods

2.1. Study population

For this imputation study, we utilized a subset of the SUPREME-DM study population that was analyzed by Raebel et al [2] in examining associations between initial antihyperglycemic therapy and patient baseline characteristics. The study population selection methods are detailed elsewhere [2]. Briefly, nonpregnant adults were required to meet either diagnosis or laboratory criteria. Incident diabetes cases were those who first met the study diabetes criteria after at least two years of health system membership with no indication of diabetes and no antihyperglycemic dispensed during those two years. The initial study population included 241,327 subjects, but for our imputation comparisons we selected 62,458 patients who had no missing data on key variables of interest within two years prior to diabetes diagnosis: A1c, FPG, RPG, age, and gender. We designate this dataset as the *complete dataset*.

2.2. Covariates

While A1c measurement is a continuous variable, in clinical and research uses, categorical values of A1c based on known cutpoints are often used [13, 14, 15]. We assessed imputation methods for both continuous A1c and a categorical classification that categorized A1c measures into six groups: A1c$\leq$6%, 6%$<$A1c$\leq$7%, 7%$<$A1c$\leq$8%, 8%$<$A1c$\leq$9%, 9%$<$A1c$\leq$10%, and A1c$>$10%. Age was a categorical variable with six groups: $<$39 years, 40-49, 50-59, 60-69, 70-79, and 80. FPG and RPG were used as continuous variables in imputing A1c. Gender was included in the analyses as a binary covariate.

Hemoglobin A1c is commonly used in clinical practice as a measurement of chronic glycemia, represents the proportion of hemoglobin A1c molecules that are glycated, and is a function of average plasma glucose levels and red blood cell turnover [16]. It is therefore reasonable to expect that A1c, FPG, and RPG would be correlated. Several studies have looked at the correlation between A1c and average glucose, including the Diabetes Control and Complications Trial [17] and the A1c-Derived Average Glucose (ADAG) study [18]. The ADAG study derived a commonly used conversion equation between A1c values and average plasma glucose values ($AG_{mg/dl}$ = 28.7 x A1C – 46.7, R^2 = 0.84), using data from continuous glucose monitoring and seven-point daily self-monitoring capillary glucose on 507 subjects [18, 19]. The ADA and the American Association for Clinical Chemistry have determined that the correlation (Pearson correlation coefficient=0.92) is strong enough to justify reporting both an A1C result and an estimated average glucose result when a clinician orders the A1C test [19].

Due to the short-term variation in plasma glucose values, one would expect the correlation between a single FPG or RPG value and A1c to be less than the correlation between a comprehensive assessment of average glucose and A1c [17, 20]. We found that only 6.5% of individuals in our cohort did not have any baseline measure of glucose (A1c, FPG, or RPG). The Pearson correlation coefficients were 0.76 between A1c and FPG, 0.64 between A1c and RPG, and -0.30 between A1c and age. Due to the high correlation between A1c, FPG, and RPG, we could not include FPG and RPG as covariates in outcome analyses with A1c due to collinearity. Instead, we were able to use FPG and RPG as auxiliary variables. Auxiliary variables are variables that are highly correlated to the variable of interest (i.e. A1c) but cannot be included as a covariate in the outcome model. We therefore explored the possibility of using FPG and RPG as auxiliary variables for imputing missing A1c [21, 22, 23, 24].

2.3. *Generate missing A1c under two assumptions*

We used the complete dataset to create datasets with different proportions of missing A1c under two mechanisms: 1) MCAR; and 2) MAR as detailed below.

MCAR*:* If the probability of an observation being missing does not depend on observed or unobserved measurements (e.g., age, gender, FPG, and RPG) in studies with A1c as a covariate, then the missing observation is classified as MCAR [12, 25]. Different proportions of missing A1c values (10%, 20%, and 33%) were generated completely at random in the complete dataset.

MAR*:* Missing at random for a covariate corresponds to the situation where the missingness depends on other observed covariates in studies with A1c as a covariate. In this case, whether or not an A1c value is missing has nothing to do with the missing value itself but this is related to the values of observed covariates (i.e., age and gender) [25]. For example, older patients may be less likely to have A1c measured while female patients may be more likely to have A1c measured.

