# Bayesian Sensitivity of Causal Inference Estimators under Evidence-Based Priors

**Nikita Dhawan**
University of Toronto, Vector Institute
nikita@cs.toronto.edu

**Daniel Shen**
University of Waterloo
d2shen@uwaterloo.ca

**Leonardo Cotta**
Vector Institute
leonardo.cotta@vectorinstitute.ai

**Chris J. Maddison**
University of Toronto, Vector Institute
cmaddis@cs.toronto.edu

## Abstract

Causal inference, especially in observational studies, relies on untestable assumptions about the true data-generating process. Sensitivity analysis helps us assess how robust our conclusions are when we alter underlying assumptions. Existing frameworks concerned worst-case changes in these assumptions. In this work, we argue that using such pessimistic criteria can often become uninformative or lead to conclusions contradicting our prior knowledge about the world. We generalize the recent $s$-value framework [16] to estimate the sensitivity of three common assumptions in causal inference and empirically demonstrate this claim. Empirically, we find that, indeed, worst-case conclusions about sensitivity can rely on unrealistic changes in the data-generating process. To overcome this limitation, we extend the $s$-value framework with a new criterion, the Bayesian Sensitivity Value (BSV), which computes the expected sensitivity of an estimate to assumption violations under priors constructed from real-world evidence.

## 1 Introduction

Causal inference offers powerful tools for decision-making across high-stakes domains, but both its internal and external validity rely on assumptions that cannot be verified from data alone [2, 30]. Sensitivity analysis [34] has emerged as a critical tool to assess the robustness of causal conclusions to violations of these assumptions. Existing frameworks typically focus on the unconfoundedness assumption [43, 45, 8] — *all confounding variables have been measured and controlled for*. However, the internal validity of causal inference also depends on other model specification assumptions [1]. As for external validity, even randomized trials are exposed to selection and outcome biases, such that participants of a study may differ systematically from the broader population of interest.

The central question of sensitivity analysis is: *To what extent can assumptions be violated before an experiment's conclusions are overturned?* This is valuable when assessing whether evidence is strong enough to support decisions such as drug approval. Practitioners may also wish to compare subpopulations and prioritize those with higher sensitivity when planning resource allocation or future data collection [31]. Since exhaustively evaluating all possible assumption violations is infeasible, existing methods typically adopt a worst-case approach, identifying the smallest violation that reverses a conclusion [8, 16]. In multi-dimensional settings, this often reduces to analyzing restricted spaces, such as one confounder at a time, a simplification that can be misleading [38]. Moreover, worst-case violations may be implausible, overly pessimistic, and uninformative in higher dimensions.

In this work, we start by formalizing and showcasing the above challenges. To that end, we extend the $s$-value method of Gupta and Rothenhäusler [16] to include other common causal assumptions,

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

including no unmeasured confounding, well-specified conditional models of outcomes given treatment and covariates, and external validity with respect to the covariate distribution. This extended framework defines an assumption variable, the space of its possible values, a divergence metric to measure violations, and constraints that enable practitioners to target sensitivity questions of interest.

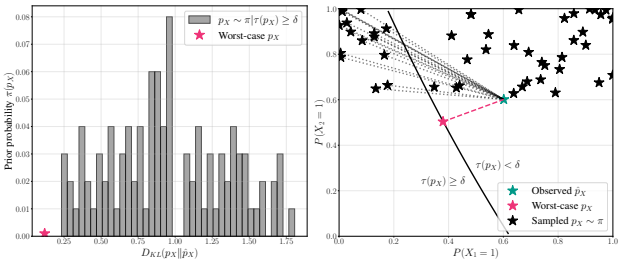



Figure 1: The closest (worst-case) distribution that reverses the causal decision has low probability under an empirical prior. More realistic distributions are further away from the observed $\hat{p}_X$ (Left) and lie in a different direction (Right).

Simulation studies show that worst-case sensitivity analyses are often overly pessimistic, since they treat all violations as equally likely and may contradict real-world knowledge (e.g., simultaneous increase in income and decrease in education is an unlikely distribution shift). To address this, we propose the Bayesian Sensitivity Value (BSV) and its Empirical Bayes variant (EBSV), which estimate expected sensitivity under priors derived from real-world evidence. Unlike existing Bayesian causal inference methods [28, 25] that place priors on treatment effects, BSV defines a posterior over changes in the data-generating process while using established frequentist estimators. EBSV further incorporates data-driven rather than arbitrary priors. Appendix E reviews related works further. As an illustrative example, fig. 1 considers sensitivity to the covariate distribution $p_X$ over two independent binary confounders. The worst-case distribution that reverses the causal decision has very low probability under an empirical prior. In contrast, distributions sampled from an empirical prior that reverse this decision indicate lower sensitivity in magnitude and lie in a different direction. Our main contributions are:

- We unify sensitivity analysis frameworks with respect to different types of assumptions in a general framework that allows practitioners to ask targeted questions about the sensitivity of a study.
- Our simulations and real-world application empirically demonstrate that worst-case analyses indeed often fail to reflect plausible scenarios, leading to overly pessimistic and uninformative results.
- We propose a new sensitivity criterion, the Bayesian Sensitivity Value (BSV), which is grounded in real-world evidence, is more realistic, and remains useful in high-dimensional assumption spaces.
- Finally, we empirically illustrate BSV's ability to address the above limitations, along with strategies to construct empirical priors from real-world evidence and practical algorithms for both criteria.

## 2 Sensitivity Analysis Framework

We consider the archetypal causal inference task of *binary treatment effect estimation*, where the goal is to make a binary decision based on this estimate. Let $T \in \{0, 1\}$ be a binary treatment and $Y(1), Y(0) \in \{0, 1\}$ be the corresponding potential outcomes under the Neyman–Rubin model [19]. In decision–focused causal inference [23, 18], the average treatment effect (ATE), $\tau := \mathbb{E}\big[Y(1) - Y(0)\big]$, is translated into binary choices such as $\mathbb{I}(\tau > \delta)$ for a user–specified threshold $\delta$. Since only $(T, Y)$ are observed in practice, consistent estimation of $\tau$ requires assumptions. Observational methods such as inverse propensity score weighting [20, 33] and outcome regression [36] achieve this by converting $\tau$ into identifiable forms via a series of assumptions on its non-identifiable components. We treat these components as a collection of functions $a$ and parameterize the ATE as $\tau(a)$. Following Lu and Ding [26, Theorem 1], for discrete covariates $X \in \mathcal{X}$ with distribution $p_X$, jointly observed with $T$ and $Y$, we have

$$\tau(\varepsilon, \mu, p_X) = \sum_{x \in \mathcal{X}} p_X(x) \big[\mathbb{E}\big[\mu_1(x) \cdot \left(T + \frac{1-T}{\varepsilon_1(x)}\right) - \mu_0(x) \cdot \left(1 - T + T\varepsilon_0(x)\right) \,\Big|\, X = x\big]\big]. \quad (1)$$

Here, $\varepsilon(x) = (\varepsilon_0(x), \varepsilon_1(x))$ with $\varepsilon_t(x) = \frac{\mathbb{E}[Y(t)|T=1, X=x]}{\mathbb{E}[Y(t)|T=0, X=x]}$ are odds ratios, $\mu(x) = (\mu_0(x), \mu_1(x))$ with $\mu_t(x) = \mathbb{E}[Y|T = t, X = x]$ are conditional outcomes, and $p_X \in \Delta^{|\mathcal{X}|-1}$ is the covariate distribution. See Lu and Ding [26] or appendix C for a proof of equivalence. Since $(\varepsilon, \mu, p_X)$ are non-identifiable or difficult to estimate, causal inference requires assumptions on their values. Appendix B discusses each function and its typically assumed value in more detail.

