On the Calibration of Conditional-Value-at-Risk

Rajeev Verma [1] **Volker Fischer** [2] **Eric Nalisnick** [1 3]

Abstract

To promote risk-averse behaviour in safety critical AI applications, Conditional-Value-at-Risk (CVaR)—a spectral risk measure—is largely being employed as a loss aggregation function of choice. We study the calibration and the refinement property of CVaR, by providing an extension of the classical proper scoring risk decomposition for CVaR. Our result suggests a trade-off: CVaR provides tail-sensitive calibration and refinement property, however this is at the cost of calibration and refinement for non-tail events. Our result calls to consider the inherent cost-benefit analysis to employ CVaR as a risk measure of choice for AI Safety.

1. Introduction

Machine learning algorithms are increasingly being used for risk prediction in consequential decision-making scenarios ranging from healthcare, public safety, cyber-security, finance, etc. These algorithms inform decision-makers about individual risk, and their predictions are then translated to specific actions based on some utility or cost structure. In such high-stakes applications with severe consequences to individuals, a recent emerging trend is to control for the "worst-case" errors.

In contrast to the traditional empirical risk minimisation, where the loss is aggregated using an expectation operator (or the empirical estimator of it), controlling for "worst-case" errors require using an alternate functional of the loss distribution, called *risk measures*, that are tailored to some specific tail-behaviour (Fröhlich & Williamson, 2023; Meng & Gower, 2023; Mehta et al., 2023), e.g. the conditional-value-at-risk (CVaR) (Serraino & Uryasev, 2013). Such risk measures have been used to promote risk-averse behaviour in critical applications (Curi et al., 2020; Levy et al., 2020; Wang & Zhou, 2023; Qin et al., 2023; Williamson & Menon, 2019). In critical applications where algorithms are used to drive decisions, calibration of the predictor is also very crucial. While calibration properties as induced by minimising an expected version of the loss are widely documented and studied, starting with the seminal work of DeGroot & Fienberg (1983), not much is known about the calibration properties induced as a result of alternate risk measures.

In this work, we study the calibration property of the predictor as a result of using CVaR as a loss functional. CVaR is a very versatile risk measure that interpolates from the expectation to the supremum, capturing the range of decision-making behaviours—from risk neutral to extremely risk averse. Our results state that CVaR results in tail-sensitive calibration and refinement behaviour, however this comes at the cost of performance for non-tail events.

2. Background

Notation Let $\mathcal{X} \times \mathcal{Y}$ be the space with some distribution P on it. We have access to the samples from this distribution with $\mathrm{x}_i \in \mathcal{X}$ denoting the input and y_i denotes the associated label, in the form of a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\mathrm{N}}$. The goal is to find a confidence predictor $g_\theta : \mathcal{X} \to \Delta^{|\mathcal{Y}|}$ for prediction tasks. This is usually accomplished via optimization of some loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+,\ \mathbf{x}, \mathrm{y}, g_\theta \mapsto \ell(\mathrm{y}, g_\theta(\mathbf{x}))$. Denoting $\mathrm{z} = \ell(\mathrm{y}, g_\theta(\mathbf{x}))$ as the incurred loss for some $(\mathbf{x}, \mathrm{y})$, we also have access to the realizations of the loss random variable z: $\{z_i\}_{i=1}^{\mathrm{N}}$. To optimise for the predictor g_θ, we employ a suitable aggregator function over the losses $\{z_i\}_{i=1}^{\mathrm{N}}$ to form a summary statistic. Traditionally, this is the average operator, leading to a vastly successful paradigm of empirical risk minimisation (ERM) as described below.

2.1. Empirical Risk Minimisation (ERM) and Calibration

In ERM, the summary statistic $\mathcal{R}_\theta(D)$ of the losses $\{z_i\}_{i=1}^{\mathrm{N}}$ is obtained using the average operator:

$$\hat{\mathcal{R}}_\theta(D) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} z_i = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \ell(y_i, g_\theta(\boldsymbol{x}_i)), \qquad (1)$$

and the resulting predictor is then obtained as $g_{\theta^*} \in \arg\min_{g \in \mathcal{H}} \hat{\mathcal{R}}_\theta(D)$. This is arguably the

[1] UvA-Bosch Delta Lab, University of Amsterdam, Netherlands [2] Bosch Center for AI, Rennigen, Germany [3] Johns Hopkins University, Baltimore, United States. Correspondence to: Rajeev Verma <rajeev.ee15@gmail.com>.

ICML 2024 Workshop on Next Generation of AI Safety. Copyright 2024 by the author(s).

foundational mechanism that underlies most of the success of machine learning systems. Due to the law of large numbers, $\hat{\mathcal{R}}_\theta\left(D\right)$ also forms an unbiased estimator for the expected population loss, also referred to as the population risk: $\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim P}\left[\ell\left(\mathrm{y}, g_\theta\left(\mathbf{x}\right)\right)\right]$, thereby lending this mechanism a specific safety certification—the resulting predictor g^*_θ will make small error on average across the population under consideration.

Besides this safety certification, decision-makers in consequential applications also care about the *confidence* of the predictor. The output of the associated function g is interpreted as the confidence of the predictor, and to meaningfully use it as a metric of reliability, decision-makers care about the *canonical calibration* of the predictor, as defined below:

Definition 2.1. *(Canonical Calibration).* Given d some divergence measure, e.g. squared error, a confidence predictor $g : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$ is said to be (perfectly) canonically calibrated if the following holds true:

$$\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim P}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g\left(\mathbf{x}\right)\,\right], g\left(\mathbf{x}\right)\right)\right] = 0. \tag{2}$$

Canonical calibration, thus, asserts that on average the confidence predictor means what it says, i.e. $\mathbb{E}\left[\mathrm{y}|g\left(\mathbf{x}\right)\right] = g\left(\mathbf{x}\right)$. When $\mathcal{Y} = \{0,1\}$, this translates that if the predictor outputs that the confidence in some event is α, among all the samples that have the same confidence α, the event will occur α times on average. By itself, it is a weak condition, and can be trivially satisfied, for example, by the average constant predictor $g\left(\mathbf{x}\right) = \mathbb{E}\left[\mathrm{y}\right]$, and there are infinitely many calibrated predictors (Vaicenavicius et al., 2019). Thus, the confidence predictor is also verified for its *refinement* as defined below:

Definition 2.2. *(Refinement error).* With d some divergence measure, the refinement error of a confidence predictor $g : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$ is defined as

$$\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim P}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g\left(\mathbf{x}\right)\,\right],\ \mathrm{y}\right)\right]. \tag{3}$$

Intuitively, refinement means that, on average, the confidence predictor is useful for predicting y, and low refinement error signals the *discriminativeness* of the confidence predictor.