Missing not at random (MNAR) for a covariate is where missing values of a covariate are associated with the dependent variable (i..e., antihyperglycemic initiation). In the study by Raebel et al (2013) [2], the antihyperglycemic initiation rates do not differ by presence of baseline A1c (39.9% for those with baseline A1c and 41.2% for those without baseline A1c). Thus we did not study this scenario.

To generate A1c MAR, we used the entire population (N=241,327) to fit a logistic regression with the dependent variable being "whether A1c was missing" and age and gender as predictor variables. Coefficients (i.e., θ_k for age group k and β for gender) were obtained. The indicator for missing A1c, m, was created in the complete dataset using the following probabilistic model:

$$\text{prob}\,(m = 1|\alpha, \beta, I_k, \text{gender}) = \frac{\exp(\alpha + I_k\theta_k + \text{gender}\,\beta)}{1 + \exp(\alpha + I_k\theta_k + \text{gender}\,\beta)}$$

where I_k is an indicator variable for age groups, I_k=1 if age=k, otherwise I_k=0 (age group 3 is the reference); $\theta_k$were coefficients for age groups with θ_1= - 0.0468, θ_2= - 0.0987, θ_4= -0.1041, θ_5= - 0.0006, and θ_6= 0.3713 indicating that the older patients were more likely to have missing A1c. The coefficient for gender (gender =1 if female) is β= - 0.0766 indicating that female patients were less likely to have missing A1c. The intercept was set to -2.40, -1.50 and -0.75 to achieve 10%, 20% and 33% missing values of A1c.

2.4. *Imputing continuous and categorical A1c*

We used linear regression to impute continuous A1c values, and the logistic regression to impute categorical A1c values in SAS procedure PROC MI (SAS 9.2) [26]. For each missing A1c, we imputed five A1c to take into account the uncertainty. We then assessed the performance of different imputation strategies and their impact on the outcome analyses with A1c as a covariate. Missing A1c values under MCAR were imputed using: 1) available FPG and 2) both available FPG and RPG. Missing A1c under MAR was imputed using 1) available FPG and 2) available FPG, age and gender.

2.5. *Evaluation*

For our evaluation of A1c as an independent variable, we analyzed the outcome variable of antihyperglycemic initiation, defined as a first dispensing of any antihyperglycemic(s) during the 182 days after cohort inclusion. Imputation performance was evaluated by 1) calculating the absolute difference between the imputed and true value and obtaining the distributions of the differences for both continuous and categorical A1c; 2) analyzing the outcome (antihyperglycemic initiation) using data with imputed categorical A1c only and comparing it to an analysis using the complete dataset. Preliminary results showed that 100 simulations were sufficient to obtain stable estimates of the effects of A1c on antihyperglycemic initiation and confidence intervals due to large sample size (the standard deviation of 100 estimates is less than 1.1% of the average of these estimates). Estimates and confidence intervals were obtained and averaged across 100 simulations. In addition, we examined the performance of different imputation strategies with only 10% of the original population. For this smaller population,1000 simulations were carried out.

3. Results

3.1. Imputation of continuous A1c under MCAR and MAR

Under the assumption of MCAR (Table 1), imputing A1c with either FPG alone or with FPG and RPG had good accuracy (unbiased A1c values) and similar precision (standard deviations). Imputation of continuous hemoglobin A1c based only on FPG can estimate an continuous A1c within about ± 1.8% (one standard deviation) of the true hemoglobin A1c value about 68.3% of the time. That is, if true A1c was 6%, the imputed estimate was between 4.2% to 7.8% about 68.3% of the time. Under the assumption of MAR (Table 2), imputing A1c with FPG alone resulted in higher values of A1c than the true values on average (by as much as 0.137%), but imputing A1c with FPG and age and gender improved accuracy of imputed A1c values (unbiased A1c values) but with similar precision (standard deviations). Similar results were observed for the smaller population (N=6,245).

Table 1. Mean and standard deviation of differences between imputed A1c and true A1c when A1c is MCAR.

		entire population (N=62,458)$^{+}$		10% of population (N=6,245)$^{++}$	
Missing %	Imputation	Mean	standard deviation	mean	standard deviation
10	FPG	<0.001	1.883	0.005	1.886
	FPG and RPG	0.005	1.703	0.019	1.705
20	FPG	0.002	1.880	0.017	1.877
	FPG and RPG	0.004	1.701	-0.003	1.691
33	FPG	0.003	1.882	0.014	1.880
	FPG and RPG	0.006	1.701	0.011	1.692

MCAR, missing completely at random; FPG, fasting plasma glucose; RPG, random plasma glucose. $^{+}$ 100 replicates; $^{++}$ 1000 replicates.

