arXiv:2602.01648v1 [stat.ME] 2 Feb 2026

揭秘双重稳健估计：重叠的作用

Chengxin Yang[1], Laine E. Thomas[1,2], 和 Fan Li[3]

[1]生物统计学与生物信息学系，杜克大学，达勒姆，北卡罗来纳州 27710[2]杜克临床研究所，达勒姆，北卡罗来纳州 27701[3]统计科学系，杜克大学，达勒姆，北卡罗来纳州 27708

2026年2月3日

摘要

研究目的： 双重稳健（DR）估计量在观察性研究中广泛用于评估因果效应。与单独使用逆概率加权（IPW）或结果建模相比，它通常被认为更具优势，因为能够针对模型设定错误提供额外保护。然而，双重稳健性是一种渐近性质，在有限样本中可能并不成立。我们研究了双重稳健估计量的有限样本表现如何取决于比较组之间协变量重叠的程度。

方法： 通过解析性示例以及在协变量重叠程度和模型设定各不相同的多种情景下进行大量模拟，我们考察了双重稳健估计量相对于逆概率加权和结局建模估计量的偏差与方差。

结果： 我们发现：（i）结果模型的设定对双重稳健估计的影响比倾向评分模型的设定更大，而且随着重叠程度降低，这种主导作用会增强；（ii）在重叠较差时，无论模型设定如何，双重稳健估计量通常都会放大极端权重的不利影响（导致较大的偏差和／或方差），其表现也往往不如逆概率加权和结局建模估计量。

结论： 作为实践指南，我们建议在应用中始终首先检查重叠程度。如果重叠较差，分析人员应考虑通过修剪或重叠加权等方法，将目标总体调整为重叠充分的亚群体。

关键词：协变量重叠；双重稳健估计量；有限样本

1 引言

在因果推断中，估计平均处理效应通常有两种方法：结局建模和逆概率加权（IPW）。第三种方法是双重稳健（DR）估计（Scharfstein 等，1999；Robins 等，1994；Lunceford 和 Davidian，2004；Bang 和 Robins，2005），它通过在 IPW 估计量中加入结果模型，或反之，将这两种方法结合起来，也称为增广逆概率加权（AIPW）估计量。如果倾向评分模型或结果模型至少有一个设定正确（但不一定两者都正确），那么双重稳健估计量对平均处理效应就是一致的，也就是说，它在渐近意义下无偏且方差有界。因此，双重稳健估计量通常更受青睐，因为它能额外防范模型设定错误。

然而，双重稳健性是一种渐近性质，在有限样本中未必成立。具体而言，双重稳健估计量中的增广项并未解决逆概率加权估计量一个众所周知的弱点，即对接近 0 或 1 的倾向评分十分敏感，这会导致极端权重，进而在有限样本中产生较大的偏差和方差。尽管关于双重稳健估计的文献十分丰富，但对双重稳健估计量有限样本表现的考察却十分少见。本文将审视在有限样本表现的背景下，双重稳健估计量优越这一看法。正如我们将要展示的，当比较组之间的协变量重叠较差时，双重稳健估计量不仅会继承极端倾向评分所导致的弱点，还会使其加剧。更一般地说，双重稳健估计量的有限样本表现很大程度上取决于协变量重叠程度。在重叠有限的情况下，转而采用关注具有充分重叠的目标总体的方法（如修剪和重叠加权），通常能够得到更稳健、有效且精确的因果效应估计。

2 方法

2.1 符号、估计目标和估计量

考虑一项包含 N 名研究对象的观察性研究。对于每名研究对象 $i(= 1、.....、 N)$，我们观测到一组处理前协变量 X_i、处理状态 $Z_i(= z)$（其中 $z = 1$ 表示接受治疗，$z = 0$ 表示对照）以及结果 Y_i。假设不存在干扰且满足一致性 Rubin（1980），亦即稳定单元处理值假设（SUTVA），则每名研究对象都有两个潜在结果 $Y_i(1)$ 和 $Y_i(0)$，分别对应治疗和对照，其中一个会被观测到 $Y_i = Y_i(Z_i)$。在全文中，我们假设 $\{Y_i(1)\ Y_i(0)\}$ 存在，因此个体处理效应为 $Y_i(1) - Y_i(0)$。倾向评分（PS）是给定协变量时接受治疗的概率：$e(X) = P(Z = 1|X)$，见 Rosenbaum 和 Rubin（1983）。在观察性研究中，PS 未知，必须进行估计，例如通过逻辑回归

$e(X,\ \beta) = 1/(1+\exp(-\beta^T X))$，其中 β 是模型参数。标准目标估计量是平均处理效应（ATE）： $\tau = E($ $Y_i(1) - Y_i(0))$，其中期望值是针对由研究样本代表的目标总体计算的。在全文中，我们采用因果推断中的两个标准假设：i）正值性，即每个个体被分配到任一处理组的概率均非零；ii）条件可交换性，即不存在未测量的混杂因素，见 Hernan and Robins（2023）。在这些假设下，我们有两个关于 τ的识别公式：

$$\tau = E[E(Y_i|Z_i=1,X_i) - E(Y_i|Z_i=0,X_i)] = E\left[\frac{Y_i Z_i}{e(X_i)} - \frac{Y_i(1-Z_i)}{1-e(X_i)}\right]. \tag{1}$$

因此，估计量有两种类型。第一种是结果模型估计量，即我们为潜在结果指定一个模型， $\mu_z(x) = E($ $Y_i(z)|X_i = x) = E(Y_i|Z_i = z,\ X_i = x)$，然后通过 $\hat{}\tau_{om} = \sum_{i=1}^{N}(\hat{}\mu_1(X_i) - \hat{}\mu_0(X_i))/N$ 估计 τ，其中 $\hat{}\mu_z(X_i)$ 是结果模型下治疗 z 的拟合结果。第二种是逆概率加权估计量，

$$\hat{\tau}_{ipw} = \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{Z_i Y_i}{\hat{e}(X_i)} - \sum_{i=1}^{N}\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}\right\},$$

其中$\hat{e}(X)$ 是基于模型估计的倾向评分，对于处理组受试者，权重为 $1/\hat{e}(X_i)$；对于对照组受试者，权重为 $1/$ $(1-\hat{e}(X_i))$。逆概率加权估计量的另一种形式称为 Hajek估计量，它对权重进行归一化，使得每个组中所有权重之和为 1。两者 $\hat{}\tau_{om}$ 和 $\hat{}\tau_{ipw}$ 都要求相应模型被正确设定，才能保证对 τ 的一致性。

双重稳健（DR）估计量为：

$$\hat{\tau}_{dr} = \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\mu}_1(X_i) + \frac{Z_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}\right\} - \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\mu}_0(X_i) + \frac{(1-Z_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}\right\}. \tag{2}$$

可以证明，只要结果模型或倾向评分模型中的任一个（但不一定两者都）设定正确， $E(\hat{}\tau_{dr}) = \tau$，因此 $\hat{}$ τ_{dr} 有两次“做对”的机会（Bang 和 Robins，2005）。

2.2 双重稳健估计量的有限样本误差

有限样本推断将处理分配 Z_i 视为唯一的随机性来源，同时将 $\{Y_i(1)\ Y_i(0)\ X_i\}_{i=1}^{N}$——当前手头的研究样本——视为非随机。因此，样本平均处理效应（SATE） $\tau_S = N^{-1}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}$ 是标准的有限样本目标估计量。相比之下，大样本框架将研究样本视为从超总体中随机抽取的样本，并以 τ（即超总体的平均处理效应（PATE））作为目标估计量，取代其有限样本对应量 τ_S。

具体而言，大样本框架包含两个层次的随机性 Miratrix 等（2018）；Abadie 等人（2020）；Ding（2024）：i）研究对象的抽样过程，它解释了 X_i 和 $\tau - \tau_S$ 中的随机性；ii）处理的随机分配。相比之下，有限样本框架仅包含后者。

在全文中，我们始终将总体平均处理效应 τ 作为目标估计量，这与双重稳健方法所植根的大样本框架，以及双重稳健方法已被广泛应用的观察性研究文献相一致。尽管如此，暂且搁置随机抽样导致的 τ 与 τ_S 之间的差异，并从有限样本视角考察双重稳健方法逼近 τ_S 的表现，仍然很有价值。正如我们很快将要介绍的，这能让我们深入了解双重稳健方法的有限样本行为及其对大样本推断的影响。此外，良好地逼近 τ_S 是科学归纳的内在要求：在将统计分析结果推广到研究样本之外的任何更广泛总体之前，结果首先应对手头的研究样本具有内部效度 Ding（2024）。

理想情况下，良好的估计量应给出接近 τ_S 的估计值。双重稳健估计量与样本平均处理效应之间的差异，即 $\Delta_{dr} = \hat{\tau}_{dr} - \tau_S$，为：

$$\Delta_{dr} := \frac{1}{N}\sum_{i=1}^{N} \frac{Z_i - \hat{e}(X_i)}{\hat{e}(X_i)}\{Y_i(1) - \hat{\mu}_1(X_i)\} + \frac{1}{N}\sum_{i=1}^{N} \frac{Z_i - \hat{e}(X_i)}{1 - \hat{e}(X_i)}\{Y_i(0) - \hat{\mu}_0(X_i)\}. \tag{3}$$

我们将 Δ_{dr} 称为双重稳健估计量的有限样本误差，以表明它是相对于 τ_S的误差。 Δ_{dr} 是基于模型的，并可分解为： $\Delta_{dr} = \frac{1}{N}\sum_i^N R_i^e R_i^y$，其中

$$R_i^e = Z_i - \hat{e}(X_i) \quad \text{and} \quad R_i^y = \frac{Y_i(1) - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} + \frac{Y_i(0) - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)},$$

分别为单元 i的倾向评分模型残差与结果模型的加权残差，下文将其统称为残差。需要说明的是，由于反事实结果缺失， τ_S、 R_i^y 或 Δ_{dr} 均不可观测。然而，这一观点以及对Δ_{dr} 的分解阐明了双重稳健估计量的有限样本行为及其相应的大样本含义，并且使我们能够通过模拟来展示并验证这些直觉；在模拟中，我们生成并追踪所有反事实结果。

重要的是，上述分解阐明了重叠在双重稳健估计中的关键作用。具体而言，即使倾向评分模型设定正确，每个单位仍然存在倾向得分残差 R_i^e ；对于实际处理状态与倾向评分预测相反的单位，该残差尤其大。每个单位的结果残差 R_i^y 是其在每种治疗下潜在结果的预测误差之和

，并按该治疗的估计概率的倒数加权。因此，在有限样本中，对于估计倾向评分接近 0 或 1 的单位，即治疗水平之间重叠较少的区域， R_i^y 会被大幅放大。尽管无论结果模型设定是否正确，极端倾向评分导致的这种误差放大 都会存在，但当结果模型设定错误时，这种现象尤其严重，因为在重叠较少的区域中，设定错误的结果模型因外推而产生的预测误差可能会过大 Yunji Zhou and Thomas（2020年）。上述讨论表明，重叠不足会给双重稳健估计带来双重风险。

总而言之，尽管理论上双重稳健估计量能够额外抵御模型设定错误，但它也继承并加剧了逆概率加权和结局建模策略中因重叠不足而产生的有限样本不利影响。因此，仅仅正确设定倾向评分模型，并不能降低双重稳健估计量的有限样本误差；正确设定结果模型仍至关重要，而这是一项颇具挑战的任务。可以说，如果能够获得正确的结果模型，人们会更倾向于选择相应的结局模型估计量，而非其增广双重稳健版本，因为前者是一致的，且渐近效率高于后者 Tan（2007年）；Robins 等人（2007年），因此也就违背了采用双重稳健估计量的初衷。

为减轻观察性研究中缺乏重叠所带来的不利影响，文献转而采用其他方法，重点关注协变量重叠充分的目标总体（以及相应的估计目标）。例如，可以使用带截尾权重的双重稳健估计量 Crump 等人（2009年）、重叠加权 Fan Li 和 Zaslavsky（2018年）；Li 等人（2018年），或匹配估计量 Li 和 Greene（2013年）；Yoshida 等人（2017年）来实现这一点。