A machine learning practitioner expects to obtain the confidence predictor with both the low calibration error and the low refinement error natively as a result of the expected risk minimisation framework. With the usual proper scoring loss functions (Gneiting & Raftery, 2007) utilised in the machine learning pipelines, this expectation is not unfounded as the population risk can be decomposed into the defined calibration and the refinement error terms (DeGroot & Fienberg, 1983; 1982; Kull & Flach, 2015), as below:

$$\begin{aligned}\mathbb{E}\left[d\left(\mathrm{y}, g\left(\mathbf{x}\right)\right)\right] &= \mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g\left(\mathbf{x}\right)\right],\ g\left(\mathbf{x}\right)\right)\right] \\ &\quad + \mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g\left(\mathbf{x}\right)\right],\ \mathrm{y}\right)\right].\end{aligned} \tag{4}$$

When ℓ is the mean squared error (MSE), the decomposition is widely known, and is stated below as an example:

$$\begin{aligned}\mathbb{E}\left[\left(\mathrm{y} - g\left(\mathbf{x}\right)\right)^2\right] = &\underbrace{\mathbb{E}\left[\left(\mathrm{y} - \mathbb{E}\left[\,\mathrm{y} \mid g\left(\mathbf{x}\right)\right]\right)^2\right]}_{\text{refinement error}} \\ &+ \underbrace{\mathbb{E}\left[\left(g\left(\mathbf{x}\right) - \mathbb{E}\left[\,\mathrm{y} \mid g\left(\mathbf{x}\right)\right]\right)^2\right]}_{\text{calibration error}},\end{aligned}$$

where the squared error is the divergence measure d, and the expectation is over $(\mathbf{x},\mathrm{y}) \sim P$. Thus, if one is to minimise a (strictly) proper scoring loss function, one expects the predictor to have low calibration error and low refinement error. Conversely, DeGroot & Fienberg (1983) show that if one is not calibrated, then one can get the risk to go down. The result is stronger: it says one can reduce the risk if and only if one is not calibrated. Thus, ERM (over the proper loss function) and calibration are intricately related.

Calibration of neural networks The connection between loss minimisation and calibration also applies to deep neural networks. Recently, Błasiok et al. (2023) showed that if the neural network's loss cannot be improved through post-processing by a simple class of smooth functions, then the network would be approximately-calibrated, and the argument goes in the reverse direction as well. The recent class of neural networks, they argue, natively satisfy this constraint, and hence they are relatively well-calibrated. This is in contrast to the common empirical wisdom that neural networks are mis-calibrated, and the usual approach has been to calibrate them post-hoc by methods like temperature scaling (Guo et al., 2017). The findings of Błasiok et al. (2023) also explain the inconsistent calibration behaviour of neural networks where they went from being empirically well-calibrated (Niculescu-Mizil & Caruana, 2005) to mis-calibrated (Guo et al., 2017), and then later to be relatively well-calibrated again (Minderer et al., 2021). The focus of this work is the native calibration property as a result of the risk minimisation, and following Błasiok et al. (2023) our results also apply to deep neural networks.

2.2. General Risk Measures and CVaR

While ERM provides some safety certifications in terms of minimum average error across the population, and the resulting notion of calibration and refinement, it is argued that marginal average guarantees are sometimes not enough in safety-critical applications, and could even mislead statistical conclusions, inter alia Simpson's paradox (Sprenger & Weinberger, 2021). Thus, in situations where the safety is the top-most priority, the emerging focus is to control the extreme errors.

Following Mehta et al. (2023), this is accomplished by the loss aggregator of the form

$$\hat{\mathcal{R}}^{\sigma}_{\theta}(D) = \sum_{i=1}^{\mathrm{N}} \sigma_i z_{(i)} = \sum_{i=1}^{\mathrm{N}} \sigma_i \ell_{(i)}\left(y_i, g_\theta(\boldsymbol{x}_i)\right), \quad (5)$$

where $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(\mathrm{N})}$ are the order statistics of the observed losses, and $0 \leq \sigma_1 \leq \ldots \leq \sigma_{\mathrm{N}} \leq 1$ is a sequence of non-decreasing weights such that $\sum_{i=1}^{\mathrm{N}} \sigma_i = 1$. The aggregator of the above form is generally known as the *L-estimator* (Shorack; Maurer et al., 2021), where σ_s are defined by the choice of a risk measure. Comparing this to Equation 1, it means suitably weighing the observed losses going from the uniform $\frac{1}{\mathrm{N}}$ for all σ_i as in the case of ERM, to $\sigma_{\mathrm{N}} = 1$ to only consider the worst observed loss.

