DETECTING HARM BEFORE DATA POOLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pooling another center's data can lower a model's accuracy on its own population when the added data carry a distribution shift. We turn *will adding this source help?* into a decision a center can make before any joint training and without exchanging raw data. The recipient either *pools* when the donor is expected to improve its deployment performance, *declines* when the donor is expected to hurt it, or *abstains* when the available evidence is insufficient to make either decision. To enable this three-way decision, we define the *recipient-specific pooling effect* as the change in the recipient's deployment performance that would result from adding the donor's data to its training set, and estimate it using a single aggregated gradient returned by the donor at the recipient's model. This allows the recipient to assess the prospective effect of pooling without accessing the donor's raw records or performing joint training. We provide guarantees for detecting harmful pooling and reducing decision costs under stated conditions. Across five datasets spanning synthetic, natural-image, and medical-imaging settings, our method more accurately identifies beneficial and harmful data additions than baseline approaches and yields reliable pool, decline, and abstain decisions across varying levels of distribution shift.

1 INTRODUCTION

Multi-center collaboration is supposed to make models better: more labeled examples, more diversity, better generalization. In healthcare, this premise has motivated a wave of federated infrastructure built on the assumption that pooling across hospitals, even without moving raw records, should help. However, the premise is **not always true**. For example, Compton et al. (2023) show that pooling a second hospital's chest X-ray data produces even *worse* performance than training on one hospital alone, driven by distribution shifts such as imaging artifacts, device signatures, and acquisition patterns that a model picks up when they differ between sites. This phenomenon was later termed the *Data Addition Dilemma* (Shen et al., 2024): adding more data does not always improve performance, because it can harmfully change the composition of the training distribution.

This raises a practical question that arises **before** collaboration takes place. Consider a healthcare center $\boldsymbol{N}$ that has a limited labeled dataset and aims to train a model for its own deployment population. Another center $\boldsymbol{M}$ holds a larger dataset that could potentially be added through a joint training procedure. Before pooling, however, center $\boldsymbol{N}$ needs to decide whether center $\boldsymbol{M}$'s data are likely to *help* or *hurt* performance on center $\boldsymbol{N}$'s deployment population. Importantly, this decision may need to be made without accessing center $\boldsymbol{M}$'s raw data: in healthcare, cross-institutional access to patient-level records can involve data-use agreements, institutional review, and other governance requirements (Rieke et al., 2020; Kaissis et al., 2020).

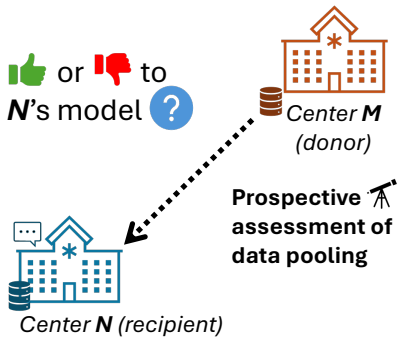

Figure 1: The data pooling problem.

Recipient-Specific Pooling Effect. The central quantity in this problem is not simply how different the donor and recipient distributions are or how valuable the donor data are in isolation. Instead, we ask a counterfactual, recipient-specific question: *what would happen to the recipient's deployment*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of approaches to assessing additional data.

Approach	Setting	Joint access[†]	Output
Gretton et al. (2012)	Compare two distributions	Yes	Distribution discrepancy
Shen et al. (2024)	Rank additional source data	Yes	Distribution discrepancy
Uncertainty sampling (Schein & Ungar, 2007)	Select unlabeled examples for labeling	No[††]	Uncertainty score
Koh & Liang (2017); Koh et al. (2019)	Reweight or remove points or groups	No[††]	Influence score
Basu et al. (2020)	Remove groups of training examples	No[††]	Influence score
Heo et al. (2026)	Group removal or addition; subset selection	No[††]	Influence score
Ours	Whether to add another center's data before joint training	No	Three-way decision: pool / decline / abstain

[†] Joint access means accessing examples from both centers within the same computation; the entries refer to the computations compared here, not impossibility results for distributed implementations.
[††] These scores admit adaptations using local computations and exchanged summaries. The cited methods do not specify our two-message protocol; the summaries required can differ, including group curvature for second-order parameter influence.

performance if the donor's data were added to the recipient's training set? We call this the *recipient-specific pooling effect*. Because this effect depends on the recipient's training distribution, model, and deployment objective, the same donor may help one center but hurt another. Existing approaches target related but different quantities: distribution discrepancy measures differences between data (Gretton et al., 2012; Shen et al., 2024), while uncertainty, influence, and data valuation characterize properties or contributions of examples, groups, or clients (Schein & Ungar, 2007; Koh & Liang, 2017; Koh et al., 2019; Basu et al., 2020; Heo et al., 2026). They therefore do not directly answer the prospective question faced by center $\boldsymbol{N}$: whether a particular donor will improve center $\boldsymbol{N}$'s deployment performance before its raw data are accessed or joint training is performed.

This motivates a decision problem rather than a ranking score. Given finite samples, the recipient should *pool* when the donor is expected to help, *decline* when it is expected to hurt, and *abstain* when the evidence is inconclusive. This three-way formulation is particularly appropriate when incorrect pooling can be costly, as in healthcare, where avoiding harmful data pooling may be preferable to making an unsupported binary decision.

Method. We propose a theoretically grounded pre-pooling assessment method. Center $\boldsymbol{N}$ (the recipient) sends its model to center $\boldsymbol{M}$ (the donor), which returns a single aggregated gradient evaluated at that model. Using this gradient and center $\boldsymbol{N}$'s own data, our method estimates the pooling effect and applies a three-way decision rule to pool, decline, or abstain, without exchanging raw records or conducting joint training. Our theory provides finite-sample guarantees for detecting harm and establishes consistency of the three-way decision rule. We also bound its error probabilities asymptotically and show that it can reduce decision costs.

Experiments. Across five datasets spanning synthetic, natural-image, and medical-imaging settings, our method outperforms baseline approaches in assessing whether additional data will help or hurt. Its three-way decision rule further provides reliable pool, decline, or abstain decisions across diverse levels of distribution shift.

2 Problem Setup

Center $\boldsymbol{N}$ trains a model that takes a feature vector $X \in \mathbb{R}^d$ as input and predicts a label $Y \in \{0, 1\}$. Center $\boldsymbol{N}$ holds n labeled examples (x, y) drawn from the joint distribution $P_{\boldsymbol{N}}$ and is considering adding a larger dataset of m examples from center $\boldsymbol{M}$, drawn from $P_{\boldsymbol{M}}$. The two centers address the same prediction task, but center $\boldsymbol{N}$ does not know whether or how $P_{\boldsymbol{M}}$ differs from $P_{\boldsymbol{N}}$. Its objective is to determine whether adding center $\boldsymbol{M}$'s data would improve performance on $P_{\boldsymbol{N}}$.

Let $\hat{\theta}_{\boldsymbol{N}}$ denote the model trained on center $\boldsymbol{N}$'s data alone. Write $\hat{\theta}_{\boldsymbol{N}\cup\boldsymbol{M}}$ for the hypothetical model that would result from applying the same training procedure to the pooled data. This pooled model cannot be obtained during our prospective assessment, since obtaining it would require the joint training whose value we seek to assess. Let $\ell(x, y; \theta)$ be the loss of the model with parameters θ on example (x, y), and write $R(\theta) = \mathbb{E}_{(X,Y)\sim P_{\boldsymbol{N}}}[\ell(X, Y; \theta)]$ for its expected loss on center $\boldsymbol{N}$'s population. The effect of adding center $\boldsymbol{M}$'s data is

$$\Delta_{\text{true}} := R(\hat{\theta}_{\boldsymbol{N}\cup\boldsymbol{M}}) - R(\hat{\theta}_{\boldsymbol{N}}). \tag{1}$$

Clearly, pooling helps when $\Delta_{\text{true}} < 0$, hurts when $\Delta_{\text{true}} > 0$, and has no effect when $\Delta_{\text{true}} = 0$. Both models are evaluated on $P_{\boldsymbol{N}}$, so the effect is specific to the recipient. This leads to our research question:

> *Can center **N** predict whether adding center **M**'s data will help or hurt its model on its own target population $P_{\boldsymbol{N}}$, before joint training and without sharing raw data, while abstaining when the evidence is insufficient?*

3 Pre-Pooling Assessment

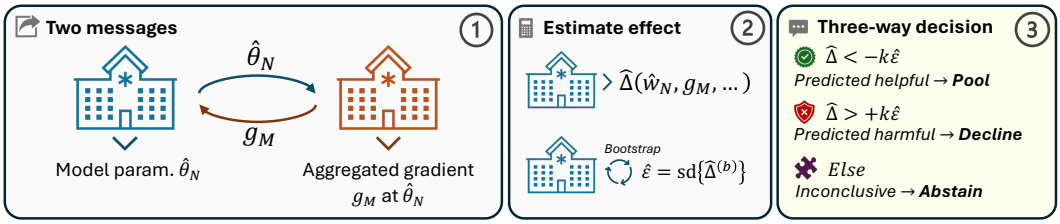

Figure 2: The pre-pooling assessment has three steps: exchange the fitted model and donor gradient; compute the pooling-effect score and its bootstrap standard deviation using recipient data with the donor message fixed; and compare the score with a variability-based threshold to pool, decline, or abstain.

Our pre-pooling assessment uses two messages to compute a pooling-effect score and decide whether to pool, decline, or abstain (Figure 2). We present it for logistic regression on a shared feature representation $x \in \mathbb{R}^d$, which may consist of raw covariates or frozen pretrained embeddings. For a label y, let $p_\theta(x) = \sigma(\theta^\top x)$, where σ is the sigmoid function, and let $\ell(x, y; \theta)$ be the binary cross-entropy loss. Center $\boldsymbol{N}$'s model is trained on its training set N_{fit} of size n_1 via $\hat{\theta}_{\boldsymbol{N}} = \arg\min_\theta \frac{1}{n_1} \sum_{i\in N_{\text{fit}}} \ell(x_i, y_i; \theta)$.

Step 1: Two Messages. Center $\boldsymbol{N}$ sends its fitted model $\hat{\theta}_{\boldsymbol{N}}$ to center $\boldsymbol{M}$. At these parameters, center $\boldsymbol{M}$ computes and returns its aggregated gradient $g_{\boldsymbol{M}}$ and sample size m to center $\boldsymbol{N}$:

$$g_{\boldsymbol{M}} = \sum_{i\in M} \nabla_\theta \ell(x_i, y_i; \hat{\theta}_{\boldsymbol{N}}) \in \mathbb{R}^d. \tag{2}$$

Step 2: Estimating the Pooling Effect. After receiving $g_{\boldsymbol{M}}$, center $\boldsymbol{N}$ computes the remaining quantities locally. It first computes the diagonal Fisher on its training set:

$$\hat{F}_{\boldsymbol{N},jj} = \frac{1}{n_1} \sum_{i\in N_{\text{fit}}} p_i(1 - p_i)x_{i,j}^2, \qquad p_i = p_{\hat{\theta}_{\boldsymbol{N}}}(x_i). \tag{3}$$

For logistic regression, these d entries form the diagonal of the empirical loss Hessian, providing a diagonal approximation to the recipient's loss curvature. On its validation set N_{val} of size n_2, center $\boldsymbol{N}$ computes the Fisher-normalized validation gradient

$$\hat{w}_{\boldsymbol{N}} = \left(\frac{1}{n_2} \sum_{i\in N_{\text{val}}} \nabla_\theta \ell(x_i, y_i; \hat{\theta}_{\boldsymbol{N}}) \right) \oslash \hat{F}_{\boldsymbol{N}}, \tag{4}$$

where $\oslash$ denotes coordinate-wise division by the Fisher diagonal. Using a separate validation set

measures sensitivity to changes in the fitted model on data not used to fit it.

Let $\alpha = m/(n_1 + m)$ be the donor's weight in the pooled training objective. An influence approximation gives the parameter change $\Delta\theta \approx -(\alpha/m)(g_{\boldsymbol{M}} \oslash \hat{F}_{\boldsymbol{N}})$. Expanding center $\boldsymbol{N}$'s risk to second order along this change gives the score

$$\hat{\Delta} = \underbrace{-\frac{\alpha}{m}\langle \hat{w}_{\boldsymbol{N}}, g_{\boldsymbol{M}} \rangle}_{\text{first-order term}} + \underbrace{\frac{1}{2}\left(\frac{\alpha}{m}\right)^2 \sum_{j=1}^{d} \frac{g_{\boldsymbol{M},j}^2}{\hat{F}_{\boldsymbol{N},jj}}}_{\text{curvature term}}. \tag{5}$$

The first term measures the effect of center $\boldsymbol{M}$'s gradient along directions in which center $\boldsymbol{N}$'s risk changes locally. The second accounts for curvature along the estimated parameter change. At a population risk minimum the first-order term vanishes, so retaining curvature is necessary to capture the risk increase from moving away from that minimum. Appendix A.2 gives the derivation and approximations. **A negative pooling effect score predicts that pooling will help, while a positive score predicts that pooling will hurt.** $\hat{\Delta} = 0$ **predicts no effect.**

To measure the score's variability under local resampling, center $\boldsymbol{N}$ independently resamples its training and validation sets with replacement, refits the baseline, and recomputes the Fisher diagonal, validation gradient, and pooling-effect score. Across a fixed number $B \geq 6$ of replicates, the bootstrap standard deviation is $\hat{\varepsilon} = \mathrm{sd}\{\hat{\Delta}^{(b)}\}_{b=1}^{B}$, using denominator B for the centered sum of squares. The original center $\boldsymbol{M}$'s message $g_{\boldsymbol{M}}$ remains fixed throughout, so this procedure requires no further communication and measures only the variability induced by local resampling. It does not resample center $\boldsymbol{M}$'s data or re-evaluate its gradient at the refitted parameters. The joint analysis below nevertheless includes the original donor sample's randomness.

Step 3: Three-Way Decision. We use $\hat{\varepsilon}$ as a local variability scale for setting the abstention threshold. For a threshold multiplier $k > 0$, the decision rule is

$$\mathcal{R}_{\text{3-way}}(k\hat{\varepsilon}) \begin{cases} \textsc{Helpful (Pool)}, & \hat{\Delta} < -k\hat{\varepsilon}, \\ \textsc{Harmful (Decline)}, & \hat{\Delta} > k\hat{\varepsilon}, \\ \textsc{Uncertain (Abstain)}, & |\hat{\Delta}| \leq k\hat{\varepsilon}. \end{cases} \tag{6}$$

Larger k widens the abstention region. Section 4 establishes consistency for fixed donors and analyzes calibration and decision cost under a local donor model. These results study the random threshold directly, without identifying $\hat{\varepsilon}$ with the finite-sample error bound used for sign recovery. We use $k = 1$ in the experiments as an abstention parameter, without assigning it a nominal confidence level. Algorithm 1 gives the procedure.

4 Theoretical Analysis

Our analysis explains why curvature is needed to detect harm, when the score identifies harmful pooling, and what the bootstrap threshold guarantees. We also establish when abstaining costs less than forcing a decision.

Let $\theta^*_{\boldsymbol{N}}$ minimize R, and let $\theta^*_{\boldsymbol{N}\cup\boldsymbol{M}}$ minimize $(1-\alpha)R + \alpha R_{\boldsymbol{M}}$ (with α as center $\boldsymbol{M}$'s fraction of the pooled training data, $\alpha = m/(n_1 + m)$), where $R_{\boldsymbol{M}}$ is the expected loss under $P_{\boldsymbol{M}}$. The population pooling effect is

$$\Delta^{\text{pop}}_{\text{true}} := R(\theta^*_{\boldsymbol{N}\cup\boldsymbol{M}}) - R(\theta^*_{\boldsymbol{N}}). \tag{7}$$

Let Δ_{pop} be the population version of the score equation 5, replacing the fitted baseline, Fisher diagonal, and empirical gradients with their population counterparts. Write $F^*_{\boldsymbol{N}} = \nabla^2 R(\theta^*_{\boldsymbol{N}})$, $D^* = \mathrm{diag}(1/F^*_{\boldsymbol{N},jj})$, and $\mu_{\boldsymbol{M}}(\theta^*_{\boldsymbol{N}}) = \mathbb{E}_{P_{\boldsymbol{M}}}[\nabla_\theta \ell(X, Y; \theta^*_{\boldsymbol{N}})]$. At center $\boldsymbol{N}$'s optimum, $\nabla R(\theta^*_{\boldsymbol{N}}) = 0$, so the first-order term vanishes and the population score consists entirely of curvature.

Proposition 1. *If $\theta^*_{\boldsymbol{N}}$ exists and $F^*_{\boldsymbol{N}}$ is positive definite, then*

$$\Delta_{pop} = \tfrac{1}{2}\alpha^2\, \mu_{\boldsymbol{M}}(\theta^*_{\boldsymbol{N}})^\top D^* \mu_{\boldsymbol{M}}(\theta^*_{\boldsymbol{N}}) \geq 0, \tag{8}$$

with equality if and only if $\mu_{\boldsymbol{M}}(\theta^*_{\boldsymbol{N}}) = 0$.

The population-level pooling effect is always nonnegative: pooling cannot improve on center $\boldsymbol{N}$'s (the recipient's) own risk minimum. Under the conditions in Appendix A.4, it is zero exactly when Δ_{pop} is zero. Thus a center $\boldsymbol{M}$ (a donor) that moves the population optimum increases recipient risk. Pooling can still help at **finite sample sizes**, where the fitted baseline differs from θ_N^* and additional data can reduce its estimation error.

Finite-Sample Sign Recovery. Both the computed score $\hat{\Delta}$ and the true pooling effect Δ_{true} depend on sampled data. The following result accounts for uncertainty in both quantities.

Theorem 1 (sign recovery, informal)**.** *For logistic fits under the conditions in Appendix A.1, $\hat{\Delta}$ and the true pooling effect Δ_{true} are both positive with probability at least $1-2\delta$, provided the population score exceeds a sufficient margin. With dimension, conditioning, comparison bound, sample-size ratios, and failure probability held fixed, this margin decreases as $O(n_1^{-1/2})$. The formal statement and proof are in Appendix A.7.*

The margin accounts for both sources of uncertainty: the score must exceed its sampling error, and population harm must remain harmful after fitting on finite data. The theorem therefore gives a sufficient resolution for detecting harm. It does not certify beneficial pooling.

Validity of the Three-Way Rule. The preceding bounds yield a theoretical threshold that certifies decline (Corollary 3); our decision rule further uses the bootstrap threshold $k\hat{\varepsilon}$. We therefore analyze this threshold directly. For a fixed donor that shifts the recipient's population optimum, the three-way rule correctly declines with probability tending to one as sample sizes grow. The reason is that the score converges to a positive value while its bootstrap variability vanishes. Proposition 2 states the conditions for this consistency result.

We next ask how to choose the bootstrap multiplier to control incorrect pooling and declining decisions. We study small donor shifts that shrink with sample size and remain comparable to estimation error, so pooling can still help or hurt. Appendix A.9.1 specifies the setting and assumptions for this analysis.

Corollary 1 (bootstrap calibration, informal)**.** *Under the model above, suppose the recipient population Fisher is diagonal in the score's feature basis. For any target error level $\beta \in (0, 1/2)$, an explicit multiplier k_β makes $\mathcal{R}_{3\text{-way}}(k_\beta\hat{\varepsilon})$ satisfy*

$$\limsup_{n_1\to\infty} \Pr[\textit{pool} \mid \Delta_{\textit{true}} > 0] \le \beta, \qquad \limsup_{n_1\to\infty} \Pr[\textit{decline} \mid \Delta_{\textit{true}} < 0] \le \beta.$$

Appendix A.9.3 gives the formal statement and proof, the multiplier k_β, and an asymptotic coverage guarantee for the interval $\hat{\Delta} \pm k\hat{\varepsilon}$.

Thus, among harmful pooling outcomes, the probability of choosing to pool is asymptotically at most β; the corresponding bound holds for declining helpful outcomes. These are not error rates among committed decisions. The experimental choice $k = 1$ is an abstention parameter, without a prescribed confidence level.

When the Three-Way Rule Reduces Decision Cost. Let $\mathcal{R}_{\text{binary}}$ pool for $\hat{\Delta} < 0$ and decline otherwise. Normalize the cost of declining a helpful donor to 1, and let r and c_a be the costs of pooling a harmful donor and abstaining, respectively.

Theorem 2 (three-way decision cost benefit, informal)**.** *Under the local model above, for each fixed B, $k > 0$, local shift, and finite $r \ge 0$, there is a strictly positive break-even abstention cost. For any fixed c_a below it,*

$$C(\mathcal{R}_{3\text{-way}}(k\hat{\varepsilon})) < C(\mathcal{R}_{\textit{binary}})$$

for all sufficiently large recipient sample sizes. This result does not require diagonal population Fisher information. The formal statement and proof are in Appendix A.9.4.

Abstention saves cost when the errors avoided on deferred trials outweigh the cost of deferral. The theorem establishes a positive range of abstention costs where this holds for our three-way rule with its bootstrap threshold. Appendix A.9 gives the joint analysis of the score, true effect, and threshold, with the formal statement and proof; it does not specify a numerical minimum sample size.

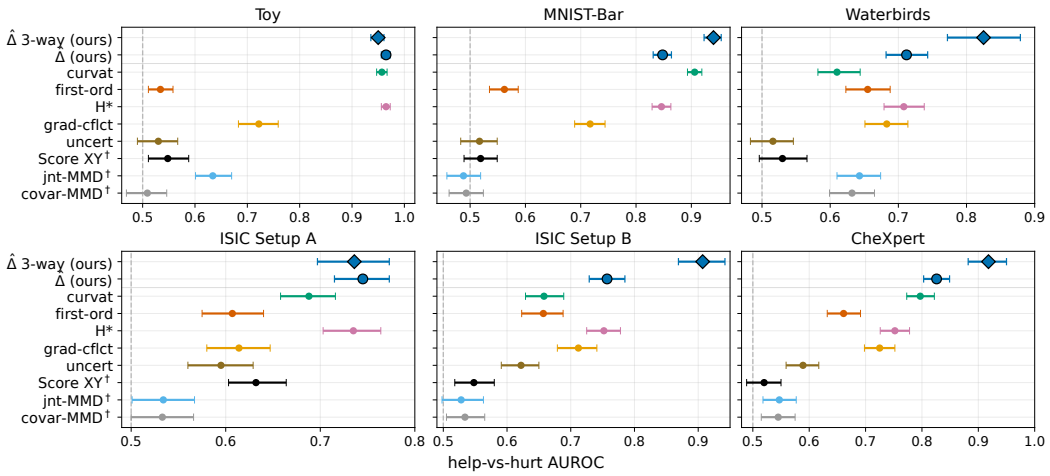

Figure 3: Help-vs-hurt AUROC per method and setup, with 95% bootstrap intervals over trials. Numbers are those of Table 8 in the appendix.

†Methods needing raw data from both centers are not admissible under the protocol. $\hat{\Delta}$ is our estimated pooling effect over all trials; $\hat{\Delta}$ *3-way* is restricted to the subset where the three-way rule commits at $k = 1$.

5 Experiments

We answer four questions in our experiments. **(EQ1)** How well does the proposed method separate donors that help from donors that hurt, measured against baselines that answer the same question by other means? **(EQ2)** Does the three-way rule abstain on less reliable predictions? **(EQ3)** When a committed prediction is wrong, is it wrong in the harmless direction? **(EQ4)** Is the curvature necessary, or would a first-order influence estimate alone predict the data pooling effect?

Datasets. We conduct experiments on six settings over five datasets, including natural images and medical data. **(1)** *Toy* is fully synthetic, with one feature coordinate that predicts the label in the donor and not in the recipient; **(2)** *MNIST-Bar* overwrites the top rows of each MNIST (LeCun et al., 1998) image with a bar whose intensity is correlated with the label in the donor and not in the recipient; **(3)** *Waterbirds* (Sagawa et al., 2020) for classifying landbirds and waterbirds; **(4-5)** *ISIC* (Ogier du Terrail et al., 2022; Tschandl et al., 2018; Codella et al., 2018; Hernández-Pérez et al., 2024) a multi-center dermatology benchmark, supplies two setups that differ only in how the donor is drawn: *Setup A* is natural heterogeneity, and *Setup B* is center-selection-induced shift. **(6)** *CheXpert* (Irvin et al., 2019) for classifying pleural effusion, where the donor shift is controlled through the correlation between view position and label. We fit a linear probe on frozen ResNet-50 (He et al., 2016) embeddings for Waterbirds, ISIC, and CheXpert. Sizes, constructions and other details are in Appendix B.1. We estimate Δ_{true} using the mean loss difference between the pooled and recipient-only models on a fresh recipient evaluation fold, disjoint from all data used for fitting or assessment, and use its sign as the help/hurt evaluation label. Reported intervals are 95% bootstrap intervals over trials.

