Journal of Clinical Epidemiology 84 (2017) 130–141

Journal of Clinical Epidemiology

Multiple imputation was a valid approach to estimate absolute risk from a prediction model based on case–cohort data

Kristin Mühlenbruch[a,b], Olga Kuxhaus[a,b], Romina di Giuseppe[c], Heiner Boeing[a], Cornelia Weikert[c,d], Matthias B. Schulze[a,b,*]

[a]*Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany*
[b]*German Center for Diabetes Research (DZD), Ingolstädter Landstr. 1, Neuherberg 85764, Germany*
[c]*Research Group Cardiovascular Epidemiology, German Institute of Human Nutrition Potsdam–Rehbruecke, Postfach 65 21 33, Berlin 13316, Germany*
[d]*Department of Food Safety, Federal Institute of Risk Assessment, Max-Dohrn-Str. 8-10, Berlin 10589, Germany*

Accepted 16 December 2016; Published online 28 January 2017

Abstract

Objective: To compare weighting methods for Cox regression and multiple imputation (MI) in a case–cohort study in the context of risk prediction modeling.

Study Design and Setting: Based on the European Prospective Investigation into Cancer and Nutrition Potsdam study, we estimated risk scores to predict incident type-2 diabetes using full cohort data and case–cohort data assuming missing information on waist circumference outside the case–cohort (~90%). Varying weighting approaches and MI were compared with regard to the calculation of relative risks, absolute risks, and predictive abilities including *C*-index, the net reclassification improvement, and calibration.

Results: The full cohort comprised 21,845 participants, and the case–cohort comprised 2,703 participants. Relative risks were similar across all methods and compatible with full cohort estimates. Absolute risk estimates showed stronger disagreement mainly for Prentice and Self & Prentice weighting. Barlow and Langholz & Jiao weighting methods and MI were in good agreement with full cohort analysis. Predictive abilities were closest to full cohort estimates for MI or for Barlow and Langholz & Jiao weighting.

Conclusions: MI seems to be a valid method for deriving or extending a risk prediction model from case–cohort data and might be superior for absolute risk calculation when compared to weighted approaches.

Keywords: Case–cohort studies; Multiple imputation; Risk prediction; Model extension; Diabetes mellitus; Type 2

1. Introduction

The case–cohort study design, initially proposed by Prentice [1], has been successfully applied in epidemiologic research, particularly in the context of expensive or unfeasible exposure measurements in the full cohort, such as genotyping or biochemical measurements [2,3]. Different weighting methods for Cox regression analyses were proposed to account for this design; the most commonly used methods being Prentice, Barlow, and Self & Prentice weighting [4,5]. However, there remains uncertainty how case–cohort designs can be used in the context of the development or extension of risk prediction models, for which valid estimation of absolute risk is an essential prerequisite. An alternative approach to weighted Cox regression of analyzing case–cohort data is multiple imputation (MI) [6,7], assuming missing at random mechanism for missing data of full cohort participants not selected for the case–cohort. This approach was proposed in 2011 along with its application in the context of risk prediction modeling [8]. Subsequently, other studies further developed this approach to determine measures of risk prediction or absolute risk estimation [9]. However, they did not provide a score equation for absolute risk estimation from a

Conflict of interest: None.

Funding: This work was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). The recruitment phase of the EPIC-Potsdam study was supported by the Federal Ministry of Science, Germany (01 EA 9401) and the European Union (SOC 95201408 05F02). The follow-up of the EPIC-Potsdam study was supported by German Cancer Aid (70-2488-Ha I) and the European Community (SOC 98200769 05F02).

* Corresponding author. Tel.: 0049-33200-882434; fax: 0049-33200-882437.

E-mail address: mschulze@dife.de (M.B. Schulze).

http://dx.doi.org/10.1016/j.jclinepi.2016.12.019

What is new?

Key findings

- Estimates from multiple imputation including relative and absolute risks, *C*-index, calibration, and net reclassification improvement were similar or superior to weighted case–cohort estimates when compared to full cohort estimates.

What this adds to what was known?

- Previous studies did not compare different weighting approaches with multiple imputation, especially when applied to 90% missing data; in contrast to previous studies with simulated data, this study is based on empirical data and measures of predictive ability, model improvement, and absolute risks were computed.

What is the implication and what should change now?

- As proof-of-concept, multiple imputation should be the preferred method for deriving or extending risk prediction models based on data from this empirical example of a case–cohort study, especially for providing an equation to calculate absolute risks.

case–cohort prognostic modeling, in terms of survival analysis; only to generally missing predictor values, MI was applied but not to missing predictor values due to a case–cohort study design [10–12]. Previous studies rather mainly focused on relative risk estimation or model fit performance measures, also including discrimination or reclassification [13–15]. However, calibration or the method how absolute risk was estimated was not sufficiently addressed for case–cohort settings. So far, studies did not show how the results can be used to provide a scoring rule and to calculate absolute risks. With regard to the amount of missing values, MI has successfully been applied to rare diseases [10] and to about 50% missing values in prognostic research; however, a maximum for deriving valid and unbiased results with MI was not determined so far [11,12].

Thus, although the application of MI to case–cohort data appears promising in the context of risk prediction modeling based on case–cohort data, a comparison of different methods with commonly large fraction of missing information is lacking. The aim of this study was therefore to systematically investigate different Cox regression weighting methods for risk prediction modeling and MI as a proof-of-concept applied to real data of the case–cohort study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study which is constructed with ~10% of the original cohort. Relative risk estimation, absolute risk estimation, discrimination, calibration, and reclassification will be compared across methods for a scenario with an assumed missingness of ~90% of a continuous predictor.

2. Methods

2.1. Study population

Data were used from the Potsdam part of the European Prospective Investigation into Cancer and Nutrition (EPIC-Potsdam) study, a prospective cohort study comprising participants from the general adult population from Potsdam and surrounding municipalities. Recruitment was performed from 1994 to 1998 achieving an overall study sample of 27,548 men and women mainly aged 35–65 years [16]. Baseline assessment included a personal interview and a lifestyle questionnaire; the diet was assessed with a validated semiquantitative food–frequency questionnaire taking into account the usual diet in the period of the last 12 months. In physical examinations in the study center, body height, body weight, waist circumference, or blood pressure were measured with a standardized protocol. A total of 26,444 participants agreed on blood drawing. Follow-up assessment was performed every 2–3 years [17] with response rates of 96%, 95%, 91%, and 90% in follow-up rounds 1, 2, 3, and 4 (by August 2005). Systematic information sources for incident cases were self-reports of a T2D diagnosis, T2D-relevant medication, and dietary treatment due to T2D during follow-up. Furthermore, we obtained additional information from death certificates or from random sources, such as the tumor centers, physicians, or clinics that provided assessments from other diagnoses. Although self-reports of T2D were generally reliable [18], by including other sources of information, we even improved the completeness of case ascertainment. Once a participant was identified as a potential case, disease status was further verified by sending a standard inquiry form to the treating physician. Only physician-verified cases with a diagnosis of T2D (International Classification of Diseases, 10th revision code: E11) and a diagnosis date after the baseline examination were considered confirmed incident cases of T2D. Family history of diabetes was assessed in the fifth follow-up round [19].

