International Journal of Epidemiology, 2020, 876–884
doi: 10.1093/ije/dyz206
Advance Access Publication Date: 16 October 2019
Original article

P Values

Big data, observational research and *P*-value: a recipe for false-positive findings? A study of simulated and real prospective cohorts

Giovanni Veronesi [1]* Guido Grassi,[2] Giordano Savelli,[3] Piero Quatto,[4] and Antonella Zambon[5]

[1]Research Center in Epidemiology and Preventive Medicine, Department of Medicine and Surgery, University of Insubria, Varese, Italy, [2]Clinica Medica, Department of Medicine and Surgery, University of Milano-Bicocca, Milano, Italy, [3]U.O. Medicina Nucleare, Fondazione Poliambulanza Istituto Ospedaliero, Brescia, Italy, [4]Department of Economics, Management and Statistics and [5]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy

*Corresponding author. Research Center in Epidemiology and Preventive Medicine, Department of Medicine and Surgery, University of Insubria, Via Rossi, 9, 21100 Varese, Italy. E-mail: giovanni.veronesi@uninsubria.it

Editorial decision 14 August 2019; Accepted 11 September 2019

Abstract

Background: An increasing number of observational studies combine large sample sizes with low participation rates, which could lead to standard inference failing to control the false-discovery rate. We investigated if the 'empirical calibration of *P*-value' method (EPCV), reliant on negative controls, can preserve type I error in the context of survival analysis.

Methods: We used simulated cohort studies with 50% participation rate and two different selection bias mechanisms, and a real-life application on predictors of cancer mortality using data from four population-based cohorts in Northern Italy ($n = 6976$ men and women aged 25–74 years at baseline and 17 years of median follow-up).

Results: Type I error for the standard Cox model was above the 5% nominal level in 15 out of 16 simulated settings; for $n = 10\,000$, the chances of a null association with hazard ratio = 1.05 having a *P*-value < 0.05 were 42.5%. Conversely, EPCV with 10 negative controls preserved the 5% nominal level in all the simulation settings, reducing bias in the point estimate by 80–90% when its main assumption was verified. In the real case, 15 out of 21 (71%) blood markers with no association with cancer mortality according to literature had a *P*-value < 0.05 in age- and gender-adjusted Cox models. After calibration, only 1 (4.8%) remained statistically significant.

Conclusions: In the analyses of large observational studies prone to selection bias, the use of empirical distribution to calibrate *P*-values can substantially reduce the number of trivial results needing further screening for relevance and external validity.

Key words: Observational studies, selection bias, big data, calibration of *P*-value, cohort studies, survival analysis

Key Messages

- An increasing number of observational studies combine large sample sizes with low participation rates. It is not yet known if standard inference can adequately control type I error in this new situation.
- In simulated cohorts mimicking 50% participation rate and $n = 10\,000$, the chances of a null association with a hazard ratio of 1.05 being statistically significant are 42.5%. Conversely, we show in simulations and in a real study that using the empirical distribution to calibrate *P*-values can maintain the type I error rate close to the desired 5% nominal level.
- The analyses of large observational studies prone to selection bias require dedicated tools to limit false-positive findings. The calibration of *P*-values can substantially reduce the number of trivial results needing further screening for relevance.

Introduction

Non-communicable diseases represent the major global health challenge of the 21st century, with 36 million deaths (63% of the total) occurring globally in 2008.[1] Nevertheless, a substantial number of non-communicable diseases can be avoided or delayed until significantly late in life by prevention,[2] especially when healthy lifestyles are adopted and maintained from early adulthood to middle age.[3,4]

In a recent editorial, 'precision prevention' has been defined as the possibility of 'providing the right intervention to the right population at the right time' before disease manifestation.[5] An example is the identification of high-risk subpopulations, such as those in low socioeconomic groups,[6–7] for tailored preventive interventions. However, the current preventive guidelines are generally designed for the 'average individual' in the population, despite standard recommendations being potentially not beneficial (or possibly harmful) to specific subgroups. For instance, the recommended levels of leisure time physical activity increased cardiovascular disease risk among men with intense occupational physical activity.[8] The availability of 'big data' from large observational studies is therefore expected to help deal with the needs of preventive medicine.[9] The large sample size allows investigation of specific subgroups with enough statistical power, and the combination of data from multiple sources allows investigation of an increasing number of new markers to improve risk stratification.[10]

