Hengartner *et al. BMC Bioinformatics* 2018, **19**(Suppl 18):485
https://doi.org/10.1186/s12859-018-2503-9

BMC Bioinformatics

RESEARCH Open Access

CAT: computer aided triage improving upon the Bayes risk through ε-refusal triage rules

Nicolas Hengartner[1*], Leticia Cuellar[1], Xiao-Cheng Wu[2], Georgia Tourassi[3], John Qiu[3], Blair Christian[3] and Tanmoy Bhattacharya[1]

From Computational Approaches for Cancer at SC17
Denver, CO, USA. 17 November 2017

Abstract

Background: Manual extraction of information from electronic pathology (epath) reports to populate the Surveillance, Epidemiology, and End Result (SEER) database is labor intensive. Systematizing the data extraction automatically using machine-learning (ML) and natural language processing (NLP) is desirable to reduce the human labor required to populate the SEER database and to improve the timeliness of the data. This enables scaling up registry efficiency and collection of new data elements. To ensure the integrity, quality, and continuity of the SEER data, the misclassification error of ML and NPL algorithms needs to be negligible. Current algorithms fail to achieve the precision of human experts who can bring additional information in their assessments. Differences in registry format and the desire to develop a common information extraction platform further complicate the ML/NLP tasks. The purpose of our study is to develop triage rules to partially automate registry workflow to improve the precision of the auto-extracted information.

Results: This paper presents a mathematical framework to improve the precision of a classifier beyond that of the Bayes classifier by selectively classifying item that are most likely to be correct. This results in a triage rule that only classifies a subset of the item. We characterize the optimal triage rule and demonstrate its usefulness in the problem of classifying cancer site from electronic pathology reports to achieve a desired precision.

Conclusions: From the mathematical formalism, we propose a heuristic estimate for triage rule based on post-processing the soft-max output from standard machine learning algorithms. We show, in test cases, that the triage rule significantly improve the classification accuracy.

Keywords: Machine learning, Classification

Introduction

The Surveillance, Epidemiology, and End Result (SEER) collects and curates cancer patient data from electronic pathology (e-path) reports. Those data are used for population-based research on cancer trends to develop cancer treatment and screening policy recommendations. That information is extracted manually by experts, which is labor intensive. This limits the geographical coverage of the database, currently, about 30% of all cancer cases in the United States are included in the SEER database.

Automating the data extraction using Machine Learning (ML) and Natural Language Processing (NLP) tools is desirable to reduce the human labor that is required to populate the SEER database, thereby increasing the geographical coverage of the database and improving the timeliness of the data. To insure the integrity, quality and general usefulness of that database, the classification algorithm needs to have a misclassification error that is no more than that of human experts. This is challenging because these experts supplement the pathology reports

*Correspondence: nickh@lanl.gov
[1]Los Alamos National Laboratory, PO Box 1663, 87545 Los Alamos, NM, USA
Full list of author information is available at the end of the article

with extra information not always available to the machine learning tools. Thus, it is implausible that, even under the best circumstances, the ML/NLP algorithms can achieve the same small error rate achieved by human experts. Differences in registry format and the desire to develop a common information extraction platform further complicate the ML/NLP tasks.

The ML/NLP task of extracting information from e-path reports can be cast as a multi-class classification problem. The misclassification error of any machine learning classification algorithm is bounded from below by the Bayes risk [1]. In some instances, that lower bound may exceed the accuracy needed by the end user, thereby making machine learning not useful.

The classification depends on both the content and the context of pathology report. This leads to heterogeneity in the degree of difficulty for classifying electronic pathology reports, with some pathology reports being easy to classify while others being harder. If one can identify which pathology reports are easy/hard to classify, i.e., have small/large expected misclassification error, one may improve upon the Bayes error by only classifying automatically the reports that have small expected misclassification error, leaving to the experts the task of classifying the more challenging ones.

We call *triage* machine learning algorithms that selectively classify item. We show in this paper that optimal triage classification rules achieve misclassification error that are lower than the Bayes classifier, at the cost of not classifying a fraction of the items. The lower misclassification rate arises because we do not get penalized for refusing to classify, since these reports will be evaluated by experts. By strategically refusing not to classify a large fraction of the items, we have the opportunity to achieve arbitrarily small misclassification rate. But there are resource constraints that limit the fraction of reports we refuse to classify. For example, ignoring the issue of building an appropriate infrastructure, if we wanted the SEER database to cover 100% of population the USA with the same manpower as now, we could refuse to evaluate at most 30% of the reports.

