综述：观察性研究中因果效应界的识别与估计

史长浩

中国人民大学统计学院

2022 年 12 月 20 日

中国人民大學
RENMIN UNIVERSITY OF CHINA

观察性研究

- 随机化试验是因果推断的金标准，然而在很多问题中，出于伦理或成本的考虑，无法控制对受试者的处理分配，这时研究者只能进行观察性研究。
- 如果没有不可检验的假定，根据观察性研究进行因果推断是不可能的，这是因为在观察性研究中，病例组和对照组中的混杂无法得到有效的平衡和控制。因此，根据观察性研究进行因果推断是一个非常具有挑战的课题。
- 因为观察性研究中的因果参数不可识别，所以很多研究者考虑为其寻找一个可识别的界，以实现因果参数的“部分识别”。本文回顾了 5 篇相关的工作，其中涉及的方法和观点在很多其他问题中也颇有启发。

病例对照研究

表 1: 2×2 列联表：病例对照研究（总体的分布）。

	$Y=1$	$Y=0$
$X=1$	$P(X=1 \mid Y=1)$	$P(X=1 \mid Y=0)$
$X=0$	$P(X=0 \mid Y=1)$	$P(X=0 \mid Y=1)$

- 在病例对照研究中，上述列联表中的参数皆是可识别的，如果知道疾病在人群中的发病率 $P(Y=1)$，则联合分布 $P(X=x, Y=y)$ 可识别。
- 这是否表明：问题的关键仅仅在于 $P(Y=1)$ 的识别与估计？

病例对照研究中的因果参数

- 事实上，我们关心的因果参数为 CRR(Causal Relative Risk) 和 CRD(Causal Risk Difference)：

$$\text{CRR} = \frac{P(Y(1)=1)}{P(Y(0)=1)}, \tag{1}$$

$$\text{CRD} = P(Y(1)=1) - P(Y(0)=1). \tag{2}$$

- 利用贝叶斯公式分解包含潜在结果的概率之后，又会出现两个不可识别的参数：$P(Y(1)=1|X=0)$ 和 $P(Y(0)=1|X=1)$，这两个参数连同 $P(Y=1)$ 一同导致了问题的复杂性。

不依从问题中处理效应的界 [Balke and Pearl, 1994, 1997]

- 作者针对试验性研究中的不依从问题，即处理分配是随机的，但受试者却不一定完美服从分配的问题，建立了平均处理效应的非参数界。

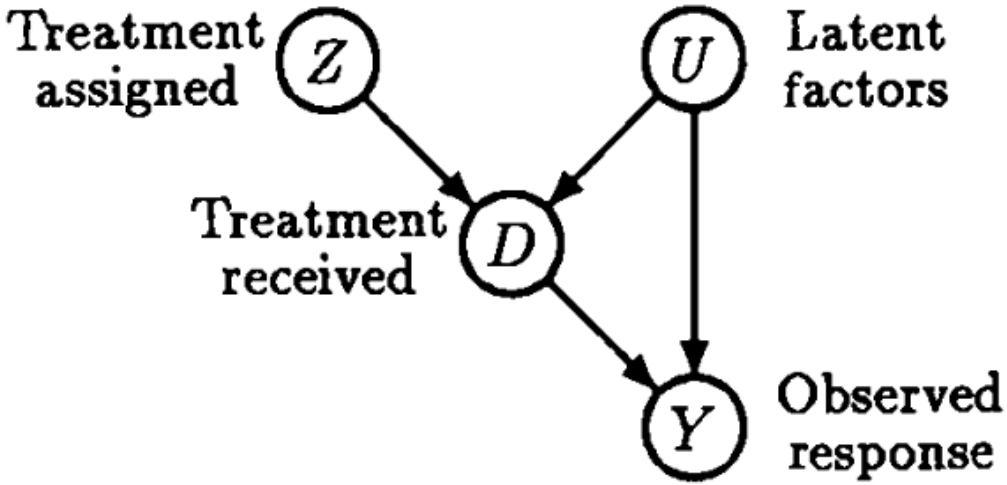


图 1: 随机临床试验中部分依从问题的因果图表示。

不依从问题中处理效应的界 [Balke and Pearl, 1994, 1997]