Table 2. Mean and standard deviation of differences between imputed A1c and true A1c when A1c is MAR.

		entire population (N=62,458)$^+$		10% of population (N=6,245)$^{++}$	
Missing %	Imputation	mean	standard deviation	mean	standard deviation
10	FPG	0.129	1.813	0.151	1.818
	FPG and age and gender	-0.002	1.775	0.025	1.779
20	FPG	0.103	1.844	0.112	1.843
	FPG and age and gender	-0.0001	1.810	0.014	1.807
33	FPG	0.077	1.869	0.087	1.871
	FPG and age and gender	0.0003	1.836	0.013	1.837

MAR, missing at random; FPG, fasting plasma glucose; RPG, random plasma glucose;$^+$ 100 replicates; $^{++}$ 1000 replicates.

3.2. Imputation of categorical A1c under MCAR and effects of A1c on initiation of antihyperglycemic therapy

Table 3 shows that when categorical A1c was MCAR, imputation with FPG alone yielded 19% of imputed A1c categories the same as the observed, and 31% of imputed A1c values only one category from the true categorical A1c. Inclusion of RPG in imputation only improved the imputation slightly: percentage of exact imputation increased 0.5% and the percentage of one category from the true increased less than 1.0%.

Table 3. Distribution of difference between imputed categorical hemoglobin A1c and true based on 100 replicates when hemoglobin A1c is MCAR using the entire original population (N=62,458).

		Percent of imputed A1c categories from their true values					
Missing %	Imputation	0	± 1	± 2	± 3	± 4	± 5
10	FPG	18.5	30.6	20.7	18.1	9.3	2.7
	FPG and RPG	18.9	31.3	20.8	18.5	8.4	2.1
20	FPG	18.5	30.6	20.6	18.1	9.4	2.8
	FPG and RPG	19.0	31.1	20.8	18.5	8.5	2.1
33	FPG	18.5	30.6	20.6	18.2	9.4	2.8
	FPG and RPG	18.9	31.2	20.8	18.6	8.4	2.1

MCAR, missing completely at random; FPG, fasting plasma glucose; RPG, random plasma glucose.

Table 4 shows that when A1c was missing completely at random, there were slight differences in the relative risks of A1c categories on risk of antihyperglycemic initiation between using FPG alone vs. FPG and RPG for the A1c imputations. Compared to the relative risks (RRs) without missing A1c (3^{rd} row in Table 4), the effects of those lower A1c categories (e.g., A1c $\leq$ 6% and 6%< A1c≤7%) increased and the effects of those higher A1c categories decreased. The difference increased when the percentage of missing A1c increased. For examples, the true RR for the group of A1c ≤6% is 0.239 comparing to the group of 7%<A1c ≤8%, the RR increased to 0.261 when 10% of A1c values were missing, the RR increased to 0.293 when 20% of A1c values were missing, the RR increased to 0.337 when 33% of A1c values were missing. For the group of 8%< A1c ≤9%, the true RR is 1.492 comparing to the group of 7%<A1c ≤8%, the RR decreased to 1.404 when 10% of A1c were missing, the RR decreased to 1.331 when 20% of A1c were missing, to 1.223 when 33% of A1c were missing. Similar results were observed for the smaller population.

Table 4. Comparison of relative risks (95% confidence intervals) of hemoglobin A1c categories with different percentages of missing hemoglobin A1c based on 100 replicates when hemoglobin A1c is MCAR using the entire original population (N=62,458).