The appearance of 'big data' in the epidemiological arena is not free from methodological concerns.[11–13] The low participation rate observed in large biobanks[12] or the adoption of self-selected, convenient samples (e.g. in studies based on Mobile-research platforms) may exacerbate the risk of selection bias generally present in observational research.[13] This bias, combined with large sample sizes, may result in type I error rate inflation above the nominal alpha level and a number of false-positive results that need further screening for clinical relevance.[11] In the context of pharmacoepidemiological studies on drug administrative records, Schuemie and colleagues introduced the 'empirical calibration of *P*-value' as a method to control type I error.[14–15] In that specific field and using a logistic regression model, the method showed a satisfactory control of type I error rate,[14–16] at the price of a substantially increased probability of type II error.[16]

In this paper, we aim to investigate the method's control of type I error rate in the presence of selection bias and in the context of survival analysis, using a comprehensive plan of simulations. In addition, we apply the method to a collaborative cohort study, to discuss its performance and practicability in real-word analyses.

Methods

Review of the empirical calibration of P-value method

The empirical calibration of *P*-value method relies on a set of variables with no known association with the study outcome ('negative controls', NCs) to obtain a distribution of effect sizes under the null hypothesis ('empirical null distribution') in the available data.[14] After being estimated from the data, the parameters of the null empirical distribution can be used to calibrate the test statistic and the *P*-value for the exposure of interest, assuming that bias arises from the same distribution (aka 'exchangeability' assumption). Full details on the method are reported in the Supplementary data available at *IJE* online.

Simulation plan

We conceived selection bias as an underlying mechanism (Z) driving participation in the study sample used to make inference on the true relationship between an exposure (X) and an outcome (Y). In other words, the vector of data (x_j, y_j) for the j-th individual in the population of interest is observed with probability p_j which depends on Z: $p_j = \text{Prob}[\text{missing}(x_j, y_j) = 1 | z_j)]$. We simulated Z and the

continuous exposure of interest X from a bivariate normal distribution with mean vector $\begin{bmatrix}\mu_Z\\ \mu_X\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$ and variance-covariance matrix $\begin{bmatrix}1 & \sigma_{Z,X}\\ \sigma_{Z,X} & 1\end{bmatrix}$, with $\sigma_{Z,X}=0.2$. In sensitivity analyses, the choice of larger values of $\sigma_{Z,X}$ did not alter our findings substantially. The failure time T was generated from the exponential model as[17] $T=-\frac{\log(U)}{\lambda\exp(\beta_X*X+\beta_Z*Z)}$, where λ is a suitable constant, U is a random number for the uniform distribution on the unit interval, β_X was 0 (in simulations for type I error), 0.1 or 0.2 (in simulations for type II error), and β_Z was 0.2 or 0.4. Censorship time was also generated from the same exponential model, but independently of X and Z. The observed follow-up time was the lowest value between failure time, censorship time and a fixed constant to mimic administrative censorship due to the end of follow-up at a given date in real studies. We set the outcome variable $Y=1$ when the observed follow-up time was equal to the failure time, and $Y=0$ otherwise. The choice of the constant values yielded to an event rate in different scenarios of about 10%. The two scenarios in Tables 1 and 2 are a combination of values for β_Z and $\sigma_{Z,X}$, and represent different amounts of bias in the data. As a third step, we simulated negative controls as a number $i=1$ to I of continuous variables ($I=10$, 30 or 50) with no role in failure time T and therefore not associated with the study outcome. Each NCi was simulated from a bivariate normal distribution with mean vector $\begin{bmatrix}\mu_Z\\ \mu_{NCi}\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$ and variance-covariance matrix $\begin{bmatrix}1 & \sigma_{Z,NCi}\\ \sigma_{Z,NCi} & 1\end{bmatrix}$.

The correlation term $\sigma_{Z,NCi}$ ranged between 0.0 and 0.3 in one set of simulations (Table 1), and between -0.3 and 0.3 in a second set of simulations (Table 2). For a fixed value of β_Z, the sign of the correlation between X and Z drives the direction of the bias for the estimator of the association between X and the outcome. Therefore, the set of values for $\sigma_{Z,NCi}$ between -0.3 and 0.3 corresponds to a situation in which the exchangeability assumption is relaxed. We set two distinct values for sample sizes, $n=5000$ and $n=20\,000$; and a number $R=200$ of replications.