CVaR The focus of this work is a specific case of risk measure, the CVaR. CVaR belongs to the class of spectral risk measures (Acerbi, 2002) that has gained special attention in the machine learning literature in recent years. Following the notation from before, z denotes the loss random variable, and assume it has some distribution function F_{z}, then a spectral risk measure is defined by a spectrum function $\sigma : [0,1] \to \mathbb{R}_+$ such that $\int_0^1 \sigma(q)\ dq = 1$ as below:

$$\mathcal{R}_\sigma\left[F_{\mathrm{z}}\right] = \int_0^1 F_{\mathrm{z}}^{-1}(q) \cdot \sigma(q)\ dq. \quad (6)$$

A monotonically increasing spectrum function, thus, models the *risk aversion* behaviour by putting high weight on the extreme values of the loss values z. For CVaR, defined by some $\alpha \in (0,1)$, the spectrum function is given as

$$\sigma(q) = \begin{cases} 0 & \text{for} \quad 0 < q < \alpha \\ \frac{1}{1-\alpha} & \text{for} \quad q \geq \alpha, \end{cases}$$

which means the values below the α quantile of the distribution function F_{z} are ignored, and loss values in the top $1-\alpha$ quantile receive the constant weight. When the involved distribution function F_{z} is continuous, CVaR has the equivalent definition as $\mathcal{R}_{\mathrm{CVaR}}[\mathrm{z}] = \mathbb{E}\left[\mathrm{z} \mid \mathrm{z} \geq F_{\mathrm{z}}^{-1}(\alpha)\right]$ for some α. It is appealing as it models a range of decision-making behaviours ranging from risk neutral ($\alpha = 0$) to the extreme risk averse behaviour ($\alpha = 1$), making it a suitable aggregator for the loss for safety-critical applications with some desired level of risk aversion.

While there is a growing body of work to replace the expectation operator with CVaR for loss aggregation to instill the predictive systems with the risk-averse behaviour for safety assurances, not much is known how does it affect the calibration properties of the resulting predictive system. While it is expected that this mechanism will control for the desired tail-sensitive errors, it is important to also study the consequences on the calibration and the resulting decision-making properties as in extremely safety-critical scenarios, the cost of the sub-optimal action would be severe.

3. Main contribution: Decomposition for CVaR

In order to understand the resulting calibration property as a result of using CVaR as the aggregator for the loss realizations, we take the theory-first approach to provide a guideline of the form of decomposition of the population risk for ERM (Equation 4). The resulting decomposition will inform what to expect when CVaR is employed as an aggregation operator in loss minimisation. For technical convenience, we assume the distribution function F_{z} is continuous, and hence the definition of CVaR is $\mathcal{R}_{\mathrm{CVaR}}[\mathrm{z}] = \mathbb{E}\left[\mathrm{z} \mid \mathrm{z} \geq F_{\mathrm{z}}^{-1}(\alpha)\right]$.

One way of achieving our goal is to realize that $\mathcal{R}_{\mathrm{CVaR}}[\mathrm{z}]$ is a tail-risk measure, and is fully-specified by the tail distribution of F_{z} beyond its α-quantile, defined as $F_{\mathrm{z}}^{\alpha} = \frac{(F_{\mathrm{z}}(z)-\alpha)_+}{1-\alpha}$, through the expectation operator, as $\mathcal{R}_{\mathrm{CVaR}}[\mathrm{z}] = \int_0^\infty z\ dF_{\mathrm{z}}^{\alpha}(z)$. This directly gives a tractable way to apply the already known decomposition results (Equation 4) in terms of the tail-distribution F_{z}^{α}, leading to the tail variants of the calibration and the refinement error. It is to be noted that the confidence predictor g is deterministic, and hence the distribution P over $\mathcal{X} \times \mathcal{Y}$ can be identifiably push-forwarded to define a distribution over z.

While this is informative, and leads to the conclusion that minimising CVaR gives corresponding tail-sensitivity to the calibration and the refinement error, we are interested in understanding the calibration (and the refinement) behaviour for the whole distribution. To achieve this, we use the *Rockafellar fundamental risk quadrangle* (Rockafellar & Uryasev, 2013) from risk management, and exploit the one-to-one connection between the risk measure and the associated deviation measure, and the connection between the deviation measure and the error measure. While the risk measure gives the aggregated summary of the random variable, a deviation measure (e.g. the standard deviation) quantifies the "non-constancy" of the said random variable, and the error measure (e.g. $\mathcal{L}^p$ norms) quantifies the "non-zeroness" of the random variable, and they are fundamentally related as stated below:

Theorem 3.1. *(Rockafellar & Uryasev (2013)). A risk measure $\mathcal{R}\left[F_{\mathrm{z}}\right]$ and the corresponding deviation measure $\mathcal{D}\left[F_{\mathrm{z}}\right]$ are related as $\mathcal{R}\left[F_{\mathrm{z}}\right] = \mathbb{E}\left[F_{\mathrm{z}}\right] + \mathcal{D}\left[F_{\mathrm{z}}\right]$[1]. And the deviation measure $\mathcal{D}\left[F_{\mathrm{z}}\right]$ can be elicited from the error measure $\mathcal{E}\left[F_{\mathrm{z}}\right]$ as $\mathcal{D}\left[F_{\mathrm{z}}\right] = \min_\kappa\{\mathcal{E}\left[\mathrm{z} - \kappa\right]\}$.*

The quantity $\kappa^* = \arg\min_\kappa\{\mathcal{E}\left[\mathrm{z} - \kappa\right]\}$ is referred to as the *statistic* associated with the risk measure. As an example, consider the $\mathcal{L}^2$ norm $||\mathrm{z}||_2 = \left[\mathbb{E}\left[|\mathrm{z}|^2\right]\right]^{1/2}$. It is easy to see that $\arg\min_\kappa ||\mathrm{z} - \kappa||_2 = \mathbb{E}\left[\mathrm{z}\right]$, or minimising

[1] We abuse the notation, and use $\mathbb{E}[\mathrm{z}]$ and $\mathbb{E}[F_{\mathrm{z}}]$ interchangeably. Similarly for others.

the expected squared error elicits the expectation. The resulting deviation measure then is the standard deviation, $\sigma(\mathrm{z}) = ||\mathrm{z} - \mathbb{E}[\mathrm{z}]||_2$. To summarise, a risk measure can be written in terms of the expectation and the deviation measure which, in turn, can be written in terms of the error measure defined in terms of the respective property it elicits.

