High-Confidence Data Programming for Evaluating Suppression of Physiological Alarms

Sydney Pugh*, Ivan Ruchkin*, Christopher P. Bonafide†, Sara B. DeMauro†,
Oleg Sokolsky*, Insup Lee*, and James Weimer*
*Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104
†Children's Hospital of Philadelphia, Philadelphia, PA 19104

***Abstract*—False alarms generated by physiological monitors can overwhelm clinical caretakers with a variety of alarms. The resulting alarm fatigue can be mitigated with alarm suppression. Before being deployed, such suppression mechanisms need to be evaluated through a costly observational study, which would determine and label the truly suppressible alarms. This paper proposes a lightweight method for evaluating alarm suppression without access to the true alarm labels. The method is based on the data programming paradigm, which combines noisy and cheap-to-obtain labeling heuristics into probabilistic labels. Based on these labels, the method estimates the sensitivity/specificity of a suppression mechanism and describes the likely outcomes of an observational study in the form of confidence bounds. We evaluate the proposed method in a case study of low SpO_2 alarms using a dataset collected at Children's Hospital of Philadelphia and show that our method provides tight and accurate bounds that significantly outperform the naive comparative method.**

I. Introduction

Alarm fatigue is a pervasive problem associated with physiologic monitoring in the hospital setting [1]. Bedside monitors continuously measuring heart rhythm, heart rate, respiratory rate, blood oxygen, and other parameters often overwhelm clinicians with very frequent non-actionable alarms. The end result is that clinicians react slowly, if at all, to alarms that have a small but nonzero probability of representing a critical patient need [2]. Ideally, the clinicians should only be alerted by the alarms that they will find informative or actionable (we call such alarms *non-suppressible*), whereas the rest of the alarms are deemed *suppressible*.

Alarm fatigue can be mitigated by reducing the number of suppressible alarms through threshold tuning, customization, integration, and other methods [3]. Researchers have proposed novel algorithms for monitoring and suppressing unnecessary alarms based on advanced data processing [4], [5], [6]. Such improvements need to be carefully balanced with the possibility of missing non-suppressible alarms. Ultimately, algorithmic methods and tuning can be seen as a *suppression system* targeting a particular type of alarm.

The clinical investigation and deployment of suppression systems is predicated on their expected performance. For example, when deploying an alarm suppression system, hospital policy makers need to confirm that its specificity to non-suppressible alarms is above certain bounds, to be confident that most non-suppressible alarms will continue to be reported. Measuring the performance of a suppression system typically requires a representative dataset of alarms labeled with their suppressibility.

It is time-consuming and expensive to create highly accurate labeled datasets for evaluation and tuning of suppression systems. A common way to do so is to perform an observational study [7] of many patients and manually label each time when an alarm would be non-suppressible. Such a study is a major commitment when it comes to an initial deployment of a novel suppression system, in part due to the significant effort of manual labeling. Furthermore, it is impractical to perform an observational study for every adjustment of the settings of a physiological monitor throughout its lifecycle. This cost can be reduced with patient simulations [8], but precise and realistic simulations of human physiology are notoriously difficult and expensive to construct.

This paper introduces a cheap and rapid method of estimating the performance of a suppression system *in the absence of a dataset with highly accurate labels*. This method can support early-stage low-cost investigations of suppression systems in a variety of ways. For example, it can prioritize observational studies of systems with higher potential to alleviate alarm fatigue so that the effort of manually labeling is spent optimally. It can also guide the tuning of the system's settings towards effective alarm suppression, reducing the risk of missing non-suppressible alarms.

A key element of our method is to probabilistically label patient data according to the recently emerging paradigm of *data programming* [9]. We start with a dataset of unlabeled patient data, typically abundant in most clinical settings, and a suppression system for some alarm type with tunable settings. We then collect clinical intuitions about this alarm type and encode them as *labeling functions* — weak classifiers of suppressible/non-suppressible alarms that can abstain and need not be comprehensive or non-contradictory. The data and the labeling functions are put together via a *generative model*, resulting in labels of varied confidence for each data point. Finally, the high-confidence subset of those labels is used to estimate the sensitivity and specificity of the suppression system.

To appropriately communicate the uncertainty of our estimates, we mathematically develop *confidence bounds* on the sensitivity and specificity on the suppression system. These bounds indicate, for a given level of confidence, the interval of possible sensitivity/specificity values that one could obtain

This work was supported in part by NSF-1915398 and NIH R18 HS026620.

978-1-6654-3965-7/21/$31.00 ©2021 IEEE
DOI 10.1109/CHASE52844.2021.00015

if they performed an observational study of a given size. These bounds account for the uncertainty of the labeling process, the randomness in the sampling of the dataset, and the amount of data available for different clinical situations.

We validated our method through a case study of low SpO_2 alarms on a 551-hour labeled dataset from Children's Hospital of Philadelphia. The proposed method was used to estimate the alarm performance for different values of the SpO_2 threshold. Our method's estimated confidence bounds almost always contain the true-label-based specificity and sensitivity — and substantially outperform the naive estimates based on each labeling function voting with an equal weight. This study demonstrated how to negotiate the sensitivity-specificity trade-off in an suppression system without investing hundreds of hours into labeling the alarm data.

In summary, this paper makes three research contributions:

- A data programming-based method for estimating the performance of alarm suppression,
- Confidence bounds on the performance estimates from the above method,
- A successful application of the above method to a case study of tuning the SpO_2 alarm threshold.

The rest of the paper is organized as follows. The next section presents the detailed motivation for low-cost estimation of alarm suppression. Section III discusses the existing ways to evaluate suppression systems. Section IV formulates the mathematical problem at the heart of our method, which is described in the following section. The case study of low SpO_2 alarms is described in Section VI, and its results are found in Section VII. The paper concludes with Section VIII.

II. Motivating Scenarios

This paper focuses on clinical alarms produced by physiological monitors in a hospital setting. A monitor takes in a combination of static inputs (*e.g.,* demographic information) and dynamic inputs (*e.g.,* 5 seconds of vital sign waveforms), and we refer to their combination as *patient data*. The *alarm-generating device* implements an algorithm that responds to patient data by either raising an alarm or not. All the inputs where the device raises an alarm are referred to as *alarm-generating inputs*, or *alarms* for short. The scope focuses on *alarm suppression systems* that deactivate the raised alarms; thus, the inputs that did *not* trigger any alarms in the first place are not considered because their relevance is nearly impossible to establish in most practical settings. Our concept of a suppression system describes both standalone algorithms deployed alongside alarm devices and any adjustments to the existing alarm device (*e.g.,* reducing the SpO_2 alarm threshold).

Suppose that some patient data is measured by or input into an alarm device, and it generates a number of alarms. Any such alarm belongs to one of the two mutually exclusive classes. A *suppressible* alarm is one that can be disregarded by the clinicians without missing any important or actionable information. The overabundance of suppressible alarms is both a cause of alarm fatigue and an opportunity for alarm suppression. A *non-suppressible* alarm is one that communicates valuable information to the clinicians and should not be missed, regardless of whether it is immediately actionable. Thus, when addressing alarm fatigue, policy makers need to carefully balance the risk versus reward of silencing suppressible alarms and missing non-suppressible alarms.

Alarm suppression systems are typically evaluated in the context of a labeled alarm dataset. The two key performance characteristics of alarm suppression are

- The *sensitivity of alarm suppression*: the proportion of suppressible alarms that were suppressed, also known as the false alarm suppression rate.
- The *specificity of alarm suppression:* the proportion of the non-suppressible alarms that were preserved (not suppressed), which can also be calculated as one minus the true alarm suppression rate.

Let us consider two clinical scenarios that motivate the problem addressed in this paper.

Scenario 1: Pre-Trial Evaluation of Suppression System. Hospital A, serving population P, is considering the deployment of an alarm suppression system that has been successful at Hospital B, which serves population Q. The system's settings can be transferred between the hospitals; however, the patient data from population Q cannot be shared. Consequently, it is unknown how well the system would perform on population P, and it is likely to require alarm suppression. Before embarking on a time-consuming and expensive clinical trial of that system, Hospital A wants to estimate whether it can plausibly deliver a sizeable fraction of the useful, non-suppressible alarms while not significantly contributing to the alarm fatigue. Hospital A executes the system on the representative patient data from population P. Ideally, Hospital A would need to label the produced alarms, but it is prohibitively expensive to construct these labels manually. In other words, Hospital A seeks to *estimate the performance* of a suppression system given *unlabeled alarms*.