Under the assumptions of no unmeasured confounding, *i.e.*, $(\varepsilon_0(x), \varepsilon_1(x)) = (\mathbf{1}, \mathbf{1})$, well-specified conditional outcome models, *i.e.*, the risk-minimizing models of outcomes $\mu^*$ are the true conditional outcome distributions $\mu$, and external validity with respect to the covariate distribution, *i.e.*, the sampled population's distribution $q_X$ is the same as the target distribution $p_X$, we have

$$\tau(\mathbf{1}, \mu^*, q_X) = \sum_{x \in \mathcal{X}} q_X(x)\big[\mu_1^*(x) - \mu_0^*(x)\big] \approx \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mu}_1(x^{(i)}) - \hat{\mu}_0(x^{(i)})\big], \tag{2}$$

which is the standard outcome imputation estimator, computed using a study population, $(x^{(i)}, t^{(i)}, y^{(i)})_{i=1}^n$, of $n$ individuals and (usually) fitted regression models $\hat{\mu}$ for conditional outcomes. An analogous parameterization for inverse propensity score weighting is in appendix D.

For a given choice of $a = (\varepsilon, \mu, p_X)$, our goal is to quantify the effect of deviations from $a$ on the decision $\mathbb{I}(\tau(a) > \delta)$. We consider these functions separately, each of which may lie in a high-dimensional space. For any function, we define:

(i) $\mathcal{A}$: the subset of possible triplets $(\varepsilon, \mu, p_X)$ over which the function ranges,
(ii) $D(a\|\hat{a})$: a divergence measuring deviation between assumed value $\hat{a}$ and some $a \in \mathcal{A}$.

Assume without loss of generality that $\tau(\mathbf{1}, \mu^*, q_X) > \delta$. Then the region $\{a \in \mathcal{A} : \tau(a) \leq \delta\}$ represents a reversal of the causal decision. A special case of this framework is the $s$-value introduced by Gupta and Rothenhäusler [16], which measures the minimum shift in $p_X$ required to change the sign of the ATE (threshold $\delta = 0$). Generalizing this, we define the worst-case sensitivity.

**Definition 1** (Worst-case Sensitivity)**.** *For a function with values in $\mathcal{A}$, assumed value $\hat{a}$, and convex divergence $D$ (*e.g. *Bregman divergences), if $\tau(\hat{a}) > \delta$, the worst-case sensitivity is*

$$\sup_{a \in \mathcal{A}} \{\exp(-D(a\|\hat{a})) \mid \tau(a) \leq \delta\}. \tag{3}$$

This definition ensures values in $[0, 1]$ for non-negative divergences, with higher values indicating higher sensitivity since smaller deviations reverse the decision. Constrained optimization algorithms for computing these values are described in appendix F.1.

## 3 Bayesian Sensitivity Value

Although sensitivity analyses are important for decision-making, they have not seen widespread adoption [41] and are often of limited practical use [38]. We take a critical view of worst-case sensitivity, highlighting pitfalls that make it uninformative or unrealistic. The following diabetes example motivates the need for an alternative.

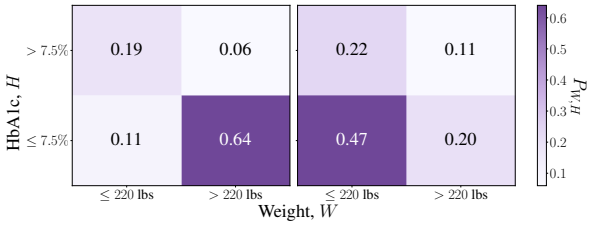



Figure 2: Worst-case optimum (Left) significantly differs from the real-world distribution recorded in NHANES (Right), contradicting prior knowledge.

**Motivating Example.** Worst-case analyses compute the solution to eq. (3), which can be compared to empirical realizations of $a$ in real populations. We did so for the joint distribution of binarized covariates `Weight` and `HbA1c` in diabetes patients (see section 4), alongside empirical data from NHANES [29]. Figure 2 shows that the empirical distribution places highest probability on low `HbA1c` and low `Weight`, consistent with the known correlation between glycated hemoglobin reduction and weight loss [15]. In contrast, the worst-case optimum assigns large probability to low `HbA1c` and high `Weight`, an implausible scenario that contradicts established relationships. To address this, we argue in favor of accounting for how likely a given assumption violation is. Concretely, we treat $(\varepsilon, \mu, p_X)$ as random variables and propose the Bayesian Sensitivity Value (BSV).

**Definition 2** (Bayesian Sensitivity Value)**.** *Given an assumption $\hat{a} \in \mathcal{A}$, divergence $D$, and a random assumption $A$ taking values in $\mathcal{A}$ with distribution $\pi_A$, if $\tau(\hat{a}) > \delta$, the Bayesian Sensitivity Value is*

$$\mathbb{E}_{A \sim \pi_A}[\exp(-D(A\|\hat{a})) \mid \tau(A) \leq \delta]. \tag{4}$$

BSV replaces the supremum in eq. (3) with an expectation under a prior, offering several advantages: (i) it is no more pessimistic than the worst-case value and usually far less so; (ii) results can be

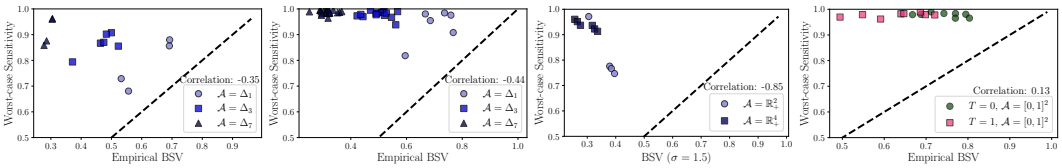



Figure 4: For all assumptions (Left to Right: $p_{X_j}$ with four covariates, $p_{X_j}$ with six covariates, $\varepsilon_{X_j}$, and $\mu_{t,X_j}$) and different choices of subsets $\mathcal{A}$, Bayesian sensitivity is uncorrelated or negatively correlated with worst-case sensitivity, revealing different trends in sensitivity rankings.

interpreted under different chosen priors $\pi_A$; (iii) large demographic or medical datasets can be used to build evidence-based priors, yielding an Empirical Bayesian Sensitivity Value (EBSV). The EBSV grounds analyses in real-world evidence and prior knowledge. For instance, demographic surveys can provide priors for $p_X$, while previous studies and subgroup analyses can inform priors over $\mu$. This quantifies the sensitivity of new studies under the prior of previous findings. Appendix F.2 describes constrained sampling algorithms to compute (E)BSVs.

## 4 Empirical Analysis

Our experiments use sensitivity analyses to compare and rank subsets of the $(\varepsilon, \mu, p_X)$ triplets, denoted $\mathcal{A}$, that differ meaningfully, so that high-sensitivity subgroups may be discovered, for instance, for resource prioritization. We computed worst-case sensitivity value, the BSV under an uniform or uninformative prior, and the EBSV under evidence-based priors, when they are available, to investigate the following questions: (i) How do different sensitivity criteria scale with dimensionality of the space of possible assumptions $\mathcal{A}$? (ii) Do different criteria yield different sensitivity rankings across $\mathcal{A}$? (iii) How does BSV perform in real-world applications?