For CVaR, the statistic κ^* is the the α quantile value, and defining $A = \{\, \mathrm{z} \mid \mathrm{z} \in (\kappa^*, \infty)\}$, and using the result from *Rockafellar fundamental risk quadrangle* gives the complete decomposition for CVaR, stated in Proposition 3.2 :

Proposition 3.2. *Denoting* $A = \{\, \mathrm{z} \mid \mathrm{z} \in (\kappa^*, \infty)\}$*, the calibration and refinement decomposition for CVaR is given as:*

$$
\begin{aligned}
\mathcal{R}_{CVaR}[\mathrm{z}] = {} & \underbrace{\mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g(\mathbf{x})\right],\ g(\mathbf{x})\right)\right]}_{\textit{average calibration error}} \\
& + \underbrace{\mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g(\mathbf{x})\right],\ \mathrm{y}\right)\right]}_{\textit{average refinement error}} \\
& + \frac{\alpha}{1-\alpha} \underbrace{\mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g(\mathbf{x})\right],\ g(\mathbf{x})\right) \mid \mathrm{z} \in A\,\right]}_{\textit{A conditional calibration error}} \\
& + \frac{\alpha}{1-\alpha} \underbrace{\mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g(\mathbf{x})\right],\ \mathrm{y}\right) \mid \mathrm{z} \in A\,\right]}_{\textit{A conditional refinement error}} \\
& - \underbrace{\mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g(\mathbf{x})\right],\ g(\mathbf{x})\right) \mid \mathrm{z} \in A^c\,\right]}_{A^c \textit{ conditional calibration error}} \\
& - \underbrace{\mathbb{E}\left[d\left(\mathbb{E}\left[\,\mathrm{y} \mid g(\mathbf{x})\right],\ \mathrm{y}\right) \mid \mathrm{z} \in A^c\,\right]}_{A^c \textit{ conditional refinement error}} + C(\kappa^*),
\end{aligned}
$$

where $\mathrm{z} = d(\mathrm{y}, g(\mathbf{x}))$*, and* d *denotes the divergence measure.* $C(\kappa^*)$ *is a constant.*

We provide the full derivation in Appendix B. The above decomposition agrees with the previous conclusion that minimising CVaR leads to tail-sensitive calibration and refinement error. However, it provides more information on the whole spectrum of the distribution of z. The crucial thing is the trade-off as stated in Corollary 3.3.

Corollary 3.3. *Define* $A = \{(\mathbf{x}, \mathrm{y}) \in \mathcal{X} \times \mathcal{Y} \mid \ell(\mathrm{y}, g(\mathbf{x})) \in (\kappa^*, \infty)\}$*, using* $\mathcal{R}_{CVaR}[\mathrm{z}]$ *improves calibration and refinement for* A *at the cost of calibration and refinement for* A^c*.*

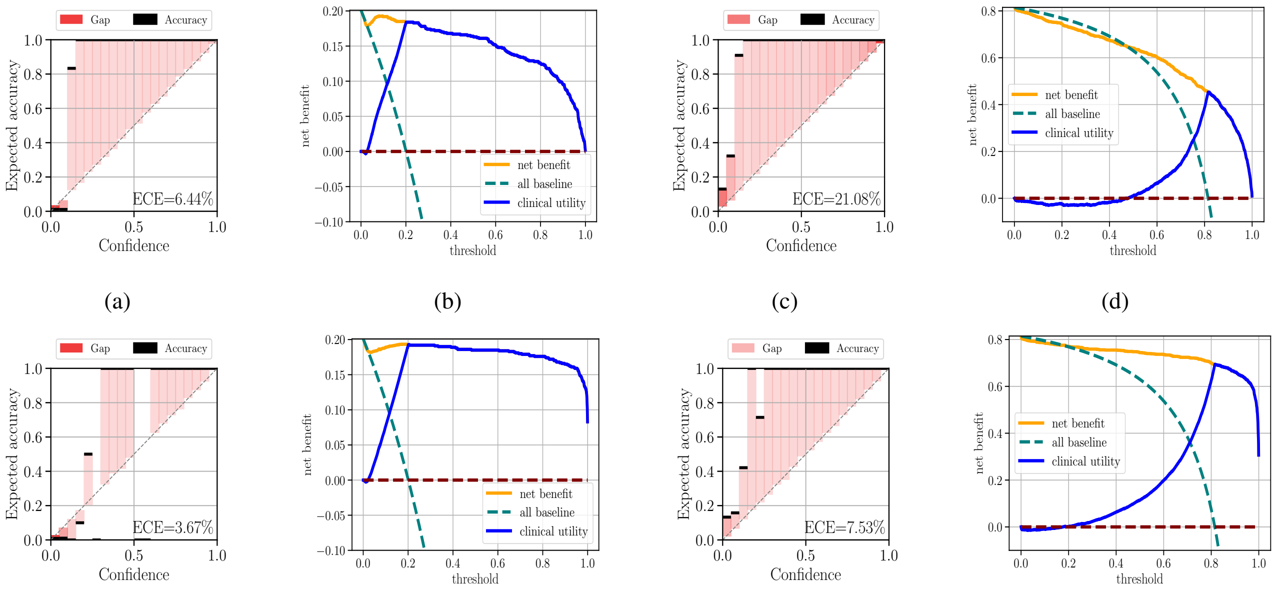
The above trade-off inherently calls to consider the cost-benefit analysis of employing CVaR as an aggregation mechanism. Interestingly, the average calibration and the refinement error would only get better.

4. Consequences: Fat Tails and Extreme Events

As artificially intelligent (AI) systems become powerful and their scope broadens in safety-critical applications, it is imperative to reconsider the unique challenges posed by such applications. One such challenge is the challenge of "fat tails" (Taleb, 2022). Many real-life critical applications conform to such distributional challenges: epidemiology, financial markets, cybersecurity, etc. The crucial challenge here is the the classical statistical wisdom does not apply to "fat tails," for example, the law of large numbers, an implicit convergence result that lends safety guarantees to the traditional ERM framework faces convergence issues (Taleb, 2022). Not incorporating such idiosyncrasies lead to fat-tailed reasoning errors. Tail-risk measures like CVaR are the emerging focus to reduce the impact of the *unknown unknowns* from the fat-tails. However, our result suggests that this does not apply broadly, and requires specific considerations.

