Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

RESEARCH ARTICLE

A test for the correct specification of marginal structural models

Alioune Sall[1,2] | Karine Aubé[2] | Xavier Trudel[2,3] | Chantal Brisson[2,3] | Denis Talbot*[2,3]

[1]Département de mathématiques et de statistique, Université Laval, Québec, Canada

[2]Unité santé des populations et pratiques optimales en santé, CHU de Québec - Université Laval research center, Québec, Canada

[3]Département de médecine sociale et préventive, Université Laval, Québec, Canada

Correspondence
*Denis Talbot, Département de médecine sociale et préventive, Faculté de médecine, Université Laval, 1050, Avenue de la Médecine, Pavillon Ferdinand-Vandry, room 2454, Québec (Québec) G1V 0A6, Canada
Email: denis.talbot@fmed.ulaval.ca

Present Address
Present address

Abstract

Marginal structural models allow estimating the causal effect of a time-varying exposure on an outcome in the presence of time-dependent confounding. The parameters of marginal structural models can be estimated utilizing an inverse probability of treatment weight estimator under certain assumptions. One of these assumptions is that the proposed causal model relating the outcome to exposure history is correctly specified. However, in practice, the true model is unknown. We propose a test that employs the observed data to attempt validating the assumption that the model is correctly specified. The performance of the proposed test is investigated with a simulation study. We illustrate our approach by estimating the effect of repeated exposure to psychosocial stressors at work on ambulatory blood pressure in a large cohort of white-collar workers in Quebec City (Canada). Code examples in SAS and R are provided to facilitate the implementation of the test.

KEYWORDS:
Causal inference; marginal structural models; model specification

1 | INTRODUCTION

Marginal structural models (MSMs) are a class of causal models that are becoming increasingly popular for the estimation of causal effects when one deals with time varying exposures in the presence of time-dependent confounding.[1] The causal parameters are often estimated using an inverse probability of treatment weight (IPTW) estimator.[2] When using this estimator, the analyst must specify an outcome model that relates the outcome to the exposure history, as well as a weighting model relating the exposure at each time point to previous potential confounders. This estimator is unbiased under the assumptions of absence of i) unmeasured confounders and ii) misspecification of both the weighting model and the outcome model.

The specification of the structural outcome model, that links the outcome to the exposure history, has been the subject of much methodological work during the last few years. For instance, it has been observed that biased inferences may be obtained when the model considers only a part of the exposure history.[3] It has also been suggested that employing a marginal stabilized weight

[0]**Abbreviations:** MSMs, Marginal structural models; IPTW, Inverse probability of treatmeant weight; ABP, Ambulatory blood pressure

IPTW estimator may provide some robustness to misspecifications in this instance.[3,4] Platt et al have proposed an information criterion for MSMs (QIC_w) inspired by Akaike's information criterion to help in selecting a best fitting model among a set of candidate specifications for the structural outcome model.[5] However, the performance of the QIC_w is mitigated in so far as, in some cases, the QIC_w selects the true model with a relatively small probability in simulation studies.[5,6] Based on the QIC_w, Taguri and Matsuyuka have presented a corrected QIC_w ($cQIC_w$) which also compares different models based on the value of the criterion.[6] The ability of this $cQIC_w$ in selecting the correct specification was also variable in simulation studies.[6] More recently, Baba et al proposed a C_P criterion.[7] In simulations, C_P was observed to perform generally better than the $cQIC_w$.[7]

The QIC_w, $cQIC_w$ and C_P are all comparative criteria. That is, their role is to help in selecting the most appropriate specification when substantive knowledge does not allow to pinpoint a single specific functional form for relating the outcome to the exposure history. Opposingly, an absolute criterion would help in determining if a given specification is appropriate or not. Such an absolute criterion would be most pertinent when prior knowledge suggests that a particular specification is appropriate. An absolute criterion could also prove useful as a complement to comparative criteria.[7] In fact, if the true specification is not among the set of candidate models, comparative criteria will inevitably fail to identify the correct model. As such, it might be interesting to test if the specification chosen with comparative criteria is correct. Unfortunately, there currently exists no absolute criterion to validate a proposed specification of a marginal structural model.

In this paper, we thus introduce a Wald-type test that seeks to detect when the proposed specification of the structural outcome model is incorrect. The paper is structured as follows. In the second section, we introduce the concepts that underlie MSMs and present the notation. Section 3 introduces our test for the correct specification of the outcome model. In Section 4, we present a simulation study that investigates the empirical properties of our test. Section 5 presents an application of our new test in which we investigate the effect of psychosocial stressors at work on ambulatory systolic and diastolic blood pressure. Finally, in Section 6, we conclude with a discussion.

2 | MSMS AND NOTATION

Marginal structural models model the expectation of the potential outcome as a function of the exposure history.[1] We consider a follow-up study with T time points and n individuals sampled from a population. For individual i ($i = 1, \dots, n$), let Y_i be the outcome at the end of the follow-up (at time T), $X_{t,i}$ be the exposure at time t and $L_{t,i}$ be the other measured risk factors of Y at time t ($t = 1, \dots, T-1$). We define $\bar{X}_{t,i} = (X_1, X_2, \dots, X_t)$ as the individual i's exposure history with $\bar{L}_{t,i}$ defined similarly. As a notational shortcut, we denote $\bar{X}_{T-1,i}$ as $\bar{X}_i$. The potential outcome $Y^{\bar{x}}$ is defined as the value that Y would have taken if the exposure history had been $\bar{x}$. Thus, the marginal structural model can be represented as $E(Y^{\bar{x}}) = f(\bar{x})$ where $f(\bar{x})$ denotes a function of the exposure history. For instance, $f(\bar{x})$ could be: $f(\bar{x}) = \beta_0 + \beta_1 x_1 + \dots, +\beta_{T-1} x_{T-1}$ or $f(\bar{x}) = \beta_0 + \beta_1 \sum_{t=1}^{T-1} x_t$.

According to the first function, the outcome depends on all the exposure history additively, while the outcome depends linearly on the total amount of exposure in the latter. The causal parameters of interest in MSMs are contrasts between the counterfactual expectations of the outcome according to different exposure history, such as $E(Y^{\bar{x}})$ vs $E(Y^{\bar{x}'})$ for $\bar{x} \neq \bar{x}'$. The parameters of interest are thus a direct function of the parameters $\boldsymbol{\beta}$ of the MSM. For the sake of simplicity, we henceforth consider binary exposures, that is $X_{t,i} = 1$ if subject i is exposed at time t and $X_{t,i} = 0$ otherwise.

Parameters of MSMs are often estimated using IPTW estimators.[2] More precisely, the parameters of MSMs are estimated with the estimated parameters of the weighted linear model for $E(Y|\bar{X} = \bar{x}) = f^*(\bar{x})$, where $f^*(\bar{x})$ is assumed to be exactly the same function as $f(\bar{x})$ and the weights are given by the inverse probability of the observed exposure history conditionally on covariates and prior exposures. The weights create a pseudo-population in which, at each time point, exposed ($X_t = 1$) and unexposed ($X_t = 0$) subjects are similar to each other. Many types of weights can be considered, including standard weights (w), stabilized weights (sw) and marginal stabilized weights (swm). Formally, these different weights for subject i are defined as follows:

$$
\begin{aligned}
w_i &= \prod_{t=1}^{T-1} \frac{1}{P(X_t = x_{t,i}|\bar{X}_{t-1} = \bar{x}_{t-1,i}, \bar{L}_t = \bar{l}_{t,i})} \\
sw_i &= \prod_{t=1}^{T-1} \frac{P(X_t = x_{t,i}|\bar{X}_{t-1} = \bar{x}_{t-1,i})}{P(X_t = x_{t,i}|\bar{X}_{t-1} = \bar{x}_{t-1,i}, \bar{L}_t = \bar{l}_{t,i})} \\
swm_i &= \prod_{t=1}^{T-1} \frac{P(X_t = x_{t,i})}{P(X_t = x_{t,i}|\bar{X}_{t-1} = \bar{x}_{t-1,i}, \bar{L}_t = \bar{l}_{t,i})}.
\end{aligned}
$$

Remark that all weights share the same denominator, only the numerator of the weights varies according to the type of weights. In fact, it has been shown that the numerator of the weights might be any function of $\bar{X}_{T-1}$ without affecting the consistency of the estimator, that is, as the sample size increases to infinity, the estimate converges to the parameter in probability under the sequential exchangeability and positivity assumptions.[1] The sequential exchangeability assumption entails that

$$
Y^{\bar{x}} \coprod \bar{X}_t|\bar{X}_{t-1}, \bar{L}_t,
$$

where $\coprod$ denotes statistical independence, whereas the positivity assumption involves that

$$
P(X_t = x_t|\bar{X}_{t-1} = \bar{x}_{t-1}, \bar{L}_t = \bar{l}_t) > 0 \text{ for all } \bar{x}_t, \bar{l}_t \text{ where } P(\bar{L}_t = \bar{l}_t) \neq 0.
$$

The covariates $\bar{L}$ are thus chosen to satisfy these conditions.

