Sampling designs for resource efficient collection of outcome labels for machine-learning, with application to electronic medical records

Wei Ling Katherine Tan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Patrick J. Heagerty, Chair

Jennifer C. Nelson

Noah R. Simon

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Sampling designs for resource efficient collection of outcome labels for machine-learning, with application to electronic medical records

Wei Ling Katherine Tan

Chair of the Supervisory Committee:
Dr. Patrick J. Heagerty
Department of Biostatistics

In leveraging data from large-scale electronic medical record systems for research, an important step is the accurate identification of key clinical outcomes. Some outcomes must be derived or predicted from both structured and unstructured data, for example using statistical machine-learning classification. Classification requires the collection of labeled data, which is a sample where actual outcome statuses are manually coded by human clinical experts. For rare outcomes, simple random sampling (SRS) for labeled data collection results in very few cases in the sample. Such outcome class imbalance results in insufficient information for classifier modeling, yet additional abstraction is often expensive and time-consuming. In this dissertation, we propose sampling designs for labeled data collection towards machine-learning, targeting the rare outcome scenario. Our proposed designs are resource efficient, requiring a smaller sample size for modeling goals compared to SRS, yet design impacts on model development and validation can be statistically characterized to be "valid". We first introduce a stratified sampling procedure based on values of enrichment surrogates, which are summaries of structured data related to the clinical outcome requiring abstraction. Next, motivated by radiology reports with multiple co-occurring findings, we discuss extensions to the multi-label setting. Finally, for scenarios where a previously developed "source" model is to be externally transferred, we propose a framework for such "new" labeled data collection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ABSTRACTION: A process where trained clinicians read unstructured clinical text (e.g. notes and radiology reports) and transcribe resulting information into codified variables.

CASE-ENRICHMENT: A sampling design with the case-enrichment property results in samples with higher case proportion compared to simple random sampling.

CLASSIFIER: A supervised machine-learning algorithm that categorizes input data into two or more categories based on pre-defined labels. A commonly used classifier in biomedical settings is logistic regression.

CLINICAL OUTCOME STATUS: The observable presentation of clinical condition status (case or control) as identified from a patient's medical record.

COHORT: The statistical population for which to draw inference from.

FEATURE: A generic term used to describe derived attributes that are used as inputs (i.e. independent variables) for a machine-learning model. Features are approximately equivalent to "predictors" in statistics, and are usually high-dimensional.

FEATURE ENGINEERING: The process of creating features, for example through counting objects and auto-encoding.

LABEL: Tags (usually 1 or 0 indicating case or control) to indicate categories of a set of documents based on human judgment. Labels are often used as the outcomes in machine-learning.

LABELED DATA: A sample of documents that has been coded with labels, often obtained through abstraction in biomedical settings. Labeled data is used to develop and validate supervised machine-learning models.

MODEL DEVELOPMENT: Model development describes the training and tuning of machine-learning model parameters; also called model training.

MODEL VALIDATION: Model validation describes the evaluation of a previously developed model in terms of its accuracy and generalizability to target settings. Examples of accuracy measures are sensitivity and specificity; also called model testing.

OUTCOME CLASS IMBALANCE: A sample with outcome class imbalance has unequal case and control proportions, usually describing outcome prevalences of 20% or lower. Outcome class imbalance is thought to negatively affect classifier learning.

PREDICTORS: Inputs or independent variables in a regression model.

RESOURCE EFFICIENCY: A sampling design with the resource efficiency property results in samples that are more informative for modeling goals compared to a simple random sample of the same size. For classification, one measure of resource efficiency is case-enrichment.

SAMPLE: A subset of data collected from the larger statistical population using a pre-defined procedure. In machine-learning samples are used for model development and model validation.

SAMPLING DESIGN: A pre-defined procedure to collect a sample from the larger statistical population.

SAMPLING RATIO: In the context of stratified sampling on two strata, we use sampling ratio to describe the ratio of sample stratum size to total sample size. Sampling ratio is related to sampling fraction (ratio of sample size to stratum size) in sampling theory through Bayes rule.

SURROGATES: Structured data elements in the electronic medical record (e.g. keywords, ICD codes) related to the true clinical outcomes of interest; may be viewed as misclassified outcomes. In this dissertation we use surrogates as sampling variables rather than to replace the true clinical outcome.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Patrick Heagerty, for his mentorship, patience, and inspiration. I am so privileged to have such a great teacher to learn from, among so many things, rigorous statistics, meaningful research, clear writing, and effective communication. Thank you, Patrick, for everything.

I sincerely thank my committee members, Jennifer Nelson, Noah Simon, Ruth Etzioni, and Robert Penfold, whose questions and comments have greatly shaped the direction of my work.

I thank the University of Washington Department of Biostatistics for the opportunity to pursue a PhD. To the faculty who have graciously taken the time and energy to mentor me, I will forever feel indebted and grateful. To the faculty, staff, and students who have together created a friendly and intellectually stimulating environment, I feel so thankful to be part of this family.

This work would not have been possible without the financial, academic, and personal support from the LIRE project team. Special thanks especially to Jerry Jarvik, Katie James, Sean Rundell, Pradeep Suri, Eric Meier, and Nancy Organ for their patience and insight during our many meetings discussing annotation, databases and NLP. I acknowledge project grants UH2 AT007766 and UH3 AR066795 for funding and data that motivated my work.

Lastly, I am so thankful for my wonderful family and friends for their love and kindness throughout the years. You guys are the best.

DEDICATION

To my parents, SiewWah and KhianKhoon, for their unconditional love and support.

Chapter 1
INTRODUCTION AND BACKGROUND

This dissertation concerns sampling designs for outcome label data collection and subsequent machine-learning, with applications to data arising from electronic medical records (EMR) databases. The focus is on valid and resource efficient designs, particularly when the outcomes are expected to be rare. In this chapter, we synthesize background material relevant to reading this thesis. We start by providing an overview of EMR databases and the Lumbar Imaging with Reporting of Epidemiology (LIRE) study [65] as motivation. Then, we provide notation and set-up, describe the rare outcome problem in machine-learning, review key concepts from epidemiology study designs, and summarize the scope of this dissertation.

1.1 Electronic medical records and scientific motivation

Electronic medical records (EMR) are large scale databases that have facilitated research involving complex data, such as linking genetic and phenotype data in genome-wide association studies, and identifying adverse events in pharmacovigilance studies [82, 75]. EMR databases contain structured elements, such as demographics, lab values, prescription medications, Current Procedural Terminology (CPT) and International Classification of Disease (ICD) codes. These structured data types are generally easily accessible, for example using techniques such as Structured Query Language (SQL) to search databases.

In addition to structured data, much of EMR data are so-called unstructured, natively stored for example as free-text clinical notes and radiology reports. Unstructured data often contain important information, such as subjects' medical conditions, but are not explicitly coded

in accessible forms. However, information from text data may be abstracted and labeled. Clinical abstraction, also known as annotation, is the process of manual chart review to transcribe selected patient information into codified variables. Typically involving highly trained medical personnel, abstraction is time consuming and expensive, but abstracted clinical outcomes provide substantial added value to EMR data. For example, these clinical outcomes may be used in research studies utilizing EMR data to investigate treatment effects or patterns of care.

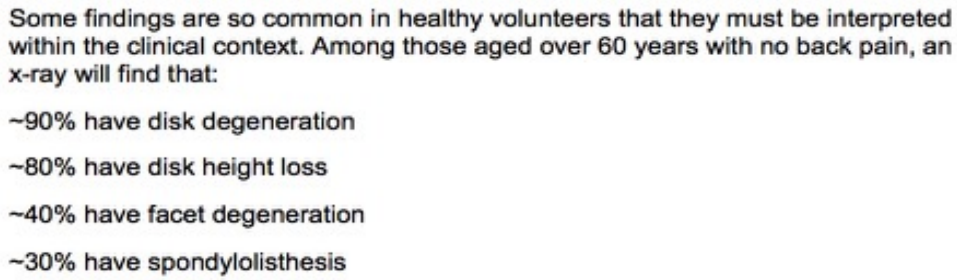
1.1.1 Motivation from the Lumbar Imaging with Reporting of Epidemiology (LIRE) study

The Lumbar Imaging with Reporting of Epidemiology (LIRE) study was a pragmatic clinical trial that studied the effect of radiology report content on subsequent treatment decisions [65]. From four health systems across the United States (Kaiser Permanente of Washington, Kaiser Permanente of Northern California, Henry Ford Health System, and Mayo Clinic Health System), the LIRE study enrolled over a quarter million adult subjects whose Primary Care Provider (PCP) ordered a lumbar spine diagnostic imaging test between October 2013 and September 2016, but had not in the prior year. The included reports represent various imaging modalities, such as x-ray and magnetic resonance imaging (MRI).

In the United States, a concern about health care spending is so called "over-utilization" of radiographic imaging: most subsequent imaging do not meaningfully improve patient outcomes [65]. A hypothesis in the LIRE study was that such unnecessary subsequent imaging may be reduced through helping PCPs interpret radiology reports. To do that, the LIRE "intervention" involved inserting text describing epidemiologic prevalences at the end of radiology reports received by participating PCPs (Figure 1.1). These intervention text are analogous to normalized values for lab tests, and served as benchmarks to aid interpretation of radiographic findings on radiology reports. Nine different versions were included, depending on imaging modality and age categories. The main scientific question in LIRE was whether inserting such epidemiologic benchmarks would reduce downstream health care

utilization, such as subsequent imaging rates.

Figure 1.1: LIRE intervention text describing estimated prevalences of four radiographic findings on the x-ray modality among patients above 60 years old.



In the LIRE study, it may be reasonable to expect different subsequent imaging rates in various sub-groups. For example, among subjects whose radiology reports contain findings specifically mentioned in the intervention text, subsequent imaging rates may be lower. Additionally, among subjects with serious "red flag" conditions, subsequent imaging rates may be high irrespective of intervention. To conduct such sub-group analyses requires reliable methods to extract radiographic finding information from unstructured text reports. Instead of traditional large-scale manual abstraction, one potential accurate and scalable approach is through the use of machine-learning classification algorithms. Towards such machine-learning model development and validation, an adequate sample of reports needs to be assembled, and true clinical outcome statuses (for instance, presence/absence of a set of radiographic findings) need to be abstracted by human clinical experts. The value of machine-learning is that, instead of manual review of all EMR records, only a smaller sample of subjects needs to be abstracted for true clinical outcome statuses. Yet, questions such as "how many do we need" and "how to plan for sample selection" are currently relatively unexplored. This dissertation is motivated by practical applications, such as the LIRE study, that where abstraction of true clinical outcomes is required for machine-learning modeling and downstream analyses.

1.2 Notation and set-up

1.2.1 Notation

As this dissertation concerns sampling designs for the purpose of machine-learning modeling, we provide notation from the dual perspectives of modeling and sampling. We follow statistical convention for notation, with X for random variables, x for values, $\tilde{X}$ for column vectors, $\mathbf{X}$ for matrices, $\{\mathcal{X}_1, \ldots, \mathcal{X}_J\}$ for sets, and $\mathcal{X}_j$ as a (generic) element of a set.

Relevant elements for machine-learning modeling include input features, outcome labels, and the classification model. Input features are multi-dimensional, where for feature matrix $\mathbf{X}$, X_{ij} indicates the jth feature for subject i, $\tilde{X}_i^T \in \mathcal{R}^p$ indicates all p features for subject i, and $\tilde{X}_j \in \mathcal{R}^n$ indicates the jth feature for all subjects. For univariate outcomes, $Y_i \in \{0, 1\}$ denotes the label for subject i, while labels for all subjects are denoted with Y, with vector notation suppressed by convention. For multi-variate outcomes $\mathbf{Y}$, $Y_{ki} \in \{0, 1\}$ is the kth label of subject i, $\tilde{Y}_i^T \in \{0, 1\}^K$ is all K labels of subject i, and $\tilde{Y}_k \in \{0, 1\}^n$ the kth label for all subjects. The classification model is function $h(.)$ that maps $\mathbf{X}$ to $E[Y|\mathbf{X}]$, where the empirical estimate $\hat{h}(.)$ is based on a observed data. For consistency of notation, we denote the fitting of $h(.)$ as model development (alternatively referred to as model training or parameter tuning), and the evaluation of $h(.)$ as model validation (alternatively referred to as model testing or classifier evaluation). From the modeling perspective, classification model development often implicitly assumes that a sample, typically representative of the population, is already available for model development and validation.

When the sample is not representative of the population, we introduce additional notation for the cohort (i.e. population), the sample, and the sampling design. The full cohort is denoted as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$. From $\mathcal{D}$, we denote $\mathbf{D}^S(n)$ as a sample of size n drawn using sampling method S. The sampling method S may be random, by convenience, or based on variables that are either observed or unobserved. Under settings that motivate this disser-

tation, we assume that the purpose of collecting $\mathbf{D}^S(n)$ for abstraction of outcome labels is so that resulting subsets may be used for model development and model validation. The distributions of cohort $\mathcal{D}$, development sample $\mathbf{D}^{dev}$, and validation sample $\mathbf{D}^{val}$ may all be different from each other, either unintentionally (e.g. different snapshots from a database) or by design (e.g. to improve prediction). It is well known in machine-learning that classifier learning severely suffers when outcomes are rare in $\mathbf{D}^{dev}$, a problem called the "outcome class imbalance" [7, 138]. Therefore, the distribution of $\mathbf{D}^{dev}$ may be intentionally altered to improve learning.

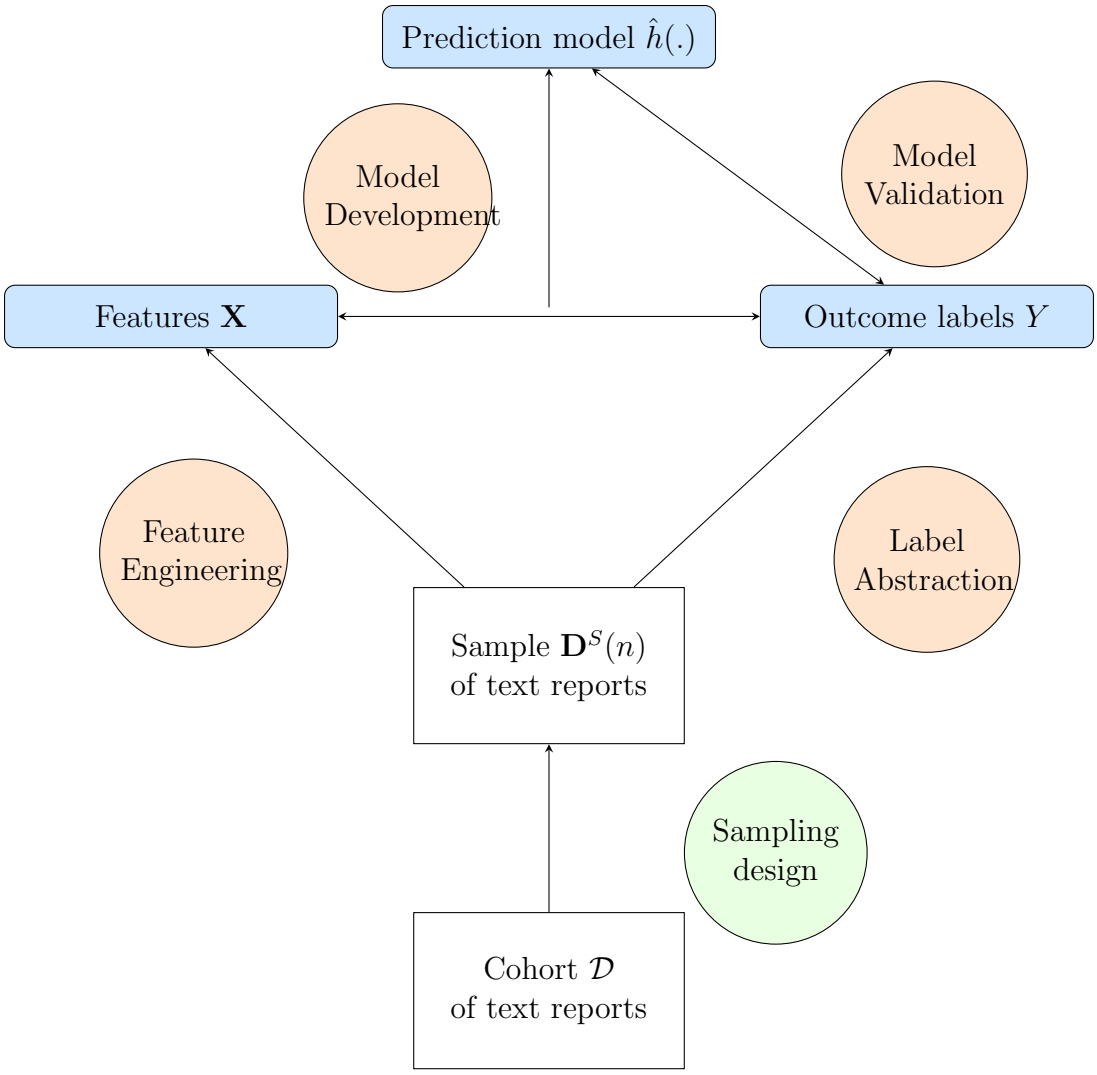
Figure 1.2 illustrates the described notation, where we had color-coded elements from the modeling perspective in blue and elements from the sampling perspective in white. This dissertation is primarily concerned with sampling designs to select samples of text reports $\mathbf{D}^S(n)$ from cohort $\mathcal{D}$ (green circle), therefore we do not provide extensive background on the other necessary processes in individual chapters. However, to fully appreciate the complexity of feature engineering, model development, and model validation methods, we next provide a brief review of key concepts.

1.2.2 Feature engineering of unstructured text into numeric features

To fit model $\hat{h}(.)$ using sample $\mathbf{D}^S(n)$, both features $\mathbf{X}$ and outcomes Y need to be available. Unstructured radiology report text may be converted into numeric feature matrices through Natural Language Processing (NLP) methods. One of the classic NLP feature engineering techniques is called bag-of-words (BOW). To understand BOW, we first define the following:

- **Word:** Word w is the basic unit of text data.
- **Document:** Document $d_i = \{w_{i,1}, \ldots, w_{i,n_i}\}$ is a sequence of n_i words, for example a single radiology report.
- **Corpus:** Corpus $\mathcal{C} = \{d_1, \ldots, d_N\}$ is a collection of N documents, for example a

Figure 1.2: Design and modeling of text data: elements from the machine-learning modeling perspective (blue boxes) and sampling design perspective (white boxes), as well as processes (circles) that connect the various elements.



database of radiology reports.

For a corpus of N documents let $T = \{t_1, \ldots, t_p\}$ denote the set of p unique terms arising from concatenating all documents, where terms t_j may be single words or sequences of words, also known as *N-grams*. Then, the binary Term Frequency $TF^{binary}(t_j, d_i)$ is

$$TF^{binary}(t_j, d_i) = I(t_j \in d_i) \tag{1.1}$$

Often, the number of terms within each document may be more informative than simple indicators, motivating the raw term frequency function. For t_j in document d_i,

$$TF^{raw}(t_j, d_i) = \sum_{k=1}^{n_i} I(w_{ik} = t_j) \tag{1.2}$$

is the number of times the jth term appears in document i. To account for a corpus with varying document length, $TF^{raw}(t_j, d_i)$ may be normalized by the total number of words in the document, and then log transformed. The (log-normalized) Term Frequency is

$$TF(t_j, d_i) = 1 + \log\left(1 + \frac{TF^{raw}(t_j, d_i)}{n_i}\right), \tag{1.3}$$

and ranges between 1 (d_i does not contain t_j) and $1 + \log(2)$ (all the terms in d_i are t_j). In addition, terms that are frequent in only a few documents (e.g. "tumor") may be more important than terms that are frequent across many documents (e.g. "pain"). The Inverse Document Frequency (IDF) captures such "between documents" term importance, and is

$$IDF(t_j) = log\left(\frac{N}{\sum_{i=1}^{N}(\sum_{k=1}^{n_i} I(w_{ik} = t_j)) > 0}\right). \tag{1.4}$$

$IDF(t_j)$ is often log-transformed, and ranges between 0 (all documents contain t_j) and ∞ (no documents contain t_j). To simultaneously capture within and between document term importance, the Term Frequency - Inverse Document Frequency (TF-IDF) multiplies TF (1.3) and IDF (1.4), where

$$\text{TF-IDF}(t_j, d_i) = TF(t_j, d_i) \times IDF(t_j). \tag{1.5}$$

Text data from document d_i can be represented with the feature vector $\tilde{X}_i^T$, with elements being any of (1.1),(1.2), (1.3) or (1.5). In comparing the various representations, TF-IDF (1.5) is generally thought of as the most "sophisticated", but the indicator function (1.1) has also been observed to be adequate for simple classification tasks and may be a more appropriate scale when models include additional non-text features, as often is the case for classification models utilizing EMR data.

Even though the widely used BOW representation is intuitive, it has been criticized for not accounting for contextual and linguistic information, such as word ordering. We briefly summarize two recent developments which are alternatives to BOW: Latent Dirichlet Allocation (LDA) [12] and Document Vectors [86]. LDA (not to be confused with the unrelated *Linear Discriminant Analysis*) takes the view that instead of representing individual words as features, similar words should be grouped together as "topics". For each document, the features generated through the LDA representation are "topic vectors", selected through fitting a generative hierarchical Bayes model that includes the number of topics (typically 50-200)

as a hyperparameter. Document Vectors can be thought of as an "autoencoder" of words to numbers by using contextual information. For each document, every word is mapped into a vector of numbers ("word vector") based on hyperparameters such as the number of surrounding words and the word vector length, where similar words have word vectors that are closer to each other based on pre-specified distance metrics. For each document, features may be obtained by averaging over all word vectors in the document. Of note, the feature representation methods of BOW, LDA, and Document Vectors were compared to each other for classifying radiology reports, and BOW was found to be as competitive as the two other more sophisticated approaches [149]. The explanation was that radiology report corpora contain fewer unique words compared to the general English language, therefore simple BOW representations may be preferable to extensive parameter fine-tuning using the other methods.

1.2.3 Classification models and evaluation metrics

To classify a single binary outcome Y, commonly used algorithms include logistic regression, decision trees, k-nearest neighbors and support vector machines [46]. For logistic regression, which is a Generalized Linear Model [88], high-dimensional features may result in model over-fitting. Assuming sparsity in features, Lasso regression [127] is an elegant extension that selects only a subset of features. Alternatively, Ridge regression [74] shrinks regression coefficients and accounts for multi-collinearity. The elastic-net [155] is a combination of Lasso and Ridge regression, and allows for simultaneous shrinkage and selection.

After model development, it is necessary to assess model generalizability to new data [46]. Model performance may be empirically estimated using a validation sample that is separate from the development sample. The modeling error may be represented using the expected prediction error (EPE) framework [46], which relates the expected value of a loss function $L(\hat{h}(.), Y)$ of predictions to actual labels, where

$$\begin{aligned} Err &= E[L(\hat{h}(.), Y)] \\ &= E^{\mathbf{D}^{dev}, \mathbf{D}^{val}}[L(\hat{h}(.), Y)]. \end{aligned} \tag{1.6}$$

Note that in (1.6), prediction error is a function of not only classifier $\hat{h}(.)$, but also development sample $\mathbf{D}^{dev}$. Therefore, classification evaluation measures can be used to compare development samples obtained using various sampling methods. Model evaluation can be based on predicted probabilities $\hat{p}$ or predicted binary classes $\hat{Y}$, where Table 1.1 summarizes common prediction accuracy metrics for evaluating binary classifiers.

Table 1.1: Commonly used evaluation metrics.

Metric name (alternative name)	Empirical estimate
Proportion Correct (Simple Accuracy)	$\frac{TP+TN}{TP+FP+TN+FN}$
Sensitivity (Recall)	$\frac{TP}{TP+FN}$
Specificity (Negative recall)	$\frac{TN}{FP+TN}$
Positive Predictive Value (Precision)	$\frac{TP}{TP+FP}$
Negative Predictive Value	$\frac{FN}{TN+FN}$
F-1 score	$2\left(\frac{Recall \times Precision}{Recall+Precision}\right)$
Area Under the ROC Curve (AUC)	$\frac{1}{n_1 n_0}\sum_{i=1}^{n}\sum_{j<i} I(\hat{h}(\tilde{X}_i^T) > \hat{h}(\tilde{X}_j^T) \mid Y_i = 1, Y_j = 0)$

For predicted classes $\hat{Y}$, evaluation metrics can be constructed based on functions of classification tables, also called confusion matrices (Table 1.2). For example, the Proportion Correct is the sum of diagonals divided by the total validation sample size, and summarizes average agreement of model predictions to true outcome labels with equal penalty on false positives (FP) and false negatives (FN).

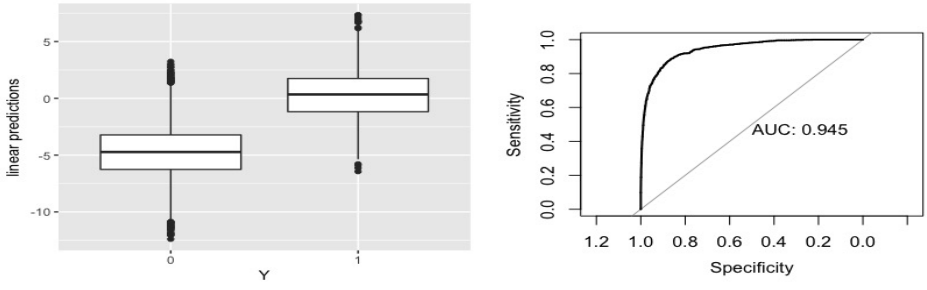
For predicted probabilities, a commonly used evaluation metric is the Area Under the Receiving Operating Characteristic Curve (AUC), which is a measure of discrimination with

Table 1.2: Classification table/Confusion Matrix

	True Y=1	True Y = 0
Predicted $\hat{Y} = 1$	TP	FP
Predicted $\hat{Y} = 0$	FN	TN

connections to the Mann-Whitney-Wilcoxon two-sample test statistic [53]. Predicted probabilities may be plotted by subject case/control status (Figure 1.3; left panel), where discrimination is higher when box plots are better separated. For a threshold k, model predicted probabilities can be thresholded to obtain the empirical Sensitivity and Specificity. The derived $(Sensitivity(k), Specificity(k))$ corresponds to a point on the $[0,1]^2$ space. Then, the Receiving Operating Characteristic (ROC) curve (Figure 1.3; right panel) may be computed by tracing the set of points $(Sensitivity(k), Specificity(k))$ for every $k \in [0,1]$. The AUC is widely used in both diagnostic testing and machine-learning to measure the ability of a test or classification model in separating true cases and controls, and is particularly attractive due to its invariance to outcome prevalence and the scale of predictions [97, 46].

Figure 1.3: Receiving Operating Characteristics (ROC) Curves.



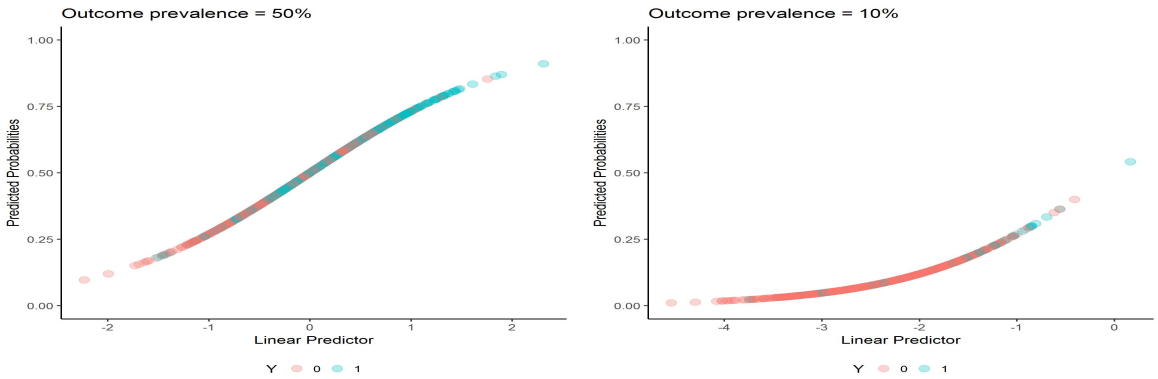
1.3 The effect of rare outcomes on classifier learning

For machine-learning classification, it is well accepted that model prediction accuracy is affected by the proportion of cases (Y=1) in the developed sample. In particular, classifying rare outcomes is more difficult than classifying outcomes with prevalences closer to 50%. This issue, known as the "outcome class imbalance" problem, has been empirically demonstrated for various classification models and evaluation metrics [7, 138]. Here, we provide some intuition for why outcome class balance may affect classifier learning.

For certain prediction accuracy metrics, the effect of rare outcomes is obvious. For example, for the Simple Accuracy metric, a trivial classifier that predicts $Y = 0$ deterministically will be artificially highly accurate, yet many more sensible classification rules may have artificially low accuracy. However, the effect of outcome class imbalance problem on classifier learning has been repeatedly noted even for prevalence-independent metrics such as the AUC [138, 7, 135]. An intuition for such a phenomenon was discussed in [138], where the authors asserted that prediction error may be decomposed by outcome class, and that the error for predicting cases (Y=1) is higher than the error for predicting controls (Y=0) when outcome classes are imbalance in training.

Why might the error for predicting cases be higher than that for predicting controls? For logistic regression in particular, an explanation can be attributed to the distribution of *predicted* probabilities. In particular, it was noted in [70] that comparing rare outcomes to more prevalent outcomes, predicted probabilities $\hat{p}_i$ are overall closer to 0 and true cases have predictions closer to 0.50 rather than 1 (an example of this claim is reproduced in Figure 1.4). The explanation in [70] related the overall low predicted probabilities to high statistical variance, but the link between variance and AUC has not been explicitly shown.

Figure 1.4: Predicted probabilities for logistic regression of a prevalent (50%) and a rare (10%) outcome.



1.4 Sample selection bias and missing data

Since outcome class imbalance is generally accepted to be a problem for classification learning, it is common practice to intentionally alter the sample case/control proportions for machine-learning model development. However, if $\mathbf{D}^{dev}$ and $\mathcal{D}$ are distributed differently, such associated "sample selection bias" [147] may compromise validity for model development and/or validation.

To understand "bias" resulting from sample selection requires the characterization of distributional differences between the collected sample and the intended population for generalizable inference. In epidemiology, a sample that is not representative of the population is said to have "selection bias". Selection bias can be due to reasons such as early study termination, loss of follow up, and personal biases of the data collector. When selection bias is due to sampling (either unintentional or by design), resulting biases are also known as "ascertainment bias" or "sample selection bias". To analyze data assumed to have sample selection bias, one perspective is to view the sample as having "missing data" from the cohort, where resulting bias may be characterized by the missing data mechanism.

The missing data mechanism comparing a sample to the population generally follows one of three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [78]. When data is assumed to follow a MCAR mechanism, the sample is statistically a random sample from the cohort, therefore analysis based on the available dataset generally results in unbiased estimates. On the other hand, samples that are assumed to be MNAR from the cohort are difficult to analyze without additional assumptions on the distribution of either the full data or the missing data. Finally, MAR describes scenarios where any missing values only depends on other observed variables, for example non-response in surveys may only depend on known subject demographics. For samples that are MAR from the intended population, resulting bias may be corrected using approaches broadly categorized as *imputation* to fill in missing values or *weighting* of the collected sample.

1.5 Two phase sampling designs

A commonly used design that results in samples with ascertainment bias, specifically bias due to MAR, is the two phase sampling design [89]. Two phase sampling was initially described for census estimation of population means and proportions, where the desired true outcome is expensive to ascertain, therefore data collection is based on values of a related cheaper variable. In the first phase, values of the cheaper variable are obtained for all subjects in the population. Then, in the second phase, subjects are drawn based on strata defined by values of the cheaper variable, and only selected subjects have their expensive variable information ascertained. The intended goal of most two phase sampling applications is to obtain efficient estimation under practical resource constraint scenarios.

Among the large body of literature associated with two phase sampling, a common scenario is when the exposure X is expensive to collect, but outcomes Y and/or related auxiliary variables V are available. Then, data collection of expensive X is based on cheaper Y and/or V. Efficiency is often described in terms of the asymptotic variance in estimating the

exposure effect using a pre-specified model (for example maximum likelihood). Even though the classic two phase sampling setting is different than the scenario studied in this dissertation (where Y is the expensive variable), certain results may be relevant to our conclusions.

1.5.1 Factors affecting two phase design efficiency

In particular, the question of "what factors affect two phase design efficiency" was described in [84], where the authors considered using binary outcomes Y and auxiliary variables V for sampling. The following three factors were identified in [84]: the true parameter values to be estimated, the method of analysis, as well as how the cheaper variable is used to define the phase 2 design also called the "design effect". Design effect on estimation efficiency can further be decomposed into the more specific questions of "what types of cheap variables are preferable for sampling" and "how to allocate the phase 2 sample based on the phase 1 values".

The effect of sampling variable characteristics on two phase design efficiency was studied in [153], where the main result was that using highly correlated and informative variables for sampling increased two phase design efficiency. In [153], analytical and empirical results were demonstrated for data following a normal distribution and estimation with maximum likelihood. The authors in [153] commented that their analytical results cannot be extended to logistic regression, but efficiency gains for general scenarios may be shown through numerical calculations.

The effect of phase 2 sample allocation was studied in [84], comparing "balanced" designs where all strata frequencies are equal, "optimal" designs based on minimizing the asymptotic variance of pre-specified models, to the baseline random sampling. The main conclusion in [84] was that if the analytic model is known in advance, then the "optimal" sampling weights are most efficient. However, their results also suggest that balanced designs are robust to using various analytic models. Of note, the balance design recommendation has also noted

in other investigations of two phase sampling [15, 105].

1.5.2 Estimating accuracy measures using samples from two phase designs

Thus far, we have described using two phase sampling in the context of efficient exposure estimation. Alternatively, samples arising from two phase designs may be used to estimate accuracy measures of "tests". In diagnostic testing, a goal is to measure the accuracy of a diagnostic test for the true clinical outcome. However if obtaining true outcomes is prohibitive, often subjects who test positive are preferentially oversampled for verification. Such "verification bias" is a special case of ascertainment bias; here we comment on bias-correction frameworks developed for verification bias, and discuss connections to our work.

Bias corrected accuracy estimators can be based on either the sampling model or the outcome model [2]. If the outcome model is known or can be accurately estimated, then imputation methods based may be based imputing the true outcomes for all subjects (full imputation), or only subjects with unobserved outcomes (mean score imputation). Alternatively, if the sampling model is known, then correction can be based on inverse probability weighting (IPW). In particular, empirical formulas for sensitivity, specificity, and AUC were derived in [2] using for both imputation and weighting methods. In the classic verification bias scenario, the test is the sampling variable, therefore sampling weights are computed based on differences between sample and cohort test distributions. Extending to the general two phase setting, the probabilities of subject inclusion may be computed based on observed sampling variables (which are not necessary the test for which accuracy is to be assessed).

1.6 Scope of this dissertation

This dissertation concerns statistical sampling designs for outcome label collection and subsequent machine-learning model development and validation, specifically for data arising from electronic medical records (EMR). Our assumptions and methods are motivated by results from machine-learning sampling and epidemiological study design. Our primary mo-

tivation comes from the outcome class balance problem in machine-learning and the need for more actual cases. However, in addition to improving empirical prediction accuracy, we are also motivated by designs that are both amenable for valid analysis and attentive to data collection costs. Therefore, it may be helpful for the reader with a general statistical background to view our proposed methods as a special case of two phase sampling, where the expensive variable to be collected is the outcome labels. This dissertation has following aims:

Aim 1: Surrogate-guided sampling designs. When Y is expected to be rare in $\mathcal{D}$, we motivate and describe a stratified sampling framework based on auxiliary variables in the EMR, so that resulting samples are more informative compared to using simple random sampling (SRS) for predictive model development. We demonstrate the resource efficiency for this class of designs, as well as statistically characterize design impact on model development and model validation.

Aim 2: Multi-label surrogate-guided sampling designs. We extend Aim 1 to handle multivariate outcomes $\mathbf{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_K)$, $Y_{ki} \in \{0, 1\}$, for $K > 1$. Radiology report processing motivates this aim where multiple non-exclusive binary findings are desired.

Aim 3: Predictive Case Control designs. When prediction model $\hat{h}(.)$ is developed in a source cohort but is then also applied to a new setting, model recalibration and revision are necessary steps for establishing external validity. For such modification learning to a new setting, we motivate and describe a class of designs based on original model predicted scores to collect "new" outcome labels.

Chapter 2

SURROGATE-GUIDED SAMPLING DESIGNS FOR CLASSIFICATION OF RARE OUTCOMES

2.1 Introduction

Electronic medical record (EMR) databases provide a potentially massive reservoir of information to help researchers understand and treat both common and rare medical conditions. Towards meaningful use of EMR data for research, an important first step is the accurate identification of key clinical outcomes from the EMR. For example, accurately extracted clinical outcomes can be used to investigate potential disease risk factors, compare treatment effects, as well as identify inherited and environmental components affecting health outcomes.

To understand how clinical outcomes may be identified using EMR data, consider categorizing data elements as either structured or unstructured [103]. Structured data is generally easily queried and analyzed; examples include demographics, lab values, and International Classification of Disease (ICD) and Current Procedure Terminology (CPT) codes. Unstructured data is usually generated at point of care and do not have fixed formatting; examples include free-text radiology reports, clinical notes, and raw image pixels. To establish research-quality clinical outcomes, querying structured data alone is insufficient, due to reasons such as over-coding for reimbursement and under-coding to protect patient privacy [93, 136]. Alternatively, unstructured data elements may be manually reviewed by highly trained clinical experts, who ascertain resulting true outcome statuses through a process called clinical abstraction. Traditional abstraction provides highly accurate clinical outcome statuses, but requires expensive clinician time and therefore is not feasible at scale for massive EMR databases.

In recent years, machine-learning classification algorithms (also called "classifiers") have shown promise towards accurate and scalable clinical outcome identification procedures. Such successes have been documented in domains ranging from radiology (natural language classification of pneumonia from chest x-ray reports [20]), rheumatology (phenotyping of rheumatoid arthritis from structured and unstructured data [18]), and dermatology (image classification of skin cancer [44]). Modeling machine-learning classifiers requires a sample that is labeled with actual case (outcome = 1) and control (outcome = 0) statuses. Such a sample, also called "labeled data", represents true clinical outcome statuses and in practice often collected through abstraction. Compared to traditional abstraction, the value of using machine-learning is that modeling requires manual review for a smaller sample instead of for all subjects. Yet, accurate classification may still require large amounts of labeled data, often in the thousands or larger. Alternatively, when available resources are constrained, classification accuracy may be improved with targeted sampling procedures. Such targeted sampling may be formulated as a study design task, where among the many currently unanswered questions include guidance for formal sampling methodologies, appropriate sampling variables, criteria to measure sample information, as well as relevant sample size calculations.

This paper is motivated by the urging need for formal statistical frameworks to guide sampling decisions for labeled data collection through abstraction towards accurate and scalable machine-learning of clinical outcomes. We specifically focus on the rare outcome scenario, where model accuracy is often rate limited by the number of cases. Our proposed framework assumes that the collected sample may be used for both model development and validation, therefore simultaneously requiring both improved outcome classification accuracy as well as characterization of sampling impact on validity and generalizability. As with conventional intuition, our proposed strategy targets sample case-enrichment of rare outcomes. The key contribution of our work is the formalization of such heuristics using sampling methodologies drawn from the fields of machine-learning and epidemiology, therefore filling a critical gap

in EMR research methods.

2.2 Background

The development of a classification model requires an adequate sample where actual outcome statuses $Y \in \{0, 1\}$ are labeled, so that patterns relating features $\mathbf{X}$ to Y may be learned from the sample for accurate prediction. Features $\mathbf{X}$ are typically high-dimensional predictors where exact definitions are domain specific; for example in natural language processing (NLP) applications feature engineering may be based on counts of individual words. When labeled data is not already available, a subset of the available population is sampled, typically through simple random sampling (SRS), and ascertained for their true outcome statuses. However, when the outcome is rare, the resulting sample may not be sufficiently informative for classifier learning, and alternative sampling procedures should be considered in order to improve ultimate prediction accuracy. Note that we use the terminology "classification" and "prediction" interchangeably referring to model outputs on yet unseen data, and do not explicitly require or discourage temporal interpretations.

2.2.1 Sampling methods in machine-learning

Classifiers have been shown to perform best when trained on samples with approximate outcome class balance, where the proportions of cases ($Y = 1$) and controls ($Y = 0$) are nearly equal. Such results have been observed regardless of the naturally occurring outcome distributions, both empirically [138, 7, 135] and to a lesser extent derived theoretically [142]. For rare outcomes, using SRS results in heavily imbalanced data sets: most individual outcomes are controls ($Y = 0$). In machine-learning applications, sampling methods such as random under-sampling (RUS) and over-sampling (ROS) are commonly used to re-balance the outcome class distribution, in order to hopefully improve model development and resulting prediction accuracy [59]. Under-sampling eliminates subjects from the majority class (typically Y=0), and has been criticized for removing data that is both costly to obtain and could be potentially informative. Over-sampling replicates subjects from the minority class

(typically Y=1), and could risk model over-fitting to exact copies in the training set.

Extensions to these basic "re-balancing" methods attempt to address the limitations of random sampling and the use of exact replicates. For example, a non-random under-sampling algorithm was introduced, where only controls with different features than cases are removed, since they do not provide additional information for learning [71]. Synthetic Minority Over-sampling Technique (SMOTE) combines random under-sampling with non-random over-sampling to avoid the exact replicate problem, where artificial cases are created based on interpolating features [24]. In addition, artificial cases can be generated adaptively, so that the decision boundary is shifted towards outcomes that are more difficult to learn [58].

The reviewed under-sampling or over-sampling strategies are more appropriately described as *analysis*-based re-sampling procedures after labeled data is collected, but prior to model development. While re-balancing outcome class distributions may allow for increased learning performance, the fundamental assumption is that an initial sample of labeled data is readily available. In most real-world EMR settings, labeled data collection requires the expensive and time-consuming process of clinical abstraction. Unfortunately, as noted in [138], most machine-learning re-sampling methods do not directly address the cost of labeled data collection, which often constrains the development of prediction models in biomedical research [16].