We then applied two distinct selection bias mechanisms. Under mechanism 1, simulated data (x_j, y_j) for the j-th individual were set at missing with probability p_j equal to 0.1, 0.5 and 0.9 according to sample tertiles of Z; i.e. p_j increased for increasing values of Z. Since Z is positively associated with the outcome, the event rate under the first mechanism was below the population value of 10%; i.e. each simulated dataset corresponds to a study setting in which healthier individuals are more likely to participate. Under mechanism 2, the probability p_j decreased for increasing values of Z (values of 0.9, 0.5 and 0.1), resulting in an event rate above the population value of 10%; i.e. each simulated dataset corresponds to a study setting in which unhealthier participants are more likely to participate. On average, in both mechanisms, the probability of missing data was 0.495, mimicking a real-word participation rate of about 50%. Finally, we also introduced the possibility of missing-at-random in the NC variables, as in real-life studies not all the variables are available for all the study participants. We estimated $\hat{\beta}_{X,\ raw}$ from Cox regression models on the available data after having applied the selection bias mechanisms (i.e. on datasets of size $n=2500$ and $n=10\,000$), and derived $\hat{\beta}_{X,\ calibrated}$ using the empirical calibration method. We computed: rejection rate, as the number of *P*-values below 0.05 on the total number R of replications; bias, as the average over R of the difference between the numerator in the t statistic for $\hat{\beta}_{X,\ raw}$ and $\hat{\beta}_{X,\ calibrated}$ with the true value of β_X; and standard deviation, as the average over R of the denominator of the t statistic for $\hat{\beta}_{X,\ raw}$ and $\hat{\beta}_{X,\ calibrated}$.

Real data application

The study sample is constituted by participants in four population-based prospective cohorts, namely three independent surveys of the MONICA-Brianza study (baseline periods: 1986–87, 1989–90 and 1993–94) and the PAMELA cohort study, with baseline visit between 1990 and 1993. A single team of researchers conducted these studies in the same area (located north of Milan) and adopted a standardized protocol for baseline visits and follow-up procedures.[8] Participation rates were between 65% and 70%. The follow-up activities for the study cohorts were approved by the local ethics committee of the University Hospital of Monza. A signed informed consent was not a requisite at the time of recruitment (1986–94). Follow-up for mortality ended in December 2008. At baseline, a fasting serum blood sample was drawn from each participant and $n=34$ markers were measured on the fresh specimens in a centralized laboratory. To investigate the association between blood markers and cancer mortality in the general population, we provided the list of available markers to a clinical oncologist (SG), initially blinded about the objective of the research. After a review of the relevant literature, SG indicated $n=21$ negative controls and $n=13$ positive associations, listed in Supplementary Table 1 and 2, available as Supplementary data at *IJE* online, respectively. We first estimated the hazard ratio of cancer mortality for 1 standard deviation-increase for each of the 34 markers from Cox regression models, adjusting for age and gender. To investigate type I error, we applied

the empirical calibration of *P*-value method to negative controls using a leave-one-out cross-validation approach.[15] Then, to assess type II error, we applied the empirical calibration method to each of the 13 positive controls, using the full set of 21 NCs.

The analyses were performed using standard procedures in the SAS Software, release 9.4; Figures 1 and 2 were drawn using R. Programming statements for the simulation study are available as Supplementary data at *IJE* online.

Results

Simulation study

Table 1 shows the simulation results for type I error when the exchangeability assumption is met. The rejection rate for the raw method (standard analysis) was above the nominal 5% level in seven out of eight simulated scenarios, with a peak of 42.5%. The rate rose with increasing bias and sample size, and it was generally higher in selection mechanism 2 (i.e. when unhealthier individuals are more likely to participate in the study) than in selection mechanism 1. Conversely, the empirical calibration of *P*-values maintained the rejection rates close to the nominal 5% level in both selection mechanisms, at the same time reducing average bias by 80–90% compared with the bias of the raw method. These figures were already achieved with a number of NC equal to 10.

When we relaxed the exchangeability assumption (Table 2), the empirical calibration of *P*-values still maintained a satisfactory control of type I error rate compared with the raw method, in all the combinations of selection mechanism, bias scenario and sample size. The method's performance generally improved for increases in the number of NCs. However, there was no effect in terms of reduction in average bias compared with the raw estimator; the lower rejection rate was obtained through an increase in the average standard deviation. Supplementary Figure 1, available as Supplementary data at *IJE* online, shows the forest plot for 30 NCs corresponding to one simulated dataset when the exchangeability assumption is met (panel a) or not met (panel b).

Figure 1 shows the rejection rates for true effect sizes β_X equal to 0.1 and 0.2 (corresponding to hazard ratios of 1.1. and 1.2 per 1 SD increase in X, respectively), from the type II error simulation study when requirements for the exchangeability assumption are met and under a mild bias. Under selection mechanism 1 (panel a), the empirical calibration of the *P*-value method is underpowered only for a sample size of 2500 and a true hazard ratio of 1.1. It is worthy of note that under the same conditions, the standard method has a rejection rate below 50% as well.

