Surrogate-guided sampling designs for classification of rare outcomes from electronic medical records data

W. KATHERINE TAN*, PATRICK J. HEAGERTY

Abstract

Scalable and accurate identification of specific clinical outcomes has been enabled by machine-learning applied to electronic medical record (EMR) systems. The development of automatic classification requires the collection of a complete labeled data set, where true clinical outcomes are obtained by human expert manual review. For example, the development of natural language processing algorithms requires the abstraction of clinical text data to obtain outcome information necessary for training models. However, if the outcome is rare then simple random sampling results in very few cases and insufficient information to develop accurate classifiers. Since large scale detailed abstraction is often expensive, time-consuming, and not feasible, more efficient strategies are needed. Under such resource constrained settings, we propose a class of enrichment sampling designs, where selection for abstraction is stratified by auxiliary variables related to the true outcome of interest. Stratified sampling on highly specific variables results in targeted samples that are more enriched with cases, which we show translates to increased model discrimination and better statistical learning performance. We provide mathematical details, and simulation evidence that links sampling designs to their resulting prediction model performance. We discuss the impact of our proposed sampling on both model development and validation. Finally, we illustrate the proposed

*To whom correspondence should be addressed at wlktan [at] uw [dot] edu.
designs for outcome label collection and subsequent machine-learning, using radiology report text data from the Lumbar Imaging with Reporting of Epidemiology (LIRE) study.

1 Introduction

Linked electronic medical record (EMR) systems provide a massive reservoir of information that can help researchers understand and treat both common and rare medical conditions. Specifically, EMR data includes both structured data, such as lab values and diagnostic codes, and unstructured data in the form of free-text medical notes and images. The majority of EMR data is natively captured in unstructured form, and this has fostered the development of learning algorithms to extract research ready variables (Pons et al. [2016], Wang et al. [2018]). Both structured and unstructured data may be used towards identification of key clinical indicators or outcomes, where accurately-derived outcomes can greatly improve downstream analyses such as the generation of prognostic models for disease risk, or predictive scores that could guide treatment. Ultimately both structured data and carefully processed unstructured data are necessary, but extracting specific findings from unstructured data is often expensive and time-consuming. The traditional approach requires highly trained clinicians or technicians who can transcribe medical notes into coded variables. The manual abstraction process is not scalable to massive EMR cohorts and has motivated using machine-learning methods, such as natural language processing (NLP) methods for medical text data (Chapman et al. [2001], Carroll et al. [2012]), and deep learning strategies for medical images (Esteva et al. [2017]), as scalable alternatives. Yet, any algorithm development relies on a base of training and validation data, and the purpose of this manuscript is to outline efficient study designs that can facilitate cost-effective data collection for the development of new prediction tools.

In order to both develop (i.e. train) and validate (i.e. test) data-driven machine-learning
algorithms requires “labeled data”, which is a sample containing both feature (predictor) and outcome information. For example, classification of imaging findings from radiology report text may use word indicators as features and clinician-defined actual case/control statuses as binary outcomes. In typical EMR settings, labeled data is not readily available, therefore a small subset of the underlying cohort needs to be selected for outcome abstraction. The information content in labeled samples is crucial towards efficient and accurate machine-learning modeling. For classification tasks, a well-known challenge for training is statistical rarity, where cases (outcome=1) are disproportionately less frequent than controls (outcome=0), and the fact that the training sample outcome class distribution may affect classification accuracy has been demonstrated both empirically [Weiss and Provost 2001, Batista et al. 2004, Wei and Dunbrack Jr 2013] and theoretically (Xue and Hall 2015). Unsurprisingly, such class distributional “imbalance” is almost always observed for clinical outcomes. One approach in the machine-learning literature that has been proposed to address statistical rarity involves re-sampling the training sample to eliminate controls (under-sampling) or replicating cases (over-sampling), in order to re-balance the effective outcome class distribution in training, and hopefully to improve ultimate model prediction accuracy (Chawla et al. 2002, He and Garcia 2009). However, such analysis-based re-sampling procedures assume that an initial labeled data sample is already available, and these strategies disregard the potential cost associated with labeled data collection (Weiss and Provost 2001).

When data collection resources are scarce, targeted sampling methods in epidemiology have offered highly efficient research designs. In contrast to analysis-based re-sampling procedures, epidemiologic sampling methods are defined at the design stage of studies prior to data collection. One well-known sampling method is the case-control design (Prentice and Pyke 1979), where collected samples may be appropriately analyzed using a logistic regression model, and has the attractive advantage that estimation proceeds as if a simple random
sample were collected (although the regression intercept is biased). Sampling designs such as the case-control, where expensive data ascertainment is based on strata defined by values of a cheaper auxiliary variable, may be viewed as special cases of the general two-phase sampling design (Neyman [1934], Chatterjee et al. [2003]). In the context of effect estimation, targeted sampling through two-phase has been shown to provide efficiency over simple random sampling (Zhao et al. [2009], McIsaac and Cook [2014]), especially when using sampling variables that are highly correlated and informative for the outcome (Zhao et al. [2012]). However, the effect of selectively sampled training data on ultimate machine-learning prediction accuracy has not been thoroughly investigated.

For clinical outcome identification using EMR data, an imperfect alternative to abstracted outcomes may be based on summaries of related structured data elements, such as International Classification of Disease (ICD) codes and simple keyword searches queried within pre-specified time frames. Such “surrogates” or “correlates” of actual clinical outcomes have been used in place of true clinical outcomes in machine-learning modeling tasks to reduce the dimensionality of EMR-generated features (Yu et al. [2016]), or directly as “noisy” imputed outcome labels for classifier development (Agarwal et al. [2016]). However, model development with misclassified outcomes may seriously compromise validity of using the resulting model predictions for downstream analyses (Sinnott et al. [2014]), therefore using surrogates to replace abstracted clinical outcomes as labels may not be justified. Alternatively, surrogates could help guide selection of subjects for labeled data abstraction. In fact, subject selection based on non-negated keywords and ICD codes has been described when assembling a labeled data sample for machine-learning classification of clinical outcomes (Pakhomov et al. [2005]). Yet, there remains little discussion of corresponding statistical rationale, and such heuristic decisions based on purposeful biased sampling may not create generalizable predictions, or valid summaries of accuracy.
This paper is motivated by the need for a formal statistical framework to guide sampling of subjects for labeled data abstraction, towards accurate and scalable machine-learning classification of clinical outcomes. We specifically focus on the rare outcome scenario, where model accuracy is often rate-limited by the number of outcome cases. As with conventional intuition, our proposed strategy targets case-enrichment of rare outcomes for selection of training data. The key contribution of our work is the formalization of heuristic sampling methodologies drawn from the fields of machine-learning and epidemiology, therefore filling a critical gap in EMR research methods. In Section 2.1, we frame the statistical problem and describe the proposed sampling framework. Then, in Section 2.2 we provide a representation of development sample composition on model discrimination and demonstrate direct connections with statistical efficiency. In Section 2.3, we describe configurations within the class of proposed designs that are best for learning, and characterize the sampling impact for both model development and validation in Section 2.4. We provide empirical evidence through simulations in Section 3. Finally, in Section 4 we illustrate the method on a data set of lumbar spine imaging reports that was obtained in a pragmatic trial of radiology decision support (Jarvik et al. [2015]), and provide a concluding discussion in Section 5.

2 Methods

2.1 Statistical motivation and proposed design

For subject $i$ denote $\tilde{X}_i \in \mathbb{R}^p$ as the feature vector and $Y_i \in \{0, 1\}$ as the binary outcome. The general classification problem is to find function $h(.)$ that maps from the features to outcomes, for example penalized logistic regression

$$\hat{\beta}_0, \hat{\beta}_X = \min_{\beta_0, \beta_X} \{- \sum_{i=1}^{n} Y_i (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) + \log(1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})) + \lambda \sum_{j=1}^{p} \|\beta_j\|_L \}.$$  (1)
where in (1), $L = 1$ refers to Lasso regression \cite{Tibshirani1996} and $L = 2$ refers to Ridge regression \cite{LeCessieVanHouwelingen1992}. For a concrete example of application of classification models such as (1), consider the task of classifying radiology reports for subject vertebral fracture status. For this natural language processing (NLP) motivated task, simple features $\tilde{X}_i^T$ may be derived using bag-of-words (BOW) representations, while outcomes $Y_i$ obtained through abstraction. More sophisticated feature engineering is common in NLP, but for the characterization of efficient designs we illustrate using a simple learning approach. For subject $i$, the BOW feature vector $\tilde{X}_i$ has binary elements $X_{ij} = I(t_j \in \text{report}_i)$ with unique terms $t_j$ obtained by concatenating all reports, while the abstracted outcome label $Y_i$ is the clinician-defined indicator of vertebral fracture. For data required towards fitting (1), note that while extracting features is relatively cheap, obtaining outcome statuses is time-consuming. Therefore, for clinical outcome identification tasks, a sample needs to be drawn from the EMR cohort $\mathcal{D}$ for outcome abstraction so that both $\tilde{X}_i$ and $Y_i$ are available for machine-learning development and evaluation.