2.2.2 Sampling designs in epidemiology

When data collection resources are scarce, targeted sampling methods in epidemiology have offered alternative research designs that have facilitated investigation into causes of diseases. For example, when the outcome is rare and collecting predictors is expensive, the case-control sampling design [101] allows estimation of predictor effects 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.

Several extensions of case-control designs have also been proposed. For example, efficient estimation of predictor effects is still possible, even when the control sub-sample is replaced by a simple random sample [100], or has misclassified outcomes [73]. Sampling can also be based on variables other than the outcome. For example, the two-phase stratified sampling design [90] collects observations according to mutually exclusive strata based on an auxiliary variable, and reduces estimation variability if there is heterogeneity among the strata.

In contrast to machine-learning sampling methods, these epidemiological sampling designs can also be thought of as *design*-based sampling procedures for data collection. However, while epidemiological sampling designs provide resource-efficient alternatives to simple random sampling, their benefits are usually discussed in the context of unbiased and efficient inference for predictor effects, instead of model prediction accuracy. Furthermore, to our knowledge, epidemiological study designs have not been formally adopted in labeled data collection efforts in EMR settings.

To summarize the similar ideas that have been developed in parallel literatures, we present a summary of sampling methods in machine-learning and epidemiology in Table 2.1.

2.2.3 Surrogates derived from EMR data elements for sampling

Assuming that clinical abstraction provides the most accurate ascertainment of true outcome statuses, ideally clinical outcomes of all subjects would be abstracted, which unfortunately is cost prohibitive for massive EMR databases. An imperfect alternative may be based on readily and easily available related structured data elements, such as ICD codes and keyword regular expressions. ICD codes are variables coding for diseases, symptoms, and abnormal findings. Keyword regular expressions are sequences of characters that can be used to define a search pattern, for example synonyms related to a medical outcome. In particular, sum-

Table 2.1: Summary of sampling methods in machine-learning and epidemiology.

	Epidemiology (Design)	**Machine learning** (Analysis)
Study Goal	Estimate effect of X on Y	Predict Y from $\mathbf{X}$
Sampling Characteristics	Pre-specified design before data collection	Re-sampling collected data set
Sampling Goal	Efficient estimation	Accurate prediction
Selected Specific Examples	Case-control Case-cohort Stratified random sampling	Under-sampling Over-sampling

maries of ICD codes and keywords are often fairly specific for clinical outcomes that may be extracted from free-text clinical reports [30, 11]. To define summaries of structured data elements, querying is often based on counts restricted within pre-specified time frames, for example the number of ICD codes related to a clinical outcome within 90 days of a subject receiving a diagnostic imaging report.

In the medical informatics literature, summaries of relevant ICD codes and/or keywords have been referred to as "surrogates" (of actual clinical outcomes) and used in machine-learning modeling tasks. For example, surrogates defined using ICD and keyword counts were used as misclassified outcomes to reduce the dimensionality of EMR-generated features [146]. The approach in [146] was demonstrated to reduce the sample size requirements for labeled data abstraction towards accurate and scalable machine-learning model development. In addition to dimension reduction, surrogates have also been directly used as "noisy" imputed outcome labels for classifier development [51, 1]. However, model development using such misclassified outcomes may seriously compromise validity of using resulting model predictions for downstream analyses [114], therefore using surrogates to replace abstracted clinical outcomes as labels may not be fully justified.

Alternatively, surrogates could help guide selection of subjects for abstraction of labeled clinical outcomes. In fact, in developing classification models for rare clinical outcomes, several papers have described the usage of ICD codes and keywords for subject selection. For example, to develop a classifier for congestive heart failure from clinical notes, subjects were selected based on coded diagnoses, as well as "term spotting", which are keywords likely to predict diagnosis [95]. Another application selected subjects based on one specific ICD code as well as non-negated keywords towards building a classification model for angina [94]. Despite some usage of such sampling strategies, there remains little discussion of corresponding statistical rationale. In particular, such heuristic decisions based on purposeful biased sampling may not create generalizable predictions, or valid summaries of accuracy.

Our contribution is to formalize a sampling strategy to select subjects for true clinical outcomes abstraction, based on using easily accessed summaries of structured data as enrichment surrogates. Such true clinical outcome statuses, instead of surrogate statuses, will then be used as labeled data for machine-learning. In Section 2.3.1, we frame the statistical problem, and the describe the proposed framework in Section 2.3.2. We demonstrate the resource efficiency of the proposed design for model development in Section 2.3.3, and discuss sampling impact for both model development and validation in Section 2.3.4. We provide empirical evidence through simulations in Section 2.4. Finally, in Section 2.5 we illustrate the method on a data set of lumbar spine imaging reports that was obtained in a pragmatic trial of radiology decision support [65], and provide a concluding discussion in Section 2.6.

2.3 Methods

2.3.1 Statistical motivation and notation

For subject i denote $\tilde{X}_i \in \mathcal{R}^p$ 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, where a commonly used model is penalized logistic regression

$$\hat{\beta}_0, \hat{\tilde{\beta}}_X = \underset{\beta_0, \tilde{\beta_X}}{min}\{-\sum_{i=1}^{n} Y_i(\beta_0 + \sum_{j=1}^{p} \beta_{X_j} X_{ij}) + \log(1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_{X_j} X_{ij})) + \lambda \sum_{j=1}^{p} ||\beta_j||_L\}. \quad (2.1)$$

The development of classification models such as (2.1) requires a sample $\mathbf{D}^S(n)$ of size n, drawn from the EMR cohort $\mathcal{D}$. In $\mathbf{D}^S(n)$, features $\tilde{X}_i^T$ and outcome Y_i need to be available for all subjects. We assume that feature engineering is relatively cheap, but obtaining outcome statuses is time-consuming and costly due to necessary abstraction. Therefore, we consider the statistical problem of drawing $\mathbf{D}^S(n)$ from $\mathcal{D}$, where subjects selected in $\mathbf{D}^S(n)$ will have their outcome status ascertained and labeled for subsequent machine-learning modeling, specifically for scenarios when Y is expected to be rare.

For concreteness, consider the task of classifying radiology reports for subject vertebral fracture status. In the rest of this paper, we discuss applications and illustrations specifically for this sub-domain of biomedical Natural Language Processing (NLP); however, our sampling framework may be generalized to machine-learning modeling tasks involving expensive labeled data collection of rare outcomes. In NLP, a common feature engineering technique involves converting unstructured free-text into numeric matrices using bag-of-words (BOW) representations. For BOW, the set of p unique terms, denoted as $T = \{t_1, \ldots, t_p\}$ is first obtained by concatenating all reports in $\mathcal{D}$. Then, for subject i the BOW feature vector $\tilde{X}_i$ has binary elements

$$X_{ij} = I(t_j \in \text{report}_i). \quad (2.2)$$

For the outcome defined as

$$Y_i = \begin{cases} 1, & \text{if the report of subject } i \text{ indicates vertebral fracture} \\ 0, & \text{otherwise} \end{cases}, \tag{2.3}$$

data processing is not as straight-forward. Realistically, simple keyword searches are inadequate as radiology reports often contain negated or uncertain findings. Therefore, the definition of (2.3) often requires abstraction, where human clinical experts interpret report text and label true outcome statuses.

Our main motivation is the burden of clinical abstraction for labeling rare outcomes. Due to the expected low number of cases, using simple random sampling (SRS) to select reports for abstraction is often inadequate. To artificially increase sample prevalence, a common procedure is oversampling, 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 prevalence can be collected by design. For subject i additionally assume that we have access to an enrichment surrogate Z_i, which is an auxiliary variable that approximates the outcome better than random noise. In practice, surrogates may created from summaries of relevant structured data elements, such as ICD and keyword counts. In the motivating problem for vertebral fracture, practical examples include

$$Z_i = I(\text{report for subject } i \text{ contains the keyword “fracture”})$$

and

$$Z_i = I(\text{subject } i \text{ has a relevant ICD code in the EMR}),$$

where examples of relevant ICD codes for vertebral fracture include `809: Fracture of bones of trunk, closed`. Note that the constructed Z is not without error in predicting Y: the keyword fracture could be negated, and ICD codes may be missed or over-utilized. Therefore, while surrogates may not be appropriate as a substitute for true clinical outcomes, they may be used as sampling variables to select subjects for clinical outcome abstraction, where resulting labeled data is subsequently used for machine-learning classifier modeling.

2.3.2 Surrogate-guided sampling (SGS) designs

For sample selection based on surrogate values, we define the resulting class surrogate-guided sampling (SGS) designs (Definition 1). Under the SGS design, all subjects in cohort $\mathcal{D}$ are divided into two strata based on surrogate values: surrogate positives with $Z_i = 1$, and surrogate negatives with $Z_i = 0$. The SGS sample $\mathbf{D}^{SGS}(n)$ is assembled by over-representing the surrogate positive stratum. Formally, let $R = P(Z = 1|S = 1)$ be the sample proportion of surrogate positives, then $R \geq 0.50$ for SGS designs. The intended benefit of SGS designs is that, for the same abstraction cost, resulting samples have higher expected outcome prevalences compared to using SRS.

Definition 1 *Surrogate-guided sampling (SGS) design class.*
Denote the surrogate-guided sampling (SGS) design class as the class of stratified sampling procedures based only on values of a binary enrichment surrogate $Z \in \{0, 1\}$.

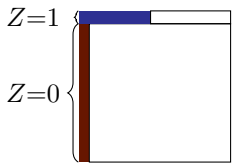
For illustration, consider an outcome prevalence of 10%, and assume that in the EMR, there exists a surrogate with 40% sensitivity and 95% specificity for the outcome of interest. For an abstraction budget of collecting $n = 500$ labels, using an SGS design with three-times as many surrogate positives as surrogate negatives (SGS 3:1) with $R = 0.75$ 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, a 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.

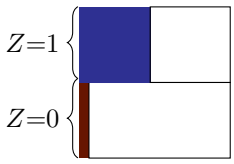
Figure 2.1 illustrates the relative case-enrichment of SGS compared to SRS, where the area of the shaded (blank) regions indicate expected case (control) proportions. As indicated by the shaded regions on the left-most subplot, 10% of a sample collected with SRS are true cases. However, the distribution of the cases are unequal across surrogate strata, where the case proportion is much higher in the surrogate positive ($Z = 1$) stratum. Under the naturally occurring proportions, as in the SRS scenario, surrogate positives only make up 8.5% of the sample. By over-representing the surrogate positive stratum, SGS designs increase the expected sample case proportion by construction.

Figure 2.1: Illustration of expected sample case proportions for simple random sampling (SRS) and surrogate-guided sampling (SGS) designs. Illustrations are based on a scenario with outcome prevalence 10%, and surrogate with sensitivity 40% and specificity 95% for the outcome of interest.

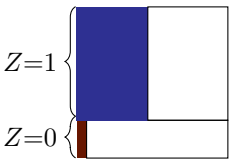
(a) SRS, $R = 0.085$.



(b) SGS 1:1, $R = 0.50$.



(c) SGS 3:1, $R = 0.75$.



After selection into the sample, subjects are then ascertained and labeled for their true outcome statuses, $Y \in \{0, 1\}$, through human clinical expert manual abstraction. In the next subsections, we characterize the proposed SGS sampling design in terms of its resource efficiency for machine-learning classification tasks, as well as sampling impact on model development and model validation.

2.3.3 Design resource efficiency for model development

We highlight a property of using SGS designs for classification model development, which we formulate as "resource efficiency". In general, design resource efficiency refers to requiring a lower cost (i.e. abstraction sample size) in order to achieve modeling goals (i.e. prediction accuracy) compared to a baseline design, typically simple random sampling (SRS). In order to formally demonstrate the design resource efficiency of using SGS designs to procure development samples, we first represent generalizable prediction accuracy in terms of development sample composition.

Model discrimination in terms of estimation variance

For tractability we focus on a specific 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 [53] in yet unseen data. Consider resulting continuous predictions as a "test" for true outcome statuses. Then, under a bi-normal assumption, the AUC has been shown to be [97]

$$AUC = \Phi(\sqrt{R_{AUC}}) = \Phi\left(\sqrt{\frac{(\mu_1 - \mu_0)^2}{\sigma_1^2 + \sigma_0^2}}\right). \tag{2.4}$$

In (2.4), μ_y and σ_y^2 are the means and variances of the "test" among the cases ($y = 1$) and controls ($y = 0$). The bi-normal AUC formula (2.4) was developed in [97] 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 classifi-

cation modeling and diagnostic testing settings. Denote $\mathbf{D}^S(n)$ 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 $\mathcal{D}$. Then, the validation AUC for model developed with $\mathbf{D}^S(n)$ may be represented using an indexing as shown in Definition 2.

Definition 2 *$AUC(Y|\mathbf{D}^S(n))$.*
Let $AUC(Y|\mathbf{D}^S(n))$ denote the validation AUC of a classification model for outcome Y developed using sample $\mathbf{D}^S(n)$, which is generated with sampling design S and has 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 ($(\mathbf{X}|Y=y) \sim N(\mu_{x|y}, \Sigma_{x|y})$). Theorem 1 (proof in Appendix A.0.1) shows that validation AUC is inversely proportional to the estimation variance and the data signal-to-noise ratio. Therefore, conditioned on using the same modeling procedure, using a design with higher statistical information as measured with 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.

Theorem 1 *Assume that in $\mathcal{D}$, for $y \in \{0,1\}$, $(\mathbf{X}|Y=y) \sim N(\mu_{x|y}^T, \mathbf{\Sigma}_{x|y})$, where $\mu_{x|y=0} = 0$ and $\mathbf{\Sigma}_{x|y=1} = \mathbf{\Sigma}_{x|y=0} = \mathbf{\Sigma}_{x|y}$. Let $\hat{\eta} = \mathbf{X}\hat{\beta}$ be the estimated linear predictor, where model coefficients $\hat{\beta}$ are estimated by logistic regression using development sample $\mathbf{D}^S(n)$. Then,*

$$AUC(Y|\mathbf{D}^S(n)) \propto \frac{1}{trace(\sigma \boldsymbol{V}(\hat{\beta}^S(n)) + \mu^T \boldsymbol{V}(\hat{\beta}^S(n))\mu}, \tag{2.5}$$

where $\boldsymbol{V}(\hat{\beta}^S(n)) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ is the approximate covariance matrix of estimating $\hat{\beta}$ using $\mathbf{D}^S(n)$, and $\mu = \mu_{x|y1}$ and $\sigma = \mathbf{\Sigma}_{x|y}$ are parameters describing the data signal-to-noise ratio.

The relationship between validation AUC and estimation variance as represented in Theorem 1 further allows application of existing results specifically for logistic regression models and more generally for two-phase sampling designs. 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 [70], and may may provide an explanation for the empirical observations in machine-learning that outcome class balance affects classifier prediction performance.

We remark that while our argument in Theorem 1 is based on logistic regression model coefficients $\hat{\beta}$, modeling bi-normal features may also proceed with Linear Discriminant Analysis (LDA) [46] where instead feature means $\mu_{x|y}$ and covariances $\Sigma_{x|y}$ are directly estimated. Under the LDA assumption, the relationship between outcome class balance and AUC has been studied analytically [142]. However, our argument is slightly different than [142], as we were more concerned with a general representation of validation AUC in terms of development sample composition, rather than investigating specifically the "best" outcome class distribution to maximize AUC. In fact, the results from 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 [97]. 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.

O_{ratio} as a measure of design effect on sample outcome prevalence

We demonstrated a relationship between development sample composition and model validation AUC through the estimation variance. Even though generalizable prediction accuracy is the ultimate measure of resource efficiency for model development, the analytical representation of the estimation variance in terms of sampling design still requires numeric approximations. However, motivated by empirical results in machine-learning, we may use sample outcome prevalence as another measure for design resource efficiency. Here, we char-

acterize the design effect of SGS on sample outcome prevalence by using a summary measure of design case enrichment that we call O_{ratio}.

Consider measuring the "effect" of a sampling design on sample outcome prevalence. 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 sample 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

$$O_{ratio} = \frac{\frac{E^{\mathbf{D}^{SGS_{(n)}}}[Y|S=1]}{1-E^{\mathbf{D}^{SGS_{(n)}}}[Y|S=1]}}{\frac{E^{\mathbf{D}^{SRS_{(n)}}}[Y|S=1]}{1-E^{\mathbf{D}^{SRS_{(n)}}}[Y|S=1]}} = \frac{Odds(cases|SGS)}{Odds(cases|SRS)}. \tag{2.6}$$

The denominator of (3) 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 case enrichment. Therefore, O_{ratio} provides a one-dimensional summary measure of the resource efficiency comparing SGS over SRS.

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 Receiving Operating Characteristics (ROC) curves, are related to positive and negative predictive values (PPV & NPV), but are invariant to outcome prevalence [27]. 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 (Equation (2.8)) shows that O_{ratio} is approximately the sum of positive and negative surrogate likelihood ratios ($LR+$ and $LR-$), weighted by the SGS sampling ratio R. In general, a "good" test requires having high values of both $LR+$ and $LR-$. However, since R may be set by design to approach 1, having a high $LR+$ alone is sufficient to achieve a high O_{ratio}, which is a measure of sample case enrichment. In other words, the properties of an variable to be a good enrichment surrogate for sampling is weaker than what may be required for a good diagnostic test.

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

$$O_{ratio}(n, R, Z) = \frac{RZ_{sens} + p_Z(1 - R - Z_{sens})}{R(1 - Z_{spec}) + p_Z(Z_{spec} - R)}. \tag{2.7}$$

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

$$O_{ratio}(n, R, Z) \approx (R)(LR+) + (1 - R)(LR-), \tag{2.8}$$

where

$$(LR+) = \frac{Z_{sens}}{1 - Z_{spec}} = \frac{\frac{P(Y=1|Z=1)}{P(Y=0|Z=1)}}{\frac{P(Y=1)}{P(Y=0)}}, \quad (LR-) = \frac{1 - Z_{sens}}{Z_{spec}} = \frac{\frac{P(Y=1|Z=0)}{P(Y=0|Z=0)}}{\frac{P(Y=1)}{P(Y=0)}}.$$

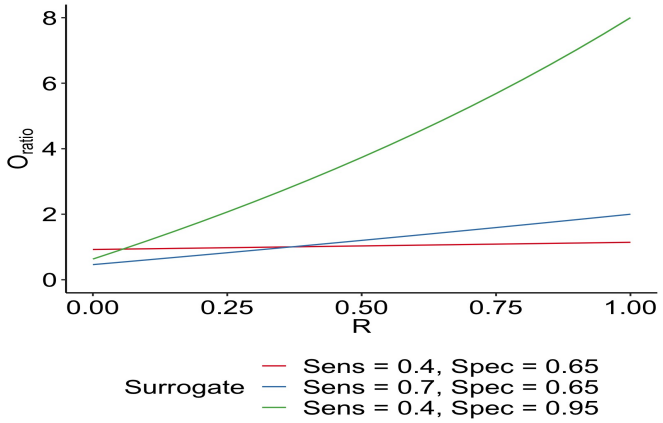
Corollary 1 *For a given Z, $O_{ratio} \propto R$. Over the set of possible Z, $O_{ratio} \propto Z_{sens}$ and $O_{ratio} \propto \frac{1}{1 - Z_{spec}}$.*

Corollary 1 follows directly from (2.8), and highlights the effect of surrogate operating characteristics on sample case enrichment. While the rate of increase in O_{ratio} is linear in Z_{sens}, it is inverse polynomial in $1 - Z_{spec}$. Therefore, a small change in specificity can have a much higher impact on O_{ratio} compared to the same change in sensitivity. Therefore, in choosing among multiple potential EMR data elements to be used as surrogates, variables with higher specificities can generally provide more enrichment. In practice, this can be achieved by setting higher thresholds when calculating summaries of structured data elements.

We provide two further illustrations to further emphasize the impact of surrogate specificity on O_{ratio}. Figure 2.2 shows O_{ratio} on the y-axis as a function of sampling ratio R on the x-axis for three different surrogates Z defined by their marginal operating characteristics, where O_{ratio} is approximately linear in R for this relatively rare outcome (prevalence = 10%). Here, notice that the surrogate with sensitivity of 0.40 and specificity of 0.65 results in low O_{ratio} for all values of R. With the same increase of 0.30, the more specific surrogate had a greater effect on overall O_{ratio} compared to the more sensitive surrogate. Figure

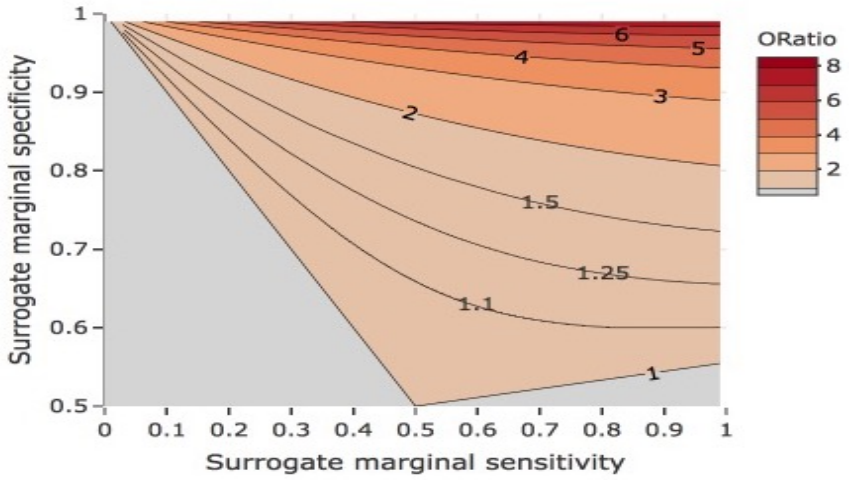
2.3 shows values of O_{ratio} indicated by different colors across possible ranges of surrogate marginal sensitivities and specificities for an SGS design with fixed $R = 0.50$. The non-gray regions of Figure 2.3 illustrates operating characteristics of surrogates that constitute good candidates for 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 2.3, 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.

Figure 2.2: O_{ratio} versus sampling ratio R when using SGS on surrogates of different operating characteristics for an outcome with prevalence of 10%.



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 sta-

Figure 2.3: 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%.



tistical learning.

2.3.4 Design impact on model development and model validation

To improve the resource efficiency of resulting samples for machine-learning, SGS designs intentionally over-represents surrogate positives. A concern, then, is if such introduced bias may impact modeling. The impact of sample characteristics on machine-learning was first formalized in [147], and can be formulated as a missing data problem [78]. Recall that sampling in SGS only depends on surrogate values Z, which are assumed to be available for all subjects in $\mathcal{D}$. Therefore, for the SGS design, sampling is independent of outcome labels conditional on surrogate values, equivalently

$$S \perp Y | Z. \tag{2.9}$$

The assumption (2.9) is also called Missing At Random (MAR) [78]. Using the MAR assumption, we now describe the impact of using SGS designs for both model development

and model validation.

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 $\mathbf{D}^S(n)$ obtained with sampling design S, [147] suggested that S may be used for developing model $\hat{h}(.)$ "validly" under the asymptotic equivalence criteria,

$$\lim_{n\to\infty} \hat{h}(\mathbf{D}^S(n)) = h(\mathcal{D}), \tag{2.10}$$

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 $\mathbf{D}^S(n)$ having outcomes MAR from $\mathcal{D}$ are "valid" for model development of classifiers based on conditional means in the asymptotic "true model" sense [147]. Note that for logistic regression, (2.9) implies that

$$logit(E[Y|\mathbf{X}, Z, S = 1]) = logit(E[Y|\mathbf{X}, Z]). \tag{2.11}$$

Therefore, machine-learning model development with logistic regression using SGS samples results in validly estimated models under this interpretation.

Design impact on model validation

Now, consider the impact of using sample $\mathbf{D}^S(n)$, where S is the SGS design, on model validation. 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 generally 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 [63] adjusts empirical estimates according to sampling weights. 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 [3]

$$AUC_{IPW} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \pi_i^{-1}\pi_j^{-1} I(\hat{p}_i > \hat{p}_j) I(Y_i > Y_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \pi_i^{-1}\pi_j^{-1} I(Y_i > Y_j)}. \tag{2.12}$$

In (2.12), $\pi_i = P(S_i = 1)$ is the sampling weight for subject i, and may be estimated from observed data for any MAR sample. For the SGS design, π_i is additionally known by construction to be

$$\begin{aligned} \pi_i = P(S_i = 1 | Z_i = z) &= \frac{P(Z_i = z | S_i = 1) P(S_i = 1)}{P(Z_i = z)} \\ &= \begin{cases} \dfrac{R}{p_Z} \times \dfrac{n}{N}, & Z_i = 1 \\ \dfrac{1-R}{1-p_Z} \times \dfrac{n}{N}, & Z_i = 0. \end{cases} \end{aligned} \tag{2.13}$$

The known sampling weights (2.13) may be directly used in IPW-corrected accuracy metrics such as (2.12). Note that the AUC indexing described in Section 2.3.3 is slightly different than the AUC estimator in (2.12). In Section 2.3.3, 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 (2.12), 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.

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 guidances 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 weights 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 in (2.11) is estimable only on the $Z = 1$ stratum and π_i in (2.13) is undefined for the $Z = 0$ stratum, without further parametric assumptions.

2.4 Simulations

To illustrate the benefit of using SGS designs for statistical machine-learning model development, we generated scenarios with features simulated according to a normal distribution (Section 2.4.1) and a Bernoulli distribution (Section 2.4.2), as well as modeling using unpenalized and penalized logistic regression. The sampling methods we compared included 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 all data generating mechanisms, we generated a cohort of size $N = 100,000$ and three

potential enrichment surrogates $Z1$, $Z2$ and $Z3$, defined with parameters so that they have operating characteristics with respect to the actual outcome Y as described in Table 2.2. Motivated by the illustrations in Figure 2.3, we defined three surrogates with expected O_{ratio} values of much greater than 2 ($Z1$), between 1 and 2 ($Z2$), and less than 1 ($Z3$). We remark that surrogate $Z1$ has very similar operating characteristics to a real-world surrogate (discussed in Section 2.5), while surrogates $Z2$ and $Z3$ may be viewed as "weaker" surrogates for sampling.

Table 2.2: Sensitivity, specificity, AUC (calculated using trapezoidal rule), and expected design O_{ratio} for the three potential enrichment surrogates.

Surrogate	Sensitivity	Specificity	AUC	O_{ratio}(SGS 1:1)	O_{ratio}(SGS 3:1)
$Z1$	0.40	0.95	0.67	3.30	5.27
$Z2$	0.67	0.66	0.67	1.75	1.89
$Z3$	0.60	0.30	0.45	0.88	0.86

For each simulated cohort, a large validation sample $\mathbf{D}^{val}$ with sample size $n_{val} = 10000$ was set aside using SRS. From the remaining examples $\mathcal{D}\backslash\mathbf{D}^{val}$, we simulated "abstraction samples" $\mathbf{D}^{S}(n)$ varying across a grid of sample sizes, and sampling methods of SRS, ROS, SGS 1:1 ($R = 0.50$) or SGS 3:1 ($R = 0.75$), where SGS may be based on surrogate $Z1$, $Z2$, or $Z3$. 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 parameter combination, $B = 500$ samples were drawn without replacement from $\mathcal{D}\backslash\mathbf{D}^{val}$ based on the specified sampling scheme. On each iteration, we fitted a classification model and applied resulting estimates to $\mathbf{D}^{val}$, calculating the empirical validation AUC

using the Wilcoxon-Mann-Whitney formula. Over all 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 [144]. In these experiments, since we compared prediction accuracy across different sampling designs conditioned on the same models and feature sets, the difference in model performance is due to differences in the sampling design that gave rise to resulting samples.

2.4.1 Features with normal distribution

We first generated features according to a bi-normal distribution to empirically demonstrate the analytical conclusions in Section 2.3.3, which is that, conditioned on the same data generating mechanism and modeling approach, model development using SGS designs results in higher validation AUC compared using an SRS sample of the same size.

Data generating mechanism

We generated two cohorts (A and B), with outcome Y having 5% and 10% prevalence respectively. For each cohort we generated $p = 50$ predictors, of which 10 had non-zero coefficients, where

$$\begin{aligned}\left(Z1_i^{mvn}, Z2_i^{mvn}, Z3_i^{mvn}, \tilde{X}_i\right) \mid (Y_i = 1) &\sim N\left(\left[1.3, 0.9, -0.3, 0.5, \ldots, 0.5, 0, \ldots, 0\right]^T, \mathbf{I}_{3+p}\right) \\ \left(Z1_i^{mvn}, Z2_i^{mvn}, Z3_i^{mvn}, \tilde{X}_i\right) \mid (Y_i = 0) &\sim N\left(\left[0, \ldots, 0\right]^T, \mathbf{I}_{3+p}\right).\end{aligned} \tag{2.14}$$

In (2.14), $\mathbf{I}_{3+p}$ indicates the indicator matrix, and appropriate cut-offs were applied to $Z1^{mvn}$, $Z2^{mvn}$, $Z3^{mvn}$ to obtain $Z1$, $Z2$, $Z3$ so that the binary enrichment surrogates had operating characteristics as described in Table 2.2. The data generating mechanism described in (2.14) implies the logistic regression model

$$logit(E[Y_i|\mathbf{X}_i^T]) = -4.74 + 1.3Z1_i + 0.9Z2_i - 0.3Z3_i \\ + 0.5X_{i1} + \ldots + 0.5X_{i10} + 0X_{i11} + \ldots + 0X_{i50}.$$

Using (2.4), we calculated the maximum theoretical AUC in this data set to be 0.94.

Learning Curves

Figure 2.4 shows the learning curves from data generated according to (2.14). Dashed lines indicate the AUC of the individual surrogates in discriminating between cases and controls, according to the trapezoidal rule. Dotted lines indicate the maximum AUC for the data set. Each of the six panels shows that AUC increases with abstraction sample size, as expected. However, the rate of increase depends on sampling design: we discuss three observations comparing learning curves within each individual panel, across rows, as well as across columns in Figure 2.4. Individual panels demonstrate comparisons using different sampling designs (SRS, ROS, SGS 1:1, SGS 3:1), rows demonstrate comparisons when using different enrichment surrogates ($Z1$, $Z2$, or $Z3$) as sampling variables for the SGS designs, and columns demonstrate comparisons for different outcome prevalences (5% or 10%).

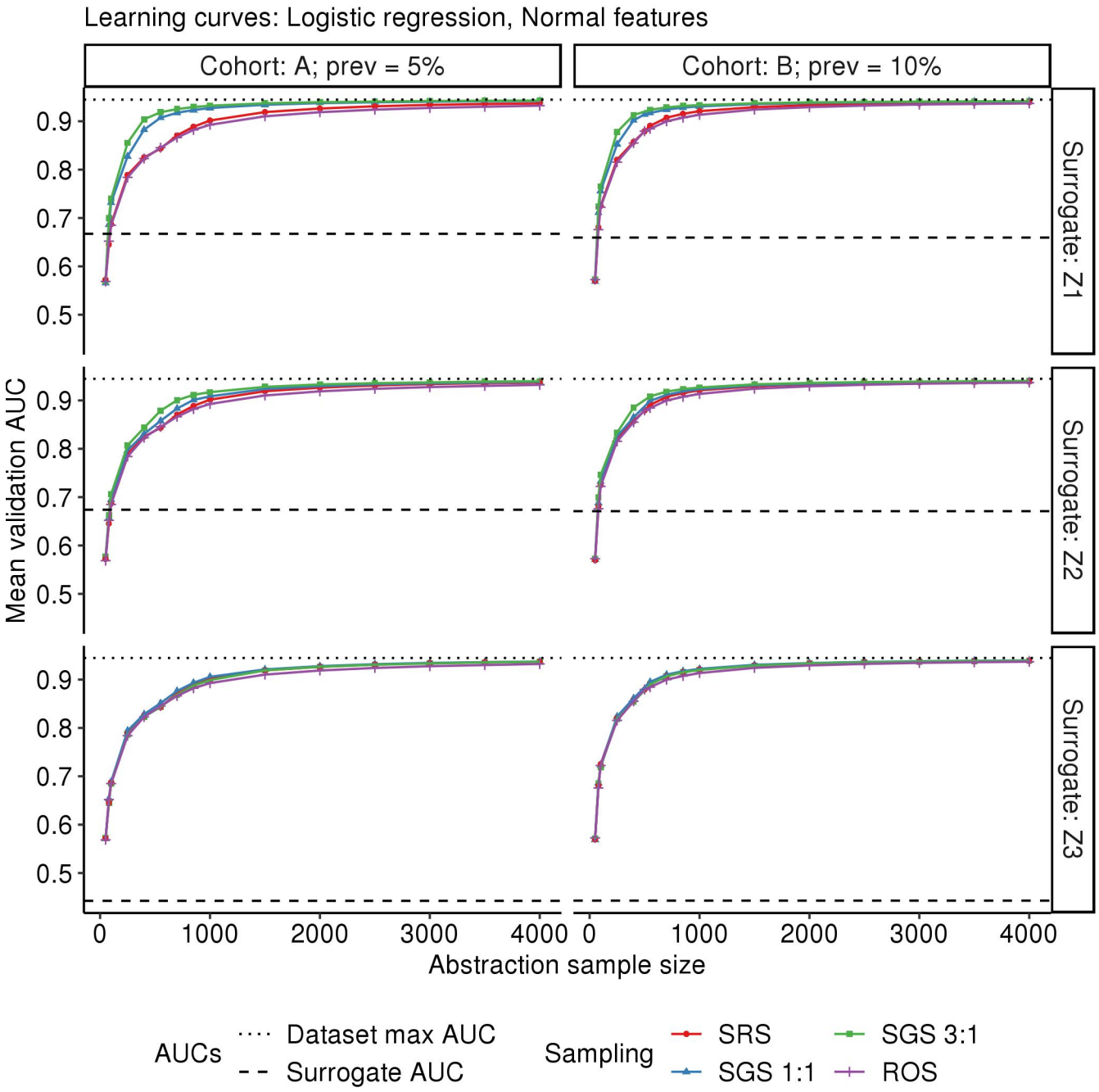
For comparisons within individual sub-plots, consider the top-left sub-plot, where $Z1$ is used as a surrogate for the SGS designs and the outcome has 5% prevalence. First, notice that over-sampling has virtually no improvement over SRS. Such an observation is unsurprising, since the correct mean model is fitted, exact replicates of observations do not provide additional information and thus estimation variance was not reduced. In fact, over-sampling results in some over-fitting at larger sample sizes. In contrast, SGS designs provided noticeably improved AUC over SRS. For example, with an abstraction sample size of 250, the average validation AUC using SRS samples was about 0.80 while the AUC using SGS samples based on $Z1$ was 0.84, when using a 1:1 ratio of sample surrogate positives to negatives.

The AUC using SGS 3:1 configuration was 0.85, a minor increase over SGS 1:1. Therefore, using an SGS design provided clear benefit over SRS. However, differences in AUC among the various sampling proportions is small relative to the difference between using an SGS or SRS design; drastically increasing the sampling ratio only provides diminishing returns.

For row comparisons, when the enrichment surrogate was $Z1$ (top rows), SGS designs provide noticeable benefit in terms of improved model discrimination at every sample size. Some benefit is seen when $Z2$ was used as an enrichment surrogate, but there was virtually no improvement over SRS samples when $Z3$ was used. This is because O_{ratio} as a measure of case enrichment is related to model discrimination. Even though $Z1$ and $Z2$ individually provide the same discrimination for Y, the higher O_{ratio} of using $Z1$ for SGS resulted in a more case-enriched sample. In fact, since the O_{ratio} of $Z3$ is less than one for all sampling ratios, using $Z3$ as an enrichment surrogate did not provide any resource-efficiency over SRS for model development.

For column comparisons, we see similar patterns for outcome prevalences of 5% and 10%, illustrating the benefit of SGS designs for outcomes with prevalences in this range. Additionally, we note that as abstraction sample size increases, the learning curves approach the same theoretical maximum, demonstrating that samples created with SGS are valid for model development under this special case of correct mean model specification.

Figure 2.4: Logistic Regression learning curves for bi-normal features, comparing simple random sampling (SRS), random over-sampling (ROS), and surrogate-guided sampling with 1:1 (SGS 1:1) or 3:1 (SGS 3:1) ratio of surrogate positives to negatives.



2.4.2 Binary features

In practical settings, clinical outcomes abstracted from unstructured text data motivates two main deviations from the scenarios presented in Section 2.4.1. First, text-derived features are often not normally distributed. Additionally, modeling approaches might utilize penalization. In this sub-section, we empirically demonstrate the effect of using SGS designs, when both the the feature distribution and modeling approaches are more typical for the real-world clinical text data setting.

Data generating mechanism

We generated two cohorts (A and B), each with outcomes Y having 5% and 10% prevalence. As most EMR data-sets contain features of high dimensionality, we set the number of features to be $p = 250$, of which only 30 were related to the outcome (non-zero coefficients). Text features can often be described as following a long upper tail distribution (Zipf's law), where the most common features are present in almost all reports, but the majority of features have very low frequencies [113]. Our simulated data set thus followed a logistic regression assumption, where features are simulated according to a Bernoulli distribution with marginal feature frequencies independently following an exponential distribution. Specifically, the data generating mechanism was

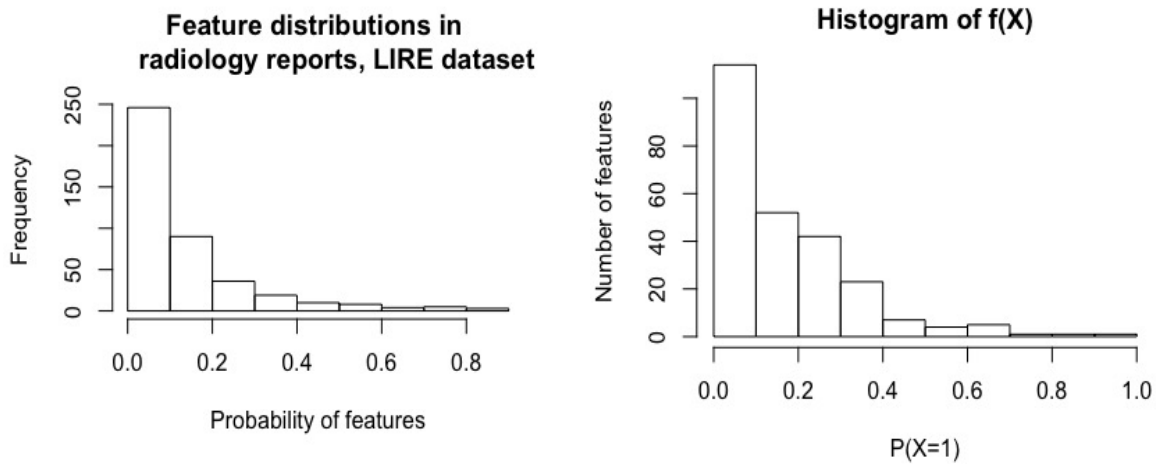
$$Y_i | \left(Z1_i, Z2_i, Z3_i, \tilde{X}_i^T \right) \sim \text{Bernoulli}(P(Y_i = 1))$$
$$logit(P(Y_i = 1)) = \beta_0 + \beta_{Z1} Z1_i + \beta_{Z2} Z2_i + \beta_{Z3} Z3_i + \sum_{j=1}^{p} \beta_j X_{ij}, \tag{2.15}$$

and the binary features $\mathbf{X}$ were generated as

$$\begin{aligned} \tilde{X}_j &\sim \text{Bernoulli}\left(p_{\tilde{x}_j}\right) \\ p_{\tilde{x}_j} &\sim \text{Exponential}\left(\text{mean} = \frac{1}{6}\right). \end{aligned} \tag{2.16}$$

The simulation parameters in (2.16) were selected so that resulting marginal feature distributions were comparable to the a real-world data set of radiology text reports from the Lumbar Imaging with Reporting of Epidemiology (LIRE) study [65] (discussed in Section 2.5; see illustration in Figure 2.5).

Figure 2.5: Histograms of number of features with binned proportions of $p_{\tilde{X}_j}$. Left plot shows distributions from BOW (unigrams) representations of LIRE radiology reports, and right plot shows simulated data using an Exponential $\left(\text{mean} = \frac{1}{6}\right)$ distribution.