3 模拟研究

3.1 模拟设计

我们在不同情景下开展模拟研究，这些情景中的组间协变量重叠程度各不相同。遵循 Li et al. Li et al.（2018年）的模拟设计，我们首先从均值为零、边际方差为单位方差且两两相关系数为 0.5 的多元正态分布中生成 V_1–V_6 。然后，将 V_1–V_3 保留为 X_1–X_3，同时通过示性函数 $I_{\{V<0\}}$ 将 V_4–V_6 二分化，以生成二元X_4–X_6，其边际期望均约为 0.5。真实倾向评分遵循逻辑斯蒂模型，logit($e(X)$) $= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6$；观测到的二元治疗则根据成功概率为 $e(X)$ 的伯努利分布独立生成。连续结局 Y服从 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \tau Z + \epsilon$，其中 ϵ 为

标准正态分布，这意味着所有研究对象的治疗效应均同质，τ。为简单起见，我们设定 $\tau = 1$。

我们模拟两种治疗患病率和两种重叠程度，构成四种情景。具体而言，我们将倾向评分模型参数设为 $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = d \times (0.2, 0.3, 0.4, -0.25, -0.3, -0.3)$。我们选择截距 α_0，使治疗患病率约为 0.4 和 0.1；并令 d 分别为 1 和 3，以代表良好和较差的协变量重叠。补充材料中提供了各情景下倾向评分分布的更多说明。对于结果模型，我们选择 $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (-0.5, -0.8, -1.2, 0.8, 0.8, 1)$ 和 $\beta_0 = 0$，使结果较低的研究对象更有可能接受治疗。在每种情景下，我们反复模拟 10,000 个不同样本量 $n \in \{100, 300, 500, 1000, 2000\}$的数据集，并基于 $n = 500$报告主要结果。由于抽样数据集中有不可忽略的比例不包含任何接受治疗的个体，因此治疗患病率为 0.1 时不采用 100 的样本量。对于每种情景，我们分别使用逆概率加权和双重稳健方法估计 τ，并考察是否在 0.05 或 0.1 处进行倾向得分截尾，同时也采用重叠加权和结果回归。所有假设的结果模型均为线性模型。模拟重复次数 10,000 通常足以获得数值稳定结果；唯一例外是，在重叠较差时，未经截尾的双重稳健方法在不同随机种子下会产生中等程度的数值差异，但我们观察到的现象和模式保持不变。

我们进一步考察模型设定错误。我们模拟了两类模型设定错误：i）函数形式错误（将 X_3 误认为 X_3^2）和 ii）遗漏变量（X_3 和 X_4）。每类模型设定错误均分别应用于倾向得分模型或结果模型，但不会同时应用于两者，因此共有四种模型设定错误。我们报告的主要结果基于函数形式错误，而补充材料则给出了遗漏变量情况下的结果。一般而言，函数形式错误是指：协变量纳入回归模型时所采用的形式不同于它在协变量与结局关系中实际起作用的正确形式。我们仅以使用错误的多项式项作为示例进行说明。

3.2 模拟结果

我们按如下方式安排模拟结果的展示：表1和表2展示了使用各种估计量估计 τ 的结果。表3及其后的所有图均展示有限样本量，以帮助理解表1和表2，并说明双重稳健估计量的有限样本行为。仅在提及估计量名称时，我们指的是未经修剪的估计量。

表 1 展示了结果

表 1：在不同重叠程度下，模型设定正确的估计量的均方根误差（RMSE）和偏差。

估计量	治疗患病率 $= 0.4$				治疗患病率 $= 0.1$			
	$d = 1$		$d = 3$		$d = 1$		$d = 3$	
	均方根误差	偏差	均方根误差	偏差	均方根误差	偏差	均方根误差	偏差
双重稳健								
不截尾	0.11	0.00	0.47	0.00	0.24	0.00	2.31	−0.01
$\delta = 0.05$	0.11	0.00	0.16	0.00	0.20	0.00	0.25	−0.01
$\delta = 0.1$	0.11	0.00	0.16	0.00	0.22	0.00	0.25	−0.01
IPW								
不截尾	0.19	0.00	1.04	−0.37	0.65	−0.05	2.00	−1.52
$\delta = 0.05$	0.16	0.00	0.20	−0.01	0.26	−0.01	0.29	−0.03
$\delta = 0.1$	0.13	0.00	0.17	0.00	0.25	−0.01	0.26	−0.02
结果回归	0.10	0.00	0.14	0.00	0.21	0.00	0.51	−0.02
重叠加权	0.10	0.00	0.14	0.00	0.16	0.00	0.20	0.00

回归模型和倾向评分模型均正确设定时各估计量的均方根误差（RMSE）和偏差。在近似平衡的设计（患病率为 0.4）且重叠充分（$d = 1$）的情况下，所有估计量的偏差都可以忽略不计，其中结果回归和重叠加权估计量的均方根误差最小。随着重叠程度下降（$d = 3$），逆概率加权（IPW）会产生偏差，且 IPW 和双重稳健（DR）会导致均方根误差急剧上升，而其他估计量总体上仍保持无偏且稳定。值得注意的是，尽管 DR 通常优于 IPW，但在所有情景中，其表现都不如结果回归，也不如聚焦于重叠性良好的总体的方法，即采用修剪的对应方法和重叠加权。

表2：倾向评分模型或结果模型之一设定错误，或两者均设定错误时各估计量的均方根误差和偏差。模型设定错误是由函数形式错误导致的（错误 X_3 乘以X_3^2）。

估计量		误设模型	治疗患病率 $= 0.4$				治疗患病率 $= 0.1$			
			$d = 1$		$d = 3$		$d = 1$		$d = 3$	
			RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
双重-稳健	不截尾	PS	0.11	0.00	0.24	0.00	0.23	0.00	0.89	−0.01
		结果	0.14	−0.02	2.11	−0.07	0.43	−0.08	4.06	−0.45
		Both	0.36	−0.33	1.07	−0.98	0.46	−0.33	1.48	−0.78
	$\delta = 0.05$	PS	0.11	0.00	0.16	0.00	0.20	0.00	0.27	−0.01
		结果	0.13	−0.01	0.19	−0.02	0.23	−0.02	0.26	−0.01
		Both	0.36	−0.33	0.86	−0.84	0.36	−0.25	0.47	−0.38
IPW	不截尾	PS	0.35	−0.30	1.22	−1.06	0.60	−0.32	2.03	−1.65
	$\delta = 0.05$	PS	0.34	−0.30	0.81	−0.78	0.38	−0.24	0.50	−0.38
结果回归		结果	0.40	−0.37	1.09	−1.08	0.47	−0.36	1.13	−0.99
重叠加权		PS	0.31	−0.28	0.76	−0.74	0.31	−0.23	0.44	−0.37

使用 $\delta = 0.1$ 修剪得到的结果与使用 $\delta = 0.05$ 的结果相似，因此未予展示。

表2展示了在函数形式错误的模型设定下各估计量的均方根误差和偏差。

遗漏变量导致模型设定错误时所呈现的模式相似，因此将其放入补充材料中讨论。尽管从理论上讲，在所有这些情况下，双重稳健估计量对于 τ 都是一致的，但其有限样本表现却存在显著差异。在具有充分重叠的平衡设计中，只要正确设定倾向评分模型或结果模型中的任一个，偏差都可以忽略不计。相反，在重叠较差和/或不平衡设计中，结果模型设定的影响远强于倾向评分模型设定，而且随着重叠程度降低，这种主导作用会进一步增强。具体而言，与两个模型均正确的双重稳健估计量相比（表 1），仅结果模型正确的双重稳健估计量所得结果相近；而仅倾向评分模型正确的双重稳健估计量则会导致偏差显著增加，尤其是均方根误差增加，在重叠较差时甚至可能达到目标估计量的数倍。这表明，较差的重叠会大幅削弱双重稳健估计量中正确倾向评分模型所提供的额外保护。

正如理论所预测的那样，在两个模型中仅有一个存在模型设定错误时，双重稳健估计量在大多数情况下都明显优于相应的设定错误的逆概率加权估计量和结果回归估计量。然而，在重叠严重不足的情况下（$d = 3$），存在例外：结果模型设定错误的双重稳健估计量，其均方根误差远高于所有其他估计量。相比之下，采用截尾权重的双重稳健估计量始终具有较小的偏差和方差。

有趣的是，在重叠性较差时，倾向评分模型和结果模型均设定错误的双重稳健估计量通常会产生较大的偏差，但令人意外的是，其均方根误差却小于仅结果模型设定错误时的对应估计量。此外，两种模型均设定错误时，修剪极端权重对双重稳健估计量表现的改善，不如仅一种模型设定错误时那么明显。总之，重叠性与模型设定错误之间显然会相互影响双重稳健估计量的有限样本表现，这与 Miratrix 等（2018）的发现相呼应。

表3：两种模型均设定正确，或倾向评分模型和结果模型之一因函数形式错误而设定错误（将 X_3 误认为 X_3^2）时，不同（子）总体中有限样本误差的均值和平均绝对值（MAV）。

模型设定		总体	治疗患病率 $= 0.4$				治疗患病率 $= 0.1$			
			$d = 1$		$d = 3$		$d = 1$		$d = 3$	
			平均	MAV	平均	MAV	平均	MAV	平均	MAV
设定正确		总体	0.00	0.07	0.00	0.22	0.00	0.18	−0.01	0.56
		Tail	0.00	0.32	0.01	0.87	0.01	0.51	0.01	1.02
		Bulk	0.00	0.08	0.00	0.12	0.00	0.20	−0.01	0.54
Mis-设定	倾向评分模型	总体	0.00	0.07	0.00	0.16	0.00	0.17	−0.01	0.49
		Tail	0.00	0.29	−0.02	0.65	0.00	0.47	0.02	0.72
		Bulk	0.00	0.08	0.00	0.11	0.00	0.20	−0.01	0.50
	结果模型	总体	−0.02	0.10	−0.07	0.67	−0.08	0.31	−0.45	1.07
		Tail	−0.14	0.85	−0.47	5.08	−0.42	1.79	−0.98	3.02
		Bulk	0.00	0.10	−0.01	0.16	−0.02	0.26	−0.25	0.80

为进一步考察双重稳健估计量的有限样本行为，我们评估了三个（子）总体的有限样本误差：i）总体，包括所有研究对象；ii）尾部，包括$\hat{e}(X_i)$位于0.05–0.95百分位数范围之外的研究对象；iii）主体部分，包括$\hat{e}(X_i)$位于0.25–0.75百分位数范围内的研究对象。根据设定的倾向评分模型，这三者分别对应总体、重叠性较差的总体和重叠性良好的总体。对于标记为 l的每个总体，我们计算特定于（子）总体 l,$\Delta_{dr,\ l} = \sum_{i:\ e(X_i)\ \in l} R_i^e R_i^y / N_l$的有限样本误差在10,000次重复模拟中的均值和平均绝对值（MAV），结果见表3。

根据表3，在所有模型设定和模拟情景下，尾部总体的 $\Delta_{dr,\ l}$ MAV 和均值均高于（或不低于）主体部分总体，而总体总体的取值 $\Delta_{dr,\ l}$ 介于两者之间。值得注意的是，结果模型设定错误会导致尾部的 $\Delta_{dr,\ l}$ 均值和 MAV 急剧增加；随着重叠程度降低以及设计变得不平衡（患病率为 0.1），这种严重程度还会进一步增加。相比之下，结果模型设定错误下，主体部分中的 $\Delta_{dr,\ l}$ 仍保持较低且稳定。另一方面，倾向评分模型设定错误不会导致所有（子）总体的 $\Delta_{dr,\ l}$ 出现明显变化。总之，尾部总体是 Δ_{dr} 的主要决定因素，并且极易受到结果模型设定错误的影响，这也印证了我们在方法部分讨论的双重稳健的双重困境。