Denote the sample $\mathcal{D}^S(n)$, having sample size $n$ and sampling design $S$. Typically, the assumption on $S$ is simple random sampling (SRS) from $\mathcal{D}$. However, due to the expected low number of cases from naturally rare clinical outcomes, using SRS to select reports for abstraction is often inadequate. In machine-learning a common procedure is oversampling to artificially increase sample prevalence where cases are randomly replicated at the analysis stage. However, oversampling does not generate new information, rather simply re-weighs existing data.

Alternatively, samples with higher outcome prevalence can be collected by design through stratified sampling. For the development of outcome classification algorithms using EMR data, recall that while true $Y_i$ requires abstraction from unstructured data, there exists other structured data elements in the database that are related to $Y_i$. For instance, true outcome
status for the example of vertebral fracture identification may be related to counts of related keywords in report text, or related to International Classification of Disease (ICD) codes recorded during the same subject visit. Denote summaries of such related structured data elements as $Z_i$, where it is reasonable to expect $Z_i$ to be associated with true $Y_i$, and therefore a “surrogate for the true outcome. However, due to the potential misclassification of $Z_i$ for $Y_i$, instead of replacing $Y_i$ by $Z_i$ we suggest using $Z_i$ as auxiliary variables for sampling to develop a prediction model.

**Definition 1** Surrogate-guided sampling (SGS) design class.

Denote the surrogate-guided sampling (SGS) design class as the set of stratified sampling procedures based only on values of a binary enrichment surrogate $Z \in \{0, 1\}$. Such designs would select an individual for sampling with probability $\pi(Z_i)$ where typically $\pi(Z_i = 1) > \pi(Z_i = 0)$ when $Z$ is positively correlated with $Y$.

The surrogate-guided sampling (SGS) design class (Definition 1) describes the class of stratified sampling designs based on values of an enrichment surrogate, and is a special case of two-phase sampling. To conduct an SGS design, all subjects in the cohort are divided into two strata based on surrogate values: surrogate positives with $Z_i = 1$, and surrogate negatives with $Z_i = 0$. Then, subjects are selected into the sample based on surrogate values, and only selected subjects have true $Y_i$ abstracted for. The intended benefit of SGS designs is that, for the same abstraction cost, resulting samples have higher expected outcome prevalences compared to using SRS. For illustration, consider an outcome prevalence of 10%, and assume that in the EMR, there exists a surrogate with 40% sensitivity and 95% specificity for the outcome of interest. For an abstraction budget of collecting $n = 500$ labels, using an SGS design with three-times as many surrogate positives as surrogate negatives yields about 185 true cases in expectation. In contrast, an SRS design would have required abstraction of almost 1850 subjects to yield 185 actual cases, abstraction burden of close to four times. Note that cases identified using SGS designs are true cases collected from the cohort, and
not replicates or synthetic data as resulting from using analysis-based re-balancing methods.

We now describe theoretical and analytical results that are key to understanding our proposed sampling framework. First, in order to demonstrate that the sampling design choice does affect learning performance, we provide a mathematical representation of how training sample composition affects prediction accuracy. Second, among the possible designs in the proposed SGS sampling class, we investigate configurations that improve sampling benefit for learning. Intuitively as illustrated in Figure 1 when the case proportion is higher in the surrogate positive \((Z = 1)\) compared to the surrogate negative \((Z = 0)\) stratum, over-representing the surrogate positive stratum provides higher expected sample case proportion by construction - we provide an analytical treatment of such intuition. Lastly, we characterize the impact of using the proposed SGS design on machine-learning performance, describing bias and modeling considerations for both model development and model validation.

2.2 Effect of training sample composition on prediction accuracy

To statistically motivate that sampling design choice does affect learning performance, we provide a mathematical representation of how training sample composition impacts prediction accuracy. For tractability we focus on a commonly used evaluation metric, the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). Model validation AUC can be interpreted as how well resulting continuous predictions discriminate between randomly selected pairs of case and control subjects in yet unseen data. Other performance metrics such as binary accuracy correspond to the sum of error values for an optimal point on the ROC curve. If we consider continuous model predictions as a “test” for true outcome statuses, then under a bi-normal assumption, Pepe 2003 has shown the AUC to be

\[
AUC = \Phi(\sqrt{R_{AUC}}) = \Phi\left(\sqrt{\frac{\left(\mu_1 - \mu_0\right)^2}{\sigma_1^2 + \sigma_0^2}}\right).
\]  

(2)
In (2), $\mu_y$ and $\sigma^2_y$ are the means and variances of the “test” among the cases ($y = 1$) and controls ($y = 0$). The bi-normal AUC formula (2) was developed in [Pepe 2003] for diagnostic testing applications, but may be generalized to the classification modeling setting. For classification model development, the continuous “test” is estimated using a development sample, but generalizable performance usually evaluated on a separate validation sample. Therefore, we introduce additional notation to express such differences between the classification modeling and diagnostic testing settings. Denote $D^S(n)$ as the development sample collected using sampling design $S$ and having sample size $n$, and assume that the validation sample is a large sample obtained through SRS from $D$. Then, the validation AUC for model developed with $D^S(n)$ may be represented using an indexing as shown in Definition 2.

**Definition 2** $AUC(Y|D^S(n))$.

Let $AUC(Y|D^S(n))$ denote the validation AUC of a classification model for outcome $Y$ developed using sample $D^S(n)$ defined with sampling design $S$ and sample size $n$.

Using the indexing as in Definition 2, we may then represent validation AUC in terms of development sample composition, assuming bi-normally distributed features ($(X|Y = y) \sim N(\mu_{x|y}, \Sigma_{x|y})$). Theorem 1 shows that validation AUC is inversely proportional to the estimation variance and the data signal-to-noise ratio. Therefore, assuming use of the same modeling procedure, using a design with higher statistical information as measured by lower estimation variance results in higher validation AUC. To our knowledge, the results in Theorem 1 are the first to directly present an indexing of validation AUC in terms of development sample composition. Note that our argument for Theorem 1 may be generalized beyond bi-normal features and logistic regression. For example, the bi-normal features assumption may be relaxed to allow for monotone transformations of normal distributions (Pepe [2003]). In addition, the results in Theorem 1 may be applied to penalized logistic regression, as long as the estimation bias and variance of resulting coefficients can be well characterized.
Theorem 1 Assume that in $\mathcal{D}$, for $y \in \{0, 1\}$, $(X|Y = y) \sim N(\mu_{x|y}, \Sigma_{x|y})$, where $\mu_{x|y} = 0$ and $\Sigma_{x|y} = \Sigma_{x|y} = \Sigma_{x|y}$. Let $\hat{\eta} = X \hat{\beta}$ be the estimated linear predictor, where model coefficients $\hat{\beta}$ are estimated by logistic regression using development sample $D^S(n)$. Then,

$$AUC(Y|D^S(n)) \propto \frac{1}{\text{trace}(\Sigma_V(\hat{\beta}^S(n))) + \mu^T \Sigma_V(\hat{\beta}^S(n)) \mu},$$

(3)

where $\Sigma_V(\hat{\beta}^S(n)) = (X^T W X)^{-1}$ is the approximate covariance matrix of estimating $\hat{\beta}$ using $D^S(n)$, and $\mu = \mu_{x|y}$ and $\sigma = \Sigma_{x|y}$ are parameters describing the data signal-to-noise ratio.