EQ1: Separation Quality Against Baselines. To evaluate how well the score ranks harmful donors above helpful ones, we report *help-vs-hurt AUROC*: the probability that the score ranks a truly harmful donor above a truly helpful one when the two are drawn independently. We compare against covariate and joint MMD (*covar-MMD* and *jnt-MMD*) (Gretton et al., 2012); Score XY (Shen et al., 2024); an aggregated margin uncertainty score (*uncert*) adapted from Schein & Ungar (2007); gradient conflict (*grad-cflct*), adapted from Yu et al. (2020) by evaluating their conflict cosine once at $\hat{\theta}_N$ between center $\boldsymbol{M}$'s aggregated gradient and center $\boldsymbol{N}$'s validation gradient and keeping it continuous rather than thresholded into a binary gate; a full-Hessian oracle (H^*), which is our own score with the diagonal Fisher in equation 5 replaced by H^{-1}, a specialization of Heo et al. (2026)'s estimator rather than a reproduction of it; and the two terms of equation 5 run separately (*first-ord* and *curvat*), then our two-term score $\hat{\Delta}$ and the subset on which the three-way rule commits ($\hat{\Delta}$ *3-way*, $k = 1$). Three of these are not admissible under the protocol and are included as references: covariate and

joint MMD need raw features from both centers and Score XY needs raw features and labels.

Figure 3 gives the result. Our method leads the whole discrepancy family on all six settings. The reason is that a discrepancy answers a different question: it measures how far apart the two centers are, in feature space and without units, while the decision needs a signed quantity in the recipient's own risk. The diagonal approximation has negligible impact on accuracy: the full-Hessian oracle's interval overlaps our raw score on five setups and is *below* ours on the sixth, CheXpert, with disjoint intervals. Restricting to the subset on which the three-way rule commits raises separation on four of the six settings with disjoint intervals (MNIST-Bar, ISIC Setup B, CheXpert and Waterbirds), and leaves it statistically unchanged on the remaining two, Toy and ISIC Setup A, whose intervals overlap the raw score's. That row answers a deployment question rather than restating the one above it.

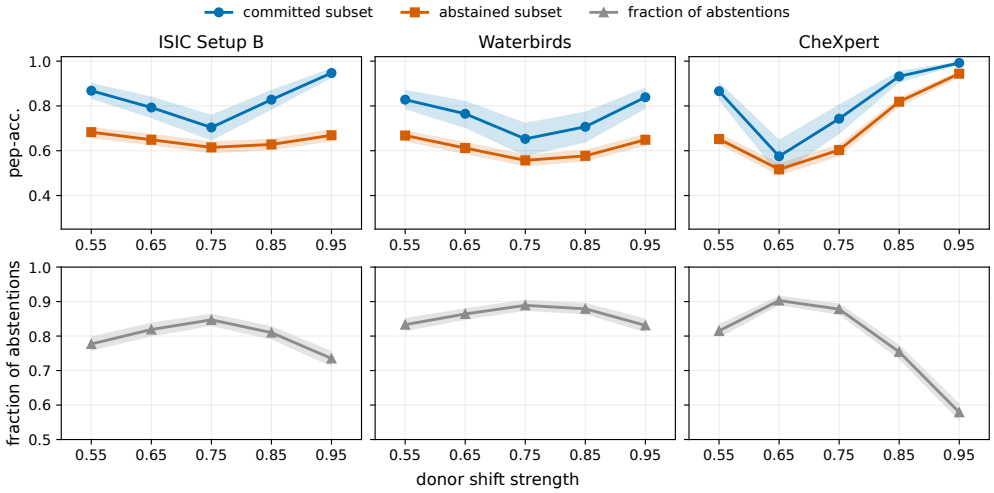

Figure 4: Pooling effect prediction accuracy (pep-acc.) on committed and on abstained trials (top row, 95% bootstrap CIs) against center $\boldsymbol{M}$ (donor) distribution shift strength, with the abstention rate (bottom row, 95% bootstrap CIs). Accuracy on the abstained subset is counterfactual; the rule does not commit there.

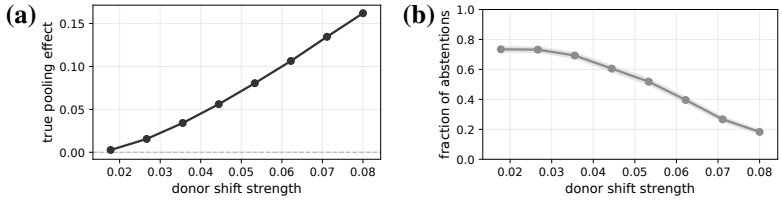

Figure 5: The abstention rate decreases as the true pooling effect becomes more strongly harmful. **(a)** The measured true pooling effect over trials increases as center $\boldsymbol{M}$ (donor) distribution shift strengthens. **(b)** At fixed sample sizes and $k = 1$, the abstention rate decreases as donor distribution shift strengthens.

EQ2: Does Abstention Identify Less Reliable Predictions? We test whether the abstention band identifies trials on which following the score's sign would be less reliable. At five fixed center $\boldsymbol{M}$ (donor) distribution shift strengths on ISIC Setup B, Waterbirds, and CheXpert, we report *pooling effect prediction accuracy* separately on committed and abstained trials, together with the abstention rate ($k = 1$). Accuracy on abstained trials is counterfactual: it measures whether the score's sign would have been correct had the rule been forced to decide. In Figure 4, committed accuracy exceeds abstained accuracy in all fifteen cells. On each setup, the highest abstention rate occurs at the distribution shift strength with the lowest committed accuracy. Thus, the band concentrates less reliable predictions while retaining more accurate decisions. Full tables, including the base rates, are in Tables 13–15 in the appendix.

Figure 5 shows how the abstention rate changes as the donor shift strengthens on MNIST-Bar. Over

the plotted range, which begins near the abstention peak, the abstention rate falls from 73.4% to 18.3%. Because the multiplier k and sample sizes are fixed, this change reflects more scores clearing their bootstrap thresholds. Stronger shift can therefore make the pooling decision easier: as harm becomes more apparent, the rule can decline more often instead of deferring.

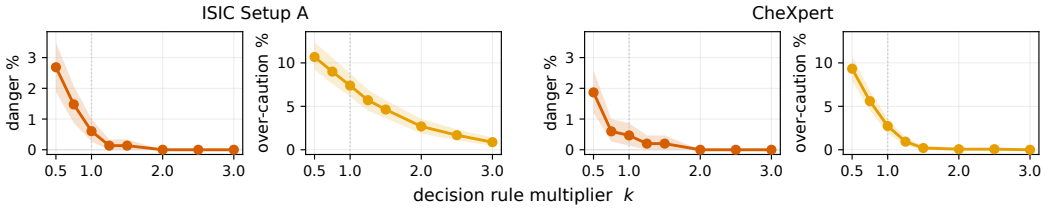

Figure 6: Danger and over-caution-type error rates, both as percentages of all trials, against the decision rule multiplier k. Danger-type error falls to zero by $k = 2$ on both setups and never exceeds over-caution, so tightening the rule buys safety by spending decisiveness.

EQ3: Fail-Safe. Committed errors come in two kinds and they are not equally bad. A *danger*-type error confidently pools a donor that truly hurts; an *over-caution*-type error confidently declines one that truly helps. We report both as percentages of all trials, so they stay comparable as the multiplier k grows. Figure 6 sweeps k on ISIC Setup A and CheXpert. At the deployed $k = 1$ the danger rate is 0.60% and 0.47%, respectively; and over-caution rates are 7.4% and 2.7%, respectively. Danger reaches zero on both by $k = 2$. The errors lie on the safe axis. Results on all setups are in Appendix B.5; the observed danger-type error rate approaches zero on all six setups as k increases.

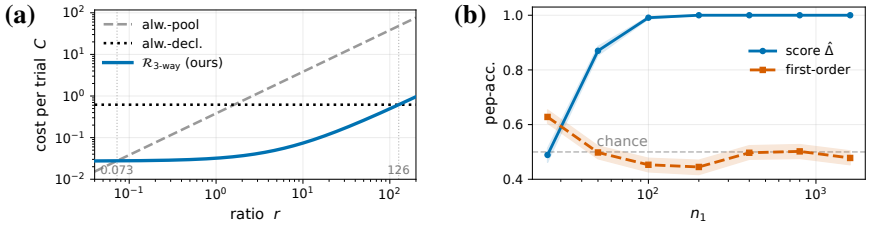

Figure 7: **(a)** With zero abstention cost, our three-way rule has lower average cost per trial than always-pool and always-decline for $r \in (0.073, 126)$ on CheXpert. **(b)** Ablating the curvature term leaves the prediction accuracy near chance at every sample size.

EQ3.1: Cost Model. Always-decline has zero danger-type error because it never pools, so the danger rate alone cannot rank decision rules. We compare rules using an asymmetric cost model, with the cost of an over-caution-type error normalized to 1 and the cost of a danger-type error equal to $r \geq 0$. Figure 7a sets the abstention cost to zero and shows that our rule has lower average cost than always-pool and always-decline on CheXpert for $r \in (0.073, 126)$. Among these rules, always-pool is cheapest at small r, while always-decline is cheapest at large r; our rule leads over the intermediate range. This comparison is restricted to the plotted rules: with free abstention, always-abstain has zero cost. Appendix B.6 reports the comparison on all setups and evaluates positive abstention costs against the same score forced to decide.

EQ4: Is the First-Order Term Enough? Figure 7b shows this on the toy setting, where the sample size can be swept over six doublings. The two-term score's ($\hat{\Delta}$, equation 5) pooling effect prediction accuracy climbs to 1.0 by $n_1 \approx 100$ to 200, while the same score with the curvature term deleted, computed from the same $\hat{w}_N$ and g_M on the same trials, stays between 0.445 and 0.628 at every size and shows no trend. More data does not rescue it.

6 Related Work

Distribution Similarity and Source Selection. Compton et al. (2023) show that combining hospital datasets can introduce spurious correlations and reduce worst-group accuracy. This is related to negative transfer: source data can degrade target performance (Wang et al., 2019). De Silva et al. (2023) show that small amounts of out-of-distribution data can help while larger amounts hurt. We

ask whether a particular donor will help before joint training and without exchanging raw data. Shen et al. (2024) address prospective source selection: their Score XY fits a classifier to distinguish sources from joint feature-label samples, then averages its source probabilities to rank candidate hospitals. MMD (Gretton et al., 2012) compares distributions through kernel mean embeddings and supports two-sample testing. Both statistics quantify distribution differences; neither directly estimates the signed recipient-risk change caused by pooling.

Influence Functions and Group Effects. Koh & Liang (2017) approximate the effect of reweighting a training example on a test prediction or loss without retraining. Koh et al. (2019) study group removal, measuring approximation accuracy and deriving conditions governing its relationship to retraining effects. These methods already provide signed, model-dependent estimates. In our setting, a first-order expansion of recipient risk at its population minimum misses the harm from moving away from that minimum. Basu et al. (2020) refine the group-induced parameter shift to second order, but their test-loss estimate still multiplies that shift by the target gradient. Our score instead retains the second-order change in the target risk itself.

In concurrent work, Heo et al. (2026) estimate the effects of group removal and addition and use pairwise interactions to guide greedy subset selection. Their experiments evaluate retraining-effect rankings and subset quality. Their analysis characterizes interactions between examples; it does not establish a three-way pooling decision with statistical guarantees against harmful pooling. We assess an entire donor before joint training, using two messages without sharing raw records, and explicitly abstain when the available evidence is insufficient. Under the conditions of Section 4, we establish finite-sample harm recovery, bootstrap consistency and asymptotic error calibration, and cost savings. Our full-Hessian baseline adapts their addition estimator to the recipient risk and pooling weight.

Theoretical Guarantees. Ben-David et al. (2010) bound target classification error using source error, domain divergence, and a joint optimal-error term, and analyze weighted combinations of labeled source and target data. Our target is the signed risk change from pooling relative to a recipient-only fit. Giordano et al. (2019) give finite-sample bounds on the parameter error of infinitesimal-jackknife approximations under controlled reweighting perturbations. Our analysis instead studies when a score computed before joint training correctly identifies harmful pooling and when abstention reduces decision cost.

Data Valuation. Data Shapley (Ghorbani & Zou, 2019) averages marginal performance contributions over subsets, while datamodels (Ilyas et al., 2022) predict model outputs from training-set composition. Their estimators use subset training or, for Gradient Shapley, sequential gradient updates. SaFE (Kumar et al., 2022) approximates institutional Shapley values in healthcare by ensembling local classifiers built on a trained federated model. LAVA (Just et al., 2023) values data through Wasserstein-distance sensitivity without specifying a learning algorithm. FedCS (Nishio & Yonetani, 2019) selects clients to accelerate federated training under resource constraints, whereas our method assesses a donor's pooling effect before joint training. FedBary (Li et al., 2024) assesses data before federated training using Wasserstein distance to a validation distribution or an estimated barycenter, exchanging interpolating distributions without raw records. Our method instead estimates the signed risk change from adding a donor to a recipient's existing fit and applies a three-way decision rule, using one returned donor gradient.

7 Conclusion

We present a method for assessing whether adding another center's data will help or hurt the recipient before joint training, using two messages without sharing raw records. A pooling-effect score supports a three-way decision to pool, decline, or abstain. Our analysis establishes finite-sample harm recovery and consistency of the three-way decision rule. Under local donor shifts, we derive asymptotic calibration with an additional curvature condition and cost savings for the three-way rule. Across six experimental settings, the score distinguishes helpful from harmful donors. The abstention experiments show higher prediction accuracy on committed than deferred trials, while the frequency and direction of errors vary across settings. The ablation demonstrates the importance of curvature for detecting harm that the first-order term misses. These results support prospective assessment before committing to collaboration and data pooling.

AI USE STATEMENT

In this work, we used generative AI tools to generate code for the synthetic data generation process (i.e., the Toy setup) based on a human-designed data-generating process, to generate some parts of the experimental code based on human-designed procedures, and to generate some parts of the code for figures based on human instructions and numerical assets provided by the authors. The generated code and figures were verified and adjusted (where applicable) by the authors, manually executed locally, and checked for correctness. No raw data from any dataset, including the Toy setup synthetic data, were exposed to or transmitted to the LLM. Additionally, we used an LLM to correct grammatical errors and improve the clarity and readability of the paper, based on text written by the authors. Its output and suggested edits were subsequently verified and incorporated only where appropriate by the authors. An LLM was also used to propose candidate paper titles, from which the authors reviewed and constructed the final title. We have reviewed all AI-assisted work, verified the generated code, and manually checked the resulting figures. We assume full responsibility for the final content of this work.

ETHICS STATEMENT

For all datasets used in this work other than the Toy setup synthetic dataset, we read and accepted the applicable licenses and data usage agreements and obtained access through formal applications where required. We strictly followed the guidelines and restrictions specified by the respective datasets. All such datasets were used locally within the permitted data usage agreements and were never exposed to or transmitted to any third party, including LLM-based systems. Consistent with these restrictions, our accompanying codebase points to the original hosts of the datasets rather than redistributing the datasets themselves. At the application level of our proposed method, i.e., the diagnostic score and the three-way decision rule, the method is designed to avoid the exchange of raw data. Nevertheless, users should consider any applicable restrictions on sharing models, gradients, or other derived information and should follow the precise ethical, legal, privacy, and data-governance requirements of their specific working scenarios.

REPRODUCIBILITY STATEMENT

We provide full proofs of all theoretical results and a detailed description of the experimental setup in the appendix, with corresponding pointers in the main paper. In addition, we provide a codebase as supplementary material to facilitate reproduction of the experimental results reported in the paper.

REFERENCES

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine Learning for Healthcare Conference*, pp. 110–127. PMLR, 2023.

Ashwin De Silva, Rahul Ramesh, Carey Priebe, Pratik Chaudhari, and Joshua T Vogelstein. The value of out-of-distribution data. In *International Conference on Machine Learning*, pp. 7366–7389. PMLR, 2023.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge University Press, 2019.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–487, 2014.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251. PMLR, 2019.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1139–1147. PMLR, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jaeseung Heo, Kyeongheung Yun, Youngbin Choi, Sehyun Hwang, Jungseul Ok, and Dongwoo Kim. Interaction-aware influence functions for group attribution. *arXiv preprint arXiv:2605.15675*, 2026.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific Data*, 11(1):641, 2024.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. LAVA: Data valuation without pre-specified learning algorithms. In *International Conference on Learning Representations*, 2023.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2 (6):305–311, 2020.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017.

Pang Wei Koh, Kai-Siang Ang, Hubert H.K. Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Sourav Kumar, Anantharaman Lakshminarayanan, Ken Chang, Feri Guretno, Ivan Ho Mien, Jayashree Kalpathy-Cramer, Pavitra Krishnaswamy, and Praveer Singh. Towards more efficient data valuation in healthcare federated learning using ensembling. In *International Workshop on Distributed, Collaborative, and Federated Learning*, pp. 119–129. Springer, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Wenqian Li, Shuran Fu, Fengrui Zhang, and Yan Pang. Data valuation and detections in federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12027–12036. IEEE, 2024.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–7. IEEE, 2019.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):119, 2020.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Andrew I Schein and Lyle H Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y. Chen. The data addition dilemma. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252. PMLR, 2024.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 180161, 2018.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11285–11294. IEEE, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

A Theoretical Appendix

Throughout, $x \in \mathbb{R}^d$ denotes the feature realization, $R(\theta)$ is center N's population risk and $R_M(\theta)$ is center M's (Section 2), $\alpha = m/(n_1 + m)$ is the mixing weight (held fixed for sign recovery and converging to α_∞ in Appendix A.9), and ρ bounds per-example gradient norms ($\|\nabla_\theta \ell\| = |p - y| \cdot \|x\| \le \rho$ under A2, since $|p - y| \le 1$). Constants $C_1, C_2, \ldots$ are absolute or depend only on ρ, m_{ratio}, the fold-proportionality constant c, and inverse powers of $\lambda_{\min}$; we track their provenance but do not optimize them. The $\lambda_{\min}$ dependence is real and worth naming rather than hiding: Lemma 1 carries $\lambda_{\min}^{-1}$, Lemma 3 carries $\lambda_{\min}^{-2}$, and the bound of Theorem 1 is therefore uniform in n_1 and δ at fixed $\lambda_{\min}$, not uniform over $\lambda_{\min} \to 0$.

A.1 Hypotheses of Proposition 1 and Theorem 1

We state assumptions A1–A4 below, followed by the additional sample-size and local-shift conditions for Theorem 1 and the assumptions used by each result.

- **(A1) Existence of the recipient optimum.** The recipient's population risk R attains its minimum at a finite parameter $\theta_N^* \in \mathbb{R}^d$. Appendix A.16 gives sufficient conditions for existence without requiring correctly specified logistic labels.
- **(A2) Bounded features.** $\|x\|_2 \le \rho$ almost surely under both P_N and P_M, so per-example gradients and Fisher contributions are bounded and the logistic risk is generalized self-concordant.
- **(A3) Conditioning.** The recipient's population loss Hessian $F_N^* = \nabla^2 R(\theta_N^*)$ has minimum eigenvalue $\lambda_{\min} > 0$. This is the population Fisher form used by the score, which need not equal the gradient covariance under misspecification.
- **(A4) No severe overestimation.** Let $\Delta_{\text{true}}^{\text{pop}} := R(\theta_{N\cup M}^*) - R(\theta_N^*)$ be the population pooling effect, where $\theta_{N\cup M}^*$ minimizes the population mixture $(1-\alpha)R + \alpha R_M$ at the protocol's fixed weight $\alpha = m/(n_1 + m)$. Define $\kappa := \Delta_{\text{true}}^{\text{pop}}/\Delta_{\text{pop}}$ for a donor with $\Delta_{\text{pop}} > 0$. By Proposition 1 and Corollary 2 the two scalars are nonnegative and vanish together, so this ratio exists and is positive for every donor: asserting that *some* $\kappa > 0$ works, or that one κ_0 serves any *finite* set of donors, would be vacuous. A4 is the substantive version, a bound $\kappa \ge \kappa_0 > 0$ holding *uniformly* over an infinite named family, which is what keeps the margin threshold $\varepsilon_{\text{res}}/\kappa_0$ bounded. We take $\mathfrak{D}(\rho, c_{\max}, s_{\text{loc,max}})$ to be the donors satisfying A2 with constant ρ, whose curvature is comparable to the recipient's ($F_M \preceq c_{\max} F_N^*$), and satisfying Lemma 8's smallness condition $s_{\text{loc}} \le s_{\text{loc,max}}$. It is not automatic over that family: $\mathfrak{D}$ is infinite and κ is a ratio of population quantities, so no finite measurement discharges it. Uniformity, not existence, is what the measurement speaks to: on the toy setup κ stays within 0.04–0.13 across a $20\times$ range of shortcut strength and both mixture ratios. Appendix A.12 separates the geometry contribution from the nonlinear correction and reports their measurements; the geometry bounds alone do not establish the assumed uniform comparison.

Sample-Size and Local-Shift Conditions. Theorem 1 further requires the mixture ratio held fixed ($m = m_{\text{ratio}}\, n_1$), the validation fold proportionally sized ($n_2 \ge c\, n_1$), the sample sizes satisfying the

fit-existence conditions of Lemmas 1 and 9 and the inverse-diagonal condition of Lemma 3, and the donor satisfying the smallness condition $s_{\text{loc}} := \alpha\rho\|\mu_M(\theta_N^*)\|/((1-\alpha)\lambda_{\min}) \le \frac{1}{6}$ of Lemma 8. Stated with the conditioning dependence that the main text suppresses, both $\varepsilon(n_1,\delta)$ and $\varepsilon_{\text{res}}(n_1,\delta)$ are $O\big(\lambda_{\min}^{-2}\sqrt{(d+\log(d/\delta))/n_1}\,\big)$ at fixed $\lambda_{\min}$. Appendix A.15 reports the realized s_{loc}.

Which Result Uses Which Hypothesis. Proposition 1 uses A1 only through the existence of θ_N^*, and A3, which makes D^* well defined since it needs $\min_j F^*_{N,jj} > 0$. Corollary 2 adds A2. Theorem 1 uses A1, A2, A3 and A4 together with the sample-size, sample-ratio, and local-shift conditions above, and Corollary 3 inherits them, except that its branches (ii) and (iii) need only Proposition 1 and the concentration bound. Appendix A.11 gives the lemma-level accounting.

Bootstrap Results. Proposition 2 uses A1–A3 with fixed donor and recipient distributions, a nonzero donor gradient, independent proportional samples, and fixed B. It does not use A4 or the sign-recovery localization condition. Theorem 3 and Theorem 2 instead use the correctly specified local model in Appendix A.9.1. Corollary 1 adds diagonal population Fisher information in the score's feature basis.

Model Specification. A1 permits misspecified logistic models: θ_N^* minimizes recipient log loss within the model class and need not generate the labels. First-order optimality centers the recipient gradient, while A2 and A3 supply the boundedness and curvature used in the concentration argument. Appendix A.16 explains this interpretation and gives a sufficient condition for A1. Theorem 2 retains its separate correctly specified logistic model in Appendix A.9.1.

A.2 Derivation of the Score

Pooling replaces N's objective with the mixture $(1-\lambda)\hat{R}(\theta) + \lambda\hat{R}_M(\theta)$, writing $\hat{R}$ for the fitting-fold empirical risk and $\hat{R}_M$ for M's. Here λ is a *free* weight, the coordinate we differentiate along; the protocol itself sits at the fixed $\lambda = \alpha = m/(n_1+m)$, which is why the two symbols are kept distinct. The path is the *empirical* one, since that is what the pooled fit actually follows: let $\theta(\lambda)$ solve $(1-\lambda)\nabla\hat{R}(\theta(\lambda)) + \lambda\nabla\hat{R}_M(\theta(\lambda)) = 0$, anchored at the fitted baseline $\theta(0) = \hat{\theta}_N$. Differentiating in λ gives, in general,

$$\theta'(\lambda) = -\big[(1-\lambda)\nabla^2\hat{R} + \lambda\nabla^2\hat{R}_M\big]^{-1}\big[\nabla\hat{R}_M(\theta(\lambda)) - \nabla\hat{R}(\theta(\lambda))\big].$$

Evaluated exactly at $\lambda = 0$, the M-side curvature term vanishes exactly (it is multiplied by $\lambda = 0$, not dropped as small) and $\nabla\hat{R}(\theta(0)) = 0$ by $\hat{\theta}_N$'s own first-order optimality on its fitting fold, leaving $\theta'(0) = -H_N^{-1}\nabla\hat{R}_M(\hat{\theta}_N)$ with $H_N := \nabla^2\hat{R}(\hat{\theta}_N)$ and no error at this step. The empirical objects are essential here: $\hat{\theta}_N$ minimizes $\hat{R}$, not the population risk R, so $\nabla R(\hat{\theta}_N) \neq 0$ in general. The population risk enters separately, in the definition of Δ_{pop} at θ_N^*, where $\nabla R(\theta_N^*) = 0$ does hold and is what Proposition 1 uses. The score then makes two Taylor steps. First, the displacement is extrapolated to the actual weight, $\theta(\alpha) - \theta(0) \approx \alpha\,\theta'(0) =: \Delta\theta$. Second, N's risk change is expanded to second order around $\hat{\theta}_N$:

$$R(\theta(\alpha)) - R(\hat{\theta}_N) \;\approx\; \nabla R(\hat{\theta}_N)^\top\Delta\theta \;+\; \tfrac{1}{2}\,\Delta\theta^\top H_{\text{val}}\,\Delta\theta, \tag{9}$$

where R is again N's population risk, its gradient estimated on the held-out validation fold and H_{val} the corresponding Hessian. Substituting $\Delta\theta = -\alpha H_N^{-1}\nabla\hat{R}_M(\hat{\theta}_N)$ and $\nabla\hat{R}_M(\hat{\theta}_N) = g_M/m$ exactly, by definition of g_M, yields the first-order term $(\alpha/m)(-\langle\hat{w}_N, g_M\rangle)$ and the curvature term $\frac{1}{2}(\alpha/m)^2 g_M^\top\hat{D}g_M$ of equation 5.