Scenario 2: Tuning of Deployed Suppression System. A hospital uses an alarm device in its ICU. The device's operation is configured with several tunable settings such as the acceptable interval for each vital sign and the minimum time spent outside of that interval to trigger the alarm. Due to reports of alarm fatigue, the hospital considers a manual adjustment of the device settings to reduce suppressible alarms. However, there is a serious risk of missing important, non-suppressible alarms as a result. To proceed with this adjustment, the hospital needs to estimate the effects of various settings: what fraction of the suppressible alarms would be silenced and what fraction of the non-suppressible alarms will continue to be raised? The hospital has abundant patient data but insufficient resources to construct a representative labeled dataset of alarms. In this situation, the hospital aims to predict the *effects of a device's settings* on its *performance* given *only patient data* and no precise information whether an alarm is suppressible.

In both scenarios, the task is to evaluate a suppression system, as illustrated in Figure 1. We are given an alarm device and can collect a dataset of representative unlabeled alarms. In an ideal situation, this data collection would be

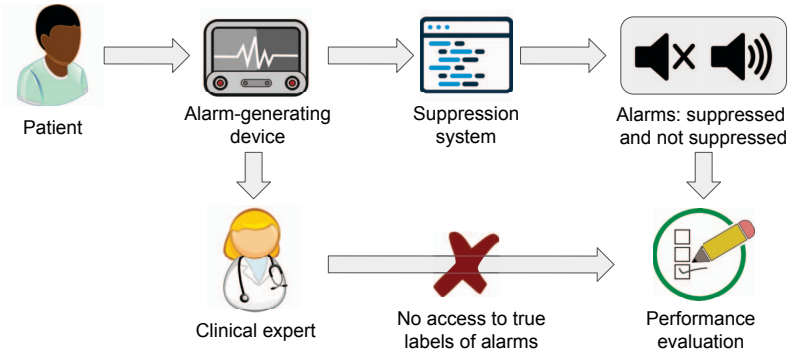


Fig. 1: A motivating scenario for this paper: a suppression system needs to be evaluated without true labels of alarms.

accompanied by patient observation, and each alarm would be annotated with a suppressible/non-suppressible label based on the clinical interpretation of the patient's circumstances. As the next section details, in practice, creating such annotations is prohibitively time-consuming and expensive. Thus, there is a need for lightweight means of predicting the performance of alarm suppression.

The next section describes the existing ways of addressing the motivating scenarios.

III. Related Work

There exists a vast literature on clinical alarm suppression and unsupervised/weakly-supervised learning. We focus on the areas particularly relevant to our setting.

Alarm fatigue is a serious and well-known problem of physiological monitors [1], [2]. A variety of approaches to suppress unnecessary alarms have been proposed based on techniques from signal processing, statistics, and machine learning [4], [5], [6]. Many of such approaches need patient data labels to be designed or trained, and all of them need the labels to be evaluated. This paper introduces a lightweight way of performing these evaluations without investing in high-quality labels. Note that the proposed method is not specific to any physiological input, unlike many alarm suppression techniques.

The gold standard for evaluating the clinical effectiveness of an alarm suppression system is an interventional study, in which the researchers deploy the system and measure its effects compared to a control group. To estimate the sensitivity and specificity of suppression, a controlled observational study would be sufficient: patient data is fed into a physiological monitor with and without suppression, the results are observed separately from the clinical context, and a comparison is made based on the desired alarms (as defined by the clinical experts). Both types of studies require substantial time and effort, in part due to the need to label the suppressibility of alarms. For example, nurses can review video feeds of patients as part of the labeling process [10]. Our work is not meant to replace either type of studies; instead, we aim to prioritize, guide, and reduce the risk of observational studies by providing an early and cheap estimation of the expected suppression performance.

High-precision methods of labeling alarm data include patient simulations and computer-aided clinical trials [11]. To provide realistic data, these methods require detailed physiological models, building which is a large investment. For clinical alarms, an appropriate model is rarely available. Our method is related to observational studies in the same way as computer-aided clinical trials are related to traditional clinical trials. That is, we perform a virtual algorithmic evaluation of suppressibility. After that, our results can provide the basis for an observational study of suppression or a clinical trial of an alarm device.

Recently, a quick and inexpensive way of labeling data has emerged, known as *data programming* [9]. A key element of data programming is a set of quantitative intuitions about how the data corresponds to labels. For example, a clinician might say, "when a patient over 60 years old has had a heart rate over 120 beats for over a minute, such an alarm is not suppressible." These intuitions, algorithmically represented as *labeling functions*, are allowed to be incomplete, sometimes incorrect, and contradictory. A labeling function returns a class label or an "abstain" verdict for any input. Given a diverse combination of many labeling functions and an unlabeled dataset, data programming algorithms produce probabilistic labels — a label and a confidence between 0 and 1 — for each sample in the dataset. A prominent data programming tool *Snorkel* [12] estimates an optimal weight for each labeling function by using a generative graphical model. Our approach encodes clinical intuitions about suppressible/non-suppressible alarms as labeling functions, feeds them along with alarm data into Snorkel, and relies on the resulting probabilistic labels to quantify the uncertainty in the suppression of an alarm device.

IV. Problem Formulation

This section states the problem addressed in this paper — first at a high-level, and then mathematically.

A. High-Level Problem Statement

A suppression system takes an alarm as input and decides whether to suppress it. The system can be configured with various settings: thresholds, timeouts, and so on. The hospital policy makers want to estimate the sensitivity and specificity of

this suppression system at various settings. For this estimation, they have collected a sample dataset of alarms; however, they do not know which of those alarms should actually be suppressed. Ideally, this information would be obtained from an observational study, but it is not carried out for various practical reasons.

The problem considered in this paper is to predict the sensitivity/specificity of a suppression system that would result from an observational study with perfect alarm annotations. We aim to make that prediction in the form of tight sensitivity/specificity bounds that would contain the observational study's estimates with high probability.

B. Technical Problem Statement

Given a set B, we write $|B|$ to be the set's cardinality and B^m to be a Cartesian product of m sets B. We write $\mathbb{1}(C)$ to be the indicator function for condition C.

Let $\mathcal{X}$ be the feature space of all possible alarms, $\mathcal{Y} = \{0, 1\}$ be the label space for alarm suppression where 1 denotes suppressible and 0 denotes non-suppressible. A single alarm is denoted as $x \in \mathcal{X}$, and we consider a finite dataset of alarms $X \subset \mathcal{X}$ generated by random variables $\tilde{X} = \{\tilde{x} \mid x \sim \tilde{x}, x \in \mathcal{X}\}$. These alarms have respective *unknown* true suppressibility labels $Y \subset \{y \in \mathcal{Y}\}$.

A *suppression system* $S : \mathcal{X} \to \mathcal{Y}$ decides whether an alarm is suppressible. For its evaluation, suppose the alarms are indexed by an *index set* $\mathcal{I} \subseteq \{1, 2, \dots\}$. The *suppression accuracy* R_j of system S on class j evaluated on $\mathcal{I}$ is defined as

$$R_j(\mathcal{I}) = \frac{\sum_{n \in \mathcal{I}} \mathbb{1}(y_n = j \wedge S(x_n) = j)}{\sum_{n \in \mathcal{I}} \mathbb{1}(y_n = j)} \tag{1}$$

The above expression is the true rate of the suppression system S on $|\mathcal{I}|$ samples with true label j. R_1 is the sensitivity of S, and R_0 is the specificity of S, as described in Section II.

A *labeling function* (LF) $\lambda : \mathcal{X} \to \hat{\mathcal{Y}}$ produces a label in the weak label space $\hat{\mathcal{Y}} = \mathcal{Y} \cup \{-1\}$ where -1 denotes an abstain. Given a finite set of labeling functions, $\Lambda \subset \{\lambda : \mathcal{X} \to \hat{\mathcal{Y}}\}$, we denote a *labeling outcome* as a tuple of labels on a given datapoint x by $L_\Lambda(x) = \big(\lambda_1(x), \dots, \lambda_{|\Lambda|}(x)\big)$. The set of all possible labeling outcomes for functions Λ is denoted as $\mathcal{L}_\Lambda = \hat{\mathcal{Y}}^{|\Lambda|}$. Thus, $L_\Lambda : \mathcal{X} \to \mathcal{L}_\Lambda$.