Hallucination of Large Language Models (LLMs) LLMs, one of the most impressive technological advancements of our times, are already suffering from the fat-tailed reasoning error. McCoy et al. (2023) show that one way hallucination manifests in LLMs is substituting actually low-probability events for the erroneous high-probability events. They show that LLMs succeed on tasks that are of high-probability in nature, with significantly sub-par performances in low-probability situations. Scaling in terms of data and compute have been the common force behind the revolutionary performance of the LLMs, however, these alone cannot be guaranteed to overcome these issues. Language inherently follows a power law distribution (Chierichetti et al., 2017), and hence require novel methodological approaches to overcome the fat-tailed reasoning errors. One promising approach could be to employ risk measures like CVaR for tail-sensitive behaviour. However, our decomposition suggests that natively this might not be the best approach for a general purpose system such as LLMs, as providing tail-sensitivity would come at the cost of non-tail performance. However, LLMs could benefit from tail-sensitive finetuning for specific safety-critical applications.

Extremile Calibration Calibrated predictions directly translate to (optimal) decision-making through the expected utility framework. (Refer to Appendix D to learn more) However, algorithms cannot be assumed to be perfectly calibrated, and hence, the interesting question is of miscalibration allocation: where it can be tolerated, and where not. Since calibration gets its usability through the associated decision-making framework, we argue that miscalibration gets undesirable when the consequences of wrong decisions is very high. In other words, for extreme events with consequential outcomes, mis-calibration is undesirable. Such is the case for *unknown unknowns*, and they will be of critical consequences in many applications of practical importance. However, the average notion of canonical calibration is not informative, and have a disparate impact across different sub-groups. Hebert-Johnson et al. (2018) argue to assert calibration in every computable sub-group, however, it is not clear how to define the reference class of *unknown unknowns*. We argue that CVaR decomposition



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1. Calibration and resulting decision-making behaviour for ERM (first row) and CVaR (second row) for a simulation on specific extreme events scenario: Blue line going below origin denotes clinical harm; we refer the reader to Appendix E to learn more. The ERM prediction system seems clinically helpful overall (Subfigures 1a and 1b), however it is severely mis-calibrated and actually clinically harmful for a wide range of utility functions on the rare disease population—a population where correct actions carry the most impact (Subfigures 1c and 1d). The bottom row show that CVaR overcomes the clinical harm on this diseased population (Subfigure 1h).

gives a tractable way to solve the reference class problem (Hájek, 2007) for *unknown unknowns* in terms of the involved predictor's loss, and to certify calibration there directly through the loss minimisation principle. We call this notion of calibration for extreme events as *extremile calibration*, and argue for its applicability by drawing connections to Fisher–Tippett–Gnedenko theorem (Basrak, 2011) in extreme value theory (EVT) (de Haan & Ferreira, 2006) in Appendix C (Corollary C.4). We devise a simulation to validate that learning with CVaR results in *extremile calibration* characteristics, and also improve decision-making for the extreme but significant events.

Simulation We consider an evolving rare disease affects a significantly small portion of the population, but could pose severe concerns to the whole population if went uncontrolled. And due to the evolving nature, it is a case of *unknown unknowns*. We refer the reader to Appendix E for the simulation setup. A medical facility utilising some risk prediction model cannot afford wrong decision-making for the diseased sub-group, else it would pose severe hazard. To aid decision-making, the facility is concerned about the calibration of the prediction model, and its role in optimal decision making for different utility functions. Using metrics like net benefit and clinical utility (Vickers et al., 2016), the medical facility concludes that the model is clinically helpful (Subfigures 1a and 1b). However, the model is actually clinically harmful for a wide range of utility functions for the rare disease sub-population (Subfigures 1c and 1d). Thus, employing this model is hazardous to the population where it matters the most, and the resulting wrong decisions have high consequences. Due to the case of extreme events, it is also computationally difficult to assert calibration in a post-hoc manner, and also due to the reference class problem (Hájek, 2007). Subfigures 1g and 1h show that minimising CVaR asserts calibration on this sub-population, and also overcomes the associated clinical harm. It also improves the average calibration on the whole population.

5. Conclusions and Future Work

We extend the popular risk decomposition for proper scoring losses into the calibration and refinement for the CVaR risk measure—a popular measure to control "worst-case" errors for AI safety applications. Our result states the trade-off, thereby calling to consider involved risk management cost-benefit analysis for its usability. We also draw connections to fat-tails and calibration for extreme events: two related concepts that thwart the safety of AI systems in real-world. While tail-risk measures like CVaR are promising, but there are considerations involved. We further argue that fat-tails pose significant challenges to the applicability of AI systems in real-world applications, and thus there is a need to adopt more robust and theoretically sound methodology to mitigate risks associated with the rare but high-impact events.

6. Acknowledgements

We thank Alexander Timans, Teodora Pandeva, Mona Schirmer, Rabanus Derr, Christian Fröhlich, and Metod Jazbec for helpful discussions. UvA-Bosch Delta Lab at the University of Amsterdam is generously supported by the Bosch Center for Artificial Intelligence.

References

Acerbi, C. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking Finance*, 2002.

Basrak, B. *Fisher-Tippett Theorem*. 2011.

Błasiok, J., Gopalan, P., Hu, L., and Nakkiran, P. When does optimizing a proper loss yield calibration? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Calster, B. V. and Vickers, A. J. Calibration of risk prediction models: Impact on decision-analytic performance. *Medical Decision Making*, 2015.

Chierichetti, F., Kumar, R., and Pang, B. On the power laws of language: Word frequency distributions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

Curi, S., Levy, K. Y., Jegelka, S., and Krause, A. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems*, 2020.

de Haan, L. and Ferreira, A. Extreme value theory : an introduction. 2006.

DeGroot, M. H. and Fienberg, S. E. Assessing probability assessors: calibration and refinement. *Statistical decision theory and related topics III*, 1982.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1983.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

Fisher, R. A. and Tippett, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1928.

Fröhlich, C. and Williamson, R. Tailoring to the tails: Risk measures for fine-grained tail sensitivity. *Transactions on Machine Learning Research*, 2023.