In this paper, we present the mathematical foundation for Computer Aided Triage (CAT) by showing how it expands on existing concepts of statistical machine learning. The paper is structured as follows: we first formalize the notation and define mathematically the ε-triage rule. We, then, prove that an ε-triage rule has monotonic decreasing classification error with increasing refusal fraction ε and characterize the optimal "Bayes" ε-triage rule, and we relate it to the classical Bayes rule. In the next Section, we apply heuristics derived from these results to post-process deep learning classification of cancerous tumor sites from electronic pathology reports to achieve a desired level of confidence. In a fourth section, we provide another application of this methodology to risk management. Finally, the interested reader will find the proofs in the appendix.

Mathematical formulation

Preliminaries

In this paper, we make use of the following notation. Denote by $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ n independent identically distributed random vectors with joint distribution

$$\mathbb{P}\left[X = dx, Y = a_j\right] = p\left(a_j, x\right) dx = p\left(a_j \middle| x\right) f(x) dx.$$

The covariates $X_i \in \mathbb{X} \subset \mathbb{R}^p$ while the response variable $Y_i \in \mathcal{A} = \{a_1, \ldots, a_m\}$ is one of m labels.

A classifier $\hat{Y}(x)$ is a function from the feature space $\mathbb{X}$ onto the set of labels $\{a_1, \ldots, a_m\}$. To any classifier $\hat{Y}(x)$, we can associate a partition $A_1, \ldots, A_m$ of the feature space $\mathbb{X}$, defined by $A_k = \left\{x \in \mathbb{X} : \widehat{Y}(x) = a_k\right\}$. The partition of the Bayes classifier $Y^\star(X)$, which minimizes the misclassification error $\mathbb{P}\left[\widehat{Y}(X) \neq Y\right]$ amount all measurable functions $\widehat{Y}(X)$, is given by

$$A_k^\star = \left\{x : p\left(a_k | x\right) \geq p\left(a_j | x\right), \quad j \neq k\right\}, \tag{1}$$

with the convention that if there exists two or more indexes for which we have equality, x is assigned to set with the lowest index to ensure that the sets $A_k^\star$ partition the feature space $\mathbb{X}$. See [1] for example.

Optimal triage rule

While the Bayes error is a lower bound for the misclassification error, the optimal triage rule will have a lower misclassification error. Let us formally define the ε-*triage* rule.

Definition 1 *A ε-triage is a function T from the feature space $\mathbb{X}$ into the extended set of labels $\{a_1, \ldots, a_m\} \cup \emptyset$, where the label $\emptyset$ represents the "no classification" category, and*

$$\mathbb{P}[\, T(X) = \emptyset] \leq \varepsilon.$$

We can associate to a triage function $\widehat{T}(x)$ the sets

$$B_k = \left\{x \in \mathbb{X} : \widehat{T}(x) = a_k\right\}.$$

Unlike the sets $A_1, \ldots, A_m$ defined for a classifier, the sets $B_1, \ldots, B_m$ do not form a partition of $\mathbb{X}$. As a result, let us define the decision set $D = \cup_{k=1}^m B_k$, and the rejection set $D^c = \{x : T(x) = \emptyset\}$.

For a triage rule, the misclassification error is only evaluated on the decision set D. That is, a refusal to classify does not get penalized. Formally, the loss function for a triage rule T is

$$L(T) = \mathbb{P}[Y \neq T(X), X \in D]. \tag{2}$$

Note that the loss $L(T)$ can be made arbitrarily small by making D small enough. To avoid this uninteresting answer, we constraint the size of the decision set D to satisfy

$$\mathbb{P}[X \in D] \geq 1 - \varepsilon.$$

Our first theorem characterizes the minimizer of (2), thus providing an analogous result to Bayes rule in the classical classification context.