We will rely on a class of *generative models* $\mathcal{H} = \{h : \mathcal{L}_\Lambda \to \mathcal{P}(\mathcal{Y})\}$, where $\mathcal{P}(\mathcal{Y})$ is a space of all probability distributions over $\mathcal{Y}$. Each such model can be understood as a pair of functions, a *predictor* $f : \mathcal{L}_\Lambda \to \mathcal{Y}$ and *confidence estimator* $g : \mathcal{L}_\Lambda \to [0, 1]$, by setting

$$f := \operatorname{argmax} h(L_\Lambda(X)) \qquad g := \max h(L_\Lambda(X)) \tag{2}$$

such that, for a datapoint x, $(f(x), g(x)) = (\hat{y}, \hat{p})$ is the label prediction $\hat{y}$ with confidence $\hat{p}$. In the event that the produced distribution over $\mathcal{Y}$ is uniform, f returns the suppressible label (*i.e.,* 1) and g assigns confidence of 0.5. [1]

[1]In practice, alarm datasets are heavily skewed with suppressible alarms. Therefore, assigning samples on which the model is uncertain to the majority class does not significantly impact results.

In our method, we will consider subsets of our data X that have high confidence g of labels f. Suppose that for class $j \in \mathcal{Y}$, we are willing to tolerate the *label uncertainty* of ϵ_j; in other words, for that class, we only use probabilistic labels with confidence of at least $1 - \epsilon_j$. We denote some set of *indices with high confidence in label* j as $\mathcal{I}_j$ and define it as

$$\mathcal{I}_j \subseteq \{n \in \mathcal{I} \mid f(x_n) = j \wedge g(x_n) \geq 1 - \epsilon_j\} \tag{3}$$

Analogously, a hypothetical observation study would manually label some samples in X and exactly determine their true labels. We refer to these samples with a *true-label index set* $\mathcal{I}_j^*$ for class j:

$$\mathcal{I}_j^* \subseteq \{n \in \mathcal{I} \mid y_n = j\} \tag{4}$$

Note that the sizes of sets $\mathcal{I}_j^*$ determine the desired numbers of samples of each class. These numbers are crucial to observation studies and are computed up-front based on power analysis and resource limitations.

We now formally state this paper's technical problem.

Problem Statement 1 (Confidence Interval Estimation): Given the following:

- Unlabeled alarms X, which have unknown true labels Y,
- Predictor $f : \mathcal{L}_\Lambda \to \mathcal{Y}$ and confidence estimator $g : \mathcal{L}_\Lambda \to [0, 1]$, which operate over labeling functions Λ.
- The sizes of index sets $|\mathcal{I}_j^*|$ for a hypothetical observation study for $j \in \mathcal{Y}$.

Our goal is, for any class $j \in \mathcal{Y}$ and confidence level p_j, to find the tightest interval C_j containing, with probability at least p_j, the observation-study estimate $R_j(\mathcal{I}_j^*)$ of the suppression accuracy on class j; that is,

$$\min_{C_j} |C_j| \text{ subject to } \mathbb{P}(R_j(\mathcal{I}_j^*) \in C_j) \geq p_j$$

V. Estimation of Suppression Accuracies

In this section we describe our approach for producing confidence bounds for the sensitivity and specificity of a suppression system. Figure 2 summarizes the steps our approach:

A. Collect unlabeled alarm and patient data
B. Elicit heuristic labeling functions from clinicians
C. Produce probabilistic labels for the alarm data
D. Estimate the suppression accuracies of the system
E. Quantify confidence bounds around those estimates

A. Unlabeled Data Collection

Our initial step is to collect a dataset of representative alarms and the corresponding patient data, on which the suppression system will be evaluated. The patient data includes the static data (demographics, disease history, etc.) and the vital signals that contextualize a raised alarm. Thus, there are two key aspects of data: determining which alarm instances to use and collecting the relevant vitals data. More formally, we produce a set of representative unlabeled alarms and patient data $X = (x_1, x_2, \dots)$ with indices $\mathcal{I}$, with each datapoint corresponding to the features of an alarm from $\mathcal{X}$.

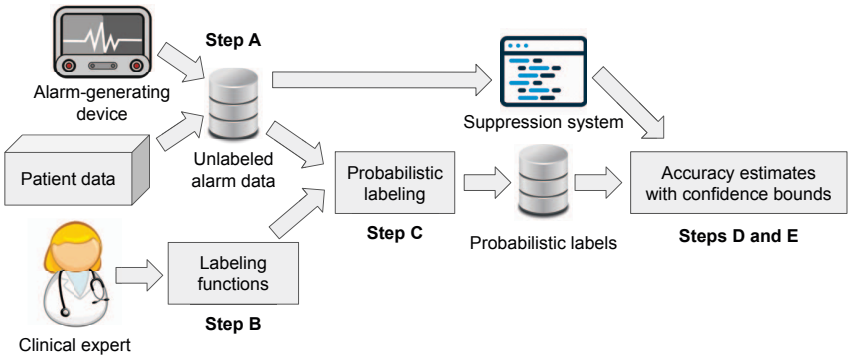


Fig. 2: Our approach of estimating the performance of suppression systems.

When choosing the alarms, our goal is to get a sample from the representative distribution in a particular clinical setting. We use the state-of-the-art approaches for appropriate sampling. Typically, alarms would be sampled from the subtypes targeted by the suppression system (*e.g.,* technical or clinical, discussed in Section VI), based on their frequency during different times of the day, and appropriately from the target patient demographics.

The collection of patient and alarm data should be fully automated and cheap as a result. Given full automation, we aim for the data to be as complete as possible in a given clinical context. More complete datasets, such as those that include more patient information and diverse vitals, allow for richer labeling functions in the next step, ultimately improving the outcomes of our approach.

B. Eliciting Labeling Functions

We ask clinical experts (*e.g.,* physicians and nurses) working in the targeted alarm suppression context to describe the guidelines that they use to make decisions on whether an alarm is suppressible or non-suppressible. Specifically, we seek *quantitative* guidelines for determining alarm suppressibility (*e.g.,* if heart rate is above 200 then non-suppressible), as opposed to *qualitative* guidelines (*e.g.,* if a child is kicking/moving then suppressible) which are often used by clinicians. Qualitative guidelines are difficult to encode into our data-driven approach and, hence, are excluded from this study — but may be explored in future work. Intuitively, guidelines are not perfect; they output noisy labels. However, very inaccurate labeling functions can negatively affect performance down the line. To address this, we ask clinicians stick to guidelines that, in their expert opinion, are *better than random chance* at identifying a suppressible or non-suppressible alarm. Better-than-random labeling functions are a common requirement in data programming [9].

Each guideline is implemented as one or more labeling functions. Each labeling function takes patient data as input and either emits a label (suppressible/non-suppressible) or abstains for each sample in the unlabeled alarm dataset. Formally, the labeling functions Λ are elicited and applied to the data X to obtain the weak labels $L_\Lambda(X)$.

C. Probabilistic Labeling

In this step, we combine the weak labels from the labeling functions into a single "strong" label. This strong label is characterized by a confidence value, indicating the level of certainty regarding the label's accuracy. Mathematically, we combine the weak labels $L_\Lambda(x)$ of each alarm x into a probabilistic strong label $f(x)$ with confidence $g(x)$. Generally, this can be achieved with a weighted combination over the weak labels with a fixed vector of weights w.

We model the weighted combination as a generative graphical model. Generative models are popular in state-of-the-art data programming literature [9]. This model leverages the agreements and disagreements of the labeling functions to estimate their accuracies. The accuracies then inform the weights (*i.e.,* relative priorities) of labeling functions.

Our goal in this step is to train a generative model from class $\mathcal{H}$ that represents a joint distribution $\mathbb{P}_w(L_\Lambda(\tilde{X}), y)$ between the random alarm variables $\tilde{X}$ and their hypothesized true labels y — without any samples of the true labels. A weighting scheme w is a function of the accuracy of each labeling function, and thus unknown. This model also takes into account the prior probability of each label occurring in the data. Prior probabilities $\mathbb{P}(\mathcal{Y})$ over classes $\mathcal{Y}$ can either be specified if known or estimated from the labeling functions. Training this model is equivalent to learning w, which is estimated by maximizing the log-likelihood of the observed labeling function outputs $L_\Lambda(X)$:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \log \sum_{y \in \mathcal{Y}^{|\Lambda|}} \mathbb{P}_w(L_\Lambda(\tilde{X}), y)$$

Using this learned weight vector w, we can use the probabilistic distribution over the labels output by the generative model,

$$h_y(L_\Lambda(x)) = \mathbb{P}_{\hat{w}}(y \mid L_\Lambda(x))$$

to encode a predictor f and confidence estimator g according to Equation 2.

