


A global goodness-of-fit test for linear structural mean models

Masataka Taguri¹ · Shizue Izumi² 

Received: 25 August 2016 / Accepted: 20 October 2016 / Published online: 30 November 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Structural mean models (SMMs) have been proposed for estimating causal treatment effects in the presence of non-ignorable non-compliance in clinical trials. To obtain a valid causal estimate, we must impose several assumptions. One of these is the correct specification of the parametric part of the SMMs. Model checking is an important task for data analysts to detect any departure of an assumed model from the true one. However, little work has been done on the goodness-of-fit (GOF) test for the SMMs. In this article, we propose a global GOF test of SMMs. Numerical studies show the proposed test can control type I errors if the SMM is correctly specified. Furthermore, the proposed test detects non-linear effect modification of continuous covariates powerfully, while an existing test does not. We apply the proposed method to data derived from a randomized trial to evaluate the impact of a primary care-based intervention on depression.

Keywords Causal inference · Effect modification · Goodness-of-fit test · Instrumental variable · Non-compliance · Structural mean models

1 Introduction

In a typical clinical trial, patients are randomly assigned to different groups with specific treatments; each patient is expected to receive that treatment throughout follow-up to assess its effect on some outcome. However, most clinical trials are not

Communicated by Joe Suzuki.

✉ Shizue Izumi
shizue-izumi@biwako.shiga-u.ac.jp

¹ Department of Biostatistics, Yokohama City University School of Medicine, Yokohama, Japan

² The Center for Data Science Education and Research, Shiga University, 1-1-1 Banba, Hikone City, Shiga 522-8522, Japan

ideal; patients often fail to adhere to their assigned treatment and switch to another trial treatment. Such non-compliance with assigned treatments is a common feature of clinical trials.

Robins (1994) developed structural mean models (SMMs) to cope with non-compliance without having to specify the mechanism of non-ignorable non-compliance (Rubin 1976) using randomization as an instrumental variable. One attractive feature of SMMs is their modeling flexibility, which allows for the expression of the causal effect of received treatments as a function of treatments and covariates through a finite number of unknown causal parameters without specifying the conditional expectation of potential outcomes under the control treatment. SMMs have now been proposed for continuous, discrete, and binary outcomes (Robins 1994; Vansteelandt and Goetghebeur 2003), and related structural distribution models have been developed for survival outcomes (Mark and Robins 1993; Loeys and Goetghebeur 2003).

To obtain a valid causal estimate, we must impose several assumptions. One of these is the correct model specification of the structural model. This can be numerically checked by evaluating the goodness of fit (GOF) of postulated SMMs to the data. For continuous outcomes, Comte et al. (2009) developed a test of the interaction between treatment and a baseline covariate, and Fischer et al. (2011) proposed a local GOF test which can detect a linear effect modification with a covariate but cannot detect a non-linear effect modification. Taguri et al. (2014) recently proposed a model selection criterion as an extension of Akaike's information criterion (Akaike 1973) for evaluating the relative fitting of candidate models using the expected Kullback–Leibler distance as a metric. However, none of them proposed a global GOF test which can detect any misspecifications of the assumed model structure.

In general, the validity of the estimating equations depends on whether the parametric part of the SMM is correctly specified. If the SMM is misspecified, the resultant estimating equations deviates in expected value from zero, thus an inconsistent estimator will be yielded. To get a valid inference, it is desirable to assess the unbiasedness of the estimating equations. Diagnostic tools such as residuals have been widely used to assess the appropriateness of a generalized linear model (Su and Wei 1991; Lin et al. 2002). However, such methods cannot apply to non-compliance data with an instrumental variable. The aim of this article is to develop a global GOF test of linear SMMs. The idea is based on testing for the unbiasedness of g-estimating equations (Robins 1994). The residual processes will be constructed in the same spirit of Su and Wei (1991), Lin et al. (2002), Pan and Lin (2005), and Chen and Qin (2014). Under the null hypothesis that g-estimating equations are unbiased, the residual processes will fluctuate about zero. Thus a large absolute value of the residuals leads to the conclusion of model misspecification. Numerical studies show that the proposed test can control type I errors if the SMM is correctly specified. Furthermore, the proposed test detects non-linear effect modification of continuous covariates with high probability, while Fischer et al.'s test does not.

