RELEASING PREDICTED PROBABILITIES: SHARP LIMITS ON EXCESS DISCLOSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Releasing predicted class probabilities can reveal sensitive information beyond what optimal prediction requires. We show that perfectly calibrated forecasts can approach the optimal expected log loss while revealing a sensitive bit independent of the true conditional class probabilities. We measure this excess disclosure as the largest reduction in optimal expected loss from observing the release instead of the true probabilities, over decisions about other attributes with losses in $[0, 1]$. We then determine the prediction cost of limiting this disclosure through randomization. For binary forecasts with log-loss regret at most ϵ, the smallest worst-case bound on both prediction regret after release and excess disclosure is $(1 + o(1))(\epsilon/12)^{1/3}$ as $\epsilon \downarrow 0$. Uniform noise after an arcsine square-root transformation attains this bound, and we prove a matching lower bound for all randomized releases. For any fixed number of classes, releasing counts of labels sampled independently from each forecast achieves the same cube-root rate. In experiments on retinal images and chest radiographs, this sampling method makes images from the same patient harder to link, at the cost of higher average prediction loss when each task uses only its own released probabilities.

1 INTRODUCTION

Sharing predicted disease probabilities lets a collaborator study disease risk without receiving patient images. The probabilities may also reveal other attributes or help identify images of the same person. The least-privilege principle suggests sharing only what the intended task requires (Saltzer & Schroeder, 1975; Stadler et al., 2024). Applying this principle to prediction requires a way to distinguish information needed for the task from additional disclosure.

Under log loss, the true class probabilities conditional on the input minimize expected loss. We call these probabilities the *population posterior*. Any release that permits the same expected loss must allow the recipient to recover them (Gneiting & Raftery, 2007; Sevetlidis, 2026). The posterior therefore provides a reference for measuring additional disclosure by a learned forecast. It may itself reveal sensitive information, as when disease risk varies with age.

Small prediction errors can reveal an attribute completely. Suppose everyone has disease probability $1/2$, while a model reports $1/2 - \delta$ for one sensitive group and $1/2 + \delta$ for another. As δ tends to zero, loss approaches its minimum, yet the report always identifies the group. Requiring calibration does not resolve this problem. We construct calibrated forecasts on a fixed population that approach optimal log loss while revealing a sensitive bit independent of the posterior. Calibration fixes the average outcome among records assigned the same probability, but that probability can still identify a sensitive group.

Randomization makes nearby forecasts harder to distinguish by giving them overlapping output distributions, at a cost in prediction accuracy. We measure *prediction regret* by expected log loss above its population minimum. We measure *excess disclosure* by the gain in optimal expected payoff for a decision about another attribute, relative to observing the posterior. The holder needs only the learned forecast to generate the randomized release.

For binary forecasts with regret at most ϵ, we minimize the larger of output prediction regret and excess disclosure in the worst case. The optimal value is asymptotic to $(\epsilon/12)^{1/3}$. Adding uniform noise after an arcsine square-root transformation attains this value. The transformation accounts for

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.

the greater sensitivity of log loss near zero and one. The lower bound holds for every randomized release, even when recipients choose prediction rules specific to their population. A release that conceals small differences between forecasts must also introduce prediction error on some exact forecasts.

For categorical prediction with a fixed number of classes, drawing labels from the forecast and releasing the class counts achieves the same cube-root rate. The bounds extend to multiple tasks and repeated releases. Experiments with known populations illustrate these guarantees. On retinal images and chest radiographs, sampled counts reduce patient linkage. To assess their prediction cost, we evaluate both direct use of the released probabilities and predictors fitted to them. These comparisons show how prediction-rule selection affects the observed loss.

2 Prediction and disclosure

2.1 Records, forecasts, and releases

A data holder predicts a categorical label Y from a record X, such as disease grade from a retinal image. A fitted model T predicts class probabilities $\widehat{Q} = q(X)$, which we call a *forecast*. A *release channel* K randomizes this forecast to produce $R \sim K(\cdot \mid \widehat{Q})$, and the recipient predicts Y using a rule $d(R)$, or *decoder*. The model and channel are public and fixed. The holder releases only R, with randomness independent of the record, label, and other attributes conditional on $\widehat{Q}$.

The log loss is $-\log q_Y$ for a reported probability vector q. The optimal prediction from the record is the vector of true conditional class probabilities,

$$Q^\star(X) = P(Y = \cdot \mid X, T), \tag{1}$$

which we call the *population posterior*. The expected loss at $Q^\star$ is the Bayes risk. Strict propriety of log loss means that a recipient attains this risk if and only if $d(R) = Q^\star$ almost surely (Gneiting & Raftery, 2007; Reid & Williamson, 2011; Sevetlidis, 2026). Thus $Q^\star$ provides a reference for measuring disclosure beyond what optimal prediction requires. Appendix A extends this argument to other strictly proper losses.