We may use the results in Theorem 1 to explain the effect of outcome class imbalance on classifier discrimination. When modeling using logistic regression, it has been noted that samples with rare outcomes tend to result in more highly variable coefficient estimates compared to that of more prevalent outcomes (King and Zeng [2001]). As demonstrated in (3), such increased estimation variance directly results in lower discrimination. To increase discrimination in validation requires using an alternative sampling design $S$ that results in lower variability in estimating $\hat{\beta}$, a higher information design. In the context of two-phase sampling designs, the choice of sampling variable as well as strata proportions may affect design information. We now turn to discussion on this point.

2.3 Quantifying SGS design benefit with $O_{ratio}$. 

For a parametric regression model we demonstrated that the development sample composition affects ultimate model validation AUC through the variance associated with the estimation of risk scores. In general, to directly quantify the effect of sampling design on estimation variance requires numerical approximations. Alternatively, motivated by empirical results in machine-learning, we turn to sample outcome prevalence as another simple measure of information. We now focus on characterizing the “effect” of a sampling design on the sample outcome prevalence. Let $S_i = 1$ denote that subject $i$ was selected for sam-
pling using design $D^S(n)$. The sample case/control odds, $\frac{E[Y|S=1]}{1-E[Y|S=1]}$, compares the expected proportion of cases to controls among sampled subjects ($S = 1$), where higher odds indicate higher prevalence. To denote the sample case enrichment comparing SGS to SRS, we propose using the case/control odds ratio, a metric we denote as $O_{ratio}$ and mathematically define in Definition 3.

**Definition 3** $O_{ratio}$.

Let $O_{ratio}$ denote the expected case/control odds ratio comparing surrogate-guided sampling (SGS) to simple random sampling (SRS), where $O_{ratio} = \frac{E[D^{SGS}(n)|Y|S=1]}{1-E[D^{SGS}(n)|Y|S=1]} / \frac{E[D^{SRS}(n)|Y|S=1]}{1-E[D^{SRS}(n)|Y|S=1]} = \frac{Odds(cases|SGS)}{Odds(cases|SRS)}$.

The denominator of $O_{ratio}$ is the expected odds of cases for samples collected with SRS, and is assumed to be less than 1 for rare outcomes. The numerator is the expected odds of cases for samples collected with SGS designs. Therefore, $O_{ratio}$ can be interpreted as the expected increase in cases comparing SGS to SRS, with higher values indicating that SGS provides more case enrichment, and $O_{ratio} > 1$ indicating improvement using SGS relative to SRS. $O_{ratio}$ has similarities and differences to the term “odds ratio” which is often used in epidemiology. The epidemiological usage of “odds ratio” compares the case/control odds of a sample drawn from the exposed group to a sample drawn from the unexposed group, and provides a single estimate of exposure effect. Similar to the exposure odds ratio, $O_{ratio}$ also compares the case/control odds of two samples drawn from the same population. However, since the samples are defined by sampling design instead of exposure statuses, the $O_{ratio}$ provides a single estimate of design effect on sample outcome prevalence. Therefore, $O_{ratio}$ provides a one-dimensional summary measure of case enrichment comparing SGS over SRS.

**2.3.1 Properties of $O_{ratio}$ and the impact of surrogate specificity**

An interesting property of $O_{ratio}$ is the connection to Likelihood Ratios (LRs) of the enrichment surrogate. Of note, LRs of a diagnostic test can be interpreted as slopes of Receiver
Operating Characteristics (ROC) curves, are related to positive and negative predictive values (PPV & NPV), but are invariant to outcome prevalence ([Choi] [1998]). Therefore, by framing enrichment surrogates Z as “prior tests” of outcome Y, we may gain insight into what types of variables are the best surrogates for sampling.

Proposition 1 Properties of O\textsubscript{ratio}.

Let a surrogate-guided sampling (SGS) design of sample size n be defined with surrogate Z and sampling ratio \( R = P(Z = 1|S = 1) \), where Z has \( p_Z := P(Z = 1) \) and operating characteristics: \( Z\text{sens} := P(Z = 1|Y = 1) \), \( Z\text{spec} = P(Z = 0|Y = 0) \). Then,

\[
O_{ratio}(n, R, Z) = \frac{RZ\text{sens} + p_Z(1 - R - Z\text{sens})}{R(1 - Z\text{spec}) + p_Z(Z\text{spec} - R)}.
\]

(4)

Additionally, if the outcome is rare \( (P(Y = 1) \approx 0) \), then

\[
O_{ratio}(n, R, Z) \approx (R)(LR+) + (1 - R)(LR-).
\]

(5)

Corollary 1 For a given Z, \( O_{ratio} \propto R \). Over the set of possible Z, \( O_{ratio} \propto Z\text{sens} \) and \( O_{ratio} \propto \frac{1}{1 - Z\text{spec}} \).

Equation (5) in Proposition 1 shows that \( O_{ratio} \) is approximately the sum of positive and negative surrogate likelihood ratios \( LR+ = \frac{Z\text{sens}}{1 - Z\text{spec}} \) and \( LR- = \frac{1 - Z\text{sens}}{Z\text{spec}} \), weighted by the sample proportion of surrogate positives. Corollary 1 follows directly from (5), and directly characterizes the effect of both strata allocation and sampling variable choice on sample case enrichment. For strata allocation, over-representing surrogate positives (higher R) results in higher values of \( O_{ratio} \). In terms of sampling variable choice, notice that for any given R, while the rate of increase in \( O_{ratio} \) is linear in \( Z\text{sens} \) it is inverse polynomial in \( 1 - Z\text{spec} \).
Therefore, a small change in specificity can have a much higher impact on $O_{ratio}$ compared to the same change in sensitivity. Another perspective on sampling variable choice may be obtained by translating sensitivities and specificities into likelihood ratios. In general, while a “good” test requires having high values of both $LR^+$ and $LR^-$, having a high $LR^+$ alone is sufficient to achieve a high $O_{ratio}$. Requirements for a variable to be a good enrichment surrogate for sampling are weaker than requirements for a good diagnostic test.

The illustration in Figure 2 emphasizes the impact of surrogate specificity on $O_{ratio}$, where values of $O_{ratio}$ are indicated by different colors across possible ranges of surrogate marginal sensitivities and specificities for an SGS design with fixed $R = 0.50$. The non-gray regions of Figure 2 illustrates operating characteristics of surrogates that constitute good candidates for stratified sampling variables. We excluded the presentation of surrogates with specificities less than 0.50, as we may redefine these surrogates to obtain a more specific variable. From Figure 2, note that when using surrogates with specificities of 0.80 or higher, case-enrichment relative to SRS can be expected even with sensitivities as low as 0.20. Our mathematical analyses convey two important practical implications. First, if there exists a dichotomous variable in the EMR that predicts the outcome better than random noise, stratified sampling based on such a variable can provide a development sample that is more enriched for cases, for the same abstraction cost of a simple random sample. Second, to improve on case enrichment, optimizing the enrichment surrogate for high specificity provides much more value compared to optimizing for high sensitivity. By stratified sampling on the values of an enrichment surrogate that is highly specific for the outcome of interest, SGS designs result in development samples with higher outcome prevalence, which may correspond to increased statistical information, lower estimation variance, and therefore improved statistical learning.
2.4 Design impact on model development and model validation

To improve the information of samples selected for machine-learning, SGS designs intentionally over-represent surrogate positives. A natural concern is whether such introduced selection bias may impact the validity of developed models. The impact of sample characteristics on machine-learning was first formalized in [Zadrozny 2004], and can be formulated as a missing data problem (Little and Rubin 2014). Recall that sampling in SGS only depends on surrogate values $Z$, which are assumed to be available for all subjects in $D$. Therefore, for the SGS design, sampling is independent of outcome labels conditional on surrogate values, equivalently $S \perp Y|Z$, an assumption also known as Missing At Random (MAR). Using the MAR assumption, we now describe the impact of using SGS designs for both model development and model validation.

2.4.1 Design impact on model development

To characterize design impact on model development, we consider distributional differences between the development sample and the cohort. For sample $D^S(n)$ obtained with sampling design $S$, [Zadrozny 2004] suggested that $S$ may be used for developing model $\hat{h}(.)$ “validly” under the asymptotic equivalence criteria, with $\lim_{n \to \infty} \hat{h}(D^S(n)) = h(D)$, where as the development sample size $n$ grows, the $\hat{h}(.)$ approaches the truth $h(.)$ as if the full cohort were available. In particular, $S$ resulting in $D^S(n)$ having outcomes MAR from $D$ are “valid” for model development of classifiers based on conditional means in the asymptotic “true model” sense ([Zadrozny 2004]). Note that for logistic regression, $\text{logit}(E[Y|X, Z, S = 1]) = \text{logit}(E[Y|X, Z])$, therefore, machine-learning model development with logistic regression using SGS samples will result in validly estimated models under this interpretation.