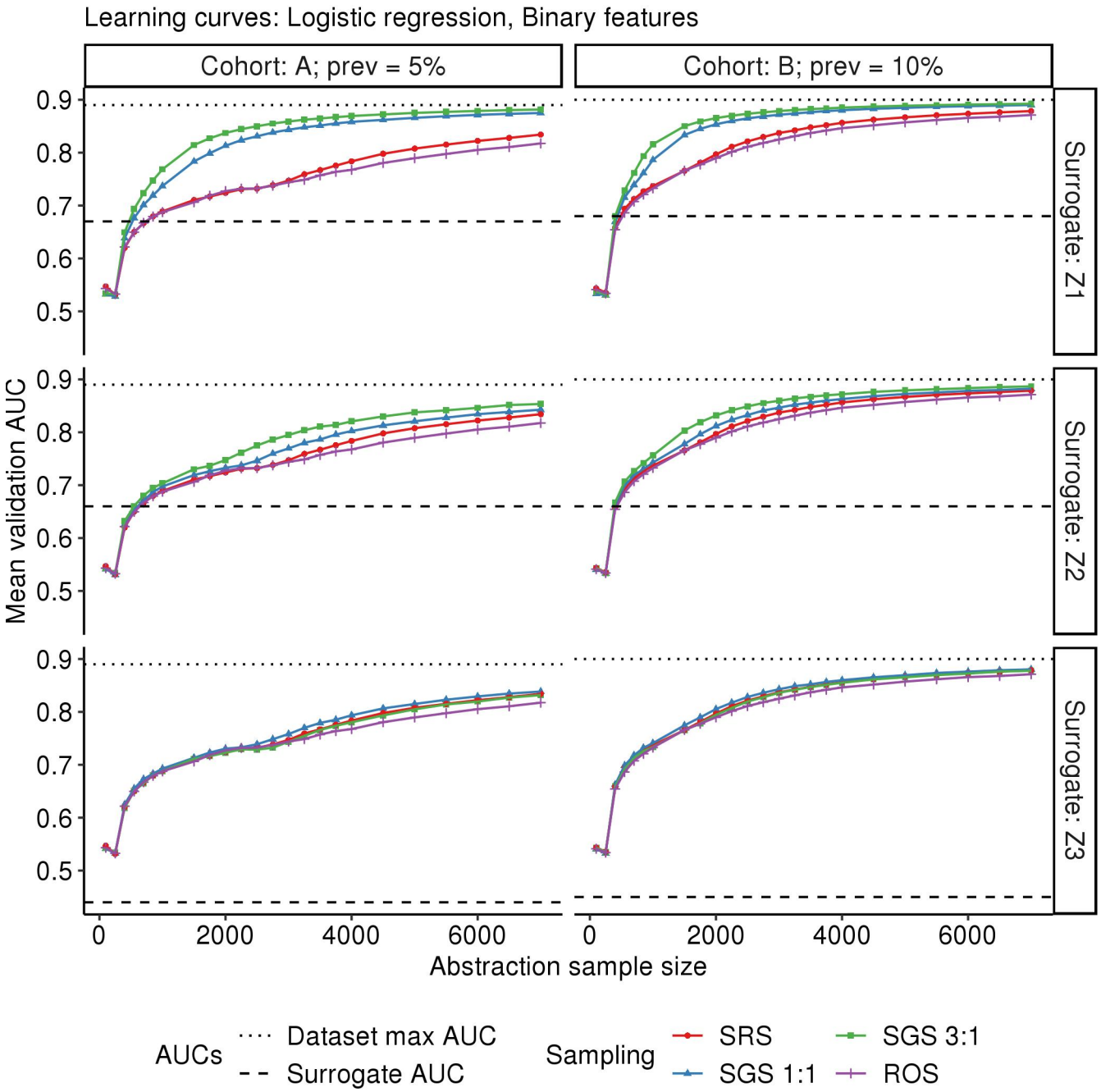
In (2.15), $\beta_j = (-0.5, 0.25, \ldots, -0.5, 0.25)$ for the first 20 most frequent features, $\beta_j = 1$ for the 10 features with frequencies closest to the outcome prevalence, and $\beta_j = 0$ for the remaining 220 features. Here, we used a simplifying assumption that the most predictive text-based features tend to occur as often as the outcome prevalence, frequent features are weakly predictive, but most features are irrelevant in predicting the outcome. Surrogates

$Z1$, $Z2$, and $Z3$ were generated as independent Bernoulli random variables with means and coefficients resulting in the operating characteristics as described in Table 2.2. The maximum AUC for this data set was about 0.90.

Learning Curves: Logistic regression

Figure 2.6 shows the learning curves from samples with features generated using (2.15). Similar to the results in Figure 2.4, validation AUC increases with sample size, but for every sample size we see differential performance depending on sampling design. Again, we see no benefit of using over-sampling but a notable benefit of using SGS compared to using SRS (comparing curves within a panel), that sampling using a design with high O_{ratio} improves discrimination (comparing rows), and that improved model performance is observed for both outcome prevalences of 5% and 10% (comparing columns). Note that the logistic regression model over-fitted for sample sizes of less than 250 due to high dimensionality. However, since our goal for this simulation study was to show the effect of sampling design and not modeling constraints, we were more concerned with model performance for development sample sizes greater than 250 for the results illustrated in Figure 2.6.

Figure 2.6: Logistic Regression learning curves for binary features, comparing simple random sampling (SRS), random over-sampling (ROS), and surrogate-guided sampling with 1:1 (SGS 1:1) or 3:1 (SGS 3:1) ratio of surrogate positives to negatives.



Learning Curves: Penalized logistic regression

Recognizing that regularization is almost always used in high-dimensional feature scenarios (where $p > n$), we next demonstrate the same learning curves but with using penalized regression models. We fitted both Lasso and Ridge regression [127, 74] using data generated according to (2.15). These experiments were run to illustrate the real-world scenario with text data, where regularization is also desired assuming feature sparsity. The regression coefficients were estimated with

$$\begin{aligned}\hat{\beta}_0, \hat{\beta}_Z, \hat{\beta} = \underset{\beta_0,\beta_Z,\beta}{min}\{&-\sum_{i=1}^{n} Y_i(\beta_0 + \beta_Z Z_i + \sum_{j=1}^{p}\beta_j X_{ij}) \\ &+ \log(1 + \exp(\beta_0 + \beta_Z Z_i + \sum_{j=1}^{p}\beta_j X_{ij})) \\ &+ \lambda\sum_{j=1}^{p}||\beta_j||_L\},\end{aligned} \tag{2.17}$$

where L was either 1 (Lasso) or 2 (Ridge). Note that in (2.17), 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. For the procedure (2.17), we estimated the regularization parameter λ based on values that maximized AUC using ten-fold cross-validation on development samples.

Figure 2.7 shows learning curves for Ridge regression. From the top-left sub-plot, we observed that SGS results in improved discrimination compared to SRS, but the increase is smaller when compared to that in Figure 2.6. This is likely due to that, unlike in the unpenalized regression setting, the estimation of Ridge regression coefficients introduced bias. Note that ROS resulted in worse performance compared to SRS in all simulation scenarios. By replicating cases, ROS may have artificially increased estimation bias without reducing estimation variance in using Ridge regression, resulting in an overall worse discrimination.

Such results are similar whether the outcome had a 5% or 10% prevalence.

Figure 2.8 shows learning curves for Lasso regression. Based on the top-left sub-plot (Surrogate $Z1$ and 5% outcome prevalence), ROS slightly outperforms SRS, consistent with empirical observations in the machine-learning literature [138, 135]. However, using SGS substantially improves prediction compared to ROS. When comparing rows, sampling using surrogates $Z2$ or $Z3$ did not provide much benefit when the outcome had a 5% prevalence, while ROS provided some benefit. However, discrimination was overall low in these scenarios, indicating that there may have been insufficient information in the data to derive any meaningful comparisons across the sampling procedures.

In summary, these simulations verified our mathematical results from Sections 2.3.3-2.3.4 and demonstrate the broad potential benefit of SGS designs to develop machine-learning predictions of clinical outcomes. Our results reinforce the value of surrogate specificity and show the benefit of using design-based strategies over analysis based re-sampling methods for resource efficient classifier learning. We now transition to an application of the proposed SGS design on real-world text data.

Figure 2.7: Logistic Ridge Regression learning curves for binary features, comparing simple random sampling (SRS), random over-sampling (ROS), and surrogate-guided sampling with 1:1 (SGS 1:1) or 3:1 (SGS 3:1) ratio of surrogate positives to negatives.

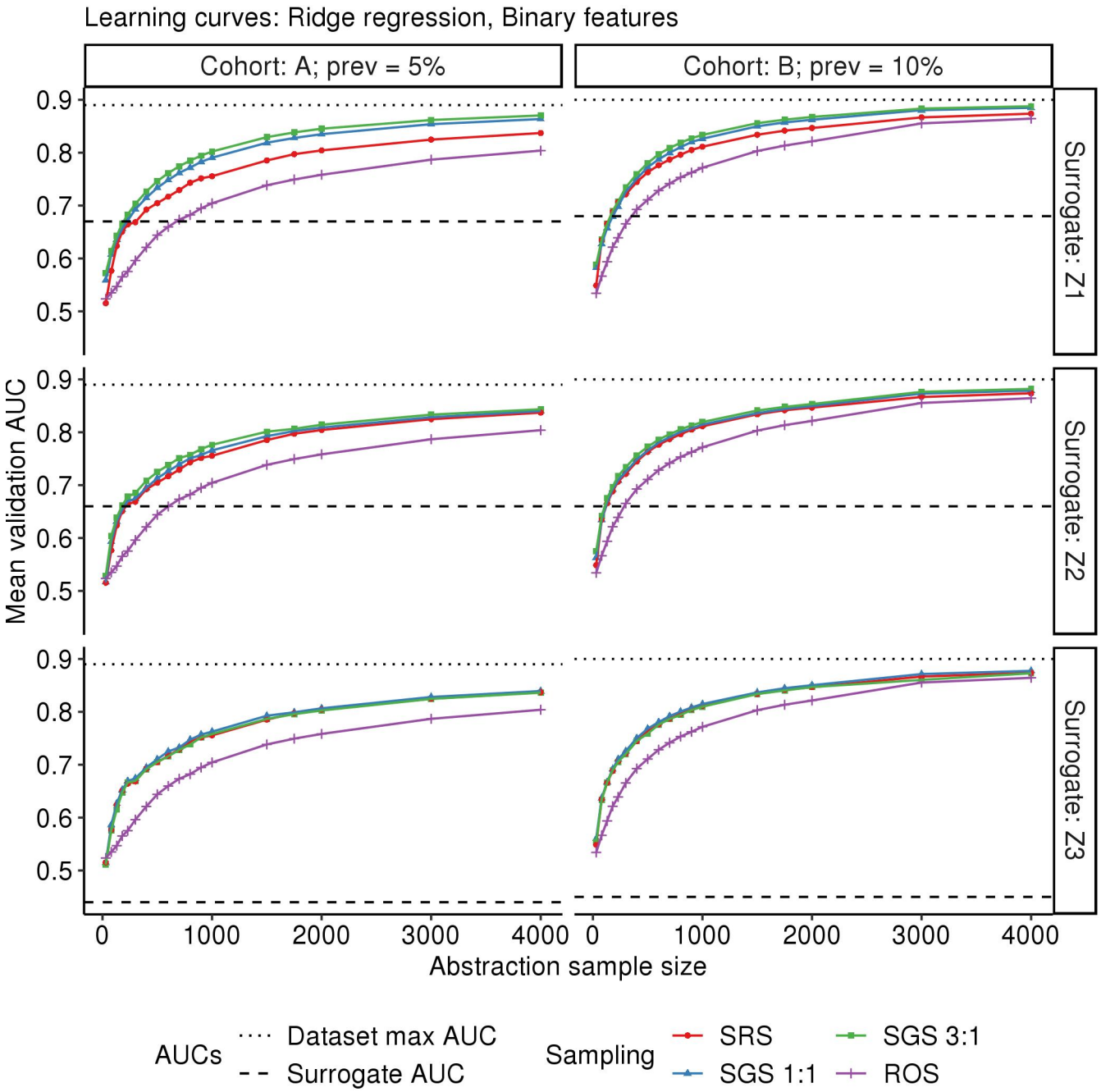
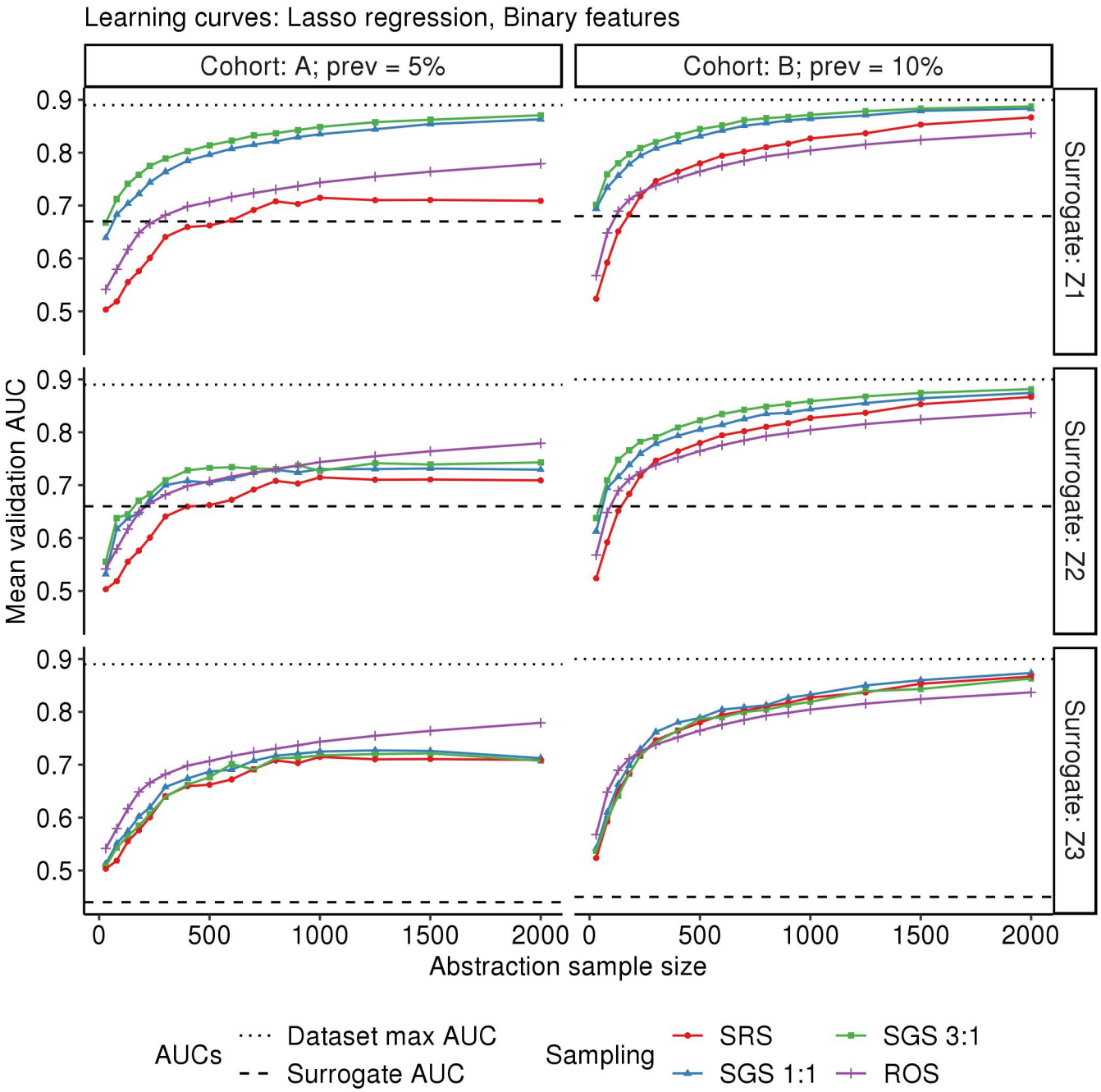


Figure 2.8: Logistic Lasso Regression learning curves for binary features, comparing simple random sampling (SRS), random over-sampling (ROS), and surrogate-guided sampling with 1:1 (SGS 1:1) or 3:1 (SGS 3:1) ratio of surrogate positives to negatives.



2.5 Illustration: Classification of vertebral fractures from radiology reports

2.5.1 Data set details

Vertebral fractures of the spine could 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 through radiographic imaging, such as plain x-ray or magnetic resonance imaging (MRI). The prevalence of vertebral fractures is estimated to be about 3-20% among primary care subjects seeking care for all reasons [134]. Accurate and scalable machine-learning identification of patient vertebral fracture statuses from EMR databases may capture potentially actionable clinical cases as well as facilitate research about prognosis and patterns of care.

The Lumbar Imaging with Reporting of Epidemiology (LIRE) study evaluated the effect of radiology report content on subsequent treatment decisions among adult subjects [65]. 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. Even though subjects from the LIRE study may be different than from the broader primary care population, the prevalence of vertebral fractures is still expected to be relatively rare. Definitive fracture status requires clinical expert abstraction of associated radiology text reports. Therefore, sampling strategies alternative to the usual SRS may be resource efficient for developing classification models for text detection algorithms in identifying vertebral fracture statuses. Using LIRE data as the "cohort", we evaluate the benefit of using SGS designs for outcome label abstraction and subsequent classification model development.

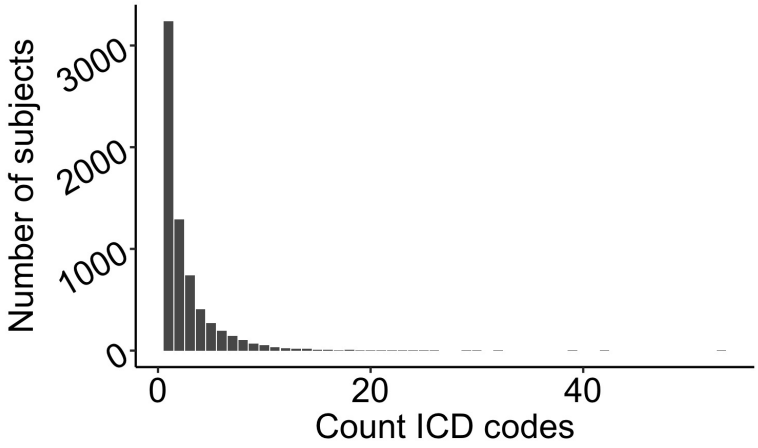
2.5.2 Sampling design: Surrogate creation and application of SGS

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 Appendix A.0.3. 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, 171592 (96%) did not have any relevant ICD codes, 3275 (1.83%) had one code, 1303 (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 in (2.18), where

$$Z_i = I(\text{count vertebral fracture ICD codes within 90 days for subject } i > 1). \quad (2.18)$$

Figure 2.9: Bar plot of the number of subjects for each count of relevant ICD codes for vertebral fracture. Counts are only shown for subjects with at least one ICD code noted within 90 days of report generation (96% did not).



This abstraction task was nested within a larger abstraction set-up in the LIRE study. The radiology reports of each selected subject were abstracted by two independent clinicians for the presence or absence of vertebral fractures. We selected 1000 reports that were abstracted for fracture, where 500 were based on SRS and 500 were based on SGS with equal values of $Z_i = 1$ and $Z_i = 0$; these abstracted reports form what we call the data marts. The SRS data mart was used for model validation, while the SGS data mart was used for

model development. Using the SRS data mart, we estimated marginal characteristics of the surrogate defined in (2.18), including its sensitivity, specificity, and AUC for true vertebral fracture status. We also estimated $LR+$, $LR-$, and O_{ratio} of the resulting SGS design which used a 1:1 sampling ratio. Sampling variability and corresponding confidence intervals were estimated by bootstrapping the SRS data mart.

2.5.3 Modeling and analysis

Features were created by processing radiology report text data using the `quanteda` package in `R`. A total of $p = 310$ features were created using bag-of-words (BOW), where the features included unigrams (i.e. single words) occurring in more than 2.5% of reports and less than 90% of all reports as well as bigrams and trigrams that represent certain negation patterns (e.g. "no fracture" and "no acute fracture"). Values in the feature matrix $\mathbf{X}$ were calculated according to the term-frequency inverse-document frequency (TF-IDF) representation, which incorporates information about the importance 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 $\mathbf{X}$ contains elements

$$\begin{aligned} X_{ij} &= TF(d_i, t_j) \times IDF(t_j) \\ TF(t_j, d_i) &= 1 + \log(1 + \frac{Count(t_j \in d_i)}{|d_i|}) \\ IDF(t_j) &= \log\left(\frac{N}{\sum_{i=1}^{N} I(t_j \in d_i)}\right). \end{aligned} \tag{2.19}$$

The representation in (2.19) is common in natural language processing (NLP) applications [106]. In addition to text-features, we also included the binary enrichment surrogate Z as a predictor. Our modeling approach was logistic regression with elastic net penalization [155],

equally penalizing the $L1$ and $L2$ norms, but imposing a zero penalty on Z:

$$\hat{\beta}_0, \hat{\beta}_Z, \hat{\beta} = \min_{\beta_0, \beta_Z, \beta} \{ -\sum_{i=1}^{n} Y_i(\beta_0 + \beta_Z Z_i + \sum_{j=1}^{p} \beta_j X_{ij}) + \log(1 + \exp(\beta_0 + \beta_Z Z_i + \sum_{j=1}^{p} \beta_j X_{ij})) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \}, \tag{2.20}$$

To investigate the design effect on model prediction accuracy, we drew $B = 100$ bootstrap samples of sizes $n = 100, 250, 500$ from the SGS data mart. To simulate the SRS design, we drew samples according to an "inverse SGS" design from the SGS data mart, where surrogate positives were instead under-included with sampling weights (2.13). To simulate the SGS design, we drew samples randomly from the SGS data mart. For each simulated sample, we fitted (2.20) where the regularization parameter λ was selected 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 (SRS data mart) to obtain estimates of the validation AUC. For each sampling design (SRS and SGS) and for each sample size, we reported bootstrap mean validation AUC and 95% bootstrap confidence intervals.

2.5.4 Data analysis results

Estimated data set characteristics are shown in Table 2.3. Note that the defined surrogate (2.18) was by itself not very discriminative for the outcome based on its AUC. 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 defined in (2.18) could be a good surrogate for sampling. In fact, while the naturally occurring outcome prevalence of vertebral fracture in the LIRE cohort was estimated to be 12%, the SGS sample outcome prevalence was estimated to be 45%.

Data analysis results are shown in Table 2.4. In general, in fitting an elastic-net logistic 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 0.93 while that of SRS was only 0.80, a difference of 0.13, suggesting that allocating a sample size of 250 is more resource efficient under SGS compared to SRS.

Table 2.3: 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_{ratio} of resulting surrogate-guided sampling (SGS) design.

Surrogate or design metric	Estimate (95% C.I.)
Z_{sens}	0.29 (0.17, 0.41)
Z_{spec}	0.99 (0.98, 0.99)
$AUC(Z)$	0.64 (0.58, 0.70)
$LR+$	29 (8.5, 41)
$LR-$	1.4 (1.2, 1.7)
$O_{ratio}(Z, R = 0.5)$	6.5 (3.5, 8.9)

Table 2.4: Average validation AUC (95% C.I.) for various training sample sizes, based on B=100 bootstrap resamples, for illustration of surrogate-guided sampling (SGS) designs on radiology reports drawn from the LIRE data set.

Training sample size	$\hat{AUC}(\mathbf{D}^{SRS}(n))$	$\hat{AUC}(\mathbf{D}^{SGS}(n))$
100	0.76 (0.50, 0.95)	0.87 (0.67, 0.94)
250	0.80 (0.50, 0.97)	0.93 (0.84, 0.98)
500	0.91 (0.50, 0.98)	0.94 (0.86, 0.98)

2.6 Discussion

Towards establishing research-quality clinical outcomes from massive EMR databases, machine-learning may provide a potential accurate and scalable alternative to traditional large-scale manual review. To that goal, we proposed surrogate-guided sampling (SGS) design, a formal statistical sampling framework to guide sample selection for such machine-learning classification of rare outcomes. SGS designs involve stratified sampling on values of enrichment surrogates, which are structured data elements related to the true clinical outcomes of interest. We demonstrated that using SGS designs may reduce sample size requirements for accurate classification. In addition, we characterized design impact on modeling, as well as provided analytic guidance for valid model development and unbiased estimation of prediction accuracies. Here, we remark that the SGS design may be viewed as a special case of the general two-phase sampling design [89].

In determining what constitutes a "good" surrogate for sampling, we have demonstrated both analytically and in simulations that sample case-enrichment and subsequent classification accuracy is most affected by surrogate specificity. A practical recommendation may be based on Figure 2.3, where a specificity of 0.95 or higher is ideal, 0.80 is very good, and 0.50 is the absolute minimum specificity a surrogate needs to achieve. To create highly specific surrogates, regular expression keyword searches may be supplemented with off-the-shelf negation tools [55], while related ICD codes may be defined with higher count thresholds within a shorter period of time. Additionally, keywords and ICD codes may 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 [3, 2].

Even though increasing surrogate specificity is expected to increase sample outcome prevalence and subsequently classification accuracy, in practice the trade-off with case representa-

tiveness may require consideration. A concern is if the true cases obtained through stratified random sampling on a highly specific surrogate are sufficiently representative of all possible subtypes. For example, in the vertebral fracture data application, requiring at least two ICD codes (instead of one) may increase surrogate specificity, but could have resulted in a sample with mostly chronic fractures. This may be a problem if acute fractures are part of the true clinical outcome definition to be learned by the machine-learning algorithm. A possible solution may implement a "tiered" surrogate, using sub-samples defined by variables to balance specificities and case representativeness (e.g. > 2, 1, 0 counts of ICD codes).

Another trade-off relates to statistical versus scientific validity for model development. Towards statistically "valid" model development, the SGS framework requires that the surrogate be included as a predictor. However, if the surrogate is defined using variables considered "downstream" of the true clinical outcome, then including the surrogate as a predictor may be scientifically invalid. One simple solution to this dilemma is to restrict surrogate creation to be based only on variables that temporarily occur before, or at the same time, of the true clinical outcome. If that is not possible, we recommend favoring scientific validity where the surrogate may be excluded from the set of predictors: relevant statistical adjustments using this approach warrant further investigation.

Anchored on the SGS design framework, future work invites a variety of methodological and practical questions related to full study planning, for example questions such as "what types of sampling variables", "how to allocate sample", as well as "how many to abstract". To address sampling variable types, we had discussed a trade-off of surrogate specificity and case representativeness; other considerations may include statistical (e.g. data completeness and distributions), scientific (e.g. data coverage, site heterogeneity), and technical (e.g. ease of querying). In terms of sample allocation, increasing sampling ratio only slightly improved classification accuracy, yet resulting inflated inverse weights may increase the variance of IPW estimators of validation accuracy measures - it may be interesting to investigate this

trade-off further. Once relevant trade-offs are formally defined, appropriate sample size calculations may then proceed taking into account the need of both model development and model validation. Other future work may include investigating the appropriateness of the SGS framework outcomes much rarer than 5%, design effects on prediction accuracy measures other than AUC, as well as best practices for sampling in the presence of site heterogeneity.

Finally, the contribution of our work is the formalization of a design framework for clinical outcome abstraction and labeled data collection, towards accurate and scalable machine-learning of research-quality clinical outcomes. We showed that many existing "ad-hoc" sample selection strategies using structured EMR data elements can be formalized as a stratified sampling design, with statistical guarantees of improved classification accuracy as well as generalizability. Ultimately, our hope is to encourage more careful statistical and study design thinking when assembling labeled data sets for machine-learning model development and validation, especially considering the non-trivial abstraction cost in obtaining such labels.

Chapter 3

MULTI-LABEL SURROGATE-GUIDED SAMPLING DESIGNS FOR MULTI-LABEL CLASSIFICATION

3.1 Introduction

Radiology reports constitute the formal documentation and communication of imaging study results by trained radiologists, and often describe the presence of multiple potentially co-occurring radiographic findings. Radiographic findings are often captured in free text, yet are important to be identified for research purposes, such as cohort definitions and subgroup analyses of scientific studies. Our scientific motivation comes from the Lumbar Imaging with Reporting of Epidemiology (LIRE) [65] pragmatic clinical trial, which captured over a quarter million of lumbar spine imaging reports. From the LIRE database, we were interested in identifying subjects with one or more of findings considered "red flag" on their radiology reports. Towards such accurate and scalable clinical outcome identification, a promising option may be based on machine-learning classification models.

In order to develop machine-learning classification models for the "red flag" findings identification task, we require the abstraction of an adequate sample of reports, which involves human medical expert coding of findings to check boxes based on interpretation of report text. In general, abstracted data can be notoriously expensive and time consuming to obtain, but is even more challenging for "red flag" findings, as there are multiple outcomes of interest requiring abstraction, many of which are rare. Such scarcity of abstracted data have constrained the advancement of machine-learning algorithms for many biomedical outcome classification tasks. For example, for the identification of critical findings on radiology reports, current state-of-the-art methods are predominantly deterministic rule-based algo-

rithms based on pre-specified search terms [72], in part due to the difficulty in obtaining large samples required for machine-learning model development.

An alternative to simply collecting large samples is to use targeted sampling designs. In particular, sampling designs targeting the rare outcome scenario have been proposed in the fields of epidemiology and machine-learning [138, 7, 101]. When true outcomes are not yet available and require abstraction, we have previously described a class of sampling designs based on stratified sampling on enrichment surrogates, which are summaries of structured data elements such as simple keyword searches and related International Classification of Disease (ICD) codes. In this work, we extend the previously proposed surrogate-guided sampling (SGS) design framework [125] to the multi-variate binary outcome setting that is typical for clinical text abstraction.

Findings on radiology reports naturally motivate a multi-label framework, which is a type of multi-variate binary outcome. We provide background for multi-label classification and multi-label datasets in Section 3.2.1, and review relevant sampling designs for multi-variate binary data from machine-learning and epidemiology in Section 3.2.2. We then describe the proposed multi-label surrogate-guided sampling design (Section 3.3.1– 3.3.2), design implementation (Section 3.3.3), as well as design impact on model development and model validation (Section 3.3.4). We provide empirical simulations of design benefit for machine-learning in Section 3.4. In Section 3.5 we illustrate the application of the design on a dataset of lumbar spine imaging reports. We provide a concluding discussion in Section 3.6.

3.2 Background

3.2.1 Multiple rare findings from radiology reports and multi-label datasets

The classification of radiographic findings from radiology report text motivates a multi-label dataset (MLD) framework, where the outcome of interest is a vector of mutually

non-exclusive binary labels, such as "vertebral fracture" and "spinal malignancy". The terminology "multi-label dataset" was first described in the context of image categorization [13], and can be viewed as a generalization from the binary outcome setting.

Multi-label classification models and classifier evaluation measures

Multi-label classification describes the class of machine-learning algorithms developed for multi-label datasets, where modeling approaches are generally based on either algorithm adaption or problem transformation [129]. Algorithm adaptation approaches re-weight the loss functions of commonly used binary classification models, for example decision trees and K-th nearest neighbors [28, 151], where weights are heuristically justified to accommodate multi-label structure. Alternatively, problem transformation involves converting the outcome vector into either individual binary or multinomial outcomes, and then applying existing algorithms to such transformed datasets.

A simple approach for multi-label classification is called Binary Representation [13], where each element of the outcome vector is modeled individually based on the same feature set. To account for potential dependencies among the outcomes, individual predictions may be cascaded along a chain as additional features to model subsequent outcomes [104]. Alternatively, multinomial classification methods can be applied to multi-label datasets, where ensemble models on randomly select subsets of outcome cross-classifications [130] may reduce computational intensity.

To evaluate multi-label classification algorithms, predictions and true classes of individual outcomes may be summarized, where commonly used metrics include sensitivity, specificity, and Area Under the Receiver Operating Characteristic Curve (AUC). Resulting vectors of evaluation metrics may also be aggregated to obtain a single summary measure, for example macro-averages or micro-averages [152]. Viewing accuracy metrics as a weighted sum of loss functions over all outcomes and all subjects, then macro-averages can be interpreted

as assigning equal weights on all outcomes while micro-averages summarize at the subject level. To compute macro-averaged measures, accuracy metrics for each outcome label is computed individually and then averaged. Obtaining micro-averaged measures is slightly more nuanced and depends on the exact accuracy measure used. Specifically for the AUC measure, while the macro-AUC is based on ROC curves of individual outcomes and can be theoretically justified [52], the micro-AUC is a more heuristic measure based on a single ROC curve with true positive and false positive rates averaged over all outcomes. For that reason, the macro-AUC may be preferred as a summary evaluation measure for multi-label classification.

Summaries of outcome "rareness" in multi-label datasets

To characterize the outcome "rareness" of multi-label data, several metrics for label sparsity and label imbalance, respectively describing the overall and relative rareness of outcome labels in a sample, have been proposed [129, 152]. For example, the measure "Density" is the average sample outcome prevalence, which is a measure of absolute outcome rareness (label sparsity). Additionally, the measure "mean imbalance ratio" (meanIR) quantifies relative outcome rareness (label imbalance), and is calculated as the average sample prevalence ratio comparing each outcome to the most common outcome. Other metrics of label sparsity and label imbalance for multi-label datasets are summarized in Table 3.1. For binary classification, the negative effect of rare outcomes on classification performance has been well documented [139, 7]. Similarly, there has been some work describing the effect of label sparsity and label imbalance on multi-label classification [22].

3.2.2 Sampling methods for rare multi-label data

When a multi-variate binary outcome is expected to be naturally rare, targeted sampling methods may provide alternatives to large random samples towards modeling goals; we review relevant sampling methods from the complementary fields of machine-learning and epidemiology.

Table 3.1: Summary of commonly used metrics for multi-label dataset outcome "rareness" in terms of label sparsity and label imbalance for outcome vector $\tilde{Y} \in \{0,1\}^K$ in a sample of size n.

Metric name	Formula	Measure type
Cardinality	$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}$	Label sparsity
Density	$\frac{1}{K}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}$	Label sparsity
IRLbl($\tilde{Y}_k$)	$\frac{\max_k \sum_{i=1}^{n} Y_{ik}}{\sum_{i=1}^{n} Y_{ik}}$	Label imbalance
meanIR	$\frac{1}{K}\sum_{k=1}^{K} \text{IRLbl}(\tilde{Y}_k)$	Label imbalance
IRLblσ	$\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\text{IRLbl}(\tilde{Y}_k) - \text{meanIR})^2}$	Label imbalance
CVIR	$\frac{\text{IRLbl}\sigma}{\text{meanIR}}$	Label imbalance

Sampling methods from machine-learning

In machine-learning, sampling methods for rare outcomes generally involve re-sampling an existing dataset, and are based on over-sampling (replicating observations) or under-sampling (deleting observations). The goal of re-sampling training samples is to increase sample case proportions, which is believed to improve classifier learning. Such re-sampling can be generalized to the multi-label setting, where observations (rows) are replicated or deleted based on the multi-labeled outcomes (columns), following either an Individual Label or Label Powerset approach.

The Individual Label Random Over-Sampling (called ML-ROS by the authors, where ML refers to "multi-label") [21] augments the original training sample with replicates of subjects selected from cases (outcome = 1) individually for each outcome. In ML-ROS, replicates are iteratively drawn from cases for each individual outcome, preferentially over-representing outcomes with low marginal sample prevalences. On the other hand, the Label Powerset Ran-

dom Over-Sampling (LP-ROS) [21] selects replicates based on multi-class cross-classification instead of the marginal outcome distributions.

Random under-sampling approaches for multi-label datasets can also be based on individual label (IL-RUS) or label powerset (LP-RUS) approaches, where controls (outcome = 0) are eliminated based on either marginal or multi-class outcome values. Another considered under-sampling approach is called "inverse random under-sampling" (I-RUS) [121], which is based on weighted bootstrap aggregation (bagging). In I-RUS, for each outcome separately, large proportions of controls are deleted until individual outcome prevalences exceed 50%, and then bagged classifiers are learned from the weighted re-samples. While the approach in [121] was empirically demonstrated to simultaneously minimize both false positive and false negative rates for classification, the use of different training samples for classifying each outcome may induce additional variation and thus compromise generalizability.

In comparing the various re-sampling approaches for machine-learning multi-label classification, Individual Label re-sampling approaches were empirically shown to outperform Label Powerset approaches [21]. The heuristic rationale, described in [22], was due to that Individual Label re-sampling directly targets improving the multi-label sample outcome "rareness" measures (shown in Table 3.1), and that such metrics represent the "information" required for multi-label classification. Note that while the reviewed multi-label re-sampling methods may improve classifier learning, it is unclear how to account for artificially introduced biases. For example, due to subject replication or deletion, empirical accuracy estimators for model validation require correction.

Sampling designs from epidemiology

Compared to the machine-learning re-sampling methods, targeted sampling for rare outcomes from epidemiology are generally framed as study designs. The goal of study designs for rare outcomes is to collect sufficient cases for statistical inference when data collection

resources are scarce. Even though the exact multi-label data formulation as in machine-learning is not common in epidemiology studies, we may consider sampling designs used for the related survival and longitudinal data types.

For survival data, consider a discrete time parameterization, where the time-to-event variable is partitioned into intervals and individual "outcomes" are defined by event indicators within each time interval. For rare diseases, the case-cohort [100] is a resource efficient sampling design for exposure estimation involving survival-motivated multiple outcomes. At baseline, a random sub-cohort is selected from all available subjects, and acts as a common control arm that is representative of the source cohort. Then, at each individual time interval, case sub-samples are drawn from subjects who are true cases (outcome = 1). The resulting case-cohort sample is the set consisting the random sub-sample and all case sub-samples. Under the discrete time survival data formulation, for each subject the resulting outcome vector can be viewed as a highly-structured multi-label outcome vector, where subjects who experienced the event during an interval do not experience subsequent events. Therefore, perhaps more similar to machine-learning multi-label data is the longitudinal binary data setting, where for each subject every element of the outcome vector make take values of case (outcome = 1) or control (outcome = 0) at individual time points. For repeated binary measures, the outcome dependent sampling [110] is a resource-efficient data collection strategy that preferentially over-samples subjects with variation in the outcome vector, and not those who are always cases or always controls.

Much of the work related to epidemiology sampling designs have focused on the valid analysis of resulting samples. For example with the case-cohort design, an issue with analysis is the double counting of subjects, which may happen when the same subject is selected both into the sub-cohort as well as into a case sub-sample. One strategy to derive appropriate statistical estimators for data arising from such epidemiology sampling designs is to frame them as two-phase sampling designs [89]. Two-phase sampling was originally described as a survey

sampling strategy to estimate population means and proportions of an expensive outcome, for example in censuses, by stratified sampling on cheaper auxiliary variables. Designs that follow the two-phase assumption may be analyzed with methods that involve *imputation* of missing variables or weighting of existing data. For example, estimators based on Inverse Probability Weighting (IPW) may be a reasonable analytic choice [63].

Therefore, contrary to machine-learning re-sampling approaches, valid and efficient analytic approaches for epidemiology sampling designs such as the case-cohort have been well documented. However, the effects of using case-cohort designs for multi-label classifier development have not been explored. Intuitively, since case-cohort designs intentionally over-include cases, it is likely that resulting samples may help classifier learning. Such an intuition motivates the idea that, in addition to re-sampling methods at the *analysis* stage, design-based sampling methods to select observations for outcome label collections may also improve classifier learning.

3.2.3 Surrogates in Electronic Medical Records (EMR) databases

Note that over-sampling, under-sampling, and case-cohort sampling approaches all assume that the true outcome statuses are already known, and may be used as sampling variables. Unfortunately, this assumption is not true for radiology reports derived from large Electronic Medical Records (EMR) databases. For multi-label outcome abstraction and subsequent machine-learning from radiology report text, true outcome statuses are generally unobserved *before* abstraction. On the other hand, unstructured text reports are usually associated with structured data elements belonging to the same subject in EMR databases, for example simple keyword searches on report text, as well as International Classification of Disease (ICD) and Current Procedure Terminology (CPT) codes, all of which tend to be more accessible for querying. Sampling based on summaries of such structured data elements, also called "surrogates", have recently been described in the univariate binary outcome scenario based on a class of designs called surrogate-guided sampling (SGS) [125].

The categorization of radiology report text with multiple labels and involving ICD codes has been investigated as a shared task described in [98]. The key difference compared to our work is that, while individual ICD codes were used as multi-label outcomes in [98], our work utilizes summaries of ICD codes as sampling variables. Interestingly, the multi-label classification shared task [98] found that while most of the top-performing Natural Language Processing (NLP) systems utilized machine-learning algorithms, some of the best NLP algorithms were purely rule-based. In [98], a possible explanation relates to the available sample size: with only $n = 978$ documents for training, yet requiring classification of 45 outcome labels having 94 distinct combinations, the training sample of the shared task is an example of a multi-label dataset with potential label sparsity and label imbalance. While it is possible that machine-learning re-sampling methods [21] may improve classifier prediction accuracy, they do not address the cost of large-scale abstracted label data collection, which was noted to be a challenge in [98]. Therefore, especially for multi-label classification tasks involving EMR data, targeted sampling at the *design* (data collection) stage may substantially reduce such so called "abstraction burden" towards obtaining high information yet valid samples for model development and validation.

3.3 Methods

3.3.1 Statistical notation and key assumptions

For subject i, denote $\tilde{X}_i \in \mathcal{R}^p$ as the feature vector, $\tilde{Y}_i \in \{0,1\}^K$ as the multi-label outcome label vector, and $\tilde{Z}_i \in \{0,1\}^K$ as the vector of surrogates related to $\tilde{Y}_i$. From the large EMR cohort $\mathcal{D}$, a sample $\mathbf{D}^S(n)$ of size n is drawn based on sampling mechanism S. The multi-label classification model is a function $H : \mathcal{R}^p \to \{0,1\}^K$ or $H : \mathcal{R}^p \to [0,1]^K$, where $\hat{H}$ is estimated using from $\mathbf{D}^S(n)$. For simplicity, we adopt the Binary Representation framework for multi-label classification, and represent H as the set of K classifiers $\{h_1, \ldots, h_K\}$. We assume the following on the overall joint distribution of the data in $\mathcal{D}$:

- **Rare outcomes:** Some of the outcome labels are rare (i.e. low "label density"), and potentially with varying prevalences (i.e. high "label imbalance").

- **Case-enriching surrogates:** Surrogate Z_k is case-enriching for outcome label Y_k. Mathematically, for the surrogate-outcome pair $(Z_k, Y_{k'})$ where $k = k'$:

$$\begin{aligned} \text{Sensitivity: } & P(Z_k = 1 | Y_{k'} = 1) > 0 \\ \text{Specificity: } & P(Z_k = 0 | Y_{k'} = 0) \approx 1, . \end{aligned} \tag{3.1}$$

  In (3.1), surrogate Z_k has non-zero sensitivity and very high specificity for outcome $Y_{k'}$ when $k = k'$, but has unrestricted sensitivity and specificity for outcome $Y_{k'}$ when $k \neq k'$. Such high specificity translates into high positive predictive value (PPV), which we define as "case-enriching" for sampling.

For multi-labeled datasets with **rare outcomes** and **case-enriching surrogates**, we can leverage such assumptions for sampling designs targeted towards multi-label case-enrichment. Specifically, sampling can be based on the surrogate vector, using a class of designs that we define as multi-label surrogate sampling.

3.3.2 Multi-label surrogate-guided sampling (mlSGS) design

We consider the statistical problem of drawing a sample $\mathbf{D}^S(n)$ from a large cohort $\mathcal{D}$, where $\mathbf{D}^S(n)$ is to be used for multi-label outcome abstraction and subsequent machine-learning model development. When sampling S is based only on functions of a pre-specified surrogate vector $\tilde{Z}$, we call the resulting class of designs multi-label surrogate sampling. The goal of using designs within the multi-label surrogate sampling class is primarily for multi-label outcome case-enrichment. Then, an example of sampling is based on full stratification on all $2^K - 1$ distinct values of $\tilde{Z}$. Due to the potentially large number of strata, full stratification is untargeted and therefore may not be ideal for case-enrichment.