Table 1. Rejection rate, average bias and average standard deviation (SD) over R = 200 replications using standard and calibrated *P*-value, for type I error simulation study under the exchangeability assumption[a]

Bias scenario	Sample size[b]	Method	Selection mechanism 1[c]			Selection mechanism 2[d]		
			Rejection rate (%)	Bias	SD	Rejection rate (%)	Bias	SD
Mild ($\beta_Z = 0.2$, corr(X, Z)=0.2)	$N = 2500$	Raw	5.0	0.024	0.068	7.0	0.025	0.061
		Calib, # NC=10	5.0	0.003	0.074	5.0	0.005	0.069
		Calib, # NC=30	3.5	0.004	0.073	3.0	0.006	0.066
		Calib, # NC=50	3.5	0.005	0.071	3.5	0.006	0.064
	$N = 10\,000$	Raw	12.5	0.023	0.034	14.5	0.025	0.030
		Calib, # NC=10	5.5	0.005	0.039	4.0	0.007	0.035
		Calib, # NC=30	3.5	0.004	0.038	4.5	0.008	0.034
		Calib, # NC=50	4.0	0.004	0.037	3.0	0.007	0.033
Moderate ($\beta_Z = 0.4$, corr(X, Z)=0.2)	$N = 2500$	Raw	8.0	0.048	0.071	17.0	0.048	0.057
		Calib, # NC=10	4.5	0.005	0.077	4.5	0.008	0.065
		Calib, # NC=30	2.5	0.006	0.076	3.0	0.006	0.062
		Calib, # NC=50	2.0	0.007	0.076	4.0	0.009	0.063
	$N = 10\,000$	Raw	28.5	0.053	0.035	42.5	0.053	0.028
		Calib, # NC=10	3.0	0.010	0.045	3.0	0.011	0.037
		Calib, # NC=30	2.0	0.009	0.044	1.5	0.011	0.037
		Calib, # NC=50	1.5	0.011	0.046	1.5	0.013	0.040

Z is the variable that drives the missing data mechanism, and X is the exposure of interest.
NC, negative controls, SD, standard deviation; calib, calibrated *P*-value method.
[a]Correlation between NCs and Z between 0 and 0.3.
[b]After having applied the selection mechanism. The true model was simulated on datasets of size 2*N.
[c]The selection mechanism reduces the event rate compared with the simulated true datasets.
[d]The selection mechanism increases the event rate compared with the simulated true datasets.

Table 2. Rejection rate, average bias and average standard deviation (SD) over R = 200 replications using standard and calibrated (calib) *P*-value, for type I error simulation study and after having relaxed the exchangeability assumption[a]

Bias scenario	Sample size[b]	Method	Selection mechanism 1[c]			Selection mechanism 2[d]		
			Rejection rate (%)	Bias	SD	Rejection rate (%)	Bias	SD
Mild ($\beta_Z = 0.2$, corr(X, Z)=0.2)	$N = 2500$	Raw	6.0	0.031	0.068	7.0	0.020	0.061
		Calib, # NC=10	3.5	0.029	0.078	6.0	0.020	0.069
		Calib, # NC=30	3.0	0.026	0.078	5.5	0.017	0.069
		Calib, # NC=50	4.0	0.028	0.076	5.0	0.017	0.068
	$N = 10\,000$	Raw	8.5	0.022	0.034	10.5	0.022	0.030
		Calib, # NC=10	3.0	0.019	0.044	7.5	0.020	0.038
		Calib, # NC=30	2.5	0.020	0.044	3.5	0.020	0.040
		Calib, # NC=50	2.0	0.021	0.044	3.5	0.021	0.040
Moderate ($\beta_Z = 0.4$, corr(X, Z)=0.2)	$N = 2500$	Raw	9.5	0.047	0.071	11.0	0.048	0.057
		Calib, # NC=10	4.5	0.035	0.090	5.0	0.041	0.080
		Calib, # NC=30	5.5	0.048	0.091	5.0	0.051	0.076
		Calib, # NC=50	5.5	0.048	0.087	4.0	0.049	0.076
	$N = 10\,000$	Raw	35.0	0.054	0.035	41.5	0.050	0.028
		Calib, # NC=10	7.0	0.047	0.075	6.5	0.045	0.056
		Calib, # NC=30	6.0	0.056	0.068	5.0	0.053	0.053
		Calib, # NC=50	4.0	0.055	0.065	3.0	0.051	0.058

Z is the variable that drives the missing data mechanism, and X is the exposure of interest.
NC, negative controls. SD, standard deviation; calib, calibrated *P*-value method.
[a]Correlation between NCs and Z between -0.3 and 0.3.
[b]After having applied the selection mechanism. The true model was simulated on datasets of size 2*N.
[c]The selection mechanism reduces the event rate compared with the simulated true datasets.
[d]The selection mechanism increases the event rate compared with the simulated true datasets.