Furthermore, it was shown that when a saturated MSM is fitted, the estimates of the causal parameters are the same regardless of the type of weights employed, but when an unsaturated model is considered, the estimates produced by the stabilized weights are different from those yielded by the standard weights.[1] The latter are more variable, but the difference is only due to sampling variability under the hypothesis that the model is correctly specified.[1] It is possible to further reduce the variance of the IPTW

estimator by also conditioning on baseline covariates ($\boldsymbol{L}_1$) in the numerator of the weights.[2] However, the outcome model of the MSM then needs to be modified to also condition on these covariates.[2] Such types of stabilized weights are thus not further considered in the current paper.

In addition to the sequential exchangeability assumption, another key assumption for unbiasedly estimating the causal contrasts $E(Y^{\bar{x}})$ vs $E(Y^{\bar{x}'})$ is the correct specification of the outcome model. That is, the function $f^*(\bar{x})$ used for relating the observed outcome to the exposure history needs to be the same as the function truly linking the counterfactual outcomes to the exposure history, $f(\bar{x})$, or to include it as a particular case. In the remainder of this paper, we will say that the outcome model is correctly specified if this is the case and misspecified otherwise. For example, if the true structural outcome model is $E(Y^{\bar{x}}) = \beta_0 + \beta_1 \sum_{t=1}^{T-1} x_t$, then the outcome models $E(Y|\bar{X}) = \beta_0 + \beta_1 \sum_{t=1}^{T-1} x_t$ and $E(Y|\bar{X}) = \beta_0 + \sum_{t=1}^{T-1} \beta_t x_t$ would both be correctly specified. In contrast, an incorrect specification could be $E(Y|\bar{X}) = \beta_0 + \beta_1 \mathbb{1}\left(\sum_{t=1}^{T} X_t > 0\right)$, where $\mathbb{1}$ is the usual indicator function that takes the value 1 if its argument is true and 0 otherwise.

3 | A TEST FOR THE CORRECT SPECIFICATION OF THE OUTCOME MODEL

MSMs provide, under certain assumptions, unbiased estimators of the causal effect of an exposure history. To ensure that the estimates obtained are unbiased, one should validate these assumptions. As previously mentioned, one of the assumptions of MSMs is that the outcome model is correctly specified. However, it is very difficult to know if the model is correctly specified or not since, so far, there is no formal way to test the validity of this assumption. Comparative criteria, such as $\boldsymbol{QIC}_w$, $\boldsymbol{cQIC}_w$ or $\boldsymbol{C}_p$ can be used to select a best-fitting specification among a set of candidates,[5,6,7] but there is no guarantee that the model chosen employing such criteria is correctly specified. For instance, a correct specification is impossible to find using these criteria if none of the model in the candidate set is correctly specified. We thus propose a statistical test that seeks to detect if the proposed specification of the outcome model is incorrect.

As was mentioned in the previous section, the parameters of MSMs can be estimated using various types of weights. When an unsaturated model is considered and the outcome model is correctly specified, the estimates may differ depending on the type of weights used, but the difference is only due to sampling variability.[1] However, if the model is misspecified, the estimators may not converge to the same values.[8] We utilize these properties of the IPTW estimator in devising a Wald-type test for the correct specification of the outcome model. That is, we want to test whether the differences observed between the estimates of the parameters using different weights can be attributed to random fluctuations. In such a case, the data are in line with the null hypothesis that the model is correctly specified. Otherwise, the data suggest that the model is incorrectly specified.

We provide details for comparing estimates obtained utilizing standard weights w and those produced by stabilized weights sw, but the same procedure can be used to compare estimates obtained with any two types of weights. The test is defined as

follows

$$H_0 : \boldsymbol{\beta}^{w} = \boldsymbol{\beta}^{sw} \text{ vs } H_1 : \boldsymbol{\beta}^{w} \neq \boldsymbol{\beta}^{sw},$$

where $\boldsymbol{\beta}^{k}$ denotes the vector of the parameters (without the intercept) estimated by using the weight $k = (w, sw)$. More precisely, $\boldsymbol{\beta}^{k}$ denotes the true parameter toward which the estimator based on weights k converges when sample size grows.

Under H_0, $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}^{w} - \hat{\boldsymbol{\beta}}^{sw}$ has a normal limit distribution with mean $\mathbf{0}$, thus $D = \hat{\boldsymbol{\delta}}\widehat{Var}(\hat{\boldsymbol{\delta}})^{-1}\hat{\boldsymbol{\delta}}' \sim \chi^2_m$ asymptotically, where m denotes the dimension of $\boldsymbol{\delta}$. Indeed, D is a sum of the square of m standard normal variables. We thus propose using D as a statistic for testing if the outcome model is correctly specified. The limit normal distribution of $\hat{\boldsymbol{\delta}}$ follows from the fact that the IPTW estimator of the parameters of MSMs is a regular asymptotically linear estimator.[9] The proof is provided in the Appendix.

To calculate our test statistic, we need to estimate the covariance matrix of $\hat{\boldsymbol{\delta}}$. We propose two alternative estimators. As is typical for Wald-type tests, for both estimators, the covariance matrix is estimated at the estimated parameter values and not at the values of the parameters under the null. We first propose a non-parametric bootstrap estimator. More precisely, n observations are sampled with replacement from the original data, B times. In each sample, the weights sw and w are first estimated. Then the parameters of the MSM are estimated employing the weights w and sw, and $\hat{\boldsymbol{\delta}}$ is computed. The covariance matrix is finally estimated by computing the empirical covariance matrix based on the B bootstrap estimates of $\boldsymbol{\delta}$. As a second estimator, we consider using the so-called sandwich estimator of a generalized estimating equation (GEE) regression. To implement this option, we first build an augmented dataset where each subject appears twice, once with weights w and once with weights sw, with an additional categorical variable indicating the type of weights. Using the GEE estimator, we then estimate the parameters of a weighted regression model for the observed outcome according to the observed exposure history, including a term for the type of weights and interaction terms between the observed exposure history and the type of weights. For example, if the postulated structural outcome model is $E(Y^{\bar{x}}) = \beta_0 + \beta_1 \sum_{t=1}^{T-1} x_t$, we would fit the following weighted regression model on the augmented dataset $E(Y|\bar{X}) = \beta_0 + \beta_1 \sum_{t=1}^{T-1} x_t + \beta_2 type + \delta type \times \sum_{t=1}^{T} x_t$. The parameters associated with the interaction terms in this model encode the difference in the estimates obtained using both types of weights. As such, the estimated covariance matrix of the estimator of these parameters can be used for performing our test. An independence working covariance matrix is chosen for the GEE estimator to mimic the situation where the IPTW estimators based on both weights are used independently.

We note that it is also possible to devise a test based on comparing the estimates from all three types of weights. The hypotheses are then

$$H_0 : \boldsymbol{\beta}^{w} = \boldsymbol{\beta}^{sw} = \boldsymbol{\beta}^{swm} \text{ vs } H_1 : \boldsymbol{\beta}^{w} \neq \boldsymbol{\beta}^{sw} \text{ and/or } \boldsymbol{\beta}^{w} \neq \boldsymbol{\beta}^{swm} \text{ and/or } \boldsymbol{\beta}^{sw} \neq \boldsymbol{\beta}^{swm}.$$

The test statistic has the same form as previously described, with $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}^{w} - \hat{\boldsymbol{\beta}}^{sw}, \hat{\boldsymbol{\beta}}^{w} - \hat{\boldsymbol{\beta}}^{swm})^{\top}$. Remark that including $\hat{\boldsymbol{\beta}}^{sw} - \hat{\boldsymbol{\beta}}^{swm}$ in $\hat{\boldsymbol{\delta}}$ would be redundant since $\hat{\boldsymbol{\beta}}^{sw} - \hat{\boldsymbol{\beta}}^{swm} = (\hat{\boldsymbol{\beta}}^{w} - \hat{\boldsymbol{\beta}}^{swm}) - (\hat{\boldsymbol{\beta}}^{w} - \hat{\boldsymbol{\beta}}^{sw})$. Both estimators of the covariance matrix are obtained

as previously described, but creating an augmented dataset with three rows (one for each type of weights) instead of two for computing the sandwich estimator.