The remainder of this article is as follows. In Sect. 2, we briefly overview the SMMs and the g-estimation procedure. In Sect. 3, we review the method by Fischer

et al. (2011) and propose a GOF test. In Sect. 4, we present a simulation study to investigate the performance of our proposed test. In Sect. 5, we apply the proposed method to data derived from a randomized trial to evaluate the impact of a primary care-based intervention on depression. Finally, in Sect. 6, we conclude with a discussion.

2 Structural mean models

We consider a randomized two-arm trial, where n patients are randomized to one of the two treatments. Let R be the indicator of treatment assignment, equal to 1 (0) for the test (control) treatment. Let A be the actual treatment whether an individual received test treatment (1: test, 0: control), \mathbf{X} is the vector of baseline covariates, and Y is the continuous outcome measured at the end of the trial. We assume the observed data $\mathbf{O}_i = (R_i, \mathbf{X}_i^T, A_i, Y_i)^T, i = 1, \dots, n$ are n independent and identically distributed random vectors. Thus, we omit the subscript i unless necessary. In contrast to the observed outcome variable Y , we define Y_{ra} with $r, a = 0, 1$ as the potential outcome (Rubin 1974) that would be observed if possibly contrary to the fact that R were set to r and A were set to a . We make the following three assumptions to estimate causal treatment effects:

- (A1) **Stable Unit Treatment Value Assumption (SUTVA)**
 The potential outcome for each patient does not depend on the treatment assigned or the treatment actually received by any other patient. SUTVA also implies the consistency assumption, which means that a patient’s potential outcome under his/her treatment is precisely his/her observed outcome. In notation, SUTVA implies that $Y = RAY_{11} + R(1-A)Y_{10} + (1-R)AY_{01} + (1-R)(1-A)Y_{00}$.
- (A2) **Exclusion restriction**
 Treatment assignment only affects the outcome through its effect on treatment received. This assumption implies that $Y_{ra} = Y_a$ with $r, a = 0, 1$. Under this assumption, $Y_{11} = Y_{01} = Y_1$ is the potential outcome under test treatment, while $Y_{10} = Y_{00} = Y_0$ is that under control treatment.
- (A3) **Randomization assumption**
 The random assignment R and Y_0 are conditionally independent given baseline covariates \mathbf{X} , i.e., $Y_0 \perp\!\!\!\perp R \mid \mathbf{X}$.

Furthermore, we assume that the average causal treatment effects follow linear SMMs (Robins 1994; Goetghebeur and Vansteelandt 2005):

$$E[Y - Y_0 \mid A, \mathbf{X}, R] = \mathbf{AZ}(\mathbf{X}, R)^T \boldsymbol{\theta}, \tag{1}$$

where $\mathbf{Z}(\mathbf{X}, R)$ is a v -dimensional ($v \geq 1$) vector that depends on (\mathbf{X}, R) and $\boldsymbol{\theta}$ is the unknown v -dimensional causal parameter vector of interest. Note that from (1), $E[Y_1 - Y_0 \mid A = 1, \mathbf{X}, R] = \mathbf{Z}(\mathbf{X}, R)^T \boldsymbol{\theta}$ is the effect of the treatment on the treated conditional on the baseline covariates and the randomization indicator (\mathbf{X}, R) . For example, when $\mathbf{Z}(\mathbf{X}, R)^T = (1, \mathbf{X}^T)$, we allow for the possibility that the average

causal effect on the treated is not constant with levels of X and changes linearly with X .

Because the full data (Y_1, Y_0, A, R, X^T) is only partially observed for each patient i , no regression methods for the complete data can be used to fit the model (1). However, from (1) and the assumption A3, it follows that $E[Y - AZ(X, R)^T \theta | X, R] = E[Y_0 | X, R] = E[Y_0 | X]$. Using this, a consistent estimator of θ can be obtained from a class of unbiased g-estimating functions (Robins 1994):