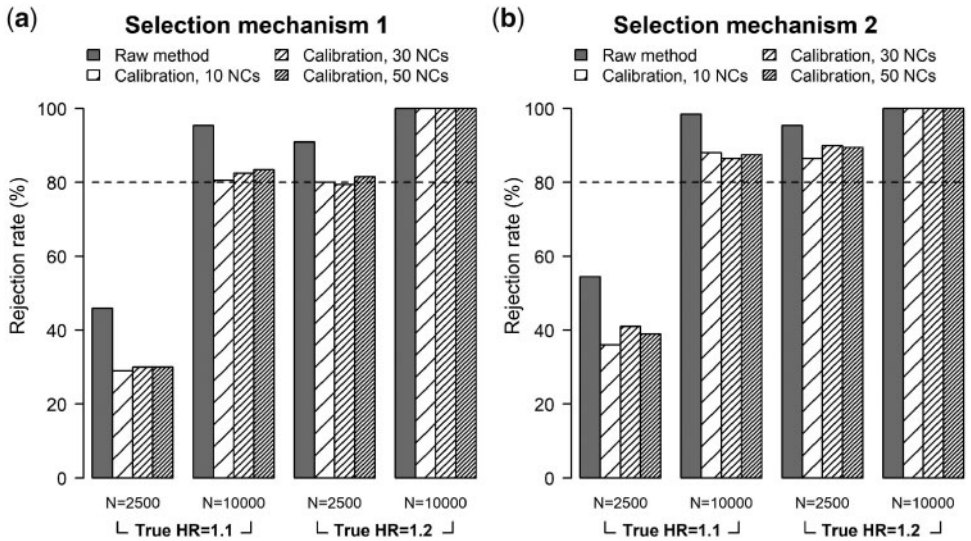

Figure 1. Rejection rates (%) for true hazard ratios of 1.1 and 1.2 in simulation studies when requirements for the exchangeability assumption are met and under a mild bias scenario. Exchangeability assumption: correlation between negative controls and Z between 0 and 0.3. Mild bias scenario: ($\beta_Z = 0.2$, corr(X, Z)=0.2). (a) The selection mechanism reduces the event rate compared with the simulated true datasets. (b) The selection mechanism increases the event rate compared with the simulated true datasets.

Rejection rates were slightly larger with selection mechanism 2 (panel b). We found similar findings in the moderate bias scenario, as well as when relaxing the exchangeability assumption (data not shown).

Real data application

The study population comprises 3464 men and 3512 women, with mean age 47.2 ± 12.2 years at baseline. Over 17 years of median follow-up, we observed $n = 470$ deaths