图1和图2展示了残差、 R_i^e和 R_i^y及其乘积 $R_i^e R_i^y$在倾向评分范围内的表现，从而深入揭示了模型设定错误及其与重叠程度之间的相互作用。具体而言，我们将真实倾向评分均匀划分为 100 个区间，每个区间长度为 0.01。对于每个区间 l，我们在每种模型设定下，基于平均残差 $\sum_{i:\ e(X_i)\in l} R_i^e / N_l$ 和 $\sum_{i:\ e(X_i)\in l} R_i^y / N_l$的 10,000 次重复模拟，计算平均绝对值（MAV）以及区间特异的有限样本误差 $\sum_{i:\ e(X_i)\in l} R_i^e R_i^y / N_l$。随后，我们通过计算每种模型设定错误下的 MAV 与正确的模型设定下 MAV 的比值，得到相对平均绝对值。因此，大于（小于）1 的相对平均绝对值表示：如果发生该模型设定错误，区间 l 中 R_i^e、R_i^y 或 $R_i^e R_i^y$ 的预期绝对值将被放大（缩小）的程度。为反映每个区间 l 对总体估计的影响，我们进一步将每个区间 l 的相对平均绝对值乘以 $\Phi(l)$ 进行缩放，其中 $\Phi(l)$ 是使用来自目标总体的大样本数据集（样本量为 100,000）进行经验估计的真实倾向评分落在区间 l 内的概率质量。残差及其乘积的缩放相对 MAV 分别显示在图1和图2中；结果模型设定错误不会改变 R_i^e，因此未予展示。

从图 1 可以看出，模型误设通常通过总体集中所在的倾向得分区域影响 Δ_{dr}，如 $\Phi(l)$ 所示。在所有模型误设和残差中，由结果模型设定错误引起的结果残差相对增幅在所有情景中最为急剧，并且

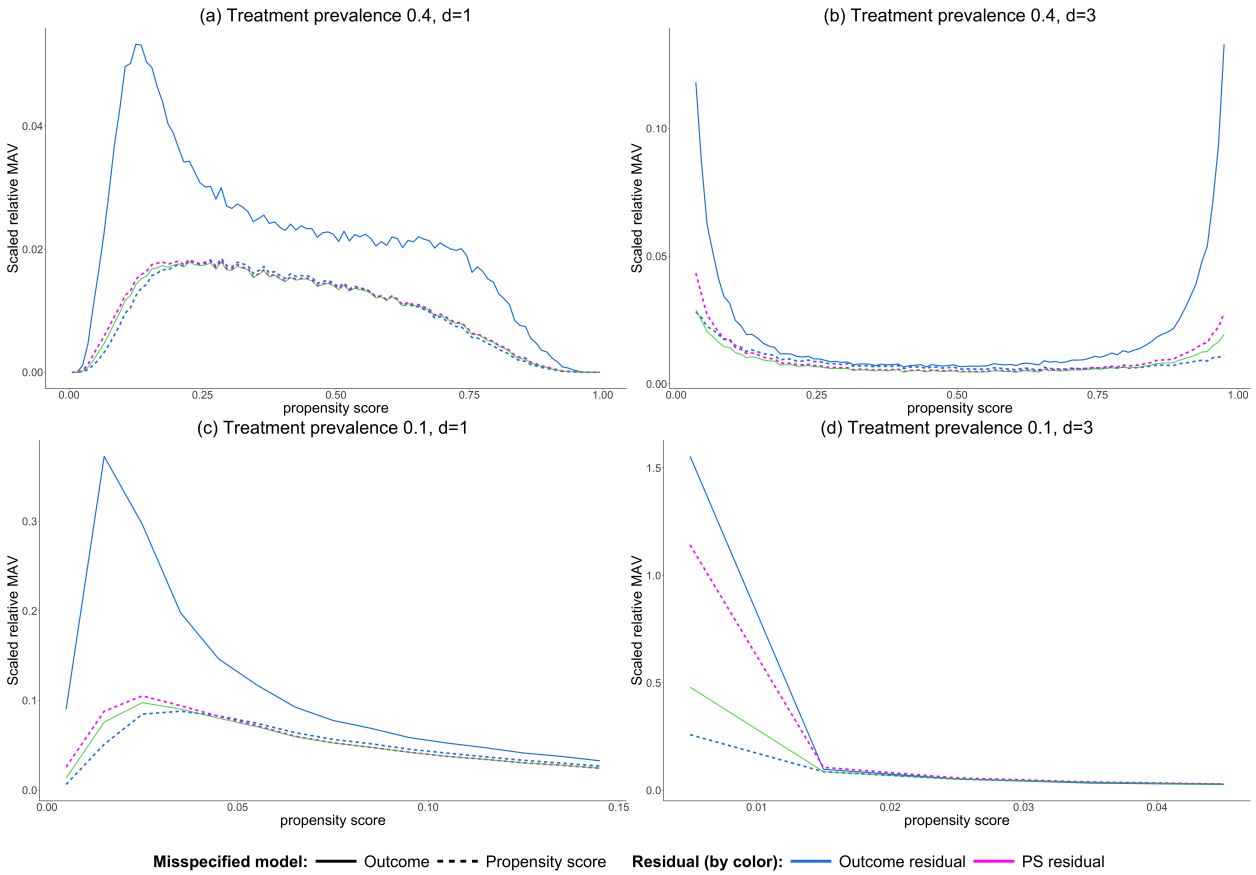


图 1： R_i^e 和 R_i^y 的缩放相对 MAV。模型设定错误由函数形式错误引起（将X_3 误认为 X_3^2）。绿色线（$\Phi(l)$）表示相对尺度没有变化。面板 (b) 尾部的较大值以及面板 (c) 和 (d) 的其余部分见补充材料。

其程度随重叠恶化而增大。更具体而言，在平衡设计下，结果模型设定错误在重叠不足时于倾向评分尾部导致的结果残差增幅，远比重叠充分时于倾向评分主体部分更为急剧；在不平衡设计下也可进行类似比较。这些观察结果证实了倾向评分接近 0 和 1 的个体存在误差放大，并阐明了在重叠较差时结果模型设定随之占据主导地位。

有趣但也值得注意的是，倾向评分模型设定错误并不总会增大结果残差——例外情况出现在倾向评分接近 0 和 1 的亚群体中，因为设定错误的倾向评分模型不一定会产生比正确的倾向评分模型更极端的逆概率权重，因此结果残差可能会被意外地 稳定下来并降低。因此，倾向评分模型设定错误并不总会增大 Δ_{dr} ，因为倾向评分残差的增加可能会被结果残差的减少所抵消。图2展示了一个例子：在重叠较差的情况下，在倾向评分接近 0 和 1 的区域，倾向评分模型设定错误并不会增大（事实上，甚至会降低）有限样本误差，这与图 1 中相应结果残差降低的情况相呼应。

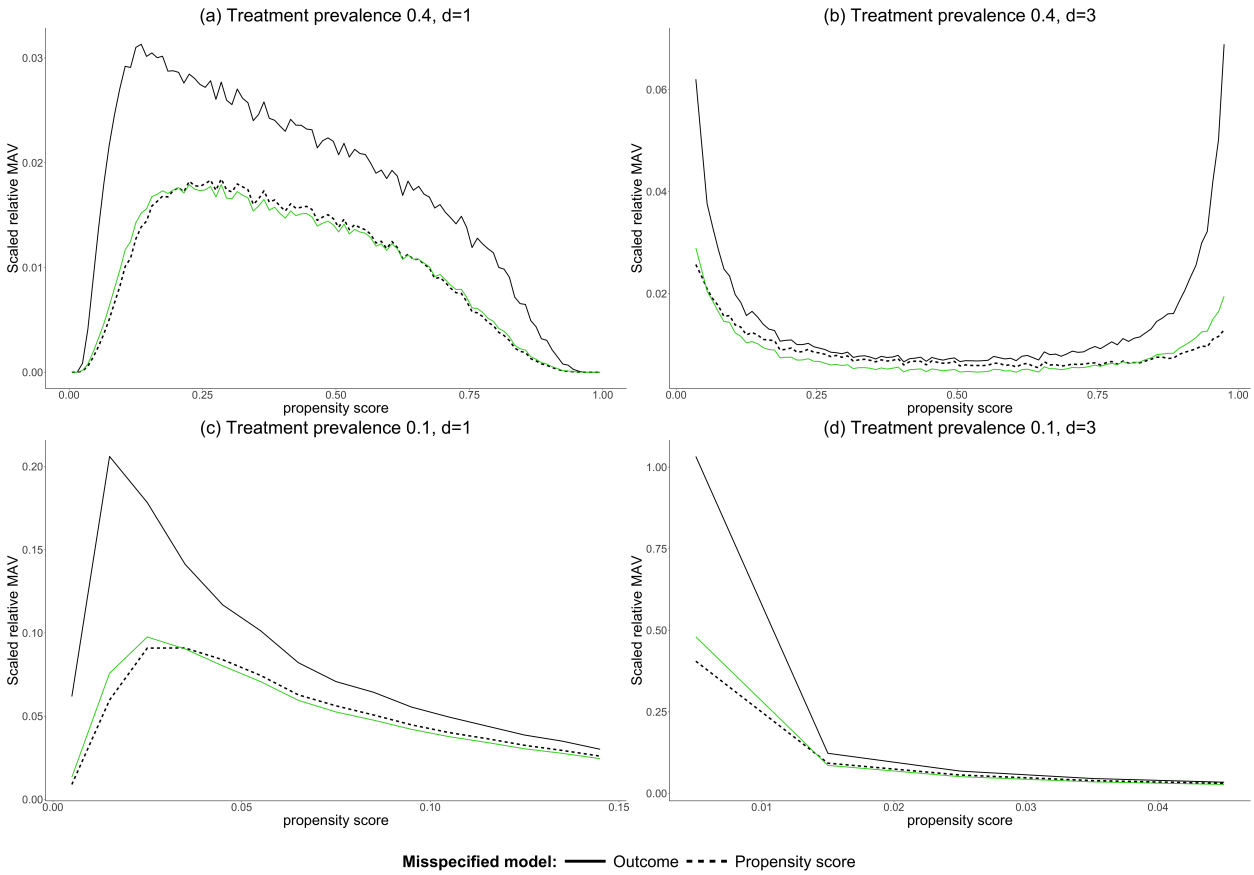


图2： $R_i^e R_i^y$ 的缩放相对 MAV。模型设定错误源于函数形式错误（将 X_3误认为 X_3^2）。绿色线（$\Phi(l)$）表示相对尺度没有变化。面板 (b) 尾部以及面板 (c) 和 (d) 其余部分中的较大值详见补充材料。

4 应用

我们分析了 COMPARE-UF 子宫肌瘤登记系统的数据，以比较子宫切除术和子宫肌瘤切除术治疗子宫肌瘤对生活质量的影响 Stewart 等（2018）；Nicholson 等（2019）；Laughlin-Tommaso 等（2020）。我们的数据集包含 1060 名患者，其中 557 名接受了子宫肌瘤切除术，503 名接受了子宫切除术。我们研究的终点是 UFS 总生活质量（UFS-QoL），并确定了 16 个混杂因素，包括人口统计学特征、病史、基线生活质量组成部分和临床特征。我们为接受子宫切除术的倾向评分指定了一个逻辑斯蒂模型，并为结局建模指定了一个线性模型，其中包含这 16 个混杂因素各自的主效应。如图 3 所示，根据估计倾向评分的直方图，两种治疗组的患者差异很大，其中相当多的患者倾向评分接近 0 或 1。

表 4 展示了不同方法估计的子宫切除术与子宫肌瘤切除术在生活质量（QoL）方面的因果效应。结果差异显著：纳入重叠较少区域中的个体，即

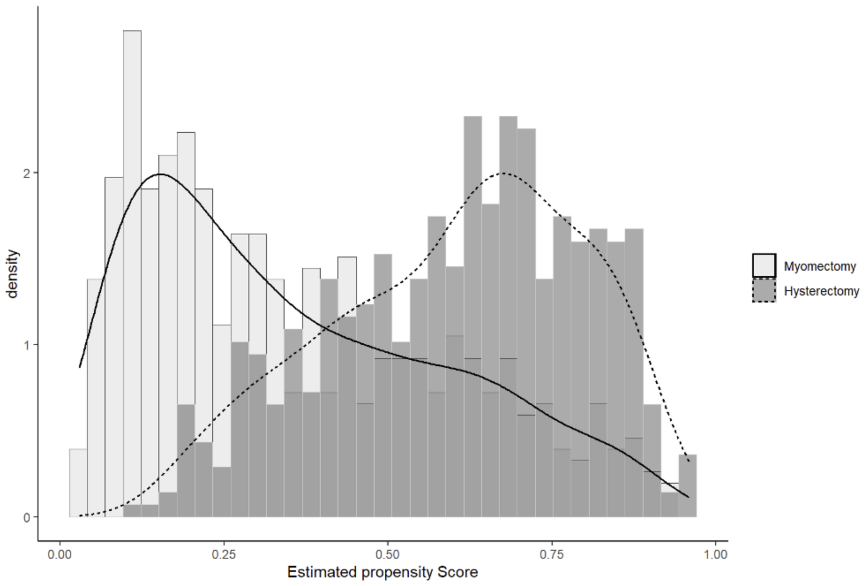


图 3：子宫肌瘤治疗中按治疗方式划分的估计倾向评分分布。

表 4：不同方法估计的 UFS-QoL 均值差。

估计量	UFS-QoL 治疗效应（95% 置信区间）（标准差） 子宫切除术 − 子宫肌瘤切除术
双重稳健	
不截尾	4.93（1.41，8.45）（1.79）
$\delta = 0.05$	4.95（1.45，8.44）（1.78）
$\delta = 0.1$	3.73（0.41，7.04）（1.69）
IPW	
不截尾	4.57（0.96，8.17）（1.84）
$\delta = 0.05$	4.77（1.16，8.38）（1.84）
$\delta = 0.1$	3.54 (0.24, 6.84) (1.69)
结果回归	3.84（0.54，7.13）（1.68）
重叠加权	3.55 (0.27，6.83) (1.67)