Definition 1 *Multi-label surrogate sampling class.*
For designs within the multi-label surrogate sampling class, sampling is based only on $\tilde{Z} \in \{0,1\}^K$.

Alternatively, recall that due to the case-enriching surrogate specificity assumption (3.1), subjects who are "surrogate positives" are highly likely to be cases. Therefore, we propose using a design that we call **multi-label surrogate-guided sampling (mlSGS)** for multi-label outcome abstraction and subsequent machine-learning. The mlSGS design consists of a set of sub-samples, drawn based on surrogate positives and simple random sampling (SRS). By sampling based on surrogate positives, mlSGS provides case-enrichment individually for each outcome. Through including a sub-sample based on SRS, mlSGS provides a comparison group with distributions similar to that of the cohort. We remark on this connection between the mlSGS design and the case-cohort design [100]: if the surrogate vector $\tilde{Z}$ is exactly the outcome vector $\tilde{Y}$, then the mlSGS design is exactly the case-cohort design where predictors are already known, and that outcome case/control statuses are available for all subjects.

Definition 2 *Multi-label surrogate-guided sampling (mlSGS) design.*
For sampling from cohort $\mathcal{D}$ *based on surrogate vector* $\tilde{Z}$*, the mlSGS sampling design is the set of sub-samples* $\{\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}_{SRS}\}$*, where sub-sample* $\mathcal{S}_k$ *is drawn from surrogate-positive units* ($Z_k = 1$)*, and* $\mathcal{S}_{SRS}$ *is based on simple random sampling.*

Figure 3.1 illustrates the mlSGS design and associated expected case-enrichment, where rows indicatate sub-samples based on either surrogate positives ($Z_1 = 1$, $Z_2 = 1$) or SRS, and columns indicate the expected case (shaded) and control (blank) proportions of individual outcomes. Figure 3.1a illustrates case-enrichment of two outcomes, each having marginal prevalences of 10%, and sampling based on associated case-enriching surrogate. Under mlSGS, sub-samples based on surrogate positives had about *four times* higher case proportions due to surrogate case-enrichment, resulting in overall higher sample prevalences compared to using SRS alone. Figure 3.1b illustrates a scenario where sampling is based

only on two surrogates but three outcomes were abstracted for true case/control statuses, since the more prevalent outcome $Y3$ does not require additional case-enrichment. Therefore, for a set of related findings with varying prevalences, assuming that the cost of abstracting additional outcomes from the same set of documents is minimal, sampling can be based on surrogates targeting case-enrichment of the rarer outcomes but all outcomes are simultaneously abstracted. We illustrate this scenario in the data application example (Section 3.5).

We identified four key characteristics of the data distribution and the sampling design that may affect overall case-enrichment when using mlSGS:

- **Surrogate operating characteristics:** Using surrogates with higher specificity leads to higher case-enrichment; analytical and empirical results were previously demonstrated in the univariate scenario [125].

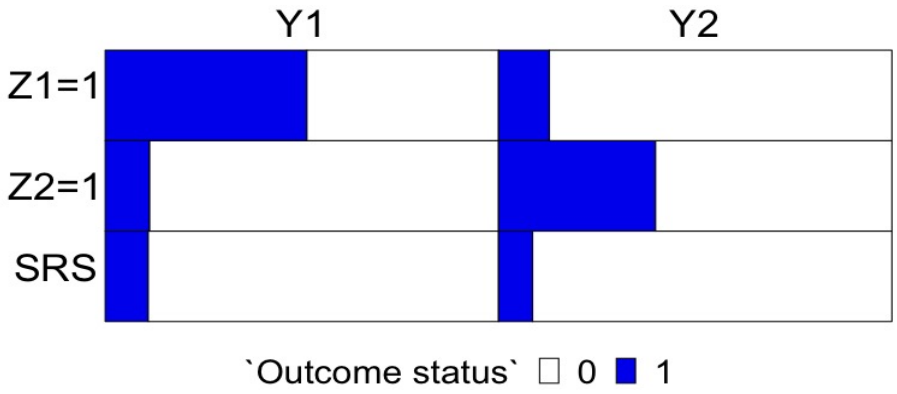
- **Sub-sample weights:** Assigning higher weights to surrogate positive sub-samples (i.e. over-including surrogate positives) leads to higher case-enrichment; we provide empirical results in Section 3.4.2.

- **Number of sub-samples:** Increasing the number of sub-samples leads to lower overall case-enrichment; we provide empirical results in Section 3.4.3. As the number of sub-samples increase, the expected sub-sample outcome prevalences remained the same, but overall sample prevalences are expected to be lower due to pooling across multiple sub-samples.

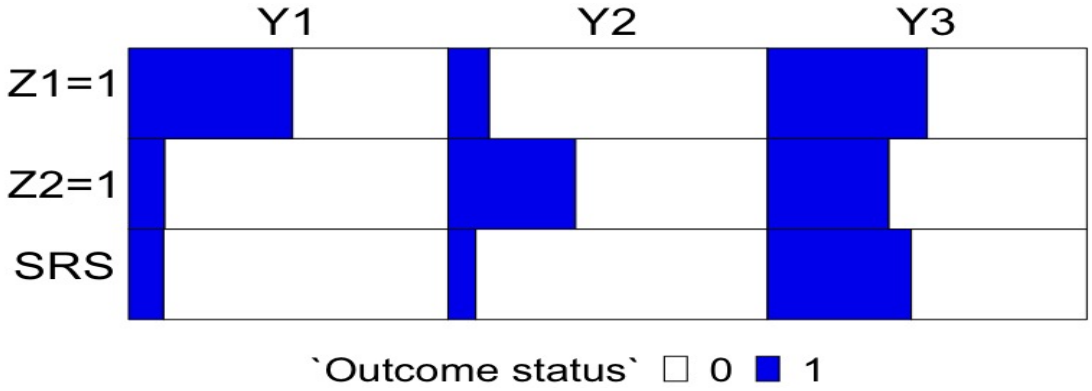
- **Outcome correlations:** Using mlSGS designs on datasets with positively correlated outcomes leads to higher case-enrichment compared to another equivalent dataset but with no correlation; we provide empirical results in Section 3.4.3.

Figure 3.1: Expected sample case-enrichment under the multi-label surrogate-guided sampling (mlSGS) design based on 2 case-enriching surrogates (individual sensitivities of 40% and specificities of 95%). Outcomes $Y1$ and $Y2$ each had marginal prevalences of 10%, while outcome $Y3$ had marginal prevalence of 50%.

(a) Outcomes Y_2, Y_2 case-enrichment under a mlSGS design with K=2 surrogates.



(b) Outcomes Y_2, Y_2, $Y3$ case-enrichment under a mlSGS design with K=2 surrogates.



3.3.3 Design Implementation

To implement the mlSGS design (Definition 2), a concern may be due to the potential of replicated subjects, for example the same subject may be selected for strata $\mathcal{S}_k$ and $\mathcal{S}_{k'}$ on the basis of being a surrogate positive for outcomes k and k'. Towards the collection of a distinct sample, we propose the sequential sampling implementation (Algorithm 1). Under sequential sampling, subjects are first drawn randomly from the cohort into the SRS sub-sample. Then, remaining sub-samples are selected from subjects who are surrogate positives (subject i such that $Z_{ki} = 1$), but excluding those who have already been selected. Finally, the resulting mlSGS sample is the set of sub-samples, which consists of unique subjects.

Algorithm 1 *Sequential sampling implementation for multi-label surrogate-guided sampling.*

Input: *Cohort $\mathcal{D}$, Sub-sample sizes $n_1, \ldots, n_K, n_{SRS}$.*

Do:

Select n_{SRS} into $\mathcal{S}_{SRS}$ from: $\mathcal{D}$; calculate $\pi_{SRS,i} = P(i \in \mathcal{S}_{SRS})$.

Select n_1 into $\mathcal{S}_1$ from: $(Z_1 = 1) \in \mathcal{D} \backslash \{\mathcal{S}_{SRS}\}$; calculate $\pi_{1,i} = P(i \in \mathcal{S}_1)$.

Select n_2 into $\mathcal{S}_2$ from: $(Z_2 = 1) \in \mathcal{D} \backslash \{\mathcal{S}_1 \cup \mathcal{S}_{SRS}\}$; calculate $\pi_{2,i} = P(i \in \mathcal{S}_2)$.

$\vdots$

Select n_K into $\mathcal{S}_K$ from: $(Z_K = 1) \in \mathcal{D} \backslash \{\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_K \cup \mathcal{S}_{SRS}\}$; calculate $\pi_{K,i} = P(i \in \mathcal{S}_K)$.

Return: *mlSGS sample $\mathbf{D}^{mlSGS}(n) = \{\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_K \cup \mathcal{S}_{SRS}\}$, Sub-sample inclusion probabilities $\{\pi_{1,i}, \ldots, \pi_{K,i}, \pi_{SRS,i}\}$.*

In Algorithm 1, sub-sample inclusion probabilities $\pi_{1,i}, \ldots, \pi_{K,i}, \pi_{SRS,i}$ may be returned together with the sample $\mathbf{D}^{mlSGS}(n)$ during the design stage, and used to calculate overall inclusion probabilities (analytical forms in Section 3.3.4).

3.3.4 Design impact on model development and model validation

The multi-label surrogate-guided sampling (mlSGS) results in an intentionally case-enriched sample to improve classification accuracy. To characterize design impact, we frame sampling as a missing data problem, similar to the formalization in [147]. Briefly, design impact on machine-learning can be characterized by distributional differences between sample and cohort, where sampling schemes following the Missing At Random (MAR) assumption [78] may still provide unbiased learning and validation.

Design impact on model development

To characterize design impact on model development, we show the asymptotic probabilistic equivalence comparing outcome model conditioned on the mlSGS design to the unconditioned model (Lemma (1)). Intuitively, since sampling into each sub-sample is based only on surrogate positives, the mlSGS sampling distribution is a function of the surrogate vector $\tilde{Z}$ which is observed for the entire cohort - precisely an MAR assumption. The immediately corollary is that any sampling function for assembling sub-samples results in probabilistically equivalent outcome models. Therefore, using the sequential implementation (Algorithm 1), even though excluding already selected subjects at every iteration, still results in valid samples for model development, as any exclusions only depend on observed data.

Lemma 1 *Statistical validity of mlSGS for model development.*
For the mlSGS design with sub-samples $\{\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}_{SRS}\}$*, let* $\tilde{S} = (S_1 \ldots S_K, S_{SRS})$ *indicate sampling into each sub-sample, and* $\tilde{Z} \in \{0,1\}^K$ *as the surrogate vector that sampling is based on. Then,*

$$\tilde{S} \perp \tilde{Y} | \tilde{Z} \tag{3.2}$$

and for any function $g(\tilde{S})$, on supp$(\tilde{Z})$,

$$f(\tilde{Y}|\tilde{Z}, g(\tilde{S})) = f(\tilde{Y}|\tilde{Z}). \tag{3.3}$$

Corollary 1 *The mlSGS design implemented with the sequential sampling results in valid samples for model development, provided estimation is conditioned on the surrogate vector $\tilde{Z}$.*

In fact, Lemma (1) implies that not only the mlSGS design, but all designs within the multi-label surrogate sampling class (Definition (1)) are conditionally valid for model development, as sampling only depends on the observed surrogate vector $\tilde{Z}$. However, modeling requires the inclusion of $\tilde{Z}$ for validity. Additionally, design validity for model development only exists on the sample support of $\tilde{Z}$. This assumption may be violated in the following example: for $K = 2$ and sampling based only on positive surrogates ($Z_1 = 1$ or $Z_2 = 1$), subjects with $(Z_1, Z_2) = (0, 0)$ are never selected, therefore compromising model generalizability to healthy controls without further assumptions. To prevent the scenario where surrogate negatives are never included, our proposed mlSGS design (Definition (2)) intentionally includes a sub-sample based on SRS.

Design impact on model validation

Often in practical scenarios, it may be easier to collect abstracted outcome labels using a single design, and then perform a "split sample" for model development and validation. Even though mlSGS designs are valid for development (Lemma 1), resulting samples have outcome distributions that are intentionally different from the cohort. Therefore, model validation is based on a sample collected with mlSGS, resulting empirical estimates of model prediction accuracy measures, such as sensitivity, specificity, and AUC, will be biased unless corrected.

Fortunately, as the sampling model is known by design, bias correction can be based on the Horvitz-Thompson estimator [63], also known as Inverse Probability Weighting (IPW). The IPW correction weights apparent empirical estimates using subject sampling weights. For subject i having true outcome values Y_{ki}, predictions $\hat{p}_{ki}$, and sampling weights π_i, the IPW-corrected estimates for the True Positive Rate (TPR) and False Positive Rate (FPR) for cut-off c are [3]

$$TPR_{IPW}(Y_k; c) = \frac{\sum_{i=1}^{n} I(\hat{p}_{ki} \geq c) Y_{ki} \pi_i^{-1}}{\sum_{i=1}^{n} Y_{ki} \pi_i^{-1}}$$
$$FPR_{IPW}(Y_k; c) = \frac{\sum_{i=1}^{n} I(\hat{p}_{ki} \geq c)(1 - Y_{ki}) \pi_i^{-1}}{\sum_{i=1}^{n} (1 - Y_{ki}) \pi_i^{-1}},$$

and a closed form of the IPW-corrected AUC has been shown as [60]:

$$AUC_{IPW}(Y_k) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{-1} \pi_j^{-1} I(\hat{p}_{ki} > \hat{p}_{kj}) I(Y_{ki} > Y_{kj})}{\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{-1} \pi_j^{-1} I(Y_{ki} > Y_{kj})}. \tag{3.4}$$

For multi-labeled outcomes, we further define the IPW-corrected macro-AUC as

$$\text{Macro-AUC}_{IPW} = \frac{1}{K} \sum_{k=1}^{K} AUC_{IPW}(Y_k) \tag{3.5}$$

The IPW estimates (3.4) and (3.5) are unbiased for the true accuracy estimates, as long as the sampling weights π_i are either known or can be correctly estimated. For the mlSGS design implemented with sequential sampling, exact analytical formulas for sub-sample and

overall sampling weights are shown in Lemma 2. Equation (3.6) shows that sub-sample sampling weights consist of two parts: the ratio of size of sub-sample $\mathcal{S}_k$ relative to the cohort size ($\frac{n_k}{N}$) and the ratio of proportion of surrogate cross-classification values in sub-sample $\mathcal{S}_k$ compared to the that in cohort ($p^{(k)}_{\tilde{Z}_i}$). Since a subject can only appear in at most one sub-sample, overall sampling weights is the sum of sub-sample sampling weights, weighted by subject surrogate statuses. (3.7) may be returned by the sampling algorithm at the design stage, or computed at the analysis stage as long as the sampling indicator vector $\tilde{S}$ is available.

Lemma 2 *Sampling weights of mlSGS.*

Let the mlSGS design be implemented with sequential sampling. For subject i the sub-sample sampling weights $\pi_{\cdot,i} := P(i \in \mathcal{S}_\cdot)$ *are*

$$\pi_{SRS,i} = \frac{n_{SRS}}{N}$$
$$\pi_{k,i} = \begin{cases} \frac{n_k}{N} \times p^{(k)}_{\tilde{Z}_i}, & Z_{ki} = 1 \\ 0, & Z_{ki} = 0 \end{cases} ; k = 1, \ldots, K, \tag{3.6}$$

where $p^{(k)}_{\tilde{Z}_i} = \frac{P(\tilde{Z} = \tilde{z} | S_k = 1, S_{k'<k} = 0)}{P(\tilde{Z} = \tilde{z})}$. *Assuming* $n << N$, $p^{(k)}_{\tilde{Z}_i} \approx P(Z_k = 1)^{-1}$. *Then the sampling weights into the overall mlSGS sample are:*

$$\pi_i = \frac{n_{SRS}}{N} + Z_{1i}\frac{n_1}{N}p^{(1)}_{\tilde{Z}_i} + \ldots + Z_{Ki}\frac{n_K}{N}p^{(K)}_{\tilde{Z}_i}. \tag{3.7}$$

Ultimately, Lemma 2 implies that, since π_i is known by design, using (3.7) in IPW-estimators such as (3.4) and (3.5), results in theoretically unbiased estimates for true AUC metrics; we demonstrate this by simulation in Section 3.4.4. Note that to obtain unbiased estimates of

model evaluation measures, using the sequential implementation is not necessary: as long as induced sampling weights can be correctly estimated, any implementation of the multi-label surrogate sampling class may be appropriately used for valid inference.

3.4 Simulations

Having established the multi-label design framework, implementation, and impact on modeling, we now present simulation results on design benefit for improved classification accuracy, as well as demonstrate certain factors that affect design benefit.

3.4.1 Simulations set-up

We simulated a cohort of $N = 100,000$ subjects, where the data consists of features $\tilde{X} \in \{0,1\}^p$, surrogates $\tilde{Z} \in \{0,1\}^K$, and outcomes $\tilde{Y} \in \{0,1\}^K$, having joint distribution factored as $f(\tilde{X}_i, \tilde{Z}_i, \tilde{Y}_i) = f(\tilde{Y}_i|\tilde{Z}_i, \tilde{X}_i) f(\tilde{Z}_i, \tilde{X}_i)$.

Surrogate and feature generation: The surrogate-feature joint distribution $f(\tilde{Z}_i, \tilde{X}_i)$ was generated by first simulating multi-variate normal variables $[\tilde{Z}^{mvn}, \tilde{X}^{mvn}]$, where

$$\begin{aligned} \begin{bmatrix} \tilde{Z}_i^{mvn} & \tilde{X}_i^{mvn} \end{bmatrix} &\sim MVN(\tilde{\mu} = \tilde{0}, \mathbf{\Sigma}), \\ \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{\Sigma}_Z & \mathbf{\Sigma}_{X,Z} \\ \mathbf{\Sigma}_{Z,X} & \mathbf{\Sigma}_X \end{bmatrix}. \end{aligned} \tag{3.8}$$

In (3.8), $\mathbf{\Sigma}_{X,Z}$, $\mathbf{\Sigma}_{Z,X}$, and $\mathbf{\Sigma}_X$ are identity matrices, while for the $K \times K$ sub-matrix $\mathbf{\Sigma}_Z$ off-diagonals indicate correlation among surrogates. Then, binary random variables $[\tilde{Z}_i, \tilde{X}_i]$ were obtained by thresholding $[\tilde{Z}^{mvn}, \tilde{X}^{mvn}]$ with appropriate cut-offs to control overall surrogate and feature proportions.

Multi-label outcome generation: The conditional multi-label outcome distribution $f(\tilde{Y}_i|\tilde{Z}_i, \tilde{X}_i)$

was simulated with the Multivariate Bernoulli (MVB) model [32, 33], which is a log-linear formulation with relationships between predictors $[\tilde{Z}_i, \tilde{X}_i]$ and outcomes $\tilde{Y}_i$ specified through the parameter matrix $\boldsymbol{B}$ having dimension $1 + K + p$ rows and $2^K - 1$ columns. Data was generated as

$$\begin{aligned} \begin{bmatrix} \tilde{Y}_i & | \tilde{Z}_i & \tilde{X}_i \end{bmatrix} &\sim MVB(\tilde{p}_i) \\ g^{-1}(\tilde{p}_i) &= [\tilde{Z}_i, \tilde{X}_i]^T \boldsymbol{B} \\ \boldsymbol{B} &= \begin{bmatrix} --- & \tilde{\beta}_0 & --- \\ \boldsymbol{B}_Z^{(1)} & \boldsymbol{B}_Z^{(2)} & \boldsymbol{B}_Z^{(H)} \\ \boldsymbol{B}_X^{(1)} & \boldsymbol{B}_X^{(2)} & \boldsymbol{B}_X^{(H)} \end{bmatrix}, \end{aligned} \tag{3.9}$$

with canonical link g^{-1} established in [33] under the Generalized Linear Model (GLM) framework. For $\boldsymbol{B}$ in (3.9), $\tilde{\beta}_0$ is the vector specifying outcome prevalences, while sub-matrices $\boldsymbol{B}_Z$ and $\boldsymbol{B}_X$ indicate surrogate and feature effects respectively on first-order individual outcomes ($^{(1)}$), pairwise outcomes ($^{(2)}$), or higher-order interactions ($^{(H)}$). We set higher-order interactions ($\boldsymbol{B}_Z^{(H)}$ and $\boldsymbol{B}_X^{(H)}$) to be zero while specifying individual and pairwise predictor relationships through the other sub-matrices in $\boldsymbol{B}$. Feature effects are chosen such that only 20% of elements in $\boldsymbol{B}_X^{(.)}$ are truly non-zero, while surrogate effects were set such that for $k = k'$ surrogate Z_k is case-enriching for outcome $Y_{k'}$.

Sampling designs: First, from the simulated cohort a large validation set of size $n_{val} = 10000$ was set aside. Then from remaining subjects development samples of various sample sizes were drawn using either SRS, or mlSGS implemented with sequential sampling (Algorithm 1).

Machine-learning modeling: For data generated using (3.9), each outcome Y_k is marginally Bernoulli [33]. Therefore, for each of the K outcomes we fitted individual logistic Lasso [127]

models following the Binary Representation framework [13], where for each Y_k, $k = 1, \ldots, K$,

$$\begin{aligned}\hat{\beta}_0, \hat{\tilde{\beta}}_Z, \hat{\tilde{\beta}}_X = \underset{\beta_0, \tilde{\beta_Z}, \tilde{\beta_X}}{min} \{ &- \sum_{i=1}^{n} Y_{ki}(\beta_0 + \sum_{k=1}^{K} \beta_{Z_k} Z_{ki} + \sum_{j=1}^{p} \beta_{X_j} X_{ji}) \\ &+ \log(1 + \exp(\beta_0 + \sum_{k=1}^{K} \beta_{Z_k} Z_{ki} + \sum_{j=1}^{p} \beta_{X_j} X_{ji})) \\ &+ \lambda \sum_{j=1}^{p} ||\beta_j||_1 \},\end{aligned} \tag{3.10}$$

For $B = 1000$ draws of development samples of each considered sample size and sampling design, model (3.10) was fitted for each outcome. The penalization parameter λ was selected with 10-fold cross-validation on the development sample using an AUC loss function; surrogate vector $\tilde{Z}$ was always kept in the model by using a zero penalty.

Classifier evaluation: For models (3.10) fitted on each simulated development samples, resulting estimated models were applied to the set-aside validation sample. The empirical validation AUC in classifying each finding was calculated with (3.4), and the macro-AUC over all models was computed with (3.7), where $\pi_i = 1$ (no re-weighting) was used since the validation sample was based on SRS.

3.4.2 Simulations: mlSGS design benefit for multi-label classification

To demonstrate the benefit of using mlSGS designs for multi-label outcome classification, we first considered a scenario with features $\tilde{X}_i \in \{0,1\}^{100}$, surrogates $\tilde{Z}_i \in \{0,1\}^2$, and outcomes $\tilde{Y}_i \in \{0,1\}^2$. The data generating mechanism was

$$\begin{aligned}\begin{bmatrix} Y_{1i}, Y_{2i} & | Z_{1i}, Z_{2i} & \tilde{X}_i \end{bmatrix} &\sim MVB(\tilde{p}_i) \\ g^{-1}(\tilde{p}_i) &= [Z_{1i}, Z_{2i}, \tilde{X}_i]^T \boldsymbol{B},\end{aligned}$$

with $[Z_{1i}, Z_{2i}, \tilde{X}_i]$ generated from thresholding standard multi-variate normal distributed variables having cut-offs of 1.5. The parameter matrix $\boldsymbol{B}$ had

$$\tilde{\beta}_0 = \begin{bmatrix} -3.7 & -2.9 & -0.35 \end{bmatrix}^T$$
$$\boldsymbol{B}_Z^{(1)} = \begin{bmatrix} 2.4 & -0.1 \\ -0.1 & 1.8 \end{bmatrix},$$

and $\boldsymbol{B}_X^{(1)}$ such that for the first two columns, the 20 features with marginal prevalence closest to 10% had values 0.70, and is zero otherwise. We did not impose any pair-wise or higher-order interactions for this simulated scenario. The simulation parameters induced a dataset where each outcome had a marginal prevalence of about 5% and that there were no correlations in either the surrogates or outcomes. For each $k = k'$, surrogate Z_k was case-enriching for outcome $Y_{k'}$, with

$$\begin{aligned} \text{Sensitivity: } & P(Z_k = 1 | Y_{k'} = 1) = 0.40 \\ \text{Specificity: } & P(Z_k = 0 | Y_{k'} = 0) = 0.95, \end{aligned} \tag{3.11}$$

and for $k \neq k'$ surrogate sensitivities and specificities were 10% and 90% respectively. (3.11) was defined based on previous results suggesting that surrogates with these operating characteristics, even though weak predictors individually, are ideal as sampling variables for case-enrichment [125].

In these simulations, we also compared the effect of using unequal sub-sample weights through two mlSGS designs: mlSGS_1 and mlSGS_5. The difference between the mlSGS design is in the ratio of surrogate positive sub-sample proportions to the SRS sub-sample proportions, assuming equal proportions for all surrogate positive sub-samples. For mlSGS_1, sub-sample weights were $(w_1, w_2, w_{SRS}) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, while for mlSGS_5, $(w_1, w_2, w_{SRS}) = \left(\frac{5}{11}, \frac{5}{11}, \frac{1}{11}\right)$.

Figure 3.2: Learning curves of mean validation AUC versus development sample size for multi-label classification of $K = 2$ outcomes, comparing the SRS, mlSGS_1 and mlSGS_5 sampling designs.

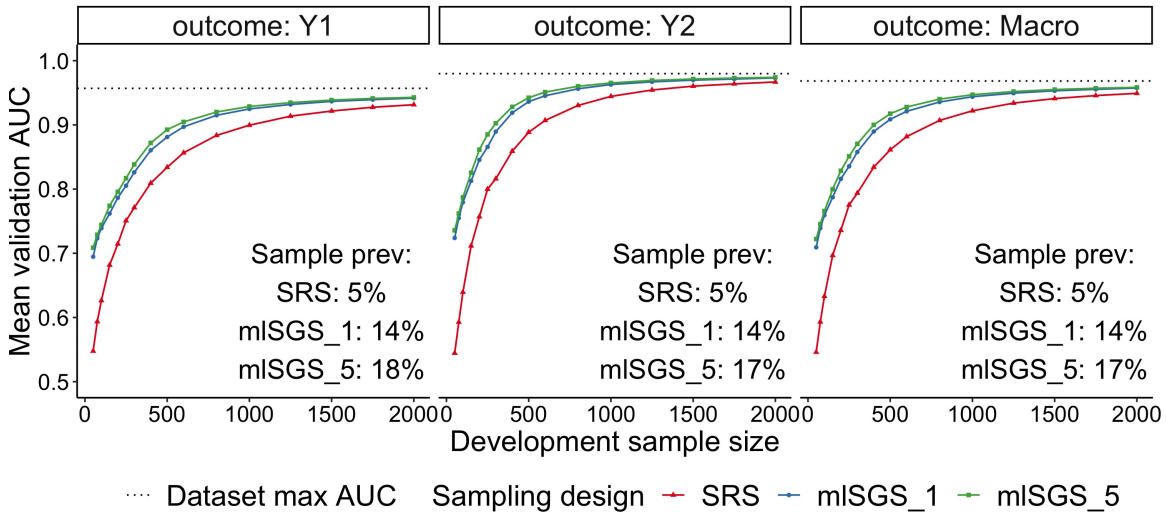


Figure 3.2 illustrates the learning curves of mean validation AUC as a measure of learning benefit versus development sample size as a measure of cost. First, consider the learning curves for outcome Y_1, where at every considered sample size, model validation AUC was higher when using mlSGS compared to using SRS. This can be attributed to the increased sample case-enrichment induced through mlSGS: average sample prevalence was 14%, a three-fold increase from the naturally occurring 5%. Note that the over-inclusion of surrogate positives in mlSGS_5 as compared to mlSGS_1 resulted in some learning benefit, but only very slightly. This is due to higher surrogate positive sub-sample weights only resulting in modestly higher overall outcome prevalence (18% versus 14%). Similar conclusions were also noted for outcome Y_2, where the learning curves were overall higher due to the higher theoretical maximum AUC (0.98 compared to 0.95 for Y_1). The macro-AUC learning curves were the average of the Y_1 and Y_2 learning curves, and therefore also exhibit similar design

effects.

3.4.3 Simulations: Effect of increasing K and presence of correlations on mlSGS benefit.

We now consider the effects of increasing the number of sub-samples K as well as the presence of correlations on mlSGS design benefit. Practically, it may be of interest to simultaneously enrich for cases and abstract for multiple outcomes under a single mlSGS design. In deciding which outcomes to include in a single sample, it may be important to also evaluate if design benefit depends on correlation. For this set of simulations, we generated $\mathbf{X} \in \{0,1\}^{90}$, $\tilde{Z} \in \{0,1\}^{10}$, and $\tilde{Y} \in \{0,1\}^{10}$ according to (3.8) and (3.9). From the "full" simulated dataset, subsets of the first 2, 4, 6, 8, or all 10 elements of $\tilde{Y}, \tilde{Z}$ were selected to illustrate the effects of increasing K. We simulated multiple versions of the "full" synthetic dataset, inducing varying levels of correlations among the surrogates and among the outcomes. Correlations among surrogates $\tilde{Z}$ were induced through setting $\mathbf{\Sigma}_Z$ in (3.8) to be

$$\mathbf{\Sigma}_Z = \sigma_Z^s \begin{bmatrix} 1 & \dots & \rho_Z \\ \vdots & 1 & \vdots \\ \rho_Z & \dots & 1 \end{bmatrix},$$

with $\rho_Z = 0, 0.30, 0.60$ to simulate varying (equal) levels of correlations among the surrogates. To induce correlation among outcomes $\tilde{Y}$, the sub-matrix $\boldsymbol{B}_Z^{(2)}$ in (3.9) was set such that $\boldsymbol{B}_Z^{(2)} = \rho_Y$, with $\rho_Y = 0, 0.30, 0.60$. Simulation parameter values were selected to illustrate general trends of correlation effects on design benefit. Similar to the $K = 2$ case, for each $k = k'$, surrogate Z_k was case-enriching for outcome $Y_{k'}$, with

$$\text{Sensitivity: } P(Z_k = 1 | Y_{k'} = 1) = 0.40$$
$$\text{Specificity: } P(Z_k = 0 | Y_{k'} = 0) = 0.95.$$

Figure 3.3 illustrates the average validation macro-AUC versus the number of multi-label outcomes, comparing development samples of size $n = 500$ collected using SRS or mlSGS with equal sub-sample sizes (mlSGS_1), for scenarios where outcomes (Figure 3.3a) or surrogates (Figure 3.3b) were correlated.

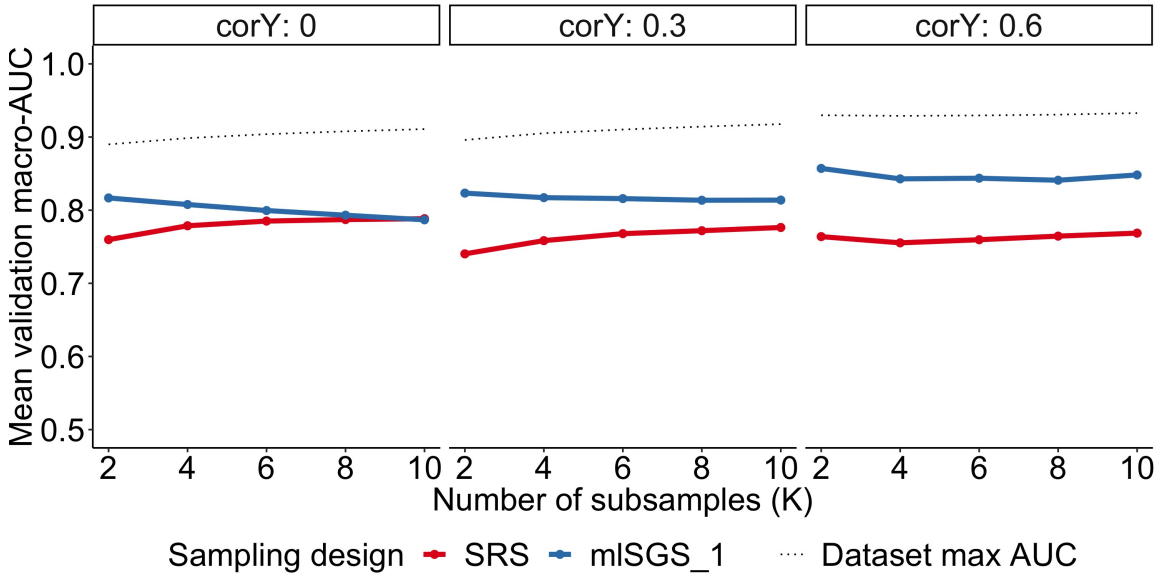
Figure 3.3a illustrates average macro-AUCs in the presence of no correlation among the surrogates, but pairs of outcomes Y_k and $Y_{k'}$ having $Cor(Y_k, Y_{k'}) = 0$, 0.30, 0.60. When outcomes were uncorrelated, the design benefit of mlSGS decreased with the increase of outcome vector dimensionality, where using mlSGS to simultaneously enrich for $K = 10$ outcomes had virtually the same performance as if SRS were used. When outcomes were slightly correlated ($Cor(Y_k, Y_{k'}) = 0.30 \; \forall \; k \neq k'$), the design benefit of mlSGS decreased with the increase in K, however the decrease was less pronounced compared to the no correlation scenario. For mlSGS design, case-enrichment for outcome Y_1 is expected in sub-sample $\mathcal{S}_1$ due to sampling on surrogate positives. When outcomes Y_1 and Y_2 are positively correlated, case-enrichment for Y_1 may also be found in $\mathcal{S}_2$, as cases for Y_2 tend to co-occur with cases for Y_1. Therefore, positive correlations among the outcomes has a synergistic effect using mlSGS designs, with case-enrichment for individual outcomes in multiple sub-samples. In fact, as demonstrated in Figure 3.3a, for highly correlated outcomes ($Cor(Y_k, Y_{k'}) = 0.60 \; \forall \; k \neq k'$), the design benefit of mlSGS over SRS was maintained even when K increased.

Figure 3.3b illustrates average macro-AUCs in the presence of no correlation among the outcomes, but pairs of surrogates Z_k and $Z_{k'}$ having $Cor(Z_k, Z_{k'}) = 0$, 0.30, 0.60. Positively correlated surrogates also had a synergistic effect on mlSGS design benefit, but were less pronounced compared to the correlated outcome scenario. For positively correlated surrogates Z_1 and Z_2, even though the selection of sub-sample $\mathcal{S}_2$ is based only on $Z_2 = 1$, surrogate positives for Z_1 may also co-occur. Assuming that Z_1 is case-enriching for outcome Y_1, then sub-sample $\mathcal{S}_2$ may also be enriched for Y_1 cases. However, except in the scenario of perfect surrogate specificity, surrogate positives do not always result in true cases. Therefore, the

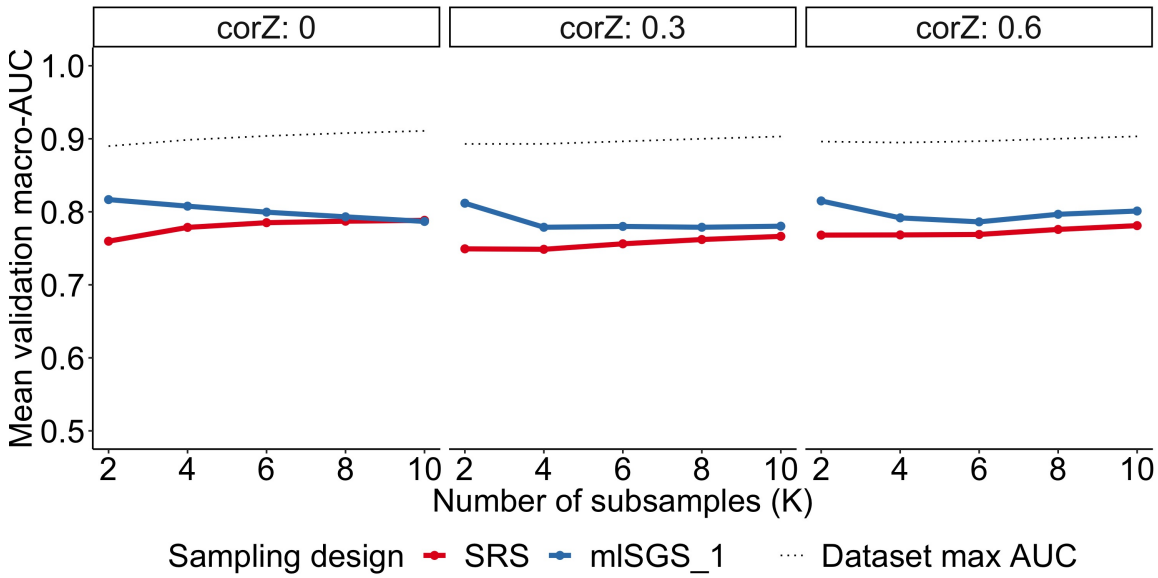
synergistic effect of positively correlated surrogates on mlSGS design benefit can be described as "indirect". As illustrated, even when using highly correlated surrogates $Cor(Z_k, Z_{k'}) = 0.60 \; \forall \; k \neq k'$), increasing the number of sub-samples eventually led to substantial decrease in mlSGS design benefit.

Figure 3.3: Mean validation macro-AUC versus number of outcomes K comparing SRS to mlSGS_1 drawn with a development sample size of $n = 500$.

(a) Effect of increasing K on mlSGS design benefit for model development, when correlations among outcomes $\tilde{Y}$ was 0, 0.30, or 0.60, but when there were no induced correlations among the surrogates.



(b) Effect of increasing K on mlSGS design benefit for model development, when correlations among surrogates $\tilde{Z}$ were 0, 0.30, or 0.60, but when there were no induced correlations among the outcomes.



3.4.4 Simulations: Using mlSGS for model validation

In this final set of simulations, we now demonstrate empirically the impact of using mlSGS designs for model validation. Specifically, we demonstrate bias of empirical accuracy measures, verify that using the derived inclusion probabilities (3.7) for the IPW-corrected AUC (3.4) and (3.5) results in unbiased estimates, and compare empirical variance of validation AUC. Using the same data generating mechanism described in Section 3.4.2, we drew a single development sample of size $n = 5000$ using mlSGS_3 and fitted (3.10). To obtain the "true" validation AUC measures, we applied the fitted models which we considered "fixed" to the remaining cohort to obtain predicted probabilities $\hat{p}_{ki}$, which we used to calculate individual AUCs ($AUC(Y_1)$, $AUC(Y_2)$) as well as the macro-AUC. These estimates, which are based on the full dataset, then served as benchmarks for illustrating the effect of sampling design for validation.

Validation samples were drawn over a grid of sample sizes from remaining subjects, using either SRS, mlSGS_1, or mlSGS_5, and AUC estimates were fitted according to:

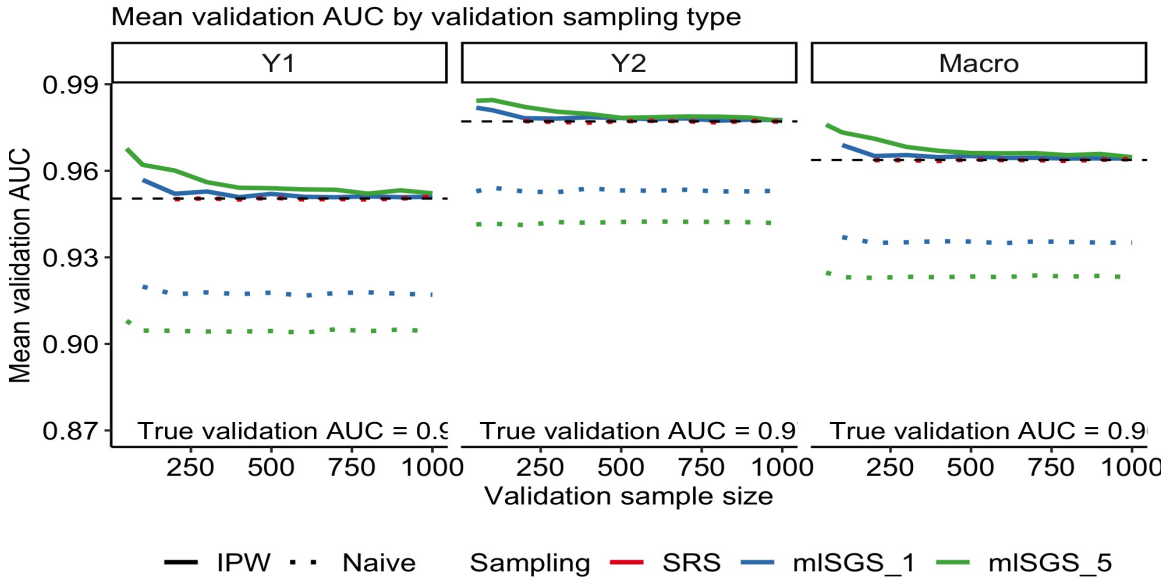
- SRS sampling design: Unweighted AUCs.
- mlSGS sampling designs: Unweighted AUCs ("naive") as well as IPW-corrected AUCs.