2.4.2 Design impact on model validation

Now, consider the impact of using sample $D^S(n)$, where $S$ is the SGS design, on model validation to assess prediction performance. This practically relevant scenario may arise, for
example, when a single sampling design is used to select subjects for outcome abstraction, and then resulting sample split into separate sub-samples for model development and model validation. On the validation sample, the developed model may be assessed for its prediction accuracy, using metrics such as sensitivity, specificity, and AUC.

In general, unless the validation sample is drawn randomly from the cohort (i.e. SRS), empirically estimated accuracy metrics are typically biased for the true values. However, for validation samples collected using SGS, due to the MAR assumption bias-correction methods are available. For example, the Inverse Probability Weighting (IPW) estimator first proposed by [Horvitz and Thompson 1952](#) adjusts empirical estimates according to inverse sampling probabilities. To estimate generalizable AUC of the model on this intentionally biased sample, for pairs of subjects \(i\) and \(j\), outcome \(Y\) and predicted probabilities \(\hat{p}\), the IPW-corrected empirical estimator is

\[
AUC_{IPW} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{-1} \pi_j^{-1} I(\hat{p}_i > \hat{p}_j)I(Y_i > Y_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{-1} \pi_j^{-1} I(Y_i > Y_j)}.
\]

In (6), \(\pi_i = P(S_i = 1)\) is the sampling probability for subject \(i\), and may be estimated from observed data for any MAR sample. For the SGS design, \(\pi_i\) is additionally known by construction to be

\[
\pi_i = P(S_i = 1|Z_i = z) = \frac{P(Z_i = z|S_i = 1)P(S_i = 1)}{P(Z_i = z)} = \begin{cases} \frac{R}{p_z} \times \frac{n}{N}, & Z_i = 1 \\ \frac{1-R}{1-p_z} \times \frac{n}{N}, & Z_i = 0. \end{cases}
\]
The known sampling probabilities (7) may be directly used in IPW-corrected accuracy metrics such as (6). Note that the AUC indexing described in Section 2.3 is slightly different than the AUC estimator in (6). In Section 2.2, we assumed that the validation sample was large and representative of the cohort, and represented the effect of development sample composition on validation AUC. Here, in using (6), we considered the developed model to be fixed, and studied the effect of validation sample composition towards unbiased estimation of true model accuracy measures of this fixed model on the target cohort. Our theoretical arguments demonstrate that any introduced bias from using SGS samples for model validation may be corrected with IPW towards unbiased estimation of model accuracy measures.

2.4.3 Theoretical requirements for design validity

By framing the proposed sampling design as a missing data problem, we have characterized sampling impact on modeling and outlined several analytic guidelines for design validity. For model development of classifiers based on conditional outcome distributions, the surrogate Z needs to be included as a predictor. For model validation, empirical accuracy measures may be corrected using IPW estimators, where required sampling probabilities are known exactly by design. For both model development and model validation, subjects representing surrogate positives (Z=1) and surrogate negatives are (Z=0) are required in the sample. For example, if only surrogate positives are available, model coefficients are estimable only on the Z = 1 stratum and \( \pi_i \) is undefined for the Z = 0 stratum, without further parametric assumptions.

3 Simulations

To illustrate the benefit of using SGS designs for statistical machine-learning model development, we conducted simulations motivated by a real-world data set of radiology text reports from the Lumbar Imaging with Reporting of Epidemiology (LIRE) study (Jarvik et al. [2015]).
- for additional information see Section 4. For the simulation study, features were generated following a long-tail distribution that is characteristic of bag-of-words text data, modeling was based on penalized regression due to high-dimensional feature assumptions, and we compared the effect of various sampling designs on validation prediction accuracy.

### 3.1 Simulation set-up

For cohorts of size $N = 100,000$, we generated conditional outcomes as independent Bernoulli random variables, having prevalence of either 5% or 10%. As most EMR datasets contain features of high dimensionality, we set the number of features to be $p = 250$, of which only 30 had non-zero coefficients. Specifically, the conditional outcome was generated as $Y_i \mid (Z_1, Z_2, \tilde{X}_i^T) \sim \text{Bernoulli}(P(Y_i = 1))$, where $\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_{Z_1} Z_1 + \beta_{Z_2} Z_2 + \sum_{j=1}^{p} \beta_j X_{ij}$, with $\beta_j = (-0.75, -0.5, 0.25, \ldots, -0.5, 0.25)$ for the first 20 most frequent features, $\beta_j = 1$ for the 10 features with frequencies closest to the outcome prevalence, and $\beta_j = 0$ for the remaining 220 features. Here, we used a simplifying assumption that the most predictive text-based features tend to occur as often as the outcome prevalence, frequent features are weakly predictive, but most features are irrelevant for predicting the outcome.

Binary features were generated as independent Bernoulli random variables, with marginal feature frequencies following an exponential distribution simulating a long upper tail distribution, where the most common features are present in almost all reports but the majority of features have very low frequencies (Sichel [1975]). Specifically, features were generated as $\tilde{X}_j \sim \text{Bernoulli}(p_{\tilde{x}_j})$, where $p_{\tilde{x}_j}$ simulated following an exponential distribution with mean $= \frac{1}{6}$ comparable to observed distributions in the LIRE dataset. For the binary enrichment surrogates, surrogate $Z1$ had a sensitivity of 0.40 and a specificity of 0.95, defined to have comparable operating characteristics with the real-world surrogate for the LIRE data set, while surrogate $Z2$ had a sensitivity of 0.67 and a specificity of 0.66, and may be viewed as a “weaker” surrogate for sampling. Note that both surrogates have the same discrimination
for the outcome (AUC = 0.67) as computed according to the trapezoidal rule.

We compared the sampling methods: simple random sampling (SRS) which we consider to be the “baseline”, surrogate-guided sampling designs (SGS), as well as random over-sampling (ROS) which is a commonly used analysis-based re-sampling procedure. For each simulated cohort, we set aside a large validation sample with sample size $n_{val} = 10000$ using SRS. From the remaining subjects, we simulated “abstraction samples” varying across a grid of sample sizes, and sampling methods of SRS, ROS, SGS 1:1 or SGS 3:1, where SGS may be based on surrogates $Z_1$ or $Z_2$. For the SRS and SGS sampling designs, the abstraction sample size is exactly the development sample size. The ROS procedure replicates cases from an SRS sample of size $n$ until the number of cases and controls are equal. Therefore, even though both SRS and ROS have the same “abstraction sample size”, ROS results in a higher development sample size due to case replication. For a fair comparison, we used abstraction sample size rather than development sample size as the unit of cost measurement.

For each iterations we fit either Lasso or Ridge classification models, but coefficients for enrichment surrogates were assigned a zero penalty, which is a modification to the usual likelihood so that the surrogate is always included in the resulting model. Regularization parameters were selected based on values that maximized AUC using ten-fold cross-validation on development samples. Then, we apply resulting model estimates to the validation sample, calculating the empirical validation AUC using the Wilcoxon-Mann-Whitney formula. Over all $B = 1000$ iterations, we calculated average validation AUCs and illustrated results in the form of learning curves. Briefly, a learning curve is a type of plot in machine-learning to show the change in model prediction accuracy (here: discrimination) when cost (here: abstraction sample size) increases. In these experiments, since we compared prediction accuracy across different sampling designs conditioned on the same models and data generating mechanism, the difference in model performance is due to differences in the sampling design that gave
3.2 Simulation results

Figure 3 illustrates simulation results when modeling with logistic lasso regression. First, consider the cohort with 5% outcome prevalence and SGS sampling using surrogate Z1 (Figure 3(a)(i)), where surrogate discrimination for outcome was low but machine-learning may improve model discrimination. Learning with samples obtained with SRS is difficult, requiring an abstraction sample size of n=3000 to achieve a validation AUC of 0.85 (94% of the maximum AUC of 0.90). In contrast, using SGS achieves such discrimination at lower sample sizes, with the SGS 1:1 design requiring n=1500 (50% of SRS cost) and the SGS 3:1 design requiring n=1000 (33% of SRS cost). Notice that while assigning a higher proportion to surrogate positives (SGS 3:1) resulted in slightly higher learning curves compared to equal proportions of surrogate positives and negatives (SGS 1:1), such differences were less substantial compared to the difference between using any SGS compared to SRS. We remark that the benefit of ROS on learning is inconsistent, where such case replication sometimes resulted in worse generalizable discrimination compared to no replication (SRS) at all.