$$\psi(\theta) = (R - p)w(X)\{U(\theta) - q(X)\}, \tag{2}$$

where $p = \Pr[R = 1 | X] = \Pr[R = 1]$ is the randomization probability known by design, $U(\theta) = Y - AZ(X, R)^T \theta$; $w(X)$ is a v -dimensional vector function, and $q(X)$ is a scalar function. For some $w(X)$ and $q(X)$, a consistent estimator of θ (called the g-estimator) is analytically obtained by solving g-estimating equations $\sum_{i=1}^n \psi_i(\theta) = 0$, where $\psi_i(\theta)$ is the i -th sample value of $\psi(\theta)$. The optimal choices for $w(X)$ and $q(X)$ from the viewpoint of efficiency that lead to a semiparametric efficient estimator of θ were derived by Robins (1994). Under the homoscedasticity assumption that the error variance of the regression of $U(\theta)$ on (R, X) is constant, these choices are given by $w_{opt}(X) = \delta_{opt}(X)E[Z(X, R)|X]$ and $q_{opt}(X) = E[U(\theta)|X]$, where $\delta_{opt}(X) = \Pr[A = 1 | R = 1, X] - \Pr[A = 1 | R = 0, X]$ is called the compliance score (Joffe and Breisinger 2003). $\delta_{opt}(X)$ upweights participants characterized by X for whom the effect of treatment assignment on the treatment received is large, thus contributing information to estimate the effect of the treatment on the outcome. Since the optimal choices are unknown functions of X , it is often assumed parametric models for $\delta_{opt}(X)$ in $w_{opt}(X)$ and $q(X)$. In our simulation and data analysis, we estimate $\Pr[A = 1 | R, X]$ in $\delta_{opt}(X)$ by a logistic regression. We assume $q_{opt}(X)$ is linear in X , which leads to an analytical estimator of $\hat{\theta}$ (Fischer et al. 2011). A consistent variance estimator of $\hat{\theta}$ is obtained as $n^{-1} \hat{\Omega}(\hat{\theta})^{-1} \hat{\Lambda}(\hat{\theta}) (\hat{\Omega}(\hat{\theta})^{-1})^T$, where $\Omega(\theta) = -E[\partial \psi(\theta) / \partial \theta^T]$, $\Lambda(\theta) = \text{var}[\psi(\theta)]$.

3 Goodness of fit tests for structural mean models

3.1 Goodness of fit test proposed by Fischer et al. (2011)

Before discussing our method for assessing the fit of the SMM (1), we briefly review the GOF test proposed by Fischer et al. (2011). Their methods are essentially based on the fact that if model (1) is correctly specified, then the expected “treatment-free” outcomes $U(\hat{\theta})$ in both arms $R = 1$ and $R = 0$ will have the same regression functions on X . The GOF test was conducted using the following linear regression model for $U(\hat{\theta})$ on (X, R) :

$$E[U(\hat{\theta}) | X, R] = \beta_0 + \beta_1^T X + \beta_2 R + \beta_3^T R X. \tag{3}$$

To assess whether the model is a good fit, we conducted the test of the following null hypothesis: $H_0 : \beta_3 = \mathbf{0}$. If we find the interaction terms significant (that is, H_0 is rejected), then there is an evidence for the lack of fit. Note that this GOF test would not detect a non-linear effect modification by \mathbf{X} because (3) includes only linear terms of \mathbf{X} . We additionally note that model misspecification will usually occur in the model (3). To see this, let θ_* be the true value of θ . For an arbitrary θ , the following equation holds by the SMM (1):

$$E[U(\theta)|\mathbf{X}, R] = E[Y_0|\mathbf{X}] + E[A|\mathbf{X}, R]Z(\mathbf{X}, R)^T(\theta_* - \theta). \tag{4}$$

The third term of (4) with $\theta = \hat{\theta}$ will be apparently nonlinear in \mathbf{X} when A is binary unless $\hat{\theta} = \theta_*$ holds. In such cases, (3) is a misspecified model. This misspecification could affect the power of the GOF test, although the size of the test should be asymptotically equal to the nominal level because the third term of (4) with $\theta = \hat{\theta}$ has asymptotically zero expectation under the correct specification of (1).

3.2 Proposed goodness of fit test

Rather than assuming a parametric model for $U(\hat{\theta})$ such as (3), we can construct a GOF test in the spirit of Su and Wei (1991), Lin et al. (2002), Pan and Lin (2005), and Chen and Qin (2014). The idea is based on testing for the unbiasedness of the g-estimating equations under the correct model specification. To check the validity of the assumed SMM (1), we consider the following statistics:

$$V_n(\mathbf{x}) = n^{-1/2} \sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)\delta_{\text{opt}}(\mathbf{X}_i)\{U_i(\hat{\theta}) - q_{\text{opt}}(\mathbf{X}_i)\}, \tag{5}$$

where \mathbf{x} is a real-valued vector of length v . Under the null hypothesis that the SMM (1) is correctly specified, (5) has zero expectation for all values of \mathbf{x} . Thus, a large value of the following omnibus test statistic $G_n = \sup_{\mathbf{x} \in R^v} |V_n(\mathbf{x})|$ leads to the conclusion of model misspecification.