Theorem 1 *The triage function $T^\star$ that minimizes $\mathbb{P}[\,Y \neq T(X), X \in D]$ subject to $\mathbb{P}[\,X \in D] \geq 1 - \varepsilon$, for a given $0 < \varepsilon < 1$, is characterized by the sets*

$$D^\star = \left\{x : \max_j p\left(a_j \middle| x\right) > b\right\} \tag{3}$$

and

$$B_k^\star = D^\star \cap \left\{p\left(a_k \middle| x\right) \geq \max_{j \neq k} p\left(a_j \middle| x\right)\right\}, \tag{4}$$

where the parameter b is the smallest value such that

$$\mathbb{P}[\,X \in D^\star] \geq 1 - \varepsilon. \tag{5}$$

Remark Observe that the sets $D^\star(b) \subset D^\star(b')$ when $b \geq b'$, or equivalently, when $\varepsilon \leq \varepsilon'$. As a result, the loss of the optimal triage rule $L(T^\star)$ is a monotone decreasing function in ε.

It is insightful to specialize Theorem 1 to characterize the optimal triage rule for binary labeled features.

Corollary 1 *Suppose that $Y \in \{0, 1\}$ takes on only two values. Then the indecision region for the optimal triage rule is*

$$D^c = \left\{x : \frac{1}{b} \leq \frac{p(1|x)}{p(0|x)} \leq b\right\}, \tag{6}$$

where b is the largest value such that $\mathbb{P}[\,X \in D] \geq 1 - \varepsilon$.

A similar "indecision set", based on the likelihood ratio, arises in sequential learning [2].

Relationship to Bayes rule

It is instructive to relate the optimal triage rule to the Bayes rule. To this end, denote by $Y^\star$ the Bayes rule.

Proposition 1 *The optimal triage rule is related to the Bayes classifier through the sets*

$$B_k^\star = D^\star \cap A_k^\star.$$

The proof is immediate by comparing the sets (2) and (4). This description allows us also write the triage rule $T^\star$ in terms of the Bayes classifier $Y^\star$. To this end, define the function

$$Z^\star(x) = \begin{cases} 1 & x \in D^\star \\ \emptyset & x \notin D^\star \end{cases},$$

with the convention that $a_k \times \emptyset = \emptyset$. The optimal triage rule can then be written as

$$T^\star = Z^\star Y^\star. \tag{7}$$

This allows us to reinterpret Theorem 1: The optimal triage rule is the Bayes rule on the restricted set where largest conditional probability $\max_j p(a_j|x) > b$. That is, the optimal triage is Bayes rule provided that the conditional probability of the winning class is large enough. As a consequence, it is possible that a triage rule does never classifies a particular class if the conditional probability $p(a_k|x) < b$ for all x. Identification of such classes is instructive, as it identifies difficult classes. to classify.

Heuristic: using soft-max to build a triage rule

The previous section describes how the optimal triage rule is a thresholding function of the (optimal) Bayes rule. In this section, we propose the heuristic triage estimator obtained by post-processing the soft-max produced by various machine learning algorithms. Looking at Eq. (7), we propose to use a classifier that produces a soft-max to define sets that mimmic the sets $D^\star$ and $B_k^\star$. If the soft-max output from a ML algorithm is a good estimate for some monotone increasing transformation of the actual conditional probabilities, then this heuristic will produce good triage rules. Other estimation strategies are possible, and will be explored in a future manuscript.

Formally, let $\mathcal{H}$ denote a collection of real-valued score functions $h(a, x)$ of the class labels $a \in \mathcal{A}$ and features $x \in \mathbb{X}$. For any function $h \in \mathcal{H}$, the soft-max is defined by

$$q_h(a, x) = \frac{e^{h(a,x)}}{\sum_{b \in \mathcal{A}} e^{h(b,x)}},$$

which can be used to define the classifier

$$\begin{aligned} Y(x) &= \arg\max_k h\left(a_k, x\right) \\ &= \arg\max_k q_h\left(a_k, x\right). \end{aligned}$$