Figure 3(b)(i) shows the same setting as described before, but SGS sampling was based on surrogate Z2 rather than Z1. Even though surrogates Z1 and Z2 had the same discrimination for the outcome (AUC = 0.675), Z2 was a worse variable for stratified sampling purposes. For a validation AUC of 0.85, using surrogate Z2 based on the SGS 1:1 and SGS 3:1 allocations required n=2500 (83% of SRS cost) and n=2000 (67% of SRS cost) respectively. Such differences may be attributed to the lower specificity of surrogate Z2 compared to Z1, resulting in overall lower sample prevalence and therefore reduced benefit for learning. Similar results on SGS design benefit and surrogate specificity were observed for the cohort having 10% outcome prevalence (Figures 3(a)(ii) and 3(b)(ii)), but SGS design benefit over SRS was less pronounced due to a less rare outcome. To achieve a validation AUC of 0.85 with lasso
regression, using SRS required an abstraction sample size of n=1000 while using SGS 1:1 sampling based on surrogate Z1 and Z2 required n=700 and n=900 respectively, compared to sample size savings of 50% and 83% respectively for the 5% outcome scenario.

Figure 4 illustrates learning curves for modeling with logistic ridge regression, with SGS using surrogate Z1 (Figure 4(a)) and Z2 (Figure 4(b)). Compared to using lasso regression, the different shapes of learning curves reflected differences in choice of modeling using variable selection versus shrinkage. To achieve a validation AUC of 0.85 for the 5% outcome prevalence cohort, learning with SRS required an abstraction sample size of at least n=4000 while SGS sampling using surrogate Z1 required about n=2500 (less than 63% of SRS cost), where using SGS regardless of stratification allocation was a consistent improvement over SRS. On the other hand, SGS sampling using surrogate Z2 had almost the same sample size requirement as with SRS, again emphasizing the importance of surrogate specificity for sampling. Similar conclusions were observed for the 10% outcome prevalence cohort. When modeling with ridge regression, ROS was consistently worse than SRS without case replication. One possible explanation is due to that using ridge regression reduces the estimation variance of classification through intentionally biased estimates. With over-sampling, while modeling bias increases, variation remains the same as case replication does not provide additional information, therefore resulting in lower generalizable prediction accuracy.

Therefore, our results suggest three consistent patterns associated with SGS sampling. First, using SGS for sampling was generally an improvement over using SRS for classification of rare outcomes, as observed for both lasso and ridge regression learning. Second, allocating higher proportions to the surrogate positive stratum resulted in improved learning compared to equal allocations, but only slightly. Third, using a more specific surrogate results in a sample with higher information and therefore improved learning compared to using a less specific surrogate. We acknowledge that quantification of the exact design benefit on learning
depends on factors such as the specific modeling choice and the outcome prevalence.

4 Application: Fracture identification from radiology reports

4.1 Data set details

Vertebral fractures of the spine can lead to spinal deformity, loss of vertebral height, crowding of internal organs, and loss of muscles, resulting in acute back pain and potentially chronic pain. Diagnosis is usually made through radiographic imaging, such as with plain x-ray or magnetic resonance imaging (MRI). In EMR systems a vertebral fracture finding is natively captured in unstructured text form, and for research a definite fracture status variable requires clinical expert abstraction of associated radiology text reports. Therefore, sampling strategies alternative to the usual SRS may be reduce the abstraction burden towards accurate and scalable machine-learning classification of vertebral fracture outcomes.

The Lumbar Imaging with Reporting of Epidemiology (LIRE) study evaluated the effect of radiology report content on subsequent treatment decisions among adult subjects (Jarvik et al. [2015]). Subjects were eligible for the LIRE study if they had a diagnostic imaging test ordered by their Primary Care Physician (PCP), so all subjects in LIRE had at least one radiology report available from the EMR database. The prevalence of vertebral fractures is estimated to be relatively rare: 3-20% among primary care subjects seeking care for all reasons (Waterloo et al. [2012]) and expected to be similar among subjects from the LIRE study. Using LIRE data as the “cohort”, we evaluate the benefit of using SGS designs for outcome label abstraction and subsequent classification model development.
4.2 Surrogate creation and sampling design application

Together with clinicians, we identified a set of 26 International Classification of Disease (ICD) codes that if present, are highly likely to indicate that a subject was diagnosed with a vertebral fracture; details are in Supplementary Material C. For each subject, we counted how many ICD codes were noted in the EMR within 90 days of cohort entry. In the cohort of 178,333 subjects, 171,592 (96%) did not have any relevant ICD codes, 3,275 (1.83%) had one code, 1,303 (0.73%) had two codes, 758 (0.42%) had three codes, and the remaining had more than three codes. Since most subjects did not have any relevant ICD codes and a count of one was the most common count, we defined the enrichment surrogate $Z$ as $Z_i = I(\text{count vertebral fracture ICD codes within 90 days for subject } i > 1)$, where 3.78% of the cohort were considered to be “surrogate positives”.

This abstraction task was nested within a larger abstraction set-up for the LIRE study. The radiology reports of each selected subject were abstracted by two independent clinicians for the presence or absence of vertebral fractures. From the available dataset, data “marts” for model development and model validation were assembled, each having a sample size of n=500. The validation data mart was selected such that it was representative of the underlying cohort, while the development data mart was selected based on an SGS 1:1 configuration. Using the validation data mart, we estimated marginal characteristics of the surrogate as well as the $O_{ratio}$ of the resulting SGS design.

4.3 Modeling and analysis

Features were created by processing radiology report text data using the quanteda package in R. Features were bag-of-words (BOW) unigrams excluding typical English stopwords as well as terms that were very rare (< 5% of all reports) or common (> 90% of all reports). We used the term-frequency inverse-document frequency (TF-IDF) representation for BOW as described in Salton and Buckley [1988], which incorporates information about the impor-
tance of terms both locally (within a single report) as well as globally (across all reports). For a collection of \( N \) reports denoted \( d_1, \ldots, d_N \), the set of \( p \) terms denoted \( T = \{ t_1, \ldots, t_p \} \) was obtained from concatenating unique words from all reports. Then the TF-IDF feature matrix \( X \) contains elements \( X_{ij} = TF(d_i, t_j) \times IDF(t_j) \), with term frequency \( TF \) defined as \( TF(t_j, d_i) = 1 + \log(1 + \frac{\text{Count}(t_j \in d_i)}{|d_i|}) \) and inverse document frequency \( IDF \) defined as \( IDF(t_j) = \log \left( \frac{N}{\sum_{i=1}^{N} I(t_j \in d_i)} \right) \). In addition to text-features, we also included the binary enrichment surrogate \( Z \) as a predictor, for a total of \( p = 298 \) features.

To investigate the design effect on model prediction accuracy, we drew \( B = 1000 \) bootstrap samples of sizes \( n = 100, 250, 500 \) from the development data mart stratified by surrogate status. To simulate the SRS design, we drew samples according to an “inverse SGS” design from the development data mart, where surrogate positives were under-included with the sampling probabilities [7]. To simulate the SGS design, we drew samples randomly from the development data mart. For each simulated sample, we fitted Lasso logistic regression selecting regularization parameter \( \lambda \) based on minimizing the average 10-fold cross-validated error using an AUC loss function. Resulting estimated model parameters were then applied to the validation sample to obtain estimates of the validation AUC. For each sampling design (SRS and SGS) and for each sample size, we reported mean validation AUC and 95% bootstrap confidence intervals.