Two plug-in substitutions are definitional, baked into both $\hat{\Delta}$ and Δ_{pop}, so their fidelity cost lives in A4 rather than in the concentration analysis. (i) The full empirical Hessian is replaced by its diagonal $\hat{F}_N$. For logistic loss the empirical Hessian equals the empirical Fisher form $\frac{1}{n_1}\sum_i p_i(1-p_i)x_i x_i^\top$ exactly, with no dependence on y_i, so diagonalization is the only approximation in this step, not an additional Hessian-versus-Fisher gap. (ii) The validation-fold Hessian H_{val} in equation 9 is replaced by the same fitting-fold diagonal $\hat{F}_N$, justified because both estimate the same population Fisher F_N^* (both folds are drawn from P_N), which collapses $\hat{F}_N^{-1}H_{\text{val}}\hat{F}_N^{-1}$ to $\hat{F}_N^{-1}$ and yields the curvature term's simple form $\sum_j g_{M,j}^2/\hat{F}_{N,jj}$.

A.3 Proof of Proposition 1

Proof. Δ_{pop} is the population version of equation 5: θ^*_N in place of $\hat{\theta}_N$, $D^* = \operatorname{diag}(1/F^*_{N,jj})$ in place of $\hat{D}$, the population validation gradient $\mathbb{E}_{P_N}[\nabla_\theta \ell(\theta^*_N)] = \nabla R(\theta^*_N)$ in place of the empirical one, and $\mu_M(\theta^*_N)$ in place of g_M/m. Since θ^*_N minimizes R, first-order optimality gives $\nabla R(\theta^*_N) = 0$ exactly, so the population first-order term vanishes identically, for every donor M. What remains is $\Delta_{\text{pop}} = \frac{1}{2}\alpha^2 \mu_M(\theta^*_N)^\top D^* \mu_M(\theta^*_N)$, a quadratic form in the positive-definite diagonal matrix D^* (each $F^*_{N,jj} = e_j^\top F^*_N e_j \geq \lambda_{\min} > 0$ by A3). Hence $\Delta_{\text{pop}} \geq 0$, with equality if and only if $\mu_M(\theta^*_N) = 0$. □

A.4 Population Sign Agreement

Corollary 2 (population sign agreement)**.** *Under A1, A2, A3, for the unpenalized fit, with Δ^{pop}_{true} and $\theta^*_{N\cup M}$ as defined in A4: $\Delta^{pop}_{true} \geq 0$, and $\Delta^{pop}_{true} = 0 \iff \Delta_{pop} = 0 \iff \mu_M(\theta^*_N) = 0$. Consequently $\operatorname{sign}(\Delta^{pop}_{true}) = \operatorname{sign}(\Delta_{pop})$ always.*

Proof. A3 implies the covariate second moment $\mathbb{E}_{P_N}[xx^\top]$ is nondegenerate (a direction annihilating it would annihilate F^*_N). For any θ the population Hessian $\nabla^2 R(\theta) = \mathbb{E}_{P_N}[\sigma'(\theta^\top x)\, xx^\top]$ is then positive definite, since $\sigma' > 0$ pointwise, so R is strictly convex. A1 gives a finite minimizer θ^*_N, which is unique by strict convexity. Continuity gives a strictly positive minimum of $R(\theta) - R(\theta^*_N)$ on the unit sphere around θ^*_N. Convexity extends this lower bound linearly along each outward ray, so R is coercive. The mixture $(1-\alpha)R + \alpha R_M$ is bounded below by a coercive convex term plus a bounded-below convex term, so it attains its minimum, and it inherits strict convexity from the N-side term alone, $(1-\alpha)\nabla^2 R + \alpha \nabla^2 R_M \succeq (1-\alpha)\nabla^2 R \succ 0$, with no condition on M's Fisher. Its minimizer $\theta^*_{N\cup M}$ is therefore unique. The mixture gradient at θ^*_N is $(1-\alpha)\cdot 0 + \alpha\, \mu_M(\theta^*_N)$, so by uniqueness $\theta^*_{N\cup M} = \theta^*_N \iff \mu_M(\theta^*_N) = 0$. Finally $\Delta^{\text{pop}}_{\text{true}} = R(\theta^*_{N\cup M}) - R(\theta^*_N) \geq 0$ by optimality of θ^*_N, with equality iff $\theta^*_{N\cup M} = \theta^*_N$ by strict convexity. Combining with Proposition 1, both scalars vanish exactly on the event $\{\mu_M(\theta^*_N) = 0\}$ and are positive off it. □

A.5 Concentration Lemmas

Fix $\delta \in (0,1)$ and write θ^*_N for N's population optimum and $F^*_N = \nabla^2 R(\theta^*_N)$ for its population loss Hessian. The fitting, validation, and donor samples are mutually independent, with i.i.d. observations within each sample.

Exceptional-Sample Conventions. To define the theoretical random variables on every draw, set each empirical fit to the zero vector when its objective has no finite unique minimizer. Set $\hat{\Delta} = 0$ if the recipient fit fails or any Fisher diagonal entry is nonpositive. In the latter case also set $\hat{D} = 0$ and $\hat{w}_N = 0$ instead of taking an undefined inverse. The score is unchanged on samples with a finite unique recipient fit and positive Fisher diagonal. Lemmas 1 and 9 establish existence on the same concentration events used for their parameter bounds, so these conventions require no additional failure-probability allocation.

Lemma 1 (parameter estimation error)**.** *Under A1, A2, A3, for a sufficiently large absolute constant C and*

$$n_1 \geq C \frac{\rho^4}{\lambda^2_{\min}} \big(d + \log((d+1)/\delta)\big),$$

with probability at least $1-\delta$, the recipient empirical risk has a finite unique minimizer and $\|\hat{\theta}_N - \theta^*_N\|_2 \leq C_1 \sqrt{d + \log(1/\delta)} \,/\, (\lambda_{\min}\sqrt{n_1})$, where C_1 is an absolute constant times ρ.

Proof. Under A2, bounded gradients justify differentiation under expectation. A1 and first-order optimality therefore give $\mathbb{E}_{P_N}[\nabla_\theta \ell(X, Y; \theta^*_N)] = 0$. Define the empirical score at this fixed population parameter by

$$S_{n_1} := \nabla \hat{R}(\theta^*_N) = \frac{1}{n_1} \sum_{i \in N_{\text{fit}}} \nabla_\theta \ell(x_i, y_i; \theta^*_N).$$

Its summands have norm at most ρ. Rectangular matrix Bernstein (Tropp, 2012, Thm. 1.6), applied to the summands as $d \times 1$ matrices, gives $\|S_{n_1}\| \leq C' \rho \sqrt{(d + \log(1/\delta))/n_1}$ with failure probability

at most $\delta/2$ in the stated sample-size regime. A second matrix Bernstein bound applies to the empirical Hessian at θ_N^*: its summands are positive semidefinite, bounded above by $(\rho^2/4)I$, and have expectation $F_N^* \succeq \lambda_{\min} I$. Thus the sample-size condition, with C sufficiently large, gives the joint event

$$\nabla^2 \hat{R}(\theta_N^*) \succeq \frac{\lambda_{\min}}{2} I, \qquad \|S_{n_1}\| \le \frac{\lambda_{\min}}{4\rho},$$

with failure probability at most δ. Both quantities are defined without assuming that an empirical minimizer exists.

The empirical logistic risk is convex, three times continuously differentiable, and generalized self-concordant with constant ρ. On this event its squared Newton decrement at θ_N^* satisfies

$$\nu^2 := S_{n_1}^\top [\nabla^2 \hat{R}(\theta_N^*)]^{-1} S_{n_1} \le \frac{2\|S_{n_1}\|^2}{\lambda_{\min}} \le \frac{\lambda_{\min}}{8\rho^2} \le \frac{\lambda_{\min}(\nabla^2 \hat{R}(\theta_N^*))}{4\rho^2}.$$

Proposition 2, Eq. (7), of Bach (2010) therefore establishes a finite unique global minimizer and bounds its squared error in the empirical Hessian norm by $16\nu^2$. Consequently,

$$\|\hat{\theta}_N - \theta_N^*\|_2 \le \frac{4\|S_{n_1}\|}{\lambda_{\min}(\nabla^2 \hat{R}(\theta_N^*))} \le \frac{8\|S_{n_1}\|}{\lambda_{\min}}.$$

Substituting the score bound proves the claim. Existence follows on the same score and curvature events, without a separate label-overlap assumption. The lemma combines their failure probabilities into δ; Appendix A.7 budgets the two elementary events separately. □

Lemma 2 (Fisher concentration)**.** *Under A1, A2, A3 and the sample-size condition of Lemma 1, with probability at least* $1 - 2\delta$, $\max_j |\hat{F}_{N,jj} - F^*_{N,jj}| \le C_2\sqrt{(d + \log(d/\delta))/n_1}$.

Proof. $\hat{F}_N$ is evaluated at the fitted $\hat{\theta}_N$, so its summands are not independent and Bernstein does not apply directly. Decompose entrywise: $|\hat{F}_{jj}(\hat{\theta}_N) - F^*_{jj}| \le |\hat{F}_{jj}(\hat{\theta}_N) - \hat{F}_{jj}(\theta^*_N)| + |\hat{F}_{jj}(\theta^*_N) - F^*_{jj}|$. The first term is bounded by the Fisher entries' Lipschitz continuity in θ (constant at most $\rho^3/4$, since $|\frac{d}{dt}\sigma(t)(1-\sigma(t))| \le \frac{1}{4}$) times Lemma 1's bound, contributing the $\sqrt{d}$. The second term has genuinely i.i.d. summands at the fixed point θ^*_N, bounded by $\rho^2/4$, and concentrates by scalar Bernstein with a union bound over the d diagonal entries, contributing the $\log(d/\delta)$. Both terms are $O(1/\sqrt{n_1})$. □

Lemma 3 (inverse propagation)**.** *Write* $\hat{D} := \mathrm{diag}(1/\hat{F}_{N,jj})$ *and* $D^* := \mathrm{diag}(1/F^*_{N,jj})$*, the inverses of the diagonals, which are the quantities the score uses (*$D^* \neq (F^*_N)^{-1}$ *in general;* Δ_{pop} *is defined with* D^**, keeping both sides of the comparison consistent). Once* $n_1 \gtrsim (d + \log(d/\delta))/\lambda^2_{\min}$ *so that Lemma 2's bound is below* $\lambda_{\min}/2$*, on that event* $\|\hat{D} - D^*\|_{\mathrm{op}} \le (2C_2/\lambda^2_{\min})\sqrt{(d + \log(d/\delta))/n_1}$.

Proof. Each $F^*_{N,jj} \ge \lambda_{\min}$ by A3, and scalar inverse perturbation $|1/a - 1/b| = |a-b|/(ab)$ with $a \ge \lambda_{\min}$, $b \ge \lambda_{\min}/2$. □

Lemmas 1–3 control the first-order term's estimation error. The curvature term $g_M^\top \hat{D} g_M$ has two further error sources, M's sampling randomness in g_M and N's Fisher error propagating through $\hat{D}$, controlled by the next three bounds.

Lemma 4 (M-side concentration)**.** *Write* $\bar{g}_M := \mathbb{E}[g_M \mid \hat{\theta}_N] = m\,\mu_M(\hat{\theta}_N)$*. Conditional on* $\hat{\theta}_N$*,* $g_M - \bar{g}_M$ *is a sum of* m *i.i.d. mean-zero vectors bounded by* 2ρ*. Then with probability at least* $1-\delta$*,* $\|g_M - \bar{g}_M\|_2 \le C_4\rho\sqrt{m(1 + \log(1/\delta))}$.

Proof. By independence and mean-zero-ness, $\mathbb{E}\|\sum_k \xi_k\|^2 = \sum_k \mathbb{E}\|\xi_k\|^2 \le 4\rho^2 m$, so $\mathbb{E}\|g_M - \bar{g}_M\|_2 \le 2\rho\sqrt{m}$ by Jensen. Replacing one example changes $\|g_M - \bar{g}_M\|_2$ by at most $\|\xi_k - \xi_k'\| \le 4\rho$, so the bounded-difference inequality gives $\Pr\big[\|g_M - \bar{g}_M\| \ge \mathbb{E}\|g_M - \bar{g}_M\| + t\big] \le \exp(-t^2/(8m\rho^2))$, and $t = 2\rho\sqrt{2m\log(1/\delta)}$ suffices. The bound is dimension-free; the coordinate-wise and matrix-embedding routes would each cost a factor of $\sqrt{d}$ or $\sqrt{\log d}$ (Appendix A.11). □

Lemma 5 (quadratic fluctuation)**.** *Conditional on N_{fit}, the mean of $(g_M - \bar{g}_M)^\top \hat{D}(g_M - \bar{g}_M)$ is* $\mathrm{tr}(\hat{D}\ \mathrm{Cov}(g_M \mid N_{fit}))$*. On the events of Lemmas 3 and 4, its deviation from this conditional mean is at most $C_5\rho^2 m(1+\log(1/\delta))/\lambda_{\min}$. After the score's $\frac{1}{2}(\alpha/m)^2$ scaling, this contribution is $O(1/n_1)$ under the fixed-ratio convention.*

Proof. Conditioning on the full fitting sample fixes both $\hat{\theta}_N$ and $\hat{D}$, and $\bar{g}_M = \mathbb{E}[g_M \mid N_{\text{fit}}] = m\mu_M(\hat{\theta}_N)$. Independence of the donor observations and the bound on their raw gradients give

$$\mathbb{E}[\|g_M - \bar{g}_M\|^2 \mid N_{\text{fit}}] \le m\rho^2.$$

On the inverse event, $\|\hat{D}\|_{\text{op}} \le 2/\lambda_{\min}$, so the conditional mean of the nonnegative quadratic form is at most $2m\rho^2/\lambda_{\min}$. On the donor concentration event the form itself is at most $2C_4^2\rho^2 m(1+\log(1/\delta))/\lambda_{\min}$. Adding these two bounds bounds the absolute deviation, with $C_5 = 2(C_4^2+1)$. This uses no independence between gradient coordinates and introduces no additional concentration event (Appendix A.11). □

Lemma 6 (cross term)**.** *By Cauchy-Schwarz and $\|\bar{g}_M\| \le m\rho$: $|2\bar{g}_M^\top \hat{D}(g_M - \bar{g}_M)| \le (4C_4\rho^2/\lambda_{\min})\, m^{3/2}\sqrt{1+\log(1/\delta)}$ on the events of Lemmas 3 and 4. After scaling this is the dominant M-side contribution, $O(1/\sqrt{n_1})$.*

N-side propagation. On Lemma 3's event, $|g_M^\top(\hat{D} - D^*)g_M| \le \|g_M\|^2\|\hat{D} - D^*\|_{\text{op}} = O(m^2)\cdot O(\sqrt{(d+\log(d/\delta))/n_1})$, which after the $\frac{1}{2}(\alpha/m)^2$ scaling is $O(1/\sqrt{n_1})$.

Two further error sources enter Stage 1 of the theorem and are collected here rather than left inside its proof.

Lemma 7 (Stage-1 residuals)**.** *Under A1, A2, A3:* (i) *conditional on N_{fit}, the validation-fold summands are i.i.d. with mean $\nabla R(\hat{\theta}_N)$ and norm at most ρ, so with probability at least $1-\delta$,*

$$\left\|\hat{g}_{val}(\hat{\theta}_N) - \nabla R(\hat{\theta}_N)\right\|_2 \ \le\ 2\rho\sqrt{2\log((d+1)/\delta)/n_2},$$

which is $O(1/\sqrt{n_1})$ precisely because $n_2 \ge c\,n_1$; (ii) *each $\nabla_\theta \ell_j$ is $(\rho^2/4)$-Lipschitz by A2, so deterministically $\|g_M(\hat{\theta}_N) - g_M(\theta^*_N)\| \le \frac{m\rho^2}{4}\|\hat{\theta}_N - \theta^*_N\|$, which on Lemma 1's event is $O(m\sqrt{(d+\log(1/\delta))/n_1}/\lambda_{\min})$; and* (iii) *the validation gradient's own systematic part obeys $\|\nabla R(\hat{\theta}_N)\| \le \frac{\rho^2}{4}\|\hat{\theta}_N - \theta^*_N\| = O(\sqrt{(d+\log(1/\delta))/n_1}/\lambda_{\min})$ on the same event. Also on the events of Lemmas 3 and 4, $\|g_M\| \le \|\bar{g}_M\| + \|g_M - \bar{g}_M\| \le m\rho + C_4\rho\sqrt{m(1+\log(1/\delta))} = O(m)$, which is the bound the N-side propagation display uses.*

Proof. For (i), embed each centered summand as a $d \times 1$ matrix and apply the rectangular matrix Bernstein inequality (Tropp, 2012) exactly as in Lemma 1, with $\|\nabla_\theta \ell\| \le \rho$ under A2. Part (ii) is the mean value inequality applied coordinate-blockwise to the m summands of g_M, using $\|\nabla^2_\theta \ell\|_{\text{op}} = \sigma'\|x\|^2 \le \rho^2/4$, followed by Lemma 1. Part (iii) is the same smoothness applied to ∇R together with $\nabla R(\theta^*_N) = 0$, which is first-order optimality; it is what makes the first-order term's *deterministic* part vanish at the parametric rate, and without it (i) alone would not control that term. The final display is the triangle inequality with $\|\bar{g}_M\| = m\|\mu_M(\hat{\theta}_N)\| \le m\rho$ under A2. □

A.6 Pooled Estimation and Resolvability

We first isolate the displacement bound, because Lemma 9 needs it and it must therefore be established independently of Lemma 9. Its proof uses only the mixture's convexity and curvature floor, no concentration, and in particular neither Lemma 9 nor Lemma 10.

Lemma 8 (localization of the pooled optimum)**.** *Under A1, A2, A3, and under the smallness condition*

$$s_{loc} \ :=\ \frac{\alpha\rho\|\mu_M(\theta^*_N)\|}{(1-\alpha)\lambda_{\min}} \ \le\ \tfrac{1}{6},$$

the pooled population optimum satisfies $r_\alpha := \|\theta^*_{N\cup M} - \theta^*_N\| \le 4\alpha\|\mu_M(\theta^*_N)\|/\big((1-\alpha)\lambda_{\min}\big)$.

Proof. If $\mu_M(\theta_N^*) = 0$, mixture stationarity and uniqueness give $r_\alpha = 0$. Otherwise, the mixture $G := (1-\alpha)R + \alpha R_M$ is strictly convex with $\nabla G(\theta_N^*) = \alpha\mu_M(\theta_N^*)$, since $\nabla R(\theta_N^*) = 0$. On the ball $\|\theta - \theta_N^*\| \le t$ the self-concordant Hessian comparison $\nabla^2 R(\theta') \succeq e^{-\rho\|\theta'-\theta\|}\nabla^2 R(\theta)$ (Bach, 2010, Prop. 1, Eq. 6) and A3 give $\nabla^2 G \succeq (1-\alpha)e^{-\rho t}\lambda_{\min} I$. Hence for θ on the sphere of radius t,

$$G(\theta) - G(\theta_N^*) \;\ge\; -\alpha\|\mu_M(\theta_N^*)\|\, t \;+\; \tfrac{1}{2}(1-\alpha)e^{-\rho t}\lambda_{\min}\, t^2,$$

which is strictly positive, forcing the minimizer strictly inside the ball, as soon as $t\, e^{-\rho t} > 2a$ with $a := \alpha\|\mu_M(\theta_N^*)\|/((1-\alpha)\lambda_{\min})$. Writing $t = \kappa_0 a$ and $s_{\text{loc}} = \rho a$, this reads $\kappa_0 > 2e^{\kappa_0 s_{\text{loc}}}$. Since $\max_{\kappa_0} \kappa_0 e^{-\kappa_0 s_{\text{loc}}} = 1/(e s_{\text{loc}})$, attained at $\kappa_0 = 1/s_{\text{loc}}$, the condition is satisfiable exactly for $s_{\text{loc}} < 1/(2e) \approx 0.184$, and $s_{\text{loc}} \le \frac{1}{6}$ admits $\kappa_0 = 4$ because $2e^{2/3} = 3.895 < 4$. Taking $t = 4a$ gives the claim. □

Lemma 9 (pooled parameter estimation)**.** *Under A1, A2, A3, with α fixed and under the smallness condition $s_{loc} \le \frac{1}{6}$ of Lemma 8, suppose the sample size satisfies the condition below. Then, with probability at least $1-\delta''$, the pooled empirical risk has a finite unique unpenalized minimizer $\hat\theta_{N\cup M}$ satisfying $\|\hat\theta_{N\cup M} - \theta^*_{N\cup M}\| \le \varepsilon_{pool} = O\big(\sqrt{(d+\log(1/\delta''))/(n_1+m)}\big) = O(1/\sqrt{n_1})$ under the fixed-ratio convention. The hidden constant carries $1/\big((1-\alpha)\lambda'_R\big)$, where $\lambda'_R := e^{-\rho r_\alpha}\lambda_{\min}$ lower-bounds $\nabla^2 R$ at $\theta^*_{N\cup M}$ and the factor $(1-\alpha)$ is the mixture weight applied once, not twice. The smallness condition keeps the exponential bounded, since $\rho r_\alpha \le 4 s_{loc} \le \frac{2}{3}$ gives $\lambda'_R \ge e^{-2/3}\lambda_{\min}$. A sufficient sample-size condition is*

$$n_1 + m \ge C\frac{\rho^4}{((1-\alpha)\lambda'_R)^2}\big(d + \log((d+1)/\delta'')\big)$$

for a sufficiently large absolute constant C.

Proof. Write $G = (1-\alpha)R + \alpha R_M$ as in Lemma 8 and $\hat G = (1-\alpha)\hat R + \alpha \hat R_M$ for its empirical counterpart. The population minimizer exists by Corollary 2. Since $\alpha = m/(n_1+m)$, mixture stationarity gives the centered decomposition

$$\nabla\hat G(\theta^*_{N\cup M}) = \frac{1}{n_1+m}\Big[\sum_{i\in N_{\text{fit}}}\big(\nabla_\theta \ell(x_i,y_i;\theta^*_{N\cup M}) - \nabla R(\theta^*_{N\cup M})\big) + \sum_{i\in M}\big(\nabla_\theta \ell(x_i,y_i;\theta^*_{N\cup M}) - \nabla R_M(\theta^*_{N\cup M})\big)\Big].$$

Each summand is centered within its own stratum, has norm at most 2ρ, and is independent of the others. Rectangular matrix Bernstein therefore bounds the score by $C'\rho\sqrt{(d+\log(1/\delta''))/(n_1+m)}$ in the stated regime. Identical distributions across the two strata are not required.

The self-concordant Hessian comparison (Bach, 2010, Prop. 1, Eq. 6) and Lemma 8 give $\nabla^2 G(\theta^*_{N\cup M}) \succeq (1-\alpha)\lambda'_R I$. The empirical Hessian averages independent matrices bounded by $(\rho^2/4)I$, with mean $\nabla^2 G(\theta^*_{N\cup M})$. Matrix Bernstein and the score bound, each assigned failure probability $\delta''/2$, imply

$$\nabla^2\hat G(\theta^*_{N\cup M}) \succeq \frac{(1-\alpha)\lambda'_R}{2} I, \qquad \|\nabla\hat G(\theta^*_{N\cup M})\| \le \frac{(1-\alpha)\lambda'_R}{4\rho}$$

on a joint event of probability at least $1-\delta''$. The pooled empirical objective is generalized self-concordant with constant ρ. Applying the Newton-decrement argument of Lemma 1, with curvature floor $(1-\alpha)\lambda'_R$, therefore establishes a finite unique pooled fit and

$$\|\hat\theta_{N\cup M} - \theta^*_{N\cup M}\| \le \frac{8\|\nabla\hat G(\theta^*_{N\cup M})\|}{(1-\alpha)\lambda'_R}.$$

The score bound gives the stated rate. The pooled score and curvature events also control nonexistence, so no additional event is needed for the fit-existence conclusion. □

Lemma 10 (resolvability)**.** *Under A1, A2, A3 and the smallness condition $s_{loc} \le \frac{1}{6}$, on the intersection of Lemma 1's and Lemma 9's events,*

$$\left|\Delta_{true} - \Delta_{true}^{pop}\right| \;\le\; \varepsilon_{res} := \tfrac{\rho^2}{4}\, r_\alpha\, \varepsilon_{pool} + \tfrac{\rho^2}{8}\left(\varepsilon_{pool}^2 + \varepsilon_1^2\right),$$

with ε_1 Lemma 1's bound and r_α bounded by Lemma 8. Consequently, if $\Delta_{true}^{pop} > \varepsilon_{res}$ then $\Delta_{true} > 0$. Under the stated sample-size conditions, $\varepsilon_{res} = O(\lambda_{\min}^{-2}\sqrt{(d + \log(d/\delta))/n_1})$ at fixed conditioning; when $r_\alpha = 0$, only the squared-error terms remain.