It is well understood while the soft-max can be interpreted as probabilities, they need not be the class conditional probability $p(a_k|x)$, even for consistent classifiers. However, the soft-max of a consistent classifier are such that the sets

$$\widehat{C}_k = \left\{x : q_{\hat{h}}\left(a_k \middle| x\right) \geq \max_{j \neq k} q_{\hat{h}}\left(a_j \middle| x\right)\right\} \tag{8}$$

converge, as the sample size increases to infinity, to $A_k^\star$. However, additional assumptions are needed to ensure that there exists a constant b' such that the set

$$\widehat{D}^\star = \left\{x : \max_k q_{\hat{h}}\left(a_k \middle| x\right) > b'\right\}, \tag{9}$$

asymptotically equals $D^\star$ as the sample size tends to infinity. To this end, consider the following definition:

Definition 2 *A soft-max estimate is said to be M-consistent (or monotone consistent) for the conditional class probability* $p\left(a_k \mid x\right)$ *if*

$$q_{\hat{h},n}\left(a_k, x\right) \longrightarrow_{n\to\infty} \phi\left(p\left(a_k, x\right)\right), \quad \forall j, x, \tag{10}$$

where ϕ *is a monotone increasing function and the convergence is in probability.*

Under that assumption, it is possible to construct, via post-precessing of the soft-max, consistent triage rule from an M-consistent soft-max estimator.

Theorem 2 *Let* $\widehat{Y}(x)$ *is an M-consistent soft-max based classifier. Then the associated triage rule defined by the sets (8) and (9) is consistent and converges to the optimal triage rule.*

Example

A first example

As an illustrative example, we combine deep learning and natural language processing algorithms to classify the primary cancer site, as described by the ICD-O-3 classification manual [3], of 22571 electronic pathology reports from the Louisiana SEER catchment area. The fitted algorithm return a soft-max value for 139 distinct primary cancer site. With minimal optimization, the false positive rate for the classifier is 24.64%, which is significantly higher than the advertised less than 5% classification error from manual classification. This presents us with an opportunity to demonstrate the practical usefulness of triage.

Figure 1 graphs the precision as a function of the percentage of pathology reports classified. First note that a precision of 99% can be achieved by selectively classifying about 20% of the pathology reports, and a performance commensurate to manual classification (95%) is achieved by selectively classifying half of the pathology report. Second, we note that the precision is (essentially) a decreasing function of the percentage classified. This implies that the precision is a monotone increasing function of $\max_j p(a_j|x)$, which is consistent with the M-consistency assumption.

Ad hoc improvement strategy

The distribution of cancer sites in the registry are not balanced. One might expect that cancer sites with more example are more aptly classified. Examination of Fig. 2, that shows the site-specific error rate as a function of sample size, indicates that this is true in some general sense. This suggests the following ad hoc modification to the triage rules considered so far: decline to classify any cancer site that has not sufficient examples to be well characterized. Figure 3 displays the precision for three choices of sample size: 100, 500 and 1000. The resulting triage rules achieve better than 99% precision on by selectively classifying 30% of the pathology reports, and if we only use the most common cancer sites, sites that have more than a 1000 example, the precision is above 99.5%.

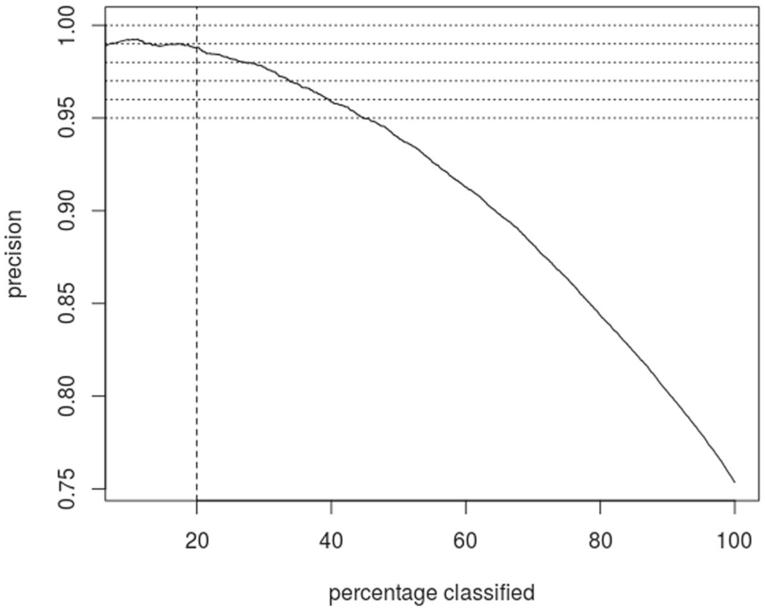

Fig. 1 Precision as a function fraction classified. By selectively classifying the 20% of pathology reports having highest max soft-max, one achieves 99% precision. Classifying pathology reports whose max soft-max exceed the median produces a triage rule that has commensurate precision to human classification