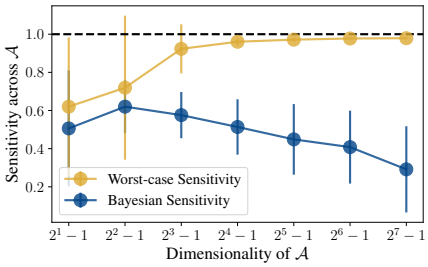
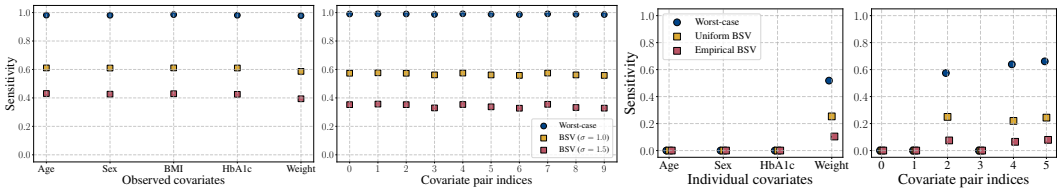



Figure 3: For high-dimensional $\mathcal{A}$, worst-case sensitivity values for $p_X$ are all close to 1 (mean $\approx 1$, variance $\approx 0$), unlike BSVs which better distinguish between different $\mathcal{A}$.

**Simulations.** We simulated data with binary covariates and heterogeneous treatment effects, introducing confounding and selection bias to reflect realistic observational settings (see appendix G.1). For BSV, we construct empirical priors for $p_X$ and $\mu$ from unbiased samples, while priors on odds ratios $\varepsilon$ are specified as truncated Gaussians. Following the comparison of sensitivity values across covariates in Gupta and Rothenhäusler [16], we constructed comparisons between different $\mathcal{A}$ for all our assumptions, where possible elements of each $\mathcal{A}$ range over particular covariates and treatments. For example, to find the pair of binary covariates whose joint distribution $\tau(\varepsilon, \mu, p_X)$ is most sensitive to, we compared sensitivity values given by eq. (3) or eq. (4) with $\mathcal{A} = \{(\mathbf{1}, \mu^*, q_{X_{-j}}, p_{X_j})\}_{p_{X_j} \in \Delta_3}$ for different pairs of covariates $X_j$, where $X_{-j}$ includes all other covariates. This results in a sensitivity ranking according to either the worst-case or the Bayesian criterion.

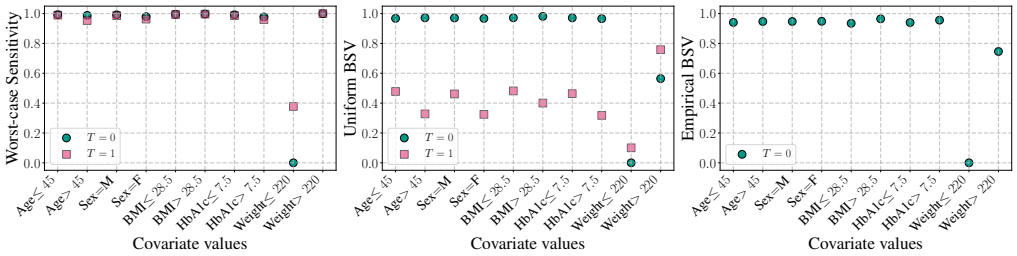
To study behavior as the dimension of $\mathcal{A}$ scales (item (i)), we varied the number of covariates and computed sensitivity values for all resulting subsets. Item (i) shows that worst-case values rapidly converge to 1 as dimensionality increases, while BSV maintains variability acrros $\mathcal{A}$. To study rankings (item (ii)), we compared sensitivity to assumptions on $p_{X_j}$, $\mu_{t,X_j}$, and $\varepsilon_{X_j}$. Worst-case values are uniformly high, whereas BSV reveals distinct and sometimes opposite trends, as shown in fig. 4. Additional comparisons are included in appendix H.

**Diabetes Application.** To answer item (iii) in a real-world application, we adapted the Semaglutide vs. Tirzepatide dataset and large-language-model (LLM) based causal estimator from Dhawan et al. [12]. This study leveraged patient-reported data from relevant subreddits, employed large language models to extract observational distributions, accounted for confounding due to `Age`, `Sex`, `BMI` (Body Mass Index), `HbA1c` (Glycated Hemoglobin), and `Weight`, and estimated the relative effect of diabetes treatments on a weight loss outcome. Given the potential selection biases in both the data collection and modeling via LLMs, informative sensitivity analyses are crucial for reliable inference in this setting. See appendix G.2 for complete details.




(a) Sensitivity to $\varepsilon_{X_j}$.

(b) Sensitivity to $p_{X_j}$.




(c) Sensitivity to $\mu_{t,X_j}$.

Figure 5: BSV for odds ratios in fig. 5a, covariate distributions in fig. 5b, and conditional outcome distributions fig. 5c in a real-world diabetes experiment can be more informative than worst-case analyses. In particular, BSV suggests relatively low sensitivity in the `Weight > 220 lbs` subpopulation, as compared to the worst-case analysis which assigns it the highest possible sensitivity value.

This setting is also useful to demonstrate the construction of empirical priors and practical instantiation of an EBSV analysis when true priors are unavailable. Here, we leveraged previous databases and studies for information relevant to the assumptions of interest, $p_X$ and $\mu$. Specifically, we extracted distributions over `Age`, `Sex`, `HbA1c`, and `Weight` across different races using the Diabetes dataset from the UCI repository [9] and used them to fit a Dirichlet distribution that serves as a prior over the covariate distribution $p_X$. We obtained distributions over the conditional outcome distributions for Semaglutide ($T = 0$) from subgroup analyses conducted in previous studies [24] and similarly fit a Dirichlet distribution to use as the empirical prior.

We conducted sensitivity analyses under the worst-case criterion, the BSV with uniform priors, and the empirical BSV, on the diabetes study, with their results visualized in fig. 5. In the analyses with respect to $p_X$ and $\mu$, shown in figs. 5b and 5c respectively, worst-case values and BSV always agree in zero-sensitivity settings as expected. However, BSV highlights informative trends in others, which can support targeted future experimentation and data collection. Notably, they suggest that the estimated ATE is less sensitive than what a worst-case analysis would indicate in the `Weight > 220 lbs` subpopulation. Similar to the simulation study, fig. 5a shows BSV for odds ratios under truncated Gaussian priors. While differences are small between choices of observed covariates, sensitivity is lower whenever `Weight` is among them.

## 5 Conclusions

In this work, we generalized the $s$-value framework to three common assumptions in causal inference and proposed a new sensitivity criterion, the Bayesian Sensitivity Value, that leverages real-world evidence. Our empirical studies showed that worst-case analyses are often overly pessimistic, especially in high-dimensional spaces, and rarely reflect realistic shifts, whereas the BSV can distinguish between different $\mathcal{A}$ and reveal trends that differ from worst-case analyses. Even so, it is not without limitations of its own, as discussed in appendix I. Exciting directions for future work include (i) constructing informative priors for different assumptions, including odds ratios, using varied sources of real-world evidence, (ii) adapting more sophisticated sampling techniques to improve efficiency of BSV computations, and (iii) deriving mathematically rigorous strategies to operationalize sensitivity analyses for future experiment design. Given the critical role of sensitivity analyses in high-stakes domains like medicine, improving their practical utility is essential. While traditional approaches emphasize worst-case violations, the BSV offers a complementary and potentially more informative criterion for real-world applications and resource prioritization.

## Acknowledgements

We would like to thank Ayoub El Hanchi for helpful discussions and feedback on the paper. Resources used in preparing this research were provided in part by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2021-03445.

## References

[1] Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

[2] Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pages 507–556.

[3] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

[4] Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

[5] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

[6] Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[7] CDC (2023). Diabetes health indicators dataset.

[8] Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67.

[9] Clore, J., Cios, K., DeShazo, J., and Strack, B. (2014). Diabetes 130-us hospitals for years 1999-2008. *UCI Machine Learning Repository*, 10:C5230J.