4 | SIMULATION STUDY

4.1 | Description of the simulation study

In this section, we detail our simulation study which includes four scenarios. This simulation study aims to evaluate the capacity of our test described in Section 3 to detect a misspecified model with different sample sizes. We compare the performance of the test according to different combinations of weights and estimator of the covariance matrix. We also investigate if truncating the weights at their $99.5th$ percentile impacts the performance of our test. The estimators of the causal parameter based on truncated weights have been observed to be less variable than those based on untruncated weights.[10] As such, we had initially hypothesized that truncating weights might improve performance.

Scenario 1

This scenario is taken from Talbot et al.[3] The relationships between the variables are as follows

$$
\begin{aligned}
&L_1 \sim \mathcal{N}(0,1),\\
&P(X_1 = 1) = expit(0.5L_1)\\
&L_1' = X_1 + L_1 + \varepsilon_{L_1'}\\
&L_2 = 0.5X_1 + \varepsilon_{L_2}\\
&P(X_2 = 1) = expit(0.5X_1 + 0.5L_1' + 0.5L_2)\\
&Y = X_2 + 0.5L_1 + L_2 + \varepsilon_Y,
\end{aligned}
$$

where $expit(a) = \frac{e^a}{1+e^a}$, $\varepsilon_{L_1}, \varepsilon_{L_1'}, \varepsilon_{L_2}, \varepsilon_Y$ are $\mathcal{N}(0,1)$ independent random variables. In this scenario, the standard, stabilized and marginal stabilized weights are defined as

$$
\begin{aligned}
w_i &= \frac{1}{P(X_1 = x_{1,i}|L_1 = l_{1,i})} \times \frac{1}{P(X_2 = x_{2,i}|X_1 = x_{1,i}, L_1 = l_{1,i}, L_1' = l_{1,i}', L_2 = l_{2,i})}\\
sw_i &= \frac{P(X_1 = x_{1,i})}{P(X_1 = x_{1,i}|L_1 = l_{1,i})} \times \frac{P(X_2 = x_{2,i}|X_1 = x_{1,i})}{P(X_2 = x_{2,i}|X_1 = x_{1,i}, L_1 = l_{1,i}, L_1' = l_{1,i}', L_2 = L_{2,i})}\\
swm_i &= \frac{P(X_1 = x_{1,i})}{P(X_1 = x_{1,i}|L_1 = l_{1,i})} \times \frac{P(X_2 = x_{2,i})}{P(X_2 = x_{2,i}|X_1 = x_{1,i}, L_1 = l_{1,i}, L_1' = l_{1,i}', L_2 = l_{2,i})}.
\end{aligned}
$$

Scenario 2

This scenario is inspired by the ones considered by Platt et al.[5]:

$$
\begin{aligned}
L_1 &\sim \mathcal{N}(10, 1)\\
P(X_1 = 1) &= expit(-2.6 + 0.25L_1)\\
L_2 &\sim \mathcal{N}(L_1 + X_1, 1)\\
P(X_2 = 1) &= expit(-2.6 + 0.25L_2 + 0.1X_1)\\
L_3 &\sim \mathcal{N}(L_2, 1)\\
P(X_3 = 1) &= expit(-2.6 + 0.25L_3 + 0.1X_2)\\
L_4 &\sim \mathcal{N}(L_3 + 2X_3, 1)\\
P(X_4 = 1) &= expit(-2.6 + 0.25L_4 + 0.1X_3)\\
Y &\sim \mathcal{N}(L_4 + 3X_4, 1),
\end{aligned}
$$

The different type of weights are defined as follows

$$
\begin{aligned}
w_i = &\frac{1}{P(X_1 = x_{1,i}|L_1 = l_{1,i})} \times \frac{1}{P(X_2 = x_{2,i}|X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i})}\\
&\times \frac{1}{P(X_3 = x_{3,i}|X_2 = x_{2,i}, X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i}, L_3 = l_{3,i})}\\
&\times \frac{1}{P(X_4 = x_{4,i}|X_3 = x_{3,i}, X_2 = x_{2,i}, X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i}, L_3 = l_{3,i}, L_4 = l_{4,i})}\\
sw_i = &\frac{P(X_1 = x_{1,i})}{P(X_1 = x_{1,i}|L_1 = l_{1,i})} \times \frac{P(X_2 = x_{2,i}|X_1 = x_{1,i})}{P(X_2 = x_{2,i}|X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i})}\\
&\times \frac{P(X_3 = x_{3,i}|X_2 = x_{2,i}, X_1 = x_{1,i})}{P(X_3 = x_{3,i}|X_2 = x_{2,i}, X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i}, L_3 = l_{3,i})}\\
&\times \frac{P(X_4 = x_{4,i}|X_3 = x_{3,i}, X_2 = x_{2,i}, X_1 = x_{1,i})}{P(X_4 = x_{4,i}|X_3 = x_{3,i}, X_2 = x_{2,i}, X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i}, L_3 = l_{3,i}, L_4 = l_{4,i})}\\
swm_i = &\frac{P(X_1 = x_{1,i})}{P(X_1 = x_{1,i}|L_1 = l_{1,i})} \times \frac{P(X_2 = x_{2,i})}{P(X_2 = x_{2,i}|X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i})}\\
&\times \frac{P(X_3 = x_{3,i})}{P(X_3 = x_{3,i}|X_2 = x_{2,i}, X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i}, L_3 = l_{3,i})}\\
&\times \frac{P(X_4 = x_{4,i})}{P(X_4 = x_{4,i}|X_3 = x_{3,i}, X_2 = x_{2,i}, X_1 = x_{1,i}, L_1 = l_{1,i}, L_2 = l_{2,i}, L_3 = l_{3,i}, L_4 = l_{4,i})}.
\end{aligned}
$$

Scenario 3

Scenario 3 is the same as Scenario 2, but with different parameter values. The variables are generated as follows

$$
L_1 \sim \mathcal{N}(10, 1)
$$

$$P(X_1 = 1) = expit(-2.6 + 0.25L_1)$$

$$L_2 \sim \mathcal{N}(L_1, 1)$$

$$P(X_2 = 1) = expit(-2.6 + 0.25L_2 + 0.1X_1)$$

$$L_3 \sim \mathcal{N}(L_2, 1)$$

$$P(X_3 = 1) = expit(-2.6 + 0.25L_3 + 0.1X_2)$$

$$L_4 \sim \mathcal{N}(L_3, 1)$$

$$P(X_4 = 1) = expit(-2.6 + 0.25L_4 + 0.1X_3)$$

$$Y \sim \mathcal{N}(L_4 + X_4, 1).$$

The weights are defined as in Scenario 2.

Scenario 4

The objective of Scenario 4 is to investigate the performance of the test when the weighting models are incorrect. The equations used to generate the data are similar to those of Scenario 2, but include non linear terms.

$$L_1 \sim \mathcal{N}(10, 1)$$

$$P(X_1 = 1) = expit(-2.6 + 0.25L_1 - 0.1(L_1 - 10)^2)$$

$$L_2 \sim \mathcal{N}(L_1 + X_1, 1)$$

$$P(X_2 = 1) = expit(-2.6 + 0.25L_2 + 0.1X_1 - 0.1(L_2 - 10)^2 - 0.05X_1(L_2 - 10))$$

$$L_3 \sim \mathcal{N}(L_2, 1)$$

$$P(X_3 = 1) = expit(-2.6 + 0.25L_3 + 0.1X_2 - 0.1(L_3 - 10)^2 - 0.05X_2(L_3 - 10))$$

$$L_4 \sim \mathcal{N}(L_3 + 2X_3, 1)$$

$$P(X_4 = 1) = expit(-2.6 + 0.25L_4 + 0.1X_3 - 0.1(L_4 - 10)^2 - 0.05X_3(L_4 - 10))$$

$$Y \sim \mathcal{N}(L_4 + 3X_4, 1)$$

The weights are defined as in Scenario 2.

