Dortmund
University Of Technology

BAYER
Science For A Better Life

Enrichment Designs and Sensitivity-preferred Classification

Dissertation

to obtain the Academic Degree of
Doktor der Naturwissenschaften

Presented to the Department of Statistics
TU Dortmund University
by

Inoncent Agueusop

Submitted in May 2014
Oral Examination held on 01.09.2014

Referees:

Prof. Dr. Katja Ickstadt
Prof. Dr. Jörg Rahnenführer
Dr. Richardus Vonk

Contents

List of Tables

List of Figures

1 Introduction

Pharmaceutical companies strive to improve quality of life and increase life expectancy by developing more effective and safer medication. They are constantly enhancing the drug development process and investing in research and innovation, which are key in the discovery and development of safer, more effective therapies. The process of research and development includes ethical and economic considerations that present additional challenges. The companies must manage the quality, or efficacy and safety, of the new drugs, their accessibility to patients in regard to price and timeliness, and their development costs. To optimally manage costs, quality, and timeliness, the drug development process must be continually enhanced, adopting both innovative methodological approaches and new technology.

Drug development is a lengthy process lasting at least 10 years. It involves drug discovery, formulation, laboratory development, animal studies, clinical development, and regulatory approval. Drug discovery, formulation, and laboratory development constitute the non-clinical/pre-clinical development stages. These involve in-vitro as well as in-vivo experiments. There are several stages in the pre-clinical development phase that evaluate the potential efficacy and safety features of the compound. Regulations require that certain pre-clinical safety assessments are successfully passed before starting the clinical development. Clinical trials, or studies conducted with humans, are carried out in three phases (phase I, phase II and phase III).

The safety and efficacy of the new drug must be established and compared with an eventual standard therapy to earn regulatory approval. Regulatory agencies such as the Food and Drug Administration (FDA) and European Medicines Agency (EMA) are responsible for new drug approvals in the USA and in the European Union, respectively. They examine the results of the studies conducted during the drug development to ensure that the new drug works and that its health benefits outweigh its known risks. A phase IV trial tends to be required after drug approval for the investigation of potential long term side effects.

Drug Development Issues

Drug development is a very costly process. The development costs estimation is complex, involving out-of-pocket expenditures and capitalization costs due to the long develop-

ment time (DiMasi et al. [2003]). Arriving at an accurate estimate of drug development costs is difficult because most of the data associated with drug development is confidential. Published estimates of average development costs from drug discovery to marketing vary depending on the literature. DiMasi et al. [2003] estimated the average costs per new drug to be $800 million. Parker et al. [2003] provided estimates between $500 millon and $2000 millon US Dollars by using a publicly available pharmaceutical projects data base. Figure 1 displays some estimates of the total development costs as well as the pre-clinical and clinical development costs in US Dollars. Drug development costs

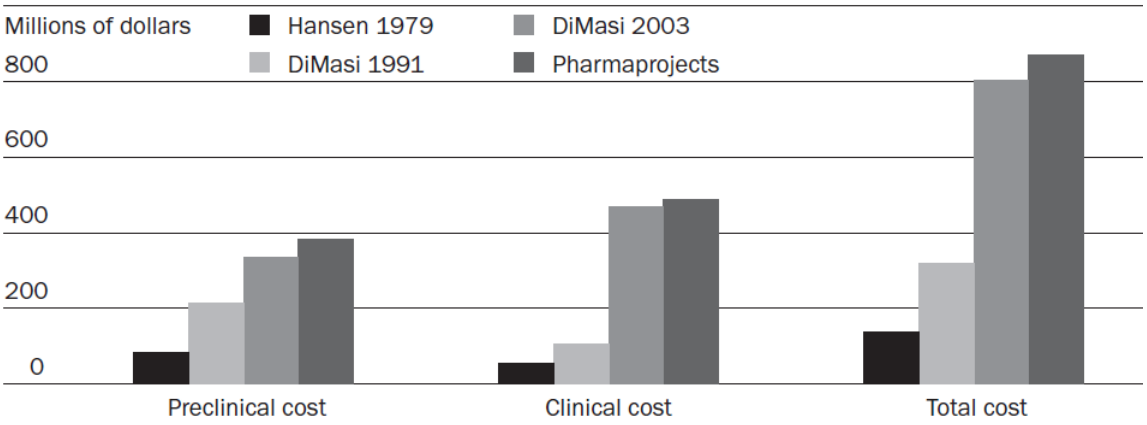

Figure 1: Capitalized preclinical, clinical, and total costs per new drug in millions of 2000 dollars. *Source*: Parker et al. [2003]. Others sources: DiMasi et al. [1991], DiMasi et al. [2003] and Chien [1979].

increase at a rate of about 7% per year, by comparing past estimates with more recent ones (Parker et al. [2003]). An accurate estimate of pre-approval development costs requires costs evaluation at the projects level and should assist investors in fixing the price of the new drug. Note that the development costs have a direct impact on drug prices and general health policy. Whilst the overall development costs have increased over the years, the number of new drug applications to the FDA's Center for Drug Evaluation and Research (CDER) have not risen. Figure 2 represents the number of applications for drug approval from 1991 to 2011 published by the FDA (Swann [2013]). The decreasing trend in the number of applications and newly approved drugs in the last years could partly be explained by the imposition of stricter approval conditions as well as the nature of current diseases. Drug development for contemporary genetic diseases such as cancers and metabolism diseases is more complicated than the infectious diseases of the past decades. To be competitive, a new drug must treat the disease at its origin, rather