For the analysis, participants with prevalent diabetes (n = 1,554), nonverified incident diabetes status (n = 13), missing follow-up (n = 589) or baseline covariate information (n = 225), implausible occupation status (n = 1), or missing information of family history of diabetes (collected during follow-up, n = 3,321) were excluded. The remaining analysis sample included 21,845 participants.

To compare full cohort with case–cohort analyses, we assumed a scenario in which waist circumference is

available only in a case–cohort and not for participants outside the case–cohort. To do so, we relied on an existing case–cohort nested within the full cohort of EPIC-Potsdam, for the measurement of biomarkers [20]. Based on the 26,444 participants who gave blood, a subcohort of 2,500 participants (about 10%) was randomly drawn, by that being representative for the full cohort [21], and all incident diabetes cases were included (n = 820). After the same exclusions as described previously (n = 529) as well as case–cohort–related exclusions of biomarker availability (n = 14), 2,078 participants in the subcohort and 693 incident diabetes cases remained for analysis; 68 cases within the subcohort (internal) and 625 cases outside the subcohort (external). All variables except for waist circumference had no missing values left.

2.2. Statistical analysis

As proof-of-concept, we compared different weighting methods and MI for the case–cohort analysis and compared results with the full cohort analysis.

The risk prediction model which was used in this analysis was the German Diabetes Risk Score (GDRS), a noninvasive diabetes risk prediction model based on data from the EPIC-Potsdam study [19,22]. The 5-year risk of developing diabetes can be calculated using information on the following risk factors: age (years), body height (cm), waist circumference (cm), prevalent hypertension (yes/no), physical activity (sports, biking, and gardening in hours/week), smoking behavior (former smoking with $<$20 cig./day, former smoking with $\geq$20 cig./day, current smoking with $<$20 cig./day, current smoking with $\geq$20 cig./day), whole grain intake (bread, rolls, or muesli in 50 g portions/day), intake of red meat (150 g portions/day), coffee consumption (150 mL portions/day), and family history of diabetes (one parent with diabetes, both parents with diabetes, a sibling with diabetes). As we have previously described for the GDRS based on a full cohort analysis, hazard ratios (HRs) and corresponding 95% confidence intervals (CIs), beta coefficients and allocated score points, mean score points, and baseline survival were estimated from Cox regression [19,23] with predictors being risk factors at baseline. The underlying time scale for the Cox model was time until diabetes diagnosis for incident cases and time under study (from baseline to last follow-up) for noncases. For full cohort analysis, individual risks were calculated using the following equation [23]:

$$P(\text{Diabetes}) = 1 - 0,99061^{\exp\left(\frac{\text{Scorepoints} - 474.17096591}{100}\right)}$$

Proportional hazards assumption was tested for all predictors in the GDRS by correlation analysis of Schoenfeld residuals with different time variables and graphical illustration. Additional interaction terms were tested in the model if necessary.

2.2.1. Differently weighted Cox proportional hazards regression for case–cohort analysis

For the case–cohort analysis, different weighting approaches were applied to Cox regression to adjust for the sampling fraction and overrepresentation of incident cases: Barlow weighting [5], Prentice weighting [1], Self & Prentice weighting [4], and the approach proposed by Langholz & Jiao [9]. The main differences between these methods are the contributing time before and at failure of cases and the weights given to cases inside and outside the subcohort. In contrast to the other approaches, Langholz & Jiao assume the same weights for all cases before or at failure and all subcohort members; weighting with the inverse of the log-transformed sampling probability is then applied to the Breslow estimator for pseudolikelihood estimation and computation of the cumulative hazard function. Based on these approaches, HRs and 95% CIs, baseline survival function, and mean score points were estimated to be able to calculate absolute risks; the equation presented previously for the full cohort risk score was adapted in terms of baseline survival and mean score points. Proportional hazards assumption was tested for all weighting methods according to Schoenfeld residuals tests and graphical evaluation as proposed by Xue et al. for case–cohort studies [24]; if unclear, interaction with time was additionally tested in the model.

2.2.2. Multiple imputation

The multiple imputation procedure was applied to the same data set as for the full cohort analysis (21,845 participants), imputing waist circumference for all participants outside the case–cohort described previously. For the imputation process, varying numbers of imputations (m = 10, 20, 30, 40) and two different imputation models were used. The minimum or ‘‘simple’’ imputation model included only the covariates of the established diabetes prediction model (GDRS risk factors) plus follow-up time and the outcome variable, whereas a more complex imputation model included additionally variables univariate correlated ($r > 0.20$) with waist circumference (body mass index, sex, beer consumption, meat, and processed meat intake) from a predefined set of potentially contributing variables (diet, anthropometry, education, or occupation). To evaluate the performance of the MI process, we calculated the fraction of missingness and relative efficiency [25] for waist circumference directly after MI was applied and, again, for all risk factors in the analysis model after analyzing the imputed data and combining them using Rubin’s rules [26]. For the combination of estimates from several imputed data sets, the application of Rubin’s rules means to calculate the average of the estimates and a variance estimate based on within-imputation variance and between-imputation variance to account for uncertainty of the imputation. To determine the individual risks for the GDRS based on MI, we summarized the MI estimates from Cox regression as the average of the baseline survival after

complementary log-log transformation over all imputations as proposed by Marshall et al. [13] and as the average of the mean score points from each imputed data set and applied these values in the presented equation. Based on these parameters, the individual risks were calculated in each imputed data set using the respective regression coefficients from the imputation. Based on these individual risks, performance of the GDRS was first determined by imputation and finally summarized as the average (Rubin's rules).

2.2.3. Calculation of performance measures

Performance was evaluated in terms of discrimination and calibration. The discrimination was evaluated in terms of the *C*-index or Harrell's C for survival estimates as proposed by Pencina et al. [27], and calibration was illustrated graphically in a calibration plot as well as calibration-in-large by comparing mean predicted risk in the full cohort with the observed event rate [28,29]. For model improvement, net reclassification improvement (NRI) for four risk categories ($<2\%$, $2-5\%$, $5-10\%$, $\geq 10\%$ risk), and *C*-indices were evaluated for a basic model without waist circumference ($C\text{-index}_1$) and an extended model with waist circumference ($C\text{-index}_2$) reflecting the GDRS. The NRI is calculated by classification into predefined risk categories using the basic model and the extended model. Proportions of upward and downward movements according to case status are summarized as the NRI [30].

The performance in the full cohort analysis was described with no need of adaptation to design; for calibration, also the *P*-value of the Hosmer−Lemeshow test (HL test) [31] was determined, and for calibration-in-the-large, observed event rate and mean predicted risk were presented.

For MI, *C*-index and NRI were determined in each imputed data set separately and combined as the mean (Rubin's rules); the calibration plot includes estimates from all imputations overlapping.