4.2 | Analysis of the simulation

For each of the scenarios described above, we generated 1000 datasets of size $n = 200$, $n = 500$, $n = 1000$ and $n = 5000$. We considered the four following specifications of the MSM

$$E(Y^{\bar{x}}) = \beta_0 + \beta_1 \sum_{t=1}^{T-1} X_t$$

$$E(Y^{\bar{x}}) = \beta_0 + \beta_T X_{T-1}$$

$$E(Y^{\bar{x}}) = \beta_0 + \mathbb{1}\left(\sum_{t=1}^{T-1} X_t > 0\right)$$

$$E(Y^{\bar{x}}) = \beta_0 + \sum_{t=1}^{T-1} \beta_t X_t,$$

where $T = 3$ in Scenario 1 and $T = 5$ in Scenarios 2, 3 and 4. In the following, these models will be designated respectively as the cumulative, the current, the indicator and the full model. They are examples of specifications encountered in the literature.[1,11,12,13]

For each model, we estimated the MSMs' parameters using the three different forms of weights described above and the probabilities forming these weights were estimated employing main effects logistic regression models. As such, the weighting models are misspecified for Scenario 4. We then applied 16 variations of our test described in Section 3. These variations are obtained by considering all possible comparisons of the weighted estimates (w vs sw, w vs swm, swm vs sw, w vs sw vs swm), estimator of the covariance matrix (bootstrap or sandwich) and untruncated or truncated weights. The bootstrap was performed with $B = 1000$ resamples using the function `boot` from the R package `boot`.[14] Then, for each combination of scenario, sample size, MSM specification and variation of the correct specification test, we calculated the proportion of p-values smaller than 5% across the 1000 generated datasets; that is, the proportion of the time the test leads to reject the null hypothesis that the model is correctly specified. When the MSM is correctly specified, this proportion should be around 5%. We have also computed the proportion of the time each specification was chosen according to QIC_w and $cQIC_w$. Note that in the case of longitudinal MSMs, $cQIC_w$ is equivalent to C_P if the weights are treated as known.[7]

As in many previous simulations of MSMs, the data generating equations were constructed to avoid non-collapsibility so that the correct specification of the outcome model can be determined easily. To capture the correct specification of a model, we proceeded by substitution. For example, in Scenario 1 we have:

$$\begin{aligned} Y &= X_2 + 0.5L_1' + L_2 + \varepsilon_Y \\ &= X_2 + 0.5(X_1 + L_1 + \varepsilon_{L_1'} + 0.5X_1 + \varepsilon_{L_2} + \varepsilon_Y \\ &= X_2 + X_1 + \varepsilon_Y^* \end{aligned}$$

where $\varepsilon_Y^* = 0.5L_1 + 0.5\varepsilon_{L_1'} + \varepsilon_{L_2} + \varepsilon_Y$. In this case, both the full and cumulative models are thus correctly specified. For Scenarios 2 and 4, only the full model is correctly specified. The current and the full models are correctly specified in Scenario 3.

4.3 | Simulation results

Contrary to our initial expectation, our proposed Wald test performed poorly when truncated weights were utilized. Indeed, rejection rates much higher than the nominal 5% rate were observed in many situations. To lighten the text, those results are thus presented as a supplementary online material and only results based on untruncacted weights are discussed below.

The results for Scenario 1 are summarized in Table 1 . The null hypothesis is rejected less than or close to 5% of the time for both the full and the cumulative models, which are the correctly specified models. For the indicator model, the rejection rates are high for most tests, even with a sample size of $n = 200$. Only the tests based on comparing w and swm offer a relatively poor performance when the sample size is 200. The tests using the sandwich estimator of the covariance matrix outperform their counterpart based on the bootstrap estimator. In particular the tests swm vs sw and w vs sw vs swm using the sandwich estimator have the best performance. Similar results are observed for the current model. The main difference is that the tests comparing w and swm yield rejection rates smaller or equal to 5% for all sample sizes. This situation can be explained by the fact that unbiased estimation of the MSM's parameter can be obtained by using either standard weights or marginal stabilized weights in such a misspecified structural model.[3] We have investigated the performance of tests using the sandwich estimator when an exchangeable working covariance structure is considered instead of the independence structure in this scenario. This resulted in poorer performance of the tests; rejection rates were sometimes much larger than 5% for correctly specified models and were generally lower than those reported in Table 1 for misspecified models (results not presented). Results of information criteria for all scenarios are presented in Table 5 . In Scenario 1, information criteria almost always select correctly specified models. $cQIC_w/C_p$ could be argued to perform better since it has a larger propensity to select the simplest of the two correctly specified model (the cumulative model). However, it also selects slightly more often misspecified models for sample sizes of $n = 200$ and $n = 500$.

The results pertaining to Scenario 2 are presented in Table 2 . For the misspecified cumulative model, all tests reject the null hypothesis that the model is correctly specified in less than or close to 5% of the replications for a sample size of $n = 200$ when the bootstrap estimator of the covariance matrix is used. Rejection rates are moderate when the sandwich estimator is used. Similarly, rejection rates are small or moderate for sample sizes of $n = 500$ or $n = 1000$, whichever estimator is used for the covariance matrix. At $n = 5000$, large rejection rates are obtained, except for tests comparing weights w and swm. As expected, the rejection rates are inferior to 5% for all sample sizes and tests for the full model, which is correctly specified when the bootstrap estimator is used. When the sandwich estimator is used, rejection rates tend to slightly, but noticeably, exceed the 5% threshold at the smaller sample sizes. The deviation from the nominal level is most important for the test w vs sw vs swm. For

the indicator model, at $n = 200$, the power to detect that the model is misspecified is small to moderate when using the bootstrap estimator and is large when the sandwich estimator is used. For larger sample sizes, the rejection rates are large for all tests, whichever covariance matrix estimator is used. The most powerful test is the one based on w vs sw vs swm using the sandwich estimator (93.9% rejection rate). Similar results are obtained for the current model. The main difference is, as in Scenario 1, that the rejection rates are lower or close to 5% for the tests comparing standard weights and marginal stabilized weights. Once again, both standard weights and marginal stabilized weights produce unbiased estimators of the MSM's parameter despite the model being misspecified. Comparative criteria QIC_w and $cQIC_w$ always select the correctly specified full model at all sample sizes.

Table 3 presents the results of Scenario 3. Concerning the cumulative model, which is misspecified, the proportion of rejection is smaller than or close to 5% for all tests and all sample sizes when considering the bootstrap estimator. When the sandwich estimator is used, the rejection rates are between 10% and 15%, except for the test w vs swm which has a rejection rate close to 5%. For the full model, as expected, the proportion of rejection of the null hypothesis that the model is correctly specified is always under or close to 5% when using the bootstrap estimator. When the sandwich estimator is used, the rejection rate tends to slightly exceed the nominal rate (up to 10.3% instead of 5%), especially at the smaller sample sizes. For the misspecified indicator model, rejection rates are inferior to 5% for all tests when using the bootstrap estimator at $n = 200$, $n = 500$ and $n = 1000$. At $n = 5000$ the rejection rates are small to moderate. When using the sandwich estimator, rejection rates are small to moderate for sample sizes $n = 200$, $n = 500$ and $n = 1000$, and are generally large for a sample size of $n = 5000$. Considering the current model, the rejection rate is under or close to 5% for all tests based, as expected since this model is correctly specified. The information criteria almost always select correctly specified models. However, they tend to favor the full model over the simpler current model. As in Scenario 1, $cQIC_w/C_p$ has a greater tendency to select the simplest of the two correctly specified model, but also selects more often misspecified models at $n = 200$.

Results of Scenario 4 are presented in Table 4 . Recall that this scenario is almost the same as Scenario 2, but is meant to investigate the performance of our test under misspecification of the weighting model. First, most versions of the test tend to reject correctly specified models much more often than the 5% nominal rate. Only the test comparing swm and sw using the bootstrap estimator has acceptable rejection rates for correctly specified models, although the rejection rate gets close to 10% in some instances at $n = 5000$. Rejection rates of incorrectly specified models follow a pattern similar to those observed in Scenario 2, but are generally larger. As in Scenario 2, information criteria always select the correctly specified model.

The results of Scenario 1, 2 and 3 indicate that the rejection rates of correctly specified models for tests using the bootstrap estimator are often much lower than the expected nominal rate of 5%, especially for small sample sizes. Tests based on the sandwich estimator seem to behave more appropriately, but rejection rates that are noticeably lower or larger than 5% are also observed for smaller sample sizes. A possible explanation for this phenomenon is that the distribution of the difference in estimates would not be normally distributed for small n, and the test would thus not follow the expected chi-square distribution

under the null. Normal QQ-plots of the differences in the estimated parameters between standard weights and stabilized weights for the full model in Scenario 2 are depicted in Figure 1 . We can notice deviations from the normal distribution in the tails which vanish as sample size increases. Similar results are obtained for the other estimates of correctly specified models (not presented). Departure from the normality distribution thus at least partly explains why rejection rates of correctly specified models do not correspond to their nominal rate. Another plausible explanation would be the incorrect estimation of the covariance matrix. For instance, it is known that the sandwich estimator may not work well for small sample sizes (for example, see reference [15]). The plausibility of this explanation is reinforced by the fact that variations in rejection rates are observed between tests using the bootstrap estimator of the covariance matrix and their counterpart based on the sandwich estimator.