Proof. Decompose

$$\Delta_{\text{true}} - \Delta_{\text{true}}^{\text{pop}} = \underbrace{\left[R(\hat\theta_{N\cup M}) - R(\theta^*_{N\cup M})\right]}_{\text{(A) pooled-fit fluctuation}} - \underbrace{\left[R(\hat\theta_N) - R(\theta^*_N)\right]}_{\text{(B) baseline excess risk}}.$$

Term (B) lies in $[0, \frac{\rho^2}{8}\varepsilon_1^2]$: nonnegative by optimality of θ^*_N, and second-order because $\nabla R(\theta^*_N) = 0$ with $\nabla^2 R \preceq \frac{\rho^2}{4} I$ under A2. This term is the finite-sample "pooling helps" channel of Section 4: pooling looks helpful relative to the noisy baseline exactly through the baseline's own excess risk, at $O(d/n_1)$, and this lemma is where that event formally lives. For term (A), the gradient of R at $\theta^*_{N\cup M}$ is not zero, but $\|\nabla R(\theta^*_{N\cup M})\| = \|\nabla R(\theta^*_{N\cup M}) - \nabla R(\theta^*_N)\| \le \frac{\rho^2}{4} r_\alpha$ by smoothness, so a Taylor expansion with the same smoothness bound gives $|(A)| \le \frac{\rho^2}{4} r_\alpha \varepsilon_{\text{pool}} + \frac{\rho^2}{8}\varepsilon_{\text{pool}}^2$ on Lemma 9's event. The bound on r_α is Lemma 8. The triangle inequality completes the bound, and $\Delta_{\text{true}} \ge \Delta_{\text{true}}^{\text{pop}} - \varepsilon_{\text{res}} > 0$ under the stated margin. □

At fixed dimension, conditioning, sample-size ratios, and failure probability, the resolvability bound has the form $\varepsilon_{\text{res}} = O(r_\alpha/\sqrt{n_1} + 1/n_1)$. For a compatible donor with $\mu_M(\theta^*_N) = 0$, Corollary 2 gives $r_\alpha = 0$, so the bound improves to $O(1/n_1)$. For a fixed donor with nonzero displacement, the bound is $O(n_1^{-1/2})$, matching the stated rate of the score-estimation bound ε. If $\Delta_{\text{true}}^{\text{pop}} \le \varepsilon_{\text{res}}$, the lemma's sufficient condition for a positive finite-sample pooling effect is not met. This does not determine the effect's sign or imply that the decision rule must abstain. Corollary 3 separately states the guarantees for decisions based on the computed score.

A.7 Formal Statement and Proof of Theorem 1

We give the precise version of the informal sign-recovery theorem in Section 4. For a donor with $\Delta_{\text{pop}} > 0$, write $\kappa := \Delta_{\text{true}}^{\text{pop}}/\Delta_{\text{pop}}$. Assumption A4 gives $\kappa \ge \kappa_0 > 0$ uniformly over its donor family.

Theorem 1 (sign recovery, formal statement)**.** *For the unpenalized logistic fits, under the conditions of Appendix A.1, with the mixture ratio held fixed and the validation set proportionally sized, there are bounds $\varepsilon(n_1, \delta)$ and $\varepsilon_{res}(n_1, \delta)$, both $O\big(\sqrt{(d + \log(d/\delta))/n_1}\big)$ at fixed conditioning, such that*

$$|\hat\Delta - \Delta_{pop}| \le \varepsilon(n_1, \delta), \qquad |\Delta_{true} - \Delta_{true}^{pop}| \le \varepsilon_{res}(n_1, \delta) \tag{10}$$

hold jointly with probability at least $1 - 2\delta$; the first bound alone holds with probability at least $1 - \delta$. Consequently, if

$$\Delta_{pop} > 2\max\left\{\varepsilon(n_1, \delta),\ \frac{\varepsilon_{res}(n_1, \delta)}{\kappa}\right\}, \tag{11}$$

then $\operatorname{sign}(\hat\Delta) = \operatorname{sign}(\Delta_{true})$ with probability at least $1 - 2\delta$.

The two bounds control score estimation and the fluctuation of the true retraining effect, respectively. Replacing the donor-specific κ by κ_0 gives a uniform sufficient margin over the donor family. The margin's constant depends on this comparison bound as well as the conditioning and sample-size ratios. Since $\Delta_{\text{pop}} \ge 0$, the conclusion recovers the sign of harmful effects. Lemma 10 in Appendix A.6 supplies the retraining-effect bound.

Seven concentration lemmas control $|\hat\Delta - \Delta_{\text{pop}}|$ in two stages, via self-concordance (Bach, 2010), matrix Bernstein for the diagonal Fisher (Tropp, 2012) and sub-Gaussian concentration for M's gradient; the sign then passes to Δ_{true} through the proved corollary, the assumed A4 inequality and the proved resolvability lemma.

Proof. The chain has four links; the first and fourth are probabilistic and proved, the second and third deterministic.

Link 1 (concentration). Two stages. Stage one carries $\hat{\Delta}$ to the intermediate quantity with M's finite-sample gradient evaluated at θ_N^*: Lemma 1 propagated through the validation gradient's Lipschitz continuity, Lemma 7(i) for that gradient's own $O(1/\sqrt{n_2})$ fluctuation (the reason the theorem requires $n_2 \geq c\, n_1$), Lemmas 2 and 3 with the N-side propagation bound, and Lemma 7(ii) for $g_M(\hat{\theta}_N) - g_M(\theta_N^*)$. Stage two carries the intermediate quantity to Δ_{pop}: Lemmas 4–6 scaled by $\frac{1}{2}(\alpha/m)^2$, with Lemma 6's cross term dominant at $O(1/\sqrt{n_1})$ and Lemma 5's fluctuation at $O(1/n_1)$. A union bound over the constituent events and a triangle inequality across the stages give $|\hat{\Delta} - \Delta_{\text{pop}}| \leq \varepsilon(n_1, \delta)$. Five events carry new randomness: the score concentration and the empirical-Hessian conditioning inside Lemma 1, whose statement absorbs both into a single δ; the Bernstein event inside Lemma 2, whose stated 2δ already reuses Lemma 1's; the validation fold's own gradient fluctuation; and Lemma 4. Lemmas 3, 5 and 6 are stated on those events and contribute no failure probability of their own. Invoking each of the five at level $\delta/6$ therefore gives total failure probability at most $5\delta/6 \leq \delta$, which is the theorem's $1 - \delta$ with one sixth of a unit of slack, and it costs a constant rather than a rate: every failure probability enters only through $\log(1/\delta)$ or $\log(d/\delta)$, and $\log(6d/\delta) = \log(d/\delta) + \log 6$, so the additive $\log 6$ is absorbed into C while the d-dependence and the $O\big(\sqrt{(d + \log(d/\delta))/n_1}\big)$ rate are unchanged. On this event the margin condition gives $\hat{\Delta} \geq \Delta_{\text{pop}} - \varepsilon > \varepsilon > 0$, so the score's sign is positive.

Link 2 (population sign agreement, proved). Corollary 2: $\Delta_{\text{pop}} > 0$ implies $\mu_M(\theta_N^*) \neq 0$ implies $\Delta_{\text{true}}^{\text{pop}} > 0$.

Link 3 (A4, the only assumed step). $\Delta_{\text{true}}^{\text{pop}} \geq \kappa\, \Delta_{\text{pop}} > 2\varepsilon_{\text{res}}$ under the combined margin condition, which is what the $\varepsilon_{\text{res}}/\kappa$ term in that condition buys.

Link 4 (resolvability, proved). Lemma 10 on its event (Lemma 9's δ''; Lemma 1's event is reused from link 1 and not recounted): $\Delta_{\text{true}} \geq \Delta_{\text{true}}^{\text{pop}} - \varepsilon_{\text{res}} > \varepsilon_{\text{res}} > 0$, so the true effect's sign is positive.

The recipient and pooled existence conclusions hold on the events already counted in links 1 and 4. Their sample-size conditions are imposed at these allocated failure levels. The two probabilistic links fail with total probability at most $\delta + \delta''$; taking $\delta'' = \delta$ gives the theorem's $1 - 2\delta$. □

A.8 Three-Way Certification and Proof of Corollary 3

Corollary 3 (three-way rule with a theoretical threshold)**.** *Under the conditions of Theorem 1, replace $k\hat{\varepsilon}$ in equation 6 with a deterministic threshold $t \geq 0$. On the joint error-bound event, of probability at least $1 - 2\delta$:*

(i) If $t \geq \varepsilon + \varepsilon_{res}/\kappa_0$, then $\hat{\Delta} > t$ implies $\Delta_{true} > 0$.

(ii) If $t \geq \varepsilon$, the condition $\hat{\Delta} < -t$ cannot occur.

(iii) If $|\hat{\Delta}| \leq t$, then $0 \leq \Delta_{pop} \leq t + \varepsilon$.

Proof. Write $E_1 = \{|\hat{\Delta} - \Delta_{\text{pop}}| \leq \varepsilon\}$ and $E_2 = \{|\Delta_{\text{true}} - \Delta_{\text{true}}^{\text{pop}}| \leq \varepsilon_{\text{res}}\}$ for the two events of Theorem 1, whose intersection has probability at least $1 - 2\delta$. Every part follows from these together with $\Delta_{\text{pop}} \geq 0$ (Proposition 1).

(ii) Pool is unreachable. On E_1, $\hat{\Delta} \geq \Delta_{\text{pop}} - \varepsilon \geq -\varepsilon \geq -t$, so $\hat{\Delta} < -t$ cannot occur. The containment $\{\hat{\Delta} < -t\} \subseteq E_1^c$ is unconditional, so a confident pool has probability at most δ for every donor. This uses neither A4 nor resolvability, only the one-sidedness of the population score and the score's concentration.

(iii) Abstention localizes the population score. On E_1, $\Delta_{\text{pop}} \leq \hat{\Delta} + \varepsilon \leq t + \varepsilon$, and $\Delta_{\text{pop}} \geq 0$ by Proposition 1, so the abstention region confines Δ_{pop} to $[0, t+\varepsilon]$. It does not bound Δ_{true}: that would need an upper companion to A4, which supplies only the lower bound $\kappa \geq \kappa_0$. Abstention therefore bounds the population score, while leaving the size of population harm unresolved. It can motivate collecting more recipient data, but does not establish that pooling is safe.

(i) Decline is certified. On E_1, $\hat{\Delta} > t$ gives $\Delta_{\text{pop}} \geq \hat{\Delta} - \varepsilon > t - \varepsilon \geq \varepsilon_{\text{res}}/\kappa_0$. By A4, $\Delta_{\text{true}}^{\text{pop}} \geq \kappa_0 \, \Delta_{\text{pop}} > \varepsilon_{\text{res}}$, and on E_2, $\Delta_{\text{true}} \geq \Delta_{\text{true}}^{\text{pop}} - \varepsilon_{\text{res}} > 0$. The trigger is no stronger than the theorem's margin condition, since $\varepsilon + \varepsilon_{\text{res}}/\kappa_0 \leq 2\max(\varepsilon, \varepsilon_{\text{res}}/\kappa_0)$; the difference is that it is stated on the observable score rather than on Δ_{pop}. □

A.9 Bootstrap Validity Under Local Donor Shifts

We analyze the recipient-only bootstrap of Section 3, keeping the original donor gradient fixed in every replicate. A joint limit for the true pooling effect, score, and bootstrap standard deviation yields both calibration under an additional curvature condition and the cost guarantee of Theorem 2. These results do not require A4. Appendix A.10 treats a fixed donor distribution separately.

A.9.1 Local Model and Bootstrap Procedure

Fix dimension $d \geq 1$, ratios $m_{\text{ratio}}, c_{\text{val}} \in (0, \infty)$, and a finite recipient parameter $\theta_N^* \in \mathbb{R}^d$. Let

$$m = \lfloor m_{\text{ratio}} n_1 \rfloor, \qquad n_2 = \lfloor c_{\text{val}} n_1 \rfloor, \qquad \alpha = \frac{m}{n_1 + m} \longrightarrow \alpha_\infty := \frac{m_{\text{ratio}}}{1 + m_{\text{ratio}}}.$$

Only positive fold sizes are considered. The ratio c_{val} is distinct from the lower bound c in $n_2 \geq c n_1$ used for sign recovery. Both centers have a common fixed covariate law P_X with $\|X\| \leq \rho$ almost surely. Recipient labels satisfy

$$Y \mid X = x \sim \text{Bernoulli}(\sigma((\theta_N^*)^\top x)), \qquad F_N^* := \mathbb{E}[\sigma'((\theta_N^*)^\top X) X X^\top] \succ 0.$$

For a fixed local shift $\mathbf{h} \in \mathbb{R}^d$, the donor has logistic parameter

$$\theta_{M,n_1}^*(\mathbf{h}) = \theta_N^* + \frac{(F_N^*)^{-1/2}\mathbf{h}}{\sqrt{n_1}}, \qquad Y \mid X = x \sim \text{Bernoulli}(\sigma(\theta_{M,n_1}^*(\mathbf{h})^\top x)). \tag{12}$$

The matrix square roots are symmetric positive definite. The three original samples are mutually independent and i.i.d. within each fold. Fits are exact empirical minimizers of logistic loss. All fitted parameters are included in the score; an intercept can be represented by a constant feature coordinate provided $F_N^* \succ 0$. This is a sequence of local experiments, with the donor approaching the recipient as n_1 grows.

Write $u = g_M/m$ and $v = \hat{g}_{\text{val}}(\hat{\theta}_N)$ for the donor and validation mean gradients, and let $\hat{D} = \text{diag}(1/\hat{F}_{N,jj})$. The score is exactly

$$\hat{\Delta} = -\alpha v^\top \hat{D} u + \frac{\alpha^2}{2} u^\top \hat{D} u, \qquad \Delta_{\text{true}} = R(\hat{\theta}_{N \cup M}) - R(\hat{\theta}_N). \tag{13}$$

Fix an integer $B \geq 6$, independent of n_1. For each $b = 1, \ldots, B$, independently resample n_1 fitting observations and n_2 validation observations with replacement. Refit the baseline, recompute the fitting Fisher diagonal and validation gradient at the refit, and retain the original u:

$$\hat{\Delta}^{(b)} = -\alpha (v^{(b)})^\top \hat{D}^{(b)} u + \frac{\alpha^2}{2} u^\top \hat{D}^{(b)} u, \qquad \hat{\varepsilon} = \left[\frac{1}{B} \sum_{b=1}^{B} (\hat{\Delta}^{(b)} - \overline{\hat{\Delta}}^{\,*})^2 \right]^{1/2}, \tag{14}$$

where $\overline{\hat{\Delta}}^{\,*} = B^{-1} \sum_b \hat{\Delta}^{(b)}$. Resampling indices are independent across folds and replicates and independent of original sampling. Probabilities below include both sources of randomness. The donor message is neither resampled nor reevaluated at the bootstrap fits.

To define the statistics on all samples, use zero as the fallback for any unavailable finite unique fit. If the baseline or any requested bootstrap fit fails, or any required Fisher denominator is nonpositive, set the score and threshold to zero and abstain. The pooled-fit fallback defines only Δ_{true} and is not checked by the assessment rule. The proof shows that all these exceptions have probability tending to zero. A procedure that skips failed replicates and abstains when at most five succeed agrees on the event that all B replicates succeed and has the same limits.

All limits hold at fixed model, dimension, sample-size ratios, B, and $k > 0$. Uniformity, where stated, is only over a fixed compact set of local shifts $\mathbf{h}$. The proofs also hold along any convergent sequence $\mathbf{h}_{n_1} \to \mathbf{h}$.

A.9.2 JOINT LIMIT OF THE EFFECT, SCORE, AND BOOTSTRAP THRESHOLD

Define

$$D^* = \mathrm{diag}(1/F^*_{N,jj}), \qquad J_{\mathrm{diag}} = (F^*_N)^{1/2} D^* (F^*_N)^{1/2}.$$

Let $\zeta_N, \zeta_M, \zeta_{\mathrm{val}}$ be independent $\mathcal{N}(0, I_d)$ vectors and $U \sim \chi^2_{B-1}$ be independent of them. Set

$$\mathsf{V}_{\mathbf{h}} = \alpha_\infty \left(\zeta_N - \mathbf{h} + \frac{\zeta_M}{\sqrt{m_{\mathrm{ratio}}}} \right), \tag{15}$$

$$\mathsf{H}_{\mathbf{h}} = -\zeta_N^\top \mathsf{V}_{\mathbf{h}} + \tfrac{1}{2} \|\mathsf{V}_{\mathbf{h}}\|^2, \tag{16}$$

$$\mathsf{S}_{\mathbf{h}} = -\left(\zeta_N + \frac{\zeta_{\mathrm{val}}}{\sqrt{c_{\mathrm{val}}}} \right)^\top J_{\mathrm{diag}} \mathsf{V}_{\mathbf{h}} + \tfrac{1}{2} \mathsf{V}_{\mathbf{h}}^\top J_{\mathrm{diag}} \mathsf{V}_{\mathbf{h}}, \tag{17}$$

$$\mathsf{Q}_{\mathbf{h}} = \sqrt{1 + c_{\mathrm{val}}^{-1}}\, \|J_{\mathrm{diag}} \mathsf{V}_{\mathbf{h}}\| \sqrt{U/B}. \tag{18}$$

The first two scalar limits are quadratic expressions in Gaussian vectors, not jointly Gaussian variables.

Theorem 3 (joint bootstrap limit)**.** *Under Appendix A.9.1, for every fixed* $\mathbf{h}$*,*

$$(n_1 \Delta_{\mathit{true}}, n_1 \hat{\Delta}, n_1 \hat{\varepsilon}) \;\Rightarrow\; (\mathsf{H}_{\mathbf{h}}, \mathsf{S}_{\mathbf{h}}, \mathsf{Q}_{\mathbf{h}}). \tag{19}$$

The same limit holds along every $\mathbf{h}_{n_1} \to \mathbf{h}$*. In particular, the bootstrap threshold has the* n_1^{-1} *scale in this local experiment. At fixed* B*, its limit retains the random factor* $\sqrt{U/B}$*.*

Proof. For brevity in this proof write $n = n_1$, $F = F^*_N$, $a = m_{\mathrm{ratio}}$, and $c_v = c_{\mathrm{val}}$. All remainder constants depend only on the fixed model, ratios, and a bound on $\|\mathbf{h}_n\|$.

Original Scores. The per-example gradient $\psi_\theta(x, y) = (\sigma(\theta^\top x) - y)x$ has norm at most ρ. The Hessian has norm at most $\rho^2/4$ and Lipschitz constant at most $\rho^3/4$. These integrable constant envelopes justify differentiation under expectation. Correct specification gives

$$\mathbb{E}_N \psi_{\theta^*_N} = 0, \qquad \mathrm{Cov}_N(\psi_{\theta^*_N}) = F. \tag{20}$$

Taylor expansion of the donor label probability yields, uniformly over bounded $\mathbf{h}_n$,

$$\mu_M(\theta^*_N) = -F^{1/2} \mathbf{h}_n / \sqrt{n} + O(n^{-1}). \tag{21}$$

The donor score covariance tends to F: its conditional label probability differs uniformly from the recipient's by $O(n^{-1/2})$, the score is bounded, and its mean outer product is $O(n^{-1})$.

Let $\bar{\psi}_N, \bar{\psi}_M, \bar{\psi}_{\mathrm{val}}$ be the sample mean gradients at θ^*_N and define

$$\zeta_{N,n} = -F^{-1/2} \sqrt{n}\, \bar{\psi}_N, \quad \zeta_{M,n} = F^{-1/2} \sqrt{m}\, (\bar{\psi}_M - \mu_M(\theta^*_N)), \quad \zeta_{\mathrm{val},n} = F^{-1/2} \sqrt{n_2}\, \bar{\psi}_{\mathrm{val}}.$$

These converge jointly to the independent standard Gaussians above. To verify the triangular-array step directly, each fixed scalar projection ξ of a centered summand is uniformly bounded. Its characteristic function at $t/\sqrt{q}$ is $1 - t^2 \mathbb{E}\xi^2/(2q) + O(q^{-3/2})$. Taking the qth power, using the logarithm near one, and using covariance convergence gives the Gaussian characteristic function. Independence gives the joint product. The multivariate characteristic-function convergence theorem applies because this limit is continuous at zero (Durrett, 2019).

Existence and Expansion of the Fits. At θ^*_N, the empirical Hessians of all original folds and the pooled sample converge to F by bounded entrywise variance and fixed dimension. Write $\lambda = \lambda_{\min}(F) > 0$. With probability tending to one, the fitting and pooled Hessians there exceed $3\lambda I/4$. On a fixed ball of radius r_0 with $(\rho^3/4) r_0 \le \lambda/4$, their Hessians exceed $\lambda I/2$. Both empirical scores at the center are $O_p(n^{-1/2})$. Thus, for either objective L_n, on the boundary,

$$L_n(\theta) - L_n(\theta^*_N) \ge -\|\nabla L_n(\theta^*_N)\| r_0 + \lambda r_0^2/4 > 0$$

with probability tending to one. A minimum on the closed ball is interior, stationary, and globally minimizing by convexity. Positive empirical Hessian implies a spanning design, so positive logistic weights make the objective strictly convex everywhere and the minimum unique.

The integrated Hessian bounds the fitted displacement by $2\|\nabla L_n(\theta_N^*)\|/\lambda = O_p(n^{-1/2})$. Its convergence to F then gives

$$F^{1/2}\sqrt{n}(\hat{\theta}_N - \theta_N^*) = \zeta_{N,n} + o_p(1),$$
$$F^{1/2}\sqrt{n}(\hat{\theta}_{N\cup M} - \theta_N^*) = (1-\alpha)\zeta_{N,n} + \alpha\mathbf{h}_n - \alpha\sqrt{n/m}\,\zeta_{M,n} + o_p(1).$$

Gradient expansion on the donor and validation folds gives

$$\begin{aligned} F^{-1/2}\sqrt{n}\,u &= \zeta_{N,n} - \mathbf{h}_n + \sqrt{n/m}\,\zeta_{M,n} + o_p(1), \\ F^{-1/2}\sqrt{n}\,v &= \zeta_{N,n} + \sqrt{n/n_2}\,\zeta_{\mathrm{val},n} + o_p(1). \end{aligned} \tag{22}$$

Also $\hat{D} \to_p D^*$. Substitution gives the score limit. The recipient-risk expansion has zero linear term, quadratic term $\mathbf{e}^\top F\mathbf{e}/2$, and remainder bounded by $\rho^3\|\mathbf{e}\|^3/24$. Applied to both fits, this gives the true-effect limit, jointly with the original scores; the rescaled cubic remainders are $o_p(1)$.

Conditional Bootstrap Fluctuations. Condition on all original observations, denoted by $\mathcal{D}_n$. At the original fitted parameter let

$$T_N^{(b)} = \sqrt{n}\{\nabla\hat{R}^{(b)}(\hat{\theta}_N) - \nabla\hat{R}(\hat{\theta}_N)\}, \qquad T_{\mathrm{val}}^{(b)} = \sqrt{n_2}\{\hat{g}_{\mathrm{val}}^{(b)}(\hat{\theta}_N) - v\}.$$

The centered resampled gradients have norm at most 2ρ. Their conditional covariances converge in probability to F. Indeed, replacing $\hat{\theta}_N$ by θ_N^* changes empirical second moments by at most a constant times $\|\hat{\theta}_N - \theta_N^*\|$. At the latter point the bounded law of large numbers gives F. The empirical fitting mean is zero on successful fits, and the validation mean is $O_p(n^{-1/2})$.

For fixed $t \in \mathbb{R}^d$, the same bounded characteristic-function expansion conditionally gives

$$\mathbb{E}_* e^{it^\top T_N^{(b)}} = \exp\{-\tfrac{1}{2}t^\top\hat{\Sigma}_N t + O(n^{-1/2})\},$$

with a uniformly bounded remainder for fixed t. The validation version uses n_2. Conditional independence across folds and replicates gives a product limit for $2B$ independent $\mathcal{N}(0,F)$ innovations.

This conditional convergence also gives independence from the original limiting scores. If $C_n(t)$ is the conditional characteristic function and $C(t)$ its deterministic limit, then $C_n \to_p C$ and $|C_n - C| \le 2$ imply

$$\mathbb{E}|C_n - C| \le \epsilon + 2\Pr[|C_n - C| > \epsilon] \to \epsilon.$$

Letting $\epsilon \downarrow 0$ gives L^1 convergence. For any original score vector Z_n,

$$|\mathbb{E}[e^{is^\top Z_n}C_n(t)] - C(t)\mathbb{E}[e^{is^\top Z_n}]| \le \mathbb{E}|C_n - C| \to 0.$$

Together with the original limit this proves the required joint convergence and independence.