Gnedenko, B. V. *On the Limiting Distribution of the Maximum Term in a Random Series*. 1992.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

Grünwald, P. Safe probability, 2016.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Hájek, A. The reference class problem is your problem too. *Synthese*, 2007.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Kull, M. and Flach, P. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2015.

Laguel, Y., Pillutla, K., Malick, J., and Harchaoui, Z. A Superquantile Approach to Federated Learning with Heterogeneous Devices. In *55th Annual Conference on Information Sciences and Systems, CISS*, 2021.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020.

Maurer, A., Parletta, D. A., Paudice, A., and Pontil, M. Robust unsupervised learning via l-statistic minimization. In *International Conference on Machine Learning*, 2021.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T. L. Embers of autoregression: Understanding large language models through the problem they are trained to solve. 2023.

Mehta, R., Roulet, V., Pillutla, K., Liu, L., and Harchaoui, Z. Stochastic optimization for spectral risk measures. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Meng, S. Y. and Gower, R. M. A model-based method for minimizing CVaR and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F. A., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 2005.

Noarov, G. and Roth, A. Calibration for decision making: A principled approach to trustworthy ml, 2024. URL https://www.let-all.com/blog/2024/03/13/calibration-for-decision-making-a-principled-approach-to-trustworthy-ml/. Blog post.

Qin, Y., van der Schaar, M., and Lee, C. Risk-averse active sensing for timely outcome prediction under cost pressure. In *Advances in Neural Information Processing Systems*, 2023.

Rockafellar, R. T. and Uryasev, S. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 2013.

Serraino, G. and Uryasev, S. *Conditional Value-at-Risk (CVaR)*. 2013.

Shorack, G. R. *Probability for statisticians*. Springer.

Sprenger, J. and Weinberger, N. Simpson's Paradox. In *The Stanford Encyclopedia of Philosophy*. 2021.

Taleb, N. N. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications, 2022.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Vickers, A. J., calster, B. V., and Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *The BMJ*, 2016.

Wang, Y. and Zhou, E. Bayesian risk-averse q-learning with streaming observations. In *Advances in Neural Information Processing Systems*, 2023.

Williamson, R. C. and Menon, A. K. Fairness risk measures. In *International Conference on Machine Learning*, 2019.

Zhao, S., Kim, M., Sahoo, R., Ma, T., and Ermon, S. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, 2021.

A. Related Work

Employing CVaR for loss aggregation is an emerging focus to instill the AI systems with safety considerations (Curi et al., 2020; Levy et al., 2020; Wang & Zhou, 2023; Qin et al., 2023), for fairness (Williamson & Menon, 2019). Mehta et al. (2023) and Meng & Gower (2023) propose gradient-based optimisation algoritms to minimise CVaR. However, none has examined the resulting calibration and refinement properties. Our work provides insights to the practitioners of the involved considerations.

B. Complete Characterization and Derivation

We follow from the *Rockafellar fundamental risk quadrangle* (Rockafellar & Uryasev, 2013) to express $\mathcal{R}_{\text{CVaR}}\left[\mathrm{z}\right]$ as $\mathcal{R}_{\text{CVaR}}\left[\mathrm{z}\right] = \mathbb{E}\left[\mathrm{z}\right] + \mathcal{D}\left[\mathrm{z}\right]$, where $\mathcal{D}\left[\mathrm{z}\right]$ is the associated deviation measure. Denoting $\mathbf{z}_+ = \max\{0, \mathbf{z}\}$ and $\mathbf{z}_- = \max\{0, -\mathbf{z}\}$, the error measure for CVaR is $\mathcal{E}\left[\mathbf{z}\right] = \mathbb{E}\left[\frac{\alpha}{1-\alpha}\mathbf{z}_+ + \mathbf{z}_-\right]$. As stated in the main text, the *statistic* κ^* associated with CVaR is the α-quantile value, where α is pre-specified. Following Theorem 3.1, the deviation measure $\mathcal{D}\left[\mathrm{z}\right]$ can be written as $\mathcal{D}\left[\mathrm{z}\right] = \mathbb{E}\left[\frac{\alpha}{1-\alpha}\left(\mathbf{z} - \kappa^*\right)_+ + \left(\mathbf{z} - \kappa^*\right)_-\right]$. Thus,

$$
\begin{aligned}
\mathcal{R}_{\text{CVaR}}\left[\mathrm{z}\right] =\ & \mathbb{E}\left[\mathrm{z}\right] + \mathbb{E}\left[\frac{\alpha}{1-\alpha}\left(\mathbf{z} - \kappa^*\right)_+ + \left(\mathbf{z} - \kappa^*\right)_-\right] \\
=\ & \mathbb{E}\left[\mathrm{z}\right] + \mathbb{E}\left[\frac{\alpha}{1-\alpha}\max\{0, \mathbf{z} - \kappa^*\} + \max 0, \kappa^* - \mathbf{z}\right] \\
=\ & \mathbb{E}\left[\mathrm{z}\right] + \frac{\alpha}{1-\alpha}\mathbb{E}\left[\,\mathbb{I}_{(\kappa^*,\infty)}\left[\mathbf{z}\right]\left(\mathbf{z} - \kappa^*\right)\right] + \mathbb{E}\left[\,\mathbb{I}_{\,(-\infty,\kappa^*)}\left[\mathbf{z}\right]\left(\kappa^* - \mathbf{z}\right)\right] \\
=\ & \mathbb{E}\left[\mathrm{z}\right] + \frac{\alpha}{1-\alpha}\mathbb{E}\left[\,\left(\mathbf{z} - \kappa^*\right)\ \mid \mathbf{z} \in \left(\kappa^*, \infty\right)\right] + \mathbb{E}\left[\,\left(\kappa^* - \mathbf{z}\right)\ \mid \mathbf{z} \in \left(-\infty, \kappa^*\right)\right].
\end{aligned}
$$

Denoting $A = \{\mathbf{z} \mid \mathbf{z} \in \left(\kappa^*, \infty\right)\}$, the $\mathcal{R}_{\text{CVaR}}\left[\mathbf{z}\right]$ is expanded into three terms: one overall $\mathbb{E}\left[\mathbf{z}\right]$, A conditional term and the A^c conditional one. We already know the calibration and refinement decomposition result for $\mathbb{E}\left[\mathbf{z}\right]$. We further use the linearity of expectation to use the conditional version of the same result on the other two terms, resulting the final decomposition in Proposition 3.2.