D. Estimation of Accuracies

Now we estimate the sensitivity and specificity of the suppression system using the probabilistic labels. In this subsection, we are interested in finding a pair of numbers for one

suppression system as its sensitivity/specificity estimates given the probabilistic labels. Formally, for each $j \in \mathcal{Y}$, we compute point estimates of $R_j(X, Y)$ for suppression system S given the data X, predictor f, and confidence estimator g. The main challenge here is to balance the labeling uncertainty in g with the sampling uncertainty related to the size of X.

First, we need to pick the data that was labeled in a trustworthy manner. In our experience, using the whole dataset X is inadvisable because low-accuracy/low-confidence labels would bias the outcome. Therefore, intuitively, we place more trust in samples that we label with high confidence (*i.e.,* with low uncertainty about their true label). Suppose that for label $j \in \mathcal{Y}$, we tolerate at most ϵ_j labeling uncertainty and, thus, only select samples with confidence at least $1 - \epsilon_j$. This leads us to consider *high-confidence index sets* $\mathcal{I}_j$ parameterized by ϵ_j as defined in Equation 3.

We can pick an arbitrary subset of high-confidence samples. We cannot, however, pick an arbitrarily small ϵ_j: very few samples would be available, and that would significantly increase the sampling uncertainty of our estimates. In short, there is a trade-off between the labeling and sampling uncertainties (demonstrated in Appendix A). We resolve this trade-off by searching for the value of ϵ_j that minimizes a combination of those uncertainties as explained in the next subsection.

Then, given sets $\mathcal{I}_j$, we estimate the suppression accuracy on each class j by applying Equation 1 to the high-confidence labels in place of y_n:

$$R_j(\mathcal{I}_j) = \frac{\sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(x_n) = j\right)}{|\mathcal{I}_j|} \tag{5}$$

This formula gives us a point estimate of sensitivity (for $j = 1$) and specificity (for $j = 0$) of the suppression system.

E. Confidence Bounds

Our accuracy estimates from the previous subsection rely on noisy labeling functions and, as a result, can be unreliable. We quantify their reliability by providing *confidence bounds* around our estimates — the intervals where the accuracy estimated from true labels would be found with some given confidence. More precisely, we interpret a confidence bound as follows: it is an interval of likely estimates of suppression accuracy of some class using the true labels of that class (from a gold-standard observation study) instead of our high-confidence probabilistic labels. That is, for class j, we aim to create an interval $C_j = [R_j(\mathcal{I}_j) - c_j, R_j(\mathcal{I}_j) + c_j]$ containing, with probability of at least p_j, the suppression accuracy $R_j(\mathcal{I}_j^*)$ estimated from manually labeled samples in $\mathcal{I}_j^*$. Notice that p_j can differ between the classes and, hence, reflect the acceptable clinical risks.

The interval size c_j depends on two factors: the sampling randomness between $\mathcal{I}_j$ and $\mathcal{I}_j^*$ and the quality of the probabilistic labels in $\mathcal{I}_j$. The former will be estimated as a function the sizes of $\mathcal{I}_j$ and $\mathcal{I}_j^*$ using the standard statistical bounds. For the latter, in our experience, with appropriate labeling functions discussed Section V-B, high-confidence labels correspond to the low-uncertainty situations in which the suppression mechanisms are relatively consistent. We formalize this intuition with the following assumption.

Assumption 1 (Consistent Accuracy across Datasets): Suppression accuracies on high-confidence sets $\mathcal{I}_j$ do not differ in expectation from those on the manually labeled sets $\mathcal{I}_j^*$ by more than the average uncertainty of the labels in $\mathcal{I}_j$:

$$\left| \frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{E}\left[S(\tilde{x}_m) = j\right] - \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{E}\left[S(\tilde{x}_n) = j\right] \right| \leq 1 - \eta_j,$$

where $\eta_j = \dfrac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} g(x_n)$ is the average label confidence in $\mathcal{I}_j$.

Leveraging the above assumption, we can derive the desired bound c_j on the difference between the estimates based on our probabilistic labels and the potential observation-study labels. The probability of exceeding that bound is given in the following theorem.

Theorem 1 (Bounded Difference of Accuracy Estimates): For any class $j \in \mathcal{Y}$, the difference between the probabilistic and manual estimates of suppression accuracy on j exceeds bound c_j with a bounded probability for any free parameter γ_j:

$$\mathbb{P}\left[\left|R_j(\mathcal{I}_j) - R_j(\mathcal{I}_j^*)\right| \geq c_j\right] \leq \\ 2\exp\left(-2|\mathcal{I}_j^*|(c_j + \eta_j - 1 - \gamma_j)^2\right) + 2\exp\left(-2|\mathcal{I}_j|\gamma_j^2\right)$$

The proof can be found in Appendix B. This result means that the chance of our estimates disagreeing with gold-standard estimates by more than c_j decreases with the increasing number of samples in $\mathcal{I}_j$ and $\mathcal{I}_j^*$, larger c_j, and the higher confidence of our estimates η_j. The parameter γ_j can be chosen as any value. This bound is contingent on the satisfaction of Assumption 1 about probabilistic labeling.

We want to guarantee that the manually labeled estimate $R_j(\mathcal{I}_j^*)$ is within c_j of our estimate $R_j(\mathcal{I}_j)$ with probability p_j. Then, by equating p_j with the bound and expressing c_j in terms of p_j, we obtain the desired interval size c_j exactly.

Corollary 1: For any desired confidence p_j and any admissible γ_j, the interval width c_j can be chosen as

$$1 - \eta_j + \gamma_j + \sqrt{\frac{\ln(2) - \ln(p_j - 2\exp(-2|\mathcal{I}_j|\gamma_j^2))}{2|\mathcal{I}_j^*|}}$$

Now we return to the problem formulation from the end of Section IV: our goal is to minimize the size of the interval c_j given a fixed confidence p_j. So we optimize for the smallest interval over the values γ_j and the choice of samples in $\mathcal{I}_j$ (by changing ϵ_j in Equation 3), which then determines the η_j. Thus, we pick the interval size c_j as follows:

$$\min_{\gamma_j \in \mathbb{R},\ \mathcal{I}_j \subseteq \mathcal{I}} 1 - \eta_j + \gamma_j + \sqrt{\frac{\ln(2) - \ln(p_j - 2\exp(-2|\mathcal{I}_j|\gamma_j^2))}{2|\mathcal{I}_j^*|}}$$

In summary, the presented results give us a way to produce uncertainty bounds for the accuracies of the suppression system by putting a confidence bound around the accuracy estimates from Section V-D. We pick the tightest interval given Theorem 1. This interval captures both the uncertainty of our probabilistic labels and the sampling uncertainty.

VI. Case Study: Low SpO_2 Alarm Dataset

To evaluate the performance of our method, we conducted a case study for low SpO_2 alarms. We consider an alarm suppression system that suppresses a low SpO_2 alarm if the SpO_2 measurement at the time of alarm is above a specified threshold, otherwise it does not suppress the alarm. Our goal is to establish and visualize the connection between the system's SpO_2 threshold and its specificity/sensitivity, given a dataset of patient vitals data and manually-annotated low SpO_2 alarm data. In this section, we overview the dataset and data preprocessing approach, introduce the labeling functions collected for labeling low SpO_2 alarms, describe method implementation details, and present a comparative approach for our analysis.

A. Data

We used a deidentified dataset originally collected as part of a study approved by the Institutional Review Board of the Children's Hospital of Philadelphia (IRB #14-010846). Researchers video-recorded 551 hours of patient care on a medical unit at Children's Hospital of Philadelphia during July 2014 to November 2015 from 100 children whose families and nurses consented. In addition, the following data was collected: patient background information, all physiologic monitoring alarms with corresponding timestamps, and continuously recorded vital signs:

- Blood oxygen saturation (SpO_2) measured by a pulse oximeter,
- Pulse rate measured by a pulse oximeter,
- Heart rate measured by a 3-lead electrocardiography (ECG),
- Cardiac rhythm measured by a 3-lead ECG
- Respiratory rate measured by a 3-lead ECG,
- Noninvasive blood pressure (NBP) measured by a cuff, from the physiologic monitoring network.