[10] Cortes-Gomez, S., Dulce, M., Patino, C., and Wilder, B. (2023). Statistical inference under constrained selection bias. *arXiv preprint arXiv:2306.03302*.

[11] de Souza, R. J., Eisen, R. B., Perera, S., Bantoto, B., Bawor, M., Dennis, B. B., Samaan, Z., and Thabane, L. (2016). Best (but oft-forgotten) practices: sensitivity analyses in randomized controlled trials. *The American journal of clinical nutrition*, 103(1):5–17.

[12] Dhawan, N., Cotta, L., Ullrich, K., Krishnan, R., and Maddison, C. J. (2024). End-to-end causal effect estimation from unstructured natural language data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[13] Fawkes, J., Fishman, N., Andrews, M., and Lipton, Z. (2024). The fragility of fairness: Causal sensitivity analysis for fair machine learning. *Advances in Neural Information Processing Systems*, 37:137105–137134.

[14] Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

[15] Gummesson, A., Nyman, E., Knutsson, M., and Karpefors, M. (2017). Effect of weight reduction on glycated haemoglobin in weight loss trials in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism*, 19(9):1295–1305.

[16] Gupta, S. and Rothenhäusler, D. (2023). The s-value: evaluating stability with respect to distributional shifts. *Advances in Neural Information Processing Systems*, 36:72058–72070.

[17] Han, L., Arfè, A., and Trippa, L. (2024). Sensitivity analyses of clinical trial designs: selecting scenarios and summarizing operating characteristics. *The American Statistician*, 78(1):76–87.

[18] Hedayat, A., Wang, J., and Xu, T. (2015). Minimum clinically important difference in medical studies. *Biometrics*, 71(1):33–41.

[19] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

[20] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

[21] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

[22] Ioannidis, J. P., Tan, Y. J., and Blum, M. R. (2019). Limitations and misinterpretations of e-values for sensitivity analyses of observational studies. *Annals of internal medicine*, 170(2):108–111.

[23] Jaeschke, R., Singer, J., and Guyatt, G. H. (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Controlled clinical trials*, 10(4):407–415.

[24] Kadowaki, T., Lee, S. Y., Ogawa, W., Nishida, T., Overvad, M., Tobe, K., Yamauchi, T., Lim, S., et al. (2024). Clinical characteristics affecting weight loss in an east asian population receiving semaglutide: A step 6 subgroup analysis. *Obesity Research & Clinical Practice*, 18(6):457–464.

[25] Li, F., Ding, P., and Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153.

[26] Lu, S. and Ding, P. (2023). Flexible sensitivity analysis for causal inference in observational studies subject to unmeasured confounding. *arXiv preprint arXiv:2305.17643*.

[27] Ma, J., Liu, M., Wang, R., Du, L., and Ji, L. (2024). Efficacy and safety of tirzepatide in people with type 2 diabetes by baseline body mass index: An exploratory subgroup analysis of surpass-ap-combo. *Diabetes, Obesity and Metabolism*, 26(4):1454–1463.

[28] McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine*, 26(11):2331–2347.

[29] National Center for Health Statistics (2023). National health and nutrition examination survey data.

[30] Pearl, J. and Bareinboim, E. (2011). External validity and transportability: a formal approach. In *JSM Proceedings*, pages 157–171.

[31] Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Piano, S. L., Iwanaga, T., Becker, W., et al. (2021). The future of sensitivity analysis: an essential discipline for systems modeling and policy support. *Environmental Modelling & Software*, 137:104954.

[32] Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

[33] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394.

[34] Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.

[35] Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[36] Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26.

[37] Ryan, D. H., Lingvay, I., Deanfield, J., Kahn, S. E., Barros, E., Burguera, B., Colhoun, H. M., Cercato, C., Dicker, D., Horn, D. B., et al. (2024). Long-term weight loss effects of semaglutide in obesity without diabetes in the select trial. *Nature medicine*, 30(7):2049–2057.

[38] Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., and Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental modelling & software*, 114:29–39.

[39] Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company.

[40] Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

[41] Tarantola, S., Ferretti, F., Piano, S. L., Kozlova, M., Lachi, A., Rosati, R., Puy, A., Roy, P., Vannucci, G., Kuc-Czarnecka, M., et al. (2024). An annotated timeline of sensitivity analysis. *Environmental Modelling & Software*, 174:105977.

[42] Tchang, B. G., Mihai, A. C., Stefanski, A., García-Pérez, L.-E., Mojdami, D., Jouravskaya, I., Gurbuz, S., Taylor, R., Karanikas, C. A., and Dunn, J. P. (2025). Body weight reduction in women treated with tirzepatide by reproductive stage: a post hoc analysis from the surmount program. *Obesity*, 33(5):851–860.

[43] VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

[44] Veitch, V. and Zaveri, A. (2020). Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in neural information processing systems*, 33:10999–11009.

[45] Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761.

# A Notation

| | |
|---|---|
| $A$ | Assumption random variable whose value must be assumed for causal inference. |
| $\pi_A$ | Prior distribution over assumption variable $A$. |
| $X$ | Random variable corresponding to features of an individual in a causal inference dataset. |
| $T$ | Random variable corresponding to treatment or intervention assigned to an individual in a causal inference dataset. |
| $Y$ | Random variable corresponding to outcome observed for an individual in a causal inference dataset. |
| $x$ | Possible instance of $X$ from its support $\mathcal{X}$. |
| $t$ | Possible instance of $T$ from its support $\mathcal{T} = \{0, 1\}$ (binary treatments). |
| $y$ | Possible instance of $Y$ from its support $\mathcal{Y} = \{0, 1\}$ (binary outcomes). |
| $a$ | Possible instance or sampled value of $A$ from its support $\mathcal{A}$. |
| $Y(t)$ | Random variable corresponding to potential outcome observed for an individual after receiving treatment $t$. |
| $\varepsilon(X)$ | Unconfoundedness assumption parameters, $(\varepsilon_0(X), \varepsilon_1(X))$. |
| $\mu(X)$ | Conditional outcome distributions given covariates and treatment, $(\mu_0(X), \mu_1(X))$. |
| $p_X$ | Distribution over covariates $X$. |
| $x^{(i)}$ | Sampled value of $X$ for individual $i$. |
| $t^{(i)}$ | Sampled value of $T$ for individual $i$. |
| $y^{(i)}$ | Sampled value of $Y$ for individual $i$. |
| $\tau$ | Average treatment effect (ATE) given by $\mathbb{E}[Y(1) - Y(0)]$, where the expectation is over some defined population of individuals. |
| $\tau(a)$ | ATE computed or estimated under the assumption $a$. |
| $\delta$ | Decision threshold on the (estimated) ATE to choose between treatments. |
| $n$ | Total number of individuals. |

## B Sensitivity Analysis Framework Details

As discussed in section 2, in decision–focused causal inference [23, 18], we translate estimates of the statistic called *average treatment effect* (ATE),

$$\tau := \mathbb{E}\big[Y(1) - Y(0)\big], \tag{5}$$

into binary choices such as $\mathbb{I}(\tau > \delta)$ for a user–specified threshold $\delta$. See a full list of notation in Appendix A.

The fundamental problem in causal inference is that we never observe the potential outcomes $(Y(0), Y(1))$. Instead, we observe the treatment $T$ and an outcome $Y = T \cdot Y(1) + (1 - T) \cdot Y(0)$. Therefore, different studies require assumptions of different forms and strengths to estimate the ATE consistently from observed data. Specifically, observational methods like inverse propensity score weighting [20, 33] and outcome regression [36] convert the eq. (5) above into identifiable forms via a series of assumptions on its non-identifiable components. In this work, we treat these components as a collection of functions $a$, and parameterize the ATE as a function of these unknown functions: $\tau(a)$.