For the case−cohort analyses, we evaluated the performance based on case−cohort individual risk estimates as described before, and additionally, a weighted version of the *C*-index was determined taking into account the sampling fraction for the case−cohort study as proposed by Ganna et al. [14]. Also, a weighted version for the NRI was proposed, however, for unstratified case−cohort study data, this is computed similarly as for full cohort studies. Instead, we excluded cases outside the subcohort who become cases after 5 years and thereby counting for noncases as proposed by Sanderson et al. [32]; by this, we account for the artificially increased risk among noncases and a more valid standard error can be calculated. For all approaches, NRI was presented as the overall value, the single components and corresponding 95% CIs as proposed by Mühlenbruch et al. [33]. To account for the overrepresentation of the cases due to design and the resulting high observed risks, evaluation of calibration by use of a calibration plot was restricted to the subcohort only.

2.2.4. Agreement with full cohort estimates

To compare the individual risks derived from all different approaches with full cohort risk estimates, we calculated the median error by subtracting the full cohort risk in each individual and averaging it over the whole study sample. A negative value indicated underestimation on average and a positive value an overestimation on average. For further evaluating the agreement with the full cohort estimates, we compared risk distributions and illustrated the agreement via plotting the respective estimates against the full cohort estimates; a diagonal line indicated perfect agreement, values below the line overestimation, and values above underestimation.

All statistical analyses were performed with SAS (version 9.4, Enterprise Guide 6.1, SAS Institute Inc., Cary, NC, USA). For the computation of the Langholz & Jiao case−cohort estimates, the published SAS macro was used [9], the imputation was performed with the SAS procedures PROC MI and PROC MIANALYZE, the *C*-index was computed with the SAS code published by Liu et al. [34], and the NRI was computed using parts of a published SAS macro %nriidi by Lars Berglund and adaptation to MI. For testing proportional hazards assumption based on Schoenfeld residuals, the SAS macro %schoen was used which was published by Mayo Clinic and created and modified by Erik Bergstralh, Terry Therneau, and Ryan Lennon [35].

3. Results

Baseline characteristics for the full cohort, random subcohort, and incident diabetes cases are shown in Table 1, showing a clear comparability of the full cohort with the subcohort and a more adverse risk profile for participants developing diabetes during the study period.

Proportional hazards assumption was neither violated in full cohort analysis nor in any of the weighted case−cohort analyses (data not shown).

3.1. Multiple imputation

From the imputation step for the ~90% of missing information of waist circumference, a fraction of missingness of 84−87% was observed when using a simple imputation model; this was decreasing with a higher number of imputations (Fig. S1A/Appendix at www.jclinepi.com). The relative efficiency increased with increasing imputations, being 92% for $m = 10$ and 98% for $m = 40$ imputations. Using a more complex imputation model improved the imputation step: the fraction of missingness decreased to 57% ($m = 40$) while the relative efficiency increased to almost 99% ($m = 40$) (Fig. S1B/Appendix at www.jclinepi.com). Because of this strong improvement, only the MI results from the more complex imputation model will be presented hereafter. Fig. S2/Appendix at www.

Table 1. Baseline characteristics of the EPIC-Potsdam study population for the full cohort, random subcohort, and incident diabetes cases

Risk factor	Full cohort (N = 21,845)	Subcohort (N = 2,078)	Incident cases (N = 693)
Age (yr)	49.0 (15.0)	49.0 (16.0)	56.0 (11.0)
Females (%)	62.7	62.7	42.9
Body height (cm)	167 (12.3)	167 (12.4)	169 (12.3)
Waist circumference (cm)	84.5 (19.0)	84.5 (19.0)	100 (15.5)
Prevalent hypertension (% yes)	30.1	30.5	57.7
Smoking behavior (%)			
Former smoker (<20 units/d)	23.2	23.6	24.4
Former smoker (≥20 units/d)	8.55	9.00	19.9
Current smoker (<20 units/d)	14.2	14.0	11.1
Current smoker (≥20 units/d)	4.95	5.25	8.37
Sports, biking, and gardening (hr/wk)	4.50 (6.50)	4.50 (6.00)	4.50 (7.00)
Whole grain intake (bread, roles, muesli; 50 g portion/d)	0.61 (1.34)	0.67 (1.39)	0.44 (1.09)
Coffee consumption (150 mL portion/d)	2.00 (2.00)	2.00 (2.00)	2.00 (2.00)
Red meat intake (150 g portion/d)	0.23 (0.20)	0.24 (0.21)	0.28 (0.27)
Family history of diabetes (%)			
One parent with diabetes	21.6	21.9	30.0
Both parents with diabetes	2.32	2.26	5.34
A sibling with diabetes	5.06	4.67	11.3

Abbreviation: EPIC-Potsdam, European Prospective Investigation into Cancer and Nutrition Potsdam.

jclinepi.com shows the HRs and corresponding 95% CIs from different numbers of imputations; there was almost no difference of the estimates. After combining the MI estimates of the whole analysis model with Rubin's rules (Figs. S3 and S4/Appendix at www.jclinepi.com), the fraction of missingness for waist circumference decreased to less than 25% and the relative efficiency increased to greater than 97% when compared to the imputation step before. For an increasing number of imputations, only slight changes could be observed, and for $m = 10$ imputations, already a good performance of the imputation procedure in terms of fraction of missingness and relative efficiency was found. Therefore, we will present these estimates exemplary for all imputations when comparing it to case–cohort and full cohort estimates further on.

3.2. Cox proportional hazards regression

In Fig. 1, HRs and 95% CIs are presented from full cohort analysis, differently weighted case–cohort analyses, and MI. The estimates were similar and of similar precision for all approaches for most continuous variables including waist circumference; only for red meat intake, the variation was stronger for the case–cohort weighting approaches. For the binary variables, a better agreement between MI and full cohort estimates was observed; weighted case–cohort estimates were generally similar to full cohort and MI estimates; however, slight to strong difference was observed for hypertension, current smoking (<20 units/day), and for the information of both parents with diabetes. However, in