采用未修剪权重或在 $\delta = 0.05$ 处修剪的方法，会得出较大的效应，但标准误也较大。相比之下，侧重于重叠较多亚群体的方法，即重叠加权和在 $\delta = 0.1$ 处修剪，得出的效应较温和、标准误较小，且与结果回归估计量一致。这些结果表明，不同亚群体之间的治疗效应可能存在显著异质性，因此在选择用于推断的目标总体时必须慎重考虑。

5 讨论

我们研究了观察性研究中用于因果推断的双重稳健方法。首先，与倾向评分模型的设定相比，结果模型的设定对双重稳健估计的影响更大，而且随着重叠程度降低，这种影响的主导地位会增强。其次，在重叠性较差的情况下，双重稳健估计量通常会产生较大的方差；如果结果模型设定错误，还会产生较大的偏差。事实上，双重稳健估计量可能在所有方面都劣于仅基于正确结果模型的估计量。我们对这些现象提供了分析性和实证性的解释。

我们强调，双重稳健估计并不能神奇地解决重叠性差的问题；事实上，它不仅继承了缺乏重叠所带来的不利后果，还会通过极端逆倾向得分权重放大结果模型中的预测误差，从而加剧这些后果。因此，正如我们的模拟所揭示的那样，结果模型设定错误时的双重稳健方法劣于正确的逆概率加权估计量。结果模型所面临的这一苛刻要求，在很大程度上抵消了重叠性较差时采用双重稳健方法的目的：除非结果模型正确，否则在逆概率加权的基础上加入结果建模（或反过来）只会加剧偏差和方差；但另一方面，如果已经知道并能够获得正确的结果模型，使用双重稳健方法又没有任何益处。因此，当研究中存在严重的重叠不足时，我们建议分析人员将目标转向具有充分重叠的目标总体，例如采用匹配、重叠权重，或使用带截尾权重的加权估计量。当改变目标总体并不理想时，可以采用带校准权重的逆概率加权，例如 Zubizarreta（2015年）提出的稳定化平衡权重，或 Imai 和 Ratkovic（2013年）提出的协变量平衡倾向得分，以减轻极端权重的影响。在某些情况下，就偏差和方差而言，结果建模也可能更为可取；此时可以转向灵活的结果建模方法，例如机器学习和 Hill（2011年）提出的贝叶斯非参数模型。

参考文献

Abadie、A.、Athey、S.、Imbens、G. W. 和 Wooldridge、J. M.（2020年）。回归分析中基于抽样的不确定性与基于设计的不确定性。《计量经济学》，88(1):265–296。

Bang、H. 和 Robins、J. M.（2005）。缺失数据与因果推断模型中的双重稳健估计。《生物统计学》，61(4):962–973。

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A.（2009）。估计平均处理效应时如何处理重叠有限的问题。*Biometrika*，96(1):187–199。

Ding, P.（2024）。因果推断入门课程。Chapman & Hall/CRC统计科学丛书。泰勒与弗朗西斯。Fan Li, K. L. M. and Zaslavsky, A. M.（2018）。通过倾向评分加权平衡协变量。美国统计学会会刊，113(521):390–400。Hernan, M. and Robins, J.（2023）。因果推断：假如。Chapman & Hall/CRC统计学与应用概率专著。泰勒与弗朗西斯。Hill, J. L.（2011）。用于因果推断的贝叶斯非参数建模。计算与图形统计学期刊，20(1):217–240。Imai, K. and Ratkovic, M.（2013）。协变量平衡倾向评分。英国皇家统计学会会刊*B*辑：统计方法学，76(1):243–263。Laughlin-Tommaso, S. K., Lu, D., Thomas, L., Diamond, M. P., Wallace, K., Wegienka, G., Vines, A. I., Anchan, R. M., Wang, T., Maxwell, G. L., et al.（2020）。COMPARE-UF 子宫肌瘤登记系统中子宫肌瘤切除术后短期生活质量。美国妇产科学杂志，222(4):345–e1。Li, F., Thomas, L. E., and Li, F.（2018）。通过重叠权重处理极端倾向评分。美国流行病学杂志，188(1):250–257。Li, L. and Greene, T.（2013）。倾向评分分析中配对匹配的加权类比方法。国际生物统计学杂志，9(2):215–234。Lunceford, J. K. and Davidian, M.（2004）。在估计因果治疗效应时通过倾向评分进行分层和加权：一项比较研究。医学统计学，23(19):2937–2960。Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F.（2018）。加权有价值吗？如何看待和使用调查实验中的权重。政治分析，26(3):275–291。Nicholson, W. K., Wegienka, G., Zhang, S., Wallace, K., Stewart, E., Laughlin-Tommaso, S., Thomas, L., Ja- coby, V. L., Marsh, E. E., Wise, L., et al.（2019）。有症状的子宫平滑肌瘤患者接受子宫切除术与子宫肌瘤切除术后的短期健康相关生活质量比较。妇产科学，134(2):261–269。Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A.（2007）。评论：当“逆概率”权重高度可变时双重稳健估计量的表现。*Statistical Science*，22(4):544 – 559。Robins, J. M., Rotnitzky, A., and Zhao, L. P.（1994）。当某些回归变量并非总能观测到时回归系数的估计。美国统计学会会刊，89(427):846–866。

ROSENBAUM, P. R. 和 RUBIN, D. B.（1983年）。倾向评分在因果效应观察性研究中的核心作用。*Biometrika*，70(1):41–55。Rubin, D. B.（1980年）。实验数据的随机化分析：关于 Fisher 随机化检验的评论。美国统计学会会刊，75(371):591–593。Scharfstein, D. O.、Rotnitzky, A. 和 Robins, J. M.（1999年）。使用半参数无应答模型调整不可忽略的脱落。美国统计学会会刊，94(448):1096–1120。Stewart, E. A.、Lytle, B. L.、Thomas, L.、Wegienka, G. R.、Jacoby, V.、Diamond, M. P.、Nicholson, W. K.、Anchan, R. M.、Venable, S.、Wallace, K.、Marsh, E. E.、Maxwell, G. L.、Borah, B. J.、Catherino, W. H. 和 Myers, E. R.（2018年）。比较子宫肌瘤管理方案：以患者为中心的结果登记研究（compare-uf）的研究依据与设计。美国妇产科学杂志，219(1):95.e1–95.e10。Tan, Z.（2007年）。评论：理解 OR、PS 和 DR。*Statistical Science*，22(4):560–568。Yoshida, K.、Hernández-Díaz, S.、Solomon, D. H.、Jackson, J. W.、Gagne, J. J.、Glynn, R. J. 和 Franklin, J. M.（2017年）。用匹配权重同时比较三种治疗组：与三向匹配的比较。*Epidemiology*，28(3):pp. 387–395。Yunji Zhou、R. A. M. 和 Thomas, L.（2020年）。重叠有限且模型设定错误时的倾向评分加权。医学研究统计方法，29(12):3721–3756。PMID: 32693715。Zubizarreta, J. R.（2015年）。用于不完整结局数据估计、能够平衡协变量的稳定权重。美国统计学会会刊，110(511):910–922。

补充材料

本补充材料与主稿配套，分为两个部分。A节提供主文中呈现的主要模拟结果的补充表格和图。B节展示次级模拟的图，这些模拟考察了不同样本量下的估计量性能，每次模拟均基于 10,000 个重复数据集。

A节：主要模拟的附加结果

表

A1. 超出 $[0.05, 0.95]$范围的倾向评分比例 A2. 通过遗漏变量导致模型设定错误时的均方根误差和偏差 A3. 通过遗漏变量导致模型设定错误时有限样本误差的均值和 MAV A4. 图 1 面板 (b)中未显示区间的缩放相对 MAV A5. 图2 面板 (b)中未显示区间的缩放相对 MAV A6. 图 A1 面板 (b)中未显示区间的缩放相对 MAV A7. 图 A2 面板 (b)中未显示区间的缩放相对 MAV

图

A1. 在遗漏变量导致模型设定错误时， R_i^e 和 R_i^y 的缩放相对 MAV A2. 在遗漏变量导致模型设定错误时， $R_i^e R_i^y$ 的缩放相对 MAV A3. 图 1 和图 A1 未显示的面板 (c) 和 (d) 中， R_i^e 和 R_i^y 的缩放相对 MAV A4. 图 2 和图 A2 未显示的面板 (c) 和 (d) 中， $R_i^e R_i^y$ 的缩放相对 MAV A5. 按治疗组划分的真实倾向评分分布 A6. 真实倾向评分分布

B节：不同样本量下的次级模拟结果

图

B1. 不同样本量下不同估计量的绝对偏差 B2. 不同样本量下不同估计量的均方根误差 B3. 函数形式错误导致模型设定错误时双重稳健估计量的绝对偏差 B4. 函数形式错误导致模型设定错误时双重稳健估计量的均方根误差 B5. 遗漏变量导致模型设定错误时双重稳健估计量的绝对偏差 B6. 遗漏变量导致模型设定错误时双重稳健估计量的均方根误差

A 主模拟的其他结果

表 A1：倾向评分的比例 在各种模拟设置下，倾向评分落在 (0.05,0.95) 范围之外的比例 .

治疗患病率	d	α_0	倾向评分落在 (0.05,0.95) 范围之外的比例（%）
0.4	1	−0.05	0
	3	0.37	36
0.1	1	−2.13	36
	3	−3.16	68

数值基于 100,000 的大样本量进行模拟。

表 A2：倾向得分模型和结果模型中有一个设定错误，或两者均设定错误时各估计量的均方根误差和偏差。模型设定错误是由遗漏变量造成的（遗漏 X_3 和X_4）。

估计量		设定错误模型	治疗患病率 $= 0.4$ $d=1$		$d=3$		治疗患病率 $= 0.1$ $d=1$		$d=3$	
			RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
双-稳健	不截尾	PS	0.11	0.00	0.22	0.00	0.23	0.00	0.95	−0.01
		结果	0.12	0.00	1.17	−0.02	0.34	−0.03	4.85	−0.36
		Both	0.39	−0.36	1.05	−1.00	0.49	−0.37	1.80	−1.14
	$\delta = 0.05$	PS	0.11	0.00	0.15	0.00	0.20	0.00	0.27	−0.01
		结果	0.11	0.00	0.18	−0.01	0.22	−0.02	0.28	−0.04
		Both	0.39	−0.36	1.00	−0.98	0.46	−0.37	1.00	−0.95
IPW	不截尾	PS	0.41	−0.37	1.40	−1.28	0.70	−0.41	2.42	−2.14
	$\delta = 0.05$	PS	0.41	−0.36	1.03	−1.00	0.48	−0.37	1.02	−0.96
结果回归		结果	0.39	−0.36	0.95	−0.93	0.47	−0.37	1.34	−1.20
重叠加权		PS	0.39	−0.36	0.98	−0.97	0.43	−0.36	0.96	−0.93