To illustrate this framework, consider the outcome imputation estimator. We start with the form of the ATE proposed in Lu and Ding [26, Theorem 1] and then state the typical assumptions that make it identifiable from data. Given discrete covariates $X \in \mathcal{X}$, drawn from a distribution $p_X$ and observed jointly with treatment $T$ and outcome $Y$, we have

$$\tau(\varepsilon, \mu, p_X) = \sum_{x \in \mathcal{X}} p_X(x) \left[ \mathbb{E}\left[ \mu_1(x) \cdot \left( T + \frac{1-T}{\varepsilon_1(x)} \right) \;\middle|\; X = x \right] - \mathbb{E}\left[ \mu_0(x) \cdot \left( 1 - T + T\varepsilon_0(x) \right) \;\middle|\; X = x \right] \right]. \tag{6}$$

This equality holds, when the odds ratio function, $\varepsilon : \mathcal{X} \to \mathbb{R}_+^2$, is given by $\varepsilon(x) = (\varepsilon_0(x), \varepsilon_1(x))$ where

$$\varepsilon_0(x) = \frac{\mathbb{E}[Y(0)|T = 1, X = x]}{\mathbb{E}[Y(0)|T = 0, X = x]}, \quad \varepsilon_1(x) = \frac{\mathbb{E}[Y(1)|T = 1, X = x]}{\mathbb{E}[Y(1)|T = 0, X = x]},$$

the conditional outcome function, $\mu : \mathcal{X} \to [0, 1]^2$, is given by $\mu(x) = (\mu_0(x), \mu_1(x))$ where

$$\mu_0(x) = \mathbb{E}[Y|T = 0, X = x], \quad \mu_1(x) = \mathbb{E}[Y|T = 1, X = x],$$

and the covariate distribution, $p_X \in \Delta^{|\mathcal{X}|-1}$ gives the probability $p_X(x)$. See Lu and Ding [26] or appendix C for a proof of the equivalence between eqs. (5) and (6). The functions $(\varepsilon, \mu, p_X)$ are either non-identifiable from observational data or challenging to estimate. For this reason, causal inference makes assumptions on the possible values of $a = (\varepsilon, \mu, p_X)$. Below, we discuss each function and its typically assumed value.

The common **no unmeasured confounding** [35, 21] assumption is that true odds ratios are identically 1, *i.e.*, $(\varepsilon_0(x), \varepsilon_1(x)) = (\mathbf{1}, \mathbf{1})$. Equivalently, potential outcomes are independent of treatment assignment, given observed covariates, *i.e.* $(Y(0), Y(1)) \perp\!\!\!\perp T|X$. Note that this assumption holds true in randomized experiments, but may be violated in observational studies due to unobserved confounders. Under the no unmeasured confounding assumption, and for the target distribution of covariates $p_X$, eq. (6) reduces to

$$\tau(\mathbf{1}, \mu, p_X) = \sum_{x \in \mathcal{X}} p_X(x) \big[ \mu_1(x) - \mu_0(x) \big]. \tag{7}$$

Assumptions of **well-specified conditional outcome models** [e.g., 36] typically state that the conditional outcome distributions $\mu$ can be learned from observed data by modeling the mapping from covariates and treatments to expected outcome, *i.e.*, the risk-minimizing models of outcomes $\mu^*$ are the true conditional outcome distributions $\mu$. This assumption may be violated due to model misspecification or reporting biases that modify the outcome distribution. Under no unmeasured confounding and correct specification of conditional outcome models, we now have

$$\tau(\mathbf{1}, \mu^*, p_X) = \sum_{x \in \mathcal{X}} p_X(x) \big[ \mu_1^*(x) - \mu_0^*(x) \big]. \tag{8}$$

**External validity** [39, 30] assumptions typically state that the sampled population is the same as the target population of interest. In particular, in the context of a population distribution over

| Assumption | Function | Space $\mathcal{A}$ | Divergence $D$ | Typical assumption |
|---|---|---|---|---|
| Unconfoundedness | $\varepsilon$ | $\mathbb{R}_+^{2d}$ | Euclidean | $\mathbf{1}$ |
| Conditional Outcome Distributions | $\mu$ | $[0,1]^{2d}$ | KL | $\mu^*$ |
| Covariate Distribution | $p_X$ | $\Delta_{d-1}$ | KL | $q_X$ |

Table 1: Typical assumptions required for ATE identification in our sensitivity analysis framework.

covariates, an assumption of external validity would state that the sampled population's distribution over covariates $q_X$ is the same as the target distribution $p_X$. This assumption can be violated due to selection biases in data collection such that the observed population is not representative of the true target population. Under all three assumptions above, we have

$$\tau(\mathbf{1}, \mu^*, q_X) = \sum_{x \in \mathcal{X}} q_X(x)\big[\mu_1^*(x) - \mu_0^*(x)\big] \approx \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mu}_1(x^{(i)}) - \hat{\mu}_0(x^{(i)})\big], \tag{9}$$

which is the standard and popular outcome imputation estimator, computed using a study population, $(x^{(i)}, t^{(i)}, y^{(i)})_{i=1}^n$, of $n$ individuals and (usually) fitted regression models $\hat{\mu}$ for conditional outcome distributions. While the example above corresponds to outcome imputation, it is straightforward to derive analogous parameterizations of assumptions for other estimators. As another example, we include the inverse propensity score weighting estimator in appendix D.

For binary outcomes and covariates that can take $d$ possible values ($|\mathcal{X}| = d$), the function of odds ratios $\varepsilon$ belong to $\mathbb{R}_+^{2d}$, conditionals $\mu$ are defined in $[0,1]^{2d}$, and the covariate distribution lies in the simplex $\Delta_{d-1}$. Hence, $\mathcal{A}$ respectively corresponds to the subsets $\{(\varepsilon, \mu^*, q_X)\}_{\varepsilon \in \mathbb{R}_+^{2d}}$, $\{(\mathbf{1}, \mu, q_X)\}_{\mu \in [0,1]^{2d}}$, and $\{(\mathbf{1}, \mu^*, p_X)\}_{p_X \in \Delta_{d-1}}$. We take $D$ as the Kullback-Leibler (KL) divergence for distributions $\mu$ and $p_X$ and Euclidean distance for the vectors $\varepsilon$. Table 1 summarizes these functions and their assumptions in our sensitivity framework.

## C ATE Identification Proof

We reproduce a proof of the result in Lu and Ding [26] for eq. (1), in our notation, for completeness.

$$\begin{aligned}
\tau &:= \mathbb{E}\big[Y(1) - Y(0)\big] \\
&= \mathbb{E}\big[Y(1) \mid T = 1\big]P(T = 1) + \mathbb{E}\big[Y(1) \mid T = 0\big]P(T = 0) \\
&\quad - \mathbb{E}\big[Y(0) \mid T = 1\big]P(T = 1) - \mathbb{E}\big[Y(0) \mid T = 0\big]P(T = 0), \\
&= \mathbb{E}\big[Y \mid T = 1\big]P(T = 1) + \mathbb{E}\big[Y(1) \mid T = 0\big]P(T = 0) \\
&\quad - \mathbb{E}\big[Y(0) \mid T = 1\big]P(T = 1) - \mathbb{E}\big[Y \mid T = 0\big]P(T = 0),
\end{aligned}$$

where the last equality follows from the definition of $Y$.