Bootstrap Refits. On successful original fitting, the bootstrap score at $\hat{\theta}_N$ is $T_N^{(b)}/\sqrt{n} = O_p(n^{-1/2})$. Conditional bootstrap Hessian fluctuations have entrywise variance $O(n^{-1})$, and their conditional mean converges to F. Conditional Chebyshev followed by averaging gives joint-law Hessian convergence. The preceding ball argument centered at $\hat{\theta}_N$ gives finite unique bootstrap fits and displacement $O_p(n^{-1/2})$. A union bound over fixed B controls all replicates and positive diagonals.

Let E_b, W_b be the respective limits of $-F^{-1/2}T_N^{(b)}$ and $F^{-1/2}T_{\mathrm{val}}^{(b)}$. They are independent standard Gaussians, also independent of the original limiting scores. Integrated-Hessian expansion gives

$$F^{1/2}\sqrt{n}(\hat{\theta}_N^{(b)} - \hat{\theta}_N) \Rightarrow E_b, \qquad F^{-1/2}\sqrt{n}(v^{(b)} - v) \Rightarrow E_b + W_b/\sqrt{c_v}.$$

For the latter expansion, the resampled validation Hessian at the original fit converges to F by conditional variance bounds, and its Lipschitz bound preserves convergence along the fitted displacement. Each $\hat{D}^{(b)} \to_p D^*$ as well. All statements hold jointly at fixed B and along every convergent local-shift sequence.

Finite-Replicate Standard Deviation. On valid samples the exact difference is

$$\begin{aligned} \hat{\Delta}^{(b)} - \hat{\Delta} = &-\alpha(v^{(b)} - v)^\top\hat{D}^{(b)}u - \alpha v^\top(\hat{D}^{(b)} - \hat{D})u \\ &+ \tfrac{1}{2}\alpha^2 u^\top(\hat{D}^{(b)} - \hat{D})u. \end{aligned}$$

The last two terms are $o_p(n^{-1})$, since $u, v = O_p(n^{-1/2})$ and the matrix difference is $o_p(1)$. Therefore, jointly over b and with the original limits,

$$n(\hat{\Delta}^{(b)} - \hat{\Delta}) \Rightarrow L_b := -(E_b + W_b/\sqrt{c_v})^\top J_{\text{diag}} \mathsf{V}_{\mathbf{h}}.$$

Given the original Gaussians, these are i.i.d. centered normals with variance $\tau_{\mathbf{h}}^2 = (1 + c_v^{-1})\|J_{\text{diag}}\mathsf{V}_{\mathbf{h}}\|^2 > 0$ almost surely. Positivity follows from $J_{\text{diag}} \succ 0$ and the nondegenerate Gaussian law of $\mathsf{V}_{\mathbf{h}}$. The conditional law of $(L_b/\tau_{\mathbf{h}})_{b=1}^B$ is fixed standard normal, so this vector is independent of the original variables. An orthogonal rotation with first axis proportional to the all-ones vector makes its centered sum of squares the sum of squares of $B-1$ independent standard normals. Thus

$$\frac{1}{B}\sum_b (L_b - \bar{L})^2 = \tau_{\mathbf{h}}^2 U/B$$

in distribution jointly with the original variables. Continuity and shift invariance of the sample-standard-deviation map prove equation 19. This argument requires no convergence of bootstrap second moments over infinitely many replicates.

Finally, exceptional-sample conventions do not change these limits. If two versions X_n, Y_n, already including any deterministic rescaling, agree outside an event E_n with probability tending to zero, then for every bounded continuous f,

$$|\mathbb{E}f(X_n) - \mathbb{E}f(Y_n)| \le 2\|f\|_\infty \Pr[E_n] \to 0.$$

For bounded decision costs the analogous bound applies to expectations. Original and bootstrap failures vanish along every convergent shift sequence. The recipient and bootstrap laws do not depend on the donor shift, and bounded shifts control the pooled gradient in the same ball argument. This also supplies the sequential property used for compact uniformity below. □

A.9.3 Calibration and Interval Coverage

Corollary 1 (formal statement)**.** *Under Appendix A.9.1, additionally suppose F_N^* is diagonal in the feature basis used by the score, so $J_{\text{diag}} = I_d$. Let F_{t_ν} be the CDF of $Z/\sqrt{W/\nu}$ for independent $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_\nu$, and define*

$$q_B(k, c_{val}) = F_{t_{B-1}}\left(-k\sqrt{(1 + c_{val})(B-1)/B}\right). \tag{23}$$

For each fixed local shift and the rule $\mathcal{R}_{3\text{-way}}(k\hat{\varepsilon})$,

$$\limsup_{n_1\to\infty} \Pr[\textit{pool}, \Delta_{\textit{true}} > 0] \le q_B \Pr[\mathsf{H}_{\mathbf{h}} > 0], \tag{24}$$
$$\limsup_{n_1\to\infty} \Pr[\textit{decline}, \Delta_{\textit{true}} < 0] \le q_B \Pr[\mathsf{H}_{\mathbf{h}} < 0].$$

Both limiting true-sign probabilities are positive. Consequently, each error probability conditional on its corresponding true sign has limsup at most q_B, and the combined probability of an incorrect committed decision has limsup at most q_B. For $\beta \in (0, 1/2)$, choosing

$$k \ge k_\beta := \frac{t_{B-1,1-\beta}}{\sqrt{(1 + c_{val})(B-1)/B}} \tag{25}$$

makes these bounds at most β. They also hold uniformly over any fixed compact set of local shifts.

For each fixed local shift, there is also the studentized limit

$$\sqrt{(1 + c_{val})(B-1)/B}\,\frac{\hat{\Delta} - \Delta_{\textit{true}}}{\hat{\varepsilon}} \Rightarrow t_{B-1}, \tag{26}$$

where the ratio is set to zero if $\hat{\varepsilon} = 0$. Thus for fixed $k > 0$,

$$\Pr\big[\Delta_{\textit{true}} \in [\hat{\Delta} - k\hat{\varepsilon}, \hat{\Delta} + k\hat{\varepsilon}]\big] \longrightarrow 1 - 2q_B(k, c_{val}). \tag{27}$$

For limiting coverage $1-\beta$, with $\beta \in (0,1)$, use the multiplier $t_{B-1,1-\beta/2}/\sqrt{(1 + c_{val})(B-1)/B}$.

The directional bounds condition on the true effect's sign, not on commitment or the selected action. Coverage concerns the random population-risk retraining effect, averaging over original sampling and bootstrap randomization. It is not a conditional guarantee given the observed donor message or a guarantee for a finite evaluation-fold estimate. Diagonal feature covariance alone does not imply diagonal logistic Fisher information.

Proof. Under $J_{\mathrm{diag}} = I_d$, let $s_{\mathbf{h}} = \|\mathsf{V}_{\mathbf{h}}\|/\sqrt{c_{\mathrm{val}}} > 0$. Then

$$\mathsf{S}_{\mathbf{h}} = \mathsf{H}_{\mathbf{h}} + s_{\mathbf{h}} Z, \qquad \mathsf{Q}_{\mathbf{h}} = s_{\mathbf{h}}\sqrt{1 + c_{\mathrm{val}}}\sqrt{U/B},$$

where Z is standard normal independent of $(\zeta_N, \mathsf{V}_{\mathbf{h}}, U)$. Indeed, the signed projection of the independent validation Gaussian onto $\mathsf{V}_{\mathbf{h}}/\|\mathsf{V}_{\mathbf{h}}\|$ has a conditional standard normal law independent of the conditioning variables. Given these variables and $\mathsf{H}_{\mathbf{h}} > 0$,

$$\Pr[\mathsf{S}_{\mathbf{h}} < -k\mathsf{Q}_{\mathbf{h}} \mid \zeta_N, \mathsf{V}_{\mathbf{h}}, U] = \Phi\Big(-k\sqrt{1 + c_{\mathrm{val}}}\sqrt{U/B} - \mathsf{H}_{\mathbf{h}}/s_{\mathbf{h}}\Big) \le \Phi\Big(-k\sqrt{1 + c_{\mathrm{val}}}\sqrt{U/B}\Big).$$

Normal symmetry gives the same bound for an erroneous decline when $\mathsf{H}_{\mathbf{h}} < 0$. Averaging over the independent U gives q_B by the defining Student ratio, while independence from the true sign gives the factors in equation 24.

To pass from the limit to decisions, note that $(\zeta_N, \mathsf{V}_{\mathbf{h}})$ has a strictly positive joint Gaussian density, since its defining map from (ζ_N, ζ_M) is invertible. Both signs of $\mathsf{H}_{\mathbf{h}}$ occur on open sets: for any nonzero $\mathbf{v}$, the choices $(\zeta_N, \mathsf{V}) = (0, \mathbf{v})$ and $(\mathbf{v}, \mathbf{v})$ give opposite signs. Conditional on $\mathsf{V} = \mathbf{v} \neq 0$, the covariance of ζ_N is $I_d/(1 + m_{\mathrm{ratio}})$, so $\Pr[\mathsf{H}_{\mathbf{h}} = 0] = 0$. Conditional normality also gives $\Pr[\mathsf{S}_{\mathbf{h}} = 0] = \Pr[\mathsf{S}_{\mathbf{h}} = \pm k\mathsf{Q}_{\mathbf{h}}] = 0$. Theorem 3 and continuity sets therefore give convergence of the relevant probabilities. Summing the two error bounds and dividing by the positive limiting sign probabilities prove the combined and conditional conclusions.

For compact uniformity, couple the limits through the same Gaussian variables and U. As $\mathbf{h}_j \to \mathbf{h}$, the limit variables converge pointwise, and the event indicators converge almost surely off the null boundaries just proved. Dominated convergence with envelope one gives continuous event probabilities. If their finite-sample convergence were not uniform on a compact set, a convergent subsequence of offending shifts would contradict the sequential version of Theorem 3 and this continuity. The limiting true-sign probabilities have positive compact-set minima. Their finite-sample counterparts are at least half those minima for sufficiently large n_1, so division preserves uniform convergence.

Finally, $\sqrt{(1 + c_{\mathrm{val}})(B-1)/B}(\mathsf{S} - \mathsf{H})/\mathsf{Q} = Z/\sqrt{U/(B-1)}$. Ratio continuous mapping applies because $\mathsf{Q} > 0$ almost surely. Moreover, for any $\epsilon > 0$,

$$\limsup_{n_1} \Pr[\hat{\varepsilon} = 0] \le \limsup_{n_1} \Pr[n_1\hat{\varepsilon} \le \epsilon] \le \Pr[\mathsf{Q} \le \epsilon],$$

which tends to zero as $\epsilon \downarrow 0$. Thus the zero-denominator convention is negligible. Continuous Student endpoints and symmetry prove equation 27 and explain the $\beta/2$ quantile for a two-sided interval. The probability tools used here are standard continuity-set, continuous-mapping, and dominated-convergence arguments Durrett (2019). □

A.9.4 Cost Savings with the Bootstrap Threshold

Normalize the cost of declining a helpful source to one, let $r \ge 0$ be the cost of pooling a harmful source, and let $c_a \ge 0$ be the cost of abstention. For fixed finite costs,

$$C(\mathcal{R}) = r\Pr[\text{danger}] + \Pr[\text{over-caution}] + c_a\Pr[\text{abstain}]. \tag{28}$$

Correct decisions and true ties incur zero error cost; every abstention costs c_a. The forced rule pools for $\hat{\Delta} < 0$ and declines otherwise. For any nonnegative threshold t, including a data-dependent one, write $\mathcal{A}_t = \{|\hat{\Delta}| \le t\}$, $G(t) = \Pr[\mathcal{A}_t]$, and, when $G(t) > 0$,

$$q_{\mathcal{D}}(t) = \Pr[\hat{\Delta} < 0, \Delta_{\text{true}} > 0 \mid \mathcal{A}_t], \qquad q_{\mathcal{O}}(t) = \Pr[\hat{\Delta} \ge 0, \Delta_{\text{true}} < 0 \mid \mathcal{A}_t].$$

Since the rules agree outside the band,

$$C(\mathcal{R}_{\text{binary}}) - C(\mathcal{R}_{\text{3-way}}(t)) = \mathbb{E}\Big[\Big(r\mathbf{1}_{\{\hat{\Delta}<0, \Delta_{\text{true}}>0\}} + \mathbf{1}_{\{\hat{\Delta}\ge 0, \Delta_{\text{true}}<0\}} - c_a\Big)\mathbf{1}_{\mathcal{A}_t}\Big], \tag{29}$$

$$= G(t)\{rq_{\mathcal{D}}(t) + q_{\mathcal{O}}(t) - c_a\} \quad \text{if } G(t) > 0. \tag{30}$$

On a true tie inside the band the integrand is $-c_a$, as required. If $G(t) = 0$, the unconditional identity gives equal costs.

In the limiting experiment define

$$\begin{aligned} g_k(\mathbf{h}) &= \Pr[|\mathsf{S}_\mathbf{h}| \le k\mathsf{Q}_\mathbf{h}], \\ p_{\mathcal{D},k}(\mathbf{h}) &= \Pr[\mathsf{H}_\mathbf{h} > 0, -k\mathsf{Q}_\mathbf{h} \le \mathsf{S}_\mathbf{h} < 0], \\ p_{\mathcal{O},k}(\mathbf{h}) &= \Pr[\mathsf{H}_\mathbf{h} < 0, 0 \le \mathsf{S}_\mathbf{h} \le k\mathsf{Q}_\mathbf{h}], \\ b_k(\mathbf{h}; r) &= \{r p_{\mathcal{D},k}(\mathbf{h}) + p_{\mathcal{O},k}(\mathbf{h})\}/g_k(\mathbf{h}). \end{aligned}$$

Their dependence on the fixed B, sample ratios, and model is suppressed.

Theorem 2 (three-way decision cost benefit, formal statement)**.** *Under Appendix A.9.1, for every fixed local shift* $\mathbf{h}$*,* $k > 0$*, and finite* $r, c_a \ge 0$*, we have* $0 < g_k(\mathbf{h}) < 1$*,* $p_{\mathcal{D},k}(\mathbf{h}) > 0$*,* $p_{\mathcal{O},k}(\mathbf{h}) > 0$*, and* $b_k(\mathbf{h}; r) > 0$*, including when* $r = 0$*. Furthermore,*

$$\Gamma_{n_1}(\mathbf{h}) := C(\mathcal{R}_{\textit{binary}}) - C(\mathcal{R}_{\textit{3-way}}(k\hat{\varepsilon})) \longrightarrow g_k(\mathbf{h})\{b_k(\mathbf{h}; r) - c_a\}. \tag{31}$$

If $c_a < b_k(\mathbf{h}; r)$*, the saving is at least half this positive limit for all sufficiently large* n_1*. If* $c_a > b_k(\mathbf{h}; r)$*, the saving is eventually negative. Equality does not determine its finite-sample sign.*

The four functions above are continuous in $\mathbf{h}$*. On any fixed nonempty compact* $\mathcal{K}$*, let* $g_{\mathcal{K},k} = \min_\mathcal{K} g_k > 0$ *and* $b_{\mathcal{K},k} = \min_\mathcal{K} b_k > 0$*. For* $c_a < b_{\mathcal{K},k}$ *there is a common finite* $n_{1,0}$ *such that*

$$\inf_{\mathbf{h}\in\mathcal{K}} \Gamma_{n_1}(\mathbf{h}) \ge \tfrac{1}{2} g_{\mathcal{K},k}(b_{\mathcal{K},k} - c_a) > 0 \qquad (n_1 \ge n_{1,0}).$$

This also bounds savings averaged over any distribution supported on $\mathcal{K}$*. Diagonal population Fisher information is not required for this theorem.*

Proof. Conditional on $\zeta_N, \mathsf{V}_\mathbf{h}, U$, the threshold is positive and finite almost surely, while $\mathsf{S}_\mathbf{h}$ is normal with mean $-\zeta_N^\top J_{\text{diag}} \mathsf{V}_\mathbf{h} + \mathsf{V}_\mathbf{h}^\top J_{\text{diag}} \mathsf{V}_\mathbf{h}/2$ and variance $\|J_{\text{diag}} \mathsf{V}_\mathbf{h}\|^2/c_{\text{val}} > 0$. The independent validation Gaussian supplies this randomness. Each conditional half-band probability is strictly positive, and the full-band probability lies strictly between zero and one. The full-support and true-sign argument in Appendix A.9.3 uses only (ζ_N, V), so both true signs have positive probability here as well. If π_-, π_+ denote these conditional half-band probabilities, the tower property gives

$$p_{\mathcal{D},k} = \mathbb{E}[\mathbf{1}_{\{\mathsf{H}>0\}}\pi_-] > 0, \qquad p_{\mathcal{O},k} = \mathbb{E}[\mathbf{1}_{\{\mathsf{H}<0\}}\pi_+] > 0.$$

Hence $0 < g_k < 1$ and $b_k > 0$, even for $r = 0$.

The boundaries $\mathsf{H} = 0$, $\mathsf{S} = 0$, and $\mathsf{S} = \pm k\mathsf{Q}$ have zero probability by the same true-effect and conditional-normal arguments, without requiring $J_{\text{diag}} = I_d$. Joint convergence therefore gives convergence of all probabilities in the exact identity equation 29 with $t = k\hat{\varepsilon}$, proving equation 31. Costs are bounded by fixed constants, so no convergence of unbounded risk expectations is needed.

Coupling through the same Gaussians and U, dominated convergence with indicator envelope one gives continuity in $\mathbf{h}$. Positivity and compactness give positive minima for g_k and b_k. The sequential limit from Theorem 3, followed by the compact-subsequence argument used for calibration, gives uniform convergence of the decision probabilities and cost difference. Its positive limiting lower bound yields the stated eventual uniform saving. The negative-limit conclusion follows in the same way. □

The theorem analyzes the random threshold $k\hat{\varepsilon}$ itself, at fixed B and k. It does not specify a numerical minimum sample size. The break-even cost depends on population quantities and the local shift, so its formula is not automatically available from the two messages. The result prices sign errors and abstention in fixed cost units, while the underlying risk effects shrink on the n_1^{-1} scale. It does not identify $\hat{\varepsilon}$ with the finite-sample error bound in Theorem 1.

A.10 Bootstrap Consistency for a Fixed Donor

The preceding local experiment keeps both helpful and harmful retraining outcomes present asymptotically. For a fixed donor, the three-way rule with its bootstrap threshold also has a pointwise consistency guarantee under A1–A3, without correct logistic label specification or A4.

Proposition 2 (fixed-donor consistency)**.** *Let P_N, P_M be fixed distributions satisfying A1–A3, with $\mu_M(\theta_N^*) \neq 0$. Use the empirical minimizers and score of Section 3, mutually independent fitting, validation, and donor samples, and the fixed-B bootstrap and positive finite sample ratios of Appendix A.9.1. For each fixed $k > 0$,*

$$\Pr[\hat{\Delta} > k\hat{\varepsilon},\ \Delta_{\textit{true}} > 0] \longrightarrow 1.$$

This assertion is pointwise in the fixed pair of distributions; it gives no uniform sample-size threshold over a donor family.

Proof. Define $R_M(\theta) = \mathbb{E}_M \ell(X, Y; \theta)$ and write $F = F_N^*$. Positive definiteness of F implies that $\mathbb{E}_N[XX^\top]$ is nondegenerate. Logistic weights are positive at every finite parameter, so R is strictly convex everywhere and its finite optimum is unique. The continuous excess risk has a positive minimum δ_0 on the unit sphere about θ_N^*. Convexity gives $R(\theta_N^* + t\mathbf{v}) - R(\theta_N^*) \geq t\delta_0$ for every unit $\mathbf{v}$ and $t \geq 1$, proving coercivity. The mixture $(1 - \alpha_\infty)R + \alpha_\infty R_M$ is continuous, coercive, and strictly convex, since $R_M \geq 0$ and $\alpha_\infty < 1$. It has a unique finite minimizer θ_P with positive definite Hessian. Its gradient at θ_N^* is $\alpha_\infty \mu_M(\theta_N^*) \neq 0$, so $R(\theta_P) - R(\theta_N^*) > 0$.

At θ_N^*, the empirical recipient gradient is $O_p(n_1^{-1/2})$ by bounded variance and mean zero, and the Hessian converges to F. The ball argument in Theorem 3 gives existence and $\hat{\theta}_N - \theta_N^* = O_p(n_1^{-1/2})$. At θ_P, the pooled empirical gradient has mean $O(n_1^{-1})$ from integer rounding of the mixture ratio, and centered fluctuations $O_p(n_1^{-1/2})$ by independent bounded observations in the two strata. Its Hessian converges to the positive definite mixture Hessian there. The same argument gives $\hat{\theta}_{N\cup M} - \theta_P = O_p(n_1^{-1/2})$. Recipient risk is globally ρ-Lipschitz, so $\Delta_{\text{true}} \to_p R(\theta_P) - R(\theta_N^*) > 0$.

Bounded laws of large numbers at θ_N^* and the gradient/Hessian Lipschitz bounds give $v \to_p 0$, $u \to_p \mu_M(\theta_N^*)$, and $\hat{D} \to_p D^*$. Therefore

$$\hat{\Delta} \to_p q_0 := \tfrac{1}{2}\alpha_\infty^2 \mu_M(\theta_N^*)^\top D^* \mu_M(\theta_N^*) > 0.$$

For each fixed bootstrap replicate, its conditional fitting score at $\hat{\theta}_N$ has mean zero and covariance $O(n_1^{-1})$. Conditional Hessian variance bounds and the same ball argument give a valid refit with displacement $O_p(n_1^{-1/2})$. Its validation gradient tends to zero and inverse diagonal tends to D^*. Holding the donor message fixed thus makes every bootstrap score converge to the same q_0. At fixed B, all replicates succeed with probability tending to one and their standard deviation tends to zero. Hence $\Pr[\hat{\Delta} > k\hat{\varepsilon}, \Delta_{\text{true}} > 0] \to 1$. □

A.11 Which Results the Lemmas Use, and Which Do Not Apply

Several lemmas above use their cited sources in a non-obvious way, or deliberately avoid a standard tool whose hypotheses fail in this setting. We record which is which, so that each proof can be checked against its sources.

Lemma 1 does not quote a single named theorem of Bach (2010). His Theorems 1 and 2 bound excess population risk rather than the parameter norm, and converting one to the other by strong convexity alone would give a slower $O(n^{-1/4})$ rate. The route above instead combines Bach's Proposition 2 (Eq. 7) with the rectangular matrix Bernstein inequality of Tropp (2012).

Lemma 2 cannot apply matrix Bernstein (Tropp, 2012) directly, because $\hat{F}_N$ is evaluated at the fitted $\hat{\theta}_N$ and its summands are therefore dependent. The entrywise decomposition in the proof isolates a term with genuinely i.i.d. summands at the fixed point θ_N^*, to which Bernstein does apply, plus a Lipschitz term controlled by Lemma 1.

Lemma 5 deliberately avoids the Hanson-Wright inequality (Rudelson & Vershynin, 2013), which would give a sharper logarithmic factor but requires a random vector with independent components. Here $g_M - \bar{g}_M$ is a sum of i.i.d. vectors whose d coordinates are correlated within each example, so that hypothesis fails. The elementary operator-norm bound used instead is weaker in $\log(1/\delta)$ and fully rigorous, and the difference is immaterial at theorem level, since this term is $O(1/n_1)$ after scaling and is dominated by Lemma 6.

Lemma 9 uses the same two ingredients as Lemma 1, applied to a stratified rather than identically distributed independent sample, plus the self-concordant Hessian-comparison inequality (Bach, 2010) to carry A3's curvature floor from θ_N^* to $\theta_{N\cup M}^*$.

Lemma 4 is elementary: Jensen plus the bounded-difference inequality. We do *not* route it through sub-Gaussian closure under independent sums (Vershynin, 2018, Prop. 2.6.1), which is a statement about *scalar* sub-Gaussian random variables and does not by itself bound the Euclidean norm of a sum of random *vectors* without paying a dimension factor. Embedding the summands as $d \times 1$ matrices and applying Tropp (2012) would also work, at the cost of a $\log(d+1)$ inside the square root. **Lemmas 3, 6, 8, 10 and Corollary 2** are elementary and use no external result beyond the Hessian comparison of Bach (2010, Prop. 1, Eq. 6) in Lemma 8.

A.12 The Comparison Factor κ

For a donor with $\Delta_{\text{pop}} > 0$, the ratio $\kappa = \Delta_{\text{true}}^{\text{pop}}/\Delta_{\text{pop}}$ measures how population harm compares with the population score. A4 assumes a positive lower bound on this ratio uniformly over its donor family. We examine the approximations that determine the ratio and estimate it on synthetic populations.

Sources of Approximation Error. Three approximations separate the score from the true pooling effect (Appendix A.2). First, the parameter displacement is approximated by extrapolating the derivative of the fitting path at zero donor weight to the specified weight α. Second, the recipient-risk expansion is truncated at second order. Third, the score replaces the full recipient Hessian with its diagonal. The factor κ captures their combined effect. A smaller lower bound κ_0 increases the sufficient margin through $\varepsilon_{\text{res}}/\kappa_0$; at fixed κ_0, it changes the constant rather than the $O(n_1^{-1/2})$ rate.

Population Measurements. On the Toy setup, we estimate population harm and the population score using two million Monte Carlo draws, integrating out labels analytically and using pathwise differencing. The reported spread across seeds is below 10^{-4}. Table 2 compares κ with κ_{full}, defined by replacing the diagonal inverse in the score with the full inverse Fisher.