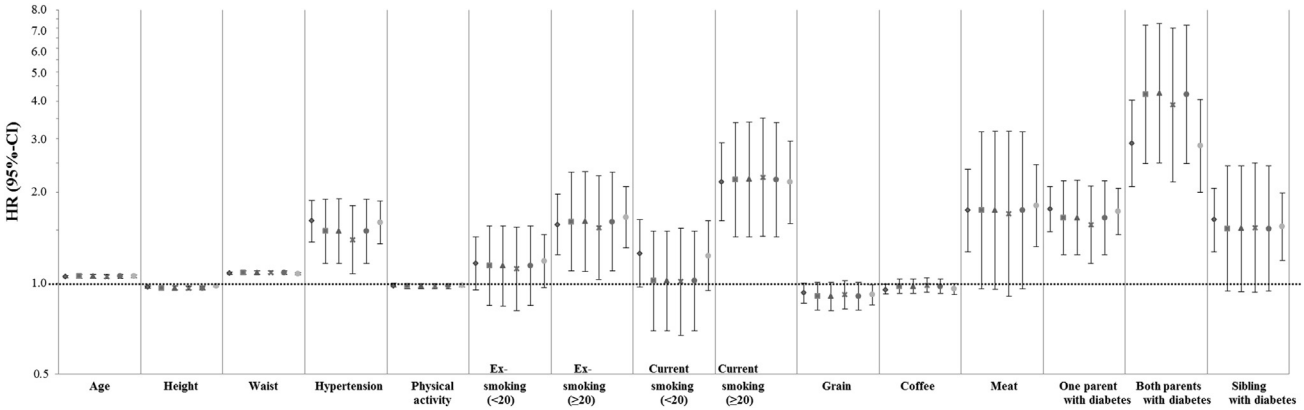

Fig. 1. Comparison of relative risks (HRs) derived from full cohort estimation, different case–cohort weighting methods, and multiple imputation. Hazard ratios (HRs) were derived from Cox regression based on full cohort analysis with all available information (diamond), case–cohort analysis with Prentice weighting (square), with Barlow weighting (triangle), with Self & Prentice weighting (star) and with Langholz & Jiao weighting (dark gray dot) and multiple imputation (MI) for missing waist circumference in the full cohort (light gray dot). MI was based on a more complex imputation model with $m = 10$ imputations assuming waist circumference information to be only available in the case–cohort study also used for weighted case–cohort analyses.

general, there was no difference between the varying weighting approaches.

3.3. Comparison of performance measures

Table 2 gives an overview of performance measures determined for the prediction models using different approaches. When comparing discrimination in terms of the *C*-index across all approaches, MI resulted in almost identical values for the original GDRS (C-index$_2$) and for a model without waist circumference (C-index$_1$) when compared to full cohort estimates. In contrast, all weighting methods resulted in estimates different from full cohort. Although Prentice, Barlow, and Self & Prentice methods provided identical estimates for discrimination, the weighted version of the *C*-index in the case–cohort setting with Langholz & Jiao weighting differed and yielded smaller deviations from the full cohort estimate compared to the other weighting methods. Differences in discrimination between C-index$_1$ and C-index$_2$ were comparable across all approaches (0.074 for full cohort, 0.075 for MI, 0.061 for Langholz & Jiao weighting, 0.060 for remaining weighting approaches). The weighted *C*-indices however showed closer estimates to full cohort and MI with differences in discrimination of 0.099 for Langholz & Jiao weighting and 0.072 for the remaining weighting approaches.

For the NRI, we observed a stronger variation across the different approaches. Full cohort analysis resulted in an overall NRI of 0.338 (0.276–0.401) with a stronger improvement among cases compared to noncases with 0.278 (0.216–0.341) and 0.060 (0.053–0.067), respectively. In contrast, Prentice and Self & Prentice weighted case–cohort analysis resulted in a comparable overall NRI with a stronger improvement among noncases compared to cases. The closest estimates provided Barlow and Langholz & Jiao weighted case–cohort analysis and MI.

The calibration plots showed very good agreement between predicted and observed risks for the full cohort analysis, MI, and for Barlow and Langholz & Jiao weighted analysis (Fig. 2A and C). Prentice and Self & Prentice weighted analyses showed strong overestimation of predicted vs. observed risks (Fig. 2B), and MI (Fig. 2B) showed similar calibration to full cohort analysis. The HL test was significant for all approaches; for MI, a combined test was not performed. Mean predicted risks derived from Langholz & Jiao weighting, Barlow weighting, and MI were 0.02, whereas Prentice and Self & Prentice weighting showed mean predicted risks of 0.15 and 0.16. The observed event rate was 0.02 (Table 2).

3.4. Agreement with full cohort absolute risk estimates

When comparing absolute risk determined in the full cohort with risks from the varying approaches based on different survival estimates and mean score points, we found very good agreement for predicted risks derived from Barlow weighting and Langholz & Jiao weighting, whereas Prentice and Self & Prentice weighting showed a severe disagreement (Fig. 3). Accordingly, the median error (p5; p95) was lowest for Barlow and Langholz & Jiao weighting with −0.0001 (−0.007; 0.014) and highest for Prentice and Self & Prentice weighting indicating strong overestimation with 0.111 (0.011; 0.642) and 0.101 (0.010; 0.599), respectively. Predicted risks were in good agreement with full cohort risk estimates (Fig. 3), independent of the number of imputations (Fig. S5/Appendix at www.jclinepi.com). However, variation was stronger for MI than for Barlow and Langholz & Jiao weighted approaches median error for MI: 0.0003 (−0.0138; 0.0103). The risk distributions from different approaches displayed in Fig. 4 further illustrate the disagreement of Prentice weighted and Self and Prentice weighted estimates with a strong disagreement of the distributions when compared to full cohort analysis; distribution of risks derived from Barlow or Langholz and Jiao weighting was similar to full cohort risk distribution. Results from MI provided a slightly steeper distribution of individual risks with a smoother tail compared to full cohort distribution of risks; still, this difference was small.

Highlighting the strong influence of the differently estimated survival functions from weighted case–cohort analyses on absolute risk calculation, we compared risk distributions and agreement with full cohort risks when using the survival estimate from the full cohort analysis. To do so, we used coefficients from weighted analyses and 0.99061 as survival from full cohort analysis in the equation for calculating absolute risks. Fig. S6/Appendix at www.jclinepi.com shows that all risk distributions for weighted case–cohort analyses were almost identical and very similar to the distribution in the full cohort. Also the agreement with full cohort estimates was very good (Fig. S7/Appendix at www.jclinepi.com), confirming the strong influence of the survival estimate for absolute risk calculation.

4. Discussion

We compared different approaches applicable to analyze case–cohort data in the context of developing and extending risk prediction models. We observed that estimates of relative risks were generally similar for differently weighted Cox regression approaches and MI for case–cohort data compared to full cohort analysis. With regard to absolute risk estimation and performance measures such as *C*-index, calibration, or NRI which were based on these estimates, we found good agreement between MI results and results from full cohort analysis, whereas weighted Cox regression approaches showed generally lower discrimination measures and only comparable results for the weighted *C*-index. Slight to large difference for NRI,

Table 2. Overview of performance measures for prediction modeling after application of full cohort, weighted case–cohort and multiple imputation methods to the data