Estimated data set characteristics are shown in Table 1. Note that the defined enrichment surrogate by itself had low AUC in discriminating the outcome. In fact, its operating characteristics were such that it was highly specific but only moderately sensitive for the finding, an overall high \( O_{ratio} \) for the resulting SGS design. Even though only a weak predictor, the surrogate is beneficial as a sampling variable. Data analysis results are shown in Table 2. In general, when fitting a logistic lasso regression, average validation AUC increases with sample size. However, for the same sample size, using samples drawn with SGS resulted in
higher average validation AUC. For example, using the same sample size of \( n = 250 \), the AUC of SGS was 0.86 while that of SRS was only 0.74, a difference of 0.12 suggesting that allocating a sample size of 250 is more resource efficient under SGS compared to SRS.

5 Discussion

In summary, motivated by sampling frameworks from epidemiology and machine-learning, we formalized a design strategy for abstraction selection and label collection of rare outcomes through a two-phase stratified sampling framework. We have demonstrated that the specificity of a sampling variable used to guide sampling greatly affects design benefit for classification accuracy. We suggest that a specificity of 0.95 or higher is ideal, 0.80 is very good, and 0.50 is the absolute minimum specificity that a surrogate needs to achieve in order to be considered as a sampling variable. To create highly specific surrogates, simple keyword searches may be supplemented with off-the-shelf negation tools for example in Harkema et al. [2009], while related ICD codes may be defined with higher count thresholds within a shorter period of time. Additionally, keywords and ICD codes may be combined with an “AND” query to further increase specificity. To estimate the specificity of a candidate surrogate, a small initial sample may be collected using SGS, where we remark that appropriate estimators may be based on those described in the verification bias literature (Alonzo [2014]).

The are two practical trade-offs of the proposed design worth considering. First, a concern may be whether sampling on a highly specific surrogate could result in a dataset that is sufficiently representative of all possible outcome subgroups. For example, in the vertebral fracture data application, while requiring at least two instead of one ICD codes may further increase surrogate specificity, such a strategy could have resulted in a sample with mostly chronic fractures and not acute fractures. A possible solution may implement a “tiered” surrogate, using sub-samples defined by variables to balance specificities and case repre-
sentativeness (e.g. > 2, 1, 0 counts of ICD codes). Another trade-off relates to strata proportions. While increasing sample surrogate positives may result in slightly improved classification accuracy, resulting inflated inverse weights may greatly increase the variance of IPW estimators of validation accuracy measures.

Anchored in the proposed SGS design framework, future work may formally investigate methodological and practical questions related to full study planning such as formal sample size calculations. Once relevant trade-offs are carefully defined, appropriate sample size calculations may then proceed taking into account the need of both model development and model validation. Other future work should include: investigating the appropriateness of the SGS framework for outcomes that are much rarer than what we considered (5%); characterizing design effects on prediction accuracy measures other than AUC; as well as determining best practices for sampling in the presence of site heterogeneity. Ultimately, our hope is to encourage careful statistical and study design thinking when assembling labeled data sets for machine-learning model development and validation, especially when considering the non-trivial abstraction cost in obtaining such labels.

References

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6):1166–1173, 2016.

Todd A Alonzo. Verification bias - impact and methods for correction when assessing accuracy of diagnostic tests. *REVSTAT-Statistical Journal*, 12(1):67–83, 2014.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

Wendy Webber Chapman, Marcelo Fizman, Brian E Chapman, and Peter J Haug. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. *Journal of biomedical informatics*, 34(1):4–14, 2001.
Nilanjan Chatterjee, Yi-Hau Chen, and Norman E Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168, 2003.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

Bernard CK Choi. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology*, 148(11):1127–1132, 1998.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Jeffrey G Jarvik, Bryan A Comstock, Kathryn T James, Andrew L Avins, Brian W Bresnahan, Richard A Deyo, Patrick H Luetmer, Janna L Friedly, Eric N Meier, Daniel C Cherkin, et al. Lumbar imaging with reporting of epidemiology (lire)protocol for a pragmatic cluster randomized trial. *Contemporary clinical trials*, 45:157–163, 2015.

Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

Michael A McIsaac and Richard J Cook. Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics*, 42(2):268–284, 2014.

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

Serguei V Pakhomov, James Buntrock, and Christopher G Chute. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. *Journal of biomedical informatics*, 38(2):145–153, 2005.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343, 2016.

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.
Herbert S Sichel. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547, 1975.

Jennifer A Sinnott, Wei Dai, Katherine P Liao, Stanley Y Shaw, Ashwin N Ananthakrishnan, Vivian S Gainer, Elizabeth W Karlson, Susanne Churchill, Peter Szolovits, Shawn Murphy, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human genetics*, 133(11):1369–1382, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sungwhan Sohn, et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49, 2018.

Svanhild Waterloo, Luai A Ahmed, Jacqueline R Center, John A Eisman, Bente Morseth, Nguyen D Nguyen, Tuan Nguyen, Anne J Sogaard, and Nina Emaus. Prevalence of vertebral fractures in women and men in the population-based tromsø study. *BMC musculoskeletal disorders*, 13(1):3, 2012.

Qiong Wei and Roland L Dunbrack Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS one*, 8(7):e67863, 2013.

Gary M Weiss and Foster Provost. The effect of class distribution on classifier learning: an empirical study. 2001.

Jing-Hao Xue and Peter Hall. Why does rebalancing class-unbalanced data improve auc for linear discriminant analysis? *IEEE transactions on pattern analysis and machine intelligence*, 37(5):1109–1112, 2015.

Sheng Yu, Abhishek Chakrabortty, Katherine P Liao, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association*, 24(e1):e143–e149, 2016.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004.

Yang Zhao, Jerald F Lawless, and Donald L McLeish. Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(1):123–136, 2009.

Yang Zhao, Jerald F Lawless, and Donald L McLeish. Design and relative efficiency in two-phase studies. *Journal of Statistical Planning and Inference*, 142(11):2953–2964, 2012.
Figure 1: Expected sample case proportions (shaded regions) for simple random sampling (SRS) and surrogate-guided sampling (SGS) designs with 1:1 and 3:1 ratio of surrogate positive to negative in the sample. Illustrations are based on a scenario with outcome prevalence 10%, and surrogate with sensitivity 40% and specificity 95% for the outcome of interest.

(a) SRS.
(b) SGS 1:1.
(c) SGS 3:1

Figure 2: $O_{ratio}$ values for surrogates of different marginal sensitivity and specificity, based on a fixed $R = 0.50$ and an outcome with prevalence of 10%.
Figure 3: Logistic Lasso Regression learning curves (outcome prevalence = 5%) comparing simple random sampling (SRS), random over-sampling (ROS), and surrogate-guided sampling (SGS) with 1:1 or 3:1 ratio of surrogate positives to negatives. Surrogate $Z_1$ had sensitivity = 0.40 and specificity = 0.95, while surrogate $Z_2$ had sensitivity = 0.67 and a specificity = 0.66.
Figure 4: Logistic Ridge Regression learning curves (outcome prevalence = 5%) comparing simple random sampling (SRS), random over-sampling (ROS), and surrogate-guided sampling (SGS) with 1:1 or 3:1 ratio of surrogate positives to negatives. Surrogate $Z_1$ had sensitivity = 0.40 and specificity = 0.95, while surrogate $Z_2$ had sensitivity = 0.67 and a specificity = 0.66.
Table 1: Estimated data set characteristics for radiology reports drawn from the LIRE data set: Sensitivity, specificity, AUC, and Likelihood Ratios of the defined surrogate, as well as the O\textsubscript{ratio} of resulting surrogate-guided sampling (SGS) design. 95% confidence intervals were based on B=1000 bootstrap resamples.

| Surrogate or design metric | Estimate (95% C.I.) |
|---------------------------|---------------------|
| Sensitivity of $Z$        | 0.27 (0.18, 0.36)   |
| Specificity of $Z$        | 0.99 (0.99, 1)      |
| AUC of $Z$                | 0.633 (0.58, 0.68)  |
| LR\textsuperscript{+} of $Z$ | 26 (18, 41)      |
| LR\textsuperscript{−} of $Z$ | 1.36 (1.22, 1.57)  |
| $O_{ratio}$ using SGS with 1:1 ratio of $Z = 1$ and $Z = 0$ | 7.13 (5.65, 9.04) |

Table 2: Average validation AUC (95% C.I.) for various training sample sizes, based on B=1000 bootstrap resamples, for illustration of surrogate-guided sampling (SGS) designs on radiology reports drawn from the LIRE data set.

| Training sample size | $A\hat{U}C(D^{SRS}(n))$ | $A\hat{U}C(D^{SGS}(n))$ |
|----------------------|--------------------------|--------------------------|
| 100                  | 0.68 (0.50, 0.90)        | 0.76 (0.63, 0.88)        |
| 250                  | 0.74 (0.50, 0.92)        | 0.86 (0.79, 0.91)        |
| 500                  | 0.83 (0.50, 0.92)        | 0.88 (0.85, 0.91)        |
A Proof of Theorem 2.1

Denote the cohort data as \( D = (X, Y) \), consisting of features \( X \) (implicitly also including the surrogate \( Z \) ), and binary outcomes \( Y \). From \( D \), units (typically subjects) are selected to form development and validation samples.