The forecast's expected loss above the Bayes risk is its *prediction regret*:

$$\epsilon = \mathbb{E} D_{\mathrm{KL}}(Q^\star \| \widehat{Q}). \tag{2}$$

We distinguish this input regret from the output regret, $\mathbb{E} D_{\mathrm{KL}}(Q^\star \| d(R))$, which also includes the effects of randomization and decoding. Here $D_{\mathrm{KL}}(p\|q) = \sum_j p_j \log(p_j/q_j)$, and kl denotes the binary case. We use natural logarithms, so losses are in nats, with standard extended values at zero probabilities. For several tasks scored by the sum of their log losses, the reference is the tuple $(Q_1^\star, \ldots, Q_J^\star)$ of individual posteriors. Predicting the joint distribution may also require information about dependence between labels.

2.2 Excess disclosure

The recipient may use the release to infer an additional attribute S. To cover different inference tasks, let $u(s, a) \in [0, 1]$ be the payoff for taking action a when $S = s$. The best expected payoff from an observation Z is

$$V_u(S; Z) = \sup_g \mathbb{E} u(S, g(Z)), \tag{3}$$

where g ranges over measurable decision rules. We define *excess disclosure* as the improvement over observing the population posterior:

$$\big[V_u(S; R) - V_u(S; Q^\star)\big]_+,$$

where $[x]_+ = \max\{x, 0\}$. With loss $1 - u$, this is the positive part of the reduction in optimal expected loss. For example, the payoff $u(s, a) = \mathbf{1}\{s = a\}$ gives classification accuracy. If S is a fair bit independent of $Q^\star$, observing the posterior gives accuracy $1/2$. A release that identifies the bit has excess disclosure $1/2$.

When auxiliary information B is available, we compare (B, R) with $(B, Q^\star)$ and require release randomness to be independent of B conditional on the forecast. We take all spaces, including action spaces, to be standard Borel and write $\mathrm{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$ for total variation.

2.3 CALIBRATION AND DETERMINISTIC RELEASES

For binary labels, calibration means $\mathbb{E}[Y \mid \widehat{Q}] = \widehat{Q}$: among records assigned the same probability, the mean label equals that probability. Calibration permits forecasts to reveal an attribute independent of the population posterior, even as their regret tends to zero.

Proposition 1 (Calibrated disclosure)**.** *There is one fixed binary prediction problem and a sequence $\widehat{Q}_m$ such that $\mathbb{E}[Y \mid \widehat{Q}_m] = \widehat{Q}_m$, $\mathbb{E}\,\mathrm{kl}(Q^\star \| \widehat{Q}_m) \to 0$, and every $\widehat{Q}_m$ identifies a sensitive bit independent of $Q^\star$.*

Draw $Q^\star$ uniformly on $[1/4, 3/4]$ and an independent fair bit S. For each value of S, partition the interval into small bins and report the midpoint of the bin containing $Q^\star$. Shift one partition so that the two sets of midpoints are disjoint. The midpoint identifies S because it belongs to only one partition. It also equals the conditional mean of $Q^\star$ within its bin, giving calibration. As the bins shrink, regret vanishes while excess disclosure stays at $1/2$. Appendix A.2 specifies the partitions and their boundary bins.

Deterministic rounding also distinguishes arbitrarily close forecasts on opposite sides of a rounding threshold. Every nonconstant deterministic map on the simplex interior has worst-case excess bit accuracy at least $1/2$ under any positive regret bound (Proposition A2). Randomization controls this distinguishability by making the output distributions overlap.

3 SHARP LIMITS FOR BINARY PREDICTIONS

3.1 MINIMAX FORMULATION

Fix an input-regret bound ϵ. Let $\mathcal{E}_\epsilon$ be the class of populations with binary Y and $\mathbb{E}\,\mathrm{kl}(Q^\star \| \widehat{Q}) \le \epsilon$. Each population specifies a joint law of (X, Y, S) and a measurable forecast based on X, which may be calibrated or uncalibrated. One channel $K : [0,1] \rightsquigarrow \mathcal{R}$ and decoder $d : \mathcal{R} \to [0,1]$ must serve this entire class, although both may depend on ϵ. Their worst-case output regret and disclosure are

$$U^\star_{\epsilon,\log}(K,d) = \sup_{\mathcal{E}\in\mathcal{E}_\epsilon} \mathbb{E}_{\mathcal{E}}\,\mathrm{kl}(Q^\star \| d(R)),$$
$$\mathfrak{L}_\epsilon(K) = \sup_{\mathcal{E}\in\mathcal{E}_\epsilon,\, S,u} [V_u(S;R) - V_u(S;Q^\star)]_+.$$

The second supremum includes every additional attribute and measurable payoff in $[0,1]$. We minimize the larger of these costs:

$$M_{\epsilon,\log} = \inf_{K,d} \max\{U^\star_{\epsilon,\log}(K,d), \mathfrak{L}_\epsilon(K)\}.$$

Theorem 2 (Sharp balanced log-loss tradeoff)**.** *Over all standard Borel output spaces, forecast-only channels, and public measurable decoders,*

$$M_{\epsilon,\log} = (1 + o(1))(\epsilon/12)^{1/3} \qquad \text{as } \epsilon \downarrow 0. \tag{4}$$

The criterion weights the two costs equally. The asymptotic value is unchanged if each recipient chooses the optimal decoder for its own population (Corollary A8). Appendix A.7 treats other disclosure weights.