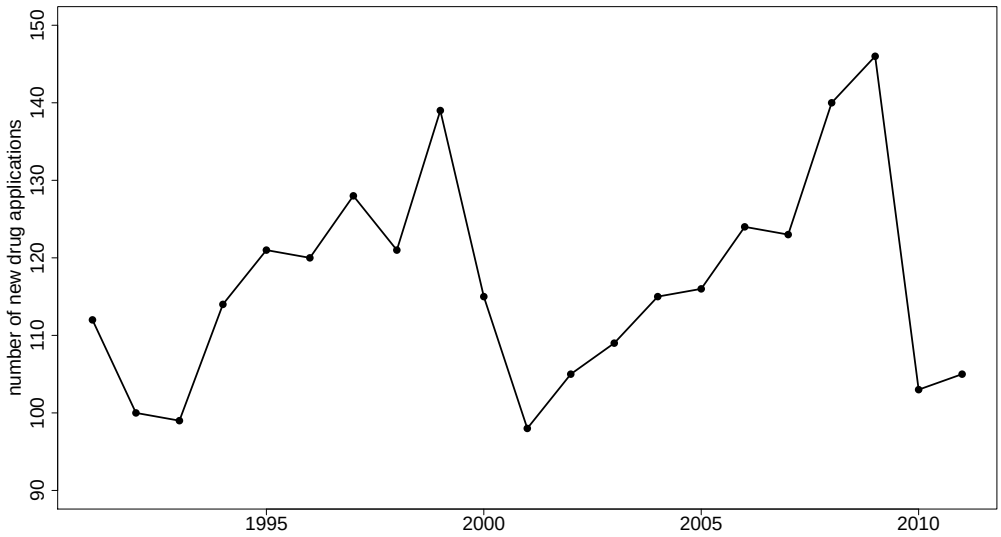

Figure 2: Number of new drugs submitted to the FDA from 1991 to 2011

than just the symptoms as in the past decades.

No drug has the same results on all patients. Some patients show no significant effects, or they cannot tolerate even those drugs proven as effective. Because genetic and/or environmental factors may explain this phenomenon, it may be crucial to find the biological criterion for the selection of non-responders and develop a more appropriate therapy for them. Interest in targeted drug development has been increasing during the last decade. A target patient population is identified in the early drug development phases as being more likely to benefit from the drug. Trials are conducted to prove the safety and efficacy of the new drug within this sub-population: This has been called *individualized medicine*.

Biomarkers

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of healthy biological processes, pathological processes, or pharmacological responses to therapeutic intervention (Atkinson et al. [2001]). Biomarkers are increasingly used in oncological research. Brünner [2009] has provided definitions of different types of biomarkers used in target drug development and cancer research. A **Prognostic biomarker** is one that provides information on the likely course of the disease in an untreated individual. **Predictive biomarkers** are those which can be used to select patients most

likely to respond to a new therapy. With predictive biomarkers, it is possible to select the therapy with the highest likelihood of efficacy to the individual patient. Predictive biomarkers are the basis for individualized or tailor-made treatment (Brünner [2009]). When the initial biomarker cannot be evaluated (e.g. lack of accurate measurement technique or high costs), a so called **surrogate biomarker** can be used. These strongly correlate with the primary markers.

If only a small fraction of those in a patients population responds well to a given drug, it may be difficult to show the drug effect in the entire population, and the trial may subsequently fail. This seems often to be the case in cancer research. In targeted drug development, the research is more focused on patients showing high effect size, who can be identified through established predictive biomarkers. The identification of predictive markers is one of the big challenge in the medical research.

Enrichment Designs

In enrichment trials, biomarkers help select a study population with higher percentage of patients more likely to respond strongly to a new drug. This proportion of high responders is pre-determined. Though authorities also often want to know what happens to the biomarker-negative patients, there are some indications in the guideline for enrichment trials (Temple and Becker [2012]) that drug development for a "target population" is gaining more acceptance from regulators (R. Vonk, personal communication Bayer Pharma AG). A higher effect within the marker-positive population leads to lower study sample size. Enrichment strategies in clinical trials may therefore reduce the drug development time and costs. However, if the prevalence of marker-positive patients is too low, the recruitment time may become excessively long resulting in very high study costs.

Issues in Conducting Enrichment Studies

Biomarker identification for disease diagnosis, drug response evaluation, identification of potential responders to a given therapy, the prediction of the disease progression and complication appearance is a fundamental to medical research. Enrichment trials used to develop the right drug for the right patient require established predictive markers. High-responders are identified (tested) using predictive biomarkers. In this thesis, enrichment trials refer to those trials for which only patients tested as marker-positive are

enrolled. However, marker-tests have the misclassification rates typical to any binary classification problem. The quality of the marker-test in selecting the study population and later in identifying the patients to be treated is an important factor with ethical consequences. Rejecting appropriate study candidates due to the lack of sensitivity of the marker-test may increase the study time and costs. Actual marker-positive patients who are tested as negative do benefit from the therapy. In addition, the enrollment of false positive patients for the study decreases the power, since the expected effect size in the study population is overestimated and the variability is underestimated.