		Case–cohort analysis with weighted estimators				Multiple imputation (MI)
Performance measure	**Full cohort analysis**	**Prentice weighting**	**Self & Prentice weighting**	**Barlow weighting**	**Langholz and Jiao method**	**MI with $m = 10$ imputations**
Discrimination						
C-index$_1$	0.767	0.708	0.708	0.708	0.708	0.766
(95% CI)	(0.688, 0.837)	(0.622, 0.787)	(0.622, 0.787)	(0.622, 0.787)	(0.622, 0.787)	(0.687, 0.837)
C-index$_2$	0.841	0.769	0.769	0.768	0.769	0.841
(95% CI)	(0.771, 0.901)	(0.687, 0.841)	(0.687, 0.841)	(0.687, 0.841)	(0.687, 0.841)	(0.770, 0.901)
Weighted C-index$_1$ (95% CI)	-	0.745 (0.718, 0.772)	0.745 (0.718, 0.772)	0.745 (0.718, 0.772)	0.745 (0.718, 0.772)	-
Weighted C-index$_2$ (95% CI)	-	0.817 (0.793, 0.841)	0.817 (0.793, 0.841)	0.817 (0.793, 0.841)	0.844 (0.819, 0.867)	-
Reclassification						
NRI (95% CI)	0.338 (0.276, 0.401)	0.380 (0.339, 0.421)	0.418 (0.377, 0.460)	0.338 (0.271, 0.406)	0.341 (0.273, 0.408)	0.326 (0.263, 0.389)
NRI$_{noncases}$ (95% CI)	0.060 (0.053, 0.067)	0.348 (0.322, 0.374)	0.393 (0.367, 0.419)	0.060 (0.038, 0.083)	0.065 (0.042, 0.087)	0.121 (0.114, 0.128)
NRI$_{cases}$ (95% CI)	0.278 (0.216, 0.341)	0.032 (0.001, 0.064)	0.026 (−0.007, 0.058)	0.278 (0.215, 0.342)	0.276 (0.213, 0.340)	0.205 (0.143, 0.268)
Calibration						
Graphical evaluation	Good	Poor	Poor	Good	Good	Good
Calibration plot		Strong overestimation across all risk groups	Strong overestimation across all risk groups	Slight underestimation in higher risk groups	Slight underestimation in higher risk groups	Very similar to full cohort
Calibration-in-the-large						
Observed event rate	0.02					
Mean predicted risk		0.16	0.15	0.02	0.02	0.02
Agreement with full cohort risk estimates	-	No agreement; severe overestimation	No agreement; severe overestimation	Very good agreement; stronger variation in the upper risks	Very good agreement; stronger variation in the upper risks	Good agreement; stronger variation in the upper risks
Median error (P5; P95)	-	0.111 (0.011; 0.642)	0.101 (0.010; 0.599)	−0.0001 (−0.007; 0.013)	−0.0001 (−0.007; 0.014)	0.0003 (−0.0138; 0.0103)
HL test	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	-

Abbreviations: CI, confidence interval; NRI, net reclassification improvement; HL, Hosmer–Lemeshow.

C-index$_1$ reflects the *C*-index for the prediction without waist circumference and *C*-index$_2$ for the original German Diabetes Risk Scores (GDRS); this refers also to the weighted *C*-index.

risk distributions, or calibration were found for Prentice and Self & Prentice weighting methods; Barlow and Langholz & Jiao weighting methods showed similar findings to full cohort and MI analysis.

For a comparison of weighted estimators with true full cohort estimators, we used an assumed scenario with missing values for waist circumference for participants not included in the case–cohort study. Relative risks (HRs) from weighted Cox regression analyses were generally similar to full cohort analysis which is in line with previous findings [8,14,15]; however, a comparison of weighting methods by Onland-Moret et al. [36] found the Prentice's method to be closest to full cohort relative risk estimates which is in contrast with our findings. The influence of the chosen method on relative risk estimates of other predictors in the model was stronger for weighted analyses than for MI, especially for binary variables. In line with previous studies [8,15,37], MI estimates were in good agreement with the full cohort estimates for all variables; however, MI was applied to a large amount of missing values for a single continuous variable, and our findings may not be generalizable to binary or categorical predictors.