TABLE 1 Rejection rates of tests at $\alpha = 0.05$ for the correct specification of the outcome model by postulated specification for Scenario 1, in percentage

		Bootstrap variance estimator				Sandwich variance estimator			
Specification	Test	n=200	n=500	n=1000	n=5000	n=200	n=500	n=1000	n=5000
Cumulative	w vs sw	0.4	1.0	2.6	4.3	2.3	4.1	4.6	4.2
	w vs swm	0.7	1.4	2.8	4.4	5.2	5.8	5.3	5.2
	swm vs sw	0.3	1.1	3.0	4.9	2.6	3.9	4.7	5.5
	w vs sw vs swm	0.1	0.2	0.5	2.3	1.9	1.8	2.0	1.9
Full	w vs sw	0.4	0.9	0.7	1.2	2.5	2.3	1.4	1.2
	w vs swm	0.4	1.0	0.8	1.3	2.7	1.9	1.3	1.5
	swm vs sw	0.1	0.3	0.6	1.3	2.6	2.8	2.2	1.8
	w vs sw vs swm	0.0	0.1	0.0	1.0	1.4	0.6	0.2	0.6
Indicator	w vs sw	64.1	90.5	98.9	100	70.8	89.8	98.6	100
	w vs swm	17.9	42.4	66.2	99.8	35.2	49.8	68.7	99.7
	swm vs sw	71.1	97.6	99.8	100	85.9	97.6	99.8	100
	w vs sw vs swm	57.0	97.5	100	100	86.5	98.9	100	100
Current	w vs sw	64.1	98.5	99.9	100	84.0	98.9	99.9	100
	w vs swm	0.0	0.0	0.0	0.0	6.7	6.8	5.2	5.1
	swm vs sw	70.4	98.8	99.9	100	86.8	99.2	99.9	100
	w vs sw vs swm	45.9	97.9	99.8	100	90.4	99.7	100	100

w = standard weights, sw = stabilized weights, swm = marginal stabilized weights. Models in bold are correctly specified.

5 | APPLICATION

In this section, data from a 5-year longitudinal cohort study of white-collar workers in Quebec City (Canada) are used to illustrate the method that we developed. More precisely, we applied our test to MSMs aimed at estimating the causal effect of repeated exposure to psychosocial stressors at work on ambulatory blood pressure (ABP) at the end of follow-up.

TABLE 2 Rejection rates of tests at $\alpha = 0.05$ for the correct specification of the outcome model by postulated specification for Scenario 2

		Bootstrap variance estimator				Sandwich variance estimator			
Specification	Test	n=200	n=500	n=1000	n=5000	n=200	n=500	n=1000	n=5000
Cumulative	w vs sw	1.8	7.4	13.8	59.0	15.1	21.6	29.6	74.9
	w vs swm	0.5	0.9	4.4	22.8	5.9	6.6	9.4	24.4
	swm vs sw	3.5	12.8	25.4	85.1	22.3	36.2	48.9	93.0
	w vs sw vs swm	1.3	8.1	18.8	79.9	19.6	29.6	43.6	91.3
Full	w vs sw	0.3	0.0	0.9	2.8	7.4	6.1	5.9	5.1
	w vs swm	0.0	0.0	0.6	2.2	7.3	5.3	5.2	5.2
	swm vs sw	0.1	0.0	1.0	2.1	7.4	5.6	5.6	4.4
	w vs sw vs swm	0.0	0.0	0.4	1.5	9.9	6.5	6.0	5.4
Indicator	w vs sw	52.4	95.0	99.9	100	90.5	99.7	100	100
	w vs swm	33.3	72.1	95.4	100	87.4	98.2	99.9	100
	swm vs sw	20.6	83.1	99.5	100	59.6	93.7	99.9	100
	w vs sw vs swm	38.4	94.4	100	100	93.9	99.9	100	100
Current	w vs sw	45.4	96.6	100	100	89.5	99.5	100	100
	w vs swm	0.0	0.2	0.3	2.2	6.2	5.8	5.4	4.8
	swm vs sw	56.1	97.7	100	100	93.2	99.9	100	100
	w vs sw vs swm	35.2	93.9	100	100	91.7	99.7	100	100

w = standard weights, sw = stabilized weights, swm = marginal stabilized weights. Models in bold are correctly specified.

TABLE 3 Rejection rates of tests at $\alpha = 0.05$ for the correct specification of the outcome model by postulated specification for Scenario 3

		Bootstrap variance estimator				Sandwich variance estimator			
Specification	Test	n=200	n=500	n=1000	n=5000	n=200	n=500	n=1000	n=5000
Cumulative	w vs sw	0.5	0.8	2.0	5.3	10.0	11.6	11.6	12.3
	w vs swm	0.0	0.2	0.2	1.6	6.2	6.0	5.5	5.4
	swm vs sw	0.9	1.2	2.1	5.4	12.3	15.0	16.0	15.0
	w vs sw vs swm	0.2	0.1	0.5	3.5	11.8	13.1	14.9	13.7
Full	w vs sw	0.0	0.0	0.2	0.2	8.7	6.5	6.1	5.0
	w vs swm	0.0	0.0	0.0	0.0	7.5	6.4	5.9	5.5
	swm vs sw	0.0	0.0	0.0	0.2	6.7	4.7	4.7	3.9
	w vs sw vs swm	0.0	0.0	0.0	0.0	10.3	6.7	7.0	5.8
Indicator	w vs sw	0.3	1.0	1.5	4.0	12.9	18.8	26.0	37.4
	w vs swm	1.0	3.0	4.4	19.2	22.8	39.2	49.9	73.1
	swm vs sw	0.1	0.3	2.4	52.7	7.8	11.2	18.3	64.9
	w vs sw vs swm	0.4	1.3	3.9	49.4	23.7	37.6	50.7	86.0
Current	w vs sw	0.0	0.1	0.6	1.7	3.6	4.0	3.7	2.9
	w vs swm	0.1	0.1	0.3	2.0	6.1	6.4	5.6	5.9
	swm vs sw	0.0	0.1	0.5	2.5	3.5	3.7	4.4	2.6
	w vs sw vs swm	0.0	0.0	0.2	1.3	5.2	5.0	4.6	3.8

w = standard weights, sw = stabilized weights, swm = marginal stabilized weights. Models in bold are correctly specified.

TABLE 4 Rejection rates of tests at $\alpha = 0.05$ for the correct specification of the outcome model by postulated specification for Scenario 4

		Bootstrap variance estimator				Sandwich variance estimator			
Specification	Test	n=200	n=500	n=1000	n=5000	n=200	n=500	n=1000	n=5000
Cumulative	w vs sw	5.0	16.6	36.1	97.4	22.7	36.2	55.3	99.4
	w vs swm	15.7	57.9	90.2	100	29.5	62.7	90.4	100
	swm vs sw	2.4	10.4	19.3	85.2	39.2	61.9	78.6	99.3
	w vs sw vs swm	9.7	49.8	86.8	100	50.6	82.0	96.7	100
Full	w vs sw	2.2	41.6	94.6	100	32.9	74.3	98.2	100
	w vs swm	5.0	52.0	96.4	100	37.2	76.3	97.9	100
	swm vs sw	0.0	0.0	0.1	1.8	8.7	13.0	21.1	50.8
	w vs sw vs swm	0.2	15.1	83.5	100	30.4	66.3	95.6	100
Indicator	w vs sw	71.3	99.0	100	100	96.8	99.9	100	100
	w vs swm	91.3	99.9	100	100	99.4	100	100	100
	swm vs sw	1.3	9.2	38.7	99.6	22.2	55.0	85.8	100
	w vs sw vs swm	81.4	99.8	100	100	99.1	99.9	100	100
Current	w vs sw	5.5	21.2	42.4	97.9	41.7	64.9	84.4	100
	w vs swm	8.4	50.6	92.6	100	38.0	74.1	96.6	100
	swm vs sw	1.0	4.7	4.6	9.8	43.1	62.3	65.7	75.6
	w vs sw vs swm	4.2	36.9	85.1	100	60.1	88.0	98.5	100

w = standard weights, sw = stabilized weights, swm = marginal stabilized weights. Models in bold are correctly specified.