3.2 UNIFORM NOISE AFTER AN ARCSINE TRANSFORMATION

A small perturbation of a probability costs more in log loss near zero or one than near $1/2$: its second-order cost is $(\Delta q)^2/[2q(1-q)]$. For $\theta(q) = \arcsin\sqrt{q}$, the derivative $\theta'(q) = 1/[2\sqrt{q(1-q)}]$ transforms this approximation to $2(\Delta\theta)^2$, independent of q. This motivates *uniform angular noise*: perturb the transformed probability uniformly and then map back to a probability.

For noise half-width $a > 0$, the release and decoder are

$$R = \theta(q) + Z, \qquad Z \sim \mathrm{Unif}[-a, a], \qquad d_a(R) = \frac{\sin^2 R + a^2/3}{1 + 2a^2/3}. \tag{5}$$

For an exact forecast away from zero and one, $\mathbb{E}Z^2 = a^2/3$ gives a second-order regret of $2a^2/3$. The correction in d_a keeps decoded probabilities away from zero and one, yielding a bound uniform over all input probabilities:

$$\sup_q \mathbb{E}\,\mathrm{kl}(q\|d_a(R)) = (2/3 + o(1))a^2.$$

Forecast error contributes a smaller-order term at the noise width used below (Appendix A.6).

Two uniform distributions of width $2a$ have total variation at most their center separation divided by $2a$. To bound disclosure, compare the release with uniform noise centered at a conditional median of $\theta(\widehat{Q})$ given $Q^\star$. A recipient can simulate this second release from the posterior alone. The gain in decision value is therefore at most the total variation between the actual and simulated releases. The median minimizes the expected absolute separation of the centers, which Lemma A4 bounds by $\sqrt{\epsilon/2}$, giving

$$\mathfrak{L}_\epsilon(K) \le \frac{\sqrt{\epsilon}}{2\sqrt{2}a}.$$

Equating this bound with $2a^2/3$ gives $a^3 = 3\sqrt{\epsilon}/(4\sqrt{2})$ and the value $(\epsilon/12)^{1/3}$. The release itself uses only the forecast and private noise. The holder may publish $d_a(R)$ directly, which preserves prediction and can only reduce disclosure.

3.3 Why every release incurs this cost

A channel that conceals small forecast differences must also incur reconstruction error. Let a fair bit choose between nearby forecasts p and q, while the true probability is the constant $(p+q)/2$. The posterior gives bit accuracy $1/2$, whereas the release gives $1/2+\mathrm{TV}(K_p, K_q)/2$, where $K_q = K(\cdot \mid q)$. When the forecasts are sufficiently close, this population meets the regret budget. A disclosure bound L therefore requires $\mathrm{TV}(K_p, K_q) \le 2L$. Subdividing an interval and applying the triangle inequality extends this restriction to more distant inputs.

Apply the same channel to exact forecasts distributed uniformly on $[1/2 - r, 1/2 + r]$, with fixed $r \in (0, 1/2)$. Averaging over this interval bounds the total variation between the reconstruction-error distribution and shifted copies of that distribution. Such a bound limits how much probability can concentrate near zero (Awan & Ramasethu, 2024). Integrating the resulting lower bound on the error tails gives a lower bound on mean squared error (Lemma A3).

By Pinsker's inequality, log-loss regret is at least twice this mean squared error. Taking $\epsilon \downarrow 0$ and then $r \downarrow 0$ gives prediction regret at least $(1 + o(1))\epsilon/(12L^2)$ near the optimum. Thus, if both costs are at most M, then $M^3 \ge (1 + o(1))\epsilon/12$. The argument bounds every decoder on the same population, including its optimal decoder. Appendix A.4 proves the result, and Table A1 gives the squared-error counterparts.

4 Categorical predictions and repeated releases

4.1 Releasing sampled class counts

For each record, draw n synthetic labels independently from its forecast q and release how many times each class occurs:

$$C \mid \widehat{Q} = q \sim \mathrm{Multinomial}(n, q). \tag{6}$$

The sample size n is fixed and public. A larger n gives a more accurate estimate of the forecast, but also makes nearby forecasts easier to distinguish. To predict from the counts, use the smoothed frequencies $D_\alpha = (C + \alpha)/(n + k\alpha)$ for k classes and $\alpha > 0$. Smoothing prevents a zero count from producing an infinite log loss.

With add-one smoothing, output regret is at most $\epsilon + \log[(n+k)/(n+1)]$ and excess disclosure is at most $\sqrt{n\epsilon/2}$ (Theorems A17 and A12). The disclosure proof compares the released counts with counts drawn from $Q^\star$, which a recipient can simulate from the posterior. The KL divergence between the count distributions grows linearly with sample size, and Pinsker's inequality gives the disclosure bound. For fixed k, choosing n of order $\epsilon^{-1/3}$ balances the $O(1/n)$ sampling cost and

$O(\sqrt{n\epsilon})$ disclosure, giving the optimal cube-root rate. The experiments use half-count smoothing, $\alpha = 1/2$, whose bound has the same order. Appendix A.12 gives the sharp squared-error result for complete binomial counts.