Table 2: Population comparison factors estimated on the Toy data-generating process at two mixture weights. The full-inverse variant κ_{full} replaces D^* by $(F_N^*)^{-1}$ in the population score. Values below one indicate that the population score overestimates population harm. This sweep uses the deployed Toy setup, which lies outside A2 and the localization condition; Table 4 reports measurements within the specified donor family.

	$m_{\text{ratio}} = 4$ ($\alpha = 0.8$)		$m_{\text{ratio}} = 1$ ($\alpha = 0.5$)	
s	κ	κ_{full}	κ	κ_{full}
0.25	0.037	0.037	0.077	0.077
0.5	0.038	0.038	0.078	0.078
1	0.042	0.041	0.082	0.082
2	0.053	0.052	0.096	0.095
3	0.068	0.067	0.110	0.108
4	0.084	0.082	0.122	0.119
5	0.098	0.095	0.131	0.126

Across the evaluated configurations, κ ranges from 0.037 to 0.131, while κ_{full} differs by less than 4%. Thus diagonalization has a small effect on the measured ratio in this sweep. These finite measurements do not establish A4's uniform bound. Moreover, the deployed Toy setup violates the bounded-feature and localization conditions defining the donor family. Appendix A.15 reports a separate sweep satisfying those conditions, with measured ratios from 0.628 to 1.000; that finite sweep also does not prove uniformity.

Effect of Curvature Mismatch. Write $F = F_N^*$, $F_M = \nabla^2 R_M(\theta_N^*)$, $\mu = \mu_M(\theta_N^*)$, and $H_G = (1-\alpha)F + \alpha F_M$. A quadratic approximation to the mixture objective gives displacement $-\alpha H_G^{-1}\mu$. Evaluating its quadratic recipient-risk change and dividing by the population score gives the geometry factor

$$\kappa_{\text{lim}} := \frac{\mu^\top H_G^{-1} F H_G^{-1} \mu}{\mu^\top D^* \mu}.$$

For $\mu \neq 0$, define the multiplicative correction $\tau_{\mathrm{nl}} := \kappa/\kappa_{\mathrm{lim}}$. This correction captures deviations from the quadratic approximation to both the mixture objective and recipient risk. In the Toy diagnostic sweep, τ_{nl} ranges approximately from 0.824 to 0.999, so the small measured κ primarily reflects the geometry factor.

If μ is proportional to a coordinate vector e_j that is a common eigenvector of F and F_M, then

$$\kappa_{\mathrm{lim}} = ((1-\alpha) + \alpha c)^{-2}, \qquad c = F_{M,jj}/F_{jj}.$$

This scalar expression need not hold for general non-diagonal matrices. In the Toy setup at $s = 0.25$, approximately 99.9% of $\|\mu\|^2$ lies in the shortcut coordinate, whose variance ratio is 6.25 by construction. With $\alpha = 0.8$, the scalar expression gives approximately 0.037, agreeing with the measured geometry factor to within 1.4%. The donor's greater curvature in this direction therefore explains most of the population score's overestimation at this operating point.

For comparison across datasets, we also evaluate the geometry expression using empirical gradients and Hessians at the fitted baseline, including the penalty in the Hessians where used. Across the five non-Toy settings in Section 5, the median values range from 0.69 to 1.98, compared with 0.093 on Toy. These empirical diagnostics suggest less overestimation from the measured geometry in those settings. They do not estimate the full population ratio κ, since the nonlinear correction and population quantities are unknown.

Bounds on the Geometry Factor. The effect of diagonalization alone can be bounded independently of the donor direction. Let $S = \mathrm{diag}(\sqrt{F_{jj}})$ and $C = S^{-1}FS^{-1}$. Since C has unit diagonal, $\lambda_{\max}(C) \leq \mathrm{tr}(C) = d$, and

$$\inf_{\mu \neq 0} \frac{\mu^\top F^{-1} \mu}{\mu^\top D^* \mu} = \lambda_{\min}(SF^{-1}S) = \frac{1}{\lambda_{\max}(C)} \geq \frac{1}{d}.$$

This bounds the ratio between the full-inverse and diagonal quadratic forms. It does not by itself bound the full comparison factor κ.

If $F_M \preceq c_{\max} F$, then $H_G \preceq \beta F$, where $\beta = (1-\alpha) + \alpha c_{\max}$. Consequently,

$$\inf_{\mu \neq 0} \kappa_{\mathrm{lim}} \geq \frac{\min_j F_{jj}\, \lambda_{\min}(F)}{\beta^2 \lambda_{\max}(F)^2} > 0.$$

Indeed, the numerator is at least $\lambda_{\min}(F)\|H_G^{-1}\mu\|^2$, which is at least $\lambda_{\min}(F)\|\mu\|^2/(\beta\lambda_{\max}(F))^2$, while the denominator is at most $\|\mu\|^2/\min_j F_{jj}$.

At fixed feature bound and recipient conditioning, donor curvature is already bounded: A2 gives $F_M \preceq (\rho^2/4)I$, and A3 therefore implies $F_M \preceq [\rho^2/(4\lambda_{\min})]F$. Specifying a smaller $c_{\max}$ can sharpen the geometry bound. The bounds above do not control τ_{nl} uniformly over the donor family.

A.13 The First-Order Term at a Fitted Baseline

The first-order term measures local changes in recipient risk around the fitted baseline $\hat{\theta}_N$, which generally differs from the population minimizer θ_N^*. Consider a fitting sample for which the baseline is finite and all Fisher diagonal entries are positive. Conditional on N_{fit}, both $\hat{\theta}_N$ and $\hat{D} = \mathrm{diag}(1/\hat{F}_{N,jj})$ are fixed. By equation 4, the conditional mean of $\hat{w}_N$ is $\hat{D}\nabla R(\hat{\theta}_N)$, while that of g_M/m is $\mu_M(\hat{\theta}_N)$. These two quantities are conditionally independent because the validation and donor samples are independent of each other and of the fitting sample. Therefore the first-order term in equation 5 satisfies

$$\mathbb{E}\Big[-\frac{\alpha}{m}\langle \hat{w}_N, g_M\rangle \,\Big|\, N_{\mathrm{fit}}\Big] = -\alpha\, \nabla R(\hat{\theta}_N)^\top \hat{D}\, \mu_M(\hat{\theta}_N).$$

This is the directional derivative of recipient risk along $-\alpha\hat{D}\mu_M(\hat{\theta}_N)$, the conditional mean of the approximate parameter displacement used by the score.

A negative conditional mean indicates a direction of decreasing recipient risk. For example, when $P_M = P_N$, the donor's population gradient equals $\nabla R(\hat{\theta}_N)$, giving

$$\mathbb{E}\Big[-\frac{\alpha}{m}\langle \hat{w}_N, g_M\rangle \,\Big|\, N_{\mathrm{fit}}\Big] = -\alpha\, \nabla R(\hat{\theta}_N)^\top \hat{D}\, \nabla R(\hat{\theta}_N) \leq 0,$$

with strict inequality whenever the recipient gradient is nonzero. This illustrates how the first-order term can indicate benefit from a donor when the baseline has estimation error. The sign of this conditional mean does not determine the sign of the realized retraining effect: sampling variability, curvature, and error in the approximate displacement also contribute.

At the population minimizer, $\nabla R(\theta_N^*) = 0$, so the population first-order term vanishes for every donor. This explains why it can provide directional information at a fitted baseline while failing to capture population harm from moving away from the recipient optimum. The first-order baseline in EQ1 and the curvature ablation in EQ4 of Section 5 evaluate its empirical contribution. The conditional interpretation above establishes neither a donor-ranking guarantee nor a guarantee that a negative first-order score identifies beneficial pooling.

A.14 Interpreting the Three-Way Decision Rule

The three-way rule acts on the sum of the first-order and curvature terms in equation 5. The first-order term can indicate either benefit or harm at a fitted baseline (Appendix A.13), while curvature contributes a nonnegative amount whenever the Fisher denominators are positive.

Remark 1 (effect of curvature on pooling decisions). If $\min_j \hat{F}_{N,jj} > 0$, the curvature term $\frac{1}{2}(\alpha/m)^2 \sum_j g_{M,j}^2 / \hat{F}_{N,jj}$ is nonnegative. For any threshold $t \geq 0$, a pool decision therefore requires the first-order term $-(\alpha/m)\langle \hat{w}_N, g_M \rangle$ to be negative with magnitude exceeding both the curvature contribution and t combined.

Holding the threshold fixed, adding curvature can prevent a pool decision or produce a decline, but cannot turn an abstention or decline into a pool decision. This algebraic property does not establish that accepted donors are helpful. A dangerous acceptance occurs when the rule pools despite $\Delta_{\text{true}} > 0$, and nonnegative curvature alone does not exclude that event. Whether harmful acceptances are rarer than declines of helpful donors is an empirical question, evaluated in EQ3 of Section 5. The comparison at a fixed threshold also does not describe an ablation that recomputes the bootstrap threshold after removing curvature.

Population nonnegativity does not determine the sign of the finite-sample pooling effect. Although pooling cannot improve on the recipient's population risk minimum, it can improve on a fitted baseline by reducing estimation error. Theorem 1 establishes a sufficient margin under which both the computed score and the finite-sample pooling effect are positive with high probability.

For fits under the conditions of Theorem 1, Corollary 3 provides a separate statistical guarantee. On its joint error-bound event, choosing $t \geq \varepsilon + \varepsilon_{\text{res}}/\kappa_0$ makes $\hat{\Delta} > t$ sufficient for $\Delta_{\text{true}} > 0$. At this threshold the pool branch cannot occur on that event, so the result does not certify beneficial pooling. Abstention gives $0 \leq \Delta_{\text{pop}} \leq t + \varepsilon$; this bounds the population score, without establishing that the true pooling effect is negligible or that pooling is safe.

The bootstrap threshold $k\hat{\varepsilon}$ has separate guarantees. Proposition 2 establishes pointwise consistency for fixed donors with nonzero population gradient under A1–A3. Under the local model of Appendix A.9.1, Theorem 3 retains the threshold's dependence on the score and retraining effect, yielding the cost benefit in Theorem 2. With diagonal population Fisher, Corollary 1 additionally gives asymptotic directional-error bounds and interval coverage. These bounds do not condition on commitment, and the interval targets the population-risk effect rather than its finite evaluation-fold estimate. They do not identify the bootstrap standard deviation with the deterministic sign-recovery bound. An abstention leaves the effect unresolved; consistency for fixed donors does not imply a finite-sample guarantee that reassessment will commit.

Finally, detecting a positive pooling effect does not establish that its magnitude is practically important. The theoretical margin describes the resolution sufficient for detecting harm under the stated assumptions. It does not specify how much degradation warrants declining a collaboration. Likewise, the decision-cost model prices incorrect decisions and abstention, rather than the magnitude of the underlying risk change. A decision based on a minimum practically important degradation would require a separately specified tolerance and an assessment of the score's accuracy for that purpose.

A.15 Synthetic Evaluation with Bounded Features and Small Donor Shifts

The Toy setup in Section 5 uses Gaussian features, which do not satisfy A2's almost-sure bound, and its ablation operates beyond the localization condition of Lemma 8. We therefore evaluate the score on a separate synthetic setup (called *Toy-Op*) with bounded features and a smaller donor shift. We examine prediction accuracy, the effect of retaining curvature, the decay of score estimation error, and the population comparison factor κ.

Experimental Setup. Both centers draw X uniformly from the unit ball in $\mathbb{R}^4$, so $\rho = 1$ holds exactly. Their logistic parameters are $\theta_N^* = (0.2, -0.1, 0.1, 0)$ and $\theta_M^* = (0.2, -0.1, 0.1, 0.5)$. We set $m = n_1/4$ and $n_2 = n_1/2$, giving $\alpha = 0.2$, and fit logistic models without an intercept. Population estimates from three independent Monte Carlo runs of 3×10^6 draws give $\lambda_{\min}(F_N^*) \approx 0.0414$, $s_{\text{loc}} \approx 0.125 < \frac{1}{6}$, and $\Delta_{\text{pop}} \approx 2.05 \times 10^{-4}$. The smaller donor weight limits the displacement while keeping the population score positive. Here the donor is smaller than the recipient, unlike the settings in the main experiments.

Prediction Accuracy and Curvature Ablation. We evaluate 600 trials per sample size across two independent seed streams. The true pooling effect is estimated on a fresh recipient sample of 100,000 observations per trial. Prediction accuracy measures agreement between the score's sign and the estimated true effect's sign, without an abstention threshold. Table 3 reports the full score's accuracy at the four largest sample sizes.

Table 3: Pooling effect prediction accuracy (pep-acc.) in the bounded-feature Toy-Op setup, over 600 trials per sample size with 95% bootstrap confidence intervals.

n_1	m	pep-acc. (raw score)	pep-acc. (committed)
9600	2400	0.725 [0.690, 0.760]	0.905 [0.863, 0.943]
19200	4800	0.757 [0.722, 0.788]	0.901 [0.863, 0.935]
38400	9600	0.878 [0.852, 0.905]	0.963 [0.943, 0.980]
76800	19200	0.943 [0.923, 0.962]	0.991 [0.981, 0.998]

Figure 8 compares the full score $\hat{\Delta}$ with its first-order term on the same trials, using the same $\hat{w}_N$ and g_M. At $n_1 = 1{,}200$ and 2,400, their accuracies are similar. At larger sizes, retaining curvature improves accuracy: at $n_1 = 76{,}800$, the full score reaches 0.943 [0.923, 0.962], compared with 0.458 [0.418, 0.498] for the first-order term alone. Proposition 1 explains why the population first-order term cannot capture harm from moving away from the recipient optimum. It does not determine the first-order score's finite-sample sign accuracy; the decline observed here is an empirical result.

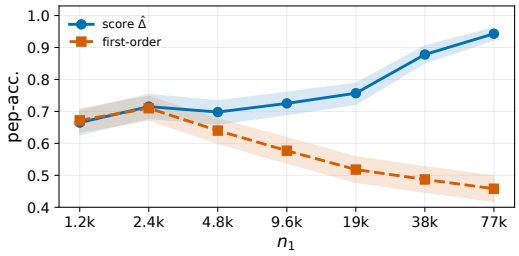

Figure 8: Effect of curvature on pooling effect prediction accuracy (pep-acc.) in the bounded-feature setup ($\rho = 1$, $\alpha = 0.2$, $s_{\text{loc}} \approx 0.125$). The full score and its first-order term are evaluated on the same 600 trials per sample size; bands show 95% bootstrap confidence intervals. At larger recipient sample sizes, retaining curvature substantially improves accuracy.

Decay of Score Estimation Error. Using a Monte Carlo population reference $\Delta_{\text{pop}} \approx 2.048 \times 10^{-4}$, we measure the median $|\hat{\Delta} - \Delta_{\text{pop}}|$ over 300 trials per size. Across $n_1 = 6{,}400$ to 102,400, its log-log slope is -0.526 $[-0.599, -0.479]$ ($R^2 = 0.991$), consistent with $n_1^{-1/2}$ scaling. Including the smaller sizes down to $n_1 = 400$ gives a steeper slope of -0.629 $[-0.656, -0.602]$, so the fitted rate depends on the sample-size range. These measurements concern median error, whereas Theorem 1 gives a high-probability upper bound. The theorem's unspecified constants also prevent identifying a numerical sample-size threshold from this experiment.

Population Comparison Factors. We vary curvature mismatch while preserving bounded features by scaling the recipient's fourth coordinate by $\beta \in \{1, 0.6, 0.4\}$ and leaving the donor's features unchanged. The recipient retains zero coefficient on this coordinate, and the donor coefficient is $s \in \{0.25, 0.5, 1\}$. Together with $\alpha \in \{0.05, 0.10, 0.20\}$, this gives 27 configurations. All satisfy A2 with $\rho = 1$; 15 also have estimated $s_{\text{loc}} \leq \frac{1}{6}$. The worst-direction curvature ratio c is estimated from F_M and F_N^* at θ_N^*, as in Appendix A.12, and ranges approximately from 1 to 6.25.

We estimate the population risks and scores using two million covariate draws per center and three independent seeds, integrating out labels analytically and evaluating recipient-risk differences on shared draws. Across the 15 configurations satisfying the localization condition, κ ranges approximately from 0.628 to 1.001. Table 4 reports the $s = 0.5$ slice, including configurations outside the localization condition for comparison.

Table 4: Population comparison factors in the bounded-feature sweep. Entries give κ at $s = 0.5$, with estimated s_{loc} in parentheses. Underlined entries violate $s_{\text{loc}} \leq \frac{1}{6}$. The recipient's fourth coordinate is scaled by β; c is the estimated worst-direction curvature ratio. All configurations satisfy $\rho = 1$.

		α		
β	c	0.05	0.10	0.20
1.0	1.00	1.000 (0.026)	1.000 (0.055)	1.000 (0.125)
0.6	2.78	0.844 (0.073)	0.721 (0.153)	<u>0.545</u> (0.345)
0.4	6.25	0.628 (0.163)	<u>0.431</u> (0.345)	<u>0.239</u> (0.775)

On the configurations satisfying the localization condition, the measured ratios are close to the scalar geometry expression $((1-\alpha)+\alpha c)^{-2}$ discussed in Appendix A.12. Replacing the diagonal inverse with the full inverse Fisher has little effect on these measurements. The larger ratios relative to the main Toy sweep also reflect the smaller donor weights: at a fixed curvature ratio $c > 1$, reducing α increases this geometry expression. The comparison therefore does not isolate the effect of bounded features.

These experiments provide evidence about score behavior with bounded covariates and small donor shifts. They do not establish A4's uniform lower bound over an infinite donor family or verify the theorem's sufficient sample-size and margin conditions at each evaluated size.

A.16 Sign Recovery Under Model Misspecification

Assumption A1 requires that R attain its minimum at a finite θ_N^*, without specifying either center's conditional label law. Under A2, differentiation under expectation and first-order optimality give

$$\mathbb{E}_{P_N}[\nabla_\theta \ell(X, Y; \theta_N^*)] = \nabla R(\theta_N^*) = 0.$$

This centers the recipient score in Lemma 1 and makes the population first-order term vanish in Proposition 1. Neither property requires logistic conditional label probabilities. Bounded gradients and the curvature properties of logistic loss used in the remaining sign-recovery argument also hold under misspecification.

A Sufficient Condition for Existence. Suppose $\|X\|_2 \leq \rho$ almost surely under P_N, $\mathbb{E}_{P_N}[XX^\top] \succ 0$, and the conditional label probability $\eta(x) := \Pr_{P_N}[Y = 1 \mid X = x]$ satisfies

$$\eta_0 \leq \eta(x) \leq 1 - \eta_0 \quad \text{almost surely,} \qquad 0 < \eta_0 \leq \tfrac{1}{2}.$$

For every scalar t, $\log(1 + e^t) - \eta(x)t \geq \eta_0|t|$. Consequently,

$$R(\theta) \geq \eta_0 \mathbb{E}_{P_N}|\theta^\top X| \geq \frac{\eta_0 \lambda_{\min}(\mathbb{E}_{P_N}[XX^\top])}{\rho}\|\theta\|_2.$$

For $\theta \neq 0$, the second inequality follows from $(\theta^\top X)^2 \leq \rho\|\theta\|_2|\theta^\top X|$; at $\theta = 0$ it is immediate. Thus $R(\theta)$ tends to infinity as $\|\theta\|_2$ grows. Continuity gives a finite minimizer, and the Hessian

$$\nabla^2 R(\theta) = \mathbb{E}_{P_N}[\sigma'(\theta^\top X)XX^\top] \succ 0$$

makes it unique. These are sufficient conditions, not necessary ones. They establish existence without first invoking A3 at an as-yet unspecified minimizer. Under bounded features, a correctly specified logistic model with finite parameter satisfies the label-probability bound, but many nonlogistic conditional probabilities satisfy it as well.

Existence of the Empirical Fits. A finite population optimum does not guarantee a finite empirical minimizer on every sample. Lemmas 1 and 9 establish empirical existence with high probability by combining score concentration, empirical curvature, and Bach (2010)'s Newton-decrement condition. Their arguments require no label-probability lower bound beyond A1–A3 and the stated sample-size and localization conditions. The exceptional-sample conventions in Appendix A.5 define the statistics outside these events. Thus the bound involving η_0 above is one sufficient way to establish A1, rather than an additional hypothesis of the sign-recovery theorem.

Interpretation and Scope. Under misspecification, θ_N^* is the minimizer of population log loss within the logistic model class. The definitions of Δ_{true} and $\Delta_{\text{true}}^{\text{pop}}$ remain unchanged. The matrix $F_N^* = \nabla^2 R(\theta_N^*)$ remains the population loss Hessian used by the analysis; it need not equal the covariance of the per-example gradient. The sign-recovery argument uses bounded-gradient concentration and Hessian conditioning, without requiring that covariance identity.

Proposition 1, Corollary 2, Theorem 1, and Corollary 3 therefore apply under the stated A1 even when the logistic model is misspecified, with each result's other conditions unchanged. In particular, the finite-sample harm guarantee still requires A2–A4, the sample-size and localization conditions, and unpenalized fits. Proposition 2 also permits misspecified labels in its fixed-donor consistency result. The local bootstrap analysis, including calibration and Theorem 2, retains correct specification for the score-covariance identity equation 20 and donor expansion equation 21.

B Empirical Appendix

B.1 Datasets and Constructions

The donor distribution shift is caused either by natural differences between centers or by a controlled spurious correlation (a shortcut). ISIC Setup A takes the donor as a genuinely different center and imposes nothing on it. The other five setups construct the shift, so its strength is a scalar we set. On the three cell-structured setups, CheXpert, Waterbirds and ISIC Setup B, that scalar is a probability s: the donor is drawn so that a fraction s of its examples fall in the cell where the shortcut variable agrees with the label, with $s = 0.5$ meaning no association and $s = 1$ a deterministic one. The recipient is always drawn at 0.5. Toy and MNIST-Bar are constructed too, but their knobs are not probabilities and are not comparable to s: Toy uses a log-odds weight and MNIST-Bar a per-class intensity separation.

How the donor is drawn from that scalar differs by experiment. Table 5 identifies the setups and Table 6 gives every operating point. Three patterns recur. EQ1 makes each cell-structured donor clean with probability one half and otherwise draws its strength from a *range* (Appendix B.2), which scores each method against a heterogeneous donor population rather than the single strength most favorable to the curvature term. EQ2 *pins* the strength at each point of a grid, which is what lets accuracy and abstention be read against strength. EQ3's fail-safe sweep and cost model (Figure 7a and Tables 11 and 12) reuse the EQ1 trials. Numbers from different rows of Table 6 are therefore not comparable cell by cell, even for the same setup.

Table 5: The six main-text setups. "Natural" means recipient and donor are two real centers with nothing imposed; "controlled" means the donor's association between shortcut and label is set by us. The recipient is shortcut-balanced in every controlled setup.

Setup	Shift source	Shortcut variable	Representation
Toy	controlled	one synthetic coordinate Z	raw features
MNIST-Bar	controlled	bar intensity in the top rows	raw pixels
CheXpert	controlled	view position, AP versus PA	encoder features
Waterbirds	controlled	image background, land vs. water	encoder features
ISIC Setup B	controlled	center of origin	encoder features
ISIC Setup A	natural	N/A	encoder features

Toy. Four-dimensional features $[X; Z]$; Z is a one-dimensional spurious feature and the donor's label model puts weight $\theta_{Z,M}$ on it, while the recipient has $\theta_{Z,N} = 0$. Features are used raw, with no encoder. EQ1 draws $\theta_{Z,M} = s \sim \mathcal{U}(-5, 5)$ per trial, so the donor pool spans compatible and contaminated sources and both outcomes occur; EQ4 instead fixes $s = 4$ and sweeps n_1, since its question is how accuracy behaves with sample size at a fixed effect.

MNIST-Bar. The top three rows of each image (84 pixels) are overwritten with a constant intensity z (i.e., a manually drawn bar on the top of each image). Clean folds draw $z \sim \mathcal{N}(0.5, 0.15^2)$ independently of the label; contaminated donors draw $z \sim \mathcal{N}(0.5 + a(2y - 1), 0.10^2)$, so a is the per-class separation of the bar intensity. EQ1 draws $a \sim \mathcal{U}(0, 0.08)$ per trial, and the EQ2 panel pins a at each of ten points on $[0, 0.08]$. Figure 9 shows examples from both centers.

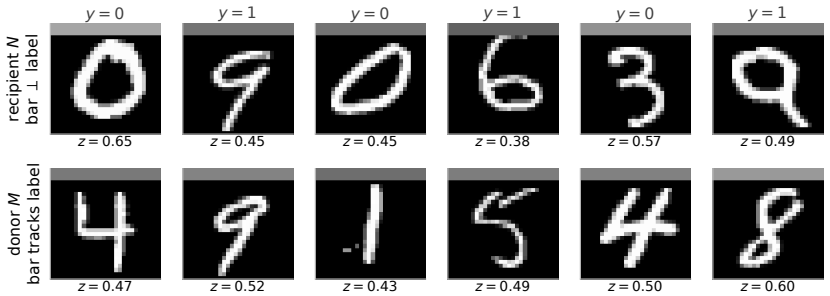

Figure 9: Example images from MNIST-Bar, with the realized bar value z printed beneath each panel. In the recipient N the bar is drawn independently of the label; in the donor M it is shifted by $\pm a$ with $a = 0.08$, so the donor's bars are lighter on y=1 on average, 0.54 against 0.47 in this draw. Each center draws its own six digits and the two rows share no images, matching the independent folds the setup samples.