- 这篇文章本来与我们的主题并不相关，然而，这篇文章却是最早研究“因果界”的工作之一。作者提出的用线性规划求解因果界的方法在之后的研究中被广泛使用。

$$
\begin{aligned}
\min \quad & q_{01} + q_{11} + q_{21} + q_{31} - q_{02} - q_{12} - q_{22} - q_{32}. \\
\text{s.t.} \quad & \sum_{j=0}^{3}\sum_{k=0}^{3} q_{jk} = 1, \\
& \bar{P}\vec{q} = \vec{p}, \\
& q_{jk} \geq 0 \text{ for } j,k \in \{0,1,2,3\}.
\end{aligned} \tag{3}
$$

不依从问题中处理效应的界 [Balke and Pearl, 1994, 1997]

- 其中

$$
\begin{aligned}
p_{00.0} &= P\left(y_0, d_0 \mid z_0\right), \quad p_{00.1} = P\left(y_0, d_0 \mid z_1\right), \\
p_{01.0} &= P\left(y_0, d_1 \mid z_0\right), \quad p_{01.1} = P\left(y_0, d_1 \mid z_1\right), \\
p_{10.0} &= P\left(y_1, d_0 \mid z_0\right), \quad p_{10.1} = P\left(y_1, d_0 \mid z_1\right), \\
p_{11.0} &= P\left(y_1, d_1 \mid z_0\right), \quad p_{11.1} = P\left(y_1, d_1 \mid z_1\right).
\end{aligned}
$$

不依从问题中处理效应的界 [Balke and Pearl, 1994, 1997]

- 及

$$
\begin{aligned}
p_{00.0} &= q_{00} + q_{01} + q_{10} + q_{11}, \\
p_{01.0} &= q_{20} + q_{22} + q_{30} + q_{32}, \\
p_{10.0} &= q_{02} + q_{03} + q_{12} + q_{13}, \\
p_{11.0} &= q_{21} + q_{23} + q_{31} + q_{33}, \\
p_{00.1} &= q_{00} + q_{01} + q_{20} + q_{21}, \\
p_{01.1} &= q_{10} + q_{12} + q_{30} + q_{32}, \\
p_{10.1} &= q_{02} + q_{03} + q_{22} + q_{23}, \\
p_{11.1} &= q_{11} + q_{13} + q_{31} + q_{33}.
\end{aligned}
$$

不依从问题中处理效应的界 [Balke and Pearl, 1994, 1997]

- 算得的结果为

$$\mathrm{ACE}(D \to Y) \geq \max \left\{ \begin{array}{c} p_{11.1} + p_{00.0} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{11.0} - p_{11.1} - p_{10.1} - p_{01.0} - p_{10.0} \\ p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1} \\ -p_{01.1} - p_{10.1} \\ -p_{01.0} - p_{10.0} \\ p_{00.1} - p_{01.1} - p_{10.1} - p_{01.0} - p_{00.0} \\ p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} \end{array} \right\},$$

不依从问题中处理效应的界 [Balke and Pearl, 1994, 1997]

- 及

$$\mathrm{ACE}(D \to Y) \leq \min \left\{ \begin{array}{c} 1 - p_{01.1} - p_{10.0} \\ 1 - p_{01.0} - p_{10.1} \\ -p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0} \\ -p_{01.1} + p_{11.1} + p_{00.1} + p_{01.0} + p_{00.0} \\ p_{11.1} + p_{00.1} \\ p_{11.0} + p_{00.0} \\ -p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0} \\ -p_{10.0} + p_{11.0} + p_{00.0} + p_{11.1} + p_{10.1} \end{array} \right\}.$$

- 类似的技术和思想在接下来的几篇文章中会多次出现。

病例对照研究中的因果界 [Kuroki, Cai and Geng, 2010]