To make a GOF test, we need to specify the distribution of $V_n(\mathbf{x})$. The cumulative-sum process $V_n(\mathbf{x})$ converges in distribution to a zero-mean Gaussian process under the null hypothesis that the SMM (1) is correctly specified. Using the Taylor expansion of (5) about $\hat{\theta}$ around the true value θ_* , we have

$$V_n(\mathbf{x}) \approx n^{-1/2} \sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)\delta_{\text{opt}}(\mathbf{X}_i)\{U_i(\theta_*) - q_{\text{opt}}(\mathbf{X}_i)\} + \boldsymbol{\eta}(\mathbf{x}; \theta)^T n^{1/2}(\hat{\theta} - \theta_*), \tag{5}$$

where $\boldsymbol{\eta}(\mathbf{x}; \theta) = n^{-1} \sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)\delta_{\text{opt}}(\mathbf{X}_i)\partial U_i(\theta)/\partial \theta$ and $A \approx B$ means $A - B = o_p(1)$. Using a similar Taylor expansion, we obtain

$$n^{1/2}(\hat{\theta} - \theta_*) = \Omega(\theta_*)^{-1} n^{-1/2} \boldsymbol{\psi}_i(\theta_*), \tag{6}$$

Thus, from (5) and (6), we have

$$V_n(\mathbf{x}) \approx n^{-1/2} \sum_{i=1}^n [I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)\delta_{\text{opt}}(\mathbf{X}_i)\{U_i(\boldsymbol{\theta}_*) - q_{\text{opt}}(\mathbf{X}_i)\} + \boldsymbol{\eta}(\mathbf{x}; \boldsymbol{\theta}_*)^T \boldsymbol{\Omega}(\boldsymbol{\theta}_*)^{-1} \boldsymbol{\psi}_i(\boldsymbol{\theta}_*)]. \tag{7}$$

Although it is hard to specify the explicit distribution form of (7), Su and Wei (1991) proposed a simulation-based method to approximate the null distribution of $V_n(\mathbf{x})$. The idea is as follows. Suppose that S_1, \dots, S_n are independent and identically distributed variables from $F(S)$, with $\mu = E[S] = 0$ and $\sigma^2 = E[S^2] < \infty$, then based on the central limit theorem, we have $n^{-1/2} \sum_{i=1}^n S_i \rightarrow N(0, \sigma^2)$. Let Z_1, \dots, Z_n be independent standard normal random variables. Then conditional on the original data S_1, \dots, S_n , $n^{-1/2} \sum_{i=1}^n Z_i S_i \sim N(0, n^{-1} \sum_{i=1}^n S_i^2) \rightarrow N(0, \sigma^2)$. That is, $n^{-1/2} \sum_{i=1}^n Z_i S_i$ has the same asymptotic distribution as that of $n^{-1/2} \sum_{i=1}^n S_i$. Using these results, for large n , the distribution of $V_n(\mathbf{x})$ is approximated by

$$\tilde{V}_n(\mathbf{x}) \approx n^{-1/2} \sum_{i=1}^n [I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)\delta_{\text{opt}}(\mathbf{X}_i)\{U_i(\hat{\boldsymbol{\theta}}) - q_{\text{opt}}(\mathbf{X}_i)\} + \boldsymbol{\eta}(\mathbf{x}; \hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{\psi}_i(\hat{\boldsymbol{\theta}})] Z_i, \tag{8}$$

where (Z_1, \dots, Z_n) is a random sample from $N(0,1)$. To approximate the null distribution of $V_n(\mathbf{x})$, we generate large number of samples (Z_1, \dots, Z_n) from $N(0,1)$ while fixing the data at their observed values.