4.2 Repeated releases and sample size

Independent count releases of fixed sizes $n_1, \ldots, n_t$ contain the same information as their sum, a count vector with $\sum_i n_i$ draws. Conversely, randomly allocating these counts among groups of the original sizes reproduces the separate releases without knowing the forecast. Returning the same previously generated release adds no information.

For several tasks, independent draws give joint disclosure at most $\sqrt{\frac{1}{2}\sum_j n_j \epsilon_j}$. Given task-regret bounds $\bar{\epsilon}_j$, choosing $\sum_j n_j \bar{\epsilon}_j \leq 2\delta^2$ therefore guarantees excess disclosure at most δ. These bounds concern regret relative to the unknown Bayes risk. Appendix A.18 extends them to multiple records and auxiliary information, and Appendix A.19 gives the regret required at a specified sample size.

5 Related work

Prediction and disclosure. Strictly proper scoring rules identify the true conditional distribution as the unique optimal forecast (Gneiting & Raftery, 2007; Reid & Williamson, 2011; Sevetlidis, 2026). This gives the reference information for our disclosure criterion. Its formulation through optimal decision payoffs follows the theory of statistical experiments (Le Cam, 1964; Torgersen, 1991). In least-privilege learning, Stadler et al. (2024) use inference from the task label as the baseline for measuring attribute disclosure. Our baseline is the population posterior, and our bound concerns the additive gain in optimal expected payoff for any decision problem with payoffs in $[0, 1]$. Feng et al. (2026) study decisions about the predicted outcome when recipients act on reported probabilities. We consider decisions about an additional attribute.

Inference from representations and outputs. Learned features can reveal attributes beyond their training task (Osia et al., 2017; Melis et al., 2019). In medical imaging, they can also support patient linkage (Nebbia et al., 2025; Packhäuser et al., 2022; Macpherson et al., 2023), although near-duplicate images can inflate measured performance (Engelmann et al., 2026). Concept-erasure methods aim to remove a specified attribute, with guarantees or evaluations tied to a chosen family of predictors (Belrose et al., 2023; Holstege et al., 2025; George et al., 2026; Chowdhury et al., 2023; Akbari et al., 2025). PriFU learns task-relevant features without training an adversary to predict sensitive labels (Bi et al., 2024). Sharing only model outputs can still reveal additional information (Fedzechkina et al., 2026). We show that an attribute independent of the true task probabilities can remain fully recoverable from forecasts with perfect calibration and vanishing prediction regret.

Randomization and optimal noise. Privacy funnels choose a randomized mapping for a specified joint distribution of useful and sensitive attributes (Calmon & Fawaz, 2012; Makhdoumi et al., 2014; Rassouli & Gündüz, 2020). Local differential privacy bounds how distinguishable outputs can be across inputs (Duchi et al., 2018; Cai et al., 2024). Bounding excess disclosure over our population class similarly limits the distinguishability of nearby forecasts. This restriction underlies our lower bound. Under quadratic cost and $(0, \delta)$ differential privacy, Geng et al. (2019) establish the optimality of uniform additive noise for small δ within a symmetric, monotone noise class. We use the interval-probability inequality of Awan & Ramasethu (2024) to obtain a lower bound for arbitrary release channels, and attain it asymptotically under log loss with uniform angular noise. Bounds for sampling, smoothing, and stochastic rounding support the other release methods we study (Mourtada, 2026; Connolly et al., 2021).

6 Experiments

We use populations with known task probabilities to measure prediction regret and excess disclosure directly. We then study prediction loss and patient linkage when releasing randomized forecasts for medical images.

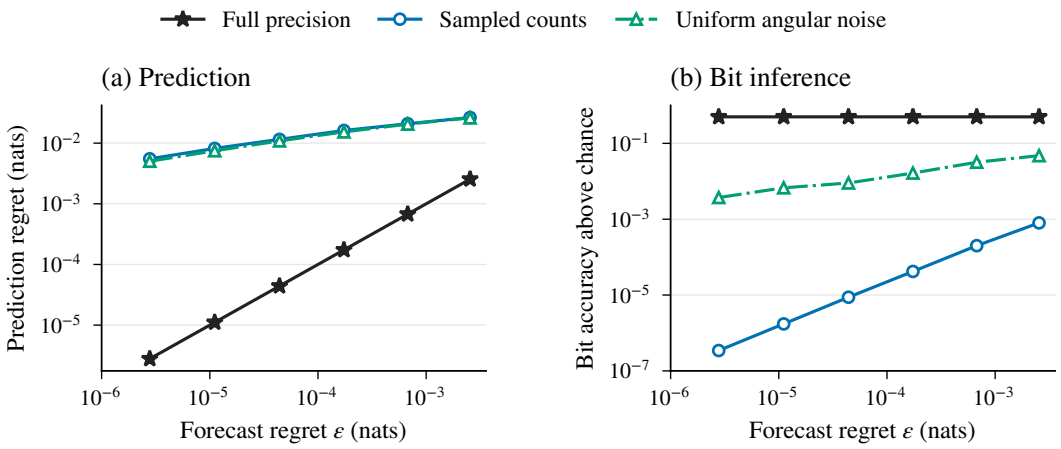

Figure 1: Prediction and disclosure for the calibrated example. The horizontal axis is the input forecast's log-loss regret. The vertical axes show the optimal recipient's prediction regret (left) and bit accuracy above chance (right). Sampled counts use $n = \lceil \epsilon^{-1/3} \rceil$, from 8 to 72 draws. Uniform angular noise uses the half-width defined in Section 3. Axes are logarithmic.