After the study, the alarms were reviewed along with the video recordings and then annotated with three alarm distinctions in mind: technical versus clinical alarms, valid versus invalid alarms, and actionable versus non-actionable alarms [10]. Technical alarms indicate an issue with a physiologic monitor or its sensors, whereas clinical alarms indicate an issue with a patient's physiologic status (*e.g.,* heart rate is too high). Valid alarms are those that correctly identify the physiologic status of a patient. Conversely, alarms that are false are considered invalid. A valid clinical alarm that results in or warrants clinical intervention or consultation can be further classified as actionable, otherwise non-actionable. Hence the alarms have the following annotations: technical alarms, invalid clinical alarms, valid actionable clinical alarms, and valid non-actionable clinical alarms.

A total of 9547 clinical alarms of 26 different types are in the dataset. Low SpO_2 generated the largest number of alarms (34% of total alarms) and the largest number of invalid alarms (81% of the low SpO_2 alarms). Hence, adjusting the settings of a low SpO_2 alarm suppression system can help reduce alarm fatigue, and we focus on these alarms in our case study.

B. Data Preprocessing

Our analysis only considers a subset of the original dataset:

- Patient age group: less than one month old, from one month to less than two month, from two month to less than six month, and six months and older;
- Patient vital signs: blood oxygen saturation, respiratory rate, heart rate measured by an ECG, and heart rate measured by a pulse oximeter — all measured at maximum sampling rate of 0.2 Hz;
- Annotated low SpO_2 alarms with corresponding timestamps and durations.

The alarms data is annotated in terms of technical/clinical, valid/invalid, and actionable/non-actionable alarms. We interpret these labels with respect to suppressibility as follows. Technical alarms, valid non-actionable clinical alarms, and invalid alarms are interpreted as suppressible, whereas only valid actionable clinical alarms are non-suppressible.

C. Labeling Functions for Low SpO_2 Alarms

In unstructured interviews with two pediatric physicians, we collected eighteen guidelines for deciding whether a low SpO_2 alarm is suppressible or non-suppressible. Six of the guidelines are excluded from this study because the dataset does not have sufficient information to implement them. The guidelines are as follows.

1) *Long alarm*: If the alarm duration is longer than t seconds, then the alarm is likely non-suppressible.
2) *SpO_2 below threshold for duration*: If SpO_2 is below threshold x for longer than t seconds since the alarm sounded, then the alarm is likely non-suppressible.
3) *Heart rate above threshold for duration*: If heart rate is above threshold x for longer than t seconds since the alarm sounded, then the alarm is likely non-suppressible.
4) *Heart rate below threshold for duration*: If heart rate is below threshold x for longer than t seconds since the alarm sounded, then the alarm is likely non-suppressible.
5) *Respiratory rate below threshold for duration*: If respiratory rate is below threshold x for longer than t seconds since the alarm sounded, then the alarm is likely non-suppressible.
6) *Repeat alarms*: If more than n alarms occurred within t seconds of the alarm, then the alarm is likely non-suppressible.
7) *Short alarm*: If the alarm duration is less than t seconds, then the alarm is likely suppressible.
8) *Immediate recovery*: If SpO_2 recovers to x within t seconds after the alarm sounds, then the alarm is likely suppressible.
9) *Heart rate technical error*: If the difference between ECG heart rate and pulse oximeter heart rate is greater than x at the time of the alarm, then the alarm is likely suppressible.

10) *Bad SpO_2 waveform*: If the SpO_2 waveform contains anomalies [2], then the alarm is likely suppressible.
11) *Bad heart rate waveform*: If the ECG heart rate waveform contains anomalies, then the alarm is likely suppressible.
12) *Bad respiratory rate waveform*: If the respiratory rate waveform contains anomalies, then the alarm is likely suppressible.

From these guidelines, we instantiated sixty-two total labeling functions for different values of parameters x, t, n, picked in consultation with the two aforementioned physicians (see Appendix C). Forty of them produce only suppressible labels, and the rest produce only non-suppressible labels.

D. Implementation

We implement the labeling functions as Python functions, taking in an alarm from the dataset and returning either a label or an abstain. To generate probabilistic labels for the low SpO_2 alarms in the dataset we use a tool called Snorkel for the generative model [12]. Snorkel is the state-of-the-art tool for weak label combination and has been applied to several applications. We use the current version at the time of this publication, version 0.9.7 (www.snorkel.org). The only hyperparameter we specify within Snorkel are the prior probabilities of labels. In interviews with physicians, it was determined that 80%/20% for suppressible/non-suppressible alarms, respectively, is a reasonable default for alarm suppression (and is also approximately consistent with our dataset). If the prior was unknown, it could be estimated from the labeling functions [13].

Estimates of sensitivity and specificity of the low SpO_2 alarm suppression system are computed on the pulse oximetry data. We consider only the timestamps for which an SpO_2 measurement is present (11300 samples in total). Then we map the labels assigned to the known low SpO_2 alarms onto these samples. For each alarm, we assign all timestamps that occur during this alarm with its own labels (true and probabilistic). Lastly, we simulate applying the alarm suppression system to the samples for SpO_2 thresholds of 0 to 100, and save the result as a label. Thus for each sample we have,

- A timestamp,
- A SpO_2 measurement,
- A ground-truth label (from the original annotations),
- A label and its confidence from the generative model,
- A label from the suppression system for each SpO_2 threshold.

For the optimization problem for finding the best confidence bounds, we use the SciPy Python library (www.scipy.org).We minimize a closed-form function with bounded parameters.

E. A Comparative Approach

We compare the performance of the confidence bounds produced by our approach with a majority vote approach. *Majority vote* is a widely-used and straightforward method for combining multiple discrete signals into one. In this case we apply it to weak labels produced by labeling functions. In this method, each labeling function is assigned equal weight and thus has equal influence on the label prediction. The label prediction is determined as the weak label that received the most votes. The confidence of a particular label is computed as the fraction of non-abstaining labeling functions that voted for this label. We assume that Assumption 1 holds for majority vote since, intuitively, as more labeling functions agree on a particular label, the more we trust that that label reflects the true unknown label. Hence in this comparative approach, steps A, B, D, and E are performed exactly as described in Section V, while step C is replaced with probabilistic labeling via majority vote.

A primary challenge of the majority vote approach is, if many of the labeling functions are inaccurate, the label prediction can often be incorrect *but still have a high confidence*. Since the labeling function accuracies cannot be known a priori due to the absence of true labels, there is no clear way of preventing this situation. Due to this challenge, we use the majority vote approach only for comparison and do not recommend using it in practice.

VII. Results

In this section we present the results of our case study for low SpO_2 alarms. Specifically, we evaluate the performance of the confidence bounds for suppression accuracies of a low SpO_2 alarm suppression system produced by our approach. A successful application of our approach would result in tight confidence bounds that contain the true suppression accuracies that would be produced in an observational study.

We consider a 5%, 10%, and 20% chance of the confidence bounds not containing the true suppression accuracies, *i.e.,* $p_j \in \{0.05, 0.10, 0.20\}$. Label uncertainty ϵ_j determines which samples are used to estimate the suppression accuracies and compute the confidence bounds, and hence is an important parameter of our approach. We considered the uncertainty of at most 10% to avoid violating Assumption 1. Now, for each p_j, we perform an optimization to find the tightest bounds with constraints $\epsilon_j \in [0.01, 0.1]$ and $\gamma_j \in [0, 1]$.

The confidence bounds for sensitivity and specificity using our approach are depicted in Figure 3 and using the comparative approach in Figure 4. We also illustrate the estimated trade-off between sensitivity and specificity in Figure 5. To draw the bounds in this figure, for each confidence level p_j, then for each SpO_2 threshold, we plot the $(\text{specificity} + c_0, \text{sensitivity} + c_1)$ for the upper-bound and $(\text{specificity} - c_0, \text{sensitivity} - c_1)$ for the lower-bound. Since we have access to true labels (*i.e.,* the labels extracted from the alarm annotations), we use them to plot the *true curve* for the sensitivity/specificity/trade-off, which represents the results of an observational study. Table I presents the average width of the sensitivity and specificity confidence bounds. Table II summarizes the percentage of these true curves that are contained in each of the confidence bounds. We note that only SpO_2 thresholds between 80 and 95 can plausibly be

[2]Waveforms with artifacts are generally unreliable. We look for anomalies (*e.g.,* spikes and outliers) in the waveform to determine if it is bad or not.

adopted into a clinical setting, and hence only portions of these curves are clinically relevant.