For $\mu_0(X) = \mathbb{E}[Y|T = 0, X]$ and $\mu_1(X) = \mathbb{E}[Y|T = 1, X]$,

$$\begin{aligned}
\mathbb{E}\big[Y \mid T = 1\big]P(T = 1) &= \mathbb{E}\big[\mathbb{E}\big[Y \mid T = 1, X\big] \mid T = 1\big]P(T = 1) \\
&= \mathbb{E}\big[\mu_1(X) \mid T = 1\big]P(T = 1). \\
\mathbb{E}\big[Y \mid T = 0\big]P(T = 0) &= \mathbb{E}\big[\mathbb{E}\big[Y \mid T = 0, X\big] \mid T = 0\big]P(T = 0) \\
&= \mathbb{E}\big[\mu_0(X) \mid T = 0\big]P(T = 0).
\end{aligned}$$

For $\varepsilon_0(X) = \dfrac{\mathbb{E}[Y(0)|T = 1, X]}{\mathbb{E}[Y(0)|T = 0, X]}$ and $\varepsilon_1(X) = \dfrac{\mathbb{E}[Y(1)|T = 1, X]}{\mathbb{E}[Y(1)|T = 0, X]}$,

$$\begin{aligned}
\mathbb{E}\big[Y(1) \mid T = 0\big]P(T = 0) &= \mathbb{E}\big[\mathbb{E}\big[Y(1) \mid T = 0, X\big] \mid T = 0\big]P(T = 0) \\
&= \mathbb{E}\left[\frac{\mu_1(X)}{\varepsilon_1(X)} \mid T = 0\right]P(T = 0). \\
\mathbb{E}\big[Y(0) \mid T = 1\big]P(T = 1) &= \mathbb{E}\big[\mathbb{E}\big[Y(0) \mid T = 1, X\big] \mid T = 1\big]P(T = 1) \\
&= \mathbb{E}\big[\mu_0(X)\varepsilon_0(X) \mid T = 1\big]P(T = 1).
\end{aligned}$$

Plugging the above back into the equation for $\tau$, we have,

$$\begin{aligned}
\tau &= \mathbb{E}\big[Y \mid T = 1\big]P(T = 1) + \mathbb{E}\big[Y(1) \mid T = 0\big]P(T = 0) \\
&\quad - \mathbb{E}\big[Y(0) \mid T = 1\big]P(T = 1) - \mathbb{E}\big[Y \mid T = 0\big]P(T = 0) \\
&= \mathbb{E}\big[\mu_1(X) \mid T = 1\big]P(T = 1) + \mathbb{E}\left[\frac{\mu_1(X)}{\varepsilon_1(X)} \mid T = 0\right]P(T = 0) \\
&\quad - \mathbb{E}\big[\mu_0(X)\varepsilon_0(X) \mid T = 1\big]P(T = 1) - \mathbb{E}\big[\mu_0(X) \mid T = 0\big]P(T = 0) \\
&= \mathbb{E}\left[\mu_1(X) \cdot \left(T + \frac{1 - T}{\varepsilon_1(X)}\right)\right] - \mathbb{E}\left[\mu_0(X) \cdot \Big(T\varepsilon_0(X) + 1 - T\Big)\right],
\end{aligned}$$

which is equivalent to eq. (1) by the law of iterated expectations.

## D Assumptions for Inverse Propensity Score Weighting

Consider the inverse propensity score weighting estimator. Equivalent to the form of the ATE proposed in Lu and Ding [26, Theorem 2], we have

$$\tau(\varepsilon, e, p_{XTY}) = \sum_{\substack{x\in\mathcal{X}\\(t,y)\in\{0,1\}^2}} p_{XTY}(x,t,y)\left[w_1(x)\cdot\left(\frac{ty}{e(x)}\right) - w_0(x)\cdot\left(\frac{(1-t)y}{1-e(x)}\right)\right], \tag{10}$$

where $w_1(x) = e(x) + \frac{1-e(x)}{\varepsilon_1(x)}$ and $w_0(x) = e(x)\varepsilon_0(x) + 1 - e(x)$.

This equality holds, when the odds ratio function, $\varepsilon : \mathcal{X} \to \mathbb{R}^2_+$, is given by $\varepsilon(x) = (\varepsilon_0(x), \varepsilon_1(x))$ where

$$\varepsilon_0(x) = \frac{\mathbb{E}[Y(0)|T=1, X=x]}{\mathbb{E}[Y(0)|T=0, X=x]}, \quad \varepsilon_1(x) = \frac{\mathbb{E}[Y(1)|T=1, X=x]}{\mathbb{E}[Y(1)|T=0, X=x]},$$

the propensity scores, $e : \mathcal{X} \to [0,1]^2$, are given by

$$e(x) = \mathbb{E}[T|X=x],$$

and the joint distribution, $p_{XTY} \in \Delta^{4|\mathcal{X}|-1}$ gives the probability $p_{XTY}(x,t,y)$.

The functions $(\varepsilon, \mu, p_{XTY})$ are either non-identifiable from observational data or challenging to estimate. So we must make the following assumptions on the possible values of $\mathbf{a} = (\varepsilon, \mu, p_{XTY})$.

Once again, the **no unmeasured confounding** assumption is that true odds ratios are identically 1, *i.e.*, $(\varepsilon_0(x), \varepsilon_1(x)) = (\mathbf{1}, \mathbf{1})$, as discussed in appendix B. Under this assumption, eq. (10) reduces to

$$\tau(\mathbf{1}, e, p_{XTY}) = \sum_{\substack{x\in\mathcal{X}\\(t,y)\in\{0,1\}^2}} p_{XTY}(x,t,y)\left[\frac{ty}{e(x)} - \frac{(1-t)y}{1-e(x)}\right], \tag{11}$$

where the propensity scores $e$ and jont distribution $p_{XTY}$ are defined as before.

Assumptions of **well-specified propensity score models** [e.g., 20] typically state that the propensity scores $e$ can be learned from observed data by modeling the mapping from covariates to treatments, *i.e.*, the risk-minimizing models of outcomes $e^*$ are the true propensity scores $e$. This assumption may be violated due to model misspecification or reporting biases that modify the treatment assignment distribution. Under no unmeasured confounding and correct specification of propensity score models, we have

$$\tau(\mathbf{1}, e^*, p_{XTY}) = \sum_{\substack{x\in\mathcal{X}\\(t,y)\in\{0,1\}^2}} p_{XTY}(x,t,y)\left[\frac{ty}{e^*(x)} - \frac{(1-t)y}{1-e^*(x)}\right]. \tag{12}$$

**External validity of** $p_{XTY}$ assumptions typically state that the sampled population is the same as the target population of interest. Here, an assumption of external validity would state that the sampled population's distribution over covariates, treatments, outcomes $q_{XTY}$ is the same as the target distribution $p_{XTY}$. This assumption can be violated due to selection biases in data collection such that the observed population is not representative of the true target population. Under all three assumptions above, we have

$$\tau(\mathbf{1}, e^*, q_{XTY}) = \sum_{\substack{x\in\mathcal{X}\\(t,y)\in\{0,1\}^2}} q_{XTY}(x,t,y)\left[\frac{ty}{e^*(x)} - \frac{(1-t)y}{1-e^*(x)}\right] \tag{13}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t^{(i)}y^{(i)}}{\hat{e}(x^{(i)})} - \frac{(1-t^{(i)})y^{(i)}}{1-\hat{e}(x^{(i)})}\right], \tag{14}$$

which is the standard and popular inverse propensity weighting estimator, computed using a study population, $(x^{(i)}, t^{(i)}, y^{(i)})_{i=1}^n$, of $n$ individuals and (usually) fitted regression models $\hat{e}$ for propensity scores.