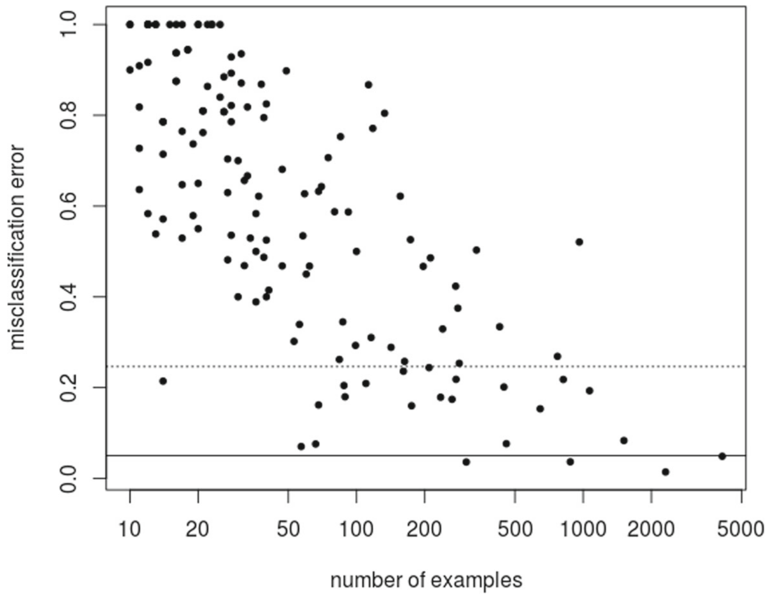

Fig. 2 Site specific error rate as a function of number of examples in the dataset. The error rate for the entire dataset is 24.64% and corresponds to the dotted horizontal line

Conclusion

This paper shows that one can improve a classifiers performance by selectively classifying items, and that the optimal triage rule can be expressed in terms of the Bayes rule. Such triage rule are useful when one seeks machine learning algorithms that achieve a prescribed precision, as is the case in automatic annotation of electronic pathology reports.

Proofs

Proof of theorem 1

The misclassification error of a triage rule T is

$$\begin{aligned} \mathbb{P}[T(X) \neq Y \cap X \in D] &= \mathbb{P}[X \in D] - \mathbb{P}[T(X) \\ &= Y \cap X \in D] = \mathbb{P}[X \in D] \qquad (11) \\ &\quad - \sum_{k=1}^{m} \int_{B_k \cap D} p\,(a_k, x)\, dx. \qquad (12) \end{aligned}$$

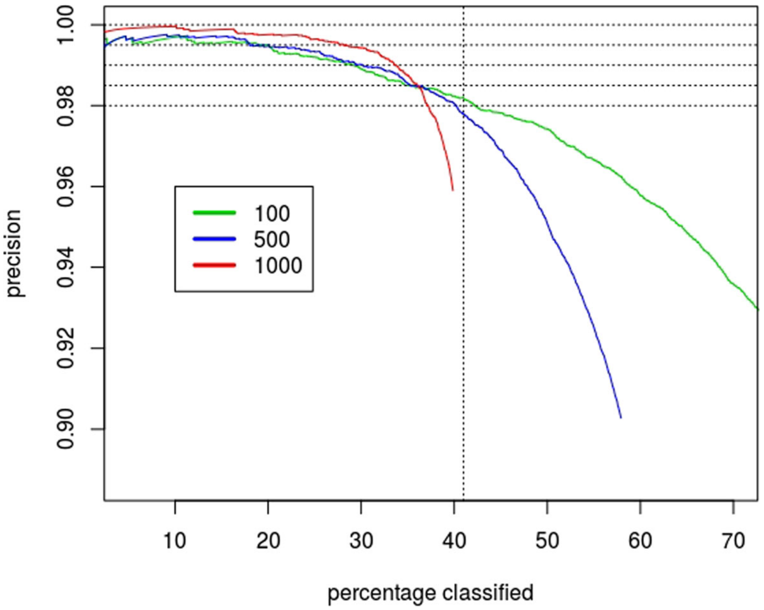

Fig. 3 Precision of triage of cancers that have low misclassification rate

For a fixed decision set D, the misclassification error is minimized by taking

$$B_k = \left\{ x \in D : p\left(a_k, x\right) \geq p\left(a_j, x\right), \quad j \neq k \right\}. \tag{13}$$

To ensure that $B_k \cap B_j = \emptyset$ for $k \neq j$, we apply the convention that if there exists two or more indices for which we have equality, x is assigned to set with the lowest index.