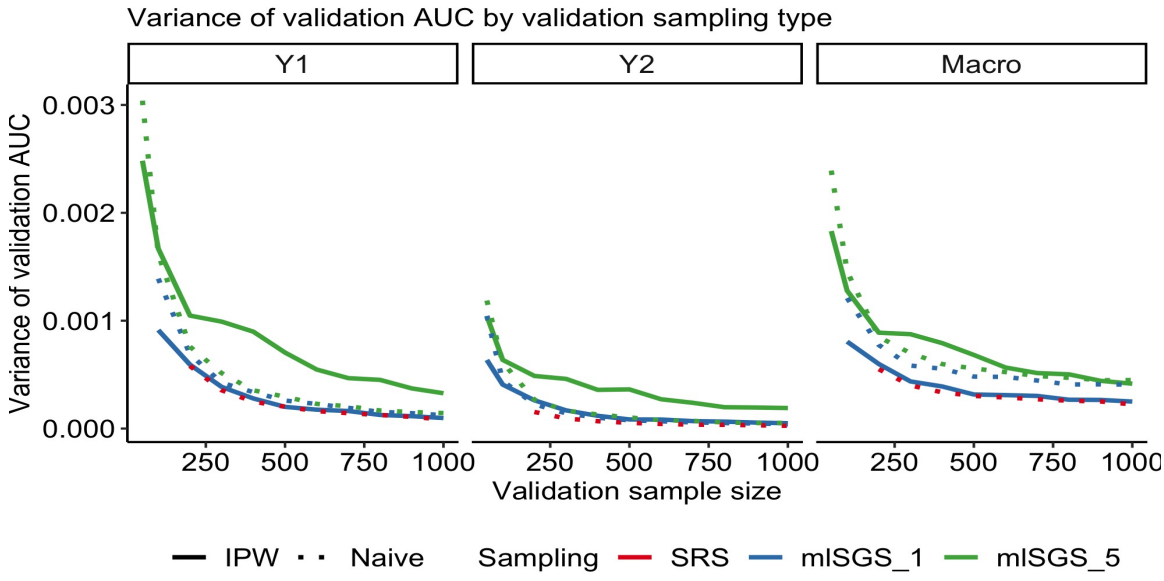
Figure 3.4 illustrates the simulation results of means and variances of resulting validation AUCs across $B = 1000$ iterations. When validation samples are collected with mlSGS, empirically estimated AUC measures were biased for the true values. However, using IPW-corrected estimates (3.4) and (3.5) based on the inclusion probabilities (3.7) results in unbiased estimates for accuracy measures. In terms of variance, due to re-weighting, using mlSGS appears to result in slightly more variable estimates compared to using SRS. In particular, using unequal sub-sample weights (mlSGS_5) results in more variable estimates compared to using equal sub-sample weights (mlSGS_1). This is due to over-inclusion of surrogate positives results in very small inclusion probabilities for subjects selected into the

SRS sub-samples: such "heterogeneity" results in increased variation of IPW estimators [63]. However, we note that the magnitude of difference is on the order of $1e-3$ for estimating a "true" validation AUC of about 0.95, which is not a substantial difference for practical purposes. In fact, for smaller sample sizes (e.g. n=200), using SRS may result in so few or even no cases, therefore making validation impossible. On the other hand, even though re-weighting introduces variation, using mlSGS is more likely to provide a reasonable number of cases for estimation.

Through our simulations, we have demonstrated that using mlSGS can provide resource efficiency over SRS for learning, where proper modeling requires the inclusion of the surrogate vector for estimation. For model development, altering sub-sample design weights did not meaningfully affect classifier learning. Increasing the number of sub-samples reduced mlSGS design benefit, but having positively correlated outcomes reduced the impact of increasing numbers of sub-samples. We have also shown that mlSGS designs may be used to obtain unbiased measures of model validation accuracy measures, as long as inclusion probabilities are accurately estimated. For model validation, using equal sub-sample weights results in the least variable estimates.

Figure 3.4: Estimated mean and validation AUC by sampling method and bias correction type.





3.5 Illustration: Multiple "red flag" findings on lumbar spine radiology reports

3.5.1 "Red flag" findings on lumbar spine radiology reports

We now return to the motivating example of selecting radiology reports for outcome label abstraction and subsequent machine-learning of multiple rare findings. The Lumbar Imaging with Reporting of Epidemiology (LIRE) pragmatic clinical trial [65] included over a quarter million patients whose primary care provider (PCP) ordered an x-ray or Magnetic Resonance (MR) imaging test. An important clinical reason for ordering diagnostic imaging tests is to identify underlying serious conditions, such as aortic aneurysms, infections, spinal malignancies, spondyloarthropathies, and vertebral fractures. A survey of the literature revealed that many of these conditions are also rare, having varying prevalences. The most common among these findings is vertebral fractures (3-20% [134]), followed by spinal malignancies (1-5% [61]). Metastasis to the spine can result in compression fracture, although fractures can also be due to other reasons such as osteoporosis. The much rarer findings were aortic aneurysm (2.2% [76]), an abnormal enlargement of the aorta where ruptures can be fatal, spondyloarthropathy (0.2-2.5% [118]), a group of inflammatory rheumatic diseases that cause arthritis, and spinal infection ($< 1\%$ [41]).

From radiology reports derived from the LIRE study, we were interested in accurate and scalable identification of such "red flag" findings. In order to use machine-learning classification towards this goal, the first step is to obtain an adequate sample of reports for model development and validation. Due to the anticipated extreme rareness of these findings, SRS for report selection is unlikely to provide sufficient cases, therefore motivating or development of alternative sampling strategies. To select lumbar spine radiology reports from the LIRE study, we had applied the proposed multi-label surrogate-guided sampling (mlSGS) design. Here, we first describe the surrogate vector creation process and mlSGS design specification. Then, we provide an analysis using the mlSGS sample, including model development

and model validation. Finally, we compare the expected number of cases under SRS to the observed number of cases under the implemented mlSGS design, and demonstrate the infeasibility of using SRS for machine-learning model development.

3.5.2 Surrogate vector creation and implemented mlSGS design

We created the surrogate vector $\tilde{Z}$ for sampling based on International Classification of Disease (ICD) codes. Surrogates were created as indicators of the presence of any related ICD codes within 90 days of the baseline radiology report text. For subject i, surrogate Z_{ki} for finding Y_k was defined as

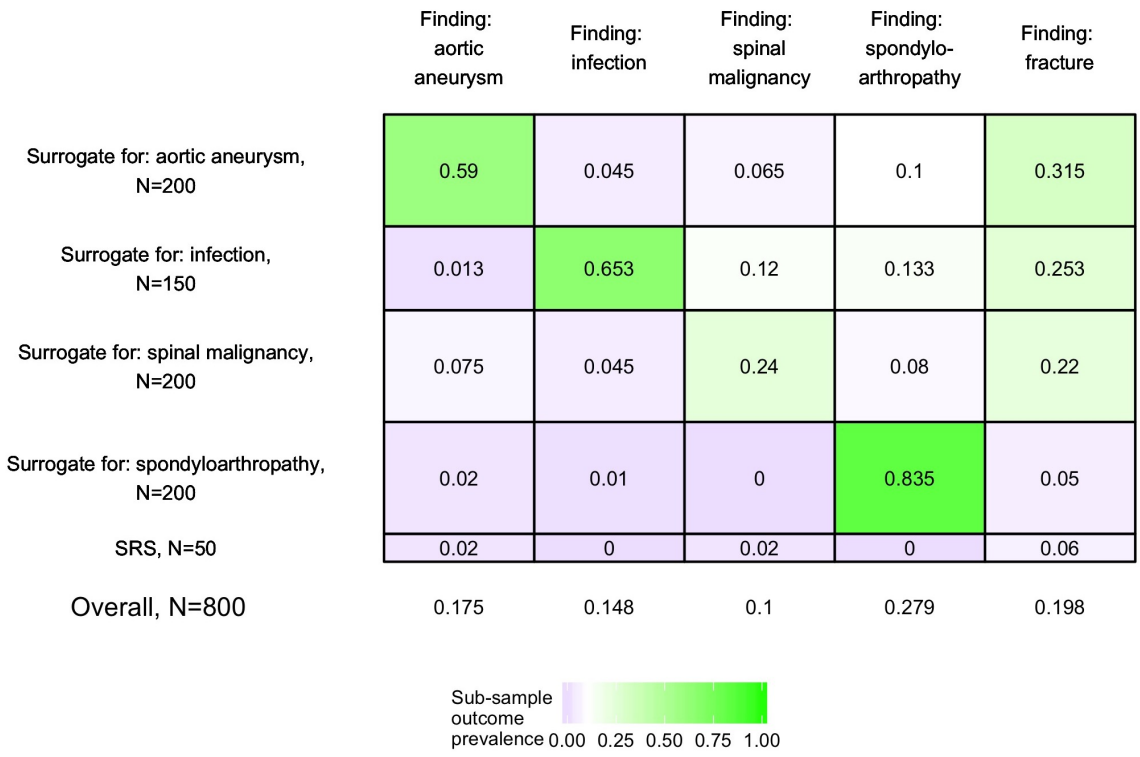
$$Z_{ki} = I(\text{count ICD}_k \text{ for subject } i \in [\text{Day 0, Day 90}] > 0)). \tag{3.12}$$

for all subjects in the cohort. The sets of ICD codes ICD_k (Table B.1 in Appendix B.0.3) were elicited from clinicians based on expert judgment.

We used an mlSGS design to simultaneously target case-enrichment of four findings: aortic aneurysm, infection, spinal malignancy, and spondyloarthropathy. We chose to not enrich for fracture, as we expect this relatively common finding to co-occur with other surrogates or findings. A total of $n = 800$ reports were selected for abstraction, where each report was independently read by two clinical experts for the true status of the 5 findings (including fracture). Figure 3.5 illustrates the sub-sample and overall sample prevalences for each abstracted finding. Note that sub-samples based on surrogate positives results in direct case-enrichment of associated findings: for example sampling surrogate positives for aortic aneurysm resulted in a sub-sample finding prevalence of 59% and an overall sample prevalence of 18%. Even though we did not over-sample surrogate positives for fracture, we found that the resulting mlSGS sample had a higher case proportion compared to what was generally expected. Such case enrichment is likely due to correlations between fracture and the other

findings.

Figure 3.5: Multi-label surrogate-guided sampling (mlSGS) design specification applied to the LIRE data set and resulting sub-sample and overall sample finding prevalences.



3.5.3 Classification model development and validation

We split the mlSGS sample of $n = 800$ into 70% development ($n_{dev} = 560$) and 30% validation ($n_{val} = 240$). Features were unigrams (single words) extracted from radiology report text, excluding common English stop-words. From the 2732 available unigrams, we further filtered out those that were too rare (in less than 10 documents) as well as too common (in more than 80% of the documents), resulting in $p = 629$ unique binary features.

For model development, we used the "baseline" Binary Representation framework for multi-label classification, fitting individual logistic Lasso classifiers (3.10) for each of the 5 findings,

based on the `glmnet` implementation in `R`, imposing a zero penalty on the surrogate vector $\tilde{Z} \in \{0,1\}^4$. We selected the penalization parameter λ that resulted in the most parsimonious model within 1 standard error of the minimum cross-validated error (`lambda.1se`), using 5 fold cross-validation on the development sample and an AUC loss function.

For model validation, we first applied the developed models on the separate validation sample to obtain predicted probabilities. Then, model discrimination for classifying each finding was summarized using the IPW-corrected validation AUC estimator AUC_{IPW} (3.4), with inclusion probabilities calculated using (3.6). The IPW macro-AUC estimator macro-AUC$_{IPW}$ was also computed using formula (3.5). To estimate variation of IPW statistics AUC_{IPW} and macro-AUC$_{IPW}$ under the implemented mlSGS design, we provided 95% confidence intervals based on the stratified empirical bootstrap, summarized over 1000 re-samples implemented with the `boot` function in `R`.

Table 3.2 shows analysis results, where we presented exponentiated surrogate coefficients (interpreted as conditional uncorrected odds ratio) to provide insights into the models. As expected, odds ratios of findings by surrogates are much greater than 1, indicating that these surrogates were indeed case-enriching for associated findings. Additionally, the surrogate for spinal malignancy was predictive for fracture cases, which is likely due to the positive correlation between the spinal malignancy surrogate and fracture finding.

IPW-corrected AUCs demonstrate good discrimination for classifying aortic aneurysm, spondyloarthropathy, and fracture from radiology reports, although some variation was due to the small validation sample size. Classification models for infection and spinal malignancy were less accurate and more variable, partly due to the lower number of cases (see Figure 3.3). The macro-AUC of 0.87 was the average AUC across all 5 models, and the variation of this statistic summarized over re-samples of the implemented design indicated a reasonable performance accuracy when using a simple analytic approach with unigram features and

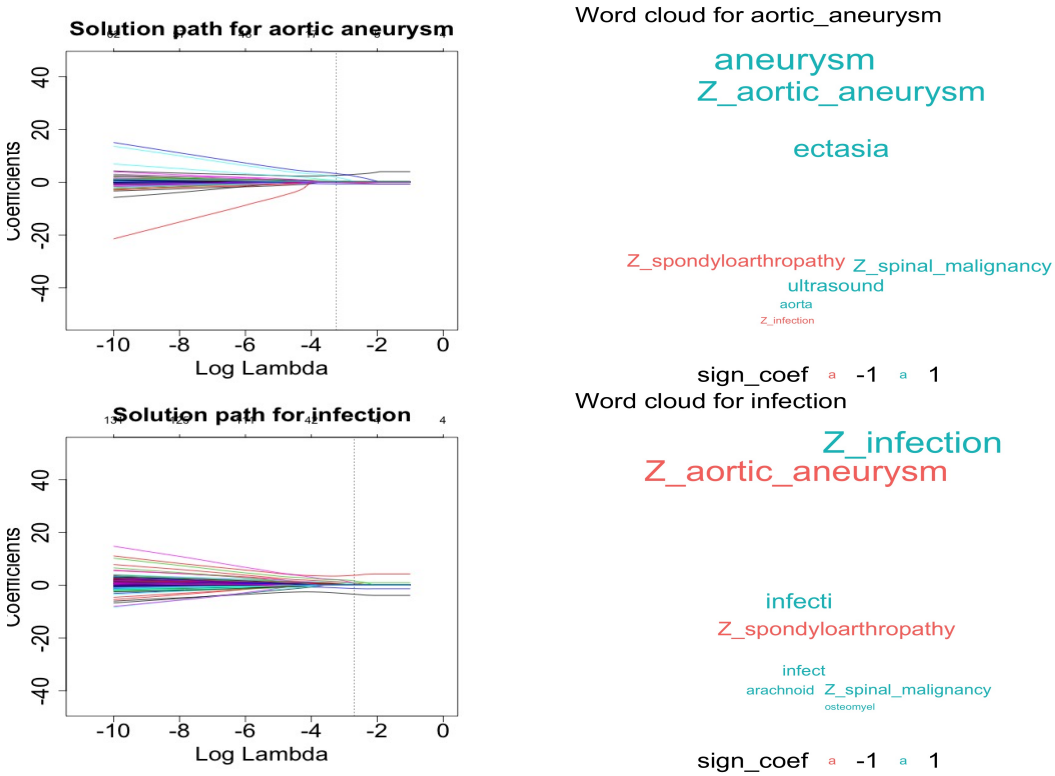
individual Lasso models.

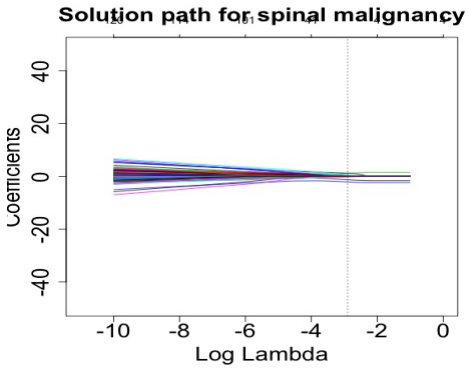
Table 3.2: Estimated odds ratios of findings by surrogates, IPW-corrected estimates for individual findings and macro-AUC, as well as 95% confidence intervals based on the stratified empirical bootstrap. OR = Odds Ratio.

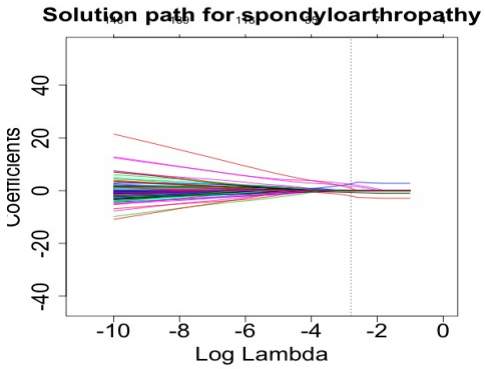
Model for finding:	OR(Z_AA)	OR(Z_Infec)	OR(Z_Mets)	OR(Z_Spondy)	AUC_{IPW} (95% C.I.)
Aortic Aneurysm (AA)	**13.44**	0.89	1.84	0.53	0.89 (0.78, 0.95)
Infection	0.03	**42.28**	2.22	0.28	0.75 (0.60, 0.84)
Spinal Malignancy (Mets)	0.28	0.80	**3.51**	0.12	0.78 (0.60, 0.96)
Spondyloarthropathy	0.48	0.13	0.65	**13.37**	0.95 (0.90, 0.98)
Fracture	1.09	0.47	7.16	0.39	0.92 (0.87, 0.98)
Macro-AUC	-	-	-	-	0.87 (0.81, 0.91)

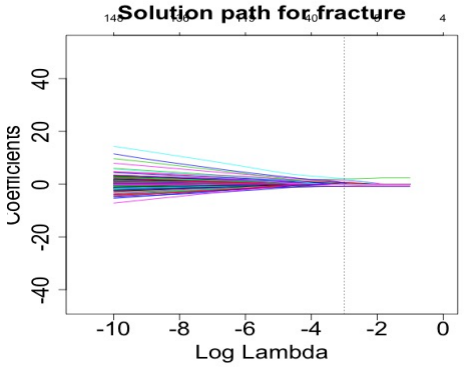
Figure 3.6 illustrates the solution paths and word clouds of selected (estimated non-zero) features for each model. Using the logistic Lasso procedure to classify each finding ultimately selected between 7 to 20 features, including the 4 surrogates. Generally, the selected unigrams seemed to meaningfully predict the findings. For example, terms such as `infect`, `fluid`, `osteomyelitis`, and `discitis`, if present on lumbar spine radiology reports, almost certainly imply the presence of spinal infection. We emphasize that these analyses were possible because of sample selection using mlSGS instead of SRS. As we next demonstrate, using SRS to select reports for outcome label abstraction results in samples that are infeasible for machine-learning due to the expected low case counts.

Figure 3.6: Solution paths and selected coefficients based on using a Logistic Lasso procedure to model the "red flag" findings. Vertical dotted lines indicate the penalization parameter λ (log scale) that resulted in the most parsimonious model within 1 standard error of cross-validated maximum AUC; labels indicate the names of unigram and surrogate features selected at this λ.

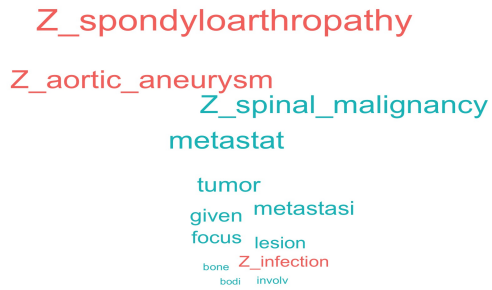








Word cloud for spinal_malignancy

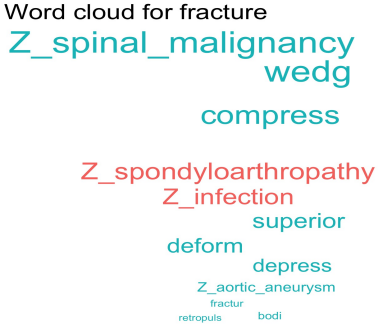


sign_coef a -1 a 1

Word cloud for spondyloarthropathy

sign_coef a -1 a 1

Word cloud for fracture



sign_coef a -1 a 1

3.5.4 Comparison to expected sample prevalences under SRS

We now compare the expected number of cases under SRS to the observed number of cases under mlSGS, based on an abstraction sample size budget of $n = 800$. We calculated the expected number of cases under SRS using estimated sample prevalences from a literature search and a data-driven approach. For the literature search, we summarized estimates obtained from primary care in developed countries based on top review papers in relevant medical epidemiology domains. For the estimation based on the collected mlSGS sample, we used the IPW-corrected prevalence estimator, which is possible due to design validity described in Section 3.3.4. For finding Y_k the IPW-corrected prevalence estimator $Prev(Y_k)_{IPW}$ is

$$Prev(Y_k)_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_{ki}}{\hat{\pi}_i}$$

where $\hat{\pi}_i$ are the estimated inclusion probabilities. We quantified the variation in estimating $Prev(Y_k)_{IPW}$ using the stratified empirical bootstrap with 1000 replicates.

Table 3.3 shows the estimated number of cases under SRS and observed number of cases under mlSGS for each finding, where the estimated SRS cases were calculated by multiplying the $Prev(Y_k)_{IPW}$ estimator by the abstraction sample size budget of 800. Note that the IPW prevalence estimates and 95% bootstrap confidence intervals were approximately within the range of prevalence estimates reported in the literature. Using these prevalence estimates, we calculated that the expected number of cases under SRS were substantially lower compared to the observed number of cases under mlSGS. For example, infection, which is the rarest finding, had an expected yield of 9 cases under SRS: model development is infeasible under this scenario. In fact, to obtain the 118 cases collected through mlSGS would have required over $10,000$ abstracted records through SRS, which is also impractical. Even for the most common finding, fracture, implementing mlSGS instead of SRS results in double the number

of cases. Therefore, for real-world scenarios where the multi-labeled outcomes are expected to be so rare that using SRS is unreasonable, mlSGS designs provide a practical alternative that results in not only more cases for model development, but also samples that are amenable to valid analysis.

Table 3.3: Number of cases estimated under SRS and observed under mlSGS for each finding. Prev^1 is based on a literature review; Prev^2 is based on estimation using Inverse Probability Weighting.

Finding	Prev^1	Prev^2 (95% C.I.)	n_{case}(SRS)	n_{case}(mlSGS)
Aortic Aneurysm	2.2% [76]	1.54% (0.88%, 2.03%)	12.34	140
Infection	<1% [41]	1.14% (0.38%, 1.77%)	9.12	118
Spinal Malignancy	1-5% [61]	3.51% (2.08%, 4.71%)	28.06	80
Spondyloarthropathy	0.2-2.5% [118]	1.53% (0.94%, 1.92%)	12.21	223
Fracture	3-20% [134]	10.77% (9.42%, 11.96%)	86.16	158

3.6 Discussion

We motivated and presented the multi-label surrogate sampling class for multi-label outcome abstraction and subsequent machine-learning of rare multi-labeled outcomes from unstructured biomedical text data. The proposed class of designs is based on sampling using surrogates, which are related summaries of structured data elements, such as simple keyword searches and ICD codes. When there exists surrogates with high specificity for the outcomes of interest, which we called the case-enriching assumption, using the multi-label surrogate-guided sampling design (mlSGS) results in higher average sample prevalence and therefore improved classification model learning. The proposed mlSGS design may be implemented with a sequential sampling algorithm, which results in distinct samples.

We demonstrated the resource efficiency of using mlSGS designs for machine-learning model development, and characterized design impact on both model development and model validation. We showed that any design and any implementation of the general multi-label surrogate

sampling class, results in valid samples for model development, and may be used for model validation with appropriate corrections such as the IPW. Using the mlSGS design under the case-enrichment surrogate assumption further results in resource efficient samples, which allows classification models to achieve higher generalizable performance compared to using another SRS sample of the same size. Such resource efficiency decreases with the number of sub-samples, where using mlSGS was ultimately no better than SRS when attempting to simultaneously enrich for 10 outcomes. On the other hand, positively correlated outcomes have a synergistic effect on mlSGS resource efficiency, where the negative effect on increasing numbers of sub-samples can be substantially reduced. Such results indicate that mlSGS designs may most benefit abstraction and machine-learning tasks targeting a moderate number (around 2-8) of positively correlated multi-labeled outcomes.

The simplest alternative to mlSGS is to use SRS, which we demonstrated was infeasible for real-world application with extremely low prevalences. Using the machine-learning resampling methods such as Individual Label or Label Powerset sampling [21] may improve classification learning. However, it is unclear how resulting replications directly benefits learning, and how to account for any introduced bias for model validation. Another commonly used approach is so-called convenience or heuristic sampling, where reports for abstraction may be selected based on what "seems" likely to be a case, with the goal of increasing the case yield for modeling [95, 94]. We have observed that many such sample selection are implicitly based on auxiliary variables, which are very similar to the "surrogates" that we have defined. Such heuristic selection can be formalized by explicitly stating the variables and criteria for sampling - where we suggest using the multi-label surrogate sampling framework - so that resulting samples may be amenable for valid analysis.

There are a few limitations to the proposed mlSGS design and the current work. The mlSGS design assumes that a case-enriching surrogate is readily available. Although certain ICD codes have been noted to be specific for the true clinical outcomes and are appropriate sur-

rogates, curating an appropriate surrogate vector may take substantial clinical expertise. Additionally, both asymptotic validity for model development as well as unbiased estimation of accuracy are only theoretically guaranteed on the support of the surrogate vector in the sample. To ensure generalizable inference to healthy controls (where $\tilde{Z} = \tilde{0}$), our proposed mlSGS design intentionally includes an SRS sub-sample, but a sub-sample consisting of only surrogate negatives would also be a reasonable alternative. In our work, we only investigated using mlSGS on individual classification models, which is also known as the Binary Representation in multi-label classification. Even though we did not explicitly demonstrate other modeling approaches, it is likely that the mlSGS design benefit extends to many other multi-label models, as was shown in the effect of using individual-label sampling on classification across a variety of algorithms [21].

In summary, through extending ideas from the complementary fields of epidemiology and machine-learning, we have developed a formal sampling framework for multi-label outcome abstraction for machine-learning model development and validation. A potential next step of this work is to investigate multi-label sampling methods that leverage surrogate correlations, which we expect will further improve the resource efficiency of the surrogate-based sampling methodologies. In addition, even though we focused on binary surrogates, certain scenarios may have quantitative surrogates - associated designs for that setting warrants attention. These further developments may allude relevant statistical trade-offs to guide formalization of sample size calculations towards rigorous project planning in practical resource constrained scenarios.

Chapter 4

PREDICTIVE CASE CONTROL DESIGNS FOR MODIFICATION LEARNING

4.1 Introduction

Clinical prediction models are often developed to help determine a specific diagnosis, to inform prognosis, and to define subgroups relevant for various clinical outcomes. Often, a clinical prediction model is developed based on data drawn from a source cohort, for example a particular hospital or research network. To facilitate sharing of learned patterns from established clinical data sources, a developed model may also be modified and applied to new settings, for example different health systems, age groups, or time frames. Before adopting model predictions in the new setting, it is important to ask whether model predictions provide an accurate representation of the true risks in the target population. For example, is there sufficient agreement between observed outcomes and model predictions? Can the predictions sufficiently distinguish between cases and controls? In an ideal situation, a model would be perfectly generalizable to the new setting without any modification. Unfortunately, a model can be valid in the source setting yet invalid when applied to new cohorts, due to reasons such as case mix differences, model over-fitting, or true differences between the populations [115]. Therefore, the developed model needs to be assessed for validity in the new setting, and modified appropriately [115, 117].

Data driven model updating or modification requires the collection of an adequate sample from the new setting. Outcome labels are often labor intensive to collect, as they require abstraction from the medical record, yet substantial sample sizes are necessary for modification learning studies. For example, systematic recalibration adjustment requires 312

observed events if predictions were 25% too extreme on the odds scale [131]. For rare outcomes, such substantial sample size requirements may be difficult to obtain without larger or targeted samples. In this paper, we motivate and describe a sampling strategy targeted for the modification learning problem. The proposed approach leverages the original model scores in order to design efficient data collection. Our proposed sampling design is motivated by strategies from experimental design and machine-learning, and may reduce the sample size requirements for model modification learning and assessment.

4.2 Background

4.2.1 Modification learning of clinical prediction models

When applying a previously developed model to a new setting, predictions based on the original model may be invalid, due to reasons such as differences in case mix or differences in regression coefficients [115]. Case mix is when the distribution of features or outcomes are different between the source and new setting. Differences in regression coefficients may occur from model over-fitting on small development datasets or from truly different populations due to different cohort selection criteria. Updating or modification learning of a prediction model to new settings should involve empirical procedures such as model recalibration, revision, and extension [115, 117].

Model recalibration involves the systematic adjustment of potentially inaccurate predictions. However, since unnecessary recalibration may introduce variation for model application, whether the original model is sufficiently calibrated in the new setting may be first formally tested. For example, testing can be based on the Hosmer-Lemeshow test [64], where estimated and observed predictions are compared across strata of grouped averages. Since the Hosmer-Lemeshow test involves variable or arbitrary binning choices, the logistic recalibration [31] test is potentially more powerful, where specific deviations from the null hypothesis indicate if recalibration is required. If model recalibration is deemed necessary, the Cox

logistic recalibration model [31] adjusts overall mean predictions as well as systematically overestimated and underestimated predictions. To increase flexibility, the linear-logistic assumption may be relaxed by modeling additional shape parameters in the sigmoid function [119] or non-linear functions of the predicted scores [34].

Model revision is a step beyond recalibration that involves re-estimating individual feature effects instead of overall systematic adjustments. Previously estimated coefficients can be refitted in a stepwise manner [116] or through shrinkage and penalized models for larger numbers of coefficients [87]. In addition, the original model may be extended to include additional features. Model revision can be done without recalibration, where the model is completely re-fitted to the new data [116]. Since complete revision may potentially be unstable and ignores past knowledge, a more strategic approach combines both recalibration and revision in the same model. If data is collected in batches from the new setting, models may also be continuously modified, where model recalibration is suggested to take precedence over model revision and/or extension [115].

4.2.2 Information criteria for modification learning

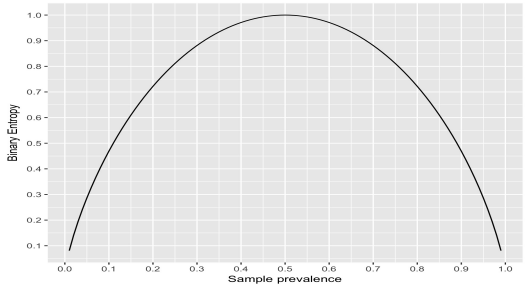
For modification learning, an adequate sample of actual outcome labels needs to be collected from the new setting. Due to the potential substantial sample size requirements [131] under simple random sampling (SRS), an alternative is to consider sampling designs specifically targeted towards the modification learning problem. To evaluate the "information" of samples for modification learning, here we review relevant statistical criteria from the related fields of information theory and optimal experimental design.

Information theoretic criteria

In information theory, objective criteria to measure information in a given sample is generally based on maximizing uncertainty or information. For example, Binary Entropy [81] measures the amount of uncertainty for a binary variable (see Figure 4.1 for an illustration). If subjects

in the samples are either all cases or all controls, there is no variation in outcome labels and thus no information for learning - the Binary Entropy is at its minimum of 0. On the other hand, if the sample contains approximately equal cases and controls, then information is maximized and Binary Entropy attains its maximum of 1. Maximizing sample Binary Entropy may help classification, since samples with approximately outcome class balance are generally accepted to improve classifier learning [138, 7, 142]. Another approach of quantifying uncertainty is based on distance to the decision boundary. For example, subjects with initial predicted probabilities close to the cut-off for binary decisions may be considered to have "uncertain" predictions, therefore collecting their true outcome statuses may provide more information.

Figure 4.1: Illustration of Binary Entropy as a function of sample outcome prevalence.



Statistical information criteria

The field of optimal experimental designs offers theoretical guidance for study design when predictors can be chosen, outcomes need to be collected, and practical constraints limit the number of experimental runs. For a given statistical model and associated parameters, information for parameter estimation can be maximized using one-dimensional summaries of the information matrix (summarized in Table 4.1). Many such criteria are considered "true" information functions, having the monotonicity, concavity, and homogeneity properties for comparing matrices ordered in the Loewner sense [102]. This means that if the difference

of two information matrices is positive semi-definite, then associated one-dimensional summaries are similarly ordered. For a fixed sample size, subjects can be allocated based on pre-specified criteria, where the unique predictor values are called "design points", and the proportions of subjects at each design point are called "design weights". Optimal statistical designs provide convenient interpretations, for example a design targeting maximizing D-optimality is one that minimizes the joint confidence region of model parameters [43].

Even though optimal design theory provides an intuitive framework for study planning, finding the optimal design in practice can be a very challenging problem. Early work discussed ''"exact" optimal designs, which is based on identifying combinations of subjects that result in maximum sample information [4]. Alternatively, approximately optimal designs places a probability distribution on the design space, thus allowing for non-discrete design points [69]. For example, in stratified designs, the design space is decomposed into non-overlapping partitions and non-zero weights are placed on each stratum. Even for approximately optimal designs, there is not a single unique "best" design, motivating using Monte Carlo and approximation-based algorithms [5]. In addition, within a class of designs, pre-specified statistical criteria can also be used to evaluate competing configurations, and higher information designs are selected for data collection.

For logistic regression, since the information matrix depends on true parameters, optimal designs need to be computed on true or assumed model parameters and therefore are only locally optimal [68, 5]. For example, in estimating a one-parameter logistic regression model, the locally D-optimal design places equal weight on two symmetric points, but design point placement depends on true parameter values. To account for a range of possible parameter values, approaches such as minimax, sequential, and Bayesian methods can be used [26, 68, 5].

Table 4.1: Selected common statistical optimality criteria, mathematical formulae, and interpretation. $\mathbf{I_m}$ is the Fisher's information matrix and $\mathbf{H}$ is the hat/projection matrix, where for logistic regression $\mathbf{I_m} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}{\mathbf{I_m}}^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$, $w_{ii} = p_i(1-p_i)$.

Criterion	Mathematical Formula	Interpretation of using criteria
A-optimality	$\underset{\zeta}{argmin}\, trace({\mathbf{I_m}}^{-1})$	Minimizes average variance of parameters.
C-optimality	$\underset{\zeta}{argmin}\, trace(c^T{\mathbf{I_m}}^{-1}c)$	Minimizes variance of a best linear unbiased estimator.
D-optimality	$\underset{\zeta}{argmax}\, det(\mathbf{I_m})$	Minimizes volume of parameter joint confidence region.
G-optimality	$\underset{\zeta}{argmin}\, \underset{\zeta}{argmax}\, Diag(\mathbf{H})$	Minimizes the maximum variance of predicted values.

4.2.3 Related problems and research gap

Modification learning of clinical prediction models is related to transfer learning and active learning from machine-learning; here we summarize these related frameworks to motivate our proposed statistical design approach.

Transfer learning

Transfer learning involves the partial modification of an original model to the new setting using source data. Transfer learning strategies may be categorized as either transductive or inductive [96], both of which have connections to modification learning. For transductive learning, similar to the case-mix assumption, feature distributions $f(x)$ between source and new scenarios are assumed to be different. Transductive learning approaches generally assume that no outcome labels are available from the new setting, therefore distributional differences are addressed by fitting a weighted model to source data. On the other hand, inductive learning assumes that conditional outcome distributions $f(y|x)$ are different between source and target, which is a generalization of the different coefficient assumption. For inductive learning, some outcome labels from the new setting are available, but the sample size is too small to generate reliable model predictions. Instead, assuming certain shared parameters, the existing model only needs to be partially modified for transfer to the new

setting. In the biomedical setting, transfer learning has been successfully applied to skin cancer classification [44], but required over 1 million outcome labels for the original model and over 100,000 labels for the partial modification to the new setting.

Active learning

Another related problem is active learning, which is the sequential modification of an original model through querying new data. Several of the previously described information theoretic and statistical information criteria have been incorporated into active learning strategies. For example, pool-based active learning sequentially selects the most informative data points for outcome labeling from a pre-specified cohort "pool". In pool-based active learning, sample selection may be based on heuristically the most uncertain predictions, for example using the Binary Entropy criterion [112], or based on variance reduction, for example using the A-optimality criterion [109]. Curiously, while some researchers advocate designs based on statistical information criteria [109], others have demonstrated that heuristic information criteria such as Binary Entropy perform well in empirical experiments [143].

Active learning has been applied to machine-learning phenotyping models [25], and for classification of cancer from radiology reports [91]. Such applications have noted some sample size savings by using active learning over "passive" learning using a random or convenient sample. However, active learning algorithms have been demonstrated to perform even worse than SRS in some settings [109]. A common criticism is that sequential myopic selection of subjects for outcome labeling without theoretical foundations may introduce unanticipated variation [112]. Furthermore, implementing active learning often requires extensive engineering infrastructure and expertise. In fact, more than half of surveyed applied machine-learning researchers revealed hesitated to use active learning strategies for annotation or abstraction tasks [128]. A concern is due to the unknown introduced bias from resulting samples: even if sampling with active learning may improve prediction accuracy, it is often difficult to use such samples for statistical analysis such as formal inference and evaluation.

Research gap and contribution

In transfer learning, an existing model is partially modified, but new outcome labels are not necessarily collected. On the other hand, the active learning framework involves sequentially procuring of additional outcome labels based on pre-specified statistical criteria, so to achieve learning goals in a resource efficient manner. However, as most active learning algorithms frame data collection as a learning rather than traditional design problem, statistical inference is usually not a goal and therefore bias in resulting samples may not be well-characterized. In the clinical prediction setting, both sample generalizability and resource efficiency are important considerations. Furthermore, modification learning of clinical prediction models are usually based on existing starting points, for example previously published model coefficients and scoring rules. Therefore, this motivates a design-based framework for new outcome label collection and subsequent modification learning, particularly for practical settings where personnel and monetary resources constraint abstraction label sample sizes.

4.3 Methods

4.3.1 Statistical notation and motivation

Denote the source cohort as $\mathcal{D}^0$ and the new cohort as $\mathcal{D}^1$. For subject i denote features as $\tilde{X}_i^T \in \mathcal{R}^p$ and binary outcome labels as Y_i. We assume that the original clinical prediction model denoted $\hat{h}(.)$ was developed using a sample drawn from $\mathcal{D}^0$. Applying $\hat{h}(.)$ to the data in $\mathcal{D}^1$ will generate original model prediction scores in the new data denoted as

$$S_i = \hat{h}(\tilde{X}_i^T). \tag{4.1}$$

The original model $\hat{h}(.)$ may need to be modified before ultimate reliable application to the new setting. Predictions may be recalibrated and model parameters potentially revised. We

consider simultaneous recalibration and revision models [116] of the form

$$logit(E[Y_i|S_i, \tilde{X}_i^T]) = \alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma}. \tag{4.2}$$

In (4.2), α_0, α_1, and $\tilde{\gamma}$ are the model modification learning parameters to be estimated. The recalibration parameters α_0 and α_1 re-adjust any systematic mis-estimation in predicted scores, while the revision parameters $\tilde{\gamma}$ identify any new predictive features or refinement of coefficients for original features. Assuming sparsity, (4.2) can be modeled using the Lasso procedure [127],

$$(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\tilde{\gamma}}) = \underset{\alpha_0, \alpha_1, \tilde{\gamma}}{argmin} \left\{ \sum_{i=1}^{n} -Y_i(\alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma} + \log(1 + \exp(\alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma})) + \lambda ||\tilde{\gamma}||_1 \right\} \tag{4.3}$$

which uses penalization for estimation. Note that in (4.3), only revision parameters $\tilde{\gamma}$ are penalized, so that original scores S are retained in the model. To fit model (4.2) requires the collection of a sample drawn from the new setting $\mathcal{D}^1$, where features $\tilde{X}_i^T$, scores S_i, and outcome labels Y_i are available for all subjects. We assume that $\tilde{X}_i^T$ is easily available, S_i can be generated using (4.1), but outcome labels require expensive and time-consuming abstraction. Motivated by information theoretic and statistical information criteria, we describe a sampling design framework based on scores S, so to select subjects from the new setting for outcome label abstraction and subsequent model modification learning.

4.3.2 Predictive Case Control (PCC) designs

For outcome label collection from the new setting, we define sampling strategies based on original model scores S as the predictive score sampling class (Definition 1). The name "predictive" is due to that, arguably among all observed variables, scores S are most predictive of unobserved true outcomes Y. Among possible ways to sample based on S, note that

subjects with higher scores tend to be "cases" ($Y = 1$), while those with lower scores tend to be "controls" ($Y = 0$).

Definition 1 *Predictive score sampling class.*
For new outcome label collection, denote the class where sampling is based only on original model scores $S \in [0, 1]$ *as the predictive score sampling class.*

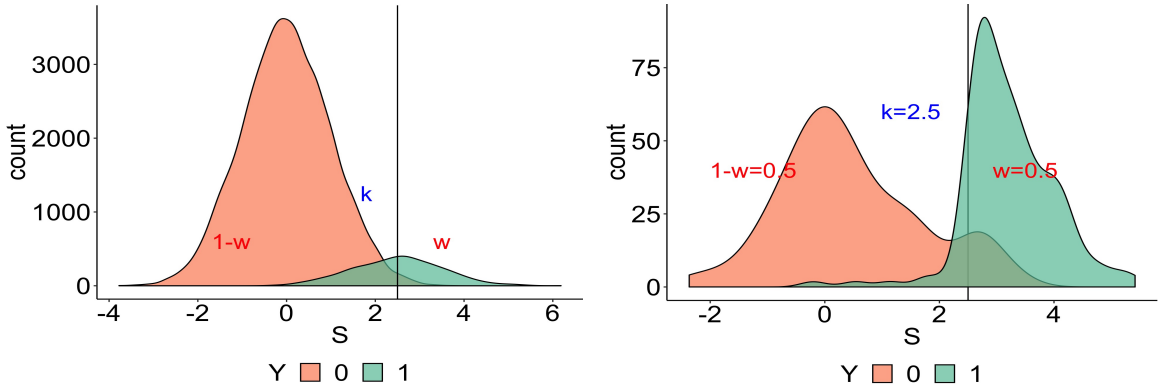
Among the class of predictive score sampling deigns, we define a stratified sampling procedure as the Predictive Case Control (PCC; Definition 2), which are indexed by design configurations defined with score cut-off k and stratum weights $w = P(S > k|\text{sampled})$. For an interpretation of PCC, assume that the score distribution $f(S)$ is a mixture of marginal outcome case/control proportions $P(Y = y)$ and conditional score distributions $f(S|Y = y)$. Then, PCC design configurations w and k can be interpreted as intentionally altering $P(Y = y)$ and $f(S|Y = y)$ in resulting samples, respectively.