4 A simulation study

In this section, the performance of the proposed method is evaluated via a simulation study. The following data (R, X, A, Y) were generated as follows. Let X be distributed as $N(0,1)$ and the treatment assignment R be generated from *Bernoulli*(0.5). Next, the received treatment A was assigned according to the logistic model $\text{logit}[\text{Pr}(A = 1|R, X, \gamma)] = -1 + 4R + X + \gamma$, where γ follows $N(0,0.25)$. Then, outcome Y was generated from $N(3X + A(k_0 + k_1X + k_2X^2) + 0.5\gamma, 0.25)$. This leads to the true SMM: $E[Y_1 - Y_0|A = 1, X] = k_0 + k_1X + k_2X^2$. The shared random effect, γ , gave rise to non-ignorable non-compliance. We set $(k_0, k_1, k_2) = (3, 0, 0)$ for no effect modification by X , $(k_0, k_1, k_2) = (3, 0, 1, 0), (3, 0, 2, 0), (3, 0, 3, 0), (3, 0, 4, 0), (3, 0, 5, 0), (3, 2, 0), (3, 5, 0)$ for linear effect modifications, $(k_0, k_1, k_2) = (-0.2, 0, 0.4), (-0.4, 0, 0.8), (-0.6, 0, 1.2), (-0.8, 0, 1.6), (-1, 0, 2), (-2, 0, 4), (-4, 0, 8)$ for quadratic effect modifications. We set the sample size $n = 500$. For each setting, we ran 1000 simulations.

For the analysis of the simulated data, we assumed the main effect model: $E[Y_1 - Y_0|A = 1, X] = \theta$. We investigated four GOF tests: (i) Fischer: Fisher et al.’s GOF test; (ii) $V_{1n}(x)$: proposed GOF test with $V_n(x) = n^{-1/2} \sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)U_i(\hat{\boldsymbol{\theta}})$; (iii) $V_{2n}(x)$: proposed GOF test with $V_n(x) = n^{-1/2} \sum_{i=1}^n I(\mathbf{X}_i \leq \mathbf{x})(R_i - p)\hat{\delta}_{\text{opt}}(\mathbf{X}_i)U_i(\hat{\boldsymbol{\theta}})$; (iv) $V_{3n}(x)$: proposed GOF test with $V_n(x) =$

Table 1 Empirical size and power of the GOF tests

Pattern of the effect	(k_0, k_1, k_2)	Method			
		Fischer	$V_{1n}(x)$	$V_{2n}(x)$	$V_{3n}(x)$
No effect modification	(3,0,0)	0.040	0.045	0.040	0.048
Linear effect modification	(3,0.1,0)	0.109	0.046	0.042	0.108
	(3,0.2,0)	0.341	0.046	0.044	0.295
	(3,0.3,0)	0.608	0.048	0.048	0.588
	(3,0.4,0)	0.837	0.049	0.049	0.803
	(3,0.5,0)	0.943	0.056	0.056	0.930
	(3,2,0)	1.000	0.081	0.146	1.000
	(3,5,0)	1.000	0.195	0.574	1.000
Quadratic effect modification	(−0.2,0,0.4)	0.715	0.046	0.039	0.579
	(−0.4,0,0.8)	0.552	0.047	0.050	0.928
	(−0.6,0,1.2)	0.510	0.047	0.081	0.976
	(−0.8,0,1.6)	0.486	0.053	0.141	0.982
	(−1,0,2)	0.522	0.054	0.234	0.986
	(−2,0,4)	0.535	0.152	0.782	0.989
	(−4,0,8)	0.535	0.420	0.984	0.990

$n^{-1/2} \sum_{i=1}^n I(X_i \leq x)(R_i - p) \hat{\delta}_{\text{opt}}(X_i) \{U_i(\hat{\theta}) - \hat{q}_{\text{opt}}(X_i)\}$. For each test, the two-sided significance level was set at 0.05.

Table 1 summarizes the empirical rejection probabilities by four methods. For no effect modification case, all of the four GOF tests kept the nominal significance level. For linear effect modification cases, the power of the all tests were increasing as the true effect was increasing. Among the four methods, Fischer et al.’s test performed the best in terms of the empirical power, although the power of the proposed test with $V_{3n}(x)$ was only slightly lower than that of Fischer et al.’s test. This is not surprising because our GOF test was an omnibus test using a Kolmogorov-type test statistic. For quadratic effect modification cases, the power of the proposed tests were also increasing as the true effect was increasing. On the other hand, the power of the Fischer et al.’s test was not monotonically increasing with the strength of the true effect. Among the three statistics for the proposed method, $V_{3n}(x)$ performed by far the best. This indicates that using the optimal nuisance functions $\delta_{\text{opt}}(X)$ and $q_{\text{opt}}(X)$ as described in Sect. 3.2 is very important for the good performance of our proposed test.