- 在本文中，作者发展了 [Balke and Pearl, 1994, 1997] 的工作，在不作任何额外假设的情况下，导出了病例对照研究中因果参数 CRD 和 CRR 的非参数界：

$$\min\{-\operatorname{pr}(x_0 \mid y_1), -\operatorname{pr}(x_1 \mid y_0)\} \leqslant \mathrm{CRD}(x;y), \tag{4}$$

$$\mathrm{CRD}(x;y) \leqslant \max\{\operatorname{pr}(x_1 \mid y_1), \operatorname{pr}(x_0 \mid y_0)\}. \tag{5}$$

$$0 \leqslant \mathrm{CRR}(x;y) \leqslant \infty. \tag{6}$$

- 其中 CRR 的非参数界不比其天然的取值域更窄，这表明若不做额外假设，数据无法提供对 CRR 有效的信息。
- 在单调性假设下，CRD 和 CRR 的下界分别取 0 和 1。

病例对照研究中的因果界 [Kuroki, Cai and Geng, 2010]

- 在计算 CRR 的非参数界时，作者援引了一个分式线性规划的定理，将一个分式非线性约束转化为线性约束，使之前的方法得以适用。
- 若假设 $0 < a \leqslant \mathrm{pr}(y_1)$，并作单调性假设，可得 CRR 有意义的因果界：

$$1 \leqslant \mathrm{CRR}(x; y) \leqslant \frac{\mathrm{pr}(x_0 \mid y_0)}{\mathrm{pr}(x_0 \mid y_1)\, a} + \frac{\mathrm{pr}(x_1 \mid y_0)}{\mathrm{pr}(x_0 \mid y_1)}. \tag{7}$$

- 作者还研究了 3 种缺失机制下因果界的估计问题，暂不在此陈述。此外上述的结果可以使用简单的代数推理得到，并不需要繁琐的规划解法。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 在本文中，作者从一个更普适的观点看待此前的问题，作者认为：病例对照研究、队列研究等观察性研究本质上都是一种所谓的“结果依赖型抽样（outcome-dependent sampling）”的特例。作者强调“选择（Selection）”这一过程，用 $S = 1$ 表示个体入选研究，只有给定 $S = 1$ 条件下的分布是可从观测数据中识别的。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 以之前讨论的病例对照研究为例，我们认为 $P(X = x \mid Y = y)$ 是可从数据中识别的分布，但在本文的观点下，$P(X = x \mid Y = y, S = 1)$ 才是可识别的，只有当“选择”仅依赖于 Y 时，两个概率才是相等的，如下图。

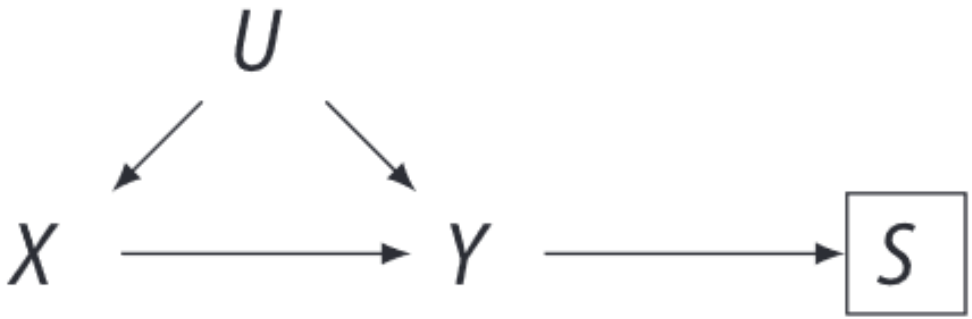
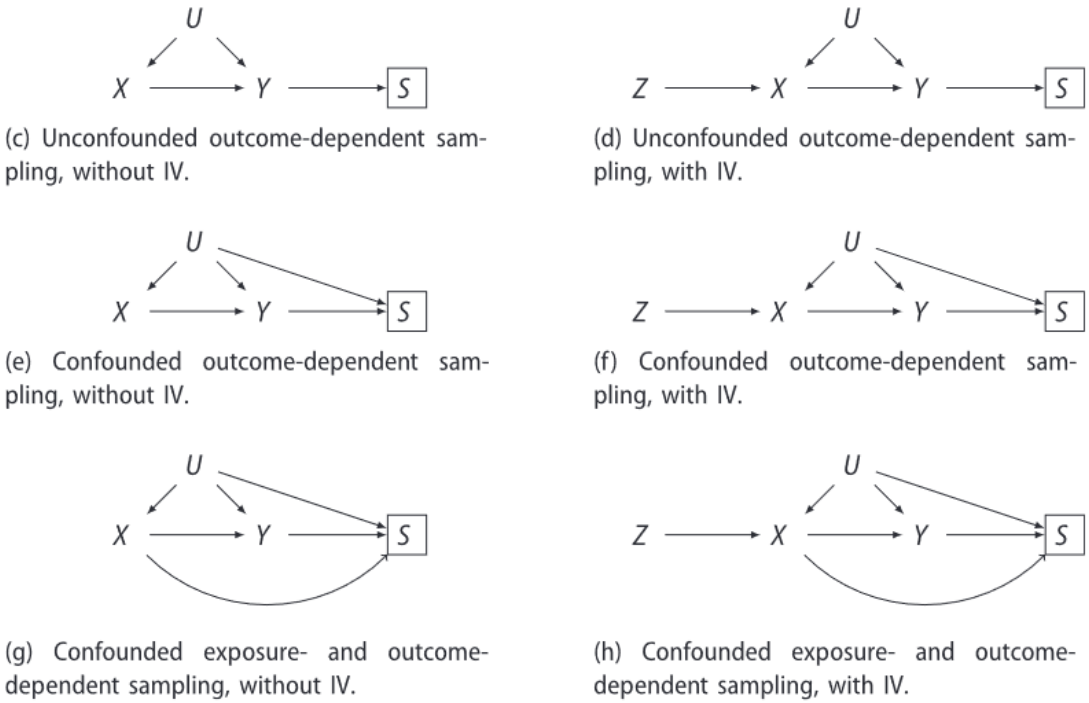