The aim of the original study was to examine the effect of psychosocial stressors at work on health. Workers completed self-administered questionnaires on work characteristics and wore oscillometric devices to assess ABP at baseline (2000-2004), 3-year follow-up (2004-2006) and 5-year follow-up (2006-2009). Systolic and diastolic ABP (in mm Hg) were assessed using validated protocols (see Trudel et al[16] for more details). ABP measures from the 5-year follow-up were used. Psychosocial stressors at work, as defined according to the Effort-reward imbalance (ERI) model, were assessed at all three time-points by a validated self-report instrument.[17] As recommended, a ratio of efforts to rewards greater than 1 was dichotomized as exposure to ERI.[17] Different repeated exposure patterns were merged together when they were expected to have a similar effect on ABP. Workers were thus classified into five categories of repeated ERI exposure: never exposed (0, 0, 0), intermittent exposure (0, 1, 0 or 1, 0, 1), exposure that ceased over follow-up (1, 0, 0 or 1, 1, 0), exposure onset (0, 1, 1 or 0, 0, 1) or chronic exposure (1, 1, 1).

The sample used included individuals that participated at all three follow-ups, had ABP measurements at the last follow-up, had adequately completed the ERI measurement at all three follow-ups, had no missing values on covariates, were not pregnant at the last follow-up and worked at least 21 hours per week (in order to prevent potential misclassification due to insufficient exposure). The final sample was composed of 1576 workers with 925 women and 651 men. Potential confounders included gender, age at baseline, education (less than college, college completed, university completed), family history of cardiovascular disease, smoking status (current smoker or not), alcohol consumption ($<$ 1 drink/week, 1-5 drinks/week, $\geq$ 6 drinks/week),

TABLE 5 Selection probabilities of information criteria by postulated specification and scenario, in percentage

Scenario	Specification	Criterion	n=200	n=500	n=1000	n=5000
Scenario 1	**Cumulative**	QIC_w	60.8	64.0	65.6	64.5
		$cQIC_w/C_P$	68.0	72.5	74.4	72.5
	Full	QIC_w	35.2	35.5	34.3	35.5
		$cQIC_w/C_P$	27.1	26.7	25.5	27.5
	Indicator	QIC_w	3.8	0.5	0.1	0.0
		$cQIC_w/C_P$	4.4	0.8	0.1	0.0
	Current	QIC_w	0.2	0.0	0.0	0.0
		$cQIC_w/C_P$	0.5	0.0	0.0	0.0
Scenario 2	Cumulative	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.0	0.0	0.0	0.0
	Full	QIC_w	100	100	100	100
		$cQIC_w/C_P$	100	100	100	100
	Indicator	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.0	0.0	0.0	0.0
	Current	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.0	0.0	0.0	0.0
Scenario 3	Cumulative	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.2	0.0	0.0	0.0
	Full	QIC_w	71.4	71.5	73.5	71.1
		$cQIC_w/C_P$	59.9	59.6	60.3	57.9
	Indicator	QIC_w	2.0	0.0	0.0	0.0
		$cQIC_w/C_P$	3.1	0.0	0.0	0.0
	Current	QIC_w	26.6	28.5	26.5	28.9
		$cQIC_w/C_P$	36.8	40.4	39.7	42.1
Scenario 4	Cumulative	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.0	0.0	0.0	0.0
	Full	QIC_w	100	100	100	100
		$cQIC_w/C_P$	100	100	100	100
	Indicator	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.0	0.0	0.0	0.0
	Current	QIC_w	0.0	0.0	0.0	0.0
		$cQIC_w/C_P$	0.0	0.0	0.0	0.0

Models in bold are correctly specified.

sedentary lifestyle (physical activity $<$ 1/week or $\geq$ 1/week), body mass index ($<$ 25, 25-26.9, $\geq$ 27 kg/m^2) and taking medication for hypertension. These covariates were selected a priori because they are factors affecting blood pressure,[18,19] and are also potentially associated with ERI exposure.[20] The three first covariates were not time-varying, but all others were.

We identified being in possible presence of time-dependent confounding. In fact, in addition to being potential confounders, some of the selected covariates were also possibly affected by exposure to psychosocial stressors at work[21]. MSMs were therefore an appropriate approach for estimating the effect of repeated exposure to ERI on ABP. The proposed outcome model specification was $E(Y^{\bar{x}}) = \beta_0 + \beta_1 IE + \beta_2 EC + \beta_3 EO + \beta_4 CE$, where Y is the ABP at the end of follow-up and IE, EC, EO and CE

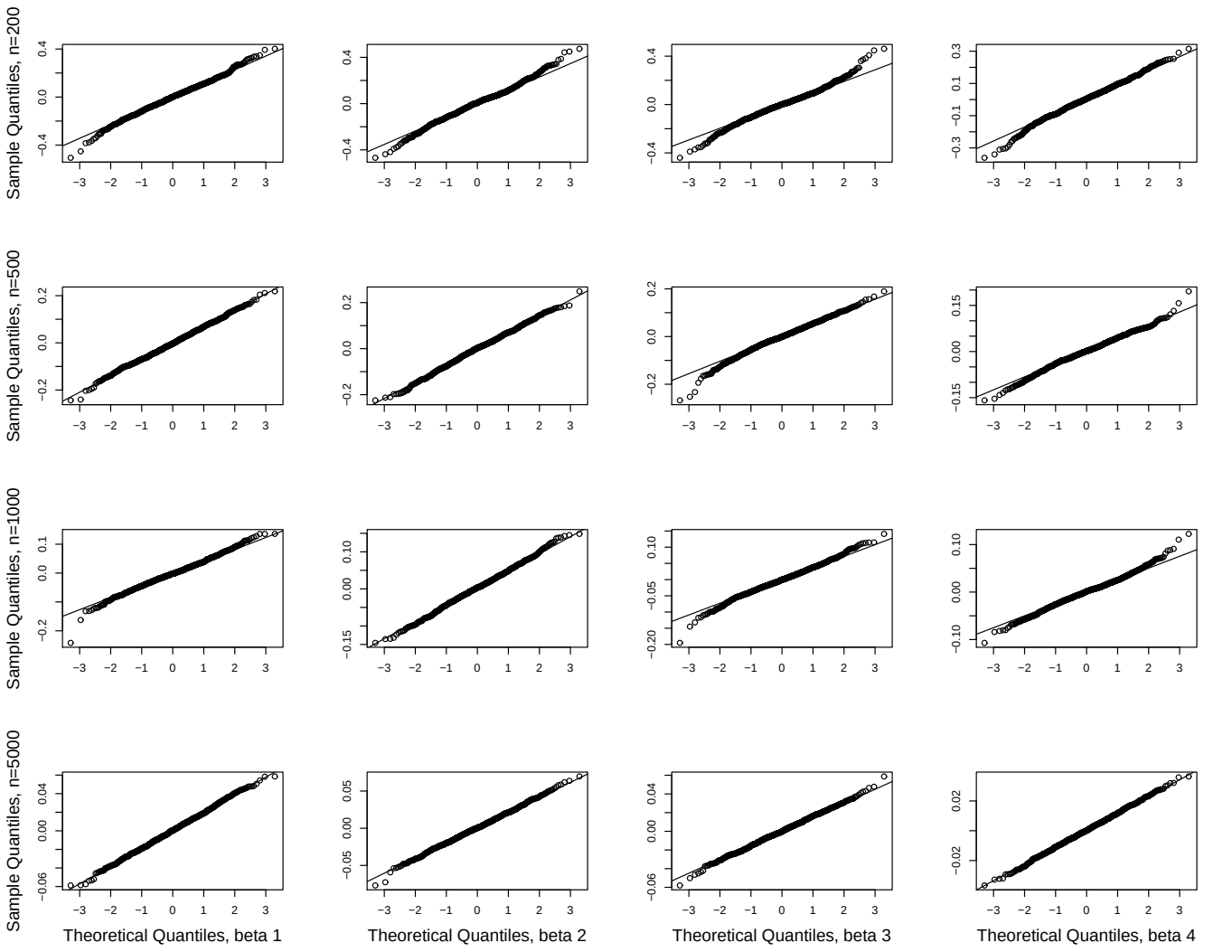

FIGURE 1 Normal QQ-plots of the differences in the estimated parameters between standard weights and stabilized weights for the full model in Scenario 2.

are indicators that the repeated ERI exposure corresponded to the intermittent exposure, exposure cessation, exposure onset and chronic exposure, respectively. The never exposed workers (0,0,0) were used as the reference category.