		Relative risks and 95% confidence intervals					
Missing %	Imputation	A1c≤6%	6%<A1c≤7%	Ref	8%<A1c≤9%	9%<A1c≤10%	A1c>10%
0	none	0.239 (0.226 0.252)	0.373 (0.359 0.387)	1	1.492 (1.423 1.565)	1.647 (1.563 1.735)	1.772 (1.704 1.843)
10	FPG	0.261 (0.247 0.277)	0.397 (0.382 0.413)	1	1.404 (1.336 1.476)	1.554 (1.471 1.642)	1.682 (1.613 1.753)
	FPG and RPG	0.256 (0.242 0.271)	0.393 (0.378 0.409)	1	1.395 (1.327 1.467)	1.545 (1.463 1.632)	1.677 (1.609 1.747)
20	FPG	0.293 (0.277 0.310)	0.423 (0.406 0.440)	1	1.331 (1.262 1.402)	1.469 (1.384 1.558)	1.599 (1.531 1.670)
	FPG and RPG	0.281 (0.266 0.298)	0.416 (0.399 0.433)	1	1.315 (1.248 1.385)	1.452 (1.370 1.539)	1.592 (1.525 1.662)
33	FPG	0.337 (0.317 0.358)	0.459 (0.440 0.478)	1	1.223 (1.155 1.296)	1.346 (1.265 1.433)	1.473 (1.406 1.544)
	FPG and RPG	0.315 (0.297 0.334)	0.445 (0.428 0.464)	1	1.201 (1.136 1.269)	1.323 (1.242 1.408)	1.456 (1.400 1.534)

MCAR, missing completely at random; FPG, fasting plasma glucose; RPG, random plasma glucose; Ref: reference group, 7%<A1c≤8%.

3.3. *Imputation of categorical A1c under MAR and effects of A1c on initiation of antihyperglycemic therapy*

Table 5 shows that when categorical A1c was MAR, imputation with FPG alone yielded 19% of A1c categories the same as the actual category, and 31% of A1c one category away from the true A1c category. Inclusion of age and gender in imputation only improved the imputation slightly, since the percentage of exact imputation increased about 1% and percentage of one category from the true increased about 0.4%.

Table 6 shows that when categorical A1c was MAR, there were no differences in the estimates of the effects of antihyperglycemic medication initiation between using FPG alone vs. FPG, age, and gender for A1c imputation. Compared to the RRs without missing A1c, in general, the effects of those lower A1c categories (e.g., A1c≤ 6% and 6%< A1c ≤ 7%) increased and the effects of those higher A1c categories decreased. The difference increased when the percentage of missing A1c increased. For examples, the true RR for the group of A1c ≤6% is 0.239 comparing to the group of 7%<A1c ≤8%, the RR increased to 0.262 when 10% of A1c values were missing, the RR increased to 0.293 when 20% of A1c values were missing, and the RR increased to 0.334 1when 33% of A1c values were missing. For the group of 8%< A1c ≤9%, the true RR is 1.492 comparing to the group of 7%<A1c ≤8%, the RR decreased to 1.402 when 10% of A1c were missing, the RR decreased to 1.327 when 20% of A1c were missing, to 1.232 when 33% of A1c were missing. Similar results were observed for the smaller population.

Table 5. Distribution of difference between imputed categorical A1c and true based on 100 replicates when A1c is MAR using the entire original population (N=62,458).

		Percent of imputed A1c categories from their true values					
Missing %	Imputation	0	± 1	± 2	± 3	± 4	± 5
10	FPG	18.6	30.7	20.6	18.1	9.3	2.8
	FPG and age and gender	19.5	30.9	20.4	17.7	9.0	2.4
20	FPG	18.5	30.7	20.6	18.1	9.3	2.8
	FPG and age and gender	19.5	31.0	20.4	17.8	9.0	2.4
33	FPG	18.5	30.6	20.7	18.1	9.3	2.8
	FPG and age and gender	19.4	31.0	20.4	17.8	9.0	2.4

MAR, missing at random; FPG, fasting plasma glucose; RPG, random plasma glucose.

Table 6. Comparison of relative risks (95% confidence intervals) of A1c categories with different percentages of missing A1c based on 100 replicates when A1c is MAR using the entire original population (N=62,458).