6.1 Controlled experiments with known probabilities

The calibrated forecasts in Proposition 1 predict a binary label and also reveal a sensitive bit independent of the true task probability. Finer partitions reduce forecast regret while leaving the bit identifiable. For $m \in \{4, 8, 16, 32, 64, 128\}$ bins in the unshifted partition, we compute the optimal prediction $\mathbb{E}[Q^\star \mid R]$ and the best achievable bit-guessing accuracy after release. Observing only the true task probability gives bit accuracy $1/2$.

Full precision reveals the bit throughout Figure 1, even as input regret falls by nearly three orders of magnitude. With sampled counts, output regret falls from 0.0265 to 0.0055 nats and bit accuracy above chance falls from 8.0×10^{-4} to 3.4×10^{-7}. Counts conceal this particular bit more effectively than uniform angular noise, whose asymptotic optimality concerns the worst case over all admissible populations. Appendix B gives further controlled experiments.

6.2 Medical data and release methods

The exploratory medical study uses 16,266 retinal images from BRSET and 112,120 radiographs from NIH ChestX-ray14 (Nakayama et al., 2024; Wang et al., 2017). BRSET targets are five-class diabetic retinopathy (DR), image quality, macular edema, and optic-disc abnormality. NIH targets are atelectasis, cardiomegaly, consolidation, edema, and effusion. We use disjoint patient sets for training, development selection, and evaluation, retaining NIH's official evaluation set. All five training seeds share this split (Appendix C).

We train adapters and task heads on a frozen DINOv3 ViT-S/16 backbone, then calibrate its predictions using training patients (Siméoni et al., 2026; Chen et al., 2022). We compare the unchanged float32 probabilities with *sampled counts*: for each image and task, we draw n independent labels from the predicted distribution and release their smoothed frequencies. A binary prediction q gives Bernoulli draws, and larger n preserves the forecast more closely. *Deterministic rounding* replaces each probability vector with a nearby vector on the same $1/n$ frequency grid, followed by the same smoothing. Comparing these methods separates the effect of random sampling from that of finite resolution.

For binary tasks, *uniform angular noise* adds noise to $\arcsin\sqrt{q}$ as in Section 3. We also evaluate *stochastic angular rounding*, which randomly rounds the same transformed value to an adjacent grid point. Both methods convert the result back to a probability. We compare all four methods on the binary targets, removing BRSET's five DR outputs before release and downstream fitting. NIH retains all five targets. Appendix D compares task probabilities with image features and feature-erasure methods.

6.3 PREDICTION AND PATIENT VERIFICATION

Recipients can use the released probabilities directly or fit predictors to them. We select each target's prediction rule by development log loss and compare two sets of candidates. The *per-task rules* use only that target's probabilities: direct use, recalibration, removal of count smoothing, and a prevalence baseline based on smoothed training class frequencies. The *full candidate set* also includes classifiers trained on all released tasks and rules that first transform the release to a coarser resolution. These candidates let recipients use relations among tasks or reduce the release's precision (Appendix C.6).

The binary-task comparison adds angular transformations to this candidate set, so selection can change even on NIH, where the forecasts stay the same. We divide each task's evaluation log loss by the prevalence baseline's loss, then average equally across tasks. A normalized loss of one matches the baseline. Within each task, we weight valid images equally, including multiple images from the same patient. The resulting loss reflects both the release and the selected rule. The theoretical optimum allows the best possible rule for the population.

Patient verification tests whether two released vectors come from the same patient (Packhäuser et al., 2022; Macpherson et al., 2023). Positive pairs use opposite eyes on BRSET and the first and last images in NIH's recorded follow-up order. Negative pairs use different patients. We also construct matched negatives by replacing one image in each positive pair with another patient's image that matches the replaced image's recorded target labels, demographics, and acquisition settings. For each pair set, we fit logistic regression, multilayer perceptron, and histogram gradient-boosting classifiers, selecting by development area under the receiver operating characteristic curve (AUROC). Verification measures the performance of these fitted classifiers, whereas excess disclosure concerns the optimal recipient (Appendix C.7).

For randomized methods, we generate three independent releases per image and training seed, then average metrics over these repetitions. We report the five-seed mean and sample standard deviation (SD). At $n = 256$, we also report 95% confidence intervals from 2,000 paired patient-bootstrap draws (Owen & Eckles, 2012). These intervals measure patient-sampling uncertainty conditional on fitted models and development selections. The SD measures training-seed variation on the shared patient split.