If only a small proportion of patients in the unselected patient population satisfies the entry criteria, the recruitment time can increase dramatically, possibly becoming unreasonable even for small study sample sizes. Recruitment time is a very important factor in planning any clinical trial. For enrichment trials, it needs particular attention, since there are more factors influencing the time than in an unselected trial. Here, unselected trials refers to normal clinical trials. Enrichment studies tend to require a smaller sample size than unselected studies due to the larger effect size in the enriched population. However, the overall study costs might rise dramatically if an exorbitant number of patients must be tested to reach the required sample size. This number of patients to test depends on the marker prevalence, the study sample size, and specific marker-test characteristics. The study costs can be divided into three categories: screening costs and marker-test costs, care costs (expenditure incurred post-recruitment) and the time costs. The overall screening and marker-test costs depend on the prevalence of marker positive patients in the unselected patients' population.

In enrichment trials, patients are first screened according to general entry criteria such as vital signs and eligible ones are additionally submit to a marker-test. Test-positive patients are then randomized to the treatment groups (see Figure 3). The screening procedure is assumed to be a deterministic process: of all patients arriving at the recruitment centers, a fixed proportion (e.g. 90%) is found eligible. The principal concerns with the use of enrichment strategies relate to the generalizability of the study results. When considering whether to use an enrichment design, it is important to consider whether the enrichment strategy can be used in practice to identify the patients the drug should be given and if the drug might be useful in a broader population than the one studied. Usually, there is no evidence that the new drug does not work on off-label patients. This is not specific to enrichment trials, but for all trials with a label restriction. Sometimes,

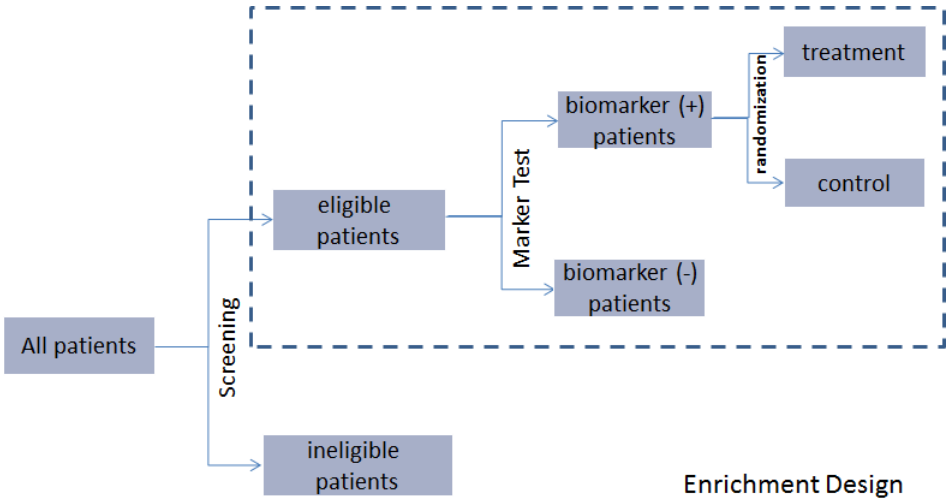

Figure 3: Eligible patients are submitted to a marker-test and only the test-positive are randomized for the trial.

drugs are used for diseases or populations for which they were not developed - this is called "off-label-use". When the treatment is a critical advance for the enriched group, it would generally be unreasonable to delay approval for the enriched group (Temple and Becker [2012]). Extension of the therapy to the marker-negative patient population is critical and requires information about this group as well as several assessments of risk-benefit.

It is difficult to pinpoint a single marker to select a study population. Usually, a list of biomarkers is compiled from a large number of predictors, such as gene expression or protein levels. These biomarkers are used to build a classification rule, since a marker-test tends to present a binary classification problem. Traditional methods in building binary classification rules usually assume equal importance between different classification errors. However, in the medical context, it is rare that classification errors have equal importance. It may be, for example, more important to classify marker-positive patients as positive, so that they can benefit from the therapy. In some situations, it might be crucial to guarantee a minimal true classification rate (e.g. 90%) in the most important class, implying a focus on the sensitivity or the specificity of the test, rather than on overall accuracy. Although such classification strategies are of great importance in the medical context, they are often ignored in research. Some techniques in this context try to increase the true classification rate in the most important class, but cannot

guarantee a pre-determined true classification rate in that class. In enrichment studies, an entire control of the misclassification rate developed as part of the classification rule could help minimize false positives who are treated though they will not benefit from effect the new drug or those rejected when they could benefit from the therapy.

Purposes of this Thesis

The first part of this thesis investigates Poisson processes for modeling patient recruitment in enrichment trials. We propose strategies in predicting patient recruitment time by deriving recruitment processes that consider the prevalence of marker-positive patients as well as the marker-test characteristics. Appropriate recruitment processes help in predicting the trial duration. We suggest costs predictive models which consider patient screening costs, marker-test costs, patients care costs during the trial, and the costs of waiting time (e.g. resources and infrastructure costs). We study the impact of misclassification at the marker-test stage. Assignment of patients to the wrong group presents an ethical burden and affects relevant factors such as the recruitment time, study costs, and the power of the study results. In the second part, we introduce a new approach to control entirely the sensitivity at the stage of building binary classifiers. Our method encourages sparsity through the L_1-norm regularization of the model parameters that enable its applicability to high-dimensional settings.