修剪结果（ δ）　　$= 0.1$ 与 $\delta = 0.05$ 的结果相似，因此　　未予显示。

表 A3：当倾向得分模型和结果模型之一因遗漏变量（遗漏 X_3 和 X_4）而设定错误时，不同（子）总体中有限样本误差的均值和平均绝对值（MAV）。

模型设定		总体	治疗患病率 $= 0.4$ $d=1$		$d=3$		治疗患病率 $= 0.1$ $d=1$		$d=3$	
			平均	MAV	平均	MAV	平均	MAV	平均	MAV
设定错误模型	倾向评分模型	总体	0.00	0.07	0.00	0.15	0.00	0.17	−0.01	0.46
		Tail	0.00	0.29	−0.02	0.58	0.00	0.47	−0.02	0.64
		Bulk	0.00	0.08	0.00	0.11	0.00	0.20	−0.01	0.46
	结果模型	总体	0.00	0.08	−0.02	0.45	−0.03	0.24	−0.36	1.17
		Tail	−0.03	0.59	−0.12	2.96	−0.09	1.10	−0.26	2.59
		Bulk	0.00	0.11	−0.01	0.16	−0.02	0.27	−0.29	1.01

表 A4：图 1 面板 (b) 中未显示区间的缩放相对 MAV。

PS 区间	残差类型	设定错误的模型	缩放相对 MAV
(0.00, 0.01]	结果	结果	1.43
		PS	0.01
	PS	PS	0.38
(0.01, 0.02]	结果	结果	0.37
		PS	0.04
	PS	PS	0.11
(0.02, 0.03]	结果	结果	0.19
		PS	0.03
	PS	PS	0.06
(0.98, 0.99]	结果	结果	0.20
		PS	0.01
	PS	PS	0.04
(0.99, 1.00)	结果	结果	0.33
		PS	0.01
	PS	PS	0.05

表 A5：图2面板 (b) 中未显示区间的缩放相对 MAV。

PS 区间	设定错误的模型	
	倾向评分模型	结果模型
(0.00, 0.01]	0.05	0.53
(0.01, 0.02]	0.04	0.16
(0.02, 0.03]	0.03	0.09
(0.98, 0.99]	0.01	0.10
(0.99, 1.00)	0.01	0.18

表 A6：图 A1 面板 (b) 中未显示的区间的缩放相对 MAV。

PS 区间	残差类型	设定错误的模型	缩放相对 MAV
(0.00, 0.01]	结果	结果	0.63
		PS	0.01
	PS	PS	0.29
(0.01, 0.02]	结果	结果	0.31
		PS	0.03
	PS	PS	0.12
(0.02, 0.03]	结果	结果	0.18
		PS	0.03
	PS	PS	0.07
(0.98, 0.99]	结果	结果	0.12
		PS	0.01
	PS	PS	0.05
(0.99, 1.00)	结果	结果	0.13
		PS	0.00
	PS	PS	0.08

表 A7：图 A2 面板 (b) 中未显示区间的缩放相对 MAV。

PS 区间	设定错误的模型	
	倾向评分模型	结果模型
(0.00, 0.01]	0.05	0.36
(0.01, 0.02]	0.03	0.14
(0.02, 0.03]	0.03	0.09
(0.98, 0.99]	0.01	0.06
(0.99, 1.00)	0.01	0.09

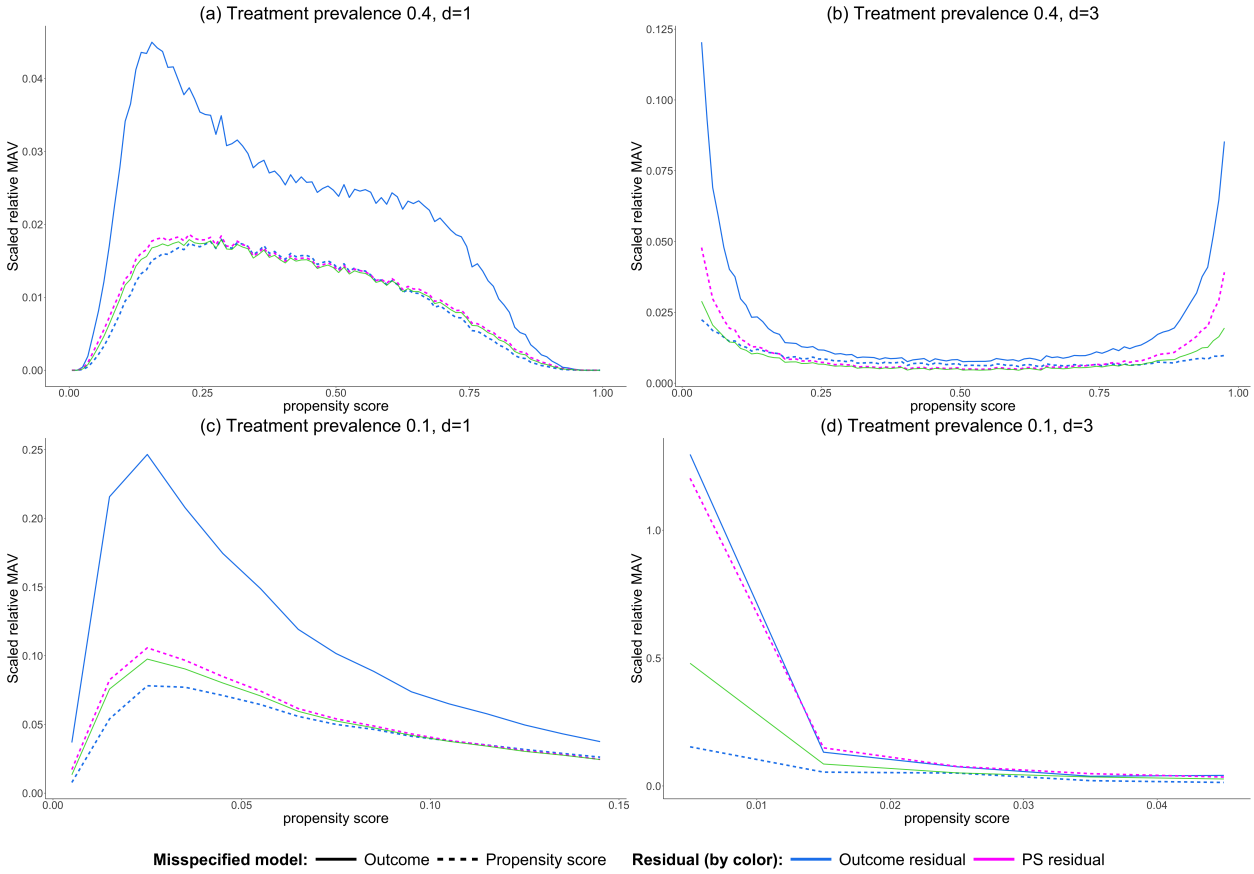


图 A1： R_i^e 和 R_i^y 的缩放相对 MAV。模型设定错误由遗漏变量引起（遗漏X_3 和 X_4）。绿色线（$\Phi(l)$）表示相对尺度没有变化。面板 (b) 尾部的较大值以及面板 (c) 和 (d) 的其余部分见 表 A6 和图 A3。

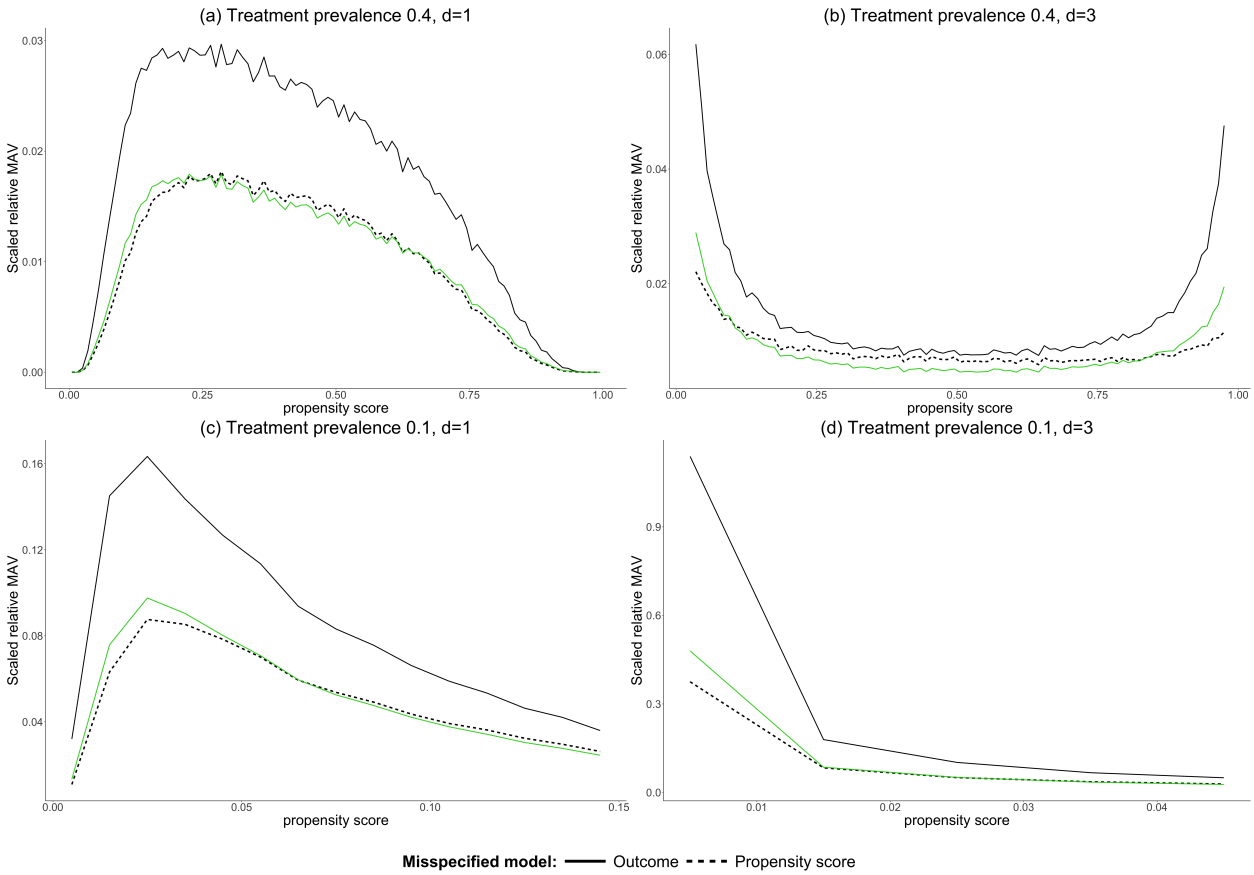


图 A2： $R_i^e R_i^y$ 的缩放相对 MAV。遗漏变量（遗漏 X_3和 X_4）会导致模型设定错误。绿色线（$\Phi(l)$）表示相对尺度没有变化。面板 (b) 尾部的大值以及面板 (c) 和 (d) 其余部分的结果见表 A7 和图 A4。