To test if the postulated model was correctly specified, we compared estimates obtained with untruncated weights w, sw and swm and using the sandwich estimator of the covariance matrix. The weights were estimated using the predicted probabilities of main effects logistic regression models. We provide example code in SAS and R for performing this test as online supplementary material. Separate models were considered for both systolic blood pressure and diastolic blood pressure. The null hypothesis that the outcome model was correctly specified was not rejected for both models, with respective p-values of 0.30 and 0.92. Since the null hypothesis was not rejected, we estimated the causal effect of the repeated ERI exposure on both the systolic and diastolic blood pressure using the stabilized weights with a robust estimator of the variance. Analyses were repeated after

stratifying for sex, since sex-specific relationships between psychosocial stressors at work and cardiovascular health have been documented.[22] Again, the null hypothesis that the outcome model was correctly specified was not rejected (all p > 0.71).

Table 6 presents the p-values of the tests for the correct specification of the postulated outcome model as well as the causal effect estimates. Repeated ERI exposure was not significantly associated with systolic ABP neither when considering all participants nor when considering women and men separately. However, statistically significant associations were observed with diastolic ABP. Among all participants, chronic and intermittent ERI exposure were respectively associated with diastolic ABP measurements that were 1.80 mm Hg (95%CI=0.33-3.27) and 1.31 mm Hg (95%CI=0.04-2.59) higher compared to never exposed workers. This association seemed to be mostly driven by an effect among women participants. Indeed, diastolic ABP was 1.99 mm Hg (95% CI=0.23-3.75) higher in women with chronic ERI exposure and 1.88 mm Hg (95% CI=0.23-3.52) higher in women with intermittent ERI exposure compared to women that had never been exposed. Effect estimates were closer to zero among men and none were statistically significant at $p = 0.05$. Given the width of the confidence intervals, these results should be interpreted as inconclusive rather than as evidence for an absence of an effect.

TABLE 6 P-values of the tests for the model correct specification and estimated causal effect on ambulatory blood pressure (in mm Hg) of the repeated exposure to effort-reward imbalance as compared to never exposed workers with their 95% confidence interval

		Ambulatory Systolic Blood Pressure				Ambulatory Diastolic Blood Pressure			
		P-value[1]	Estimate[2]	95% CI		P-value[1]	Estimate[2]	95% CI	
All participants	Intermittent exposure	0.30	1.49	-0.33	3.31	0.92	1.31	0.04	2.59
	Exposure cessation		1.12	-0.48	2.70		0.59	-0.51	1.69
	Exposure onset		0.19	-1.49	1.88		0.24	-0.99	1.47
	Chronic exposure		1.64	-0.32	3.59		1.80	0.33	3.27
Women	Intermittent exposure	0.71	1.50	-0.95	3.95	0.94	1.88	0.23	3.52
	Exposure cessation		1.74	-0.26	3.74		1.12	-0.35	2.59
	Exposure onset		0.90	-1.39	3.19		1.00	-0.69	2.68
	Chronic exposure		1.77	-0.85	4.38		1.99	0.23	3.75
Men	Intermittent exposure	0.80	1.74	-0.52	4.00	1.00	0.70	-1.40	2.80
	Exposure cessation		0.87	-1.72	3.46		0.07	-1.68	1.81
	Exposure onset		-0.33	-2.95	2.29		-0.74	-2.83	1.36
	Chronic exposure		1.47	-1.41	4.35		1.14	-1.29	3.58

[1]P-value for the test that the model is correctly specified

[2]The reference category is never exposed workers

6 | DISCUSSION

The correct specification of the structural outcome model is one of the hypotheses that needs to be satisfied for estimating the causal effect of an exposure history using MSMs. The existing comparative criteria and the new test we have introduced in this paper are complementary tools to help analysts in specifying their MSMs. The formers are most useful when many candidate specifications are plausible, while the latter may be helpful to validate either an a priori specification or the specification chosen using comparative criteria.

The performance of different versions of our test was investigated and compared in a simulation study. Considering these results, we first recommend that our test should not be performed utilizing truncated weights. In fact, tests based on truncated weights have been observed to reject correctly specified models much more often than expected. We also recommend not using versions comparing standard weights with marginal stabilized weights, since the comparison of such weights is unable to detect when the current model is incorrectly specified. This is because both of these weights yield unbiased estimators of the parameters of the MSM under some types of misspecifications.[3,4] If the goal of the analysis is to compare outcomes according to the full exposure history, that is to compare $E(Y^{\bar{x}})$ vs $E(Y^{\bar{x}'})$, then biased comparisons would be obtained using either weights when the model is misspecified, despite the parameter of the MSM being unbiasedly estimated. We note that if the goal is instead to compare outcomes according to only part of the exposure history, then considering history-restricted marginal structural models would be preferable.[23]

The results of Scenario 4 indicate that all versions of the test reject correctly specified models more often than the nominal rate when the weighting model is misspecified. Although this is not the expected behavior for the test, we do not see this feature as undesirable. Indeed, this means that significant results can be interpreted as evidence that something is wrong with how the parameters of the MSM were estimated. Since the misspecification of either the weighting model or of the outcome model can yield biased estimators, significant results for the test should invite analysts to revise their models. To reduce the risk of misspecifying the weighting model, machine learning approaches such as the Super Learner can be considered.[24]

When the weighting model is correctly specified, the test that had the best general performance (rejection rate close to 5% when the model is correctly specified and among the largest rejection rates when the model is misspecified) is the one based on comparing estimates produced by untruncated standard weights, stabilized weights and marginal stabilized weights, employing a sandwich estimator for the covariance matrix. We have provided an example of SAS and R code for implementing this version of the test as an online supplementary material to facilitate its application. This version of the test was able to detect misspecified models in $\geq 86\%$ of replications even at a small sample size of $n = 200$ in Scenario 1. In Scenario 2, the rejection rate of two out of three misspecified model was also large ($\geq 91\%$) at a sample size of $n = 200$. Although rejection rates were moderate for smaller sample sizes for the misspecified cumulative model in this scenario, high rejection rates were obtained at $n = 5000$.

Despite being the best performing test, its absolute performance was admittedly somewhat disappointing in Scenario 3. Another limitation of this version of the test is that it tends to reject correctly specified models slightly too often for smaller sample sizes. When analyzing small data, one might thus prefer using another version of the test, although there is no clear second best version in our opinion.

To better understand the results we have observed in Scenario 3, we have simulated a very large sample of size $n = 1,000,000$ and have observed that the parameter estimates were very similar for the misspecified models, regardless of the type of weights considered. As such, although the estimators *can* converge to different values depending on the type of weights when the model is misspecified, this may not always be the case. And even when there is a difference, it might be small. In such situations, the test we have introduced will inevitably have poor performance. This indicates that the results of our test shall appropriately be used as evidence against a given specification of the outcome model, but not as evidence in favor of its correct specification. However, this limitation needs not prevent the use of the test. Other tests share the same limitations and are still routinely used in practice. For instance, the Chi-Square test for Structural Equation Models is largely used for evaluating the model fit while it has a number of limitations similar to our test's limitations.[25]

In our application for estimating the effect of repeated exposure to psychosocial stressors at work on blood pressure, we used the test identified as the best among those investigated. The test did not provide evidence that the a priori chosen specification was incorrect. Overall, the results we obtained provide support that repeated exposure to psychosocial stressors at work negatively affect diastolic blood pressure in women. Associations observed with chronic and intermittent exposure were particularly large. These results, along with those of previous studies,[26,27] thus suggest that implementing interventions and policies to sustainably reduce psychosocial stressors at work may help improve blood pressure, and thus have a beneficial effect on cardiovascular health.

Finally, our empirical investigation of the performance of our test based on a simulation study has limitations. First, only four scenarios were considered. These scenarios have allowed us to make important observations concerning the performance of the different versions of our test and to recommend that the test be performed comparing untruncated standard weights, stabilized weights and marginal stabilized weights, using a sandwich estimator for the covariance matrix. Further research is nevertheless warranted to investigate the behavior of the test in other situations. For instance, investigation of the performance of the test when considering repeated measures MSMs or marginal structural Cox models could be conducted. Additional efforts could also be devoted to better understanding situations in which our proposed Wald test has low or no power to detect misspecifications. Such investigations, however, are outside the scope of the current paper.