A major portion of expenditures in drug development come from unnecessary waiting. This waste of time can be lowered through accurate deadline planning, leading to an optimal use of resources. This requires accurate patient recruitment time and trial duration predictions which pass through appropriate modeling of the recruitment process. The number of recruited marker-positive patients (e.g. per day, week, month) is modelled as a random variable following a Poisson distribution. Patient recruitment models in traditional unselected clinical trials are well discussed in the literature. Anisimov and Fedorov [2007] is one of the best elaborated works in this context. Other interesting papers on this theme include Carter [2004], Tang et al. [2012], Mijoule et al. [2012], Anisimov [2008], Anisimov [2009] among others. These authors assumed that the number of patients recruited follows a Poisson process. Poisson processes are standard in modeling count data. Here, the number of patients enrolled per time unit is assumed to follow a Poisson distribution.

Patient recruitment processes allow researchers to predict the recruitment time before

starting the study. This is important for decision making. Recruitment time prediction in ongoing stages leads to more accurate estimates, since more information is available. Patient recruitment processes as count processes may be reasonably modeled through other processes of counts, such as Poisson autoregressive processes, which are largely discussed in the literature (Ferland et al. [2006], Fokianos et al. [2009]). Poisson autoregressive processes consider past information as the intensity of the Poisson process depends on past observations. They are relatively flexible in modeling special events in the process such as level shift and outliers (Fokianos and Fried [2010], Fried et al. [2013], Fokianos and Fried [2012]). We derive our models for patient recruitment in enrichment studies based on the existing processes of counts and provide analytic distributions of the recruitment time in most cases. We follow a Bayesian approach for model parameters estimation, as well as a Bayesian detection of level-changes within the process.

Most publications on drug development costs estimation focus on the estimation of the total drug development costs (Parker et al. [2003], DiMasi et al. [1991], DiMasi et al. [2003], and Chien [1979]). We however, group the expenditures in an enrichment trial in patient screening costs, marker-test costs, traditional care costs (e.g. visits, drug supply) and time costs. Marker-test costs, screening costs and waiting time may increase the overall study costs depending on the nature of the marker-test (e.g. blood pressure, blood glucose, gene profile), the marker prevalence, and indirectly on marker-test characteristics. Note that erroneously testing actual marker-positive patients as negative because the tests' lack of sensitivity increases the number of patients who must be tested, therefore raising screening costs, test costs, and recruitment time. This is difficult to control in practice, since we cannot identify the true- and false-positive patients.

In a statistical context, a biomarker test to assign patients into marker-positive and marker-negative groups is a two-class classification problem. In a broader sense, any effort to assign patients into subgroups (e.g. diseased and healthy) is a binary classification problem with different misclassification costs. Accounting for the costs of incorrect patient assignment to a group remains a crucial issue in building binary classifiers. In some, if not most, diagnostic situations, it could be essential to include the control of the true positive rate in the most important class (sensitivity), and reject the classifiers, which lead to a sensitivity less than a pre-determined lower bound of admissible values (for example 90%). The second part of this thesis deals with building binary classifiers in high-dimensional settings by different importances. Some approaches to this have

been suggested: The resampling method (Japkowicz and al. [2000]), weighting observations (Liu and Tan [2008]), thresholding the decision score (Sheng and Ling [2006]) and cost-sensitive learning (Elkan [2001]). However, while these approaches are designed to increase sensitivity, they cannot guarantee a pre-determined value. To guarantee a predetermined sensitivity value, the corresponding cut-off must be selected after computing the decision score, as investigated in Jung et al. [2010]. Here, decision scores, like the probabilities of having a disease, are computed, then a cut-off leading to the specified sensitivity is selected. That means the important information requiring a high sensitivity value is not considered in the optimization procedure and the achieved specificity is not necessarily the largest.

The new technique works by optimizing any loss function of binary classification under the constraint that the sensitivity belongs to a predetermined interval of high values, e.g. between 90 and 100%. Here the sensitivity denotes the true classification rate in the most important class. We illustrate this approach by considering the Bernoulli likelihood function in the case of logistic regression on the one hand, and the Youden-index as an objective function of classification on the other hand. Both functions are optimized subject to the constraint on the sensitivity and the L_1-norm of the model parameters to select simultaneously relevant predictors. This optimization strategy provides the best classifier with the sensitivity in the pre-determined interval.

The first part of this thesis investigates three issues in conducting enrichment studies. In section 2.1, we propose appropriate Poisson-based processes for patient recruitment and a strategy for recruitment time prediction in enrichment trials. Section 2.2 presents Bayesian techniques for updating the suggested models and identifying change points in the recruitment process. Techniques for costs and power management are investigated in section 2.3. The second part of this thesis deals with the problem of classification with different class importances also faced by researches when conducting enrichment studies. Section 3.1 and section 3.2 present the data material and backgrounds. In section 3.3, we propose a new approach for entirely controlling the sensitivity in building binary classifiers. An overview of techniques in dealing with class importance in building binary classifiers is given in section 3.4, and the results are presented in section 3.5. We end with conclusions in section 4 and attach references and an Appendix including R-codes to the new classification strategies.