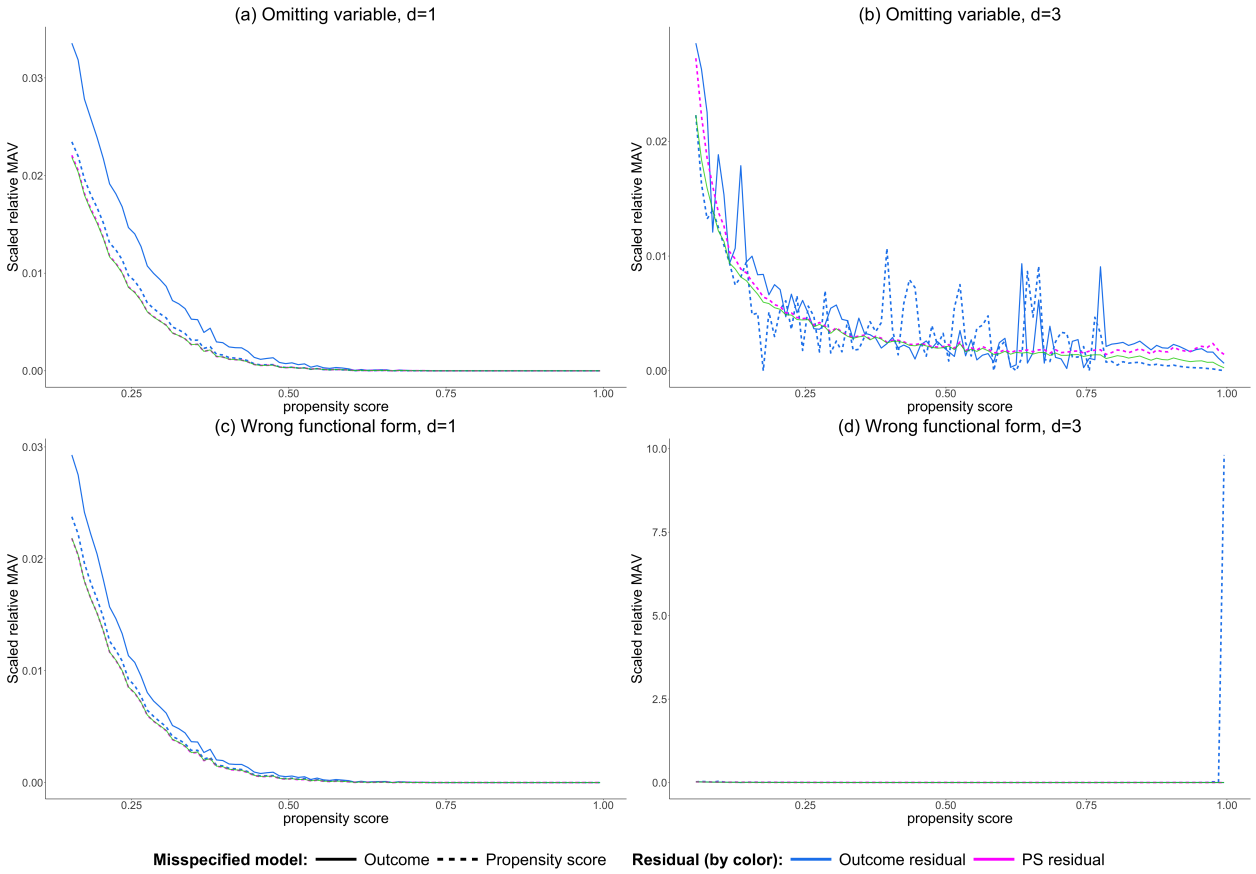


图 A3：未在图 1 和图 A1 中展示的面板 (c) 和 (d) 中，R_i^e 和 R_i^y 的缩放相对 MAV。$d = 1$ 对应图 1 和图 A1 的面板 (c)，$d = 3$ 对应图 1 和图 A1 的面板 (d)。

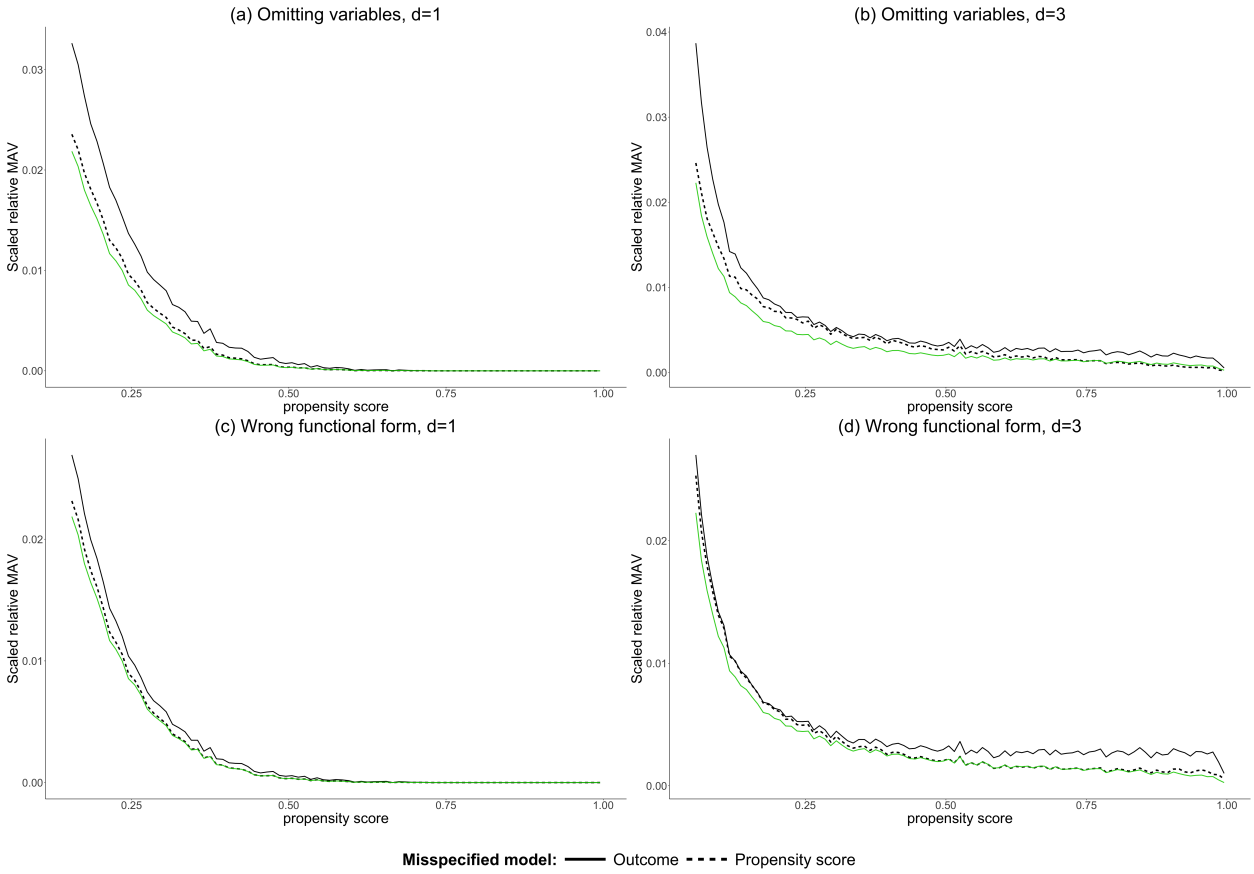


图 A4： $R_i^e R_i^y$ 的缩放相对 MAV，展示图2和图 A2中未显示的其余面板 (c) 和 (d)。 $d = 1$ 对应图2和图 A2的面板 (c)， $d = 3$ 对应图2和图 A2的面板 (d)。

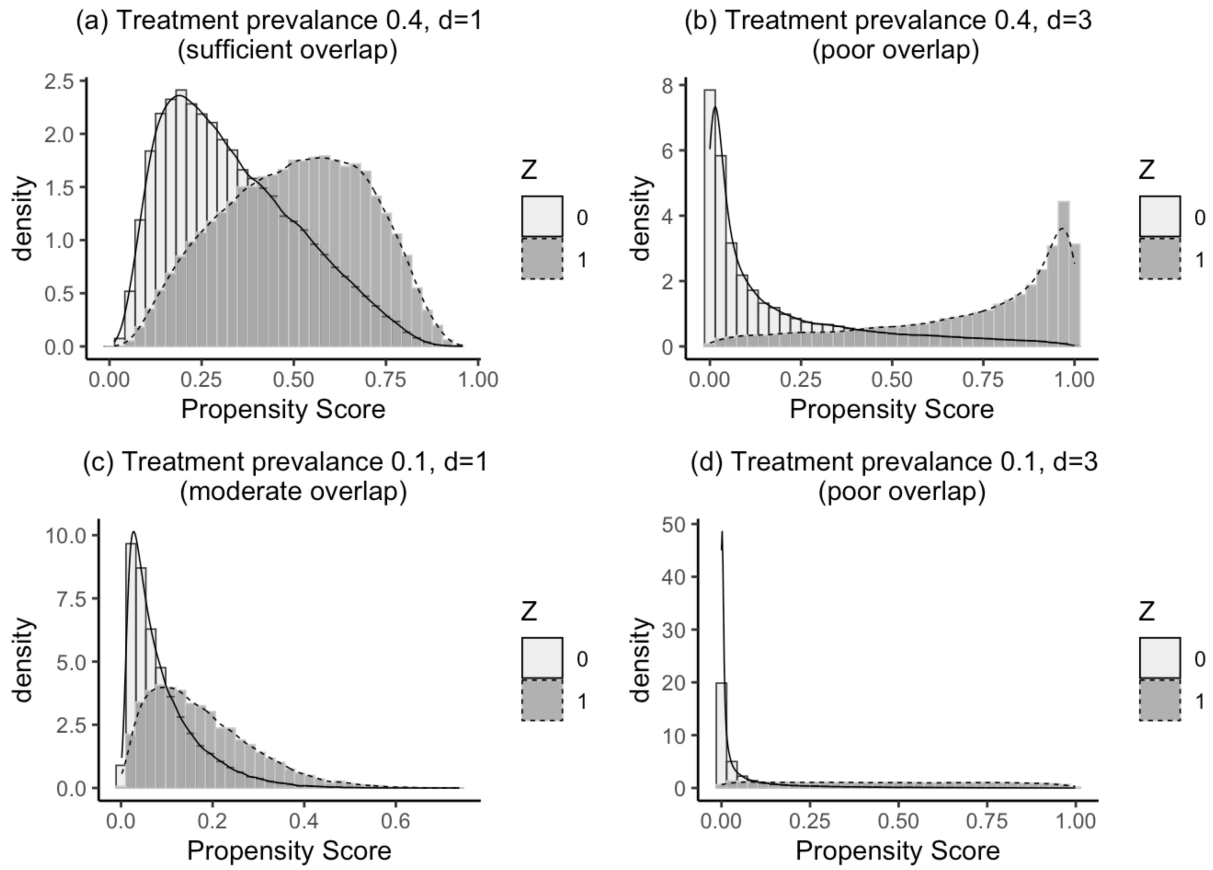


图 A5：按治疗组划分的真实倾向得分分布（基于样本量为 100,000 的大样本数据集进行评估）。

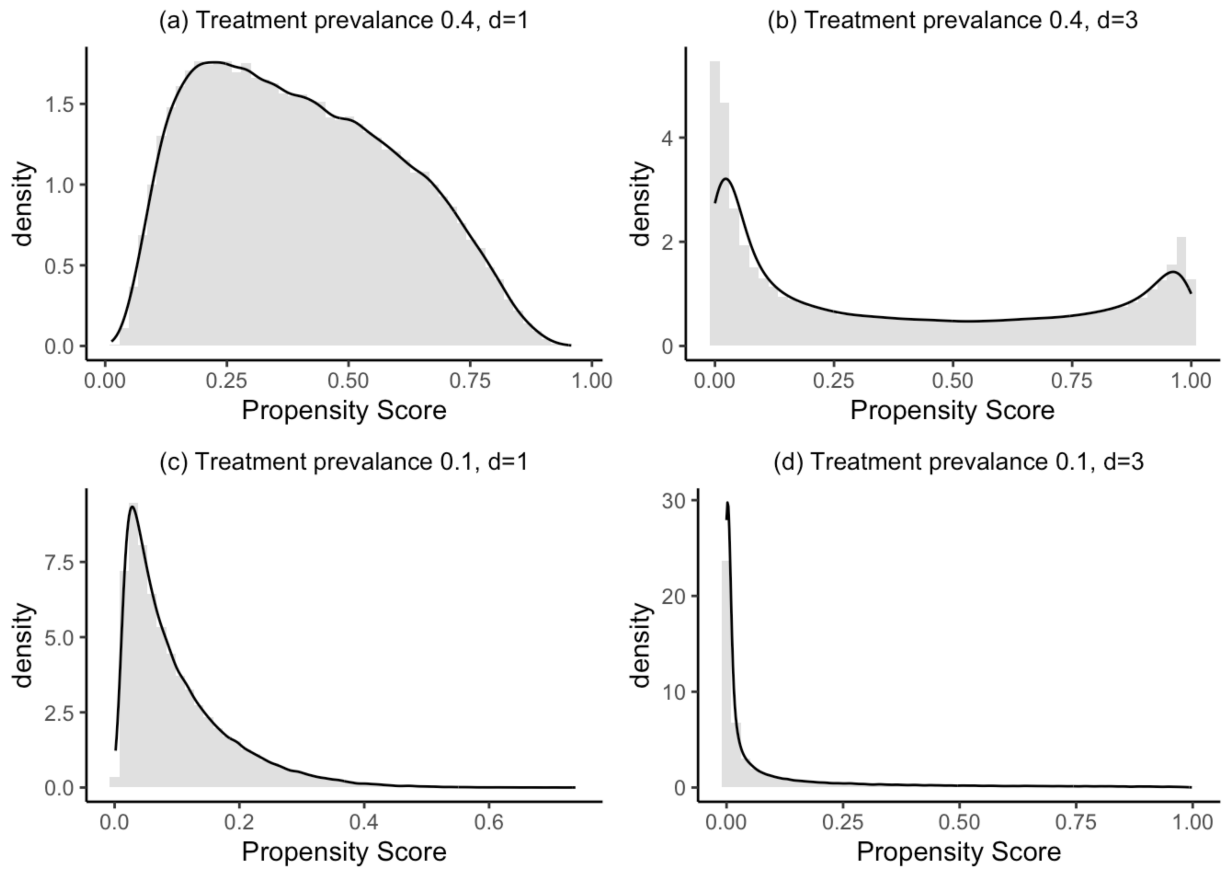


图 A6：分布　　真实倾向得分的分布（通过大小为 100,000 的大样本数据集进行评估）　　.