Definition 2 *Predictive case control (PCC) design.*
The predictive case control design is a stratified sampling procedure based on original model scores $S \in [0, 1]$*, where for a fixed abstraction sample size* n *and selected design configurations defined with score cut-off* k *and stratum weights* $w = P(S > k|sampled)$*, the procedure is*

$$\textit{Select:} \begin{cases} n \times w \ \textit{ subjects} & \textit{from those with scores} \quad S > k \\ n \times (1 - w) \ \textit{ subjects} & \textit{from those with scores} \quad S \leq k. \end{cases}$$

Figure 4.2 illustrates the effect of using PCC design on induced sample score distribution, where sample scores may have different marginal outcome prevalence as well as conditional score distributions compared to that in the cohort. The intended benefit of using PCC is to reduce sample size requirements for modification learning and evaluation, through selecting subjects with "high information" for modeling goals. In particular, information functions of

Figure 4.2: Left: Score distribution by outcome classes of a simulated cohort with $N = 10,000$; Right: Score distribution by outcome classes of a sample ($n = 300$) drawn from the cohort using a Predictive Case Control design with configurations $k = 2.5$, $w = 0.50$.



scores motivated by the dual modification learning goals of recalibration and revision may be used to select appropriate design configurations.

Recalibration goals: Statistical power and D-optimality

For recalibration, since unnecessary readjustments may introduce additional variation, a first step is to test whether model recalibration is required. From the modification learning model (4.2), recalibration parameters α_0 and α_1 that deviate from 0 and 1 indicate potential mis-calibration in the new setting. For such recalibration testing, which may be based on Likelihood Ratio Tests in Table 4.2, we consider maximizing statistical power as a relevant modeling goal.

Maximizing the local power of likelihood ratio statistical tests has been noted to be statistically equivalent to maximizing the determinant of the information matrix under the null [132]. Therefore, under the assumption that

$$\alpha_0 = 0,\ \alpha_1 = 1,\ \tilde{\gamma} = \tilde{0}, \tag{4.4}$$

Table 4.2: Model recalibration hypothesis tests. LRT = Likelihood Ratio Test. α_0, α_1 respectively indicate the recalibration intercept and slope.

	Recalibration intercept	Recalibration slope	Logistic recalibration
Null hypothesis, H_0	$\alpha_0 \vert (\alpha_1 = 1) = 0$	$\alpha_1 = 1$	$(\alpha_0, \alpha_1) = (0, 1)$
Alternative hypothesis, H_A	$\alpha_0 \vert (\alpha_1 = 1) \neq 0$	$\alpha_1 \neq 1$	$(\alpha_0, \alpha_1) \neq (0, 1)$
Degrees of freedom using LRT	1	1	2

we propose using the D-optimality criterion [69] as a function to summarize information in sample scores. Define the score information function based on D-optimality criterion, $\phi^D(S)$, as

$$\phi^D(S) := \log\left(det\left(\mathbf{I}_m(S)\right)\right). \tag{4.5}$$

where in (4.5), $\mathbf{I}_m(S) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{mi}(S)$ is the sample partial information matrix for recalibration parameters (α_0, α_1), where for $p_i = \text{expit}(S_i)$,

$$\mathbf{I}_{mi}(S) = \begin{bmatrix} p_i(1-p_i) & S_i p_i(1-p_i) \\ S_i p_i(1-p_i) & S_i^2 p_i(1-p_i) \end{bmatrix}$$

Note that $\phi^D(S)$ is always defined, as $\mathbf{I}_m(S)$ is by construction positive semi-definite. Designs with higher values of $\phi^D(S)$ may be interpreted as having lower generalized variances and smaller joint confidence regions of recalibration parameter estimates [43].

Revision goals: Support recovery and Binary Entropy

Also using the modification learning model (4.2), revision parameters $\tilde{\gamma}$ that are non-zero indicate that model revision is necessary. Assuming sparsity for the revision parameters, an evaluation criterion for model revision performance can be based on the support recovery of

truly predictive features. Denote the true revision parameters γ_j^*, $j = 1, \ldots, p$, where each γ_j^* can be either truly predictive ($\gamma_j^* \neq 0$) or non-predictive ($\gamma_j^* = 0$), belonging to one of the mutually exclusive sets $\boldsymbol{\gamma}_{\boldsymbol{S}}^*$ or $\boldsymbol{\gamma}_{\boldsymbol{S}^c}^*$, defined as

$$\begin{aligned} \boldsymbol{\gamma}_{\boldsymbol{S}}^* &= \{\gamma_j^* : \gamma_j^* \neq 0\} \\ \boldsymbol{\gamma}_{\boldsymbol{S}^c}^* &= \{\gamma_j^* : \gamma_j^* = 0\}. \end{aligned} \tag{4.6}$$

Based on fitting Lasso procedure (4.3), denote the estimated revision parameters as $\hat{\gamma}_j$, $j = 1, \ldots, p$, where each estimate $\hat{\gamma}_j$ is either predictive or non-predictive, with corresponding sets defined as

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{\boldsymbol{S}} &= \{\hat{\gamma}_j : \hat{\gamma}_j \neq 0\} \\ \hat{\boldsymbol{\gamma}}_{\boldsymbol{S}^c} &= \{\hat{\gamma}_j : \hat{\gamma}_j = 0\}. \end{aligned} \tag{4.7}$$

Then, a measure for how well $\hat{\gamma}_j$ estimates true γ_j^* can be represented using the False Discovery Rate (FDR) and False Exclusion Rate (FER) measures, where:

$$\begin{aligned} FDR &= 1 - P(\hat{\boldsymbol{\gamma}}_{\boldsymbol{S}} \in \boldsymbol{\gamma}_{\boldsymbol{S}}^*) \\ FER &= 1 - P(\hat{\boldsymbol{\gamma}}_{\boldsymbol{S}^c} \in \boldsymbol{\gamma}_{\boldsymbol{S}^c}^*). \end{aligned} \tag{4.8}$$

In (4.8), FDR can be interpreted as the false positive rate (Type I error), and FER the false negative rate (Type II error), in identifying the set of truly predictive features for model revision. Perfect support recovery is when both FDR and FER are zero. Under certain assumptions on the data generating mechanism[1], support recovery errors tend to decrease with sample size.

[1]assumptions include that model dimensionality p and sparsity ratio $f = \dfrac{k}{p}$, $k = |\boldsymbol{\gamma}_{\boldsymbol{S}}^*|$ do not vary with sample size

We take the view of improving support recovery for a fixed sample size. Note that the modification learning model (4.2) may be framed as a classification model, where a key factor that affects classifier "learning" is outcome class balance in the development sample. Much research on outcome class balance on model prediction accuracy have focused on metrics such as the proportion correct (simple accuracy) or the Area Under the Receiving Operator Characteristic curve (AUC) [138, 7, 142]. However, outcome class balance may additionally affect support recovery accuracy measures. Therefore, we propose using the Binary Entropy criterion as another summary of information based on scores. Define the information function based on Binary Entropy, $\phi^B(S)$, as

$$\phi^B(S) = -\bar{p}(S)log_2(\bar{p}(S)) - (1-\bar{p}(S))log_2(1-\bar{p}(S)), \tag{4.9}$$

where $\bar{p}(S) = \frac{1}{n}\sum_i^n p_i$ with $p_i = expit(S_i)$. Note that $\phi^B(S)$ attains its maximum of 1 if the sample outcome prevalence is exactly 50%. Designs with higher values of $\phi^B(S)$ may be interpreted as having more predicted outcome class balance and therefore better learning of revision parameters.

4.3.3 A computational framework to evaluate PCC design configurations

We now describe a computational framework to evaluate and select configurations (k, w) for PCC designs. The proposed empirical framework is based on Monte Carlo of expected sample score information summarized with $\phi^D(S)$ and $\phi^B(S)$ under considered design configurations. Then, resulting estimated information response surfaces are visualized with pairs of contour plots, which allows for convenient evaluation of information from competing design configurations.

Details of the proposed computational framework are in Algorithm 1. First, two grids varying

in cut-offs k and stratum weights w are specified. Then, for each considered configuration as well as under SRS, expected values of sample information functions are computed using Monte Carlo. Finally, resulting estimated information response surfaces are visualized with pairs of contour plots, where we illustrate an example in Figure 4.3.

Algorithm 1 *Computational framework for PCC design configuration evaluation.*

Input *Scores S; grid $k \in supp(S)$; grid $w \in [0, 1]$.*

Do *Over the grids of $k \in supp(S)$, $w \in [0, 1]$:*

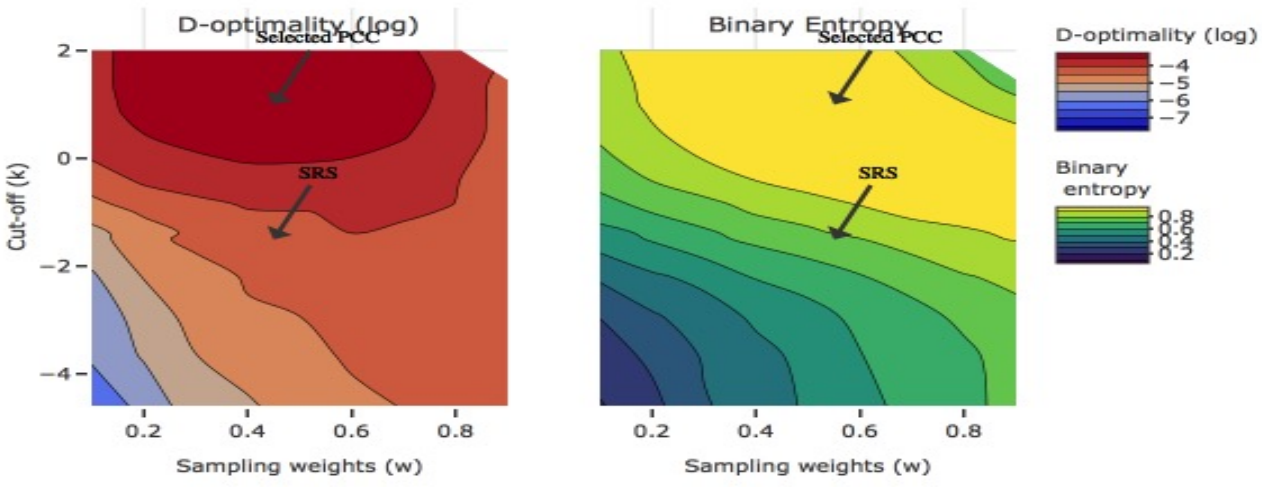
1. *Draw sample of S using configuration (k, w).*
2. *Calculate information functions $\phi^D(S|k, w)$ and $\phi^B(S|k, w)$ based on sample data.*
3. *Repeat 1. and 2. a total of B times.*

Return *Matrices of information response surface for $\phi^D(S)$ and $\phi^B(S)$.*

To illustrate the information response surfaces computed from Algorithm 1, we simulated scores with distribution $S \sim N(-1.5, 1)$. Figure 4.3 shows resulting sample information response surfaces computed under the D-optimality (left) and Binary Entropy (right) information functions, where warmer colors represent better designs. For the example in Figure 4.3, within the class of considered PCC design configurations, expected sample D-optimality $\phi^D(S)$ ranged between -8 and -3 and expected sample Binary Entropy $\phi^B(S)$ ranged between 0.20 and 0.99. We indicated with arrows expected sample information surfaces under SRS as well with a "selected" PCC design with configuration $(k, w) = (1, 0.50)$, where qualitatively the "selected" configuration results in higher information compared to SRS as measured with D-optimality and Binary Entropy.

For a quantitative interpretation of information function comparisons, note that comparing PCC^* having $(k, w) = (1, 0.50)$ to SRS

Figure 4.3: Pairs of contour plots for expected sample information response surfaces calculated using D-optimality (left) and Binary Entropy (right) information functions, based on simulated scores distributed as $S \sim N(-1.5, 1)$.



$$\frac{\phi^D(S|PCC^*)}{\phi^D(S|\text{SRS})} = \frac{-3.12}{-4.12} \quad \text{and} \quad \frac{\phi^B(S|PCC^*)}{\phi^B(S|\text{SRS})} = \frac{0.99}{0.77}.$$

Then, by exponentiating $\phi^D(S)$ the ratio of determinant of information functions comparing designs is

$$\frac{det(\mathbf{I}_m(S|PCC^*))}{det(\mathbf{I}_m(S|SRS))} = \frac{exp(-3.12)}{exp(-4.12)} \approx 2.72,$$

and by reversing $\phi^B(S)$ the ratio of sample outcome prevalences is

$$\frac{\bar{p}(S|PCC^*)}{\bar{p}(S|SRS)} = \frac{0.49}{0.23} \approx 2.13.$$

Therefore, comparing PCC^* to SRS, the expected recalibration parameter confidence region is about 2.72 times smaller, indicating an effective sample size savings about three times. In addition, the expected sample outcome prevalence comparing PCC^* to SRS is about 2.13 higher towards outcome class balance. Selecting configurations resulting in increased information as measured by $\phi^D(S)$ and $\phi^B(S)$ ultimately improves the dual modification learning goals of recalibration power and revision support recovery, as we demonstrate through simulation in Section 4.4. Next, we remark on the design impact of PCC designs on modification learning.

4.3.4 Design impact on modification learning

To characterize design impact on modification learning, we focus on the criteria of design validity and design resource efficiency. For a given sampling design, "validity" for modeling refers to whether using resulting samples may provide valid inference, while "resource efficiency" refers to whether using resulting samples can achieve modeling goals at a lower cost compared to a "baseline" design (usually SRS). For the modification learning problem, we describe the design validity of PCC designs and the design resource efficiency compared to SRS.

Design validity

Design validity may be framed as a missing data problem [147], where distributional differences comparing resulting samples to the "complete" cohort determines whether the design is valid. Since PCC designs is a member of the general predictive score sampling class (Definition 1), sampling is based only on original model scores S. Denote Δ as the indicator of a subject being included in a sample. Then, for all designs within the predictive score

sampling class

$$\Delta \perp Y|(S, \mathbf{X}), \tag{4.10}$$

implying $f(Y|S, \mathbf{X}, \Delta) = f(Y|S, \mathbf{X})$, therefore establishing design validity. Due to (4.10), fitting the modification learning model (4.2) using samples drawn with the PCC design results in asymptotically equivalent inference as if the entire cohort were available.

The condition (4.10) is also known as the (outcome) Missing At Random (MAR) property from the missing data literature [78]. Samples that have outcome MAR are intentionally biased from the cohort; mathematically this means that $f(\Delta|S) \neq f(\Delta)$, but any differences may be characterized. For the PCC design, sampling depends on a specific function of the score, which is whether scores exceed threshold k with stratum frequency w. Therefore, the sampling distribution under PCC is essentially a re-weighting of that under SRS, where weights are specified within strata defined by the fixed threshold k (Lemma 1).

Lemma 1 *Equivalence of SRS and strata frequency weighted PCC sampling distributions. For the PCC design based on stratum frequency w for strata defined by scores S exceeding cut-off k, then for fixed $k \in supp(S)$ the sampling weights for subject i are*

$$P_{PCC}(\Delta_i = 1) = \begin{cases} P_{SRS}(\Delta_i = 1) \times \dfrac{w}{P(S_i > k)}, & S_i > k \\ P_{SRS}(\Delta_i = 1) \times \dfrac{1-w}{P(S_i \leq k)}, & S_i \leq k \end{cases}, \tag{4.11}$$

and when $w = P(S_i > k)$,

$$P_{PCC}(\Delta_i = 1) = P_{SRS}(\Delta_i = 1).$$

Design resource efficiency

From Lemma 1, it is clear that PCC design over-represents subjects with higher scores in resulting samples. A natural question then, is if such re-weighting provides higher sample "information". Intuitively, stratified sampling that up-weights the more informative strata results in overall higher information. In particular, for the Binary Entropy information function $\phi^B(S)$, we show a direct correspondence between stratum weights and sample information in Lemma 2. By over-representing subjects with scores $S > k$ more than under the SRS distribution, resulting samples have higher outcome class balance, assuming that outcome probabilities p_i are monotone increasing with scores S_i.

Lemma 2 *PCC design configurations for higher sample Binary Entropy.*
For the PCC design with design configurations (k, w), assume that outcome probabilities p_i are monotone in scores S. Then, for fixed cut-off $k \in supp(S)$, using stratum weights w such that $w > P(S > k)$ results in

$$\phi^B(S|PCC) > \phi^B(S|SRS).$$

It is also possible to show that the D-optimality function may be increased by re-weighting sample scores. However, required PCC design configurations are more complicated than simple fixed cut-offs and higher stratum weights as shown for the Binary Entropy function. As the information functions $\phi^D(S)$ and $\phi^B(S)$ are intermediaries of true modification learning goals, we omit extensive discussion, and instead remark that ordering of the D-optimality information function may be shown through the monotonicity property [102]. To demonstrate that design configuration evaluation and selection using the PCC framework may improve

on the ultimate modification learning goals of recalibration parameter testing power and revision parameter support recovery, we next provide empirical evidence through Monte Carlo simulation.

4.4 Simulations

4.4.1 Data generating mechanism to simulate model modification learning

We generated synthetic data ($n = 100,000$) to demonstrate the benefit of using PCC designs for model modification learning to a new setting. True outcomes Y were generated using model (4.2) across a range of scenarios, through specifying various modification learning parameters α_0, α_1 and $\tilde{\gamma} \in \mathcal{R}^{100}$. Features $\tilde{X}$ and scores S were considered "fixed" across the various data scenarios, generated using the Linear Discriminant Analysis (LDA) assumption. Conditioned on the original outcome prevalence π^0, features were generated as

$$\tilde{X}_i^T | (Y_i^{initial} = y) \sim N(\tilde{\mu}_y, \mathbf{\Sigma}_y),$$

with $\tilde{\mu}_y$ and $\mathbf{\Sigma}_y$ set such that

$$logit(E[Y_i^{initial} | \tilde{X}_i^T]) = \beta_0 + \tilde{X}_i^T \tilde{\beta}$$

$$\tilde{\beta}_j = \begin{cases} 0.7 & j = 1, \ldots, 10 \\ -0.7 & j = 11, \ldots, 20 \\ 0 & j = 21, \ldots, 100. \end{cases}$$

The original scores were defined as $S_i = \beta_0 + \tilde{X}_i^T \tilde{\beta}$, and the true mean model was thus simulated as

$$logit(E[Y_i|S_i, \tilde{X}_i^T]) = \alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma}$$

From each simulated cohort, samples drawn using PCC and SRS were compared based on metrics for recalibration (statistical power) and revision (support recovery). We used the same design, PCC^{sim} with $(k, w) = (-1, 0.50)$ for consistency across the various simulated scenarios.

4.4.2 Design benefit for recalibration

For all recalibration simulations, the power of recalibration tests was compared between using samples drawn with SRS or PCC^{sim}. The initial outcome prevalence was set as $\pi^0 = 0.10$, the true recalibration intercept α_0 was set to be one of $-\log(2)$, $-\log(1.5)$, or 0, the true recalibration slope α_1 was set to be one of 0.6, 0.8, or 1, and the true revision parameters were set to be $\tilde{\gamma} = \tilde{0}$ (no revision). Therefore, the synthetic data simulated scenarios where the outcome is relatively rare, that outcome prevalence is expected to be even lower in the new cohort, and that the model was over-fitted in the source cohort. For each scenario and over a grid of sample sizes, the empirical power was calculated across $B = 500$ simulations. These recalibration parameters were selected based on previous work illustrating the sample size requirements for model external validation studies [131].

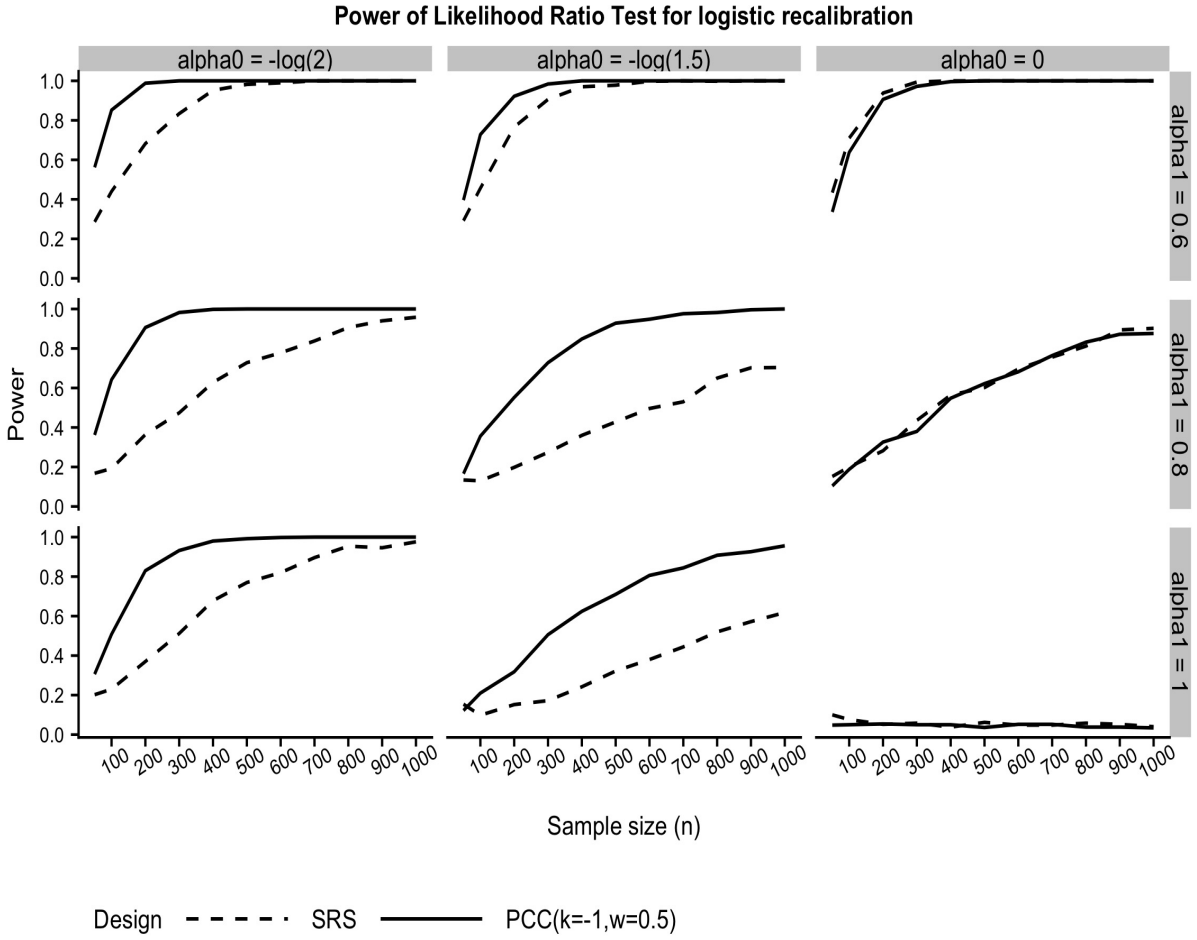
Figure 4.4 illustrates the simulation results for the logistic recalibration test; additional simulation results for testing the recalibration intercept and slope are shown in Appendix C.0.3. Overall, Figure 4.4a shows that using PCC^{sim} provided equivalent or higher power compared to SRS across the range of investigated recalibration parameters. For scenario-specific comparisons, we compare three power curves in Figure 4.4a with the true D-optimality information response surfaces in Figure 4.4b.

First, consider the true data generating mechanism of $\alpha_0 = -\log(2)$ and $\alpha_1 = 0.80$ (left column middle row), where using PCC^{sim} resulted in a higher power curve compared to using SRS. This observation may be explained by the D-optimality contour surfaces, where the joint confidence region for estimating the recalibration parameters was about 1.5 times smaller using PCC^{sim} compared to SRS. For true $\alpha_0 = -\log(1.5)$ and $\alpha_1 = 0.60$ (middle column top row), using PCC^{sim} provided some improvement over SRS. However, the SRS power curve was already sufficiently high, as shown by the overall high D-optimality values across all PCC designs (including the SRS-equivalent configuration). For $\alpha_0 = -\log(1.5)$ and $\alpha_1 = 1$ (middle column bottom row), PCC^{sim} provided some improvement over SRS but both power functions were low, again demonstrated by the overall low D-optimality values for this set of recalibration parameters. Note that when no model recalibration is needed ($\alpha_0 = 0$ and $\alpha_1 = 1$), using both SRS and PCC^{sim} samples provided the correct size.

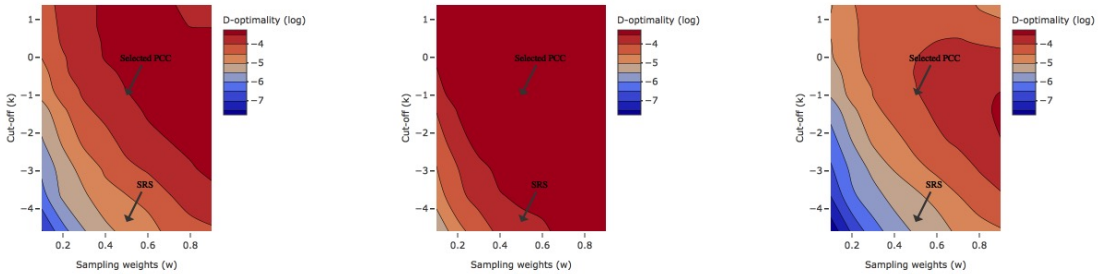
The power functions in Figure 4.4a may be used for sample size calculation and study planning. For example, consider the scenario where true $\alpha_0 = -\log(1.5)$ and $\alpha_1 = 0.80$ (middle column middle row), to achieve 80% power for the logistic recalibration test requires close to $n = 1000$ abstracted samples when using SRS. For an outcome with 10% prevalence this corresponds to an effective sample size of abut 100, comparable to what was found in [131]. However, to achieve this same power using PCC^{sim} only requires the abstraction of $n = 200$ samples, a cost savings of almost five-fold.

Figure 4.4: Simulation results for model recalibration.

(a) Average empirical power of the Likelihood Ratio Test for logistic recalibration, comparing PCC^{sim} with $(k, w) = (-1, 0.50)$ to SRS for a range of sample sizes under various recalibration parameters, over $B = 500$ simulations.



(b) Contour plots of D-optimality (log transformed) information functions of sample scores computed using true modification learning parameters for $(\alpha_0, \alpha_1) = (-\log(2), 0.80)$ (left), $(\alpha_0, \alpha_1) = (-\log(1.5), 0.60)$ (middle), and $(\alpha_0, \alpha_1) = (-\log(1.5), 1)$ (right). Simulation averages were computed based on a sample size of $n = 100$.



4.4.3 Design benefit for revision

For all revision simulations, the average False Discovery Rate (FDR) and False Exclusion Rate (FER) (4.8) in selecting truly predictive features were compared between using SRS and PCC. The original outcome prevalence was set as either $\pi^0 = 0.10$ or $\pi^0 = 0.25$, the true recalibration parameters as $\alpha_0 = -\log(3)$ and $\alpha_1 = 0.90$, and the true revision parameters with effect size $|\gamma_j| = 0.60$ and sparsity ratio (proportion of non-zero revision parameters) of $f = 0.05$. The simulated data compares the effect of using PCC designs on support recovery measures for data with two different outcome prevalences. For each scenario and over a grid of sample sizes, estimated revision parameters $\hat{\gamma}(\lambda)$ were fitted using (4.3) for the solution path defined by $\lambda \in [-\log(8), -\log(2)]$. Resulting estimates were compared against true revision parameters, and empirical FDR and FER calculated across $B = 500$ simulations.

Figure 4.5 illustrates the simulation results. Figure 4.5a illustrates the simulated average 1-False Exclusion Rate (1-FER) versus False Discovery Rate (FDR), where curves towards the top-left corner of the plot indicate better designs. For each sample size (n = 250, 500, 750) and each sampling design (SRS or PCC), the point at $(0, 0)$ indicates support recovery of model coefficients estimated at a fixed high penalty, where both the false positives and true positives are zero since all coefficients are excluded. As penalty decreases, coefficients enter the solution path, increasing both false positives and true positives in estimating true model coefficients. Finally, when all coefficients are included in the model both false positives and true positives tend to 1.

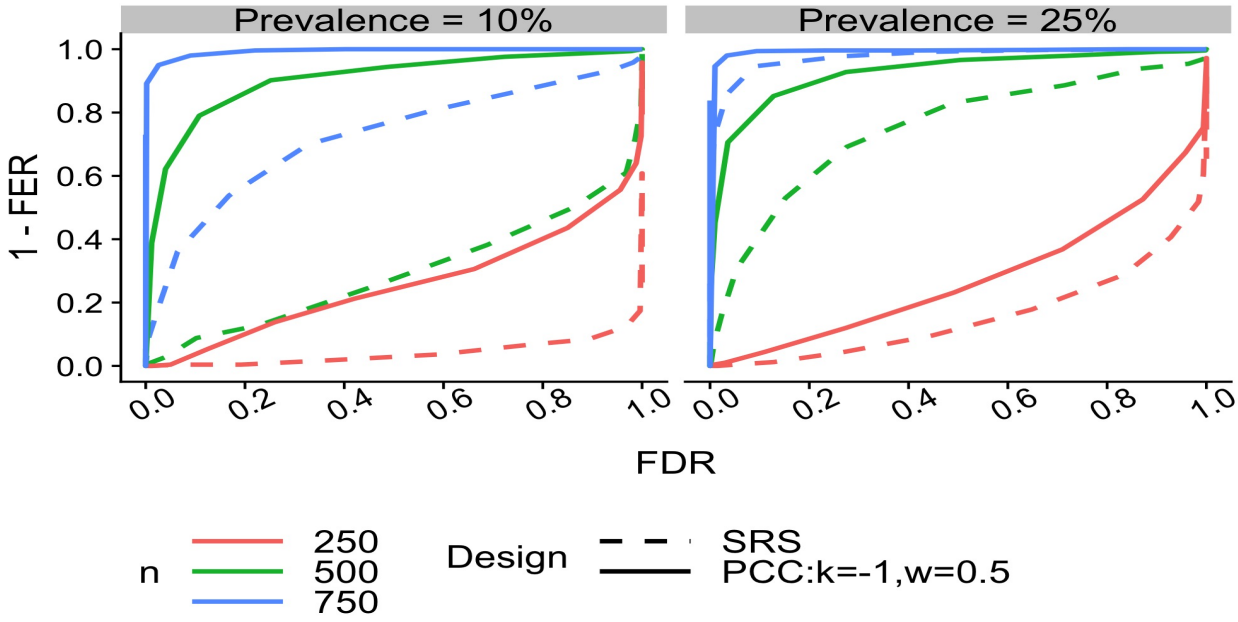
For each sample size, we may compare resulting support recovery curves between PCC^{sim} and SRS. For an outcome prevalence of 10%, when $n = 250$, using SRS results in models that exclude almost all predictive features, unless all features were selected, but PCC^{sim} may allow recovery of about 20% truly predictive features at FDR=0.40. Support recovery improved with sample size, where the improvement was faster for PCC^{sim}: at $n = 500$ for

FDR=0.20 using PCC^{sim} recovers 80% while SRS only recovers 10% of truly predictive features, and at $n = 750$ using PCC^{sim} results in almost perfect support recovery, but using SRS may still result in false negatives and false positives. Similar patterns are observed for the 25% outcome prevalence scenario, where for every sample size support recovery curves using PCC^{sim} were higher compared to SRS, although separation of PCC^{sim}/SRS support recovery curves were smaller compared to the 10% prevalence scenario.

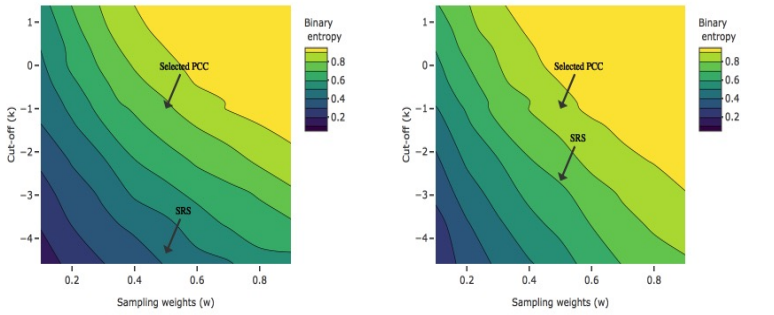
To explain this observation, we turn to the sample Binary Entropy as illustrated in Figure 4.5b. For the 10% outcome prevalence, sample Binary Entropy values for SRS and PCC^{sim} were 0.43 (sample outcome prevalence = 9%) and 0.80 (sample outcome prevalence = 25%) respectively. For the 25% outcome prevalence, sample Binary Entropy values for SRS and PCC^{sim} were 0.77 (sample outcome prevalence = 23%) and 0.85 (sample outcome prevalence = 28%) respectively. Thus, while using the specified PCC^{sim} design more than doubled sample outcome prevalence for the 10% prevalence scenario, it only slightly increased outcome class balance for the 25% prevalence scenario. Such differences in sample Binary Entropy values corroborated the support recovery curves illustrated in Figure 4.5a. Therefore, through selecting samples targeted towards higher Binary Entropy, PCC^{sim} designs improved support recovery of predictive features - such effect was more pronounced for rarer outcomes.

Figure 4.5: Simulation results for model revision.

(a) Empirical average False Exclusion Rate (FER) versus False Discovery Rate (FDR) comparing PCC^{sim} with $(k, w) = (-1, 0.50)$ to SRS for a range of sample sizes, over $B = 500$ simulations.



(b) Contour plots of Binary Entropy information functions of sample scores computed using true modification learning parameters for outcome prevalence of $\pi^0 = 0.10$ (left) and $\pi^0 = 0.25$ (right). Simulation averages were computed based on a sample size of $n = 300$.



4.4.4 Design local robustness

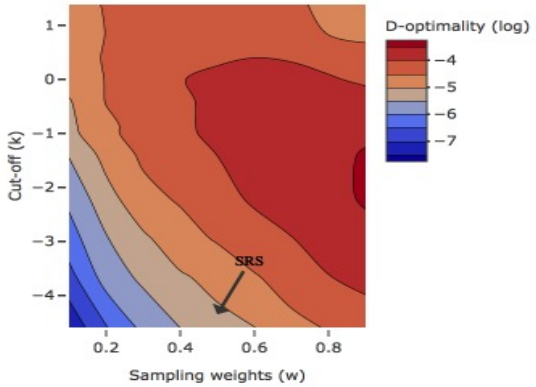
Our simulations thus far demonstrated the benefit of using PCC^{sim} for model recalibration and revision, where design configurations were evaluated conditioned on true modification learning parameters. Here, we investigate by simulation whether the local robustness to parameter mis-specification but assuming model (4.2).

Figure 4.6 shows sample D-optimality contour surfaces, where configurations with warmest colors (red) provide highest sample D-optimality values. Note that warmest color regions coincide regardless of whether computations were based on the null assumption of $\alpha_0 = 0$, $\alpha_1 = 1$, $\tilde{\gamma} = \tilde{0}$ (Figure 4.6a) or under the actual recalibration parameters (Figure 4.6b). When true (α_0, α_1) deviates from $(0, 1)$, systematic shifting of D-optimality values were observed, where contour surfaces were shifted upwards when true $\alpha_0 < 0$, rotated counter-clockwise when true $\alpha_1 \in [0, 1]$, and rotated clockwise when true $\alpha_1 > 1$. However, conclusions based on sample score information computed under any considered assumptions are comparable, where for the illustration in Figure 4.6 designs to the right of the contour plots indicate high information designs. Figure 4.7 shows a similar story for sample scores summarized with the Binary Entropy information function, where contour plots were systematic shifted, rotated, and/or scaled but overall conclusions are locally robust to slight parameter mis-specifications.

However, in contrast to mis-specification to parameter mis-specification, score distributions may affect design evaluation conclusions. Figure 4.8 illustrates the effect of using different score distributions (Figure 4.8a) on sample scores summarized with the D-optimality (Figure 4.8b) and Binary Entropy (Figure 4.8c) information functions. Depending on cohort score distribution, high information designs as evaluated by D-optimality and Binary Entropy differ. However, such conclusions are less concerning, as true score distributions are assumed to be known for all subjects in the new cohort.

Figure 4.6: Contour plots of sample scores summarized with the D-optimality (log) information function, based on data generated based on model (4.2) using Linear Discriminant Analysis (LDA) features with $\pi^0 = 0.10$.

(a) D-optimality contour surfaces computed using the assumption $\alpha_0 = 0$, $\alpha_1 = 1$, $\tilde{\gamma} = \tilde{0}$.



(b) D-optimality contour surfaces computed using true $(\alpha_0, \alpha_1) = (-\log(2), 1)$ (left); $(\alpha_0, \alpha_1) = (0, 0.8)$ (middle); $(\alpha_0, \alpha_1) = (0, 1.25)$ (right). The true revision parameters were $\tilde{\gamma} = \tilde{0}$ for all scenarios.

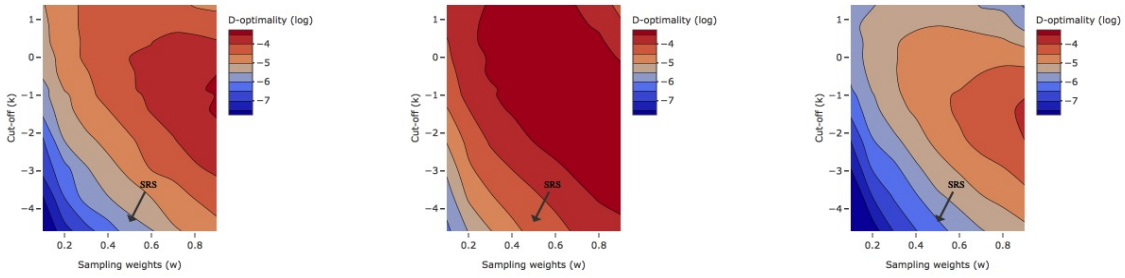
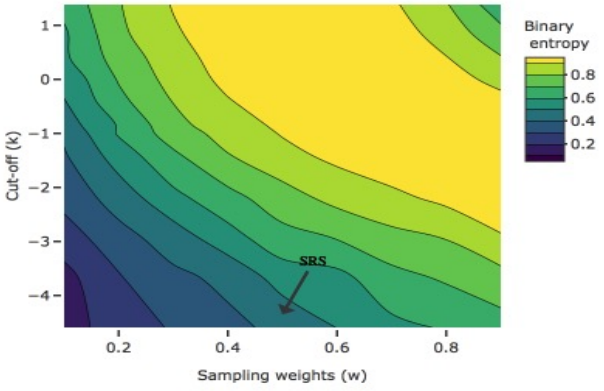


Figure 4.7: Contour plots of sample scores summarized with the Binary Entropy information function, based on data generated based on model (4.2) using Linear Discriminant Analysis (LDA) features with $\pi^0 = 0.10$.

(a) Binary Entropy contour surfaces computed using the assumption $\alpha_0 = 0$, $\alpha_1 = 1$, $\tilde{\gamma} = \tilde{0}$.



(b) Binary Entropy contour surfaces computed using using true $(\alpha_0, \alpha_1) = (0, 1)$, $|\gamma_j| = 0.60$ (left); $(\alpha_0, \alpha_1) = (-\log(3), 0.90)$, $\tilde{\gamma}$, $|\gamma_j| = 0.60$ (middle); $(\alpha_0, \alpha_1) = (0, 1)$, $|\gamma_j| = 1.50$ (right). f=5% of the revision parameters are non-zero with effect size $|\gamma_j|$; remaining revision parameters are non-predictive.

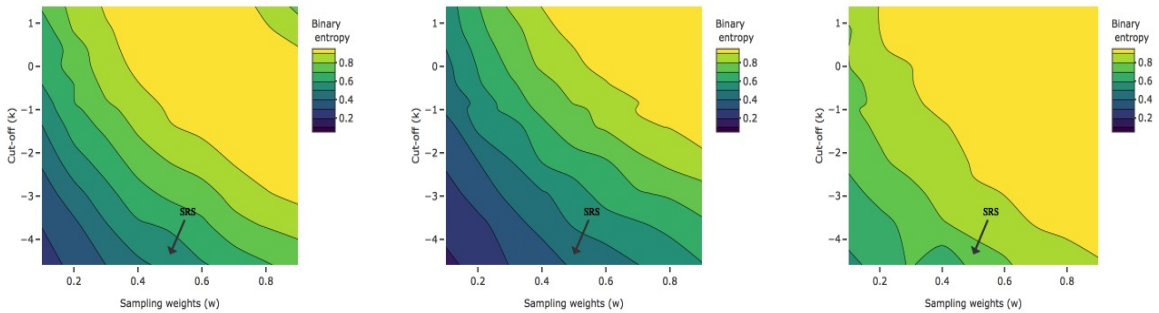
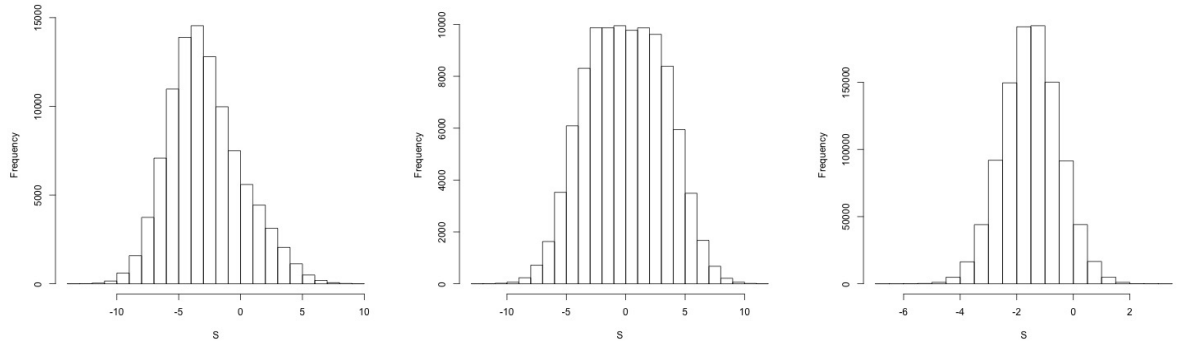
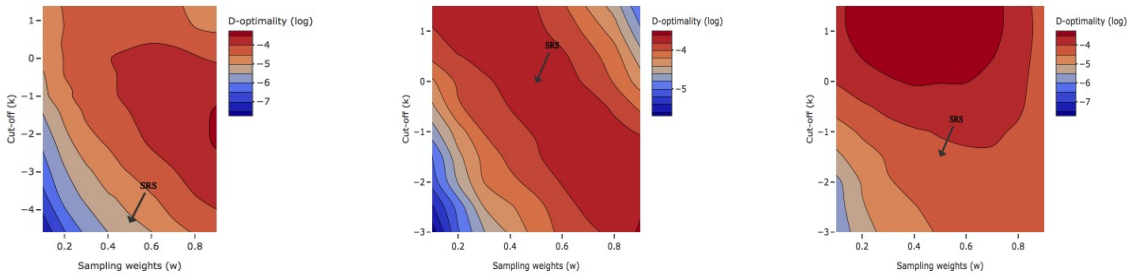


Figure 4.8: Effect of score distribution on D-optimality and Binary Entropy information functions of sample scores. Data generating mechanisms are: scores based on LDA features with $\pi^0 = 0.10$ (left); LDA features with $\pi^0 = 0.50$ (middle); $S \sim N(1.5, 1^2)$ (right). All contour surfaces are computed based on the assumption $\alpha_0 = 0$, $\alpha_1 = 1$, $\tilde{\gamma} = \tilde{0}$. LDA: Linear Discriminant Analysis; π^0: the original outcome prevalence assumed in the source cohort.