C. Connections to Extreme Value Theory

Extreme Value Theory (EVT) (de Haan & Ferreira, 2006) concerns with the asymptotic distribution of $\max\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ as $N \to \infty$ where each $\mathbf{x}_i$ is sampled in an i.i.d. fashion from some distribution. The theory is inspired from the classical central limit theorem (CLT) that gives a tractable asymptotic behaviour for the $\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_N$ in terms of the standard normal distribution. Similar to the CLT where the asymptotics are given in terms of the normal distributions, EVT define the asymptotic distribution for the extreme in terms of the *extreme value distributions*

Theorem C.1. *(Extreme Value Distribution). (de Haan & Ferreira, 2006; Fisher & Tippett, 1928; Gnedenko, 1992). The class of extreme value distribution functions is* $G_\gamma\left(\mathbf{x}\right)\left(a\mathbf{x} + b\right)$ *with* $a > 0, b \in \mathbb{R}$ *defined as*

$$
G_\gamma\left(\mathbf{x}\right) = \exp\{-\left(\,1 + \gamma\mathbf{x}\right)^{-1/\gamma}\},\ 1 + \gamma\mathbf{x} > 0,
$$

where $\gamma \in \mathbb{R}$*, is called as the extreme value index. For* $\gamma = 0$*, the* $G_\gamma\left(\mathbf{x}\right)$ *is taken as* $\exp\{-e^{-\mathbf{x}}\}$.

Elucidating on EVT in not our focus, however, EVT states that the extreme value(s) in a sample would asymptotically follow some form of the extreme value distribution whose nature will be defined by the extreme value index γ. EVT is attractive as it allows one to use this result to fit models to extrapolate for the nature of extreme events for safety assurances. One can further characterise different distributions depending on the value of the extreme value index γ, as below:

Proposition C.2. *(de Haan & Ferreira, 2006). The extreme value index* γ *can be used to characterise different class of distributions:*

1. $\gamma > 0$*:* $G_\gamma\left(\mathbf{x}\right) < 1$ *for all* $\mathbf{x}$*, i.e. the distribution extends to infinity. Also* $\mathbf{x} \to \infty$ *implies* $1 - G_\gamma\left(\mathbf{x}\right) \sim \gamma^{-1/\gamma}\mathbf{x}^{-1/\gamma}$*, i.e. the distribution has a rather heavy right tail; or the moments of order greater than or equal to* $1/\gamma$ *do not exist.*

2. $\gamma = 0$*: the distribution again extends to infinity. However, the distribution is light-tailed as* $1 - G_0(\mathbf{x}) \sim e^{-\mathbf{x}}$ *as* $\mathbf{x} \to \infty$*. And all moments exist.*

3. $\gamma > 0$*: the right endpoint of the distribution is* $-1/\gamma$*, and hence a short tail.*

Thus, in terms of the *unknown unkowns*, distributions that approach extreme value distributions with the extreme value index in $\gamma \in [0, \infty)$ pose challenges with further challenges posed by $\gamma \in (0, \infty)$ due to the heavy-tailed nature. Due to the moments not existing beyond $1/\gamma$ further poses significant challenges to the standard statistical machinery, and Taleb (2022) refers to this class of distributions as the *Extremistan*. This is also the class that carry the maximum impact, and covers many crucial safety-critical applications like finance and wealth, epidemiology, natural disasters and environmental risks. In terms of risk and mis-calibration allocation, wrong or sub-optimal decisions here carry the severe consequences. It turns out employing tail-risk measures like CVaR can natively consider this allocation. This directly follows from the convergence result from Mehta et al. (2023), restated below for completion:

Proposition C.3. *(Mehta et al., 2023). Given* $z_1, z_2, \ldots, z_N$ *realizations from the associated loss distribution with distribution function* F_z*, then the empirical version of the spectral risk measure* $\hat{\mathcal{R}}^\sigma[D] = \sum_{i=1}^{N} \sigma_i \cdot z_{(i)}$ *for the spectrum function* $\sigma(q), q \in (0,1)$*, and its population counterpart* $\mathcal{R}^\sigma[F_z]$*, and assume that* $\mathbb{E}|\mathrm{z}|^p < \infty$ *for some* $p > 2$*, and* $||\sigma||_\infty = \sup_{\sigma \in (0,1)} |\sigma(q)| < \infty$*, then the following convergence holds true:*

$$\mathbb{E}|\hat{\mathcal{R}}^\sigma[D] - \mathcal{R}^\sigma[F_\mathrm{z}]|^2 \leq \frac{2||\sigma||_\infty^2 \left(\frac{p}{p-2}\right)^2 \mathbb{E}[|\mathrm{z}|^p]^{\frac{2}{p}}}{N}.$$

The above convergence result states that, under the assumption of "tame" tails: $\mathbb{E}|\mathrm{z}|^p < \infty$ for some $p > 2$ and the finiteness of the assumed spectrum function (in terms of the supremum norm), the finite sample version of the risk estimator and the population version converge as $N \to \infty$. The next corollary establishes the notion of "tame" tails with the extreme value distribution for a range of *extremistan* distributions.

Corollary C.4. *L-estimators of the form in Equation 5 with risk measures with finite spectrum function in terms of the supremum norm covers a sub-class of heavy-tailed distributions with extreme value index* $\gamma \in [0, 0.5)$.

Thus, tail-risk measures like CVaR form a suitable aggregators for risk managment for some class of *extremistan* problems. While CVaR is motivated to control "worst-case" errors in the machine learning literature, the above result strengthens their use for some mild cases of heavy-tailed critical applications. The proposed risk management further lends to provide calibration allocation to control for sub-optimal actions in such applications. We are not aware of this discussion appearing in other contemporary machine learning works. In the future, it is interesting to use results from the EVT to further expand the range of covered extreme tail index γ.