5 Application

We now apply the proposed method to data derived from the PROSPECT (Prevention of Suicide in Primary Care Elderly: Collaborative Trial) (Bruce and Pearson 1999; Bruce et al. 2004). Data are available at <http://research.bmh.manchester.ac.uk/biostatistics/research/data>. PROSPECT was a multi-site

prospective, randomized trial designed to evaluate the impact of a primary care-based intervention on reducing major risk factors (including depression) for suicide in later life. Participants were recruited from 20 primary care practices in New York City, Philadelphia and Pittsburgh regions. Ten pairs of practices were matched by region (urban vs suburban/rural), affiliation, size, and population type. Within these 10 pairs, practices were randomly allocated to one of the two conditions. The two conditions were either (a) an intervention based on treatment guidelines tailored for the elderly with care management including antidepressant medication ($R = 1$) or (b) treatment as usual ($R = 0$). For illustration purposes, here, we analyzed the data as if interventions were randomly assigned at the individual level. We use these data to assess the effect of antidepressant medication ($A = 1$: presence; $A = 0$: absence) on the change of the Hamilton Depression Rating Scale (HDRS) (Hamilton 1960) score at four months after randomization from baseline (Y). We use the baseline score of the HDRS as a baseline covariate (X), and it is centered with the mean value of the entire sample for estimation of SMMs.

Table 2 summarizes the analysis results. We started with an Intention-to-treat (ITT) analysis and it indicated that the HDRS score at four months was significantly lower in the intervention group than it was in the control group (ITT effect = $\hat{E}[Y|R = 1] - \hat{E}[Y|R = 0] = -3.62$, 95 % confidence interval: -5.29 to -1.95). However, those who did not comply with the assigned treatment comprised 15.2 % (22/145) of the intervention group and 45.4 % (69/152) of the control group. Thus, the ITT effect would substantially underestimate the true causal effect of the treatment (that is, antidepressant medication). Then, we applied the following two SMMs for estimation of the causal treatment effect on the treated: (i) a one parameter SMM including the main effect only, that is, $Z(X, R) = 1$; (ii) a

Table 2 Summary statistics and estimation results in the suicide prevention trial (PROSPECT)

	Intervention ($n = 145$)	Control ($n = 152$)	p value of GOF test	
			Fischer	Proposed
Antidepressant medication: number (%)	123 (84.8)	69 (45.4)		
HDRS score: mean \pm SD				
At baseline	18.93 \pm 6.17	17.36 \pm 5.62		
At follow-up	11.51 \pm 7.38	13.55 \pm 8.35		
Change from the baseline	-7.42 \pm 6.60	-3.80 \pm 7.90		
ITT effect (95 % CI)	-3.62 (-5.29, -1.95)			
1 parameter SMM (95 % CI)				
Main effect	-7.92 (-12.52, -3.32)		0.770	0.854
2 parameter SMM (95 % CI)				
Main effect	-8.09 (-12.78, -3.40)		0.974	0.825
Effect modification with baseline HDRS score	-0.13 (-1.04, 0.78)			

CI confidence interval, *HDRS* Hamilton depression rating scale, *ITT* intention-to-treat, *PROSPECT* prevention of suicide in primary care elderly: collaborative trial, *SD* standard deviation, *SMM* structural mean model

two parameter SMM assuming the effect modification with X , that is, $Z(X, R) = (1, X)$. The baseline covariate X was centered; thus, the main effect parameter for the model (ii) was interpreted to represent the treatment effect at the mean value for the covariates. As shown in Table 2, the two SMMs gave much larger effect estimates than the ITT analysis did, as expected. From the estimation result of the two parameter SMM, the treatment effect was slightly larger for those with higher baseline levels of the baseline HDRS score, although the effect was not statistically significant. We then applied the proposed GOF test using the test statistics $V_{3n}(x)$ in Sect. 4 as well as the test proposed by Fischer et al. (2011). The p value of the GOF test was large for the one parameter model (i) for both methods, indicating good fitting of the main effect model. No difference was observed between the two GOF tests in this analysis. As noted in Taguri et al. (2014), the larger model (two parameter model) gave the larger p value for the Fischer et al.'s test.