图 2: 因果图：仅依赖 Y 的采样。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 类似地，作者又讨论了 $S = 1$ 依赖于混杂 U 和处理 X 的情况，并考虑了存在可用的工具变量时，相应的因果效应界，作者关心的因果参数为 $\text{CRD} = P(Y(1) = 1) - P(Y(0) = 1)$。



(c) Unconfounded outcome-dependent sampling, without IV.

(d) Unconfounded outcome-dependent sampling, with IV.

(e) Confounded outcome-dependent sampling, without IV.

(f) Confounded outcome-dependent sampling, with IV.

(g) Confounded exposure- and outcome-dependent sampling, without IV.

(h) Confounded exposure- and outcome-dependent sampling, with IV.

图 3: 因果图：结果依赖型采样。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 因为 $S = 1$ 的存在，约束不再是线性的，因此此前的线性规划方法不能直接使用，作者的办法是分两步走：首先将目标函数合理地拆成可观测部分和不可观测部分，可观测部分直接使用前人得到的界；下一步集中精力处理不可观测部分（不是很容易，作者写了 29 页证明），最后把两部分整合起来就得到最后的结果。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 但这样得到的结果也越来越复杂...

$$\theta \geq \max \left\{ \begin{array}{l} p_{11.11} r_1 + p_{00.01} r_0 - 1 \\ p_{11.01} r_0 + p_{00.11} r_1 - 1 \\ (p_{11.01} - p_{10.01} - p_{01.01}) r_0 - (p_{11.11} + p_{01.11}) r_1 - \frac{B(0,0,1)}{1+B(0,0,1)} (1 - r_0) - (1 - r_1) \\ (p_{11.11} - p_{10.11} - p_{01.11}) r_1 - (p_{11.01} + p_{01.01}) r_0 - \frac{B(0,1,1)}{1+B(0,1,1)} (1 - r_1) - (1 - r_0) \\ -(p_{10.11} + p_{01.11}) r_1 - \max \left\{ \frac{1}{1+B(1,1,1)}, \frac{B(0,1,1)}{1+B(0,1,1)} \right\} (1 - r_1) \\ -(p_{10.01} + p_{01.01}) r_0 - \max \left\{ \frac{1}{1+B(1,0,1)}, \frac{B(0,0,1)}{1+B(0,0,1)} \right\} (1 - r_0) \\ (p_{00.11} - p_{10.11} - p_{01.11}) r_1 - (p_{10.01} + p_{00.01}) r_0 - \max \left\{ \frac{1}{1+B(1,1,1)}, -\frac{1-B(0,1,1)}{1+B(0,1,1)} \right\} (1 - r_1) - (1 - r_0) \\ (p_{00.01} - p_{10.01} - p_{01.01}) r_0 - (p_{10.11} + p_{00.11}) r_1 - \max \left\{ \frac{1}{1+B(1,0,1)}, -\frac{1-B(0,0,1)}{1+B(0,0,1)} \right\} (1 - r_0) - (1 - r_1) \end{array} \right\}$$