2 Enrichment Trials: Recruitment Time, Costs Management and Power

We begin by investigating strategies to predict patient recruitment time at the initial stage of the recruitment process, as well as in the ongoing stage after the recruitment has started. Accurate prediction of recruitment time requires the definition of stochastic processes that best describe the recruitment process of marker positive patients. We then propose and analyse appropriate Poisson processes that consider factors specific to enrichment studies, such as the prevalence of marker-positive patients in the unselected patient population, as well as the sensitivity and specificity of the marker-test. In most cases, we derive an analytical distribution of the recruitment time also used in evaluating the study costs. For example, if the arrival process of patients at recruitment centers is assumed to be a Poisson process with a constant or a Gamma distributed rate, the time needed to enroll a given number of patients follows the Erlang and Pearson type VI distribution, respectively. Traditional factors affecting recruitment time, such as the number of study centers and the center capacities, are also considered.

Patient recruitment modeling in clinical trials is largely discussed in the literature. Almost all suggested models rely on Poisson distributed observations. Carter [2004] suggested homogeneous Poisson processes, assuming that eligible patients arrive at recruitment centers according to Poisson processes with constant intensity. Homogeneous Poisson processes assume that the variability in the process is equal to the expectation. However, patient recruitment processes usually present a larger variability (overdispersed) due to additional variability in the intensity. Poisson processes with constant intensity are then inaccurate in this circumstance (Anisimov and Fedorov [2007], Mijoule et al. [2012]). To overcome this limitation, Anisimov and Fedorov [2007] suggested Poisson processes with a Gamma distributed intensity (Gamma-Poisson processes). Equivalently, the observations of the process are Negative Binomial distributed, which better models the larger variability in the recruitment process. A comparable approach has been investigated by Mijoule et al. [2012], who investigated Poisson processes with a Pareto distributed intensity (Pareto-Poisson). The Pareto distribution is the probabilistic formulation of the well-known fact that around 80% of enrollments are made by 20% of centers (Mijoule et al. [2012]). Pareto-Poisson processes may apply better than Gamma-Poisson processes when a large number of centers each have a small capacity, as the authors argued. However, the Pareto distribution is more difficult to handle

computationally than the Gamma distribution and forces to use Monte Carlo techniques.

In practice, patient recruitment processes can be divided in three different phases to be considered in modeling the process for more accuracy. In the first phase, centers are progressively activated, meaning that the overall recruitment rate summed over the recruitment centers increases as a function of time. The second phase lasts longest, starting after all centers are initialized: the process becomes stable and oscillates randomly. A large number of models assume stable processes from the beginning of the recruitment process. In the last phase of the recruitment process, a level shift may occur, when the investigators are informed of an upcoming closure date (Tang et al. [2012]). They suggested to model the intensity at the first phase as an increasing function of time, then constant in the later stages. Patient recruitment processes are count processes and should benefit more from research advances in this area. Other well-established works in count process modeling are Poisson-autoregressive models (e.g. Ferland et al. [2006], Fokianos et al. [2009], Fokianos and Fried [2010]). Here, the intensity of the Poisson process is a function of the past observations and the past values of the intensity itself. However, none of these well elaborated works is designed for patient recruitment in enrichment trials, so the impact of prevalence and marker-test characteristics on recruitment time cannot be studied.

The suggested Poisson recruitment processes and cost models are explicitly designed for enrichment studies and can be used for any trial through parametrization. An unselected trial is equivalent to enrichment trials with a marker prevalence of 100%. Poisson processes offer mathematical and computational flexibility in multi-center enrollment insofar as the sum of independent Poisson processes is also a Poisson process. We assume that the recruitment centers are independent and the different processes are independent. This allows us to consider a unique Poisson process representing the sum of the different processes observed in the different recruitment centers. An estimate of the recruitment time at the initial stage before starting the study is crucial to investigating the feasibility of the trial. This requires estimates of center capacities and recruitment rates that may not be available at the initial stage. Here, information about past and comparable studies are usually combined with expert knowledge to derive an estimate of the overall recruitment rate. This estimate can be updated after some data are collected by using, for example, Bayesian methods. In enrichment trials, an estimate of the marker prevalence must be provided. This may be in the form of point- or interval estimates, or as a prob-

ability distribution, such as the beta distribution. Updates of the remaining recruitment time in the ongoing stage are useful in deciding if more centers should be activated to accomplish the study in time. We propose a Bayesian approach for updating our models.

Enrichment studies require a smaller study sample size since the effect size (standardized effect difference) is larger. Such trials are then expected to be cheaper than unselected ones. However, the study costs may be considerably affected by screening and marker-test procedures. This occurs when the marker-prevalence is very low (e.g. 10%), which implies that a large number of patients must be screened and tested to obtain the required sample size. The total test and screening costs may rise considerably for some types of marker-tests (e.g. gene profile evaluation). The costs may incrementally increase even if the screening and marker-test costs are very low: waiting and wasting time is decisive (resources costs). Note that the smaller the marker prevalence, the larger the uncertainty in the recruitment time, and the more difficult to plan deadlines. Enrichment studies for target drug development are increasingly necessary and popular. Hence, developing costs prediction modeling is increasingly important. We derive an analytical distribution of study costs by considering marker-prevalence and marker-test characteristics. Parker et al. [2003], DiMasi et al. [1991], DiMasi et al. [2003] and Chien [1979] focused on estimating total development costs. The cost model we suggest is designed for costs evaluation in enrichment studies. It represents the sum of the screening, marker-test costs, care costs, and duration costs, whereas the recruitment time is derived from the underlying recruitment process.