		Relative risks and 95% confidence intervals					
Missing %	Imputation	A1c≤6%	6%<A1c≤7%	Ref	8%<A1c≤9%	9%<A1c≤10%	A1c>10%
0	none	0.239 (0.226 0.252)	0.373 (0.359 0.387)	1	1.492 (1.423 1.565)	1.647 (1.563 1.735)	1.772 (1.704 1.843)
10	FPG	0.262 (0.248 0.277)	0.396 (0.380 0.411)	1	1.402 (1.333 1.473)	1.549 (1.465 1.637)	1.682 (1.614 1.752)
	FPG and age and gender	0.257 (0.243 0.271)	0.392 (0.377 0.408)	1	1.403 (1.335 1.474)	1.554 (1.471 1.641)	1.684 (1.616 1.754)
20	FPG	0.293 (0.276 0.311)	0.421 (0.404 0.438)	1	1.327 (1.259 1.398)	1.464 (1.397 1.553)	1.600 (1.532 1.672)
	FPG and age and gender	0.282 (0.266 0.299)	0.415 (0.398 0.432)	1	1.328 (1.261 1.400)	1.469 (1.387 1.556)	1.605 (1.537 1.676)
33	FPG	0.334 (0.315 0.355)	0.455 (0.436 0.474)	1	1.232 (1.164 1.303)	1.352 (1.270 1.441)	1.489 (1.422 1.559)
	FPG and age and gender	0.315 (0.297 0.346)	0.446 (0.427 0.465)	1	1.231 (1.165 1.301)	1.358 (1.276 1.444)	1.493 (1.425 1.563)

MAR, missing at random; FPG, fasting plasma glucose; RPG, random plasma glucose; Ref: reference group, 7%<A1c≤8%.

4. Discussions

In clinical databases, A1c information may be missing for several reasons among individuals with diabetes: not having an A1c measured within a certain time window, having an A1c measured but not available electronically, or not having healthcare encounters over a specified timeframe. Results from this study suggest that under the MCAR assumption imputation of A1c based only on FPG 1) estimates an continuous A1c within about ± 1.8% of the actual value 68.3% of the time; 2) estimates a categorical A1c within ± one category 50% of the time. This is not very precise when considered from a clinical standpoint, but it may still be useful in research where the alternative is to exclude those individuals with missing A1c and possibly introduce substantial bias.

For continuous A1c, the accuracy did not improve by including RPG in imputation under MCAR assumption. However, including gender and age in imputation under MAR assumption improved the accuracy of imputed continuous A1c. For categorical A1c, the accuracy did not improve by including RPG, age and gender in imputation. Adding RPG values in addition to FPG values did not significantly improve the validity of the imputation, perhaps because RPG are more widely dispersed depending on hours since eating and other factors, relative to FPG. In addition, the accuracy of imputation did not vary much as a function of the proportion of A1c values that were imputed. Similar results were observed when the size of population decreased nearly ten-fold.

Moreover, it is encouraging that an analysis of A1c as a predictor of medication initiation in these newly diagnosed diabetes patients showed that results were not altered by up to 33% of A1c values being imputed using this simple method. This finding supports the use of the simple A1c imputation model we evaluated in analyses for applications where A1c is used to predict a different dependent variable, such as a clinical action by providers. These data neither support nor refute the use of this simple imputation model for A1c when it is the dependent variable in an analysis. However, in such a scenario, more sophisticated imputation models for A1c might be considered.

Limitations to our work include that the validity of multivariable imputation of A1c can be improved by including additional variables we omitted by design (e.g., number of medications, prior A1c values), our analysis included only incident cases of diabetes mellitus, and the A1c we imputed was the first A1c value at the time of or within the two years preceding diabetes identification. The distribution of A1c values at new diagnosis of diabetes is bimodal, which increases the difficulty of accurate imputation [27].

Several other factors constrain the interpretation of these data. First, while our study used a large sample of adults with incident diabetes receiving care at one of 11 medical groups in the U.S., generalizability of our findings to other populations, especially to uninsured patients, is not assured. Second, the precision of A1c imputation may be greater in situations where antecedent A1c values, medications, BMI, and race/ethnicity are also included in the imputation process. Additional work to assess the impact of these variables on precision of A1c imputation is needed.

In light of the design of the study and the constraints of the data, we conclude that imputation of A1c based on FPG, age, and gender is reasonably accurate for analyses in which A1c is being used as an independent variable to predict outcomes such as medication initiation or intensification. However, the lack of precision of imputed A1c values augers poorly for the use of these basic imputation methods when A1c is a dependent variable. In such a scenario, more sophisticated methods of multivariate imputation may be needed, and when these include antecedent A1c values, the likelihood of precise A1c imputation may increase.