A.1 Preliminaries

In \( D \), let the features follow a bi-normal distribution, so that for \( y \in \{0, 1\} \)

\[
(X|Y = y) \sim N(\mu_{x|y}, \Sigma_{x|y})
\]  

(1)

This is equivalent to a Linear Discriminant Analysis (LDA) setting, where

\[
\logit(E[Y|X]) = \beta_0 + \beta^T X
\]

\[
\beta_0 = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\mu_{x|1} + \mu_{x|0})^T \Sigma_{x|y}^{-1}(\mu_{x|1} - \mu_{x|0})
\]

\[
\beta^T = \Sigma_{x|y}^{-1}(\mu_{x|1} - \mu_{x|0})
\]

(2)

The parameters in (2) are true parameters in \( D \). To estimate regression coefficients, a sample \( D^S(n) \) needs to be drawn from \( D \). Then, based on theory from generalized linear models, the resulting estimate \( \hat{\beta} \) has the following first and second moments:

\[
E^{D^S(n)}[\hat{\beta}] = \beta + \text{Bias}^{D^S(n)}(\hat{\beta})
\]

\[
\text{Var}^{D^S(n)}(\hat{\beta}) = (X^T W X)^{-1}
\]

(3)

In (3), \( W = \text{Diag}(p_i(1 - p_i)) \), where \( p_i = P(Y_i = 1|X, S_i = 1; \beta) \) estimates the average probabilities resulting from the sigmoidal transformation of development sample linear predictions. The terms in (3) are accurate up to second order approximations. In estimating the regression parameters, denote the bias \( \text{Bias}^{D^S(n)}(\hat{\beta}) \) as \( B(\hat{\beta}^S(n)) \) and variance \( \text{Var}^{D^S(n)}(\hat{\beta}) \) as \( V(\hat{\beta}^S(n)) \), then both \( B(\hat{\beta}^S(n)) \) and \( V(\hat{\beta}^S(n)) \) depend on the development sample \( D^S(n) \) through sample size \( n \) and sampling design \( S \). To evaluate the resulting classification model, we use a large validation sample, obtained using simple random sampling from \( D \). Denote the true linear predictions in the validation sample as \( \eta := X^v \beta \), with distribution

\[
X^v \beta \sim N(\mu_y, \sigma_y^2)
\]

\[
\mu_y = \mu_{x|y}^T \beta; \quad \sigma_y^2 = \beta^T \Sigma_{x|y} \beta
\]

for \( y \in \{0, 1\} \), where \( \mu_{x|y} \) and \( \Sigma_{x|y} \) were defined in (1). Under the bi-normal ROC assumption [Pepe 2003], the AUC is
\[
AUC = \Phi(\sqrt{R_{AUC}}) = \Phi \left( \sqrt{\frac{(\mu_1 - \mu_0)^2}{\sigma_1^2 + \sigma_0^2}} \right).
\]

In the classification setting, coefficients are estimated from the development sample \(D^S(n)\), where \(D^S(n)\) is generated with sampling design \(S\) and with development sample size \(n\). We use \(AUC(Y|D^S(n))\) to denote an indexing of resulting validation AUC, where

\[
AUC(Y|D^S(n)) = \Phi(\sqrt{R_{AUC}(D^S(n))})
\]

\[
R_{AUC}(D^S(n)) = \frac{(\hat{\mu}_1 - \hat{\mu}_0)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}.
\]

In (5), the notation \(\hat{\cdot}\) and \(D^S(n)\) indicates that the estimation of \(\hat{\beta}\) is from \(D^S(n)\). This proof outlines \(AUC(D^S(n))\) in terms of development sample composition.

### A.2 Mean and variances of validation sample linear predictions

In the large and representative validation sample, for \(y \in \{0, 1\}\), the mean of the estimated linear predictions is

\[
\hat{\mu}_y = E^{D^S(n)|X^v}[X^v\hat{\beta}|Y^v = y]
= E^{X^v}[E^{D^S(n)|X^v}[X^v\hat{\beta}|Y^v = y]]
= E^{X^v}[X^v(\beta + B(\hat{\beta}^S(n)))]Y^v = y]
= \mu^T_x(\beta + B(\hat{\beta}^S(n))).
\]

where the double expectation is due to the dependence on validation sample features \(X^v\) as well as development sample estimated coefficients \(\hat{\beta}\). Similarly, the variance of the estimated linear predictions is

\[
\hat{\sigma}_y^2 = Var^{D^S(n)|X^v}[X^v\hat{\beta}|Y^v = y]
= Var^{X^v}[E^{D^S(n)|X^v}[X^v\hat{\beta}|Y^v = y]] + E^{X^v}[Var^{D^S(n)|X^v}[X^v\hat{\beta}|Y^v = y]]
\]

The first part of the right hand side of (6) is

\[
Var^{X^v}[E^{D^S(n)|X^v}[X^v\hat{\beta}|Y^v = y]] = Var^{X^v}(X^v(\beta + B(\hat{\beta}^S(n)))]Y^v = y)
= (\beta + B(\hat{\beta}^S(n)))^T\Sigma_{x|y}(\beta + B(\hat{\beta}^S(n))),
\]

and the second part of the right hand side of (6) is
\[ E^{X^v}[Var^{D^S(n)}X^v(X^v\hat{\beta}|Y^v = y)] = E^{X^v}[X^vT V(\hat{\beta}^S(n))X^v|Y^v = y] = \text{trace}(V(\hat{\beta}^S(n))\Sigma_{x|y}) + \mu_{x|y}^T V(\hat{\beta}^S(n))\mu_{x|y}, \]

thus the numerator in (4) is the square of

\[ \sigma_y^2 = Var^{X^v}(E^{D^S(n)}X^v[X^v\hat{\beta}|Y^v = y]) + E^{X^v}[Var^{D^S(n)}X^v(X^v\hat{\beta}|Y^v = y)] = (\beta + B(\hat{\beta}^S(n)))^T \Sigma_{x|y}(\beta + B(\hat{\beta}^S(n))) + \text{trace}(V(\hat{\beta}^S(n))\Sigma_{x|y}) + \mu_{x|y}^T V(\hat{\beta}^S(n))\mu_{x|y}. \]

A.3 Classifier validation AUC in terms of estimation variance

Now we plug in values for (4). WLOG assume that \( \mu_{x|y} = 0 \) and that \( \Sigma_{x|y} = \Sigma_{x|y} \). Then, the means and variances of validation sample linear predictions among cases \((Y=1)\) and controls \((Y=0)\) are respectively

\[
\hat{\mu}_1 = \mu_{x|y}^T (\beta + B(\hat{\beta}^S(n))) \\
\hat{\mu}_0 = 0 \\
\hat{\sigma}_1^2 = (\beta + B(\hat{\beta}^S(n)))^T \Sigma_{x|y}(\beta + B(\hat{\beta}^S(n))) + \text{trace}(\Sigma_{x|y}) + \mu_{x|y}^T V(\hat{\beta}^S(n))\mu_{x|y} \\
\hat{\sigma}_0^2 = (\beta + B(\hat{\beta}^S(n)))^T \Sigma_{x|y}(\beta + B(\hat{\beta}^S(n))) + \text{trace}(V(\hat{\beta}^S(n))\Sigma_{x|y}).
\]

Thus, the numerator in (4) is the square of

\[ \hat{\mu}_1 - \hat{\mu}_0 = \mu_{x|y}^T (\beta + B(\hat{\beta}^S(n))), \]

while the denominator in (4) is

\[ \hat{\sigma}_1^2 + \hat{\sigma}_0^2 = 2((\beta + B(\hat{\beta}^S(n)))^T \Sigma_{x|y}(\beta + B(\hat{\beta}^S(n))) + \text{trace}(V(\hat{\beta}^S(n))\Sigma_{x|y})) + \mu_{x|y}^T V(\hat{\beta}^S(n))\mu_{x|y}. \]