Some Practical Considerations

The inclusion of sensitivity and specificity of the marker-test into the prediction of the study costs and recruitment time helps to visualize the unseen impact of marker-test characteristics on enrichment studies. The marker-test must have reasonable test characteristics for ethical and economic reasons. In practice, these characteristics are usually unknown, since it is difficult, if not impossible, to identify the true positive and true negative patients in a given patient population. This means the marker prevalence is simply the proportion of marker-test positive patients in a given sample of unselected patients. The expected effect size in the marker positive group used for sample size calculation is estimated from a mixed population of marker-test positive patients (true positive and false positive) and may be underestimated. This dilution effect goes through each stage of the study (e.g. sample size calculation, observed effect, study results), and there is

no need of extra consideration of test characteristics in evaluating the time and costs. However, marker-test characteristics should be considered if a very expensive marker-test is replaced by a cheaper test to reduce development costs.

Another important factor in planning an enrichment study is the study's possible loss of potency due to the enrollment of false positive patients. The impact of the sensitivity and specificity of the marker-test on the power of the study must be considered to better adjust the sample size. Note that the study sample size is computed under the assumption that effect size should be larger in the marker population. If the marker-test is of poor specificity, a considerable proportion of marker-negative patients may be enrolled. False positive will not show the expected effect, possibly not benefiting at all from the new therapy, and this may decrease the power of the study.

2.1 Poisson Processes for Patient Recruitment in Enrichment Studies

This section presents a new strategy in modeling patient recruitment in enrichment studies. The main objective is an accurate prediction of the recruitment time at the initial stage before starting the recruitment process, as well as the remaining time in the ongoing stage. The simplest way to access prediction of the recruitment time in clinical trials is to use a deterministic approach.

Example 2.1 *Deterministic prediction of patient recruitment time*

a) ***Unselected study:*** *Let us consider a multicenter unselected clinical trial that requires a sample size of 200 patients. For 10 centers we estimate to recruit 8 eligible patients per week. Then, in the further definition of prevalence we can include this as well. By using the deterministic approach, the recruitment time is simply estimated by 200/8 = 25 weeks.*

b) ***Enrichment study:*** *Let us assume now that each eligible patient is tested according to a given biomarker and only those found as positive are recruited. By a marker prevalence of* 50%*, 4 patients are assumed to be marker-positive and then 50 weeks are required to recruit 200 biomarker positive patients.*

In the context of patient recruitment for clinical trials, deterministic approaches for study duration planning are very simple but less realistic than stochastic approaches.

The number of patients arriving at recruitment centers varies considerably during the recruitment process. It is well established that the amount of recruited patients per time unit follows a discrete probability distribution such as the Poisson distribution.

In enrichment trials, it is relevant to establish the impact of the marker prevalence and the sensitivity and specificity of the marker-test on the recruitment process. As mentioned in the introduction, an enrichment trial may differ from an unselected trial by the fact that a larger proportion of marker-positive patients are required. Without loss of generality, we assumed that only marker-test positive patients are enrolled for the trial. If the study population must consist of N^+ marker-positive and N^- marker-negative patients, with large N^+ (enriched), then the process of negative patients is also to be considered. This process will be obtained by changing the entry probability θ by $1-\theta$ in our models. In practice, eligible patients are tested until both N^- negative and N^+ positive patients are recruited. We stop the recruitment of, for example, test-negative patients once N^- is reached. Considering only the process of test-positive ($N^- = 0$) is not a restriction, since the processes of test-positive and test-negative patients only differ in the entry probabilities.

Proposition 2.1 *Thinning property of Poisson processes*
Let $\{N_t; t \in [0,\infty)\}$ be a Poisson process with intensity $\lambda(t)$. If each event or observation is kept with probability θ and rejected with probability $1-\theta$ independently from event to event, then the process of selected events (kept events) thus obtained is again a Poisson process. Its intensity is equal to the intensity of $\{N_t; t \in [0,\infty)\}$, decreased by a factor equal to the selection probability θ: $\lambda^+(t) = \theta\lambda(t)$ is the intensity of the selected process (Cont and Tankov [2004]).

Marker-positive patients are selected from the population of eligible patients with a probability equal to the probability of testing a patient as positive. We follow this idea and use Proposition 2.1 to define appropriate Poisson processes for patient recruitment in biomarker studies that consider the marker prevalence and the marker-test characteristics.

2.1.1 Formulation of the Recruitment Model

We consider a multi-center enrichment study, where patients are recruited in $K > 1$ centers. The process of unselected patients is denoted by $\{N_t, t \in [0,\infty)\}$ and $\{N_t^+, t \in [0,\infty)\}$ denotes the process of marker-test positive patients (selected process). Let the

intensities of these processes in center $k = 1, 2, \cdots, K$ be defined by $\lambda_t^{(k)} > 0$ and $\lambda_t^{(+k)} > 0$, respectively. The overall processes $\{N_t, t \in [0, \infty)\}$ and $\{N_t^+, t \in [0, \infty)\}$ are Poisson processes with intensities $\lambda_t = \sum_{k=1}^K \lambda_t^{(k)}$ and $\lambda_t^+ = \sum_{k=1}^K \lambda_t^{(+k)}$, respectively. This allows for more visibility in the formulas without loss of generality, since the analysis methods can be applied to centers separately.