ACKNOWLEDGEMENTS

This research was funded by the Natural Sciences and Engineering Research Council of Canada grant # 2016-06295 and the

Fondation du CHU de Québec. A. Sall was also supported by a scholarship of Excellence from the Faculté de sciences et de génie of Université Laval. Dr Talbot is a Fonds de recherche du Québec – Santé Chercheur-Boursier.

DATA AVAILABILITY STATEMENT

The R code used for performing the simulation study is openly available as a supplementary online material.

How to cite this article: Sall A., K. Aubé, X. Trudel, C. Brisson, and D. Talbot (2018), A test for the correct specification of marginal structural models, *Statistics in Medicine*, *2018;00:1–6*.

APPENDIX

In this section, we provide a proof of the asymptotic normal distribution of the estimator of the difference in parameters estimated with different weights based on the property that estimators of the parameters are regular and asymptotically linear (RAL).

Proof. An estimator $\hat{\Psi}$ of an m-dimensional parameter Ψ is RAL if

$$\sqrt{n}\left(\hat{\Psi}-\Psi\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i^{\psi}+o_p(1),$$

where ϕ^{ψ} is the influence function of the estimator $\hat{\Psi}$, $\mathbb{E}[\phi^{\psi}]=\mathbf{0}$, $Var[\phi^{\psi}]$ is finite and nonsingular, $o_p(1)$ is a term that converges in probability to zero as n goes to infinity, and $\hat{\Psi}$ meets some mild regularity condition.[28] Such estimators have a normal limit distribution.[28]

Since both $\hat{\boldsymbol{\beta}}^{sw}$ and $\hat{\boldsymbol{\beta}}^{w}$ are regular and asymptotically linear,[9] we can write

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\right)&=\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{w}-\boldsymbol{\beta}^{w}\right)-\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{sw}-\boldsymbol{\beta}^{sw}\right)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i^{\beta^{w}}+o_p(1)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i^{\beta^{sw}}-o_p(1)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i^{\delta}+o_p(1),\end{aligned}$$

where $\phi_i^{\delta}=\phi_i^{\beta^{w}}-\phi_i^{\beta^{sw}}$ is the influence function of the estimator $\hat{\boldsymbol{\delta}}$.

To demonstrate that $\hat{\boldsymbol{\delta}}$ is an RAL, it remains to show that $\mathbb{E}[\phi^{\delta}]=0$ and that $Var[\phi^{\delta}]$ is finite and nonsingular. This directly follows from the fact that the influence functions $\phi^{\beta^{w}}$ and $\phi^{\beta^{sw}}$ are both elements of the Hilbert space $\mathcal{H}$ of m-dimensional functions with mean $\mathbf{0}$ and finite and non-singular variances, with covariance inner product.[28] Because Hilbert spaces are linear spaces, any linear combination of its elements is also an element of the Hilbert space. Thus, $\hat{\boldsymbol{\delta}}$ is an RAL estimator and it has a normal limit distribution. Note that even if even if the regularity conditions necessary for the estimator to be an RAL would not hold, the limit distribution of an asymptotically linear estimator is also normal.[28] □

References

1. Robins James M, Hernán Miguel Angel, Brumback Babette. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550-560.

2. Cole Stephen R, Hernán Miguel A. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology.* 2008;168(6):656–664.

3. Talbot Denis, Atherton Juli, Rossi Amanda M, Bacon Simon L, Lefebvre Geneviève. A cautionary note concerning the use of stabilized weights in marginal structural models. *Statistics in medicine.* 2015;34(5):812–823.

4. Taguri Masataka. Comments on 'A cautionary note concerning the use of stabilized weights in marginal structural models' by D. Talbot, J. Atherton, AM Rossi, SL Bacon, and G. Lefebvre. *Statistics in medicine.* 2015;34(8):1438–1439.

5. Platt Robert W, Brookhart M Alan, Cole Stephen R, Westreich Daniel, Schisterman Enrique F. An information criterion for marginal structural models. *Statistics in medicine.* 2013;32(8):1383–1393.

6. Taguri Masataka, Matsuyama Yutaka. Comments on 'An information criterion for marginal structural models' by RW Platt, MA Brookhart, SR Cole, D. Westreich, and EF Schisterman. *Statistics in medicine.* 2013;32(20):3590–3591.

7. Baba Takamichi, Kanemori Takayuki, Ninomiya Yoshiyuki. A criterion for semiparametric causal inference. *Biometrika.* 2017;104(4):845–861.

8. Neugebauer Romain, Laan Mark. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference.* 2007;137(2):419–434.

9. Robins James M. Marginal structural models versus structural nested models as tools for causal inference In: Halloran M Elizabeth and Berry Donald ed. *Statistical models in epidemiology, the environment, and clinical trials*. In: New-York:Springer 2000 (pp. 95–133).

10. Xiao Yongling, Moodie Erica EM, Abrahamowicz Michal. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods.* 2013;2(1):1–20.

11. Sampson Robert J, Laub John H, Wimer Christopher. Does marriage reduce crime? A counterfactual approach to within-individual causal effects. *Criminology.* 2006;44(3):465–508.

12. Hernán Miguel Angel, Brumback Babette, Robins James M. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5):561-570.

13. Lewis Michael A, MacRae Kenneth D, Kuhl-Habich1 Doorthe, Bruppacher Rudolf, Heinemann Lothar AJ, Spitzer* Walter O. The differential risk of oral contraceptives: the impact of full exposure history. *Human Reproduction.* 1999;14(6):1493–1499.

14. Canty Angelo, Ripley BD. boot: Bootstrap R (S-Plus) Functions.2017. R package version 1.3-19.

15. Drum M, McCullagh P. Comment on the paper by Fitzmaurice, Laird & Rotnitzky. *Statistical Science.* 1993;8:300–301.

16. Trudel Xavier, Brisson Chantal, Milot Alain, Masse Benoit, Vézina Michel. Effort–reward imbalance at work and 5-year changes in blood pressure: the mediating effect of changes in body mass index among 1400 white-collar workers. *International archives of occupational and environmental health.* 2016;89(8):1229–1238.

17. Siegrist Johannes. Adverse health effects of high-effort/low-reward conditions. *Journal of occupational health psychology.* 1996;1(1):27-41.

18. Chobanian Aram V, Bakris George L, Black Henry R, et al. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the JNC 7 report. *Journal of the American Medical Association.* 2003;289(19):2560–2571.

19. Vargas Clemencia M, Ingram Deborah D, Gillum Richard F. Incidence of hypertension and Educational attainment the NHANES I epidemiologic Followup study. *American Journal of Epidemiology.* 2000;152(3):272–278.

20. Kouvonen Anne, Kivimäki Mika, Virtanen Marianna, et al. Effort-reward imbalance at work and the co-occurrence of lifestyle risk factors: cross-sectional survey in a sample of 36,127 public sector employees. *BMC Public Health.* 2006;6(1):24.

21. Siegrist Johannes, Rödel Andreas. Work stress and health risk behavior. *Scandinavian journal of work, environment & health.* 2006;32(6):473–481.

22. Kivimäki Mika, Virtanen Marianna, Elovainio Marko, Kouvonen Anne, Väänänen Ari, Vahtera Jussi. Work stress in the etiology of coronary heart disease–a meta-analysis. *Scandinavian journal of work, environment & health.* 2006;32(6):431–442.

23. Neugebauer Romain, Laan Mark J, Joffe Marshall M, Tager Ira B. Causal inference in longitudinal studies with history-restricted marginal structural models. *Electronic journal of statistics.* 2007;1:119-154.

24. Karim Mohammad Ehsanul, Platt Robert W, group BeAMS. Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural Cox model context. *Statistics in medicine.* 2017;36(13):2032–2047.

25. Hooper Daire, Coughlan Joseph, Mullen Michael. Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods.* 2008;6(1):53-60.

26. Trudel Xavier, Brisson Chantal, Gilbert-Ouimet Mahée, Milot Alain. Psychosocial Stressors at Work and Ambulatory Blood Pressure. *Current cardiology reports.* 2018;20(12):127.

27. Gilbert-Ouimet Mahée, Trudel Xavier, Brisson Chantal, Milot Alain, Vézina Michel. Adverse effects of psychosocial work factors on blood pressure: systematic review of studies on demand-control-support and effort-reward imbalance models. *Scandinavian journal of work, environment & health.* 2014;40(2):109–132.

28. Tsiatis Anastasios. *Semiparametric theory and missing data*. New-York: Springer Science & Business Media; 2007.