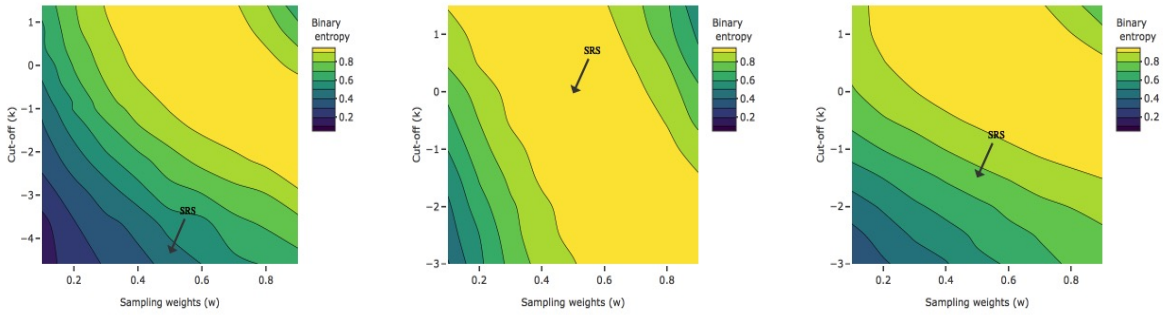
(a) Score distributions in the simulated cohort.



(b) D-optimality (log) contour surfaces.



(c) Binary Entropy contour surfaces.



4.5 Illustration: Modification learning on radiology report modality

4.5.1 Radiology reports and modification learning across imaging modalities

To illustrate the proposed PCC design on reducing sample size requirements for modification learning, we consider radiology reports arising from different imaging modalities. Radiology reports constitute the formal communication of imaging study results by trained radiologists, contain important information about radiographic findings, but often presented as unstructured data via free-text. Therefore, the collection of radiographic findings outcome labels require abstraction by highly-trained human clinical experts, which is a labor intensive and costly process. Such costs motivate using modification learning modeling and resource efficient sampling approaches for data collection.

The application data set comes from radiology reports derived from the Lumbar Imaging with Reporting of Epidemiology (LIRE) study [65]. The LIRE study was a randomized pragmatic clinical trial that studied the effect of radiology report content on subsequent treatment decisions. Adult subjects were considered for study inclusion if their primary care provider (PCP) ordered either an x-ray or Magnetic Resonance (MR) imaging test of the lumbar spine. A finding of interest was vertebral fracture, which is visible on both x-ray and MR imaging modalities. To modify a "source" model for fracture previously developed using x-ray reports for application to MR reports, a sample of "new" outcome labels needs to be abstracted from MR reports. Towards the resource efficient collection of such a sample, we illustrate the benefit of using PCC designs compared to SRS, as evaluated based on the dual modification learning goals of model recalibration and model revision.

4.5.2 Illustration set-up

Our approach for illustration is through simulation on the full data, defining true source and modification learning models based on all available data, and demonstrating design effects through sub-sampling. Our illustrations require text processing and classification modeling,

where we used existing routines from the `quanteda` and `glmnet` packages in `R`, respectively. To reduce any potential between-site variation, we restricted analysis to the single largest site from the LIRE study. For all subjects from this single site, we obtained a large number of simulated abstracted outcomes using a rule-based natural language processing algorithm [122]. A total of $N^0 = 158,405$ labels (prevalence = 10%) were obtained from source x-ray reports, and $N^1 = 41418$ labels (prevalence = 7%) were possible from MR reports. Full details of the source and target model definition procedures are in Appendix C.0.4; here we provide an overview of the feature engineering, source model definition, and target model definition processes:

Feature engineering: Features $\tilde{X}_i^T$ were created for all x-ray and MR reports and included the following for a total of $p = 131$ features:

- Text features: Bag-of-words stemmed unigrams indicators, excluding rare (occurring in $< 5\%$ of documents) and common (occurring in $> 80\%$ of documents) terms.

- Demographic features: Gender (male or female) and age category (< 40, 40-60, > 60 years).

Source model definition: The source classification model was estimated using the Logistic Lasso procedure with 10-fold cross-validation based on an empirical AUC loss function. Model development was based on a random sample of $n^1 = 5000$ drawn from N^1, where 23 non-zero coefficients were selected.

Modification learning model definition: To define the modification learning model, first scores were computed by applying the source model to features from the MR cohort. Then, modification learning parameters were estimated as follows:

- Recalibration parameters α_0 and α_1: Estimated by fitting the modification learning model (4.2) conditioned on no revision ($\tilde{\gamma} = \tilde{0}$).

- Revision parameters $\tilde{\gamma}$: Estimated by fitting modification learning model (4.2) with the Lasso procedure (4.3) to penalize features but not scores.

This results in a "true" modification learning model with parameters

$$logit(E[Y|S, \mathbf{X}]) = -1.03 + 0.89S + \mathbf{X}\tilde{\gamma}$$

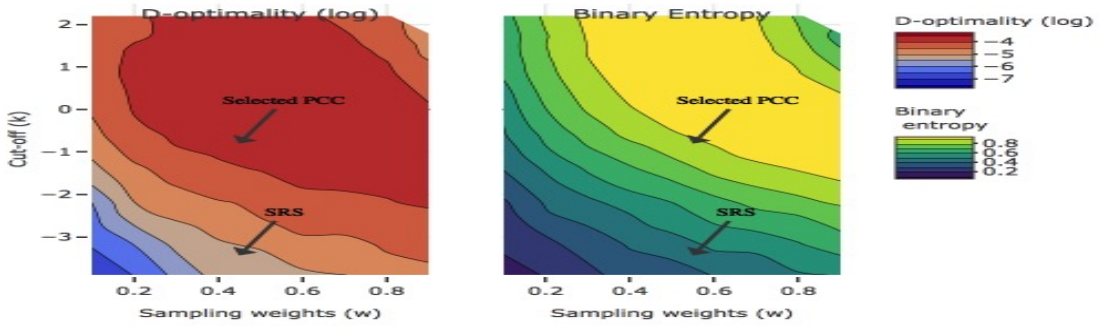
where 6 features in $\tilde{\gamma}$ were estimated to be non-zero: stemmed unigrams sublux = 1.42; deform = 1.15; fractur = 0.78; scoliosi = 0.78; normal = -0.53; desicc = -0.57.

4.5.3 Evaluating and selecting a PCC design configuration

Using the proposed computational framework (Algorithm 1 in Section 4.3.3), we evaluate and select, score cut-off k and stratum weights w that increase expected sample information within PCC class design configurations. Using original scores computed for MR reports based on the source x-ray model, we calculated resulting D-optimality (4.5) and Binary Entropy (4.9) information functions with Monte Carlo.

The pairs of contour plots of information functions are shown in Figure 4.9. As illustrated, there were a range of PCC design configurations that may provide improved information for model modification learning compared to using SRS. However, for this practical illustration we were additionally constrained by the total available sample size. For example, while using the configuration $(k, w) = (2, 0.50)$ may be ideal based purely on values of information functions, there were only 105 subjects with scores exceeding $k = 2$ resulting in a maximum sample size of 202. Alternatively, as there are 461 subjects with scores exceeding $k = 0.80$, the configuration of $(k, w) = (0.80, 0.50)$ is potentially more reasonable as it allows up to a maximum of 5526 subjects to be selected. Therefore, considering both potential design enrichment as well as practical constraints, we selected PCC^{illus} with $(k, w) = (0.80, 0.50)$ to illustrate the benefit of using PCC over SRS through sub-sampling.

Figure 4.9: Pairs of contour plots for D-optimality and Binary Entropy information functions of sample scores for LIRE data application example.



4.5.4 Benefit of PCC for modification learning

The selected PCC^{illus} was then compared to SRS through sub-sampling, with statistical power and support recovery as evaluation metrics for recalibration and revision, respectively. Over a grid of sample sizes, sub-samples of MR reports were drawn with either SRS or PCC^{illus}. Recalibration comparisons were based on the empirical power of likelihood ratio tests for the recalibration intercept, slope, and logistic recalibration computed over the sub-samples, fitting the modification learning model (4.2) conditioned on no revision ($\tilde{\gamma} = \tilde{0}$). Since the true recalibration parameters were estimated to be $\alpha_0 = -1.03$ and $\alpha_1 = 0.89$ we expected empirical power to approach 1 as sample size increases, but the rate of power increase may differ depending on sampling design.

Revision comparisons were based on empirical support recovery curves. However, as some estimated revision parameters had small coefficients, we used alternative definitions for FDR and FER, where

$$
\begin{aligned}
FDR^{alt} &= 1 - P(\hat{\boldsymbol{\gamma}}_{\boldsymbol{S}} \in \boldsymbol{\gamma}^*_{\boldsymbol{H}} \cup \boldsymbol{\gamma}^*_{\boldsymbol{L}}) \\
FER^{alt} &= 1 - P(\hat{\boldsymbol{\gamma}}_{\boldsymbol{S^c}} \in \boldsymbol{\gamma}^*_{\boldsymbol{L}} \cup \boldsymbol{\gamma}^*_{\boldsymbol{S^c}}).
\end{aligned}
\tag{4.12}
$$

In (4.12), estimated coefficients from the sub-samples $\hat{\boldsymbol{\gamma}}_{\boldsymbol{S}}$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{S^c}}$ were defined similarly as with convention. However, true coefficients estimated from the full cohort were divided into high, low, and no signal categories $\boldsymbol{\gamma}^*_{\boldsymbol{H}}$, $\boldsymbol{\gamma}^*_{\boldsymbol{L}}$ and $\boldsymbol{\gamma}^*_{\boldsymbol{S^c}}$, with

$$
\begin{aligned}
\boldsymbol{\gamma}^*_{\boldsymbol{H}} &= \{\gamma^*_j : |\gamma^*_j| > 0.50\} \\
\boldsymbol{\gamma}^*_{\boldsymbol{L}} &= \{\gamma^*_j : 0 < |\gamma^*_j| \leq 0.50\} \\
\boldsymbol{\gamma}^*_{\boldsymbol{S^c}} &= \{\gamma^*_j : |\gamma^*_j| = 0\}.
\end{aligned}
\tag{4.13}
$$

In (4.13), FDR^{alt} is defined similar to the usual FDR: any feature that was selected but had a truly non-zero coefficient was considered a false positive. However using FER^{alt} allows features with "low signal"($\boldsymbol{\gamma}^*_{\boldsymbol{L}}$) to be excluded yet not count as a false negative, therefore avoiding penalizing the non-selection of low signal features which could result in artificially high FER. The threshold 0.50 in (4.13) was chosen to keep the feature sparsity ratio around $f = 5\%$.

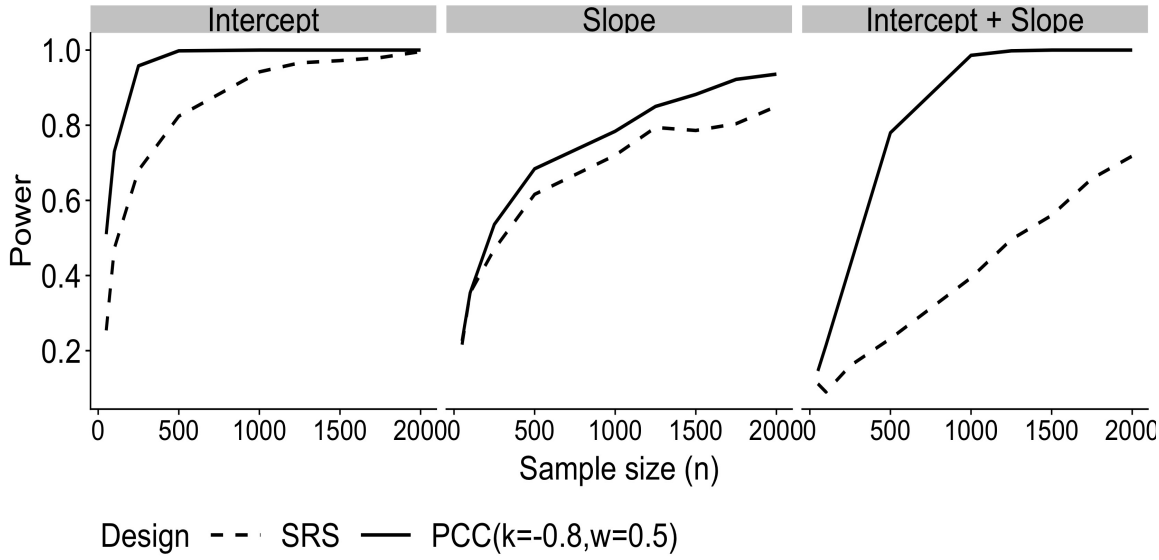
Figure 4.10a shows the design effect on recalibration statistical power. As illustrated, the power functions for all three recalibration tests were higher under PCC^{illus} compared to under SRS. For example, to detect mis-calibration of average predictions (recalibration intercept test) with a 80% power controlling test size at 0.05, using SRS requires $n = 500$ but using PCC^{illus} only requires $n = 250$. Therefore, for this specific external validation study, using PCC^{illus} was about twice as cost efficient compared to using SRS.

Figure 4.10b shows the design effect on revision support recovery. For the three illustrated

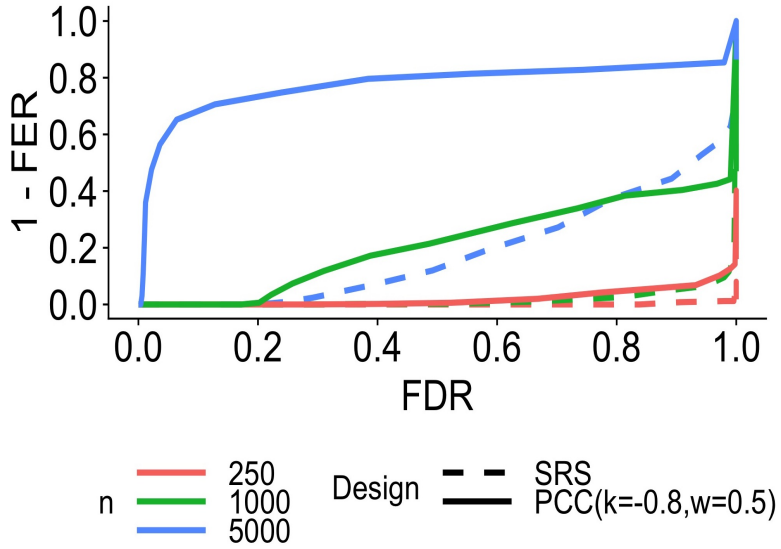
sample sizes, using PCC^{illus} resulted in better support recovery compared to using SRS. For a sample size of $n = 250$, support recovery was low regardless of sampling design. For $n = 1000$, while using SRS did not provide any meaningful support recovery, using PCC^{illus} may recover 30% of truly predictive features for FDR=0.50. For $n = 5000$, using PCC^{illus} results in the recovery of 70% truly predictive features for FDR=0.10, however when using SRS both FDR and FER remained very high. These illustrations demonstrate that, even when the abstraction budget can be as high as $n = 5000$, data collection using SRS could still lead to inaccurate model revision as measured by support recovery.

Figure 4.10: Results from data example illustration.

(a) Empirical power of Likelihood Ratio Tests comparing PCC^{illus} to SRS over $B = 500$ simulations, where true recalibration and slope were respectively $\alpha_0|(\tilde{\gamma} = \tilde{0}) = -1.03$ and $\alpha_1|(\tilde{\gamma} = \tilde{0}) = 0.89$.



(b) Empirical support recovery curves comparing PCC^{illus} to SRS over $B = 500$ simulations, using the alternative definitions of False Discovery Rate (FDR^{alt}) and False Exclusion Rate (FER^{alt}). The revision parameter vector length was $p = 131$, of which 6 features were truly non-zero.



4.6 Discussion

We demonstrated that using design-based principles for new outcome label collection can substantially reduce the sample size requirement for model modification learning. The proposed class of designs is based on sampling using original model predicted scores, which we showed to be amenable for valid analysis. By stratified sampling on strata defined by score values and over-representing subjects with "high information" scores, the resulting Predictive Case Control (PCC) design is additionally resource efficient for model recalibration and model revision. We developed a computational framework to visualize and compare design configurations within the PCC class.

The proposed PCC sampling design is easy to communicate and implement in practice. Intuitively, over-representing subjects with higher scores can be interpreted as over-including "likely cases", establishing an equivalence of the proposed method and the well-known case-control study design. Implementing the stratified PCC design is straightforward, as it only depends on two design configurations: score cut-off k and stratum weights w. Design configuration selection is based on information functions of the scores which are directly related to the dual modification learning goals. Our proposed framework worked well for scenarios even with relatively small effect sizes for recalibration tests ($\alpha_0 \approx -\log(1.5)$, $\alpha_1 \approx 0.80$) and relatively rare outcomes (5% - 10%).

The simple alternative to PCC is to use SRS, which while results in a representative sample for external validation, may not provide sufficient effective sample sizes. More sophisticated alternatives include active and transfer learning procedures. However, since such machine-learning procedures are specifically developed towards resource efficient learning, resulting labeled samples obtained through such procedures may not be amenable for valid analysis. We remark that it is possible to view the described modification learning model as a special formulation of transfer learning, and the proposed PCC design as a pre-specified sampling

design for that transfer learning task.

There are a few limitations of the proposed PCC design. First, we assumed that initial scores are easily computed in the new setting though applying the developed model to the "new" features. However, in practice it is possible that the features used for the original model development are missing in the new setting due to reasons such as incomplete data capture in the EMR. The extent of the impact of missing features on our proposed method is yet unexplored. Our simulation results showed that targeting samples for D-optimality and Binary Entropy improves the true model updating goals of recalibration and revision. While the relationship between D-optimality and statistical power is known in the statistical literature [132], theoretical relationships between Binary Entropy and support recovery are relatively unexplored.

Future work directions include the translation of PCC for use in study design planning, for example sample size calculation through user-supplied parameter values. Towards that end, we plan on integrating simulation calculations and graphical displays for information response surfaces, power functions and support recovery curves into an interactive `RShiny` application: this process is under development at time of this writing. In addition, it may be interesting to explore using alternative information functions for PCC design configuration selection on the resource efficiency of modification learning. For example, the D-optimality information function may be replaced by other "true" information functions such as A-optimality or G-optimality that are more focused on average and prediction variance reduction, and Binary Entropy information function by other measures of outcome class balance. Ultimately, we hope to inspire design-based thinking for "new" outcome label collection, especially when using SRS dictates insurmountable sample size requirements.

Chapter 5

FUTURE WORK

In this dissertation, we had described sampling frameworks for outcome label abstraction and subsequent machine-learning of rare outcomes. We demonstrated sampling design frameworks for labeled data collection and classification of univariate binary outcomes (Chapter 2), multi-labeled outcomes (Chapter 3), as well as modification learning in a new setting (Chapter 4). The general theme of our proposed methods is sampling based on strata defined by "surrogates", which are assumed to be cheaper to obtained than true clinical outcomes yet highly informative of outcome statuses, for example related summaries of structured data elements or original model predicted risk scores. For each of our proposed designs, we have demonstrated both design resource efficiency for the relevant modeling applications as well as formally characterized design impact on modeling. We note that since the proposed designs are stratified sampling procedures on a cheaper variable to select subjects for a more expensive variable, they may be viewed as special cases of two phase sampling [89].

The immediate next step is to implement the described sampling frameworks towards formal study design planning for machine-learning of clinical outcome identification. For example, certain questions of interest include sample size calculations and "optimal" design specifications for this task. Towards this goal, we plan on packaging empirical sample size calculations in the form of learning curves, power functions, and support recovery curves into an interactive `RShiny` application, which is work-in-progress at the time of this writing. In addition, practical scenarios may require combining multiple surrogates for sampling, for example two existing but potentially outdated phenotype definitions, or case type heterogeneity arising from different structured data elements (e.g. different hospitals use a different set of codes).

One option for combining designs may be to use a multi-label sampling framework even for univariate outcomes, where factors affecting design benefit on machine-learning warrants further investigation.

In terms of methodology development, future work include designs based on richer classes of surrogates, designs for non-binary outcomes, and methods to synthesize data from sample and cohort to improve machine-learning of clinical outcomes. For example, quantitative surrogates may arise when the most informative variable for sampling is continuous, for example relevant biomarker values. In such scenarios, it would be interesting to study sampling methods based on quantitative variables. Second, even though the proposed sampling designs are described for binary outcomes, such designs may be extended for the resource efficient data collection for machine-learning of other outcome types such as longitudinal and survival data. Here, an important first step is to identify meaningful and easily calculated measures of information that can help define sampling designs, similar to sample prevalences in the binary setting. Finally, abstracted true clinical outcomes may be combined with mis-classified surrogate outcomes to obtain a larger labeled data set for machine-learning. Intuitively, our pre-specified designs may allow statistical characterization of important information required for data synthesis, such as relationships between surrogates and outcomes; formalization of relevant models requires future work.

Lastly, even though we described applications to radiology reports, our method can be applied to broad machine-learning and Natural Language Processing (NLP) tasks arising from EMR data. For example, towards monitoring of drug safety, a priority for the Federal Drug and Administration (FDA) Sentinel is the accurate identification of relevant Health Outcome of Interests (HOI). Many HOI, such as anaphylaxis (a serious allergic reaction that is rapid in onset and may cause death) [133], even though considered medically "common" adverse effects are statistically "rare". Traditional clinical outcome definitions based on querying structured data (e.g. ICD codes) alone is thought to be potentially inaccurate, but the ex-

tent of mis-classification is often unknown. Alternative, a promising accurate and scalable approach is to use machine-learning classification for such clinical outcome identification. When selected subjects for true clinical outcome abstraction, ideally the collected sample may be appropriately used for both goals: the evaluation of the traditional ICD-only definitions, and serve as reference-standard labels for contemporary machine-learning model development and validation. Our proposed designs are ideally suited for these tasks, where exact design implementations will be project specific.

Ultimately, we have demonstrated the potential of using design-based sampling methods towards resource efficient outcome label abstraction and machine-learning modeling of rare outcomes. With the formal framework established in this dissertation, we hope to guide future development of relevant sampling designs that are statistically valid for analysis, resource efficient for machine-learning, and easily communicable in practice.

BIBLIOGRAPHY

[1] 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.

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

[3] Todd A Alonzo and Margaret Sullivan Pepe. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):173–190, 2005.

[4] AC Atkinson. Developments in the design of experiments, correspondent paper. *International Statistical Review/Revue Internationale de Statistique*, pages 161–177, 1982.

[5] Anthony C Atkinson and David C Woods. Designs for generalized linear models. *Handbook of Design and Analysis of Experiments*, pages 471–514, 2015.

[6] Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415, 1975.

[7] 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.

[8] Colin B Begg and Robert A Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, pages 207–215, 1983.

[9] Yves G Berger. Rate of convergence for asymptotic variance of the horvitz–thompson estimator. *Journal of Statistical Planning and Inference*, 74(1):149–168, 1998.

[10] Flavia Cristina Bernardini, Rodrigo Barbosa da Silva, Rodrigo Magalhaes Rodovalho, and Edwin Benito Mitacc Meza. Cardinality and density measures and their influence to multi-label learning methods. *Dens*, 1:1, 2014.

[11] Elena Birman-Deych, Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, and Brian F Gage. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors. *Medical care*, 43(5):480–485, 2005.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[13] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[14] Norman E Breslow and Nilanjan Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999.

[15] Norman E Breslow, Thomas Lumley, Christie M Ballantyne, Lloyd E Chambless, and Michal Kulich. Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49, 2009.

[16] David S Carrell, Scott Halgrim, Diem-Thy Tran, Diana SM Buist, Jessica Chubak, Wendy W Chapman, and Guergana Savova. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *American journal of epidemiology*, 179(6):749–758, 2014.

[17] Robert J Carroll, Anne E Eyler, and Joshua C Denny. Naïve electronic health record phenotype identification for rheumatoid arthritis. In *AMIA annual symposium proceedings*, volume 2011, page 189. American Medical Informatics Association, 2011.

[18] 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.

[19] Abhishek Chakrabortty, Matey Neykov, Raymond Carroll, and Tianxi Cai. Surrogate aided unsupervised recovery of sparse signals in single index models for binary outcomes. *arXiv preprint arXiv:1701.05230*, 2017.

[20] 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.

[21] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. Resampling multilabel datasets by decoupling highly imbalanced labels. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 489–501. Springer, 2015.

[22] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015.

[23] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Mlsmote: approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015.

[24] 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.

[25] Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):e253–e259, 2013.

[26] Hugh A Chipman and William J Welch. D-optimal design for generalized linear models. *Unpublished*, 1996.

[27] 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.

[28] Amanda Clare and Ross King. Knowledge discovery in multi-label phenotype data. *Principles of data mining and knowledge discovery*, pages 42–53, 2001.

[29] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.

[30] Mike Conway, Richard L Berg, David Carrell, Joshua C Denny, Abel N Kho, Iftikhar J Kullo, James G Linneman, Jennifer A Pacheco, Peggy Peissig, Luke Rasmussen, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. In *AMIA annual symposium proceedings*, volume 2011, page 274. American Medical Informatics Association, 2011.

[31] David R Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565, 1958.

[32] David R Cox. The analysis of multivariate binary data. *Applied statistics*, pages 113–120, 1972.

[33] Bin Dai, Shilin Ding, Grace Wahba, et al. Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483, 2013.

[34] Jarrod E Dalton. Flexible recalibration of binary clinical prediction models. *Statistics in medicine*, 32(2):282–289, 2013.

[35] André de Carvalho and Alex Freitas. A tutorial on multi-label classification techniques. *Foundations of Computational Intelligence Volume 5*, pages 177–195, 2009.

[36] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.

[37] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[38] Holger Dette. Designing experiments with respect to standardizedoptimality criteria. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):97–110, 1997.

[39] Holger Dette, Frank Bretz, Andrey Pepelyshev, and Jose Pinheiro. Optimal designs for dose-finding studies. *Journal of the American Statistical Association*, 103(483):1225–1237, 2008.

[40] Richard A Deyo, Daniel C Cherkin, and Marcia A Ciol. Adapting a clinical comorbidity index for use with icd-9-cm administrative databases. *Journal of clinical epidemiology*, 45(6):613–619, 1992.

[41] Rui M Duarte and Alexander R Vaccaro. Spinal infection: state of the art and management algorithm. *European Spine Journal*, 22(12):2787–2799, 2013.

[42] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001.

[43] L Eriksson, E Johansson, N Kettaneh-Wold, C Wikström, and S Wold. Design of experiments. *Principles and Applications, Learn ways AB, Stockholm*, 2000.

[44] 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.

[45] Ming Fang, Yuqi Xiao, Chongjun Wang, and Junyuan Xie. Multi-label classification: Dealing with imbalance by combining labels. In *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, pages 233–237. IEEE, 2014.

[46] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[47] Jochen Garcke and Thomas Vanck. Importance weighted inductive transfer learning for regression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 466–481. Springer, 2014.

[48] Andrés Felipe Giraldo-Forero, Jorge Alberto Jaramillo-Garzón, José Francisco Ruiz-Muñoz, and César Germán Castellanos-Domínguez. Managing imbalanced data sets in multi-label problems: a case study with the smote algorithm. In *Iberoamerican Congress on Pattern Recognition*, pages 334–342. Springer, 2013.

[49] Ira Goldstein, Anna Arzumtsyan, and Özlem Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2007, page 279. American Medical Informatics Association, 2007.

[50] Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, pages 1491–1523, 1964.

[51] Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. Using anchors to estimate clinical state without labeled data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 606. American Medical Informatics Association, 2014.

[52] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

[53] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[54] John G Hanly, Kara Thompson, and Chris Skedgel. The use of administrative health care databases to identify patients with rheumatoid arthritis. *Open access rheumatology: research and reviews*, 7:69, 2015.

[55] 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.

[56] Radoslav Harman. Multiplicative methods for computing d-optimal stratified designs of experiments. *Journal of Statistical Planning and Inference*, 146:82–94, 2014.

[57] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[58] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328. IEEE, 2008.

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

[60] Hua He, Jeffrey M Lyness, and Michael P McDermott. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in medicine*, 28(3):361–376, 2009.

[61] Nicholas Henschke, Christopher G Maher, RW Ostelo, HC De Vet, Petra Macaskill, Les Irwig, et al. Red flags to screen for malignancy in patients with low-back pain. *Cochrane Database Syst Rev*, 2(2), 2013.

[62] Tim Holland-Letz and Annette Kopp-Schneider. Optimal experimental designs for dose–response studies with continuous endpoints. *Archives of toxicology*, 89(11):2059–2068, 2015.

[63] 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.

[64] David W Hosmer, Trina Hosmer, Saskia Le Cessie, Stanley Lemeshow, et al. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, 16(9):965–980, 1997.

[65] 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.

[66] Jeffrey G Jarvik, Laura S Gold, Katherine Tan, Janna L Friedly, Srdjan S Nedeljkovic, Bryan A Comstock, Richard A Deyo, Judith A Turner, Brian W Bresnahan, Sean D Rundell, et al. Long-term outcomes of a large, prospective observational cohort of older adults with back pain. *The Spine Journal*, 2018.

[67] Elizabeth FO Kern, Miriam Maney, Donald R Miller, Chin-Lin Tseng, Anjali Tiwari, Mangala Rajan, David Aron, and Leonard Pogach. Failure of icd-9-cm codes to identify patients with comorbid chronic kidney disease in diabetes. *Health services research*, 41(2):564–580, 2006.

[68] André I Khuri, Bhramar Mukherjee, Bikas K Sinha, and Malay Ghosh. Design issues for generalized linear models: A review. *Statistical Science*, pages 376–399, 2006.

[69] Jack Kiefer and Jacob Wolfowitz. Optimum designs in regression problems. *The Annals of Mathematical Statistics*, pages 271–294, 1959.

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

[71] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

[72] Paras Lakhani and Curtis P Langlotz. Automated detection of radiology reports that document non-routine communication of critical or significant results. *Journal of digital imaging*, 23(6):647–657, 2010.

[73] Tony Lancaster and Guido Imbens. Case-control studies with contaminated controls. *Journal of Econometrics*, 71(1):145–160, 1996.

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

[75] Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigam H Shah. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6):547–555, 2013.

[76] Xi Li, Ge Zhao, Jian Zhang, Zhiquan Duan, and Shijie Xin. Prevalence and trends of the abdominal aortic aneurysms epidemic in general population-a meta-analysis. *PLoS One*, 8(12):e81260, 2013.

[77] Mary J Lindstrom and Douglas M Bates. Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

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

[79] Mei Liu, Eugenia Renne McPeek Hinz, Michael Edwin Matheny, Joshua C Denny, Jonathan Scott Schildcrout, Randolph A Miller, and Hua Xu. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association*, 20(3):420–426, 2012.

[80] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

[81] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[82] Catherine A McCarty, Rex L Chisholm, Christopher G Chute, Iftikhar J Kullo, Gail P Jarvik, Eric B Larson, Rongling Li, Daniel R Masys, Marylyn D Ritchie, Dan M Roden, et al. The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*, 4(1):13, 2011.

[83] Andrew McDavid, Paul K Crane, Katherine M Newton, David R Crosslin, Wayne McCormick, Noah Weston, Kelly Ehrlich, Eugene Hart, Robert Harrison, Walter A Kukull, et al. Enhancing the power of genetic association studies through the use of silver standard cases derived from electronic medical records. *PloS one*, 8(6):e63481, 2013.

[84] 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.

[85] Olli Miettinen. Design options in epidemiologic research: an update. *Scandinavian journal of work, environment & health*, pages 7–14, 1982.

[86] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[87] KGM Moons, A Rogier T Donders, EW Steyerberg, and FE Harrell. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *Journal of clinical epidemiology*, 57(12):1262–1270, 2004.

[88] John Ashworth Nelder and R Jacob Baker. *Generalized linear models*. Wiley Online Library, 1972.

[89] 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.

[90] Jerzy Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

[91] Dung HM Nguyen and Jon D Patrick. Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, 21(5):893–901, 2014.

[92] Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. 2009.

[93] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

[94] Serguei Pakhomov, Susan A Weston, Steven J Jacobsen, Christopher G Chute, Ryan Meverden, Véronique L Roger, et al. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care*, 13(6 Part 1):281–288, 2007.

[95] 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.

[96] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

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

[98] John P Pestian, Christopher Brew, Paweł Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[99] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[100] Ross L Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11, 1986.

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

[102] Friedrich Pukelsheim. *Optimal design of experiments*, volume 50. siam, 1993.

[103] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):3, 2014.

[104] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

[105] Marie Reilly, Anna Torrång, and Åsa Klint. Re-use of case–control data for analysis of new outcome variables. *Statistics in medicine*, 24(24):4009–4019, 2005.

[106] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[107] Suchi Saria, Gayle McElvain, Anand K Rajani, Anna A Penn, and Daphne L Koller. Combining structured and free-text data for automatic coding of patient outcomes. In *AMIA Annual Symposium Proceedings*, volume 2010, page 712. American Medical Informatics Association, 2010.

[108] Tosiya Sato. Risk ratio estimation in case-cohort studies. *Environmental health perspectives*, 102(Suppl 8):53, 1994.

[109] Andrew I Schein and Lyle H Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

[110] Jonathan S Schildcrout and Patrick J Heagerty. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, 9(4):735–749, 2008.

[111] Jonathan S Schildcrout, Paul J Rathouz, Leila R Zelnick, Shawn P Garbett, and Patrick J Heagerty. Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *The annals of applied statistics*, 9(2):731, 2015.

[112] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[113] Herbert S Sichel. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547, 1975.

[114] 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.

[115] Ewout W Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008.

[116] Ewout W Steyerberg, Gerard JJM Borsboom, Hans C van Houwelingen, Marinus JC Eijkemans, and J Dik F Habbema. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in medicine*, 23(16):2567–2586, 2004.

[117] Ewout W Steyerberg and Yvonne Vergouwe. Towards better clinical prediction models: seven steps for development and an abcd for validation. *European heart journal*, 35(29):1925–1931, 2014.

[118] Carmen Stolwijk, Annelies Boonen, Astrid van Tubergen, and John D Reveille. Epidemiology of spondyloarthritis. *Rheumatic Disease Clinics*, 38(3):441–476, 2012.

[119] Thérèse A Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.

[120] Ting-Li Su, Thomas Jaki, Graeme L Hickey, Iain Buchan, and Matthew Sperrin. A review of statistical updating methods for clinical prediction models. *Statistical methods in medical research*, page 0962280215626466, 2016.

[121] Muhammad Atif Tahir, Josef Kittler, and Fei Yan. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750, 2012.

[122] W Katherine Tan, Saeed Hassanpour, Patrick J Heagerty, Sean D Rundell, Pradeep Suri, Hannu T Huhdanpaa, Kathryn James, David S Carrell, Curtis P Langlotz, Nancy L Organ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Academic radiology*, 2018.

[123] W Katherine Tan and Patrick J Heagerty. Multi-label surrogate-guided sampling designs for multi-label classification and validation. *Manuscript in preparation*.

[124] W Katherine Tan and Patrick J Heagerty. Predictive case control designs for modification learning. *Manuscript in preparation*.

[125] W Katherine Tan and Patrick J Heagerty. Surrogate-guided sampling designs for classification of rare outcomes from electronic medical records data. *Manuscript in preparation*.

[126] Jozef L Teugels. Some representations of the multivariate bernoulli and binomial distributions. *Journal of multivariate analysis*, 32(2):256–268, 1990.

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

[128] Katrin Tomanek and Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 45–48. Association for Computational Linguistics, 2009.

[129] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. A review of multi-label classification methods. In *Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006)*, pages 99–109, 2006.

[130] Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. *Machine learning: ECML 2007*, pages 406–417, 2007.

[131] Yvonne Vergouwe, Ewout W Steyerberg, Marinus JC Eijkemans, and J Dik F Habbema. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of clinical epidemiology*, 58(5):475–483, 2005.

[132] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

[133] Kathleen E Walsh, Sarah L Cutrona, Sarah Foy, Meghan A Baker, Susan Forrow, Azadeh Shoaibi, Pamala A Pawloski, Michelle Conroy, Andrew M Fine, Lise E Nigrovic, et al. Validation of anaphylaxis in the food and drug administration's mini-sentinel. *Pharmacoepidemiology and drug safety*, 22(11):1205–1213, 2013.

[134] 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.

[135] 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.

[136] Wei-Qi Wei and Joshua C Denny. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome medicine*, 7(1):41, 2015.

[137] Xuan Wei, Daniel Dajun Zeng, and Junming Yin. Multi-label annotation aggregation in crowdsourcing. *arXiv preprint arXiv:1706.06120*, 2017.

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

[139] Gary M Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.

[140] Weng-Kee Wong and Peter A Lachenbruch. Designing studies for dose response. *Department of Statistics, UCLA*, 2011.

[141] Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. *arXiv preprint arXiv:1609.00288*, 2016.

[142] 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.

[143] Yazhou Yang and Marco Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401–415, 2018.

[144] Louis E Yelle. The learning curve: Historical review and comprehensive survey. *Decision sciences*, 10(2):302–328, 1979.

[145] Meliha Yetisgen-Yildiz, Martin L Gunn, Fei Xia, and Thomas H Payne. A text processing pipeline to extract recommendations from radiology reports. *Journal of biomedical informatics*, 46(2):354–362, 2013.

[146] 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.

[147] 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.

[148] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616, 2001.

[149] John Zech, Margaret Pain, Joseph Titano, Marcus Badgeley, Javin Schefflein, Andres Su, Anthony Costa, Joshua Bederson, Joseph Lehar, and Eric Karl Oermann. Natural language–based machine learning models for the annotation of clinical radiology reports. *Radiology*, 287(2):570–580, 2018.

[150] Yi Zhai and Zhide Fang. Locally optimal designs for some dose–response models with continuous endpoints. *Communications in Statistics-Theory and Methods*, pages 1–17, 2017.

[151] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[152] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[153] 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.

[154] Xiao-Hua Zhou, Donna K McClish, and Nancy A Obuchowski. *Statistical methods in diagnostic medicine*, volume 569. John Wiley & Sons, 2009.

[155] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix A

APPENDIX FOR CHAPTER 2

A.0.1 Proof of Theorem 1

Denote the cohort data as $\mathcal{D} = (\mathbf{X}, Y)$, consisting of features $\mathbf{X}$ (implicitly also including the surrogate Z), and binary outcomes Y. From $\mathcal{D}$, units (typically subjects) are selected to form development and validation samples. The ideas in our proof are related to results from [97] and [142].