We observe that our approach successfully produced narrow confidence bounds with high containment, whereas the comparative approach produced narrow bounds that suffer from low containment. Our approach's confidence bounds for sensitivity are tighter than that of the comparative approach (4-5% and 5-7% in average width, respectively). Furthermore, our bounds contain all of the true curve (with the exception of 6% of the curve's length for $p_j = 0.2$), whereas the comparative bounds contain only 80-81%. For specificity, our approach produced looser bounds (18-21% average width) than the comparative approach (6-8% average width). However, our approach achieved full containment of the true specificity curve as opposed to the comparative approach which contained 76-78%. The difference in true curve containment between the approaches is even more exaggerated in the clinically relevant region. Most of the true sensitivity and specificity curves in this region are not contained by the comparative approach's bounds.

In practice, hospital policy makers would select an SpO_2 threshold for this suppression system based on the trade-off between its sensitivity and specificity (illustrated in Figure 5). The advantage of using our approach over the comparative approach is clear here. While the comparative approach's bounds are tighter than our approach's bounds, they contain less than 12% of the true trade-off curve, which can lead to a misguided policy. On the other hand, the bounds from our approach have low-to-moderate width (which appropriately indicates the uncertainty) and contain the entire true curve. The region of the plot where specificity is greater than 50% corresponds to the clinically relevant region, and even here, our approach outperforms the comparative approach.

Since the bounds from our approach effectively capture the true sensitivity/specificity trade-off of the suppression system, a policy maker could use our bounds to select the system's SpO_2 threshold. A good SpO_2 threshold would produce specificity close to one (*i.e.,* not suppress any non-suppressible alarms) while maximizing sensitivity (*i.e.,* silence as many suppressible alarms as possible). This corresponds to the lower-right region of Figure 5a. Suppose policy makers decide to allow a minimum of 90% specificity. Our approach determines that an SpO_2 threshold of at least 92 is required which can suppress up to 6% of false alarms (based on the sensitivity) Using the true curve, a minimum SpO_2 threshold of 91 is required and at most 3% of false alarms would be suppressed.

Trade-off curves generally bow inward, but we observe in Figure 5b that the comparative approach's confidence bounds bow outward. If we consider flipping the labels that majority vote outputs in our comparative approach, the bounds would go inward and exhibited slightly improved containment of the true curve. This implies that majority vote, on the samples it labeled suppressible/non-suppressible, was mostly incorrect with high-confidence.

Limitations: our confidence bounds are accurate when suppression accuracy is relatively consistent on different high-confidence labels, as stated in Assumption 1. This assumption may be violated in contexts with few available samples or when high-confidence labeling is particularly biased/inaccurate — and then our theoretical guarantees might not hold. Our case study has been performed on a dataset collected from pediatric patients on a medical floor in a hospital, and the alarms were labeled for being actionable. To apply our method to a different setting, one may need to elicit different/more labeling functions, and so the tightness and accuracy of the confidence bounds may vary.

	p_j	Sensitivity	Specificity
Our Approach	0.05	0.049	0.211
	0.10	0.045	0.201
	0.20	0.041	0.187
Comparative Approach	0.05	0.065	0.078
	0.10	0.060	0.072
	0.20	0.054	0.065

TABLE I: Average width of the confidence bounds.

	p_j	Sensitivity	Specificity	Trade-off
Our Approach	0.05	1.0	1.0	1.0
	0.10	1.0	1.0	1.0
	0.20	0.940	1.0	1.0
Comparative Approach	0.05	0.810	0.780	0.115
	0.10	0.800	0.780	0.115
	0.20	0.800	0.760	0.109

TABLE II: Percentage of the true sensitivity/specificity/trade-off curve contained in the confidence bounds.

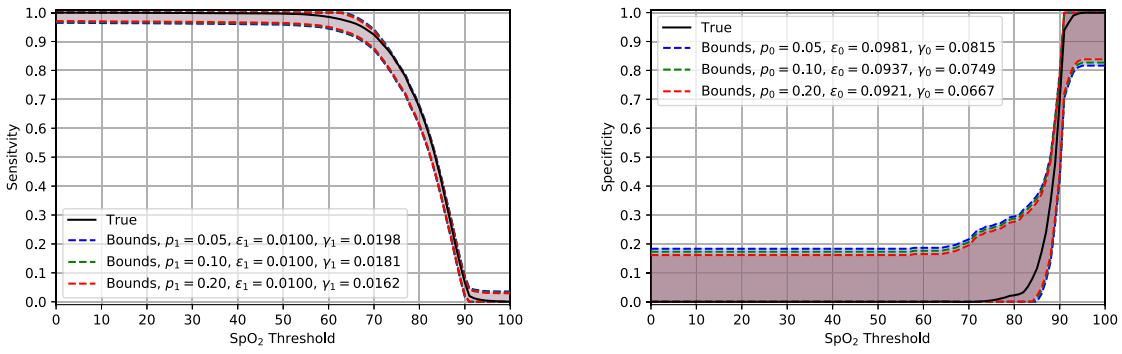
VIII. Conclusion

In this paper, we proposed an approach for estimating the performance of a physiologic alarm suppression system with access only to unlabeled data. Generative modeling is used to produce probabilistic labels that serve as proxy to the unknown ground-truth labels when computing suppression accuracy estimates. We then provide a confidence bound on these accuracy estimates. Finally, we evaluated our method in a case study for low SpO_2 alarms and showed that we find mostly tight confidence bounds that contain the true curve almost always.

This work suggests a handful of directions for future work. First, we plan to automate the extraction of weak labeling functions to satisfy the consistency assumption of generative models, which will likely require explicitly encoding the dependencies between labeling functions. Second, over-confident (poorly-calibrated) probabilistic labels can have adverse effects on the results of our method, hence we plan to explore unsupervised calibration for data programming and/or develop alternative approaches to producing probabilistic labels. Finally, we also seek to validate our approach on other alarm types (*e.g.,* tachycardia and high/low respiratory rate).

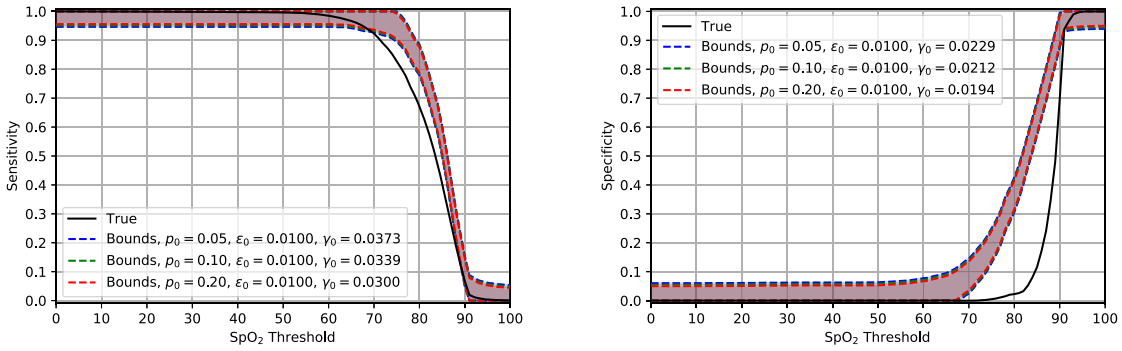
References

[1] C. W. Paine, V. V. Goel, E. Ely, C. D. Stave, S. Stemler, M. Zander, and C. P. Bonafide, "Systematic review of physiologic monitor alarm characteristics and pragmatic interventions to reduce alarm frequency," *Journal of Hospital Medicine*, vol. 11, no. 2, pp. 136–144, 2016.