6.4 PREDICTION COST AND PATIENT VERIFICATION

Sampled counts lower patient-verification AUROC in both datasets (Figure 2). At $n = 256$ with all target tasks, verification AUROC falls from 0.788 to 0.746 on BRSET and from 0.716 to 0.685 on NIH. It also falls with matched negative pairs, from 0.689 to 0.652 and from 0.670 to 0.626, respectively. All eight confidence intervals for AUROC changes in the full-task and binary-task comparisons are below zero (Table 1).

Sampling increases prediction loss in both datasets when selection uses only per-task rules: normalized loss rises by $+0.0077$ on BRSET and $+0.0089$ on NIH, with both confidence intervals above zero. With the full candidate set, loss still increases on NIH ($+0.0084$), but the mean change on BRSET is -0.0015 and its interval includes zero (Table 1). The binary-task comparisons follow the same pattern. The measured prediction cost therefore depends on the recipient's choice of rule.

6.5 WHY THE SELECTED PREDICTION RULE MATTERS

Selecting a prediction rule can improve on using coarse probabilities directly. For BRSET probabilities rounded at $n = 8$, normalized loss is 0.793 with direct use, 0.608 with per-task rules, and 0.599 with the full candidate set.

Development selection can also favor predictors that generalize poorly. For full-precision BRSET forecasts, selection from the full candidate set raises normalized loss from 0.580 to 0.590 relative to the per-task rules. In training seed 4, the selected macular-edema rule rounds the probabilities at resolution 128 and removes smoothing. It has lower development loss than the per-task choice (0.0529 versus 0.0546 nats), but higher evaluation loss (0.0850 versus 0.0649). With selection from the full candidate set at $n = 256$, sampling raises loss in four BRSET seeds, but the reduction of

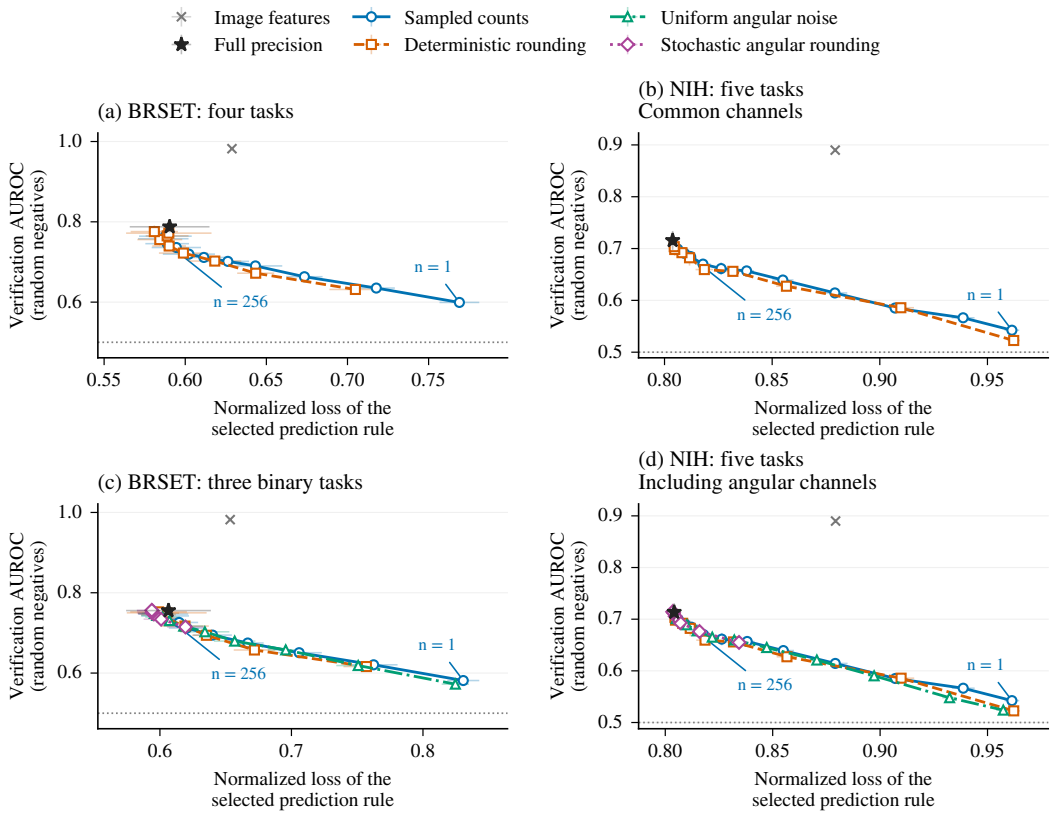

Figure 2: Prediction loss after development selection from the full candidate set, and verification with random negative pairs. Points show five-seed means and bars one sample SD, with randomized releases averaged within seed. Labels give sampled-label counts per image and task. Both NIH panels use the same five targets. Lines connect parameter settings specified in Appendix C; the dotted line marks AUROC 0.5.

0.0306 in seed 4 reverses the mean difference. Figure A4 shows the loss changes for every seed in all four comparisons.

Average loss also conceals variation among diseases. On NIH, deterministic rounding at $n = 16$ raises normalized loss by 0.0147 on average and by 0.0421 for cardiomegaly. Appendix F.2 reports every target separately.