B 不同样本量下的次要模拟结果

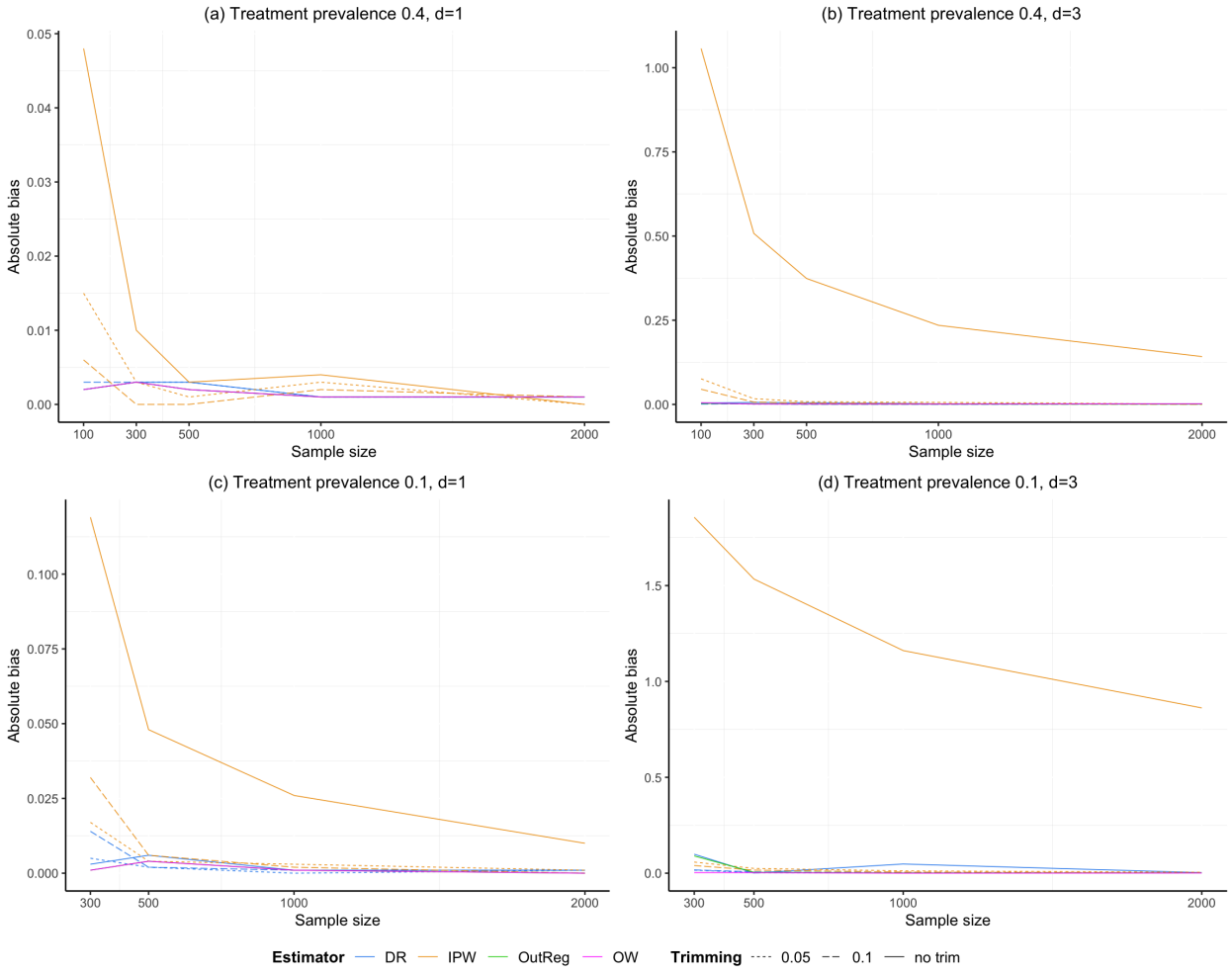
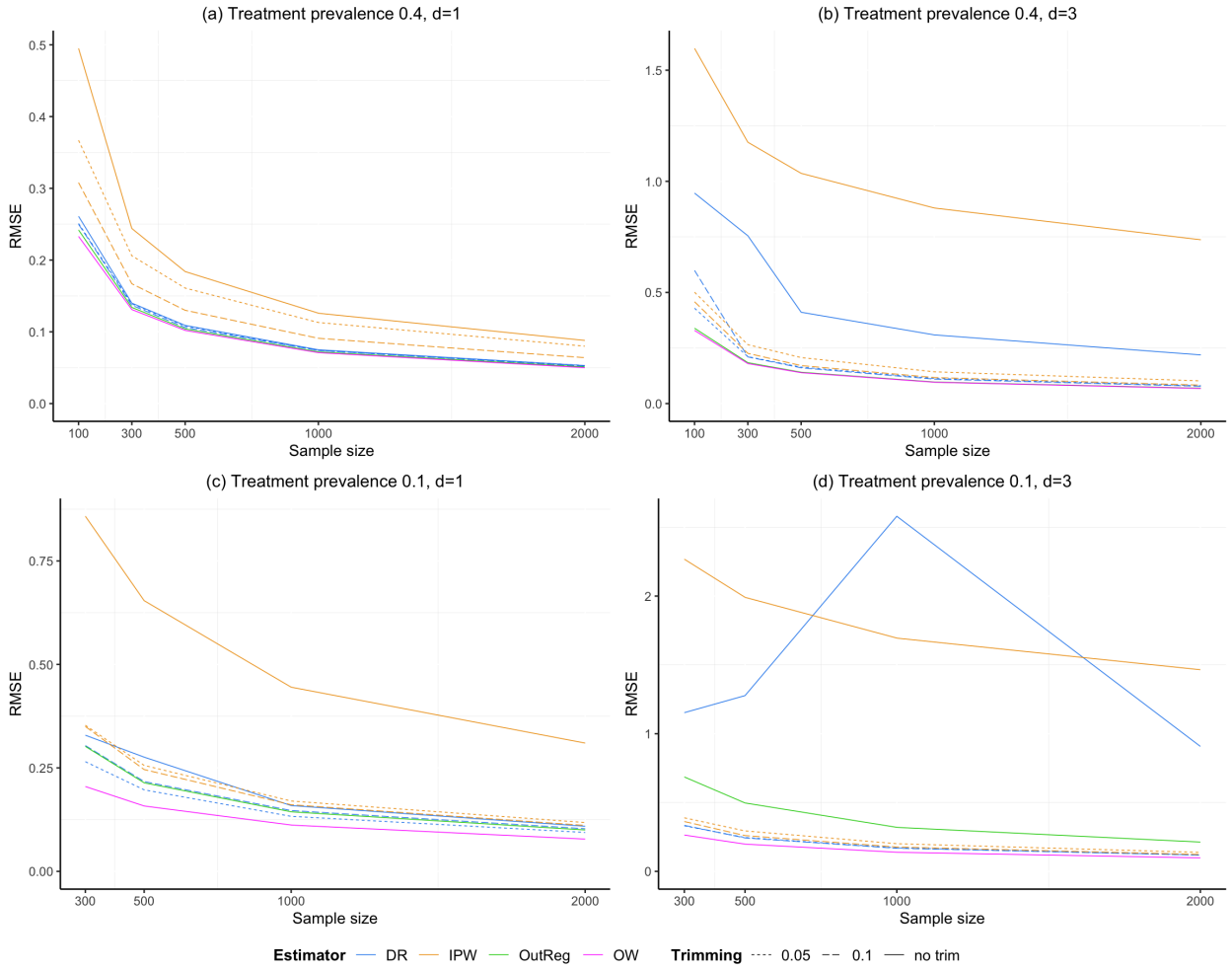


图 B1：不同样本量下各估计量的绝对偏差。



Figu关于 B2：不同样本量下各估计量的均方根误差（均方根误差） e.

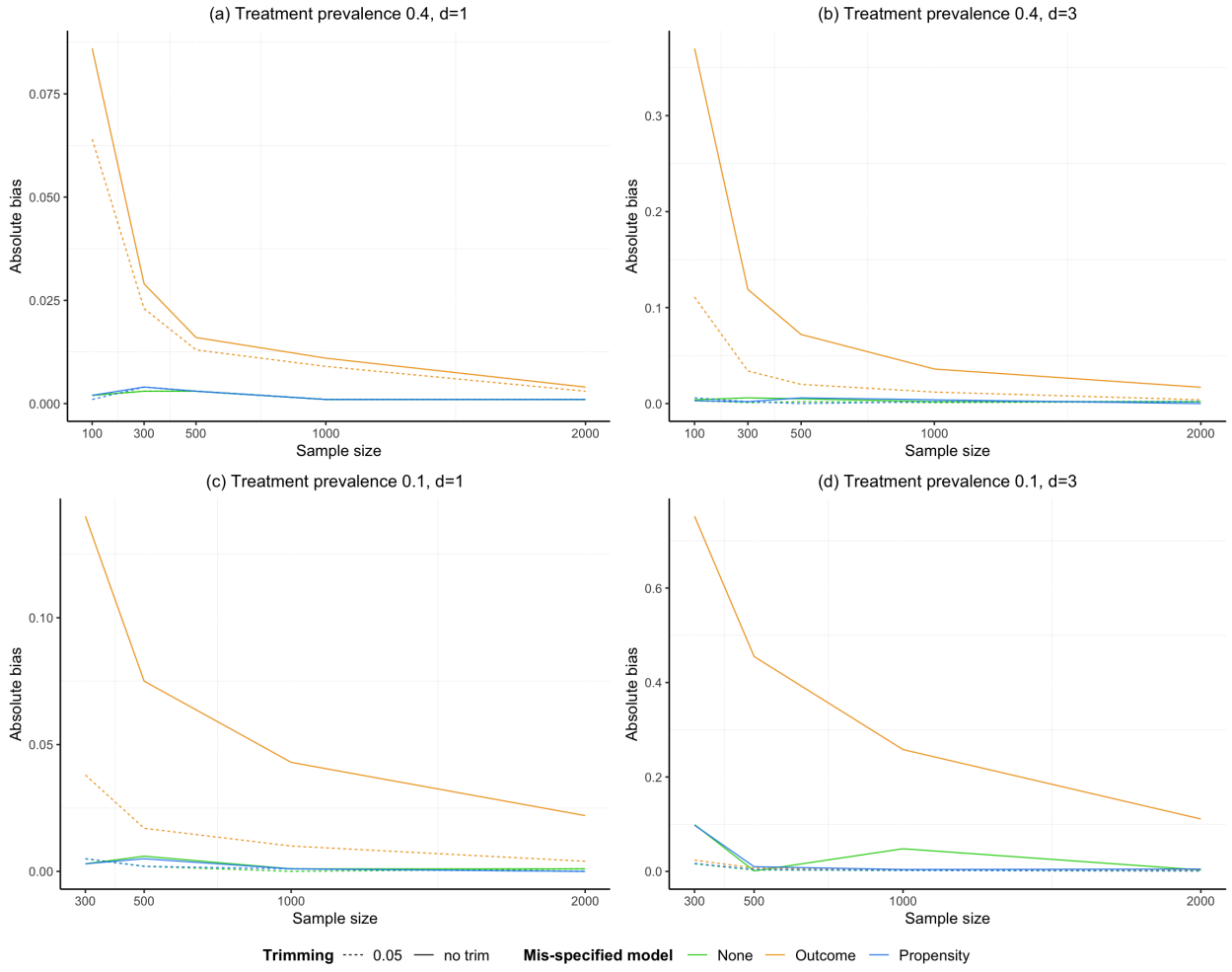
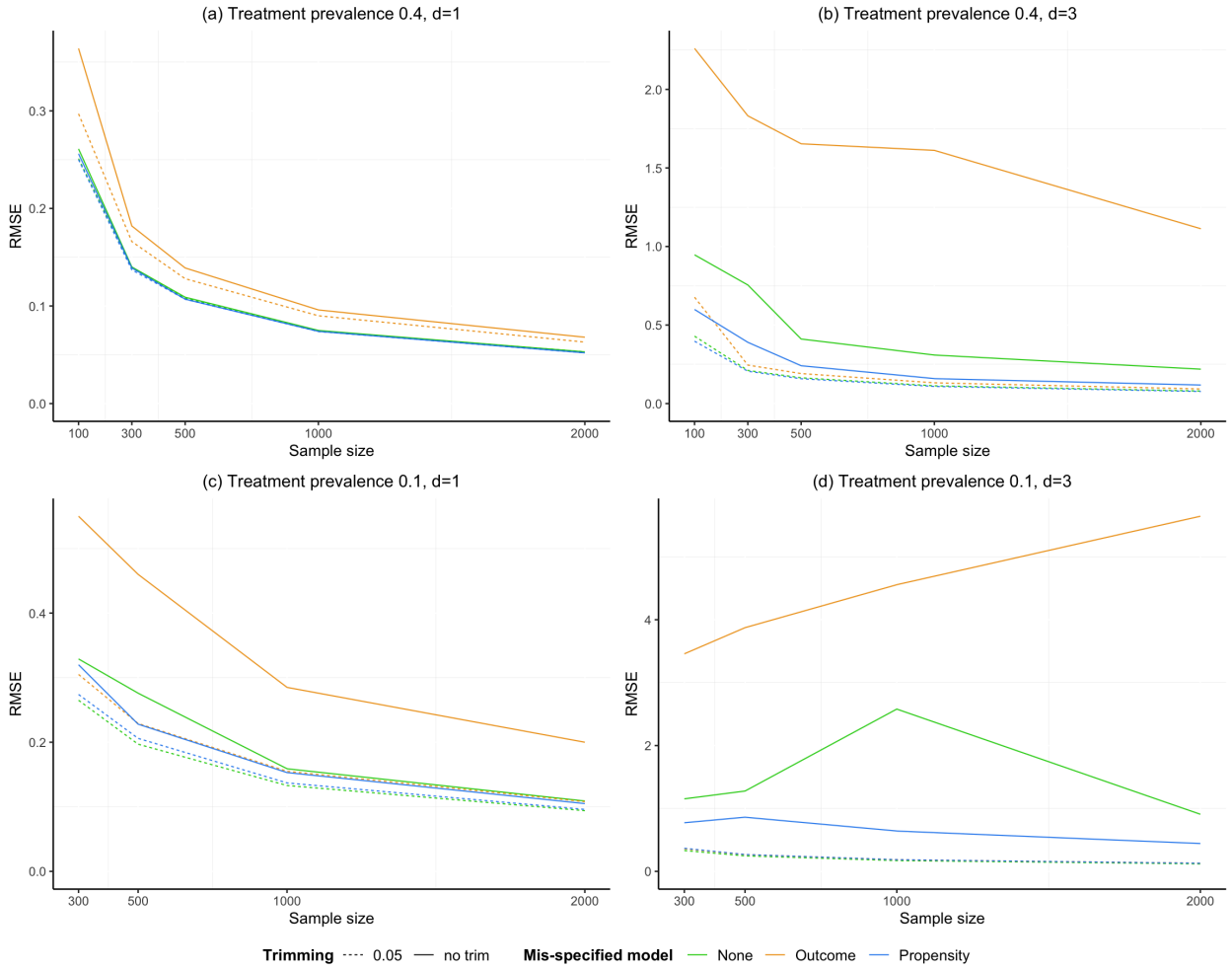