To determine the decision set D that minimizes the misclassification error subject to the constraint that $\mathbb{P}[X \in D] \geq 1 - \varepsilon$, consider all triage rules that satisfy (13)

$$\begin{aligned} \mathbb{P}[T(X) &\neq Y \cap X \in D] = \mathbb{P}[X \in D] \quad &(14)\\ &- \sum_{k=1}^{m} \int_{B_k \cap D} p\left(a_k, x\right) dx \quad &(15)\\ &= \sum_{k=1}^{m} \int_{B_k \cap D} f(x) - \max_j p\left(a_j, x\right) dx \quad &(16)\\ &= \int_D f(x) - \max_j p\left(a_j, x\right) dx \quad &(17) \end{aligned}$$

Set

$$\psi(x) = f(x) - \max_j p\left(a_j, x\right)$$

and proceed as in the Neyman-Pearson lemma [4] to conclude that optimal decision set D is of the form

$$D = \left\{ \frac{\psi(x)}{f(x)} < b \right\}, \tag{18}$$

where the constant b is the smallest value for which the constraint $\mathbb{P}[X \in D] \geq 1 - \varepsilon$ is satisfied. The conclusion follows by expressing the ratio

$$\begin{aligned} \frac{\psi(x)}{f(x)} &= 1 - \max_j \frac{p\left(a_j, x\right)}{f(x)} \quad &(19)\\ &= 1 - \max_j p\left(a_j \middle| x\right) \quad &(20) \end{aligned}$$

Acknowledgements
The authors would like to thank Lynne Penberthy from NIH from bringing to ur attention the SEERS classification problem.

Funding
NH and TB were supported by the National Nuclear Security Administration's Advanced Simulation and Computation program. LC and NH were also supported by CRADA with Ernst and Young. This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Los Alamos National Laboratory under Contract DE-AC5206NA25396 and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. Publication costs were funded by the JDACS4C program.

Availability of data and materials
The raw epath reports were accessed as parts of a Data Use Agreement.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 19 Supplement 18, 2018: Selected Articles from the Computational Approaches for Cancer at SC17 workshop. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-18.

Disclaimer
This report was prepared as an account of work sponsored by an agency of the U.S. Government. Neither Los Alamos National Security, LLC, the U.S. Government nor any agency thereof, nor any of their employees make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by Los Alamos National Security, LLC, the U.S. Government, or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of Los Alamos National Security, LLC, the U.S. Government, or any agency thereof. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Authors' contributions
NH and TB conceived of this project; NH, LC, GT, JQ, BC, and TB carried out the research; GT, X-CW, and TB made the data available. NH wrote the paper with TB, all authors agreed to the text. All authors read and approved the final manuscript.

Ethics approval and consent to participate
The data was obtained from the Louisiana registry under data use agreements, and the research was approved by the institutional review boards.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Los Alamos National Laboratory, PO Box 1663, 87545 Los Alamos, NM, USA. [2]Louisiana State University, 2020 Gravier Street, 3rd Floor, 70112 New Orleans, LA, USA. [3]Oak Ridge National Laboratory, PO Box 2008, 37831 Oak Ridge, TN, USA.

Published: 21 December 2018

References

1. Devroye L, Gyorfi L, Lugosi G. A Probabilitic Theory of Pattern Recognition. New York: Springer; 1996.
2. Wald A. Sequential Analysis. New York: Wiley; 1947.
3. Organization WH. International Classification of Diseases for Oncology, Third edition, first revision. Geneva: WHO press; 2013.
4. Neyman J, Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. In: Philosophical Transactions of the Royal Society of London. Series A, Vol. 231. 1993. p. 289–337.