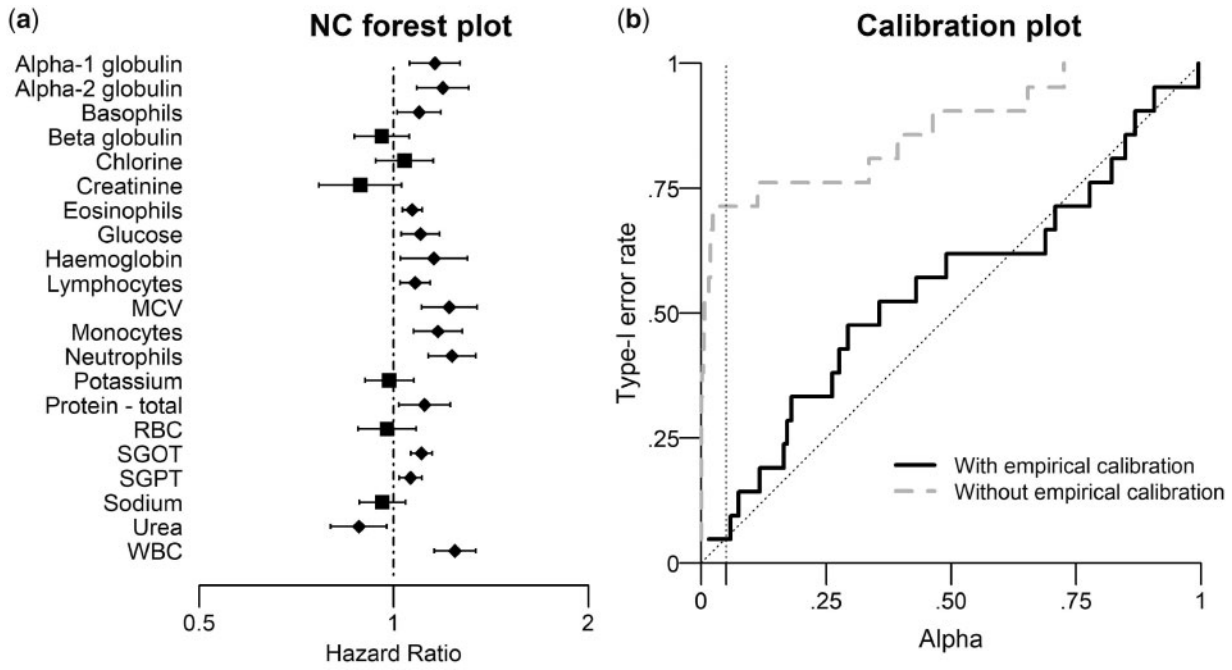

Figure 2. Age- and sex-adjusted hazard ratios (with 95% confidence intervals) of cancer mortality for the set of $n = 21$ negative controls (a), and calibration plot for the type I error rate (b), with and without the empirical calibration of the *P*-value method. Men and women, 25 to 74 years old, participants in the MONICA-Brianza and the PAMELA cohort studies ($n = 6979$). Hazard ratios are for 1-SD increase in the blood marker. MCV, mean corpuscular volume; RBC, red blood cells; SGOT, serum glutamic-oxaloacetic transaminase; SGPT, serum glutamic pyruvic transaminase; WBC, white blood cells. Diamonds indicate a $P < 0.05$.

due to cancer, corresponding to an event rate of 7%. Of these: 39% were due to gastrointestinal cancer; 25% to respiratory tract cancer (30% in men, 13% in women); and 7% to neoplasm of lymphatic and haematopoietic tissue. About 19% of cancer deaths in women were attributed to breast cancer.

The age- and sex-adjusted hazard ratios of cancer death for negative controls are presented in Figure 2a as well as in Supplementary Table 1, available as Supplementary data at *IJE* online. The association was statistically significant for 15 out of 21 NCs, corresponding to a type I error equal to 71%. The hazard ratios ranged from 0.88 to 1.24, being above 1 in 14/15 cases and indicating a potential positive bias. With the empirical calibration method, only one variable remained statistically significant, namely the one with hazard ratio below 1 (Supplementary Table 1), corresponding to a type I error of 4.8%, close to the nominal 5% level, as illustrated by the calibration plot (Figure 2b). Out of the 13 positive controls, six were associated with cancer mortality in our data (Supplementary Table 2), with hazard ratios ranging from 0.89 to 1.18; three of them had hazard ratios above 1. Only three out of the six associated positive controls remained so after the application of the empirical calibration method (Supplementary Table 2); all three had hazard ratios below 1.

Discussion

The analysis of 'big data' in observational research may rely on the observed data rather than on pre-specified hypotheses; therefore, the external validation and the replication of findings are advised.[10] The replicability of scientific results is a complicated process,[18] particularly when they come from complex data combining multiple and heterogeneous sources. In this framework, the control of type I error is of paramount importance to limit false-positive discoveries that will require further screening for biological plausibility and clinical relevance.[9,12] In our simulation study that reproduced the presence of selection bias mechanisms in the observed data, the type I error rate for the raw estimator in a standard Cox model was above the nominal 5% level in 15 out of the 16 simulation settings. For any given scenario characterizing the amount of bias in the data, the error rate increased as a function of the sample size: with $n = 10\,000$, a null association with a hazard ratio as low as 1.05 had a 42.5% chance of being statistically significant. The poor results suggest that the standard use of traditional inference tools is not adequate to control type I error in large databases prone to selection bias. These situations may become more and more common in the near future. Since the 1980s, population-based cohorts have experienced a drastic reduction in participation rates, which are currently close to 50%[19] (as in our simulation

plan) or even lower.[20] Compared with non-responders, participants have more favourable health behaviours[21] and lower mortality rates.[22] Recent studies using smartphones and M-research platforms are prone to selection bias due to the self-selection of study participants and the fast decline in retention rates.[13] Of the almost 49 000 individuals who consented to participate in a cardiovascular health study on lifestyles, less than 10% completed 7 days of motion data collection through the study app.[23] In another study, patients with asthma were more likely to use the study app and upload their data on days with more severe symptoms.[24] All these are examples of our selection bias mechanism 1[21–23] and 2.[24] respectively.