CheXpert. The shortcut is radiographic view position, AP versus PA, a real acquisition attribute whose association with the pleural-effusion label is $\phi \approx 0.31$ in the patients carrying explicit labels for both. That natural association is what makes view a plausible shortcut here. Nothing is injected into the images, and the donor is instead drawn from real images so as to raise its own view-label alignment to s; the recipient is drawn at 0.5. The setup is therefore controlled in the same sense as ISIC Setup B, and we do not present it as evidence about a naturally arising donor.

Waterbirds. The shortcut is the image background, land versus water, and s is the fraction of donor examples whose background agrees with the bird label. The label marginal is held at one half in every fold, so a discrepancy measure sees no prior shift.

ISIC Setup A. The dataset with no construction. The recipient is one of four real centers and the donor is a different real center, sampled at that center's own natural malignancy rate rather than balanced. Nothing is injected and no pairing is imposed, so whatever makes the donor help or hurt is whatever actually differs between two hospitals: case mix, imaging equipment, and a genuine prior shift, since the centers differ in how often lesions are malignant. That prior shift is why Setup A is the one setup where a discrepancy measure looks competitive, and it is the reason Setup B exists alongside it: Setup B pins the label marginal so that the same centers can be compared with the prior shift removed.

ISIC Setup B. The shortcut is center of origin. Two real centers are chosen, and the donor is drawn so that a fraction s of its examples come from the center that agrees with the malignancy label. The features are real and per-center throughout, and only the pairing between center and label is constructed, which is why we call this setup controlled rather than synthetic.

B.2 Baseline Comparison

This subsection gives the details behind Figure 3: what each compared method needs from each center (Table 7), how each score is computed and how it was adapted to our setting, what the full-Hessian oracle costs next to our score, which donors each setup draws, and the full numbers (Table 8). Within a setup, every method except the Committed row is scored on the same trials. Every score is oriented so that a larger value means the method predicts that pooling hurts, so a

Table 6: n_1 is N's fitting fold, n_2 its validation fold, n_{eval} the fresh disjoint fold on which Δ_{true} is measured, and m the donor's size; $\alpha = m/(n_1 + m)$ is the weight M's data carries in the pooled fit.

Experiment	Setups	n_1	n_2	n_{eval}	m	α
EQ1, EQ3 (Figs. 3, 6, 7a)	Toy	400	200	20000	1600	0.80
	MNIST-Bar	400	200	2000	1600	0.80
	CheXpert, Waterbirds, ISIC Setup B	100	50	300	200	0.67
	ISIC Setup A	120	60	150	300	0.71
EQ2 (Figs. 4, 5)	ISIC Setup B, Waterbirds, CheXpert	100	50	300	200	0.67
	MNIST-Bar	400	200	2000	1600	0.80
EQ4[‡] (Fig. 7b)	Toy	sweep	$\max(50, n_1/2)$	20000	$4n_1$	0.80

[‡]EQ4 sweeps n_1 over $\{25, 50, 100, 200, 400, 800, 1600\}$; α is held at 0.80 while the sample size varies.

single help-vs-hurt AUROC compares them all.

Table 7: What each compared method requires from each center. For the last five rows only two quantities cross between the centers: $\hat{\theta}_N$, which N sends so that M can compute g_M, and g_M itself. The other entries in the N column are computed and kept at N.

Method	From M	From N
MMD	raw features	raw features
Score XY	raw features + labels	raw features + labels
Margin uncertainty	uncertainty score	fitted model
Full-Hessian oracle	aggregated gradient	full Hessian H
Gradient conflict	aggregated gradient	validation gradient
First-order term (ablation)	aggregated gradient	$\hat{w}_N$
Curvature term (ablation)	aggregated gradient	$\hat{F}_N$
Ours	aggregated gradient	$\hat{w}_N$, $\hat{F}_N$

Discrepancy Measures. Covariate MMD is the unbiased squared maximum mean discrepancy (Gretton et al., 2012) between the features of N's validation fold and those of M, with a Gaussian kernel whose bandwidth is set by the median heuristic, computed on at most 150 examples from each side. Joint MMD is the same statistic with the label appended to the features. Score XY (Shen et al., 2024) fits a logistic classifier that separates M's examples from N's validation examples using both features and label, and scores the donor by the classifier's mean predicted probability, over M's examples, that an example comes from M. For all three, a larger value means the two centers are easier to tell apart. All three need raw data from both centers, so they are not admissible under the protocol and are included as references.

Margin Uncertainty. Uncertainty sampling (Schein & Ungar, 2007) picks individual points near a model's decision boundary. We adapt it to score a whole donor: M averages $1 - |2p_{\hat{\theta}_N}(x) - 1|$ over its own examples, using N's fitted model, and returns that single number. A larger value means N's model is less certain on M's data. It needs no exchange of raw data, so it is admissible. It uses M's features but never its labels.

Gradient Conflict. Following the conflict measure of Yu et al. (2020), we score the donor by $-\cos(g_M, \bar{g}_{\text{val}})$, where $\bar{g}_{\text{val}}$ is the mean gradient of N's validation loss at $\hat{\theta}_N$. We evaluate it once, at $\hat{\theta}_N$, and keep it continuous rather than thresholding it into a gate. Up to the normalization of the two vectors, it is our first-order term without the division by the Fisher diagonal.

Full-Hessian Oracle. This is our score with the coordinate-wise division by $\hat{F}_N$ replaced by multiplication by H^{-1}, where H is the Hessian of N's mean fitting-fold loss at $\hat{\theta}_N$ with the same ridge term on its diagonal as $\hat{F}_N$:

$$-\frac{\alpha}{m}\,\bar{g}_{\text{val}}^{\top} H^{-1} g_M + \frac{1}{2}\left(\frac{\alpha}{m}\right)^2 g_M^{\top} H^{-1} g_M.$$

The diagonal of H equals $\hat{F}_N$, so the two scores coincide when H is diagonal. Heo et al. (2026) use two Hessians, one for the training loss and one for the target function. Setting both to H gives this score, so it is a special case of their estimator rather than a reproduction of it. Because N forms H from its own fitting fold, the oracle sends the same two messages as our score.

The Two Terms Alone. The First-order only and Curvature only rows score the donor by each term of equation 5 on its own.

Computational Cost. All gradient-based methods share the cost of fitting $\hat{\theta}_N$, of the mean validation gradient, $O(n_2 d)$, and of M computing g_M, $O(md)$, and each message has $O(d)$ entries. They differ in what N computes next. Our score builds the Fisher diagonal in $O(n_1 d)$ and then scores a donor in $O(d)$. The full-Hessian oracle builds H in $O(n_1 d^2)$ and solves a $d \times d$ system in $O(d^3)$, for $O(n_1 d^2 + d^3)$ in total; if one factorization of H is reused across donors, each further donor costs $O(d^2)$. These are the costs of our exact implementation.

Donors in Each Setup. The sample sizes are the EQ1 rows of Table 6. The donor's shift strength is drawn afresh in every trial. On Toy, $\theta_{Z,M} = s \sim \mathcal{U}(-5, 5)$. On MNIST-Bar, $a \sim \mathcal{U}(0, 0.08)$. On CheXpert, Waterbirds and ISIC Setup B, the donor is clean ($s = 0.5$) with probability one half and otherwise has $s \sim \mathcal{U}(0.65, 0.90)$, $\mathcal{U}(0.70, 0.95)$ and $\mathcal{U}(0.65, 0.95)$ respectively. ISIC Setup A imposes nothing: each of the four recipient centers is paired with each of the five other centers as the donor, with 75 draws per pair, and 11 draws were skipped because a fold could not be filled. In the column order of Table 8 this gives 2000, 2000, 1489, 1500, 1500 and 1500 trials, and the fraction of trials in which pooling truly hurts is 0.869, 0.802, 0.780, 0.305, 0.382 and 0.261.

Table 8 reports the numerical results corresponding to Figure 3 in the main paper.

The Committed Row. The Committed row uses the same trials as the Full score row. For each trial we also compute the bootstrap standard deviation $\hat{\varepsilon}$ of the score ($B = 40$), drawing the bootstrap from its own random stream so that the trials themselves are unchanged, and keep the trials on which the three-way rule commits, $|\hat{\Delta}| > k\hat{\varepsilon}$ with $k = 1$. The row is therefore the full score's AUROC on a subset of the trials behind the row above it. The rule commits on 94.5%, 45.4%, 64.1%, 20.5%, 20.9% and 15.6% of trials in column order, so the committed values on ISIC Setup B, CheXpert and Waterbirds rest on 307, 313 and 234 trials.

Table 8: Help-vs-hurt AUROC with 95% bootstrap intervals over trials, setups and methods.

Method	Toy	MNIST-Bar	ISIC Setup A	ISIC Setup B	CheXpert	Waterbirds
MMD (covariate)	.509 [.469, .546]	.493 [.462, .524]	.533 [.500, .566]	.534 [.505, .565]	.545 [.515, .575]	.632 [.599, .665]
MMD (joint)	.634 [.601, .670]	.488 [.458, .519]	.534 [.501, .567]	.528 [.498, .563]	.547 [.518, .577]	.643 [.610, .674]
Score XY	.548 [.511, .588]	.519 [.489, .549]	.632 [.603, .664]	.548 [.518, .580]	.520 [.489, .550]	.530 [.496, .566]
Margin uncertainty	.530 [.490, .567]	.517 [.483, .549]	.595 [.560, .629]	.622 [.591, .650]	.589 [.559, .617]	.516 [.483, .546]
Gradient conflict	.722 [.683, .759]	.717 [.689, .744]	.614 [.580, .647]	.712 [.679, .741]	.725 [.698, .752]	.683 [.651, .714]
Full-Hessian	.965 [.956, .973]	.846 [.829, .863]	.735 [.703, .764]	.752 [.725, .778]	.752 [.726, .778]	.708 [.679, .738]
First-order only	.534 [.511, .558]	.562 [.535, .587]	.607 [.575, .640]	.657 [.623, .688]	.661 [.632, .691]	.655 [.623, .688]
Curvature only	.957 [.947, .967]	.906 [.893, .919]	.688 [.658, .716]	.658 [.629, .689]	.797 [.773, .822]	.610 [.582, .644]
Full score (ours)	.965 [.956, .973]	.848 [.831, .864]	.745 [.715, .773]	.757 [.729, .785]	.826 [.803, .849]	.712 [.682, .743]
Committed (ours)	.950 [.936, .961]	.940 [.923, .954]	.736 [.697, .773]	.907 [.869, .942]	.918 [.882, .950]	.825 [.772, .879]

B.3 Task Performance of the Recipient Models

We evaluate the task performance of the recipient-only classifiers whose pooling effects are assessed in EQ1. Each trial's model $\hat{\theta}_N$ uses that setup's fitting size n_1 (Table 6), features, and training procedure. We evaluate it on the same disjoint recipient fold used to obtain the measured true pooling effect. Table 9 reports the mean task AUROC across these trials.

Task AUROC measures each fitted model's discrimination between positive and negative examples for the setup's prediction task. Help-vs-hurt AUROC instead measures the pooling effect score's discrimination between harmful and helpful donor additions across trials. The two metrics evaluate different predictions.

Mean task AUROC ranges from 0.639 on CheXpert to 0.929 on Waterbirds, indicating above-chance discrimination on every evaluated task. These results characterize the recipient models used in the pooling experiments; they do not measure the accuracy of the pooling effect predictions, which is evaluated separately in EQ1.

B.4 Encoder Differences

Whether a shortcut is visible to the score is a property of the representation, not of the dataset alone, and we vary the encoder while holding everything else fixed on the CheXpert setup of EQ1. Table 10

Table 9: Mean task AUROC of the recipient-only models $\hat{\theta}_N$ on the EQ1 trials, with 95% bootstrap intervals for the mean obtained by resampling trials. Values are rounded to three decimals; identical printed interval endpoints on Toy reflect rounding.

Setup	Task	Task AUROC
Toy	synthetic label	$0.857\ [0.857, 0.857]$
MNIST-Bar	digit ≥ 5 vs < 5	$0.756\ [0.756, 0.757]$
ISIC Setup A	malignant vs benign	$0.701\ [0.698, 0.703]$
ISIC Setup B	malignant vs benign	$0.661\ [0.659, 0.663]$
CheXpert	pleural effusion	$0.639\ [0.637, 0.641]$
Waterbirds	land vs water bird	$0.929\ [0.928, 0.930]$

repeats the setup on four pretrained encoders: ResNet-50 (He et al., 2016), CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and MedCLIP (Wang et al., 2022). Figure 10 plots the two ablation terms and the full score across them.

Table 10: Encoder differences on CheXpert at the EQ1 design, 1500 trials per encoder, fresh-fold ground truth. $|\Delta_{\text{hurt}}|$ is the mean size of the measured true effect on trials where pooling hurts. *AUROC* is the help-vs-hurt AUROC. The three general-purpose encoders surface the harm; the domain-specialized MedCLIP abstracts it away, and its harm is about eight times smaller.

Encoder	$\lvert\Delta_{\text{hurt}}\rvert$	AUROC $\hat{\Delta}$	AUROC curvature	AUROC first-order
ResNet-50	0.0396	$0.826\ [0.803, 0.849]$	$0.797\ [0.773, 0.822]$	$0.661\ [0.632, 0.691]$
CLIP	0.0310	$0.835\ [0.814, 0.857]$	$0.812\ [0.789, 0.835]$	$0.648\ [0.619, 0.676]$
DINOv2	0.0301	$0.776\ [0.750, 0.801]$	$0.694\ [0.664, 0.722]$	$0.673\ [0.641, 0.703]$
MedCLIP	0.0048	$0.619\ [0.590, 0.647]$	$0.103\ [0.088, 0.118]$	$0.737\ [0.711, 0.762]$

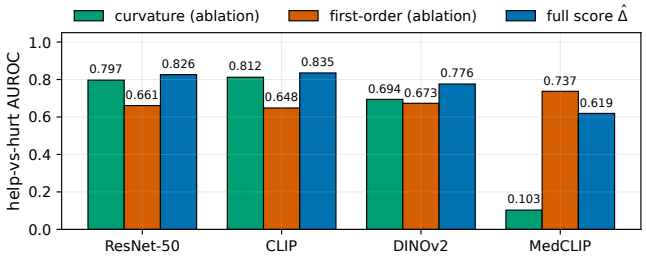

Figure 10: Curvature, first-order, and score $\hat{\Delta}$ help-vs-hurt AUROC by encoder on CheXpert.

The MedCLIP row is the informative one and we read it as a confirmation rather than a failure. Its curvature AUROC of 0.103 is far below chance, which would be alarming if there were an effect to detect, but the true harm on that encoder is $|\Delta_{\text{hurt}}| = 0.0048$, roughly an eighth of ResNet-50's 0.0396. A medical-domain encoder that represents effusion directly rather than through view position leaves almost no shortcut-driven harm, and a number computed on an effect of that size is noise around zero. The ResNet-50 row is the CheXpert run of EQ1 itself and matches Table 8. The other encoders use the same design and random seed.

B.5 Low Dangerous Error Rate on All Setups

Figure 11 gives the complete results behind Figure 6. The ISIC Setup A and CheXpert rows are the panels shown in the main text.

Setup. The sweep uses the EQ1 trials of Appendix B.2, whose sizes are in Table 6, and the same bootstrap estimate $\hat{\varepsilon}$ ($B = 40$) as the committed row of Table 8. No new trials are drawn. Changing k only changes which trials commit.

Results. The observed danger-type error rate approaches zero on all six setups as k increases. At $k = 3$, it is at most 0.05% of all trials on every setup.

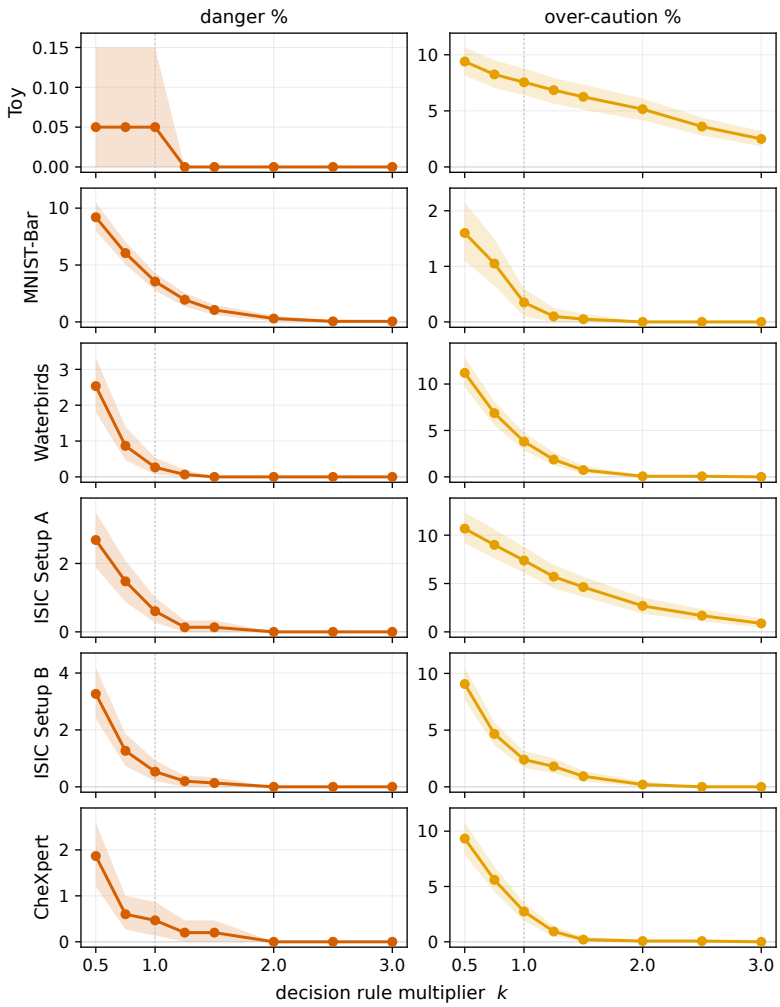

Figure 11: Danger and over-caution error rates, both as percentages of all trials, against the decision rule multiplier k, one setup per row. Shaded bands are 95% bootstrap intervals over trials. Each panel has its own y-axis range.

B.6 Cost Model

Always-decline has zero danger-type error because it never pools, so the danger rate alone cannot rank decision rules. We compare costs against three trivial rules. *Always-pool* and *always-decline* commit on every trial. For hurt rate h, always-pool has danger rate h and over-caution rate 0, while always-decline has danger rate 0 and over-caution rate $1 - h$. *Random-commit* commits independently of the trial with probability equal to our rule's overall commitment fraction, then chooses pool or decline with equal probability. This is a random policy with the same commitment rate; the comparison evaluates our selection of trials and decision direction together.

The cost of declining a helpful donor is normalized to 1, the cost of pooling a harmful one is r, and the cost of one abstention is c_a. We estimate the expected cost in equation 28 using

$$C = \frac{r \cdot \text{danger}\% + \text{over-caution}\% + c_a \cdot \text{abstain}\%}{100}.$$

Here the three rates are percentages of all trials, and C is the average cost per trial. We estimate this cost on the EQ1 trials at $k = 1$, using the sign of the measured true pooling effect to label help and hurt. The figure reports average cost per trial; the tables display error and abstention rates as percentages.

Trivial Rules at Zero Abstention Cost. Table 11 sets $c_a = 0$, as in Figure 7a, and gives for each

setup the range of r over which our rule is cheaper than all three listed trivial rules. The CheXpert row gives the range shown in the main text. This is not an optimality claim over all decision rules: always-abstain has zero cost when $c_a = 0$. We therefore examine positive abstention costs below.

For a positive danger-type error rate, the crossing with always-decline occurs at

$$r = \frac{(1-h) - \text{over-caution}}{\text{danger}},$$

where both error rates are fractions for our rule. Above this crossing, always-decline is cheaper. MNIST-Bar has the highest danger-type error rate (3.55%) and, with $h = 0.802$, one of the lowest always-decline costs (0.198). Together with its over-caution rate of 0.35%, these give a crossing near $r = 5.5$; the table rounds the upper endpoint inward to 5.4. The upper endpoints are sensitive to small counts of danger-type errors, including 7 of 1500 trials on CheXpert. Its upper endpoint has a 95% bootstrap interval of $[69, 430]$.

Table 11: Our rule against trivial baselines at $k = 1$ and $c_a = 0$, on the EQ1 trials. Rates are percentages of all trials, and h is the hurt rate. Always-pool (danger $100h$, over-caution 0) and always-decline (danger 0, over-caution $100(1-h)$) are fixed by h and are not listed. Random-commit selects trials independently at our overall commitment rate and chooses pool or decline with equal probability; its listed error rates average over this randomization. The last column is the range of r over which our rule is cheaper than all three trivial rules, rounded inward.

Setup	h	commit %	ours		random-commit		r range
			danger %	over-c. %	danger %	over-c. %	
Toy	0.869	94.5	0.05	7.55	41.0	6.2	(0.087, 111)
MNIST-Bar	0.802	45.4	3.55	0.35	18.2	4.5	(0.005, 5.4)
ISIC Setup A	0.780	64.1	0.60	7.39	25.0	7.0	(0.096, 24)
ISIC Setup B	0.305	20.5	0.53	2.40	3.1	7.1	(0.080, 125)
CheXpert	0.382	20.9	0.47	2.73	4.0	6.4	(0.073, 126)
Waterbirds	0.261	15.6	0.27	3.80	2.0	5.8	(0.147, 262)

Pricing Abstention. We next compare the three-way rule with the forced rule $\mathcal{R}_{\text{binary}}$, which pools when $\hat{\Delta} < 0$ and declines otherwise. The two rules agree on every trial where the three-way rule commits. On each abstained trial, the three-way rule pays c_a, while the forced rule pays r, 1, or 0 according to its error. Its average error cost over the abstained subset is $rq_{\mathcal{D}} + q_{\mathcal{O}}$, where $q_{\mathcal{D}}$ and $q_{\mathcal{O}}$ are the observed fractions of that subset on which following the sign would pool a harmful donor or decline a helpful one. Since every setup has a nonempty abstained subset, the empirical cost identity equation 30 gives lower cost for the three-way rule exactly when

$$c_a < rq_{\mathcal{D}} + q_{\mathcal{O}}.$$

At equality, the two rules have the same empirical cost. Table 12 reports this break-even value, which is positive on every setup. At $r = 1$, the values range from 0.33 to 0.57 in units of the cost of declining one helpful donor, and they increase with r on these trials. The reported 95% intervals use 1000 bootstrap resamples of completed trials, retaining each trial's score, bootstrap threshold, and evaluation label.

Theorem 2 establishes eventual cost savings for the bootstrap threshold $k\hat{\varepsilon}$ under its local donor model. The comparisons here evaluate the same decision rule on the experimental distributions; they do not verify the theorem's assumptions or identify its sufficient sample size.

B.7 Three-Way Decisions Across Donor Shift Strength

Tables 13–15 give the numerical results behind Figure 4, together with hurt rates and directional error rates. On ISIC Setup B, Waterbirds, and CheXpert, we fix donor shortcut-label alignment at each $s \in \{0.55, 0.65, 0.75, 0.85, 0.95\}$. The recipient remains label-balanced and shortcut-balanced, and the sample sizes remain $n_1 = 100$, $n_2 = 50$, $m = 200$, and $n_{\text{eval}} = 300$. Each setting uses 1500 trials and $k = 1$. Unlike the baseline comparison, which samples donor strengths from a range, this sweep evaluates decisions separately at each strength.

Metrics. A trial is committed when $|\hat{\Delta}| > \hat{\varepsilon}$ and abstained otherwise. Pooling effect prediction accuracy (pep-acc.) is the fraction of trials in the indicated subset for which the score's sign agrees with

Table 12: Break-even abstention cost c_a on the EQ1 trials at $k = 1$, in units of the cost of declining one helpful donor. The three-way rule is cheaper than the forced rule below this value and equally costly at it. *Abstain %* is the percentage of trials on which the three-way rule abstains. The interval at $r = 1$ is a 95% bootstrap interval over trials.

Setup	abstain %	break-even c_a		
		$r = 0.1$	$r = 1$	$r = 10$
Toy	5.5	0.56	0.57 [0.47, 0.65]	0.65
MNIST-Bar	54.6	0.10	0.37 [0.34, 0.40]	3.05
ISIC Setup A	35.9	0.22	0.38 [0.34, 0.42]	2.04
ISIC Setup B	79.5	0.25	0.35 [0.33, 0.38]	1.42
CheXpert	79.1	0.26	0.33 [0.30, 0.36]	0.99
Waterbirds	84.4	0.27	0.36 [0.33, 0.38]	1.25

the sign of the measured true pooling effect. The hurt rate is the fraction with a positive measured true pooling effect. On committed trials, danger counts pool decisions with a positive measured true pooling effect. On abstained trials, both pep-acc. and danger are counterfactual: they describe following the score's sign despite the rule's abstention. These danger rates are percentages *within each subset*, whereas EQ3 reports percentages of all trials. They cannot be added directly; converting a conditional rate to a percentage of all trials requires multiplying by that subset's fraction.