图 4: 因果图 (d) 对应的因果下界。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 但这样得到的结果也越来越复杂...

$$\theta \leq \min \left\{ \begin{array}{l} 1 - p_{10.11} r_1 - p_{01.01} r_0 \\ 1 - p_{10.01} r_0 - p_{01.11} r_1 \\ (p_{10.11} + p_{00.11}) r_1 + (p_{11.01} + p_{00.01} - p_{10.01}) r_0 + (1 - r_1) + \max \left\{ \frac{1-B(0,0,1)}{1+B(0,0,1)}, \frac{B(1,0,1)}{1+B(1,0,1)} \right\} r_0 \\ (p_{11.11} + p_{00.11} - p_{10.11}) r_1 + (p_{10.01} + p_{00.01}) r_0 + \max \left\{ \frac{1-B(0,1,1)}{1+B(0,1,1)}, \frac{B(1,1,1)}{B(1,1,1)} \right\} (1 - r_1) + (1 - r_0) \\ (p_{11.11} + p_{00.11}) r_1 + \max \left\{ \frac{1}{1+B(0,1,1)}, \frac{B(1,1,1)}{1+B(1,1,1)} \right\} (1 - r_1) \\ (p_{11.01} + p_{00.01}) r_0 + \max \left\{ \frac{1}{1+B(0,0,1)}, \frac{B(1,0,1)}{1+B(1,0,1)} \right\} (1 - r_0) \\ (p_{11.11} + p_{00.11} - p_{01.11}) r_1 + (p_{11.01} + p_{01.01}) r_0 + \max \left\{ \frac{1}{B(0,1,1)}, -\frac{1-B(1,1,1)}{1+B(1,1,1)} \right\} (1 - r_1) + (1 - r_0) \\ (p_{11.01} + p_{00.01} - p_{01.01}) r_0 + (p_{11.11} + p_{01.11}) r_1 + \max \left\{ \frac{1}{B(0,0,1)}, -\frac{1-B(1,0,1)}{1+B(1,0,1)} \right\} (1 - r_0) + (1 - r_1) \end{array} \right\}$$