6 Discussion

In this article, we have proposed a new global GOF test for the parametric part of the SMMs. The proposed model-checking method is an objective and informative approach for numerically checking the function form of covariates in SMM. Simulation studies demonstrate that the proposed test works well in terms of the type errors and power for both linear and non-linear effect modifications.

Although SMMs and g -estimation always provide a valid test of the no treatment effect in the presence of non-compliance (Robins 1994), the correct model specification is a fundamental assumption for consistently estimating the causal treatment effect. In this regard, assessing the GOF of the candidate SMMs is very important. Our GOF test and the model selection criterion proposed by Taguri et al. (2014) can be used as complementary approaches, with the GOF test evaluating the overall fit and the model selection criterion evaluating the relative fit of candidate models.

SMMs have been used to handle repeated measures over time as structural nested mean models (Robins 1994) and related structural distribution models have been developed for survival outcomes (Mark and Robins 1993; Loeys and Goetghebeur 2003). Recently, Wallace et al. (2016) proposed a model assessment technique which can detect misspecifications of nuisance functions in SMMs for dynamic treatment regimens using the property of double robustness in observational studies. It is interesting to investigate as to how to extend our method to these problems.

Acknowledgements This work was partially supported by Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (15K15951) to MT and (C25330039) to SI.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Casaki F (eds) Second international symposium on information theory. Akadémiai Kiadó, Budapest, pp 267–281
- Bruce ML, Pearson JL (1999) Designing and intervention to prevent suicide. *Dialogues Clin Neurosci* 1:100–110
- Bruce ML, Ten Have TR, Reynolds CF, Katz II, Schulberg HC, Mulsant BH, Brown GK, McAvay GJ, Pearson JL, Alexopoulos GS (2004) Reducing suicidal ideation and depressive symptoms in depressed older primary care patients—a randomized controlled trial. *J Am Med Assoc* 291:1081–1091
- Chen B, Qin J (2014) Test the reliability of doubly robust estimation with missing response data. *Biometrics* 70:289–298
- Comte L, Vansteelandt S, Toussef E, Baxter G, Vrijens B (2009) Linear and loglinear structural mean models to evaluate the benefits of an on-demand dosing regimen. *Clin Trials* 6:403–415
- Fischer K, Goetghebeur E, Vrijens B, White IR (2011) A structural mean model to allow for non-compliance in a randomized trial comparing two active treatments. *Biostatistics* 12:247–257
- Goetghebeur E, Vansteelandt S (2005) Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Stat Methods Med Res* 14:397–415
- Hamilton M (1960) A rating scale for depression. *J Neurol Neurosurg Psychiatry* 23:56–62
- Joffe MM, Breisinger C (2003) Weighting in instrumental variables and G-estimation. *Stat Med* 22:1285–1303
- Lin DY, Wei LJ, Ying Z (2002) Model-checking techniques based on cumulative residuals. *Biometrics* 58:1–12
- Loeys T, Goetghebeur E (2003) A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics* 59:100–105
- Mark SD, Robins JM (1993) A method for the analysis of randomized trials with compliance information: an application to the multiple risk factor intervention trial. *Control Clin Trials* 14:79–97
- Pan Z, Lin DY (2005) Goodness-of-fit methods for generalized linear mixed models. *Biometrics* 61:1000–1009
- Robins JM (1994) Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat Theory Methods* 23:2379–2412
- Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66:688–701
- Rubin DB (1976) Inference and missing data. *Biometrika* 68:581–592
- Su JQ, Wei LJ (1991) A lack-of-fit test for the mean function in a generalized linear model. *J Am Stat Assoc* 86:420–426
- Taguri M, Matsuyama Y, Ohashi Y (2014) Model selection criterion for causal parameters in structural mean models based on a quasi-likelihood. *Biometrics* 70:721–730
- Vansteelandt S, Goetghebeur E (2003) Causal inference with generalized structural mean models. *J Roy Stat Soc B* 65:817–835
- Wallace MP, Moodie EE, Stephens DA (2016) Model assessment in dynamic treatment regimen estimation via double robustness. *Biometrics* 72:855–864