Bootstrap Procedures. Within each trial, $B = 40$ recipient resamples estimate the score's variability $\hat{\varepsilon}$, with the donor message held fixed as in Section 3. Separately, 1000 bootstrap resamples of the completed trials give the reported 95% percentile intervals. Each resampled trial retains its score, threshold, and evaluation label. The subset statistics are recomputed from these resampled trials, without refitting models or recomputing their thresholds.

Table 13: ISIC Setup B: numerical results for Figure 4, with 1500 trials per donor alignment s. Brackets give 95% bootstrap intervals over trials. The labels *cmt.* and *abst.* denote committed and abstained subsets. Hurt rates, abstention fractions, and pep-acc. are fractions; danger rates are percentages within the indicated subset. Pep-acc. and danger on abstained trials are counterfactual.

Quantity	$s = 0.55$	$s = 0.65$	$s = 0.75$	$s = 0.85$	$s = 0.95$
hurt rate (all)	.095 [.080, .109]	.177 [.158, .195]	.409 [.383, .433]	.669 [.647, .692]	.885 [.868, .901]
frac. abstain	.777 [.757, .799]	.819 [.799, .839]	.847 [.829, .865]	.810 [.790, .829]	.735 [.712, .757]
pep-acc. | cmt.	.868 [.832, .903]	.793 [.745, .842]	.704 [.645, .761]	.828 [.783, .872]	.947 [.925, .969]
danger% | cmt.	3.3 [1.5, 5.4]	6.3 [3.5, 9.4]	7.4 [4.1, 11.0]	4.6 [2.2, 7.2]	2.0 [0.8, 3.5]
pep-acc. | abst.	.683 [.656, .705]	.649 [.624, .676]	.615 [.588, .642]	.628 [.602, .654]	.669 [.642, .696]
danger% | abst.	6.3 [4.9, 7.7]	7.7 [6.3, 9.2]	15.0 [13.2, 17.1]	20.9 [18.7, 23.2]	25.7 [23.1, 28.1]

Table 14: Waterbirds: numerical results for Figure 4, using the same protocol and row definitions as Table 13. Brackets give 95% bootstrap intervals over trials.

Quantity	$s = 0.55$	$s = 0.65$	$s = 0.75$	$s = 0.85$	$s = 0.95$
hurt rate (all)	.085 [.072, .100]	.136 [.119, .155]	.283 [.262, .307]	.509 [.486, .537]	.761 [.739, .781]
frac. abstain	.833 [.814, .852]	.864 [.847, .881]	.889 [.873, .905]	.879 [.863, .896]	.831 [.811, .851]
pep-acc. | cmt.	.828 [.785, .872]	.765 [.701, .823]	.653 [.577, .725]	.707 [.636, .775]	.839 [.787, .882]
danger% | cmt.	4.0 [1.8, 6.2]	2.9 [0.9, 5.4]	3.0 [0.6, 5.7]	2.2 [0.5, 4.5]	3.1 [1.2, 5.4]
pep-acc. | abst.	.668 [.641, .691]	.612 [.585, .639]	.557 [.530, .583]	.577 [.552, .604]	.649 [.622, .676]
danger% | abst.	6.2 [4.9, 7.5]	7.6 [6.2, 9.1]	12.7 [10.9, 14.4]	16.1 [14.5, 18.1]	20.2 [18.0, 22.5]

Accuracy and Abstention. Committed pep-acc. exceeds counterfactual abstained pep-acc. at all fifteen evaluated points. The differences in point estimates range from 0.048 to 0.278. This supports the main paper's observation that the rule defers less reliable sign predictions in these experiments. The tables report intervals for each accuracy separately, rather than intervals for their difference, so this pattern does not establish a statistically significant gap at every point.

The highest observed abstention fraction occurs at the evaluated strength with the lowest committed pep-acc. on each setup: $s = 0.75$ for ISIC Setup B and Waterbirds, and $s = 0.65$ for CheXpert. Beyond these points, abstention decreases as donor alignment strengthens. The relationship is therefore not monotone over the full sweep. At fixed sample sizes and k, these changes reflect how often scores clear their trial-specific bootstrap thresholds.

Directional Errors. The conditional danger rate is lower on committed trials than on abstained

Table 15: CheXpert: numerical results for Figure 4, with donor shift controlled by view-label alignment. The protocol and row definitions match Table 13; brackets give 95% bootstrap intervals over trials.

Quantity	$s = 0.55$	$s = 0.65$	$s = 0.75$	$s = 0.85$	$s = 0.95$
hurt rate (all)	.085 [.071, .099]	.239 [.217, .262]	.576 [.549, .599]	.884 [.867, .900]	.990 [.985, .995]
frac. abstain	.815 [.795, .835]	.903 [.889, .917]	.878 [.862, .895]	.754 [.732, .775]	.579 [.555, .605]
pep-acc. | cmt.	.866 [.824, .907]	.575 [.490, .650]	.743 [.674, .807]	.932 [.904, .957]	.992 [.984, .998]
danger% | cmt.	3.2 [1.4, 5.4]	2.0 [0.0, 4.6]	0.6 [0.0, 1.8]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
pep-acc. | abst.	.652 [.627, .678]	.517 [.489, .545]	.603 [.574, .629]	.819 [.795, .839]	.944 [.927, .957]
danger% | abst.	5.0 [3.8, 6.2]	8.7 [7.3, 10.2]	10.2 [8.6, 11.8]	7.9 [6.5, 9.6]	4.8 [3.6, 6.3]

trials at every evaluated point. For example, at $s = 0.95$ the rates are 2.0% versus 25.7% on ISIC Setup B and 3.1% versus 20.2% on Waterbirds. These are empirical comparisons with the counterfactual action of following the sign on deferred trials. They do not establish a statistical guarantee against harmful acceptance, and the difference does not increase monotonically with strength on every setup.

B.8 Selecting Among Many Donors

The theorem and every experiment above concern a single candidate donor. A recipient may face several at once. Whether our method also orders donors by how much each one would help is a different question from the one this paper studies, and no theoretical result speaks to it. We ran a small experiment on that question, and report it here as a preliminary study that falls outside the scope of the paper and outside what our analysis supports.

The setup uses the *eICU* (Pollard et al., 2018) cohort of 33 hospitals holding at least 1,200 stays. Each hospital can serve as either center $\boldsymbol{N}$ (a recipient) or center $\boldsymbol{M}$ (a donor), with each recipient being scored against the remaining 32 candidate donors. The experiment runs at $n_1 = 200$, $n_2 = 100$, $m = 800$, an evaluation fold of $n_{\text{eval}} = 600$ stays, and twelve repeats per recipient and donor pair. A donor's true effect is the change in the recipient's log-loss on the fresh evaluation fold, as elsewhere in the paper. We summarize one recipient's ranking by the Spearman correlation between the score $\hat{\Delta}$ and that change, taken across its 32 candidates.

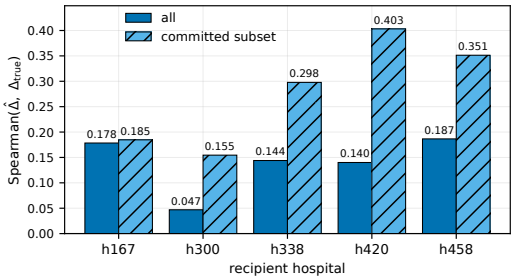

Figure 12: Ranking candidate donors on eICU, for five recipient hospitals. Each bar is the Spearman correlation between our score and the fresh-fold change in the recipient's log-loss, taken across that recipient's 32 candidate donors and averaged over twelve repeats. Solid bars use all 32 candidates. Hatched bars use only the candidates the three-way rule commits at $k = 1$.

Each of the twelve repeats is treated as one deployment. The recipient scores every candidate once, applies the three-way rule (at $k = 1$) to that repeat's own score and its own $\hat{\varepsilon}$, and ranks the donors the rule resolves; we then average the twelve correlations. Figure 12 shows five recipient hospitals. On each of the five the committed value is the larger of the two.

B.9 Differential Privacy on the Donor's Message

Exchanging aggregated gradients avoids sharing raw records but does not itself provide differential privacy. We evaluate a private donor message obtained by clipping each per-example gradient to L_2 norm C and adding Gaussian noise:

$$g_M^{\text{clip}} = \sum_{i \in M} \Pi_C\Big(\nabla_\theta \ell(x_i, y_i; \hat{\theta}_N)\Big), \qquad \tilde{g}_M = g_M^{\text{clip}} + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d),$$

where $\Pi_C(v) = v \min\{1, C/\|v\|_2\}$ for $v \neq 0$ and $\Pi_C(0) = 0$. The recipient substitutes $\tilde{g}_M$ for g_M in the score and bootstrap procedure.

Adjacency and the Donor Count. We use a fixed public donor capacity m, declared independently

of the records. The message sums over m slots, with an empty slot contributing zero. Neighboring datasets have the same capacity and differ only by replacing one record with an empty slot, or conversely. The clipped sum then has L_2 sensitivity at most C, and releasing the declared capacity does not distinguish neighbors. The recipient model is fixed with respect to donor records, so subsequent scoring, recipient resampling, and decision-making are post-processing and preserve the message's donor-side privacy guarantee.

All slots are occupied in the experiments: $m = 1600$ on Toy and $m = 300$ on ISIC Setup A. Thus the declared capacity equals the actual donor sample size, and $\alpha = m/(n_1 + m)$ is the usual pooling weight. With empty slots, this weight would instead use capacity and would no longer equal the actual donor fraction in a pooled training sample. The guarantee is specific to this neighboring relation. Releasing the exact number of occupied slots would violate ordinary add/remove privacy, while replacing an arbitrary record with another requires a sensitivity bound of $2C$.

Privacy Scope and Calibration. We fix $C = 2$ on Toy and $C = 6$ on ISIC Setup A; the observed clipping fractions are 5% and 14%, respectively. We use

$$\sigma = \frac{C\sqrt{2\ln(1.25/\delta_{\mathrm{DP}})}}{\varepsilon_{\mathrm{DP}}}, \qquad \delta_{\mathrm{DP}} = 10^{-5}.$$

Although the classical calibration is stated for $\varepsilon_{\mathrm{DP}} < 1$ (Dwork & Roth, 2014), the exact Gaussian-mechanism condition of Balle & Wang (2018) confirms every finite budget in Table 16 at sensitivity C. The chosen noise standard deviations are 1.01 to 1.58 times the minimum required at the corresponding budgets. Under arbitrary record replacement, the same noise levels give weaker budgets.[1] The accounting covers one release per donor and claims no privacy amplification from subsampling. Repeated releases would require composition accounting.

Experimental Design. Within each setup, all noise levels use the same sampled trials and clipped gradients. The $\varepsilon_{\mathrm{DP}} = \infty$ row is a clipped, noiseless baseline, so the sweep measures the effect of adding noise after clipping. For each trial, we reuse the same recipient bootstrap refits across noise levels and recompute their scores using the corresponding noisy message, held fixed across those refits. Their standard deviation supplies $\hat{\varepsilon}$, and the decision rule uses $k = 1$. The threshold therefore depends on the noisy message, but does not measure variability across fresh privacy-noise draws. The true effect remains the fresh-fold loss change from retraining on the original donor sample.

Table 16: Effect of donor-message noise on Toy and ISIC Setup A, with 95% bootstrap intervals over trials (1500 trials, $k = 1$). help-vs-hurt AUROC uses all trials; committed accuracy uses only trials on which the rule acts. Abstention is a fraction of all trials, and both error rates are percentages of all trials. Privacy budgets use fixed-capacity, record-versus-empty-slot adjacency with $\delta_{\mathrm{DP}} = 10^{-5}$.

$\varepsilon_{\mathrm{DP}}$	σ	help-vs-hurt AUROC	abstain	committed acc.	danger %	over-caution %
Toy (clip $C = 2$, 5% of gradients clipped)						
∞	0	0.985 [.978, .992]	0.33 [.30, .35]	0.790 [.764, .815]	0.00 [0.00, 0.00]	14.1 [12.4, 16.0]
8	1.2	0.985 [.978, .992]	0.33 [.30, .35]	0.789 [.763, .813]	0.00 [0.00, 0.00]	14.3 [12.5, 16.1]
4	2.4	0.985 [.978, .992]	0.32 [.30, .35]	0.787 [.761, .810]	0.07 [0.00, 0.20]	14.4 [12.7, 16.3]
2	4.8	0.985 [.977, .991]	0.32 [.30, .35]	0.787 [.761, .810]	0.00 [0.00, 0.00]	14.5 [12.8, 16.3]
1	9.7	0.985 [.977, .992]	0.32 [.30, .35]	0.784 [.756, .809]	0.00 [0.00, 0.00]	14.6 [12.9, 16.5]
0.5	19.4	0.986 [.979, .992]	0.30 [.28, .33]	0.760 [.734, .787]	0.07 [0.00, 0.20]	16.7 [14.7, 18.5]
0.25	38.8	0.983 [.974, .990]	0.28 [.26, .31]	0.721 [.694, .748]	0.07 [0.00, 0.20]	19.9 [17.9, 21.9]
0.1	96.9	0.983 [.975, .990]	0.12 [.10, .13]	0.579 [.551, .605]	0.00 [0.00, 0.00]	37.2 [34.9, 39.7]
ISIC Setup A (clip $C = 6$, 14% of gradients clipped)						
∞	0	0.741 [.716, .769]	0.38 [.35, .40]	0.875 [.855, .896]	0.40 [0.13, 0.74]	7.4 [6.0, 8.7]
8	3.6	0.741 [.715, .769]	0.37 [.35, .40]	0.872 [.851, .894]	0.47 [0.13, 0.87]	7.5 [6.2, 8.9]
4	7.3	0.742 [.717, .771]	0.38 [.35, .40]	0.868 [.845, .890]	0.47 [0.13, 0.81]	7.7 [6.4, 9.1]
2	14.5	0.739 [.714, .767]	0.37 [.34, .39]	0.859 [.836, .882]	0.54 [0.20, 0.94]	8.4 [6.9, 9.8]
1	29.1	0.731 [.706, .759]	0.35 [.32, .37]	0.856 [.833, .877]	0.54 [0.20, 0.94]	8.9 [7.5, 10.3]
0.5	58.1	0.702 [.675, .730]	0.27 [.25, .29]	0.822 [.798, .844]	0.07 [0.00, 0.20]	13.0 [11.3, 14.7]
0.25	116	0.667 [.637, .700]	0.10 [.08, .11]	0.794 [.771, .817]	0.00 [0.00, 0.00]	18.6 [16.6, 20.8]
0.1	291	0.560 [.524, .593]	0.00 [.00, .01]	0.771 [.749, .792]	0.00 [0.00, 0.00]	22.8 [20.7, 25.1]

The ∞ row retains clipping and adds no noise. Rows within each setup use the same trials and clipped gradients. A $[0.00, 0.00]$ bootstrap interval indicates zero observed errors; for reference, the approximate one-sided 95% upper bound for zero errors in 1500 independent trials is $3/1500 = 0.20\%$.

[1] At sensitivity $2C$ and $\delta_{\mathrm{DP}} = 10^{-5}$, the minimum budgets for rows labeled $8, 4, 2, 1, 0.5, 0.25, 0.1$ are approximately $18.899, 7.914, 3.511, 1.610, 0.751, 0.353, 0.130$, respectively. These values use the unrounded noise levels.

Observed Utility. At $\varepsilon_{\mathrm{DP}} = 1$, Toy's help-vs-hurt AUROC remains 0.985, while ISIC Setup A's changes from 0.741 to 0.731. Committed accuracy decreases from 0.790 to 0.784 on Toy and from 0.875 to 0.856 on ISIC Setup A. Thus ranking changes little at this budget, although decision metrics also change. With stronger noise at $\varepsilon_{\mathrm{DP}} = 0.1$, Toy retains an AUROC of 0.983 but committed accuracy falls to 0.579. On ISIC Setup A, AUROC falls to 0.560 $[0.524, 0.593]$. Its committed accuracy of 0.771 is close to the hurt rate of 0.78, with only 0.4% of trials abstained on, so this accuracy alone does not demonstrate useful discrimination.

Across the sweep, abstention falls from 0.33 to 0.12 on Toy and from 0.38 to approximately zero on ISIC Setup A. Over-caution errors increase from 14.1% to 37.2% and from 7.4% to 22.8%, respectively. The observed dangerous-error rate is at most 0.54% in either setup. These are empirical outcomes, not a guarantee that privacy noise cannot produce a harmful acceptance.

Effect of Noise on the Score. To isolate the Gaussian contribution, condition on the recipient data and clipped donor sum, and write $\hat{\Delta}(g)$ for the score evaluated with donor message g. Let $\hat{D}$ denote the diagonal inverse used in that calculation . Since the first-order term is linear and the curvature term is quadratic in the message,

$$\mathbb{E}_{\xi}\left[\hat{\Delta}(\tilde{g}_M) - \hat{\Delta}(g_M^{\mathrm{clip}})\right] = \frac{1}{2}\left(\frac{\alpha}{m}\right)^2 \sigma^2 \operatorname{tr}(\hat{D}) \geq 0.$$

This gives an upward shift in the conditional mean score. It does not imply that every noisy score increases: the first-order term and the quadratic cross term can change in either direction. Nor does it imply a monotone change in decisions, since the bootstrap threshold also depends on the noisy message. The nonnegative curvature property in Remark 1 does not exclude harmful acceptances under noise.

Subtracting the displayed expectation is post-processing and uses no additional privacy budget. It removes the conditional mean inflation relative to the clipped, noiseless score, but leaves clipping bias and random noise. We report the uncorrected score; the effect of this correction on accuracy and abstention has not been evaluated.

Scope. The privacy guarantee covers donor records under the stated adjacency. The recipient's transmitted model has no corresponding privacy guarantee; protecting it would require a separate private fitting or release mechanism. Theorem 1 and the bootstrap analysis do not establish decision-accuracy guarantees for the clipped, noisy score. The results here measure its empirical utility under donor-message privacy.

C Extended Reading

C.1 Algorithm

For logistic regression, write $p_\theta(x) = (1 + e^{-\theta^\top x})^{-1}$ and

$$\ell(x, y; \theta) = -y \log p_\theta(x) - (1 - y) \log(1 - p_\theta(x)), \qquad \nabla_\theta \ell(x, y; \theta) = (p_\theta(x) - y)x.$$

Algorithm 1 writes out the computations of Section 3. The recipient holds N_{fit} and N_{val}, and the donor holds M; these datasets remain at their respective centers. Each bootstrap index list retains repeated draws, so its sums count observations with their resampled multiplicities.

C.2 Extended Related Work

Mechanisms by Which More Data Can Hurt. Double descent concerns non-monotonic test error as model complexity varies (Belkin et al., 2019); increasing sample size can also worsen performance near the interpolation threshold (Nakkiran et al., 2021). The benign-overfitting literature studies when interpolating noisy data still permits accurate prediction (Bartlett et al., 2020). Our sign-recovery analysis uses a fixed-dimensional logistic model with sufficient samples and controlled conditioning. Its sample-size requirements are separate from A3's population Fisher conditioning assumption. Although donor-specific shortcuts motivate our setting, the sign-recovery results do not require identifying shortcut coordinates.

Implementation of Interaction-Aware Influence. Heo et al. (2026) use damped Gauss–Newton

Algorithm 1 Pre-pooling assessment with two messages and recipient-only bootstrap.

Require: Recipient folds $N_{\text{fit}}, N_{\text{val}}$; donor data M kept on the donor side; multiplier $k > 0$; fixed integer $B \geq 6$

Ensure: HELPFUL (POOL), HARMFUL (DECLINE), or UNCERTAIN (ABSTAIN)

$n_1 \leftarrow |N_{\text{fit}}|, n_2 \leftarrow |N_{\text{val}}|$

$\hat{\theta}_N \leftarrow \arg\min_\theta \frac{1}{n_1} \sum_{i \in N_{\text{fit}}} \ell(x_i, y_i; \theta)$

Step 1: Two messages

Recipient sends $\hat{\theta}_N$ to donor

Donor computes $m \leftarrow |M|$ and $g_M \leftarrow \sum_{i \in M} (p_{\hat{\theta}_N}(x_i) - y_i) x_i$

Donor returns (g_M, m) to recipient

Step 2: Estimating the pooling effect

$\alpha \leftarrow m/(n_1 + m)$

$p_i \leftarrow p_{\hat{\theta}_N}(x_i)$ for $i \in N_{\text{fit}} \cup N_{\text{val}}$

$\hat{F}_{N,jj} \leftarrow \frac{1}{n_1} \sum_{i \in N_{\text{fit}}} p_i(1 - p_i) x_{i,j}^2, j = 1, \ldots, d$

$\hat{w}_{N,j} \leftarrow \left(\frac{1}{n_2} \sum_{i \in N_{\text{val}}} (p_i - y_i) x_{i,j} \right) / \hat{F}_{N,jj}, j = 1, \ldots, d$

$\hat{\Delta} \leftarrow -\frac{\alpha}{m} \sum_{j=1}^d \hat{w}_{N,j} g_{M,j} + \frac{1}{2} \left(\frac{\alpha}{m} \right)^2 \sum_{j=1}^d \frac{g_{M,j}^2}{\hat{F}_{N,jj}}$

for $b = 1, \ldots, B$ **do**

  Independently draw n_1 indices $I^{(b)}$ from N_{fit} and n_2 indices $J^{(b)}$ from N_{val}, with replacement

  $\hat{\theta}_N^{(b)} \leftarrow \arg\min_\theta \frac{1}{n_1} \sum_{i \in I^{(b)}} \ell(x_i, y_i; \theta)$

  $p_i^{(b)} \leftarrow p_{\hat{\theta}_N^{(b)}}(x_i)$ for indices in $I^{(b)}$ and $J^{(b)}$

  $\hat{F}_{N,jj}^{(b)} \leftarrow \frac{1}{n_1} \sum_{i \in I^{(b)}} p_i^{(b)}(1 - p_i^{(b)}) x_{i,j}^2, j = 1, \ldots, d$

  $\hat{w}_{N,j}^{(b)} \leftarrow \left(\frac{1}{n_2} \sum_{i \in J^{(b)}} (p_i^{(b)} - y_i) x_{i,j} \right) / \hat{F}_{N,jj}^{(b)}, j = 1, \ldots, d$

  $\hat{\Delta}^{(b)} \leftarrow -\frac{\alpha}{m} \sum_{j=1}^d \hat{w}_{N,j}^{(b)} g_{M,j} + \frac{1}{2} \left(\frac{\alpha}{m} \right)^2 \sum_{j=1}^d \frac{g_{M,j}^2}{\hat{F}_{N,jj}^{(b)}}$

end for

$\overline{\hat{\Delta}}^* \leftarrow \frac{1}{B} \sum_{b=1}^B \hat{\Delta}^{(b)}$

$\hat{\varepsilon} \leftarrow \left[\frac{1}{B} \sum_{b=1}^B (\hat{\Delta}^{(b)} - \overline{\hat{\Delta}}^*)^2 \right]^{1/2}$

Step 3: Three-way decision

if $\hat{\Delta} < -k\hat{\varepsilon}$ **then**

  return HELPFUL (POOL)

else if $\hat{\Delta} > k\hat{\varepsilon}$ **then**

  return HARMFUL (DECLINE)

else

  return UNCERTAIN (ABSTAIN)

end if

curvature, EK-FAC, and low-rank gradient projection. Their general formulation permits expansion points that do not minimize the target risk. Their example-selection procedure also requires candidate-specific quantities beyond an aggregate group gradient. Appendix B.2 describes our adaptation of their group estimator; we do not reproduce their selection algorithm.

C.3 EXTENDED LIMITATIONS AND FUTURE WORK

Representational Scope. Our analysis concerns logistic regression on a shared, fixed feature representation. The imaging experiments use linear probes on frozen encoders, so the assessed pooling effect concerns retraining the probe while keeping the representation fixed. Both centers must use compatible feature definitions, preprocessing, and parameter coordinates; sharing an encoder alone does not ensure this compatibility. The protocol assumes these choices are aligned and does not verify them. We have not established that a probe's pooling assessment predicts the effect of jointly fine-tuning an entire network. Training the representation introduces nonconvex optimization and falls outside the logistic-regression analysis developed here.

Correctness of the Donor Message. The method assumes that the donor computes the requested aggregate correctly. A fabricated message can distort the score and its bootstrap threshold, and the statistical guarantees do not cover such messages. The donor-message privacy mechanism in Appendix B.9 protects donor records under its stated adjacency; it does not verify the donor's com-

putation. We have not evaluated adversarial donors. Verifying donor computations while retaining the communication and data-sharing constraints is a direction for future work.

Adaptive Pooling Weights. The current rule assesses pooling at the fixed weight α and chooses among pooling, declining, and abstaining. The same donor message and recipient quantities also permit evaluating the quadratic score in equation 5 at alternative weights, without further communication. This suggests choosing how much weight to assign to the donor. However, minimizing that approximate score does not guarantee a reduction in recipient risk. Selecting a weight from the data would require analysis of approximation error and statistical uncertainty across candidate weights, beyond the guarantee at a fixed weight established here. We have not evaluated this extension empirically.