图 5: 因果图 (d) 对应的因果上界。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

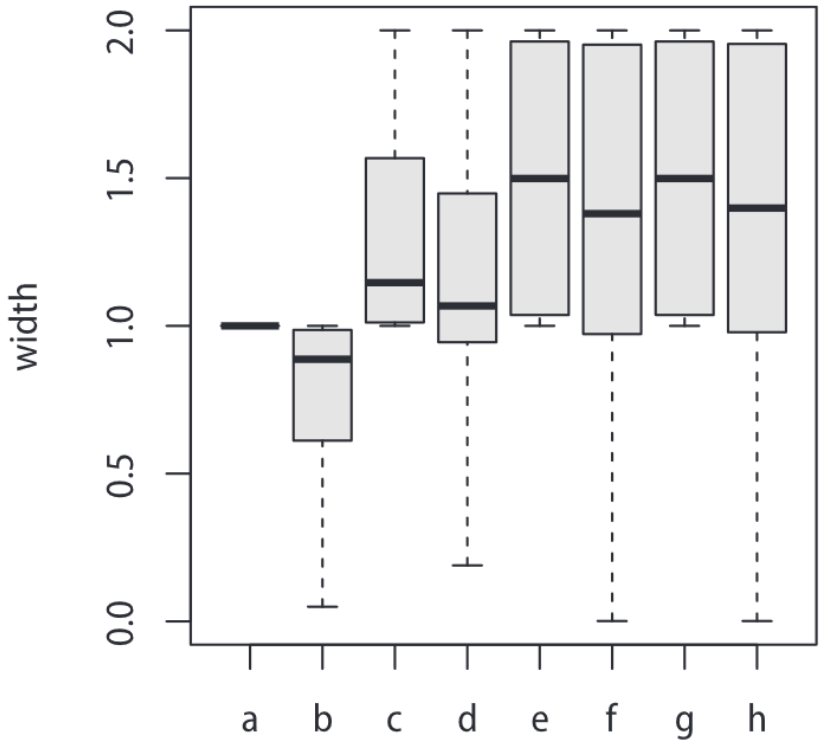


图 6: 数值模拟：(c)(d)(e)(f)(g)(h)6 种情形对应的因果界宽度。

结果依赖型采样的因果界 [Gabriel, Sachs and Sjölander, 2022]

- 值得指出的是，作者认为个体入选研究的概率 $P(S=1)$ 是已知的或者可估计的，具体来说，被纳入研究的 n 个个体通常来自一个更大的容量为 N 有限样本，而这个有限样本可以视作超总体的一个随机样本，因此可以将 $\frac{n}{N}$ 作为 $P(S=1)$ 的估计值。
- 对此我认为拆成下面两句话去理解更容易些：(1) 人类的认识是有边界的，我们所做的任何统计推断都被天然地限制在某个有限总体中，所有因果分析的结论仅在这个有限总体范围内是有效的。(2) 如果该有限总体是来自超总体的一个随机样本，那么，关于该有限总体的统计结论便可外推至整个超总体中。

结果依赖型采样的因果界：基于单调性假设 [Jun and Lee, 2021]

- 本文研究的重点是在单调性假设下做因果推断，关心的因果参数是用协变量分层后的 CRR 和 CRD：

$$\theta(x) := \frac{\mathbb{P}\{Y^*(1) = 1 \mid X^* = x\}}{\mathbb{P}\{Y^*(0) = 1 \mid X^* = x\}},$$

$$\theta_{\mathrm{AR}}(x) := \mathbb{P}\{Y^*(1) = 1 \mid X^* = x\} - \mathbb{P}\{Y^*(0) = 1 \mid X^* = x\}.$$

- 作者证明了在单调性假设下有 $1 \le \theta(x) \le \mathrm{OR}(x)$。
- 此外，作者定义了一些所谓的聚合参数如：

$$\beta(y) := \int \log \mathrm{OR}(x) dF_{X|Y}(x \mid y),$$

并构造了 $\beta(y)$ 的有效估计等。

讨论

- 在观察性研究中，因果参数不可识别，研究者通过寻找其可识别的边界来实现部分识别性。但一种更受欢迎的办法是找一个随机区间以一定的概率覆盖感兴趣的因果参数，但这是更加困难的一个问题。
- 最后，关于因果效应界的应用问题，可以参考 Gabriel, Sachs and Sjölander, 2021, 2022] 和 [Jun and Lee, 2021] 等，这些文献中提供了基于因果效应界的有意义的案例。在实际中，这方面的工作可以帮助研究者明确正效应和负效应的边界，以做出更明智的判断。

Thanks!

参考文献

[1] Holland, P. W., and Rubin, D. B. (1987), "Causal inference in retrospective studies," ETS Research Report Series, 1987, 203–231.

[2] Balke, A., and Pearl, J. (1994), "Counterfactual Probabilities: Computational Methods, Bounds and Applications," in Uncertainty Proceedings 1994, Elsevier, pp. 46–54.

[3] Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects from Studies with Imperfect Compliance," Journal of the American Statistical Association, 92, 1171–1176.

[4] Kuroki, M., Cai, Z., and Geng, Z. (2010), "Sharp bounds on causal effects in case-control and cohort studies," Biometrika, 97, 123–132.

参考文献

[5] Gabriel, E. E., and Sachs, M. C. (2021), "On the Use of Nonparametric Bounds for Causal Effects in Null Randomized Trials," American Journal of Epidemiology, 190, 2231–2231.

[6] Jun, S. J., and Lee, S. (2021), "Causal Inference under Outcome-Based Sampling with Monotonicity Assumptions," arXiv.

[7] Gabriel, E. E., Sachs, M. C., and Sjölander, A. (2022), "Causal Bounds for Outcome-Dependent Sampling in Observational Studies," Journal of the American Statistical Association, 117, 939–950.