D. Calibration and Decision Making

In this section, we elaborate on the connection between calibration and decision-making. Following the notation from before, we also have some (finite) action space $\mathcal{A}$ and a utility function $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}_+$ that assigns the utility for the action $a \in \mathcal{A}$ when the outcome is y. Under the expected utility framework, the optimal action for $\mathbf{x} = \boldsymbol{x}$ is $a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\mathrm{y} \sim p^*(\boldsymbol{x})}[u(a, \mathrm{y})]$, i.e. the action that maximizes the expected utility under $p^*(\boldsymbol{x}) = P(\mathrm{y} \mid \mathbf{x} = \boldsymbol{x})$. Thus, to be able to take the optimal action, one needs access to $p^*(\boldsymbol{x})$. However, it is generally unavailable due to computational and statistical estimation issues. The question is that if one can use the calibrated predictor $g(\mathbf{x})$, instead, for decision-making. It is now widely (Dwork et al., 2021; Zhao et al., 2021; Grünwald, 2016) known that one certainly can for any arbitrary utility function. For completion, we state the result from Noarov & Roth (2024):

Proposition D.1. *(Calibration and Decision-making). (Noarov & Roth, 2024). Assume* $g(\mathbf{x})$ *is canonically calibrated. Then for any agent with some utility function* u*, the expected utility decision policy* $\hat{a} = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\mathrm{y} \sim g(\mathbf{x})}[u(\mathrm{y}, a)]$*, is the optimal policy among all the decision policies from predictions to actions.*

Thus, having a calibrated predictor is desirable as one can employ it to make correct actions, when $p^*(\boldsymbol{x})$ is not available.

E. Simulation Setup

We describe the setup for the simulation study in the main text.

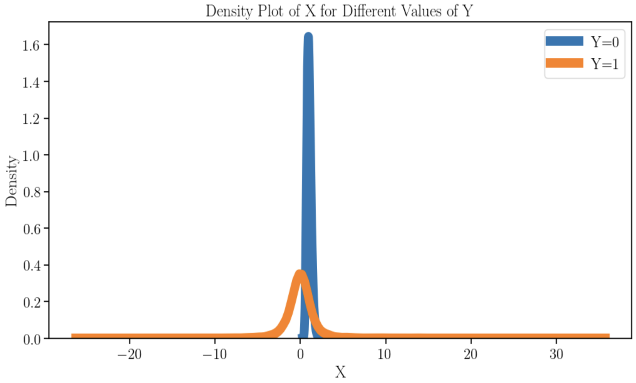


Figure 2. Density plot for the simulated data for the simulation study. $Y = 1$ denotes diseased sub-population and $Y = 0$ denotes the non-diseased sub-population.

Data In order to have control over the tails, we chose to work with the synthetic data. We consider the setup for a rare evolving disease that is currently prevalent in a small portion of the population. We set the base rate of this disease to be 20%. We sample features for the diseased sub-population from the standard t-distribution with 3 degrees of freedom. For the healthy sub-population, the features are drawn from the log-normal distribution with $\mu = 0$ and $\sigma = 0.25$. We draw 5000 samples in total, with the density plot shown in Figure 2. Since the diseased sub-population forms a case of the *unknown unknowns*, the density for the diseased population features $Y = 1$ has richer tail behaviour. We split this simulated dataset into $80 - 20\%$ training-test split. We verify the marginal properties of the model on the test split of the dataset. We further sample from the diseased population distribution to verify the sub-population property on it.

Model We fit a regular Logistic regression model to this dataset. The resulting predictor has the AUC-ROC score of 0.90 which denotes good discriminativeness property. For fitting CVaR, we use $\alpha = 0.30$ and use the off-the-shelf optimisation framework: sqwash[2] (Laguel et al., 2021), and the resulting predictor has the AUC-ROC score of 0.95. Thus, by accounting for the extreme errors, the CVaR minimisation improves the discriminativeness of the predictor as well as the resulting calibration (as shown in Subfigures 1a and 1e).

Metrics While AUC-ROC informs the decision-makers about the discriminativeness of the resulting predictor, it does not consider the connection with the involved utility function a decision-maker has in mind, and the consequences involved. A medical facility could be working with a range of utility functions, hence, the usability of the risk predictor is studied with respect to the clinical utility (Vickers et al., 2016) for a range of utility functions. When the action space $\mathcal{A}$ is binary, the decision-making behaviour reduces to thresholding the predicted risk based on some threshold that depends on the utility structure. Since utility functions are arbitrary, the clinical utility of a predictor is determined by its usability for all the thresholds $t \in [0, 1]$—the risk prediction model is clinically helpful if it has positive clinical utility for all the decision-making utility considerations. In Subfigures 1b, 1d, 1f, 1h, the x-axis denotes different thresholds that correspond to different utility function structure. Roughly, the lower the threshold, the more risk-averse the decision-maker is, and vice-versa. *Net Benefit* (Vickers et al., 2016) is a metric that considers the consequences of using a certain risk prediction algorithm in real-world. The metric inherently considers how many false positives a decision-maker is willing to tolerate to reach one true positive, with a more risk-averse individual willing to handle more false-positives for one true positive. For a threshold $t \in [0, 1]$, the net benefit is defined as

$$\text{Net Benefit} = \frac{\text{True Positives}}{N} - \frac{\text{False Positives}}{N} \cdot \frac{t}{1-t}.$$

An even more risk-averse individual can choose to treat (or take favourable action) for all the individuals, and this is referred to as the *all baseline*. A risk prediction algorithm is referred to as useful if the net benefit of using it exceed the all baseline, with clinical utility for a threshold t defined as:

$$\text{Clinical utility(t)} = \text{Net Benefit(t)} - \text{Net Benefit(All Baseline)}.$$

[2] https://github.com/krishnap25/sqwash

A model is called clinically helpful if the clinically utility is positive for all thresholds t, which means the model can safely be used in the real-world for a range of arbitrary utility functions. We refer the reader to (Vickers et al., 2016) to learn more. It is also shown that mis-calibration affects the clinical utility of the risk prediction algorithm (Calster & Vickers, 2015), which our simulation also confirms.