Model without Marker-test Characteristics

Let θ be the marker prevalence and μ the proportion of patients screened as eligible. The screening and marker-test procedures are assumed to be independent. The probability of an unselected patient to be screened as eligible and tested as marker-test positive is given by $\mu\theta$. Including a patient or not is a Bernoulli experiment with probability of success $\mu\theta$. This corresponds to the marker prevalence θ in the absence of screening conditions ($\mu = 1$). By using the thinning property of Poisson processes, we derive the process of marker-test positive patients $\{N_t^+, t \in [0, \infty)\}$ as a Poisson process with intensity $\mu\theta\lambda_t$,

$$N_t^+ \sim Pois(\mu\theta\lambda_t). \tag{1}$$

For $\theta = \mu = 1$, the model (1) reduces to a traditional Poisson process for patient recruitment in an unselected clinical trials. This is only for theoretical considerations, since an enrichment study with very large prevalence is unnecessary. Very large prevalence means, there is no relevant difference between the marker-positive population and the unselected population. Therefore, it would not be worth the trouble to use such a strategy which would increase the study time and costs. Without loss of generality we assume that $\mu = 1$.

Model with Marker-test Characteristics

Recall that the marker-test characteristics are rarely available. If an estimate of the sensitivity and specificity of the marker-test is provided, the impact of misclassification on the recruitment process can be evaluated. The probability of an unselected patient to be tested as marker positive depends on the sensitivity and specificity of the marker-test. Let us consider the events: T^+ = "an unselected patients is tested as positive"; T^- = "an unselected patients is tested as negative"; P^+ = "the patient is actually marker-positive" and P^- be the event "the patient is actually marker-negative." $P(T^+)$ represents the probability of an unselected patient to be enrolled for an enrichment trial. The probability of a marker-positive patient to be tested as positive ($P^+ \cap T^+$) is given by

Bayes' formula

$$P(P^+ \cap T^+) = P(T^+|P^+) \cdot P(P^+),$$

where $P(P^+)$ represents the probability of being marker-positive in the unselected patient population (prevalence) and $P(T^+|P^+)$ denotes the sensitivity ($Sens$) of the marker-test. We derive the probability $P(T^+)$ by using Bayes' formula and the fact that T^+ is a disjoint union of $T^+ \cap P^+$ and $T^+ \cap P^-$. In other words,

$$\begin{aligned} P(T^+) &= P(T^+|P^+) \cdot P(P^+) + P(T^+|P^-) \cdot P(P^-) \\ &= P(T^+|P^+) \cdot P(P^+) + [1 - P(T^-|P^-)] \cdot [1 - P(P^+)] \\ &= Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta) \end{aligned}$$

Note that $P(T^-|P^-)$ represents the specificity ($Spec$) of the marker-test. $P(T^+)$ denotes the entry probability of patients in enrichment trial and corresponds to the marker prevalence if the sensitivity and specificity of the marker-test is assumed to be equal to 100%. We use the thinning property of Poisson processes given by proposition 2.1 and formulate the process as follows

$$N_t^+ \sim Pois\left(\left[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)\right]\lambda_t\right). \tag{2}$$

This is the general form of Poisson processes for patient recruitment in enrichment studies and corresponds to the traditional Poisson process through parametrization ($\theta = Sens = Spec = 1$). Different Poisson processes can be derived for different assumptions on the distribution of λ_t and θ. The recruitment process of test-negative patients can be derived from model (2) by

$$N_t^- \sim Pois\left(\left[1 - Sens \cdot \theta - (1 - Spec) \cdot (1 - \theta)\right]\lambda_t\right).$$

We consider the case where λ_t is constant as suggested by Carter [2004] in modeling recruitment process of unselected trials. We also investigate the case where λ_t is assumed to be Gamma distributed as investigated by Anisimov and Fedorov [2007] as well as when λ_t is differently defined in the different stages of the process (Tang et al. [2012]). Poisson processes with intensity depending on past observations may be a good alternative in modeling patient recruitment processes in clinical trials. The recruitment rate may not only depends on the randomness but may include information on its past values or past observations. We consider a Poisson-autoregressive process (Ferland et al. [2006],Fokianos et al. [2009]). We start by assuming that θ is constant. Then we investigate the case where θ follows a Beta distribution.

2.1.2 Model with constant Recruitment Rate

In this section, the arrival process of unselected patients at the recruitment center k is assumed to be a Poisson process with constant intensity (λ_k). We assumed that the centers start at the same time and focus on the impact of the marker prevalence on the process and the recruitment time. The overall process without considering the sensitivity and specificity of the marker-test is defined as

$$N_t^+ \sim Pois(\lambda\theta), \tag{3}$$

where $\sum_{k=1}^{K} \lambda_k = \lambda$ represents the overall arrival rate of unselected patients and K the number of centers. At the initial stage, λ should be provided by the experts, for example, on the basis of past studies. The process (3) is a homogeneous Poisson process with intensity $\lambda\theta$ (see Appendix A.1.1).

Recruitment Time

The recruitment time $T^+(n^+)$ required to enroll n^+ marker-positive patients is a sum of n^+ independent exponentially distributed random variables corresponding to the waiting time between jumps of the process, which follows an Erlang distribution as given in Appendix A.1.1. The distribution of $T^+(n^+)$ is given by

$$T^+(n^+) \sim Erlang(n^+, \theta\lambda), \tag{4}$$

with mean $E[T^+(n^+)] = n^+/\theta\lambda$ and variance $Var[T^+(n^+)] = n^+/(\theta\lambda)^2$. Let us consider an unselected trial with sample size $n_0 \geq n^+$ as a reference. The corresponding recruitment process of unselected patients is defined as $N_t \sim Pois(\lambda)$, and thus, the recruitment time needed to enroll n_0 patients is $T(n_0) \sim Erlang(n_0, \lambda)$. The comparison of the first and second moments of the distributions of recruitment time $T^+(n^+)$ and $T(n_0)$ leads to the following remark.