## E Related Work

Sensitivity analyses have long been recognized as critical for causal inference, particularly in observational studies [33]. Traditional approaches, that typically consider worst-case violations of the unconfoundedness assumption, include Tan [40], VanderWeele and Ding [43], Cinelli and Hazlett [8], Veitch and Zaveri [44]. Lu and Ding [26] also proposed a flexible analysis for the unconfoundedness assumption that may be applied to different estimators. There are also Bayesian approaches to causal inference [25] and sensitivity analysis [28] in the context of this assumption. These works typically employ expert-elicited or uninformative priors. Gupta and Rothenhäusler [16] tackled sensitivity to covariate distribution shifts and proposed the $s$-value framework, which we generalized to other assumptions, including those on conditional outcome distributions and unconfoundedness.

Causal sensitivity analyses have been adapted to several applications in fairness [13], optimization of operating characteristics of clinical trial design [17] and evaluation of constrained selection biases [10]. Recent critiques have highlighted limitations and misinterpretations associated with sensitivity analyses in practice, emphasizing a need for more realistic and informative methodologies [22, 38, 11]. Our work takes a step towards addressing these limitations by grounding sensitivity analyses in real-world evidence.

Several large databases, collected and curated nationally or globally, provide empirical data on demographic and medical variables [9, 7]. Additionally, post-hoc analyses of subgroup effects are commonly found in medical literature [42, 24, 27, 37]. Our proposed Bayesian Sensitivity Value provides a way to leverage these sources of information to understand sensitivity with respect to $p_X$ and $\mu_X$, respectively, while representing real-world populations.

# F Practical Algorithms

Computation of the worst-case sensitivity value in eq. (3) and the BSV in eq. (4) requires solving constrained optimization and constrained sampling problems, respectively. Gupta and Rothenhäusler [16] exploited the linearity of $\tau$ in $p_X$ to derive closed-form solutions for sensitivity to $p_X$. We can generalize their solutions by noticing that the constrained problem in eq. (3) is convex in each of $(\varepsilon, \mu, p_X)$ (or a one-to-one transformation of it), even if it isn't linear. We accordingly adapted existing optimization strategies via Lagrange multipliers to compute worst-case sensitivity values. Similarly, we used existing constrained sampling schemes for the BSV, though more efficient samplers may be derived by exploiting the convexity of eq. (4).

## F.1 Constrained Optimization

We formulate eq. (3) as an equivalent minimization problem, with the following Langrangian.

$$\max_{\lambda} \min_{a \in \mathcal{A}} D(a\|\hat{a}) + \lambda(\tau(a) - \delta), \tag{15}$$

where $\lambda \geq \mathbf{0}$ is a vector of Lagrange multipliers enforcing the threshold constraint on the ATE. We include any assumption-specific constraints on $a$, *e.g.* $\varepsilon_0(x) \geq 0, \varepsilon_1(x) \geq 0$ for all covariate sets $x$, with their own Lagrange multipliers. Due to its convexity, the above problem is solved exactly by an ascent-descent algorithm that alternates between updating the primal variable $a$ and performing gradient ascent on the dual variable $\lambda$ [6, 5]. For assumptions that represent probability distributions ($\mu$ and $p_X$), we used Entropic Mirror Descent [4] to solve the optimization problem under simplex constraints efficiently. For odds ratios $\varepsilon \in \mathbb{R}_+^{2d}$, we folded the non-negativity constraints into eq. (15) with its own Lagrange multipliers and used Gradient Descent to update $a$.

## F.2 Constrained Sampling

We compute a rejection-based Monte Carlo estimate [32] of the conditional expectation in eq. (4). Specifically, we drew samples from the given prior distribution, $A^{(i)} \sim \pi_A$ i.i.d., accepted them if they satisfy the constraint $\tau(A^{(i)}) \leq \delta$, and repeated this procedure until we obtain $M$ accepted samples, for some $M$. In all our experiments, we set $M = 5000$. Then, the BSV estimate is given by

$$\frac{1}{M} \sum_{i=1}^{N} \exp(-D(A^{(i)}|a)) \cdot \mathbb{I}(\tau(A^{(i)}) \leq \delta) \text{ where } N = \min \left\{ n \;\middle|\; \sum_{i=1}^{n} \mathbb{I}(\tau(A^{(i)}) \leq \delta) \geq M \right\}. \tag{16}$$

Hence, samples of the assumption parameters are accepted with the prior probability that they reverse the causal decision. Note that if this probability is very small, rejection sampling can be very inefficient. In this case, more sophisticated sampling techniques like Markov Chain Monte Carlo methods [14] may be adapted, though we found rejection sampling sufficient for our experiments.

# G Dataset Details

## G.1 Simulation Study

We simulated datasets with four binary covariates and binary treatments and outcomes, as follows.

1. Sample covariates independently as $X_i \sim \texttt{Ber}(\cdot)$, where there the Bernoulli parameter for each covariate is in $[0.4, 0.5, 0.6, 0.7]$ for $i \in \{1, 2, 3, 4\}$.
2. Simulate observational data by setting a propensity score between 0 and 1 that depends on covariates $X$, given by `ps = expit ( X.dot(t-coeff) )` and sampling treatments as $T \sim$ `Ber(ps)`. Here, we set `t-coeff` $= [0, -3, 0, 0]$ and `expit` denotes the logistic sigmoid function.
3. Assign outcomes based on treatments and covariates via a logistic model that includes non-linear interactions between treatment and covariates. Outcome $Y \sim$ `Ber ( expit (logits) )` where `logits = beta * T + X.dot(gamma) X.dot(delta) * T`. Here, we set `beta` $= 4$, `gamma` $= [1, -1, 1, 0]$, and `delta` $= [-2, -3, -1, -2]$.
4. To further simulate real-world selection biases, we sampled a mask $S \sim$ `Ber ( expit (sel-logits) )`, with `sel-logits = delta-y * Y + delta-t * T + X.dot(delta-x)`, and selected only datapoints for which $S^{(i)} = 1$. Here, we set `delta-y` $= 2$, `delta-t` $= 1$, and `delta-x` $= [10, 10, 5, 1]$.

For the setting with six covariates, we take $X_5$ and $X_6$ as copies of $X_3$ and $X_4$, respectively, and similarly repeat all the corresponding coefficients that determine relationships with $T$, $Y$, and $S$. Note that our choices of coefficients here ensure that treatment effects are heterogeneous across subgroups.

We used settings with varying numbers of covariates $(4, 6, 8)$ to explore how the dimensionality of the problem affects sensitivity analyses. For BSV computations, we considered an idealistic setting and constructed empirical prior distributions for $p_X$ and $\mu$ using samples from the true data-generation mechanism without confounding or selection biases. We note that constructing empirical priors on odds ratios $\varepsilon$ from data remains a challenge since they are inherently unobserved quantities. Hence, we use user-defined priors which are, in this case, truncated Gaussians, centered at $\mathbf{1}$, with scales $\sigma \in \{1.0, 1.5\}$.

## G.2 Diabetes Application

We followed the dataset curation and set-up of the Semaglutide vs. Tirzepatide experiment in Dhawan et al. [12].

1. There are five binary covariates constructed by binning values into discrete categories:
   - `Age`: $[\leq 45, > 45]$ years,
   - `Sex`: [Male,Female],
   - `BMI` (Body Mass Index): $[\leq 28.5, > 28.5]$ kgs per meter squared,
   - `HbA1c` (Glycated Hemoglobin): $[\leq 7.5, > 7.5]$ %, and
   - `Weight`: $[\leq 220, > 220]$ lbs.
2. Possible treatments are Semaglutide ($T = 0$) and Tirzepatide ($T = 1$).
3. Outcome is whether the user lost 5% or more of their intital weight ($Y = 1$) or not ($Y = 0$).

This dataset was constructed using unstructured text data found in the PushShift collection [3] of Reddit posts upto December 2022. It has been curated to contain submissions and comments that describe users' lived experiences with the treatments and outcome above. Hence, it is naturally prone to the selection biases typical of such a platform, and exemplifies a real-world use-case of the sensitivity analyses discussed in this work. Further, observational distributions were estimated by computing conditional probabilities given by a large language model (LLM), before plugging them into standard causal estimators. This leaves the estimated distributions prone to biases of the LLM as well.