The empirical calibration of the *P*-value method has been introduced to control bias and type I error inflation in pharmacoepidemiology.[14–15] Our study has made a relevant contribution to the recent debate on its routine use in observational research.[16,25] To begin with, in survival analysis of data subject to selection bias from different mechanisms, by using simulations and a real study we showed that empirical calibration is able to keep type I error close to the 5% nominal level, independently of sample size and even in case of departure from the exchangeability assumption. Type I error preservation can also be achieved when negative controls are hard to find and their number is limited, as shown in simulations. Therefore, we considerably extended the range of real-life situations to which the method may be applied, with the aim of reducing the number of false findings. Second, researchers may want to use negative controls to adjust for selection bias in the point estimate of the exposure of interest. Our study suggests that the possibility of using the empirical calibration method to this extent depends on the choice of the set of negative controls. Sanderson *et al.* came to a partially similar conclusion for bias due to measurement error.[25] In the context of selection bias, the exchangeability assumption is not verifiable, since correlation between X and Z, as well as between the NCs and Z, cannot be estimated. However, as illustrated in Supplementary Figure 1 (available as Supplementary data at *IJE* online), under the exchangeability assumption the majority of negative controls are associated with the outcome in the same direction (panel a), compared with a mixed situation when the assumption is not met (panel b). In the real-life analysis, we were able to reproduce a forest plot suggestive of exchangeability assumption by selecting negative controls that all shared the same measurement source (e.g. blood specimen in our case). This seems a reasonable suggestion for deriving an appropriate set of negative controls capable of estimating selection bias for a given exposure. However, further studies are desirable to derive 'rules-of-thumb' to assist practitioners in the choice of negative controls. Third, Gruber *et al.* lamented a drastic increase in type II error when the empirical calibration of the *P*-value method was applied to the context of residual unmeasured confounding.[16] Our simulations confirmed these concerns for the combination of small sample size ($n = 2500$) and effect size [hazard ration (HR) = 1.1]. The analysis of positive controls in our real=life application with a sample size of about 7000 subjects suggests that when the association of interest and bias have opposite directions, the empirical calibration method may not have power issues. In data characterized by positive bias, a hazard ratio estimate below 1 is likely an underestimate of the true effect; the empirical calibration method confirmed 100% of these situations. Conversely, when the association of interest and bias have the same direction, the empirical calibration can provide guidance towards the minimum effect size needed to confirm the association as statistically significant, given the amount of bias present in the data.

Other alternative methods may be adapted to control type I error in the presence of large samples and selection bias, such as false-discovery rate (FDR),[26] FDR based on mixture models[27] or instrumental variables.[28] FDR is sensitive to sample size due to the reduced standard error.[14] Since we observed the same behaviour for the raw estimator in our simulations, we may expect similar consequences on the control of type I error. The identification of instrumental variables is difficult in real-world settings and it may require the use of additional data. Furthermore, the performance of the method relies on strong assumptions, including the strength of the association with the exposure variable.[28] In contrast, the empirical calibration method is not sensitive to sample size and does not require additional data. However, further studies are desirable to fully compare these alternative approaches, paying particular attention to mixture models.[27]

We acknowledge that this study has some limitations. Although we planned a quite comprehensive set of scenarios representative of the epidemiological literature, their number is still limited when compared with specific situations that may occur in observational studies. In addition, we assumed no other sources of bias, such as residual confounding or measurement error. In real data analysis, these may affect the exposure variable and the negative controls differently. However, not all the negative controls are required to have the same source of bias.[15] Finally, we only considered both continuous exposure and negative controls.

To conclude, in observational research with the contemporary presence of large sample sizes and selection bias, the standard application of traditional inference tools may fail to adequately control type I error. The empirical calibration of *P*-value is a robust method that can be applied

from available data to reduce the number of false-positive discoveries to be further screened for relevance and external validity.

Supplementary Data

Supplementary data are available at *IJE* online.

Acknowledgements

The authors would like to thank Professor Marco M Ferrario (Università degli Studi dell'Insubria) and Professor Giancarlo Cesana (Università di Milano-Bicocca), Principal Investigators of the MONICA-Brianza cohorts, for their permission to use their data and for their valuable comments on selection bias in observational research.

Author Contributions

G.V., A.Z. and P.Q. raised the research question, planned simulation settings and real-life analyses and interpreted simulations and real-life results. G.V. performed the statistical analyses for the simulation study and the real-life application, and drafted the manuscript. G.G. is the PI of one of the study cohorts and contributed to the interpretation of results from the real-life analysis. S.G. reviewed relevant literature to select negative controls for blood markers and cancer mortality, and contributed to the interpretation of the results from the real-life analysis. All authors reviewed manuscript drafts and added important contributions to study methods, results and discussion. G.V. is the guarantor for the paper.

Conflict of interest: None declared.