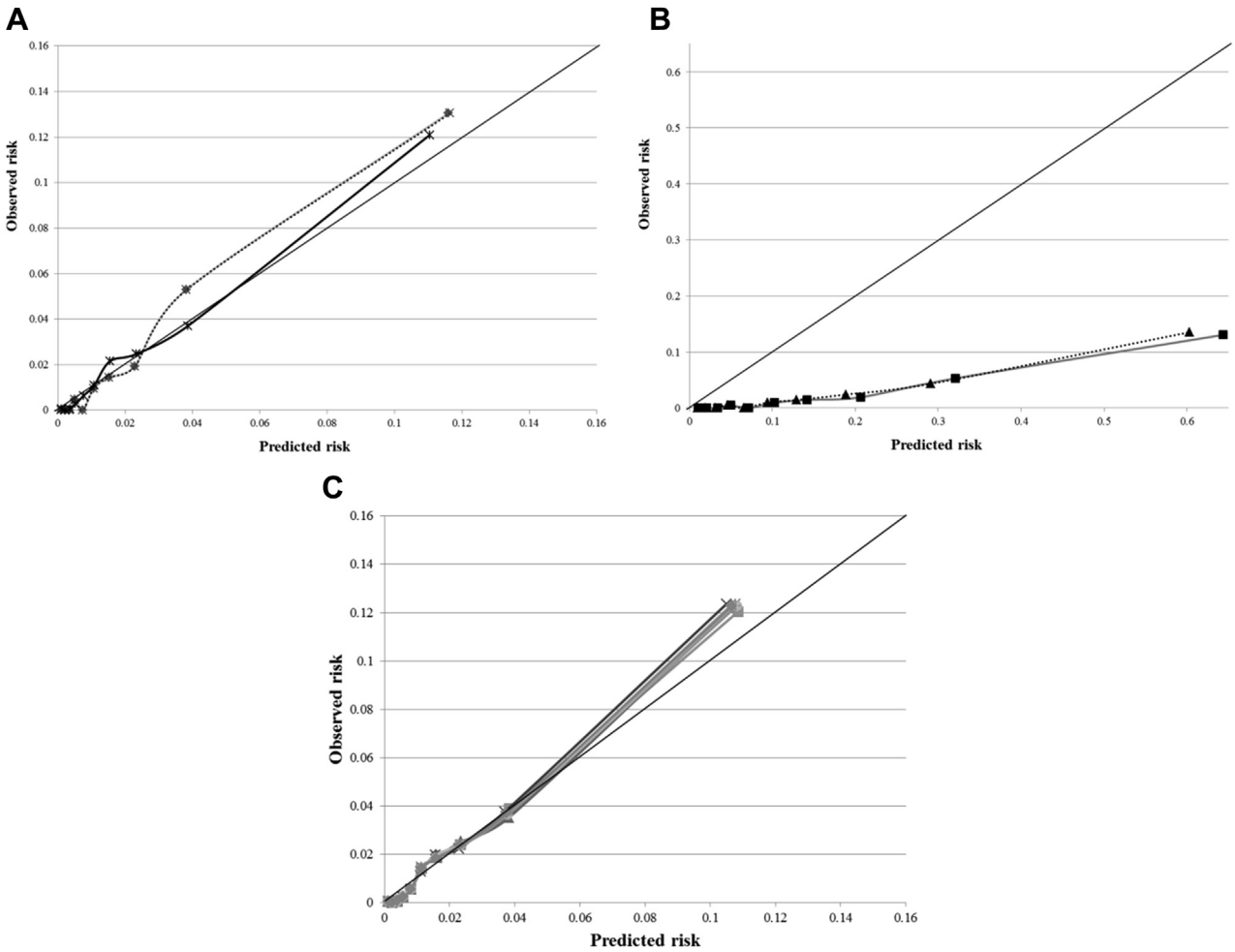

Fig. 2. Calibration plots for the GDRS determined for differently weighted risks from case–cohort analysis, full cohort risks, and risks derived from multiple imputation. (A) Predicted 5-year risks were determined using score points, mean score points, and the baseline survival derived from full cohort analysis (black solid line), a case–cohort analysis with Barlow weighting (black dotted line), and with Langholz & Jiao weighting (gray solid line). Predicted 5-year risks were plotted against observed risks in 10 equally sized groups. For weighted analysis, calibration was determined in the subcohort. (B) Predicted 5-year risks were determined using score points, mean score points, and the baseline survival derived from case–cohort analysis with Prentice weighting (dark gray solid line), with Self and Prentice weighting (black dotted line). Predicted 5-year risks were plotted against observed risks in 10 equally sized groups. For weighted analysis, calibration was determined in the subcohort. (C) Predicted 5-year risks were determined using score points, mean score points, and the baseline survival derived from a full cohort analysis using multiple imputation with $m = 10$ imputations with the more complex imputation. GDRS, German Diabetes Risk Score.

Comparing MI and weighted case–cohort analyses, previous findings rather indicate more precision and more likely unbiased estimates with MI than with weighted case–cohort analyses [8,15,37]. Although the authors of these studies recommended building an analysis model based on weighted estimators and reanalyzing the data using MI [8,15], they also highlighted the comparability and usefulness of MI in the context of risk prediction modeling. Later on, Keogh and White suggested applying full cohort analyses with MI for case–cohort studies for survival analysis [37]. In line with this, we could confirm the usefulness of MI in our proof-of-concept study, especially when providing an equation for the calculation of absolute risks for practical usage. When calculating absolute risks, the main driving factor appears to be the baseline survival; this could be illustrated in the present study by comparing calibration plots, risk distributions, and evaluating the agreement with full cohort risks between the different approaches. Although discrimination was identical across all weighting methods indicating that the baseline survival does not affect discrimination, differences were found for calibration and reclassification between the varying weighting approaches. By further comparing risk distributions and agreement of full cohort and weighted estimates based on survival estimates from case–cohort and from full cohort analysis, we showed that agreement was very good when using full cohort survival rather than case–cohort survival, highlighting the impact of the correct baseline survival estimation. In contrast to our findings, Ganna et al. found unbiased estimates and a good agreement with full cohort

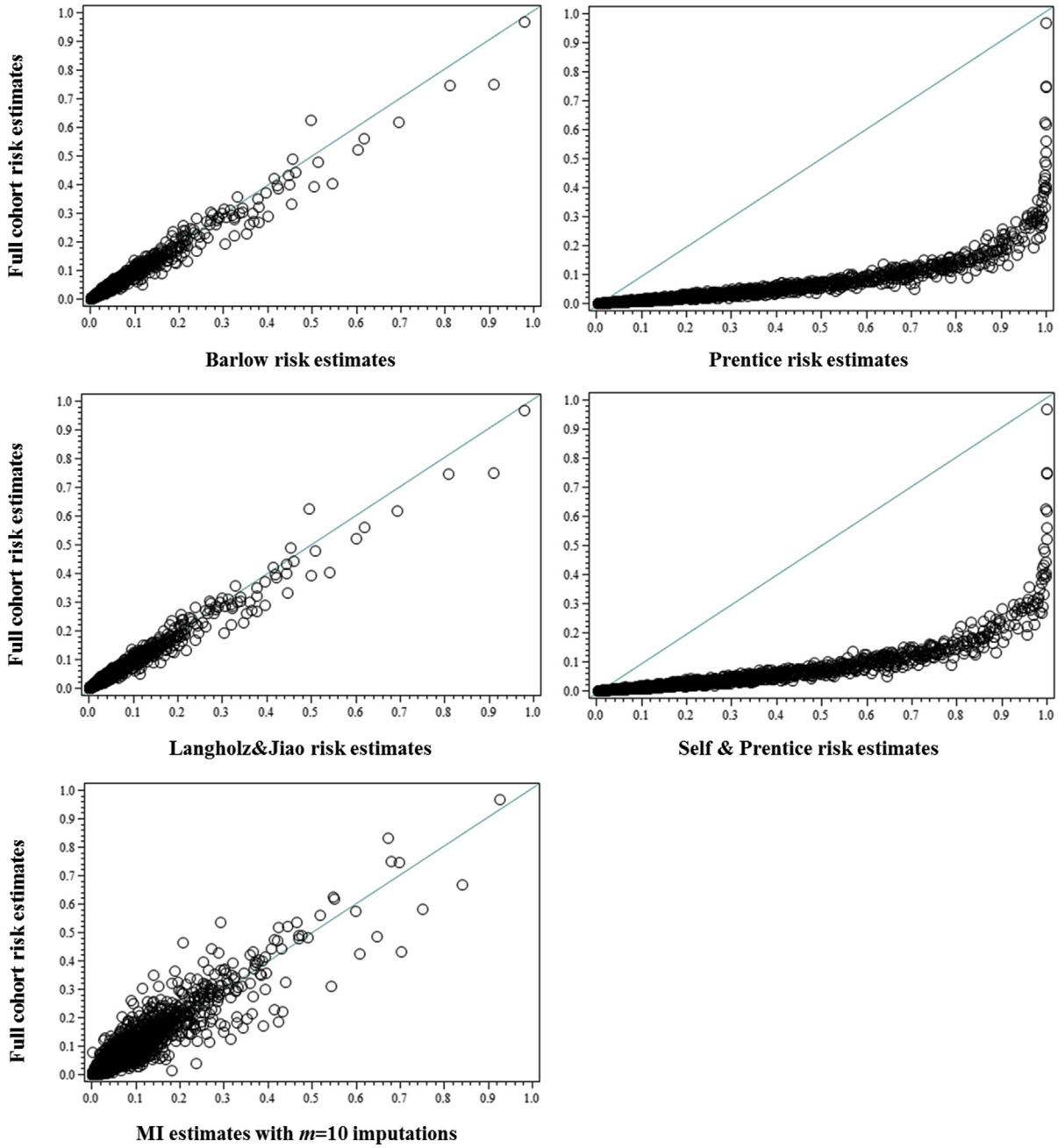

Fig. 3. Agreement of absolute 5-year risk estimates from full cohort analysis with differently weighted case–cohort estimates and multiple imputation ($m = 10$). Absolute 5-year risks were determined using score points derived from beta coefficients multiplied with risk factor values (linear predictor) and survival estimates from Cox regression using full cohort, differently weighted case–cohort analyses, or multiple imputation (MI, $m = 10$) based on the more complex imputation model including more variables than the analysis model. For MI, results were combined with Rubin's rules. The following equation was used for absolute risk estimation: $P(\text{diabetes}) = 1 - S_0^{\exp(\text{score points} - \text{mean score points})}$ with S_0 being the estimated baseline survival function of the respective approach. Individual 5-year risks from case–cohort or mean individual 5-year risks from 10 for MI were plotted against the full cohort risk estimates. A diagonal line implies perfect agreement; values above imply underestimation and values below overestimation.