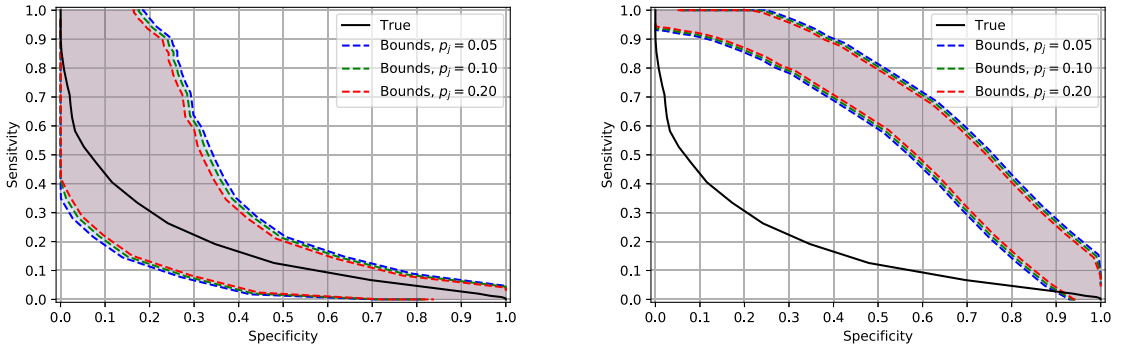
(a) (b)

Fig. 3: Our approach's confidence bounds for (a) sensitivity and (b) specificity. The bounds are tight for specificity and moderately-sized for sensitivity, and both bounds contain most, if not all, of their respective true curves.



(a) (b)

Fig. 4: Comparative approach's confidence bounds for (a) sensitivity and (b) specificity. The bounds for both sensitivity and specificity are tight, but demonstrate low containment of the true curve, particularly in the SpO$_2$ threshold region of 70 to 90.



(a) Our Approach (b) Comparative Approach

Fig. 5: Sensitivity/Specificity trade-off confidence bounds produced by (a) our approach and (b) the comparative approach. Our approach produced low-to-moderate width bounds that capture the entire true trade-off curve, whereas the comparative approach produces narrow bounds that contain a very low percentage of the true curve.

[2] B. J. Drew, P. Harris, J. K. Zègre-Hemsey, T. Mammone, D. Schindler, R. Salas-Boni, Y. Bai, A. Tinoco, Q. Ding, and X. Hu, "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients," *PloS one*, vol. 9, no. 10, p. e110274, 2014.
[3] B. D. Winters, M. M. Cvach, C. P. Bonafide, X. Hu, A. Konkani, M. F. O'Connor, J. M. Rothschild, N. M. Selby, M. M. Pelter, B. McLean *et al.*, "Technological distractions (part 2): a summary of approaches to manage clinical alarms with intent to reduce alarm fatigue," *Critical Care Medicine*, vol. 46, no. 1, pp. 130–137, 2018.
[4] P. Lameski, E. Zdravevski, S. Koceski, A. Kulakov, and V. Trajkovik, "Suppression of intensive care unit false alarms based on the arterial blood pressure signal," *IEEE Access*, vol. 5, pp. 5829–5836, 2017.
[5] H. Nguyen, S. Jang, R. Ivanov, C. Bonafide, J. Weimer, and I. Lee, "Reducing pulse oximetry false alarms without missing life-threatening events," *Smart Health*, vol. 9-10, 07 2018.
[6] W.-T. M. Au-Yeung, A. K. Sahani, E. M. Isselbacher, and A. A. Armoundas, "Reduction of false alarms in the intensive care unit using an optimized machine learning based approach," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–5, 2019.
[7] C. P. Bonafide, A. R. Localio, J. H. Holmes, V. M. Nadkarni, S. Stemler, M. MacMurchy, M. Zander, K. E. Roberts, R. Lin, and R. Keren, "Video analysis of factors associated with response time to physiologic monitor alarms in a children's hospital," *JAMIA Pediatrics*, vol. 171, no. 1, pp. 524–531, 2017.
[8] L. Kobayashi, J. W. Gosbee, and D. L. Merck, "Development and application of a clinical microsystem simulation methodology for human factors-based research of alarm fatigue," *HERD: Health Environments Research & Design Journal*, vol. 10, no. 4, pp. 91–104, 2017.
[9] A. Ratner, C. De Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," *Advances in neural information processing systems*, vol. 29, p. 3567, 2016.
[10] M. MacMurchy, S. Stemler, M. Zander, and C. P. Bonafide, "Research: Acceptability, feasibility, and cost of using video to evaluate alarm fatigue," *Biomedical instrumentation & technology*, vol. 51, no. 1, pp. 25–33, 2017.
[11] K. Jang, J. Weimer, H. Abbas, Z. Jiang, J. Liang, S. Dixit, and R. Mangharam, "Computer aided clinical trials for implantaule cardiac devices," vol. 2018, 07 2018, pp. 1–4.
[12] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *The VLDB Journal*, vol. 29, no. 2, pp. 709–730, 2020.
[13] A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré, "Training complex models with multi-task weak supervision," 2018.
[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.
[15] S. Jang, R. Ivanov, I. lee, and J. Weimer, "Confidence calibration with bounded error using transformations," 2021. [Online]. Available: https://arxiv.org/abs/2102.12680

Appendix A
Over-Confident Probabilistic Labels

Neural networks have been shown to be over-confident on their predictions [14]. This is problematic, especially in safety-critical applications like medicine, because the predictor can be wrong with high confidence. Our preliminary analysis shows that state-of-the-art data programming is also generally over-confident in its probabilistic labels. High-confidence, mislabeled samples have the potential to negatively impact our suppression accuracy estimates and confidence bounds.

We demonstrate miscalibration in data programming on the low SpO_2 alarm dataset from our case study. Figure 6 shows the average confidence and actual accuracy of high-confidence alarm subsets of the dataset generated via a generative model. For suppressible alarms, we observe over-confidence in the labels of approximately 10% for ϵ less than 0.42, and under-confidence of approximately 2% for ϵ greater than 0.44.

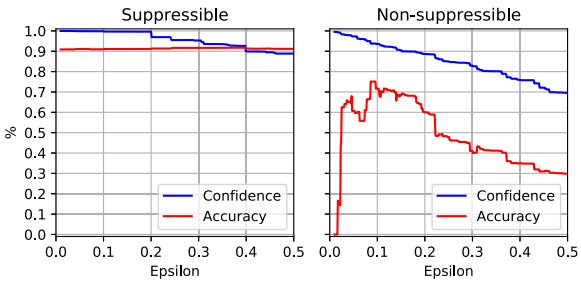


Fig. 6: Average confidence versus accuracy for high-confidence data subsets of varying label uncertainty epsilon.

For non-suppressible alarms, we observe significant over-confidence for all choices of ϵ.

While there exist state-of-the-art methods to calibrate supervised models like neural networks [14], [15], there are no well-established calibration techniques for unsupervised or even weakly-supervised models. We aim to explore this direction in future work.

Appendix B
Theorem and Corollary Proof

We start with some indices $\mathcal{I}$ of datapoints X. We pick some subsets of true-labeled indices and high-confidence indices:

$$\mathcal{I}_j^* \subseteq \{n \in \mathcal{I} \mid y_n = j\}$$
$$\mathcal{I}_j \subseteq \{n \in \mathcal{I} \mid f(x_n) = j \wedge g(x_n) \geq 1 - \epsilon_j\}$$

Recall our assumption of consistent suppression accuracy:

$$\left| \frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{E}\left[S(\tilde{x}_m) = j\right] - \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{E}\left[S(\tilde{x}_n) = j\right] \right| \leq 1 - \eta_j$$

where $\eta_j = \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} g(x_n)$.