图 B3：当倾向评分模型或结果模型之一设定错误时，双重稳健估计量（未修剪和在 0.05 处修剪倾向得分）的绝对偏差。模型设定错误是由函数形式错误导致的（将 X_3 误认为 X_3^2）。



图B4：倾向评分或结果模型之一设定错误时，双重稳健估计量（未修剪及在 0.05 处修剪倾向评分）的均方根误差。模型设定错误是由函数形式错误造成的（将X_3 误认为 X_3^2）。

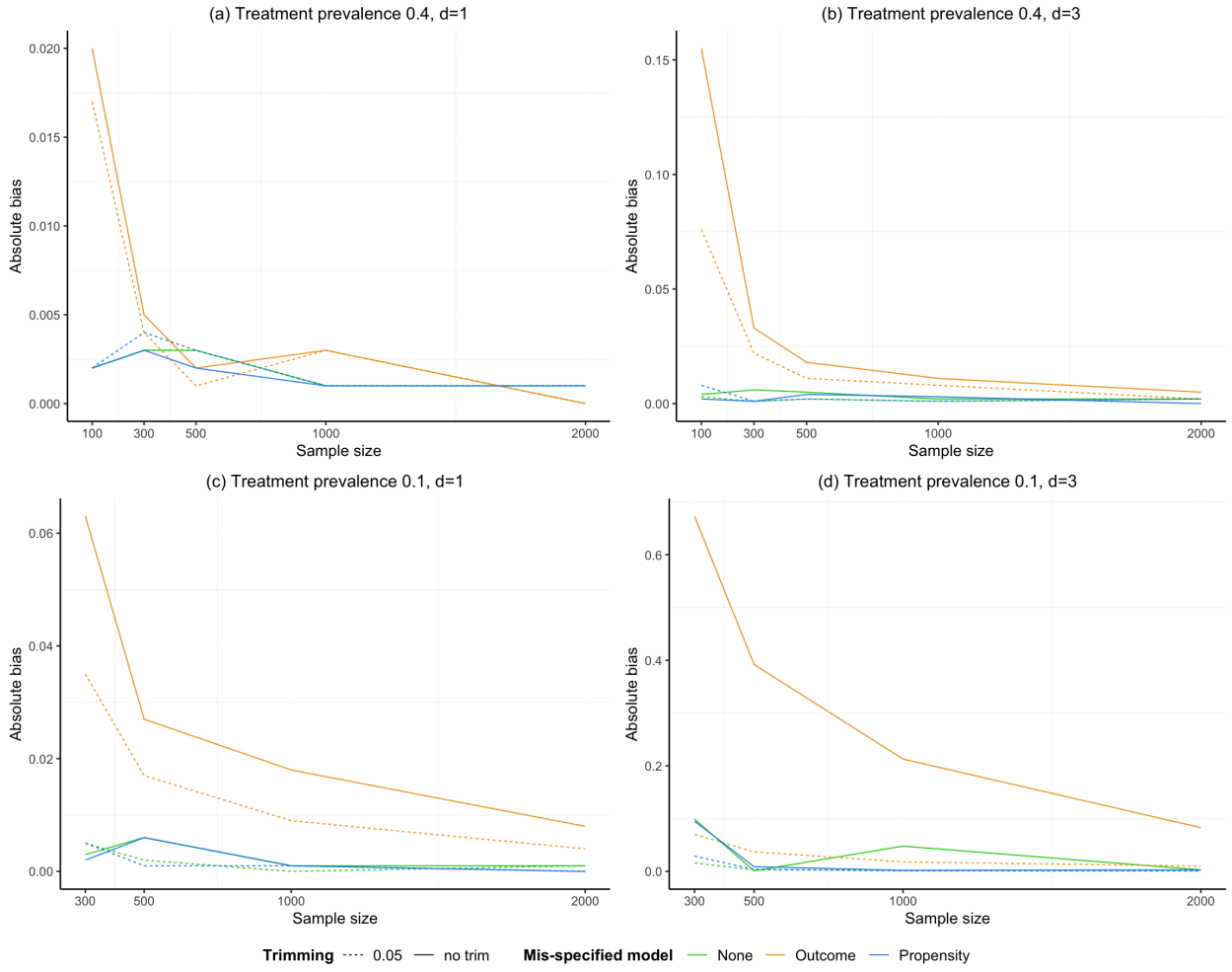
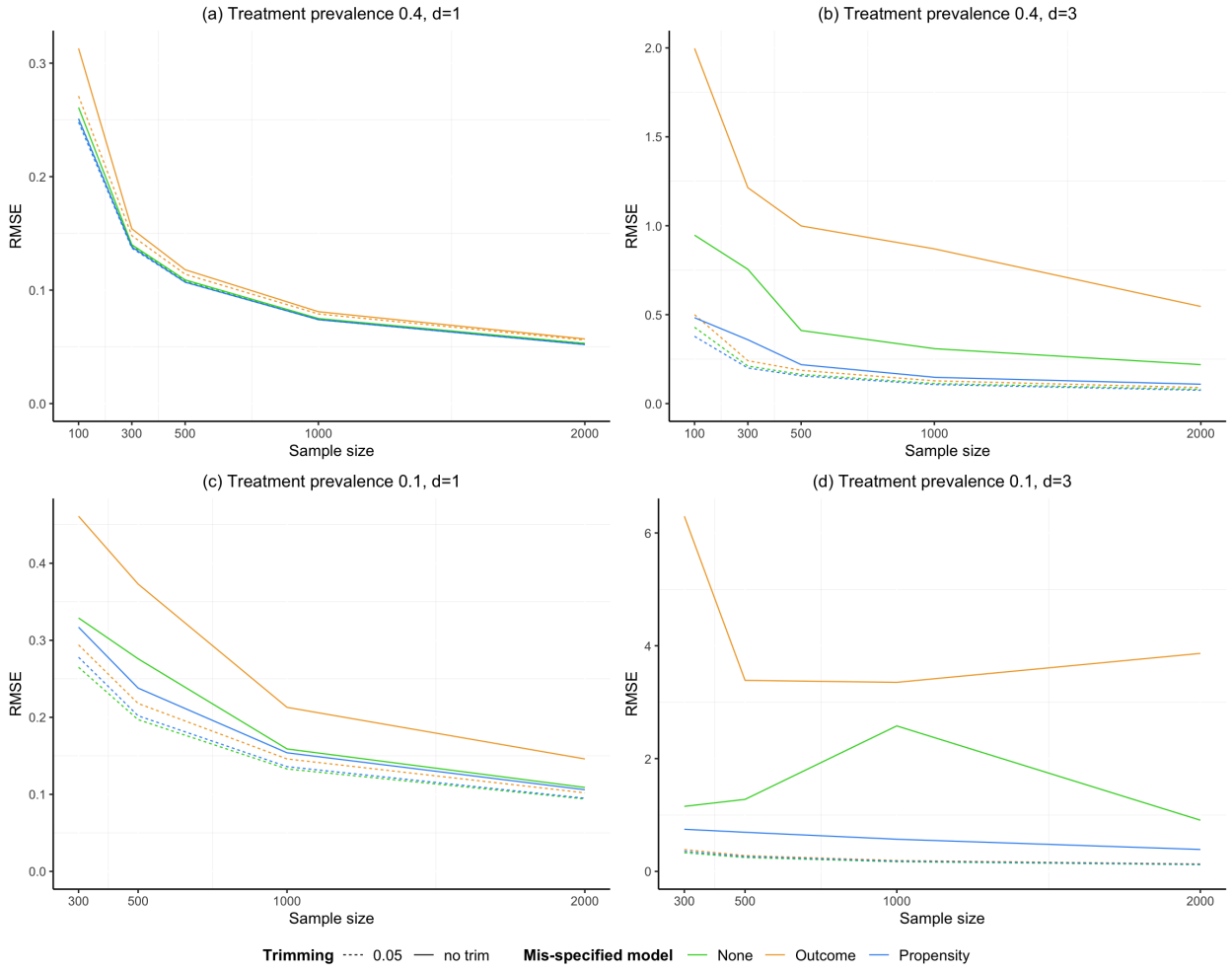


图 B5：当倾向评分模型或结果模型之一存在模型设定错误时，双重稳健估计量在不进行倾向得分截尾和以 0.05 为阈值进行倾向得分截尾时的绝对偏差。模型设定错误由遗漏变量导致（遗漏 X_3 和 X_4）。



图B6：当倾向评分模型或结果模型之一设定错误时，双重稳健估计量（倾向得分截尾阈值为 0.05，未截尾和截尾）的均方根误差。通过遗漏变量（遗漏 X_3和 X_4）造成模型设定错误。