estimates from weighted analyses [14]. Nevertheless, they suggest to calculate absolute risks based on the baseline hazard or survival from the randomly sampled subcohort, similar to the approach proposed by Langholz and Borgan for case–control settings [38]. The most appropriate methods, however, appear to be the computations proposed by Barlow or by Langholz & Jiao [9] or MI. These methods yielded results close to full cohort analysis for the *C*-index, calibration, calibration-in-the-large, NRI, risk distribution, and agreement with mean errors close to 0. However, adaption to the methods was applied such as restricting calibration to the subcohort only and excluding cases appearing

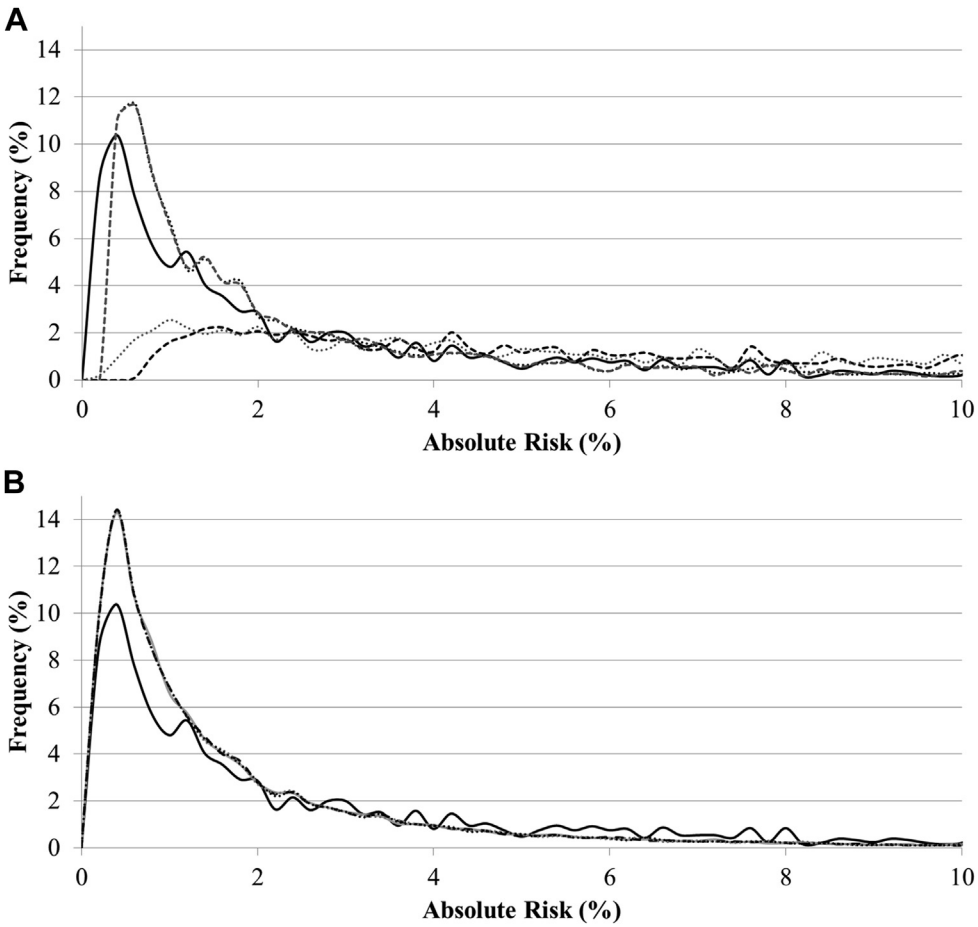

Fig. 4. Distribution of absolute 5-year risks derived from full cohort estimation and different case–cohort weighting methods (A) and from multiple imputation (B). (A) Absolute 5-year risks were derived from beta coefficients and survival estimates from Cox regression in a full cohort analysis (black solid line), a case–cohort analysis with Barlow weighting (black dotted), with Prentice weighting (black dashed line), with Self and Prentice weighting (gray dotted line), and with Langholz & Jiao weighting (gray dashed line) multiplied with risk factor values (linear predictor). (B) Absolute 5-year risks were derived from beta coefficients and survival estimates from Cox regression in a full cohort analysis (black solid line), and a full cohort analysis using multiple imputation with $m = 10$ (gray solid line), $m = 20$ (black dashed line), $m = 30$ (gray dotted line), and $m = 40$ (black dotted line) imputations with the more complex imputation model including additional variables when compared to the analysis model. The following equation was used: $P(\text{diabetes}) = 1 - S_0^{\exp(\text{score points} - \text{mean score points})}$ with S_0 being the estimated baseline survival function of the respective approach. The illustration was limited to 10% of 5-year risk as the distribution became stable and for discriminating the separate risk distributions.

after 5 years, who count as noncases for reclassification, and increase the absolute risks for this group artificially [32]. Compared to MI, Prentice and Self & Prentice weighting methods showed severe disagreement compared to full cohort analysis with systematic overestimation of individual risks resulting in poor calibration and calibration-in-the-large and stronger deviation of NRI when compared to full cohort estimates. A good calibration was observed for Barlow and Langholz & Jiao weighting, being similar to MI and the full cohort estimates; this was confirmed with evaluation of calibration-in-the-large.

With regard to discrimination and reclassification, Ganna et al. [14] found similar estimates for weighted performance measures and real estimates from the full cohort, though differences were observed in the study by Marti et al. [15] and in our study. In particular, although Marti et al. reported higher predictive abilities from case–cohort estimation, in the present study, we observed lower values for discrimination with the weighted *C*-index being closer to full cohort estimates than the unweighted. Similar results for the weighted and full cohort *C*-index were also reported by Sanderson et al. [32]. The differences between the studies might be explained by the different scenarios, that is, empirical vs. simulated data, different subcohort sizes or unstratified vs. stratified sampling. This highlights the influence of the study itself and the extent to which results can or cannot be transported to other situations.