## H Further Simulation Results

In our simulation study, we visualize subpopulation comparisons of sensitivity values with respect to $\varepsilon$, $\mu$, and $p_X$ in figs. 6 to 8, respectively. When comparing distributions over covariate pairs or triplets in fig. 7, the empirical BSV uncovers trends that are significantly different from not only worst-case sensitivity but also the BSV under uniform priors, highlighting the usefulness of priors that reflect real populations.

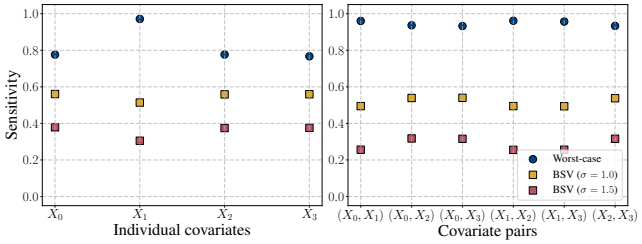
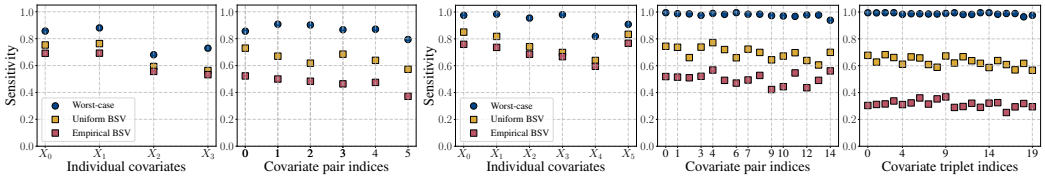



Figure 6: Worst-case sensitivity to unconfoundedness parameter $\varepsilon$ increases as number of observed confounders, and hence, dimension of $\mathcal{A}$, increases. BSV with respect to $\varepsilon(X)$ in the simulation study reveals a different trend.




(a) Simulation with four covariates

(b) Simulation with six covariates

Figure 7: Worst-case sensitivity to shifts in $p_X$ over single, pairs, or triplets of covariates becomes increasingly uninformative as the dimensionality of the corresponding assumption parameter spaces increases. Bayesian sensitivity is more informative than the worst-case, with empirical priors revealing different trends than uninformative ones.

Figures 6 and 7 show increasingly high sensitivity when considering joint distributions over more covariates and can not meaningfully distinguish between different sets of covariates, especially as the dimensionality of the corresponding assumption parameter space increases. In fact, worst-case sensitivity to unconfoundedness is greater in settings with pairs of observed covariates, relative to single observed covariates, which is unintuitive and unlikely to reflect real scenarios. We attribute this limitation to the large dimension of assumption parameter spaces, where the space of possible violations grows exponentially and a worst-case analysis becomes increasingly pessimistic, regardless of the plausibility of the worst-case violation in practice. However, BSV does not suffer as much from this curse of dimensionality and better distinguishes between subpopulations even in high-dimensional parameter spaces.

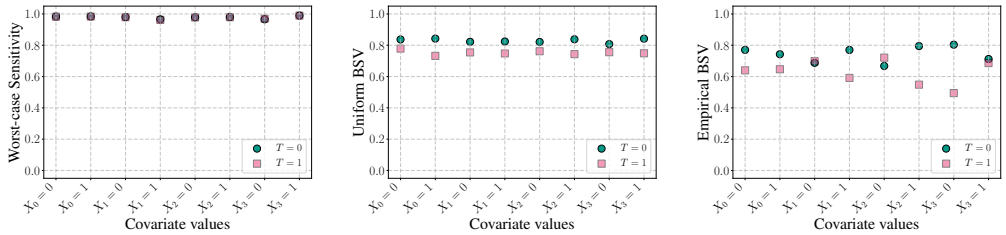



Figure 8: In the simulation study, BSV with respect to conditional outcome distributions in $\mu(X)$ attributes lower sensitivity to $T = 0$, while worst-case analyses attribute maximum possible sensitivity to all settings, making it impossible to meaningfully distinguish between subpopulations.

High sensitivity scenarios are precisely the ones where we would like to identify subpopulations for further experimentation or data collection. The assumption on conditional outcome distributions is especially prone to exhibiting high sensitivity because shifts in $\mu$ very easily shift the ATE across the decision threshold $\delta$. Figure 8 shows worst-case sensitivity values for this parameter for different covariate and treatment values, indicating highest possible sensitivity in every case. Without incorporating any information on how likely the shifts in this parameter space are in practice, the worst-case analysis tends to be binary in nature, attributing either zero or full sensitivity to different subpopulations. This would be uninformative for practitioners trying to decide where to allocate resources for more robust ATE estimation. In contrast, BSV also reveals lower overall sensitivity for the treatment $T = 0$, whereas worst-case sensitivity values are too pessimistic to find this insight.

# I Limitations

While this work has taken a critical view of worst-case sensitivity analyses, the proposed Bayesian Sensitivity Value is not without limitations in its real-world application. Our experiments suggest that the empirical BSV under data-driven priors can be practically useful. However, constructing such priors on odds ratios $\varepsilon$ from data remains fundamentally challenging since they are inherently unobservable quantities. Next, we considered and compared different subsets $\mathcal{A}$ of possible assumptions that were of the same form and admitted the same divergence metric, *e.g.* KL divergence for distributions lying in a simplex. However, combining different spaces of possible assumptions, along with their corresponding divergence metrics may allow practitioners to ask more interesting questions about sensitivity and capture interactions between different assumptions. Finally, the practical implementation of BSV computations is a simple rejection-based sampling technique, which can be very inefficient for small prior probability of reversing a causal decision. More involved and efficient techniques would help improve the adoption of this sensitivity criterion in practice.

Any sensitivity analysis framework has important societal implications for causal inference in high-stakes domains like healthcare. While the BSV can improve decision-making by providing more realistic sensitivity analyses and a better understanding of treatment effects across subpopulations, it also carries risks: the quality of analyses depends critically on the representativeness of real-world evidence used to construct priors, and the method could be misused to justify decisions that appear more robust than they actually are. To mitigate these risks, we recommend transparent documentation of prior construction and use of multiple sensitivity criteria, including our Bayesian criterion as well as traditional worst-case analyses.

## NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The derivation of our sensitivity analysis yields a more comprehensive and general framework and the empirical motivations and results support the claims about worst-case vs Bayesian analyses.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Limitations and directions for future improvement are briefly described in section 5, with more detailed discussion in appendix I.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: All the assumptions we consider are stated in section 2, with proofs deferred to appendices B to D.
   Guidelines:
   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: All dataset details and algorithm implementations to reproduce our experimental results are described in appendices F and G.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [No]
   Justification: The code will be released upon publication.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: All dataset details and algorithm implementations to reproduce our experimental results are described in appendices F and G.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: Wherever there is source of variability, variation is explicitly shown, as in fig. 3.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: The computations for all our experiments can be run on a single CPU.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: We have reviewed the NeurIPS Code of Ethics and our work conforms to it.
   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: Limitations and directions for future improvement are briefly described in section 5, with more detailed discussion of broader impacts in appendix I.
    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).
11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
    Answer: [NA]
    Justification: [NA]
    Guidelines:
    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: All contributors of existing assets used in this work are cited appropriately.
Guidelines:
- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [NA]
Justification: [NA]
Guidelines:
- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [NA]
Justification: [NA]
Guidelines:
- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification: [NA]
Guidelines:
- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.