REFERENCES

1. Nichols G A, Desai J, Lawrence J M, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. *Preventing chronic disease*, 2011, 9: E110-E110.
2. Raebel RA, Xu S, Goodrich GK, et al. Predictors of Initial Antihyperglycemic Therapy among Adults with Newly Identified Diabetes in the SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM) Cohort. *The Annals of Pharmacotherapy*, 2013, 47(10):1280-1291.
3. Gonen B, Rachman H, Rubenstein AH, Tanega SP, Horwitz DL. Hemoglobin A1c as an indicator of the degree of glucose intolerance in diabetics. *Lancet* 1977, 2:734 -737.
4. Nathan DM, Singer DE, Hurxthal K, Goodson JD. The clinical information value of the glycosylated hemoglobin assay. *N Engl J Med* 1984, 310:341-346.
5. Singer DE, Coley CM, Samet JH, Nathan DM. Tests of glycemia in diabetes mellitus: their use in establishing a diagnosis and treatment. *Ann Intern Med* 1989, 110 :125-137.
6. 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 (14): 977-986.
7. Diabetes Trials Unit. Oxford University. United Kingdom Prospective Diabetes Study, 1999.
8. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* 2006, 59(10):1102-1109.
9. Janssen KJ, Donders AR, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, Moons KG. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol* 2010, 63(7):721-727.
10. Masica AL, Ewen E, Daoud YA, Cheng D, Franceschini N, Kudyakov RE, Bowen JR, Brouwer ES, Wallace D, Fleming NS and West SL. Comparative effectiveness research using electronic health records: impacts of oral antidiabetic drugs on the development of chronic kidney disease. *Pharmacoepidemiology and Drug Safety* 2013 (in press).
11. Hung AM, Roumie CL, Greevy RA, Liu X, Grijalva CG,Murff HJ, and Griffin MR. Kidney function decline in metformin versus sulfonylurea initiators:assessment of time-dependent contribution of weight, blood pressure, and glycemic control. *Pharmacoepidemiology and Drug Safety* 2013 (in press).
12. Rubin DB. Inference and missing data (with discussion). *Biometrika* 1976, 63:581-592.

13. The International Expert Committee. International Expert Committee Report on the Role of the A1c Assay in the Diagnosis of Diabetes. *Diabetes Care* 2009, 32:1327-1334.
14. Lu ZX, Walkerm KZ, O'Dea K, Sikaris KA, and Shaw JE. A1C for Screening and Diagnosis of Type 2 Diabetes in Routine Clinical Practice. *Diabetes Care* 2010, 33: 817-819.
15. Choi SH, Kim TH, Lim S, Park KS, Jang HC, and Cho NH. Hemoglobin A1c as a Diagnostic Tool for Diabetes Screening and New-Onset Diabetes Prediction: A 6-year community-based prospective study. *Diabetes Care* 2011, 34(4): 944-949.
16. Kahn R, Fonseca V. Translating the A1C Assay. *Diabetes Care* 2008, 31:1-4.
17. Rohlfing CL, Wiedmeyer HM, Little RR, England JD, Tennill A, Goldstein DE. Defining the relationship between plasma glucose and HbA1c: analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial. *Diabetes Care* 2002, 25:275-278.
18. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. A1c-Derived Average Glucose Study Group. Translating the A1C assay into estimated average glucose values. *Diabetes Care* 2008, 31(8):1473-1478.
19. American Diabetes Association. Standards of Medical Care in Diabetes-2013. *Diabetes Care* 2013, Supplement 1.
20. Koenig R, Peterson CM, Kilo C, Cerami A, and Williamson JR. Hemoglobin A1c as an indicator of degree of glucose intolerance in diabetes. *Diabetes* 1976, 25: 230-232.
21. Little RJ. Regression with missing X's: a review. *J Am Stat Assoc* 1992, 87:1227-1237.
22. Little RJ, Rubin DB. Statistical analysis with missing data. New York: Wiley; 2002.
23. Enders CK. Analyzing structural equation models with missing data. In G. R.Hancock and R. O. Mueller (Eds.), Structural Equation Modeling: A Second Course (pp.313-342). 2006. Greenwich, CT: Information Age Publishing.
24. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010, 29:2920-2931.
25. Schafer JL. Analysis of Incomplete Multivariate Data. 1997. Chapman and Hall.
26. Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1987. New York: John Wiley and Sons.
27. O'Connor PJ, Gregg E, Rush WA, Cherney LM, Stiffman MN, Engelgau MM. Diabetes: how are we diagnosing and initially managing it? *Ann Fam Med* 2006, 4(1):15-22.