Preliminaries

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

$$(\mathbf{X}|Y = y) \sim N(\mu_{x|y}, \mathbf{\Sigma}_{x|y}) \tag{A.1}$$

This is equivalent to a Linear Discriminant Analysis (LDA) setting [46], where

$$\begin{aligned} logit(E[Y|\mathbf{X}]) &= \beta_0 + \beta^T \mathbf{X} \\ \beta_0 &= log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\mu_{x|y1} + \mu_{x|y0})^T \mathbf{\Sigma}_{x|y}^{-1}(\mu_{x|y1} - \mu_{x|y0}) \\ \beta^T &= \mathbf{\Sigma}_{x|y}^{-1}(\mu_{x|y1} - \mu_{x|y0}). \end{aligned} \tag{A.2}$$

The parameters in (A.2) are true parameters in $\mathcal{D}$. To estimate regression coefficients, a sample $\mathbf{D}^S(n)$ needs to be drawn from $\mathcal{D}$. Then, based on theory from generalized linear models [88], the resulting estimate $\hat{\beta}$ has the following first and second moments:

$$\begin{aligned} E^{\mathbf{D}^S(n)}[\hat{\beta}] &= \beta + Bias^{\mathbf{D}^S(n)}(\hat{\beta}) \\ Var^{\mathbf{D}^S(n)}(\hat{\beta}) &= (\mathbf{X}^{sT}\mathbf{W}\mathbf{X}^s)^{-1}. \end{aligned} \tag{A.3}$$

In (A.3), $\mathbf{W} = Diag(p_i(1 - p_i))$, where $p_i = P(Y_i = 1|\mathbf{X}, S_i = 1; \beta)$ estimates the aver-

age probabilities resulting from the sigmoidal transformation of development sample linear predictions. The terms in (A.3) are accurate up to second order approximations [88]. In estimating the regression parameters, denote the bias $Bias^{\mathbf{D}^S(n)}(\hat{\beta})$ as $\boldsymbol{B}(\hat{\beta}^S(n))$ and variance $Var^{\mathbf{D}^S(n)}(\hat{\beta})$ as $\boldsymbol{V}(\hat{\beta}^S(n))$, then both $\boldsymbol{B}(\hat{\beta}^S(n))$ and $\boldsymbol{V}(\hat{\beta}^S(n))$ depend on the development sample $\mathbf{D}^S(n)$ 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 $\mathcal{D}$. Denote the true linear predictions in the validation sample as $\eta := \mathbf{X}^v\beta$, with distribution

$$\begin{aligned} \mathbf{X}^v\beta &\sim N(\mu_y, \sigma_y^2) \\ \mu_y = \mu_{x|y}^T\beta; &\qquad \sigma_y^2 = \beta^T\boldsymbol{\Sigma}_{x|y}\beta \end{aligned}$$

for $y \in \{0, 1\}$, where $\mu_{x|y}$ and $\boldsymbol{\Sigma}_{x|y}$ were defined in (A.1). Under the bi-normal ROC assumption [97], 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 $\mathbf{D}^S(n)$, where $\mathbf{D}^S(n)$ is generated with sampling design S and with development sample size n. We use $AUC(Y|\mathbf{D}^S(n))$ to denote an indexing of resulting validation AUC, where

$$\begin{aligned} AUC(Y|\mathbf{D}^S(n)) &= \Phi(\sqrt{R_{AUC}(\mathbf{D}^S(n))}) \\ R_{AUC}(\mathbf{D}^S(n)) &= \frac{(\hat{\mu}_1 - \hat{\mu}_0)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}. \end{aligned} \tag{A.4}$$

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

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

$$\begin{aligned}\hat{\mu}_y &= E^{\mathbf{D}^S(n),\mathbf{X}^v}[\mathbf{X}^v\hat{\beta}|Y^v = y] \\ &= E^{\mathbf{X}^v}E^{\mathbf{D}^S(n)|\mathbf{X}^v}[\mathbf{X}^v\hat{\beta}|Y^v = y] \\ &= E^{\mathbf{X}^v}[\mathbf{X}^v(\beta + \boldsymbol{B}(\hat{\beta}^S(n)))|Y^v = y] \\ &= \mu_{x|y}^T(\beta + \boldsymbol{B}(\hat{\beta}^S(n))).\end{aligned} \tag{A.5}$$

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

$$\begin{aligned}\hat{\sigma}_y^2 &= Var^{\mathbf{D}^S(n),\mathbf{X}^v}(\mathbf{X}^v\hat{\beta}|Y^v = y) \\ &= Var^{\mathbf{X}^v}(E^{\mathbf{D}^S(n)|\mathbf{X}^v}[\mathbf{X}^v\hat{\beta}|Y^v = y]) + E^{\mathbf{X}^v}[Var^{\mathbf{D}^S(n)|\mathbf{X}^v}(\mathbf{X}^v\hat{\beta}|Y^v = y)]\end{aligned} \tag{A.6}$$

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

$$\begin{aligned}Var^{\mathbf{X}^v}(E^{\mathbf{D}^S(n)|\mathbf{X}^v}[\mathbf{X}^v\hat{\beta}|Y^v = y]) &= Var^{\mathbf{X}^v}(\mathbf{X}^v(\beta + \boldsymbol{B}(\hat{\beta}^S(n)))|Y^v = y) \\ &= (\beta + \boldsymbol{B}(\hat{\beta}^S(n)))^T\boldsymbol{\Sigma}_{x|y}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))),\end{aligned} \tag{A.7}$$

and the second part of the right hand side of (A.6) is

$$\begin{aligned}E^{\mathbf{X}^v}[Var^{\mathbf{D}^S(n)|\mathbf{X}^v}(X^v\hat{\beta}|Y^v = y)] &= E^{\mathbf{X}^v}[\mathbf{X}^{vT}\boldsymbol{V}(\hat{\beta}^S(n))\mathbf{X}^v|Y^v = y] \\ &= trace(\boldsymbol{V}(\hat{\beta}^S(n))\boldsymbol{\Sigma}_{x|y}) + \mu_{x|y}^T\boldsymbol{V}(\hat{\beta}^S(n))\mu_{x|y},\end{aligned} \tag{A.8}$$

where we have used properties of the expectation of a quadratic form: for $\epsilon \sim (\mu, \mathbf{\Sigma})$, $E[\epsilon^T \Lambda \epsilon] = trace(\Lambda\mathbf{\Sigma}) + \mu^T \Lambda \mu$. Therefore, combining (A.7) and (A.8), the variance of η is

$$\begin{aligned}\hat{\sigma}_y^2 &= Var^{\mathbf{X}^v}(E^{\mathbf{D}^S(n)|\mathbf{X}^v}[\mathbf{X}^v\hat{\beta}|Y^v = y]) + E^{\mathbf{X}^v}[Var^{\mathbf{D}^S(n)|\mathbf{X}^v}(\mathbf{X}^v\hat{\beta}|Y^v = y)] \\ &= (\beta + \boldsymbol{B}(\hat{\beta}^S(n)))^T \mathbf{\Sigma}_{x|y}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))) + trace(\boldsymbol{V}(\hat{\beta}^S(n))\mathbf{\Sigma}_{x|y}) + \mu_{x|y}^T \boldsymbol{V}(\hat{\beta}^S(n))\mu_{x|y}.\end{aligned} \tag{A.9}$$

Classifier validation AUC in terms of estimation variance

Now we plug in values for (A.4). WLOG assume that $\mu_{x|y0} = 0$ and that $\mathbf{\Sigma}_{x|y=1} = \mathbf{\Sigma}_{x|y=0} = \mathbf{\Sigma}_{x|y}$. Then, the means and variances of validation sample linear predictions among cases (Y=1) and controls (Y=0) are respectively

$$\begin{aligned}\hat{\mu}_1 &= \mu_{x|y1}^T(\beta + \boldsymbol{B}(\hat{\beta}^S(n))) \\ \hat{\mu}_0 &= 0 \\ \hat{\sigma}_1^2 &= (\beta + \boldsymbol{B}(\hat{\beta}^S(n)))^T \mathbf{\Sigma}_{x|y}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))) + trace(\mathbf{c}\mathbf{\Sigma}_{x|y}) + \mu_{x|y1}^T \boldsymbol{V}(\hat{\beta}^S(n))\mu_{x|y1} \\ \hat{\sigma}_0^2 &= (\beta + \boldsymbol{B}(\hat{\beta}^S(n)))^T \mathbf{\Sigma}_{x|y}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))) + trace(\boldsymbol{V}(\hat{\beta}^S(n))\mathbf{\Sigma}_{x|y}).\end{aligned} \tag{A.10}$$

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

$$\hat{\mu}_1 - \hat{\mu}_0 = \mu_{x|y1}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))), \tag{A.11}$$

while the denominator in (A.4) is

$$\hat{\sigma}_1^2 + \hat{\sigma}_0^2 = 2\{(\beta + \boldsymbol{B}(\hat{\beta}^S(n)))^T \mathbf{\Sigma}_{x|y}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))) + trace(\boldsymbol{V}(\hat{\beta}^S(n))\mathbf{\Sigma}_{x|y})\} + \mu_{x|y1}^T \boldsymbol{V}(\hat{\beta}^S(n))\mu_{x|y1} \tag{A.12}$$

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

$$AUC(\mathbf{D}^s(n)) = \frac{(\mu_{x|y1}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))))^2}{2((\beta + \boldsymbol{B}(\hat{\beta}^S(n)))^T \boldsymbol{\Sigma}_{x|y}(\beta + \boldsymbol{B}(\hat{\beta}^S(n))) + trace(\boldsymbol{V}(\hat{\beta}^S(n))\boldsymbol{\Sigma}_{x|y})) + \mu_{x|y1}^T \boldsymbol{V}(\hat{\beta}^S(n))\mu_{x|y1}}.$$

When $\boldsymbol{B}(\hat{\beta}^S(n)) \approx 0$, then since β, $\mu_{x|y}$ and $\boldsymbol{\Sigma}_{x|y}$ are assumed to be "fixed" quantities in a large validation sample,

$$AUC(\mathbf{D}^s(n)) \propto \frac{1}{trace(\boldsymbol{V}(\hat{\beta}^S(n))\boldsymbol{\Sigma}_{x|y})) + \mu_{x|y1}^T \boldsymbol{V}(\hat{\beta}^S(n))\mu_{x|y1}}.$$

A.0.2 Derivations and approximations of O_{ratio}

Derivation of Proposition 1

For $Y \in \{0,1\}$, $E[Y] = P(Y = 1)$. Denote subjects where $S = 1$ as those included in $\mathbf{D}^{SGS}(n)$, 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

$$
\begin{aligned}
Odds(cases|SGS) &= \frac{E^{\mathbf{D}^{SGS}(n)}[Y|S=1]}{1-E^{\mathbf{D}^{SGS}(n)}[Y|S=1]} = \frac{P(Y=1|S=1)}{P(Y=0|S=0)} \\
&= \frac{P(Y=1|S=1,Z=1)P(Z=1|S=1)+P(Y=1|S=1,Z=0)P(Z=0|S=1)}{P(Y=0|S=1,Z=1)P(Z=1|S=1)+P(Y=0|S=1,Z=0)P(Z=0|S=1)} \\
&= \frac{P(Y=1|Z=1)P(Z=1|S=1)+P(Y=1|Z=0)P(Z=0|S=1)}{P(Y=0|Z=1)P(Z=1|S=1)+P(Y=0|Z=0)P(Z=0|S=1)} \\
&= \frac{P(Y=1)}{P(Y=0)} \frac{R\dfrac{P(Z=1|Y=1)}{P(Z=1)}+(1-R)\dfrac{P(Z=0|Y=1)}{P(Z=0)}}{R\dfrac{P(Z=1|Y=0)}{P(Z=1)}+(1-R)\dfrac{P(Z=0|Y=0)}{P(Z=0)}} \\
&= \frac{P(Y=1)}{P(Y=0)} \frac{R(1-P(Z=1))Z_{sens}+P(Z=1)(1-R)(1-Z_{sens})}{R(1-P(Z=1))(1-Z_{spec})+P(Z=1)(1-R)(Z_{spec})} \\
&= \frac{P(Y=1)}{P(Y=0)} \frac{RZ_{sens}+p_Z(1-R-Z_{sens})}{R(1-Z_{spec})+p_Z(Z_{spec}-R)}
\end{aligned}
$$

where

$$
\begin{aligned}
R &= P(Z=1|S=1) \\
p_Z &= P(Z=1) \\
Z_{sens} &= P(Z=1|Y=1) \\
Z_{spec} &= P(Z=0|Y=0).
\end{aligned}
$$

The expected case odds in samples collected using SRS is

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

Then, the case/control odd ratio of samples obtained with SGS compared to that of SRS is:

$$\begin{aligned} O_{ratio} &= \frac{E^{\mathbf{D}^{SGS}(n)}[Y|S=1]}{1-E^{\mathbf{D}^{SGS}(n)}[Y|S=1]} \Big/ \frac{E^{\mathbf{D}^{SRS}(n)}[Y|S=1]}{1-E^{\mathbf{D}^{SRS}(n)}[Y|S=1]} \\ &= \frac{E^{\mathbf{D}^{SGS}(n)}[Y|S=1]}{1-E^{\mathbf{D}^{SGS}(n)}[Y|S=1]} \Big/ \frac{P(Y=1)}{P(Y=0)} \\ &= \frac{RZ_{sens}+p_Z(1-R-Z_{sens})}{R(1-Z_{spec})+p_Z(Z_{spec}-R)}. \end{aligned} \tag{A.1}$$

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

$$\begin{aligned} O_{ratio} &= \frac{RZ_{sens}+p_Z(1-R-Z_{sens})}{R(1-Z_{spec})+p_Z(Z_{spec}-R)} \\ &= \frac{\dfrac{R}{1-P(Y=1|Z=1)}(LR+)+\dfrac{1-R}{P(Y=0|Z=0)}(LR-)}{\dfrac{R}{1-P(Y=1|Z=1)}+\dfrac{1-R}{P(Z=0|Y=0)}} \\ &\approx (R)(LR+)+(1-R)(LR-) \end{aligned} \tag{A.2}$$

where

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

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

$LR+$ and $LR-$ are the likelihood ratios of the surrogate Z in predicting the outcome Y among surrogate positives and negatives, respectively.

A.0.3 Details of enrichment surrogate for data application

Table A.1 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

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

Table A.1: 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

Appendix B

APPENDIX FOR CHAPTER 3

B.0.1 Proof of Lemma 1: Statistical validity of mlSGS for model development

For the multi-label surrogate-guided sampling (mlSGS) design consisting of sub-samples $\{\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}_{SRS}\}$, let $\tilde{Z} \in \{0,1\}^K$ indicate the binary surrogate vector and $\tilde{S} = (S_1 \ldots S_K, S_{SRS})$ denote the vector of sampling into each sub-sample. For each $k = 1, \ldots, K$, since sampling into the sub-samples only depends on surrogate positives, $S_k \perp Z_k | (\tilde{Y}, S_{k'})$. Therefore, factoring the joint distribution to products of conditional distributions,

$$\begin{aligned} P(\tilde{S}|\tilde{Z},\tilde{Y}) &= P(S_1, \ldots, S_K, S_{SRS}|\tilde{Z},\tilde{Y}) \\ &= P(S_K|S_{SRS}, S_1, \ldots, S_{K-1}, \tilde{Z}, \tilde{Y}) \times \ldots \times P(S_1|S_{SRS}, \tilde{Z}, \tilde{Y}) \times P(S_{SRS}|\tilde{Z}, \tilde{Y}) \\ &= P(S_K|S_{SRS}, S_1, \ldots, S_{K-1}, \tilde{Z}) \times \ldots \times P(S_1|S_{SRS}, \tilde{Z}) \times P(S_{SRS}|\tilde{Z}) \\ &= P(\tilde{S}|\tilde{Z}), \end{aligned}$$

where

$$P(\tilde{S}|\tilde{Z},\tilde{Y}) = P(\tilde{S}|\tilde{Z}) \implies \tilde{S} \perp \tilde{Y}|\tilde{Z}.$$

Therefore, by definition

$$f(\tilde{Y}|\tilde{Z},\tilde{S}) = f(\tilde{Y}|\tilde{Z}),$$

and for any function $g(\tilde{S})$,

$$f(\tilde{Y}|\tilde{Z}, g(\tilde{S})) = f(\tilde{Y}|\tilde{Z}).$$

B.0.2 Proof of Lemma 2: Sampling weights of mlSGS with sequential sampling implementation

For the multi-label surrogate-guided sampling (mlSGS) design consisting of sub-samples $\{\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}_{SRS}\}$ each having size $\{n_1 \ldots n_K, n_{SRS}\}$, let $\tilde{S} = (S_1 \ldots S_K, S_{SRS})$ denote the vector of sampling into each sub-sample. Denote overall sampling weights under the mlSGS design with sequential sampling implementation as

$$\begin{aligned}
\pi_i &= P(\tilde{S}^T\tilde{1} = 1|\tilde{Z}_i = \tilde{z}) \\
&= P(i \in \{\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}_{SRS}\}) \\
&= P(i \in \mathcal{S}_{SRS}) \cup P(i \in \mathcal{S}_1) \cup \ldots \cup P(i \in \mathcal{S}_K).
\end{aligned}$$

For the SRS sub-sample $\mathcal{S}_{SRS}$: From the cohort of N subjects, n_{SRS} are drawn randomly. Therefore, the sub-sample sampling weights are:

$$\begin{aligned}
\pi_{SRS,i} &:= P(i \in \mathcal{S}_{SRS}) \\
&= P(S_{SRS} = 1|\tilde{Z}_i = \tilde{z}) \\
&= \frac{n_{SRS}}{N}
\end{aligned}$$

For the sub-sample $\mathcal{S}_1$: From the remaining $N - n_{SRS}$ subjects, n_1 are drawn from those with $Z_1 = 1$. Therefore, the sub-sample sampling weights are:

$$
\begin{aligned}
\pi_{1,i} &:= P(i \in \mathcal{S}_1) \\
&= P(S_1 = 1 | \tilde{Z}_i = \tilde{z}) \\
&= \sum_{s \in \{0,1\}} P(S_1 = 1 | S_{SRS} = s, \tilde{Z}_i = \tilde{z}) P(S_{SRS} = s | \tilde{Z}_i = \tilde{z}) \\
&= P(S_1 = 1 | S_{SRS} = 0, \tilde{Z}_i = \tilde{z}) P(S_{SRS} = 0 | \tilde{Z}_i = \tilde{z}) \\
&= \frac{P(\tilde{Z}_i = \tilde{z} | S_1 = 1, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z} | S_{SRS} = 0)} \frac{P(\tilde{Z}_i = \tilde{z} | S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z})} \times P(S_1 = 1 | S_{SRS} = 0) P(S_{SRS} = 0) \\
&= \frac{P(\tilde{Z}_i = \tilde{z} | S_1 = 1, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z})} \times \left(\frac{n_1}{N - n_{SRS}} \right) \left(\frac{N - n_{SRS}}{N} \right) \\
&= \begin{cases} \dfrac{P(\tilde{Z}_i = \tilde{z} | S_1 = 1)}{P(\tilde{Z}_i = \tilde{z})} \times \dfrac{n_1}{N}, & Z_{1i} = 1 \\ 0, & Z_{1i} = 0 \end{cases}
\end{aligned}
$$

since $S_{SRS} \perp \tilde{Z}$, $S_1 = 1$ only when $S_{SRS} = 0$, and that $S_1 = 1$ is based only on $Z_1 = 1$.

For the sub-sample $\mathcal{S}_2$: From the remaining $N - n_{SRS} - n_1$ subjects, n_2 are drawn from those with $Z_2 = 1$. Therefore, the sub-sample sampling weights are:

$$
\begin{aligned}
\pi_{2,i} &:= P(i \in \mathcal{S}_2) \\
&= P(S_2 = 1 | S_1 = 0, S_{SRS} = 0, \tilde{Z}_i = \tilde{z}) \\
&= P(S_2 = 1 | S_1 = 0, S_{SRS} = 0, \tilde{Z}_i = \tilde{z}) P(S_1 = 0, S_{SRS} = 0 | \tilde{Z}_i = \tilde{z}) \\
&= \frac{P(\tilde{Z}_i = \tilde{z} | S_2 = 1, S_1 = 0, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z} | S_1 = 0, S_{SRS} = 0)} \frac{P(\tilde{Z}_i = \tilde{z} | S_1 = 0, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z})} \\
&\times P(S_2 = 1 | S_1 = 0, S_{SRS} = 0) P(S_1 = 0, S_{SRS} = 0) \\
&= \frac{P(\tilde{Z}_i = \tilde{z} | S_2 = 1, S_1 = 0, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z})} \times \left(\frac{n_2}{N - n_{SRS} - n_1} \right) \left(\frac{N - n_{SRS} - n_1}{N} \right) \\
&= \begin{cases} \dfrac{P(\tilde{Z}_i = \tilde{z} | S_2 = 1, S_1 = 0)}{P(\tilde{Z}_i = \tilde{z})} \times \dfrac{n_2}{N}, & Z_{2i} = 1 \\ 0, & Z_{2i} = 0 \end{cases}
\end{aligned}
$$

since $S_{SRS} \perp \tilde{Z}$, $S_2 = 1$ only when $S_{SRS} = 0, S_1 = 0$, and that $S_2 = 1$ is based only on $Z_2 = 1$.

Generalizing to the sub-sample $\mathcal{S}_k$: From the remaining $N - n_{SRS} - \sum_{k'<k}^{K} n_k$ subjects, n_k are drawn from those with $Z_k = 1$. Therefore, the sub-sample sampling weights are:

$$
\begin{aligned}
\pi_{k,i} &:= P(i \in \mathcal{S}_k) \\
&= P(S_k = 1 | S_{k'<k} = 0, S_{SRS} = 0, \tilde{Z}_i = \tilde{z}) P(S_{k'<k} = 0, S_{SRS} = 0, \tilde{Z}_i = \tilde{z}) \\
&= \frac{P(\tilde{Z}_i = \tilde{z} | S_k = 1, S_{k'<k} = 0, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z} | S_{k'<k} = 0, S_{SRS} = 0)} \frac{P(\tilde{Z}_i = \tilde{z} | S_{k'<k} = 0, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z})} \\
&\times P(S_k = 1 | S_{k'<k} = 0, S_{SRS} = 0) P(S_{k'<k} = 0, S_{SRS} = 0) \\
&= \frac{P(\tilde{Z}_i = \tilde{z} | S_k = 1, S_{k'<k} = 0, S_{SRS} = 0)}{P(\tilde{Z}_i = \tilde{z})} \times \left(\frac{n_k}{N - n_{SRS} - \sum_{k'<k}^{K} n_k} \right) \left(\frac{N - n_{SRS} - \sum_{k'<k}^{K} n_k}{N} \right) \\
&= \begin{cases} p_{\tilde{Z}_i}^{(k)} \times \dfrac{n_k}{N}, & Z_{ki} = 1 \\ 0, & Z_{ki} = 0 \end{cases}
\end{aligned}
$$

with $p_{\tilde{Z}_i}^{(k)} := \dfrac{P(\tilde{Z}_i = \tilde{z} | S_k = 1, S_{k'<k} = 0)}{P(\tilde{Z}_i = \tilde{z})}$, since $S_{SRS} \perp \tilde{Z}$, $S_k = 1$ only when $S_{SRS} = 0, S_{k'<k} = 0$, and that $S_k = 1$ is based only on $Z_k = 1$.

Computation and approximation of $p_{\tilde{Z}_i}^{(k)}$: The ratio of surrogate cross-classification values comparing subsample $\mathcal{S}_k$ to cohort $\mathcal{D}$, $p_{\tilde{Z}_i}^{(k)}$, may be computed empirically. Alternatively, assuming that the total mlSGS sample size n is much smaller than cohort size N such that any perturbations in distributions when cascading down the subsample sequence is negligible, we claim that

$$p_{\tilde{Z}_i}^{(k)} \approx P(Z_k = 1)^{-1}.$$

We demonstrate this claim for $K = 2$ and note that it generalizes with additional algebra. For $\tilde{Z}_i^T \in \{0,1\}^K$, K=2, let p_{Z1} and p_{Z2} denote the marginal proportions of $Z1 = 1$ and $Z2 = 1$ respectively in the cohort. Since there are only 2^K -1 = 3 degrees of freedom in the

cross-classifications of $\tilde{Z}_i^T$, then observe the following cross-classification proportions:

$(Z1, Z2)$	$p_{\tilde{Z}_i}$ in $\mathcal{D}$	$p_{\tilde{Z}_i}$ in $\mathcal{S}_1$	$\frac{P(\tilde{Z}_i = \tilde{z}\|S_1 = 1)}{P(\tilde{Z}_i = \tilde{z})}$	$p_{\tilde{Z}_i}$ in $\mathcal{S}_2$	$\frac{P(\tilde{Z}_i = \tilde{z}\|S_2 = 1)}{P(\tilde{Z}_i = \tilde{z})}$
(1,1)	$p_{Z1,Z2}$	$\frac{p_{Z1,Z2}}{p_{Z1}}$	$\frac{1}{p_{Z1}}$	$\frac{p_{Z1,Z2}}{p_{Z2}}$	$\frac{1}{p_{Z2}}$
(1,0)	p_{Z1} - $p_{Z1,Z2}$	$\frac{p_{Z1} - p_{Z1,Z2}}{p_{Z1}}$	$\frac{1}{p_{Z1}}$	0	0
(0,1)	p_{Z2} - $p_{Z1,Z2}$	0	0	$\frac{p_{Z2} - p_{Z1,Z2}}{p_{Z2}}$	$\frac{1}{p_{Z2}}$
(0,0)	1-p_{Z1} - p_{Z2} + $p_{Z1,Z2}$	0	0	0	0

Therefore, since

$$\frac{P(\tilde{Z}_i = \tilde{z}|S_k = 1)}{P(\tilde{Z}_i = \tilde{z})} = \frac{1}{P(Z_k = 1)},$$

assuming that $n << N$ so that $P(\tilde{Z}_i = \tilde{z}|S_k = 1) \approx P(\tilde{Z}_i = \tilde{z}|S_k = 1, S_{k'<k} = 0)$ then

$$p_{\tilde{Z}_i}^{(k)} \approx \frac{1}{P(Z_k = 1)}.$$

Overall sampling weights: Finally, the overall sampling weight for subject i is

$$\pi_i = \frac{n_{SRS}}{N} + Z_{1i}\frac{n_1}{N}p_{\tilde{Z}_i}^{(1)} + \ldots + Z_{Ki}\frac{n_K}{N}p_{\tilde{Z}_i}^{(K)}.$$

B.0.3 Details of surrogate creation in the data application example

Table B.1: International Classification of Disease (ICD) codes and long description of ICD codes used to create surrogate variables for the multi-label surrogate-guided sampling design (mlSGS) illustrated in the data application example.

(a) ICD codes used to create surrogate for aortic aneurysm

ICD code	Long description
441.4	Abdominal aneurysm without mention of rupture
441.2	Thoracic aneurysm without mention of rupture
441.9	Aortic aneurysm of unspecified site without mention of rupture
447.72	Abdominal aortic ectasia
441.7	Thoracoabdominal aneurysm, without mention of rupture
441.01	Dissection of aorta, thoracic
441.1	Thoracic aneurysm, ruptured
441.02	Dissection of aorta, abdominal
441.03	Dissection of aorta, thoracoabdominal
441.3	Abdominal aneurysm, ruptured
441.5	Aortic aneurysm of unspecified site, ruptured

(b) ICD codes used to create surrogate for infection

ICD code	Long description
730.28	Unspecified osteomyelitis, other specified sites
324.1	Intraspinal abscess
730.08	Acute osteomyelitis, other specified sites
324.9	Intracranial and intraspinal abscess of unspecified site
730.18	Chronic osteomyelitis, other specified sites
730.88	Other infections involving bone in diseases classified elsewhere, other specified sites
730.98	Unspecified infection of bone, other specified sites
39	Cutaneous actinomycotic infection
478.24	Retropharyngeal abscess
711.8	Arthropathy associated with other infectious and parasitic diseases, site unspecified
730.19	Chronic osteomyelitis, multiple sites
730.29	Unspecified osteomyelitis, multiple sites
730.89	Other infections involving bone in diseases classified elsewhere, multiple sites
730.9	Unspecified infection of bone, site unspecified

(c) ICD codes used to create surrogate for spinal malignancy

ICD code	Long description
733.13	Pathologic fracture of vertebrae
V54.27	Aftercare for healing pathologic fracture of vertebrae
239.2	Neoplasm of unspecified nature of bone, soft tissue, and skin
733.20	Cyst of bone (localized), unspecified
198.5	Secondary malignant neoplasm of bone and bone marrow
V54.29	Aftercare for healing pathologic fracture of other bone
225.2	Benign neoplasm of cerebral meninges
225.3	Benign neoplasm of spinal cord
731.0	Osteitis deformans without mention of bone tumor
225.4	Benign neoplasm of spinal meninges
238.0	Neoplasm of uncertain behavior of bone and articular cartilage
733.29	Other bone cyst
170.9	Malignant neoplasm of bone and articular cartilage, site unspecified
V10.81	Personal history of malignant neoplasm of bone
170.2	Malignant neoplasm of vertebral column, excluding sacrum and coccyx
192.2	Malignant neoplasm of spinal cord
170.6	Malignant neoplasm of pelvic bones, sacrum, and coccyx
213.9	Benign neoplasm of bone and articular cartilage, site unspecified
213.2	Benign neoplasm of vertebral column, excluding sacrum and coccyx
192.3	Malignant neoplasm of spinal meninges
213.6	Benign neoplasm of pelvic bones, sacrum, and coccyx
209.73	Secondary neuroendocrine tumor of bone
733.21	Solitary bone cyst
192.1	Malignant neoplasm of cerebral meninges
237.6	Neoplasm of uncertain behavior of meninges

(d) ICD codes used to create surrogate for spondyloarthropathy

ICD code	Long description
720.2	Sacroiliitis, not elsewhere classified
720.0	Ankylosing spondylitis
720.9	Unspecified inflammatory spondylopathy
718.50	Ankylosis of joint, site unspecified
718.55	Ankylosis of joint, pelvic region and thigh
718.58	Ankylosis of joint, other specified sites
718.51	Ankylosis of joint, shoulder region
720.89	Other inflammatory spondylopathies

Appendix C

APPENDIX FOR CHAPTER 4

C.0.1 Proof of Lemma 1

For subject i let

$$\Delta_i = \begin{cases} 1, & \text{sampled} \\ 0, & \text{otherwise} \end{cases}$$

Under the PCC design, for strata defined by scores S exceeding threshold k, sampling is based on re-weighting strata frequencies. Thus by construction

$$P(S_i > k|\Delta_i = 1) = w$$
$$P(S_i \leq k|\Delta_i = 1) = 1 - w.$$

Therefore,

$$\begin{aligned} P_{PCC}(\Delta_i = 1) &= \begin{cases} P(\Delta_i = 1|S_i > k), & S_i > k \\ P(\Delta_i = 1|S_i \leq k), & S_i \leq k \end{cases} \\ &= \begin{cases} \dfrac{P(S_i > k|\Delta_i = 1)P(\Delta_i = 1)}{P(S_i > k)}, & S_i > k \\ \dfrac{P(S_i \leq k|\Delta_i = 1)P(\Delta_i = 1)}{P(S_i \leq k)}, & S_i \leq k \end{cases} \\ &= \begin{cases} P(\Delta_i = 1) \times \dfrac{w}{P(S_i > k)}, & S_i > k \\ P(\Delta_i = 1) \times \dfrac{1 - w}{P(S_i \leq k)}, & S_i \leq k \end{cases} \end{aligned}$$

Let $w = P(S_i > k)$. Then,

$$P_{PCC}(\Delta_i = 1) = \begin{cases} P(\Delta_i = 1) \times \dfrac{w}{w}, & S_i > k \\ P(\Delta_i = 1) \times \dfrac{1-w}{1-w}, & S_i \leq k \end{cases}$$
$$:= P_{SRS}(\Delta_i = 1)$$

C.0.2 Proof of Lemma 2

Define $p_i = \text{expit}(S_i)$ so p is monotone increasing in S. Since PCC is a stratified sampling strategy,

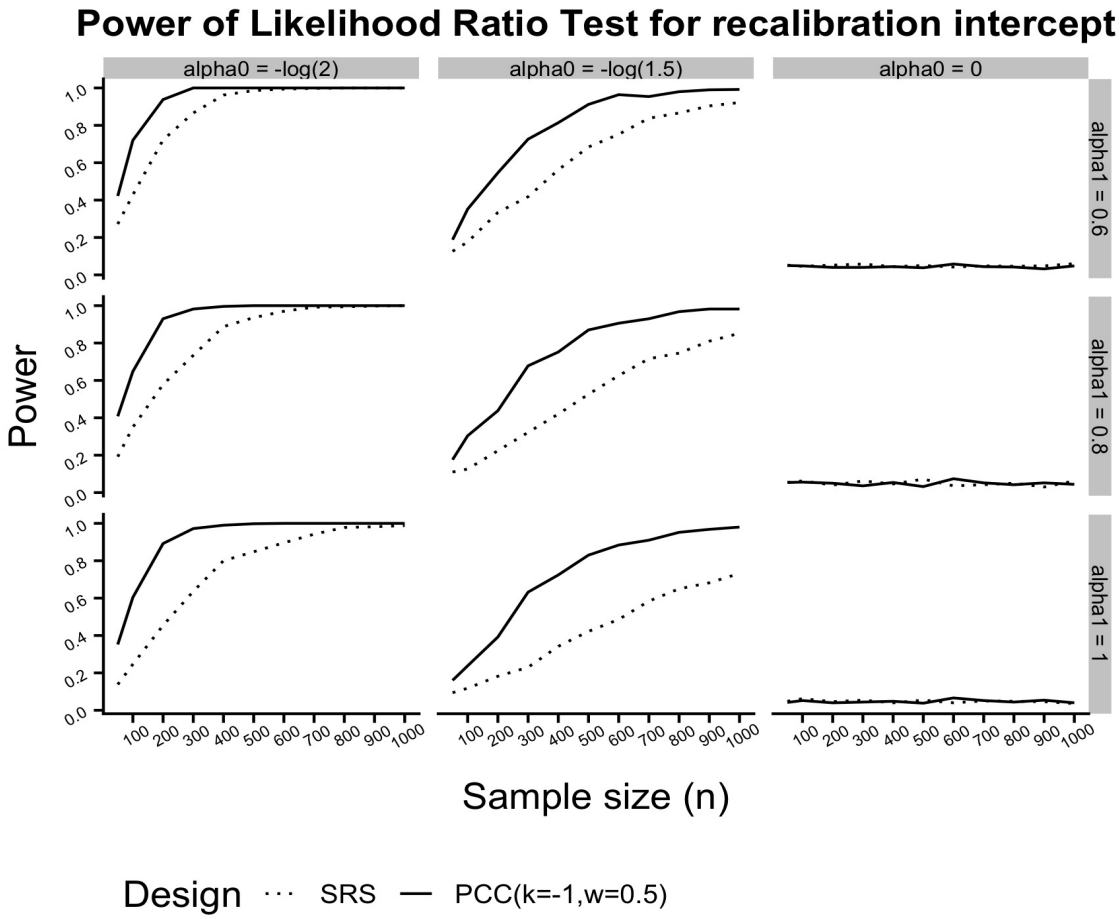
$$E_{PCC}[p] = E[p|I(S > k)]w + E[p|I(S \leq k)](1 - w)$$

and recall that for fixed k, $E_{SRS}[p] = E_{PCC}[p]$ when $w = P(S > k)$. For the Binary Entropy criterion defined as $\phi^B(x) = -x\log_2(x) - (1-x)\log_2(1-x)$, $\phi^B(p(S)) = f(S)$ is a function of the scores. For the same fixed k, consider the alternative PCC configurations of either $w > P(S > k)$ or $w < P(S > k)$.

Case: $w > P(S > k)$ Using $w > P(S > k)$ implies that $E_{PCC}[p] > E_{SRS}[p]$. Therefore, $\phi^B(p(S)|PCC) > \phi^B(p(S)|SRS))$ in the region of $E[p] < 0.50$ as $H(x)$ is monotone increasing in x when $x < 0.50$. Therefore for rare outcomes using higher stratum weights results in higher sample Binary Entropy.

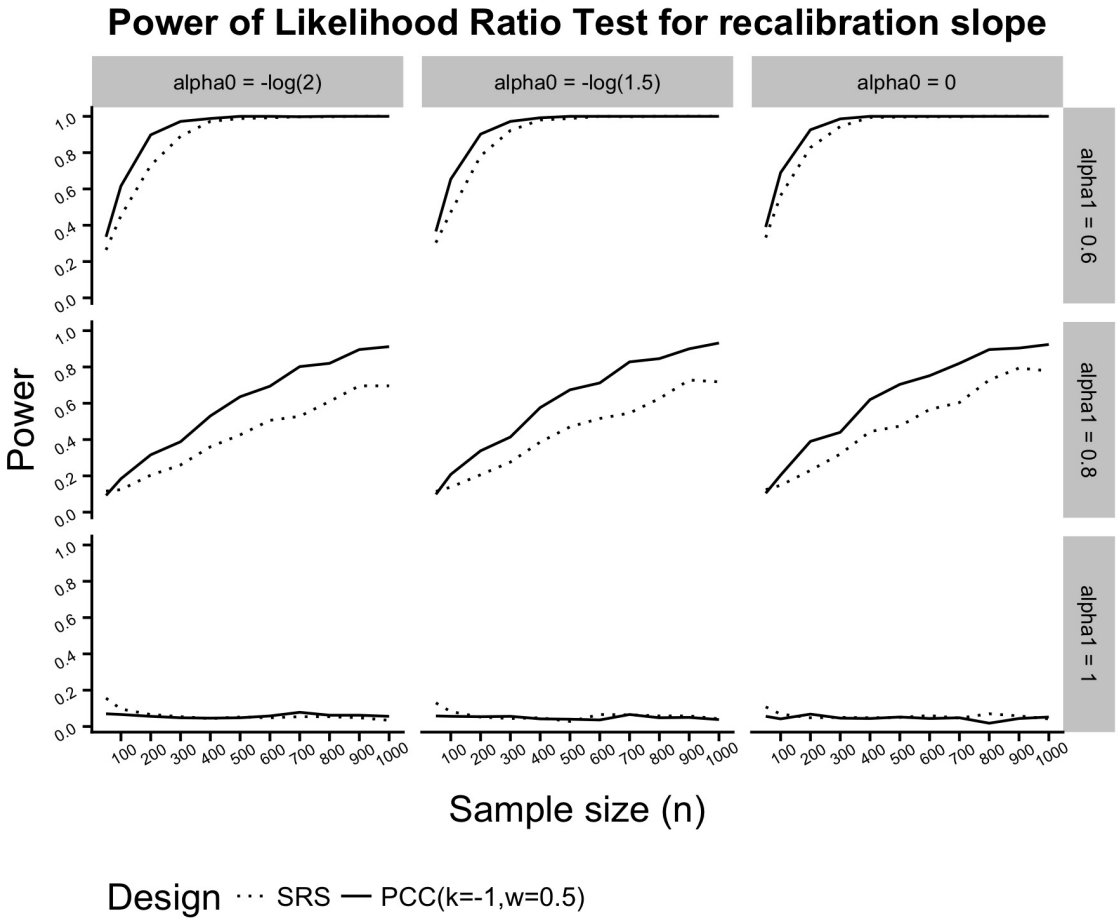
Case: $w < P(S > k)$ Using $w < P(S > k)$ implies that $E_{PCC}[p] < E_{SRS}[p]$. Therefore, $\phi^B(p(S)|PCC) > \phi^B(p(S)|SRS))$ in the region of $E[p] < 0.50$ as $\phi^B(x)$ is monotone decreasing in x when $x < 0.50$. Therefore for very prevalent outcomes using lower stratum weights results in higher sample Binary Entropy.

C.0.3 Additional simulation results for model recalibration



Comments on results:

- The power function under the PCC design is generally higher than under the SRS design.
- PCC benefit more pronounced for smaller effect size $\alpha_0 = -\log(1.5)$ compared to $-\log(2)$.
- Both designs provide correct test size.



Comments on results:

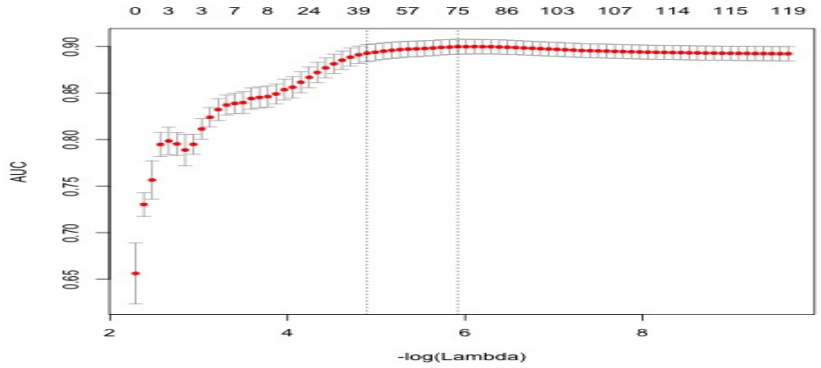
- The power function under the PCC design is generally higher than under the SRS design.
- PCC benefit more pronounced for smaller effect size $\alpha_1 = 0.80$ compared to 0.60.
- Both designs provide correct test size.

C.0.4 Details of true model definitions for data application example

The source model was estimated with the Logistic Lasso procedure, with source model coefficients $\hat{\tilde{\beta}}$ estimated using:

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta}}{argmin} \left\{ \sum_{i=1}^{n} -Y_i(\beta_0 + \tilde{X}_i^T\tilde{\beta} + \log(1 + \exp(\beta_0 + \tilde{X}_i^T\tilde{\beta})) + \lambda||\tilde{\beta}||_1 \right\}.$$

The penalty parameter λ was selected with 10-fold cross-validation using the AUC loss function, using value within 1 standard error (`lambda.1se`); cross-validation error plot:



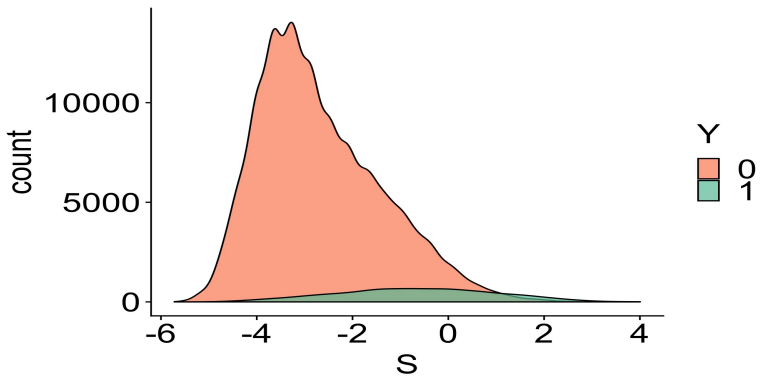
From the selected 38 features, by thresholding to 0.25 we obtained a smaller subset of 23; these coefficients together defined the "source model", which were:

Features	Estimate
Intercept	-2.92
anterior	1.55
compress	1.45
deform	1.20
endplat	0.83
not	0.81
fractur	0.67

slight	0.50
none	0.42
age_category60+	0.41
bodi	0.40
bone	0.34
sever	0.33
desicc	0.31
later	-0.26

spur	-0.28
evid	-0.29
miner	-0.30
intact	-0.31
sclerosi	-0.32
degen	-0.38
preserv	-0.48
maintain	-0.51
no	-1.09

The scores were computed as $S_i = \tilde{X}_i^T\hat{\tilde{\beta}}$. Note that the AUC in discriminating cases (Y=1) and controls (Y=0) by using S as the single test was 0.83. The score distribution by case/control status was:



The recalibration parameters were estimated to be $\alpha_0 = -1.00$, $\alpha_1 = 0.89$ using $logit(E[Y_i|S_i]) = \alpha_0 + \alpha_1 S_i$. The revision parameters were estimated with $logit(E[Y_i|S_i, \tilde{X}_i^T]) = \alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma}$ with

$$(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\tilde{\gamma}}) = \underset{\alpha_0, \alpha_1, \tilde{\gamma}}{argmin} \left\{ \sum_{i=1}^{n} -Y_i(\alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma} + \log(1 + \exp(\alpha_0 + \alpha_1 S_i + \tilde{X}_i^T \tilde{\gamma})) + \lambda ||\tilde{\gamma}||_1 \right\},$$

From the selected 64 features, by thresholding to 0.50 we obtained a smaller subset of 6; these coefficients together defined the "target model", which were:

Feature	Estimate
sublux	1.42
deform	1.15
fractur	0.78
scoliosi	0.78
normal	-0.53
desicc	-0.57