With regard to the MI procedure, compared to a simple imputation model, we observed that with a well-defined imputation model, the relative efficiency increases, the fraction of missingness decreases, and the number of imputations needed decreases. However, we investigated only a limited set of variables potentially contributing to the

imputation of waist circumference based on prior knowledge rather than using all information from the whole data set; we also used an arbitrary cutoff for the correlation coefficient with 0.20. A more detailed analysis of potentially useful information with the aim to find the perfect imputation model would require much more time and computation effort. The interdependency of the number of imputations and the fraction of missingness or relative efficiency was already displayed by Graham et al. [25]. Marti et al. also highlighted that the relative efficiency is a function of the fraction of missingness and the number of imputations [8]. In many situations, 5−10 imputations appear to be sufficient to get valid results. In our study, with a very large proportion of missing values (~90%), no further increase in accuracy or validity was observable after 10 imputations. Similarly, Keogh and White found no substantial differences for more than 10 imputations, although their study was based on a small subcohort sampling rate (5%). However, the major drawback of applying MI is that the result of the imputation process is highly sensitive to changes in any specification of this process, for example, the definition of a correct imputation model and the order of the variables in the imputation model for computation of the imputation process. We ordered the variables by increasing amount of missing values so that MI is based on less already imputed data for other predictors. We cannot rule out that our specifications had an impact on prediction. Yet, results were similar compared to a more simple imputation model, and therefore, we assumed the impact to be limited. Still, our results as proof-of-concept for this empirical study cannot be evaluated from a more general perspective but rather need to be specifically interpreted based on all definitions for the imputation process. Therefore, it is pivotal to specify all assumptions and each step for the imputation process beforehand.

One limitation of this study is that we used an empirical example to compare the different estimates with true estimates rather than using simulated data. This might limit the extent to which the results can be generalized to other studies. However, using this approach on empirical data, we were able to compare the different methods with true data and this is also frequently used for analyses in contrast to simulated data. Furthermore, only the unstratified case−cohort study design was displayed, and the application to other study designs such as stratified case−cohort or case−control study cannot be evaluated. For NRI, we also need to take into account the several concerns [39] related to the number of risk categories, choice of risk cutoffs [40,41], the problematic statistical test and reliance on the *P*-value, or the inflated positive values for NRI due to overfitting [33,42,43]. In conclusion, these findings support the application of MI to facilitate a full cohort analysis for risk prediction modeling when information is collected in case−cohort settings, usually with a high amount of missing information. Especially for the calculation of absolute risks and the following evaluation of discrimination, reclassification, or calibration, MI might be superior to weighted case−cohort analysis.

Acknowledgments

The authors thank Dr. Manuela Bergmann who was responsible for the methodological and organizational work of data collections of exposures and outcomes and Wolfgang Fleischhauer for his medical expertise that was used in case ascertainment and contacts with the physicians and Ellen Kohlsdorf for data management.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.12.019.

References

[1] Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1−11.
[2] Thorand B, Schneider A, Baumert J, et al. [Case-cohort studies: an effective design for the investigation of biomarkers as risk factors for chronic diseases−demonstrated by the example of the MONICA/KORA Augsburg Case-Cohort Study 1984-2002]. Gesundheitswesen 2005;67 Suppl 1:S98−102.
[3] Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice—experiences from the MORGAM Project. Epidemiol Perspect Innov 2007;4:15.
[4] Prentice RL, Self SG. Aspects of the use of relative risk models in the design and analysis of cohort studies and prevention trials. Stat Med 1988;7:275−87.
[5] Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. J Clin Epidemiol 1999;52:1165−72.
[6] Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research−part 2: multiple imputation. Acad Emerg Med 2007;14:669−78.
[7] Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
[8] Marti H, Chavance M. Multiple imputation analysis of case-cohort studies. Stat Med 2011;30:1595−607.
[9] Langholz B, Jiao J. Computational methods for case-cohort studies. Comput Stat Data Anal 2007;51:3737−48.
[10] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003;56:28−37.
[11] Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092−101.
[12] Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol 2010;63:205−14.
[13] Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 2009;9:57.
[14] Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. Am J Epidemiol 2012;175:715−24.

[15] Marti H, Carcaillon L, Chavance M. Multiple imputation for estimating hazard ratios and predictive abilities in case-cohort surveys. BMC Med Res Methodol 2012;12:24.
[16] Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43(4):205−15.
[17] Bergmann MM, Bussas U, Boeing H. Follow-up procedures in EPIC-Germany−data quality aspects. European Prospective Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43(4):225−34.
[18] Bergmann MM, Jacobs EJ, Hoffmann K, Boeing H. Agreement of self-reported medical history: comparison of an in-person interview with a self-administered questionnaire. Eur J Epidemiol 2004;19:411−6.
[19] Mühlenbruch K, Ludwig T, Jeppesen C, et al. Update of the German Diabetes Risk Score and external validation in the German MONICA/KORA study. Diabetes Res Clin Pract 2014;104:459−66.
[20] Schulze MB, Weikert C, Pischon T, et al. Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. Diabetes Care 2009;32:2116−9.
[21] Stefan N, Fritsche A, Weikert C, et al. Plasma fetuin-A levels and the risk of type 2 diabetes. Diabetes 2008;57:2762−7.
[22] Schulze MB, Hoffmann K, Boeing H, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care 2007;30:510−5.
[23] Mühlenbruch K, Joost H-G, Boeing H, Schluze MB. Risk prediction for type 2 diabetes in the German population with the updated German Diabetes Risk Score (GDRS). Ernahrungs Umschau 2014;61:90−3.
[24] Xue X, Xie X, Gunter M, et al. Testing the proportional hazards assumption in case-cohort analysis. BMC Med Res Methodol 2013;13:88.
[25] Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 2007;8:206−13.
[26] Rubin DB. Multiple imputation for nonresponse in surveys. New York: J. Wiley & Sons; 1987.
[27] Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004;23:2109−23.
[28] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 2013;13:33.
[29] Steyerberg EW. Clinical prediction models. A practical approach to development, validation, and updating. New York: Springer Science+Business Media; 2009.
[30] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157−72. discussion 207−112.
[31] Hosmer DWJ, Lemeshow S. Applied logistic regression. [Chapter 5]. New York: John Wiley & Sons; 2000.
[32] Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. BMC Med Res Methodol 2013;13:113.
[33] Mühlenbruch K, Kuxhaus O, Pencina MJ, Boeing H, Liero H, Schulze MB. A confidence ellipse for the Net Reclassification Improvement. Eur J Epidemiol 2015;30:299−304.
[34] Liu L, Forman S, Barton B (2009) Fitting Cox model using PROC PHREG and beyond in SAS. Proceedings of SAS Global Forum 2009 Paper 236−2009.
[35] Bergstralh E, Therneau T, Lennon R (2011). Available at http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros. Accessed February 17, 2017.
[36] Onland-Moret NC, van der AD, van der Schouw YT, et al. Analysis of case-cohort data: a comparison of different methods. J Clin Epidemiol 2007;60:350−5.
[37] Keogh RH, White IR. Using full-cohort data in nested case-control and case-cohort studies by multiple imputation. Stat Med 2013;32:4021−43.
[38] Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. Biometrics 1997;53:767−74.
[39] Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology 2014;25:114−21.
[40] Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. Am J Epidemiol 2010;172:353−61.
[41] Mühlenbruch K, Heraclides A, Steyerberg EW, Joost HG, Boeing H, Schulze MB. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. Eur J Epidemiol 2013;28:25−33.
[42] Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets. Stat Biosciences 2015;7:282−95.
[43] Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? J Natl Cancer Inst 2014;106:dju041.