Then, let $\delta_j = 1 - \eta_j$, and for any suppression $S : \mathcal{X} \to \mathcal{Y}$, the following holds:

$$\begin{aligned}
&\mathbb{P}\left[\left|R_j(\mathcal{I}_j) - R_j(\mathcal{I}_j^*)\right| \geq c_j\right] = \\
&\mathbb{P}\left[\left|\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(x_m) = j\right) - \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(x_n) = j\right)\right| \geq c_j\right] \\
=&\mathbb{P}\left[\left|\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(x_m) = j\right) - \mathbb{E}\left[\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(\tilde{x}_m) = j\right)\right] + \right. \\
&\mathbb{E}\left[\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(\tilde{x}_m) = j\right)\right] - \mathbb{E}\left[\frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(\tilde{x}_n) = j\right)\right] + \\
&\mathbb{E}\left[\frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(\tilde{x}_n) = j\right)\right] - \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(x_n) = j\right)\Bigg| \\
&\left. \geq (c_j - \delta_j - \gamma_j) + \delta_j + \gamma_j \right] \\
\leq\ &\mathbb{P}\left[\left|\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(x_m) = j\right) - \mathbb{E}\left[\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(\tilde{x}_m) = j\right)\right]\right| \geq c_j - \delta_j - \gamma_j\right] \\
&+ \mathbb{P}\left[\left|\mathbb{E}\left[\frac{1}{|\mathcal{I}_j^*|} \sum_{m \in \mathcal{I}_j^*} \mathbb{1}\left(S(\tilde{x}_m) = j\right)\right] - \mathbb{E}\left[\frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(\tilde{x}_n) = j\right)\right]\right| \geq \delta_j\right] \\
&+ \mathbb{P}\left[\left|\mathbb{E}\left[\frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(\tilde{x}_n) = j\right)\right] - \frac{1}{|\mathcal{I}_j|} \sum_{n \in \mathcal{I}_j} \mathbb{1}\left(S(x_n) = j\right)\right| \geq \gamma_j\right]
\end{aligned}$$

$$= \mathbb{P}\left[\left|\frac{1}{|\mathcal{I}_j^*|}\sum_{m\in\mathcal{I}_j^*}\mathbb{1}\left(S(x_m)=j\right)-\mathbb{E}\left[\frac{1}{|\mathcal{I}_j^*|}\sum_{m\in\mathcal{I}_j^*}\mathbb{1}\left(S(\bar{x}_m)=j\right)\right]\right|\geq c_j-\delta_j-\gamma_j\right]$$
$$+\mathbb{P}\left[\left|\frac{1}{|\mathcal{I}_j|}\sum_{n\in\mathcal{I}_j}\mathbb{1}\left(S(x_n)=j\right)-\mathbb{E}\left[\frac{1}{|\mathcal{I}_j|}\sum_{n\in\mathcal{I}_j}\mathbb{1}\left(S(\bar{x}_n)=j\right)\right]\right|\geq\gamma_j\right]$$
$$\leq 2\exp\left(-2|\mathcal{I}_j^*|(c_j-\delta_j-\gamma_j)^2\right)+2\exp\left(-2|\mathcal{I}_j|\left(\gamma_j\right)^2\right)$$

The steps taken above are justified as follows:

- The first step rewrites the expression based on the definition of R_j.
- The second step equivalently adds and subtracts several expressions.
- The third step uses a triangle inequality:

$$\mathbb{P}\left[|A+B+C|\geq a+b+c\right]\leq$$
$$\mathbb{P}\left[|A|\geq a\right]+\mathbb{P}\left[|B|\geq b\right]+\mathbb{P}\left[|C|\geq c\right]$$

- The fourth step eliminates the second probability due to our consistency assumption.
- The fifth step applies the Hoeffding's inequality twice to sums of Bernoulli variables and their expectations.

For the corollary, we are given a desired confidence p_j:

$$p_j = 2\exp\left(-2|\mathcal{I}_j^*|(c_j-1+\eta_j-\gamma_j)^2\right)+2\exp\left(-2|\mathcal{I}_j|\left(\gamma_j\right)^2\right)$$

We solve the above in terms of c_j, obtaining the expression for the bound size:

$$c_j = 1-\eta_j+\gamma_j+\sqrt{\frac{\ln(2)-\ln(p_j-2\exp(-2|\mathcal{I}_j|\gamma_j^2))}{2|\mathcal{I}_j^*|}}$$

Appendix C
Case Study Labeling Functions

In this section, we describe how the guidelines from our case study are encoded as sixty-two labeling functions.

1) `LF-long-alarm-T` labels non-suppressible if the alarm duration is at least T seconds, otherwise it abstains. LFs 1 to 3 use $T = 60, 65,$ and 70 respectively.
2) `LF-spo2-aboveX-belowY-overT` labels non-suppressible if SpO_2 is in range $(X, Y]$ for longer than T seconds since the alarm start, otherwise it abstains. LFs 4 to 9 use parameter tuples $(X, Y, T) = (80, 85, 120)$, $(0, 80, 120)$, $(70, 80, 100)$, $(60, 70, 90)$, $(50, 60, 60)$, and $(0, 50, 30)$ respectively.
3) `LF-hr-aboveX-overT` labels non-suppressible if heart rate is above X for longer than T seconds, otherwise it abstains. LF 10 uses $X = 220$ and $T = 10$.
4) `LF-hr-aboveX-belowY-overT` labels non-suppressible if heart rate is in range $(X, Y]$ for longer than T seconds, otherwise it abstains. LFs 11 to 14 use parameter tuples $(X, Y, T) = (0, 50, 10)$, $(40 \cdot \alpha, 50 \cdot \alpha, 120)$, $(30 \cdot \alpha, 40 \cdot \alpha, 60)$, and $(0, 30 \cdot \alpha, 0)$ respectively, where α is a scaling age factor taking value of 3.833 for less than one month, 3.766 for one month to less than two month, 3.733 for two month to less than six month, 3.533 for six months and older.
5) `LF-rr-aboveX-belowY-overT` labels non-suppressible if respiratory rate is in range $(X, Y]$ for longer than T seconds, otherwise it abstains. LFs 15 to 18 use parameter tuples $(X, Y, T) = (0, 10, 120)$, $(40 \cdot \alpha, 50 \cdot \alpha, 120)$, $(30 \cdot \alpha, 40 \cdot \alpha, 60)$, and $(0, 30 \cdot \alpha, 0)$ respectively, where α is a scaling age factor taking value of 0.933 for less than one month, 0.9 for one month to less than two month, 0.866 for two month to less than six month, 0.8 for six months and older.
6) `LF-repeat-Xalarms-inT` labels non-suppressible if there has been at least X other low SpO_2 alarms within T seconds of the alarm, otherwise it abstains. LFs 19 to 22 use parameter pairs $(X, T) = (1, 15)$, $(1, 30)$, $(1, 60)$, and $(10, 300)$ respectively.
7) `LF-short-alarm-T` labels suppressible if the alarm duration is at most T seconds, otherwise it abstains. LFs 23 to 25 use $T = 5, 10,$ and 15 respectively.
8) `LF-recoverX-inT` labels suppressible if SpO_2 recovers by more than X points within T seconds of the alarm, otherwise it abstains. LFs 26 and 27 use parameter pairs $(X, T) = (20, 10)$ and $(20, 15)$ respectively.
9) `LF-hr-tech-error-X` labels suppressible if the absolute difference between ECG heart rate and pulse oximeter heart rate is greater than X at the time of alarm, otherwise it abstains. LFs 28 and 29 use $X = 20$ and 30 respectively.
10) `LF-bad-spo2-waveform-X-T` labels suppressible if there exists an outlier with value larger than X within a T seconds window of the alarm in the SpO_2 waveform matrix profile, otherwise it abstains. [3] LFs 30 to 40 use parameter pairs $(X, T) = (8.4, 120)$, $(7.8, 110)$, $(7.2, 100)$, $(6.6, 90)$, $(6.0, 80)$, $(5.3, 70)$, $(4.6, 60)$, $(3.8, 50)$, $(2.9, 40)$, $(2.1, 30)$, and $(1.0, 20)$ respectively.
11) `LF-bad-hr-waveform-X-T` labels suppressible if there exists an outlier with value larger than X within a T second window of the alarm in the heart rate waveform matrix profile, otherwise it abstains. LFs 41 to 51 use $(X, T) = (9.0, 120)$, $(8.5, 110)$, $(7.8, 100)$, $(7.3, 90)$, $(6.7, 80)$, $(6.0, 70)$, $(5.4, 60)$, $(4.7, 50)$, $(3.9, 40)$, $(3.1, 30)$, and $(2.1, 20)$ respectively.
12) `LF-bad-rr-waveform-X-T` labels suppressible if there exists an outlier with value larger than X within a T second window of the alarm in the respiratory rate waveform matrix profile, otherwise it abstains. LFs 52 to 62 use $(X, T) = (8.7, 120)$, $(8.1, 110)$, $(7.6, 100)$, $(7.1, 90)$, $(6.5, 80)$, $(6.0, 70)$, $(5.4, 60)$, $(4.7, 50)$, $(3.9, 40)$, $(3.0, 30)$, and $(2.0, 20)$ respectively.

[3] To find anomalies in time-series (waveform) data we analyze their matrix profiles. At a high-level, a matrix profile represents the dissimilarity between each vital sign measurement in the data and the rest of the data. Hence large values in the matrix profile correspond to outliers. We use the matrixprofile-ts Python library (github.com/matrix-profile-foundation/matrixprofile).