References

1. WHO. *Global Action Plan for the Prevention and Control of NCDs 2013–2020*. 2013. http://www.who.int/nmh/events/ncd_action_plan/en/ (29 August 2018, date last accessed).
2. WHO. The Challenge of Cardiovascular Disease–Quick Statistics. 2016. http://www.euro.who.int/en/health-topics/noncommunicable-diseases/cardiovascular-diseases/data-and-statistics. (29 August 2018, date last accessed).
3. Weintraub WS, Daniels SR, Burke LE *et al*. Value of primordial and primary prevention for cardiovascular disease: a policy statement from the American Heart Association. *Circulation* 2011;**124**:967–90.
4. Lloyd-Jones DM, Dyer AR, Wang R, Daviglus ML, Greenland P. Risk factor burden in middle age and lifetime risks for cardiovascular and non-cardiovascular death (Chicago Heart Association Detection Project in Industry). *Am J Cardiol* 2007;**99**:535–40.
5. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. *Am J Prev Med* 2016;**50**:398–401.
6. Diderichsen F, Hallqvist J, Whitehead M. Differential vulnerability and susceptibility: how to make use of recent development in our understanding of mediation and interaction to tackle health inequalities. *Int J Epidemiol* 2019;**48**:268–74.
7. Veronesi G, Tunstall-Pedoe H, Ferrario MM, for MORGAM Project. *et al*. Combined effect of educational status and cardiovascular risk factors on the incidence of coronary heart disease and stroke in European cohorts: Implications for prevention. *Eur J Prev Cardiol* 2017;**24**:437–45.
8. Ferrario MM, Roncaioli M, Veronesi G *et al*. Differing associations for sport versus occupational physical activity and cardiovascular risk. *Heart* 2018;**104**:1165–72.
9. Psaty BM, Dekkers OM, Cooper RS. Comparison of 2 treatment models: precision medicine and preventive medicine. *JAMA* 2018;**320**:751–52.
10. Dinov ID, Heavner B, Tang M *et al*. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One* 2016;**11**:e0157077.
11. Mooney SJ, Westreich DJ, El-Sayed AM. Epidemiology in the era of big data. *Epidemiology* 2015;**26**:390–94.
12. Bracken MB, Baker D, Cauley JA *et al*. New models for large prospective studies: is there a risk of throwing out the baby with the bathwater? *Am J Epidemiol* 2013;**177**:285–89.
13. Dorsey ER, Yvonne Chan YF, McConnell MV, Shaw SY, Trister AD, Friend SH. The use of smartphones for health research. *Acad Med* 2017;**92**:157–60.
14. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct P-values. *Stat Med* 2014;**33**:209–18.
15. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of P-values using observational data. *Stat Med* 2016;**35**:3883–88.
16. Gruber S, Tchetgen Tchetgen E. Limitations of empirical calibration of P-values using observational data. *Stat Med* 2016;**35**:3869–82.
17. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005;**24**:1713–23.
18. Nuzzo R. How scientists fool themselves - and how they can stop. *Nature* 2015;**526**:182–85.
19. Tolonen H, Ahonen S, Jentoft S, Kuulasmaa K, Heldal J. European Health Examination Pilot Project. Differences in participation rates and lessons learned about recruitment of participants: the European Health Examination Survey Pilot Project. *Scand J Public Health* 2015;**43**:212–19.
20. Doherty A, Jackson D, Hammerla N *et al*. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank study. *PLoS One* 2017;**12**:e0169649.
21. Christensen AI, Ekholm O, Gray L, Glümer C, Juel K. What is wrong with non-respondents? Alcohol-, drug- and smoking-related mortality and morbidity in a 12-year follow-up study of respondents and non-respondents in the Danish Health and Morbidity Survey. *Addiction* 2015;**110**:1505–12.
22. Harald K, Salomaa V, Jousilahti P, Koskinen S, Vartiainen E. Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972-92. *J Epidemiol Community Health* 2007;**61**:449–54.
23. McConnell MV, Shcherbina A, Pavlovic A *et al*. Feasibility of obtaining measures of lifestyle from a smartphone app: The MyHeart Counts Cardiovascular Health Study. *JAMA Cardiol* 2017;**2**:67–76.

24. Chan YY, Bot BM, Zweig M *et al*. The asthma mobile health study, smartphone data collected using ResearchKit. *Sci Data* 2018;**5**:180096.
25. Sanderson E, Macdonald-Wallis C, Davey Smith G. Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *Int J Epidemiol* 2018;**47**:587–96.
26. Strimmer K. A unified approach to false-discovery rate estimation. *BMC Bioinform* 2008;**9**:303.
27. Efron B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Institute of Mathematical Statistics Monographs*. Cambridge, UK: Cambridge University Press, 2010.
28. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014;**33**:2297–340.