7 Discussion

The disclosure guarantee requires a bound on forecast regret relative to the population Bayes risk. This risk is unknown in the medical studies. For example, with 256 sampled labels per record and task and a bound $\bar{\epsilon}$ on summed task regret, excess disclosure for one record is at most $\min\{1, \sqrt{128\bar{\epsilon}}\}$. A sufficient condition for a bound of 0.05 is $\bar{\epsilon} \leq 1.953125 \times 10^{-5}$ nats (Appendix A.19). The medical experiments do not establish this regret bound and hence do not certify that disclosure level.

Downstream predictors learn from cross-fitted training forecasts but receive forecasts from the final image model and calibrator at development and evaluation time. The image-model and calibration cross-fitting steps are not nested, so a training patient's labels can indirectly affect that patient's release. Neither fitting stage uses evaluation labels. Both the indirect label dependence and the change in forecasting models can affect downstream generalization.

Table 1: Changes from full precision to 256 sampled labels per image and task. Each estimate has a 95% paired patient-bootstrap interval below it. Prediction rules and verification attacks are selected on development data. Positive loss changes indicate worse prediction; negative AUROC changes indicate less successful verification. Common channels are sampled counts and deterministic rounding. The lower block also permits angular transformations during prediction and verification rule selection. Both NIH rows use the same five target forecasts.

	Change in normalized loss		Change in verification AUROC	
Target tasks	Per-task rules	Full candidate set	Random negatives	Matched negatives
Common channels				
BRSET (four tasks)	$+0.0077$ $[+0.0028, +0.0113]$	-0.0015 $[-0.0127, +0.0075]$	-0.042 $[-0.051, -0.033]$	-0.036 $[-0.048, -0.025]$
NIH (five tasks)	$+0.0089$ $[+0.0083, +0.0096]$	$+0.0084$ $[+0.0076, +0.0091]$	-0.031 $[-0.040, -0.023]$	-0.044 $[-0.056, -0.033]$
Including angular channels				
BRSET (three tasks)	$+0.0101$ $[+0.0069, +0.0138]$	-0.0026 $[-0.0110, +0.0045]$	-0.013 $[-0.023, -0.004]$	-0.011 $[-0.020, -0.001]$
NIH (five tasks)	$+0.0089$ $[+0.0083, +0.0096]$	$+0.0080$ $[+0.0072, +0.0088]$	-0.029 $[-0.038, -0.020]$	-0.044 $[-0.056, -0.033]$

From development to evaluation, NIH cardiomegaly prevalence rises from 2.04% to 4.18%, effusion from 9.79% to 18.20%, and the anteroposterior-view share from 35.03% to 56.65%. Differences of this size could affect calibration and the generalization of selected predictors. Our comparisons establish neither the cause of these shifts nor their contribution to the observed errors.

8 Conclusion

Accurate, calibrated forecasts can disclose information beyond the true task probabilities. Our lower bound shows that limiting this disclosure also incurs prediction error on some exact forecasts. We characterize this tradeoff sharply for binary log loss and attain the optimal asymptotic value with uniform angular noise. For categorical predictions, sampled label counts attain the same cube-root rate and give repeated-release bounds. In both medical datasets, sampled counts reduce patient linkage and increase prediction loss when each task uses only its own released probabilities. The broader prediction-rule comparison shows why measured loss also depends on fitting and selection: a rule preferred on development data can perform worse on evaluation data. Assessing a release therefore requires specifying how recipients predict from it.

ETHICS STATEMENT

We use BRSET v1.0.2 under PhysioNet's credentialed access and data-use terms (Pollard et al., 2026; Nakayama et al., 2026), and the de-identified NIH ChestX-ray14 release with its 2020 metadata. BRSET's original collection received UNIFESP institutional approval and a consent waiver. We obtained no separate institutional review for this secondary analysis. The supplement contains aggregate results without patient images, identifiers, or record-level predictions.

REPRODUCIBILITY STATEMENT

The appendix provides proofs, dataset partitions, fitting procedures, and additional results. Code and aggregate results for reproducing the figures and tables are available in an anonymous repository.[1] Repeating model fitting requires the licensed data and model checkpoints.

AI USE STATEMENT

We used Codex to find related work, polish the writing, help with LaTeX formatting, and write and debug code. We also asked it to check our proofs and improve their exposition and advise on research and implementation choices. We developed the research ideas, interpreted the results, and designed the figures ourselves. We read and checked all AI-assisted work and made the final decisions. We take responsibility for the paper and supplementary materials.

[1] https://anonymous.4open.science/r/predicted-probabilities-7D7F

REFERENCES

Ramin Akbari, Milad Afshari, and Vishnu Boddeti. Obliviator reveals the cost of nonlinear guardedness in concept erasure. In *Advances in Neural Information Processing Systems*, volume 38, pp. 160707–160739, 2025. doi: 10.52202/085713-4847.

Jordan Awan and Aishwarya Ramasethu. Optimizing noise for f-differential privacy via anti-concentration and stochastic dominance. *Journal of Machine Learning Research*, 25(351):1–32, 2024. URL https://jmlr.org/papers/v25/23-1624.html.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023.