Thus, based on (4), (11), and (12), since \( \Phi(.) \) and \( \sqrt(.) \) are monotone transformations,

\[ AUC(D^S(n)) = \frac{(\mu_{x|y}^T (\beta + B(\hat{\beta}^S(n))))^2}{2((\beta + B(\hat{\beta}^S(n)))^T \Sigma_{x|y}(\beta + B(\hat{\beta}^S(n))) + \text{trace}(V(\hat{\beta}^S(n))\Sigma_{x|y})) + \mu_{x|y}^T V(\hat{\beta}^S(n))\mu_{x|y}}. \]

When \( B(\hat{\beta}^S(n)) \approx 0 \), then since \( \beta \), \( \mu_{x|y} \) and \( \Sigma_{x|y} \) are assumed to be “fixed” quantities in a large validation sample,
AUC(D^s(n)) \propto \frac{1}{\text{trace}(V(\hat{\beta}^S(n))\Sigma_{x|y}) + \mu_{x|y_1}^T V(\hat{\beta}^S(n)) \mu_{x|y_1}}.
B Derivation of Proposition 2.1

For $Y \in \{0,1\}$, $E[Y] = P(Y = 1)$. Denote subjects where $S = 1$ as those included in $D^{SGS}(n)$, the SGS sample selected from the cohort only based on values of $Z$. Thus, $S \perp Y | Z$. The expected case odds in samples collected using SGS is

$$Odds(cases|SGS) = \frac{E^{D^{SGS}(n)}[Y|S = 1]}{1 - E^{D^{SGS}(n)}[Y|S = 1]} = \frac{P(Y = 1|S = 1)}{P(Y = 0|S = 0)}$$

$$= \frac{P(Y = 1|S = 1, Z = 1)P(Z = 1|S = 1) + P(Y = 1|S = 1, Z = 0)P(Z = 0|S = 1)}{P(Y = 0|S = 1, Z = 1)P(Z = 1|S = 1) + P(Y = 0|S = 1, Z = 0)P(Z = 0|S = 1)}$$

$$=\frac{P(Y = 1|Z = 1)P(Z = 1|S = 1) + P(Y = 1|Z = 0)P(Z = 0|S = 1)}{P(Y = 0|Z = 1)P(Z = 1|S = 1) + P(Y = 0|Z = 0)P(Z = 0|S = 1)}$$

$$= \frac{P(Y = 1)P(Z = 1|Y = 1) + (1 - R)P(Z = 0|Y = 1)}{P(Y = 0)P(Z = 1|Y = 0) + (1 - R)P(Z = 0|Y = 0)}$$

$$= \frac{P(Y = 1)R(Z_{sens}) + P(Z = 1)(1 - R)(1 - Z_{sens})}{P(Y = 0)R(Z_{spec}) + P(Z = 0)(1 - Z_{spec})}$$

where

$$R = P(Z = 1|S = 1)$$
$$p_Z = P(Z = 1)$$
$$Z_{sens} = P(Z = 1|Y = 1)$$
$$Z_{spec} = P(Z = 0|Y = 0).$$

The expected case odds in samples collected using SRS is

$$Odds(cases|SRS) = \frac{E^{D^{SRS}(n)}[Y|S = 1]}{1 - E^{D^{SRS}(n)}[Y|S = 1]} = \frac{P(Y = 1)}{P(Y = 0)}.$$

Then, the case/control odd ratio of samples obtained with SGS compared to that of SRS is:
\[
O_{\text{ratio}} = \frac{E^{DG}_{n}[Y|S=1]}{1 - E^{DG}_{n}[Y|S=1]} / \frac{E^{DR}_{n}[Y|S=1]}{1 - E^{DR}_{n}[Y|S=1]}
\]

\[
= \frac{E^{DG}_{n}[Y|S=1]}{1 - E^{DG}_{n}[Y|S=1]} / P(Y=1) / P(Y=0)
\]

\[
= \frac{RZ_{\text{sens}} + pZ(1 - R - Z_{\text{sens}})}{R(1 - Z_{\text{spec}}) + pZ(Z_{\text{spec}} - R)}.
\]

Assume that the outcome is rare, so \(P(Y=1) \approx 0\). Then, a linear approximation of (1) is

\[
O_{\text{ratio}} = \frac{RZ_{\text{sens}} + pZ(1 - R - Z_{\text{sens}})}{R(1 - Z_{\text{spec}}) + pZ(Z_{\text{spec}} - R)}
\]

\[
= \frac{R}{1 - P(Y=1|Z=1)}(LR+) + \frac{1 - R}{1 - P(Y=0|Z=0)}(LR-)
\]

\[
\approx (R)(LR+) + (1 - R)(LR-)
\]

where

\[
LR+ = \frac{P(Z=1|Y=1)}{P(Z=1|Y=0)} = \frac{Z_{\text{sens}}}{1 - Z_{\text{spec}}}
\]

\[
= \frac{P(Y=1|Z=1)}{P(Y=0|Z=1)}
\]

\[
LR- = \frac{P(Z=0|Y=1)}{P(Z=0|Y=0)} = \frac{1 - Z_{\text{sens}}}{Z_{\text{spec}}}
\]

\[
= \frac{P(Y=1|Z=0)}{P(Y=0|Z=0)}
\]

\(LR+\) and \(LR-\) are the likelihood ratios of the surrogate \(Z\) in predicting the outcome \(Y\) among surrogate positives and negatives, respectively.
C Details of enrichment surrogate for data application

Table 3 shows details of the set of ICD codes used to construct an enrichment surrogate which is used for collecting reports that are more likely to contain vertebral fracture. The enrichment surrogate was defined as

\[ Z_i = I((\text{count vertebral fracture ICD codes in Table 3 within 90 days for subject } i) > 1). \]

Table 3: Set of International Classification of Disease (ICD) codes used to define enrichment surrogate

| ICD code | Long description                                                                 |
|---------|----------------------------------------------------------------------------------|
| 806.25  | Closed fracture of T7-T12 level with unspecified spinal cord injury               |
| 806.26  | Closed fracture of T7-T12 level with complete lesion of cord                      |
| 806.27  | Closed fracture of T7-T12 level with anterior cord syndrome                       |
| 806.28  | Closed fracture of T7-T12 level with central cord syndrome                        |
| 806.29  | Closed fracture of T7-T12 level with other specified spinal cord injury           |
| 806.35  | Open fracture of T7-T12 level with unspecified spinal cord injury                 |
| 806.39  | Open fracture of T7-T12 level with other specified spinal cord injury             |
| 806.4   | Closed fracture of lumbar spine with spinal cord injury                            |
| 806.5   | Open fracture of lumbar spine with spinal cord injury                             |
| 806.6   | Closed fracture of sacrum and coccyx with unspecified spinal cord injury          |
| 806.61  | Closed fracture of sacrum and coccyx with complete cauda equina lesion           |
| 806.62  | Closed fracture of sacrum and coccyx with other cauda equina injury               |
| 806.69  | Closed fracture of sacrum and coccyx with other spinal cord injury                |
| 806.8   | Closed fracture of unspecified vertebral column with spinal cord injury           |
| 806.9   | Open fracture of unspecified vertebral column with spinal cord injury             |
| 733.13  | Pathologic fracture of vertebrae                                                 |
| 805.4   | Closed fracture of lumbar vertebra without mention of spinal cord injury         |
| 805.5   | Open fracture of lumbar vertebra without mention of spinal cord injury           |
| 805.6   | Closed fracture of sacrum and coccyx without mention of spinal cord injury       |
| 805.7   | Open fracture of sacrum and coccyx without mention of spinal cord injury         |
| 805.8   | Closed fracture of unspecified vertebral column without mention of spinal cord injury |
| 805.9   | Open fracture of unspecified vertebral column without mention of spinal cord injury |
| 809     | Fracture of bones of trunk, closed                                               |
| 809.1   | Fracture of bones of trunk, open                                                 |
| V54.17  | Aftercare for healing traumatic fracture of vertebrae                            |
| V54.27  | Aftercare for healing pathologic fracture of vertebrae                            |