Xiuli Bi, Yang Hu, Bo Liu, Weisheng Li, Pamela Cosman, and Bin Xiao. PriFU: Capturing task-relevant information without adversarial learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3104–3112, 2024. doi: 10.1145/3664647.3681216.

Zhongteng Cai, Xueru Zhang, and Mohammad Mahdi Khalili. Privacy-aware randomized quantization via linear programming. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 499–516, 2024.

Flavio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1401–1408, 2012.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678, 2022. doi: 10.52202/068431-1212.

Somnath Basu Roy Chowdhury, Nicholas Monath, Kumar Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43284–43306, 2023. doi: 10.52202/075280-1875.

Michael P. Connolly, Nicholas J. Higham, and Theo Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing*, 43(1):A566–A585, 2021. doi: 10.1137/20M1334796.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018. doi: 10.1080/01621459.2017.1389735.

Justin Engelmann, Siegfried K. Wagner, Yukun Zhou, and Pearse A. Keane. Matters arising: Near-identical images, not foundation models, explain purported re-identification of patients from medical imaging. *npj Digital Medicine*, 9:159, 2026. doi: 10.1038/s41746-026-02440-9.

Masha Fedzechkina, Eleonora Gualdoni, Rita Ramos, and Sinead Williamson. What do your logits know? (the answer may surprise you!). *arXiv preprint arXiv:2604.09885*, 2026.

Yiding Feng, Liuhan Qian, and Wei Tang. Is this predictor more informative than another? a decision-theoretical comparison. *arXiv preprint arXiv:2507.12094v2*, 2026. URL https://arxiv.org/abs/2507.12094v2.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Optimal noise-adding mechanism in additive differential privacy. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 11–20, 2019. URL https://proceedings.mlr.press/v89/geng19a.html.

Daniel George, Charles Yeh, Daniel Lee, and Yifei Zhang. From measurement to mitigation: Quantifying and reducing identity leakage in image representation encoders with linear subspace removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3284–3293, 2026.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30563–30603, 2025. doi: 10.52202/085713-0907.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964. doi: 10.1214/aoms/1177700372.

Matthew S. Macpherson, Charles E. Hutchinson, Carolyn Horst, Vicky Goh, and Giovanni Montana. Patient reidentification from chest radiographs: An interpretable deep metric learning approach and its applications. *Radiology: Artificial Intelligence*, 5(6):e230019, 2023. doi: 10.1148/ryai.230019.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *IEEE Information Theory Workshop*, pp. 501–505, 2014. doi: 10.1109/ITW.2014.6970882.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy*, pp. 691–706, 2019. doi: 10.1109/SP.2019.00029.

Jaouad Mourtada. Estimation of discrete distributions in relative entropy, and the deviations of the missing mass. *Mathematical Statistics and Learning*, 9(1/2):69–144, 2026. doi: 10.4171/MSL/55.

Luis Filipe Nakayama, David Restrepo, Joao Matos, Lucas Zago Ribeiro, Fernando Korn Malerbi, Leo Anthony Celi, and Caio Saito Regatieri. BRSET: A Brazilian multilabel ophthalmological dataset of retina fundus photos. *PLOS Digital Health*, 3(7):e0000454, 2024. doi: 10.1371/journal.pdig.0000454.

Luis Filipe Nakayama, Mariana Goncalves, Lucas Zago Ribeiro, Helen Santos, Daniel Ferraz, Fernando Malerbi, Leo Anthony Celi, and Caio Regatieri. A Brazilian multilabel ophthalmological dataset (BRSET). PhysioNet, version 1.0.2, 2026. URL `https://physionet.org/content/brazilian-ophthalmological/1.0.2/`.

Giacomo Nebbia, Sourav Kumar, Stephen Michael McNamara, Christopher Bridge, J. Peter Campbell, Michael F. Chiang, Naresh Mandava, Praveer Singh, and Jayashree Kalpathy-Cramer. Re-identification of patients from imaging features extracted by foundation models. *npj Digital Medicine*, 8(1):469, 2025. doi: 10.1038/s41746-025-01801-0.

Seyed Ali Osia, Ali Shahin Shamsabadi, Ali Taheri, Kleomenis Katevas, Hamid R. Rabiee, Nicholas D. Lane, and Hamed Haddadi. Privacy-preserving deep inference for rich user data on the cloud. *arXiv preprint arXiv:1710.01727*, 2017.

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. doi: 10.1214/12-AOAS547.

Kai Packhäuser, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. *Scientific Reports*, 12(1):14851, 2022. doi: 10.1038/s41598-022-19045-3.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Borzoo Rassouli and Deniz Gündüz. Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 15:594–603, 2020. doi: 10.1109/TIFS.2019.2903658.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939.

Vasileios Sevetlidis. Bayes-sufficient representations in supervised learning. *arXiv preprint arXiv:2606.04045*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, et al. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Theresa Stadler, Bogdan Kulynych, Michael Gastpar, Nicolas Papernot, and Carmela Troncoso. The fundamental limits of least-privilege learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46393–46411, 2024.

Erik Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511666353.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.