Remark 2.2
If $\theta \geq n^+/n_0$, the recruitment time of the enrichment study is expected to be less than or equal to that of the unselected study.
If $\theta < n^+/n_0$, the recruitment time of the enrichment study is expected to be longer.
There is a larger variability in the prediction of recruitment time in an enrichment study

than in an unselected study, since

$$Var[T^+(n^+)] = \frac{n^+}{[\theta \sum_{k=1}^{K} \lambda_k]^2} = \frac{n^+}{\theta^2[\sum_{k=1}^{K} \lambda_k]^2}$$
$$= \frac{Var[T(n^+)]}{\theta^2} \geq Var[T(n^+)]. \quad (\theta \leq 1)$$

The increasing variability is due to the additional randomness at the marker-test stage which does not exist in traditional trials. $Var[T(n^+)]$ *denotes the variance of the recruitment time needed to recruit* n^+ *unselected patients.*

Example 2.3

We consider a large phase III trial with sample size 600 marker-positive patients for visibility of figures. Let us assume an overall arrival rate of 10 unselected patients in the recruitment centers (per time unit such as week).

The distributions of the recruitment time required to enroll the 600 marker positive patients are represented in Figure 4 for different values of the prevalence. Figure 4

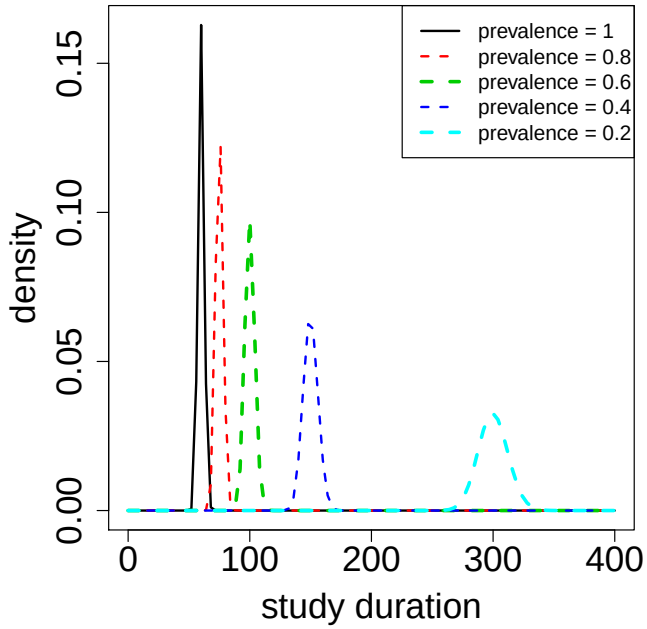

Figure 4: Density of the recruitment time of 600 marker positive patients for difference values of the marker prevalence given a recruitment rate of 10 patients in the overall recruitment centers.

illustrates the relationship between patient recruitment time and marker prevalence in an

enrichment study. The recruitment time increases when the marker prevalence decreases. The variability in the distribution of the recruitment time increases when the prevalence decreases as illustrated by the width of these distributions (see Figure 4). That means, the smaller the prevalence, the larger the uncertainty in predicting the recruitment time as compared to the unselected trial. Table 1 provides the mean of the recruitment time for different values of the prevalence and the corresponding 5^{th} and 95^{th} percentiles. According to Table 1, 300 time units are required to enroll 600 marker-positive patients

prevalence	recruitment time (mean)	5^{th} percentile	95^{th} percentile
0.1	600	560.28	640.85
0.2	300	280.14	320.42
0.3	200	186.76	213.61
0.4	150	140.07	160.21
0.5	120	112.057	128.17
0.6	100	93.38	106.80
0.7	86	80.04	91.55
0.8	75	70.03	80.10
0.9	67	62.25	71.20

Table 1: Patient recruitment time for different values of the marker prevalence assuming a constant arrival rate of unselected patients

if the marker prevalence is equal to 20% while for a marker prevalence of 70%, only 86 time units will be required to enroll the 600 patients. The marker prevalence is a crucial time component in planning an enrichment trial because one of the objectives of such a trial is to accelerate the drug development process.

2.1.3 Model with constant Recruitment Rate and Marker-test Characteristics

A Poisson recruitment process with a constant rate for patient recruitment in enrichment studies can be derived from model (2). The model is defined as

$$N_t^+ \sim Pois\Big(\Big[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)\Big]\lambda\Big). \tag{5}$$

The form of the distribution of the recruitment time is not affected by the sensitivity and specificity, which remains an Erlang distribution given by

$$T^+(n^+) \sim Erlang\Big(n^+, [Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda\Big). \tag{6}$$

This means $E[T^+(n^+)] = n^+/([Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda)$ and the variance $Var[T^+(n^+)] = n^+/([Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda)^2$. Let us consider example 2.3, where 600 marker-positive patients must be recruited. Figure 5 shows the variation of the recruitment time with the sensitivity and specificity of the marker-test. Each line on

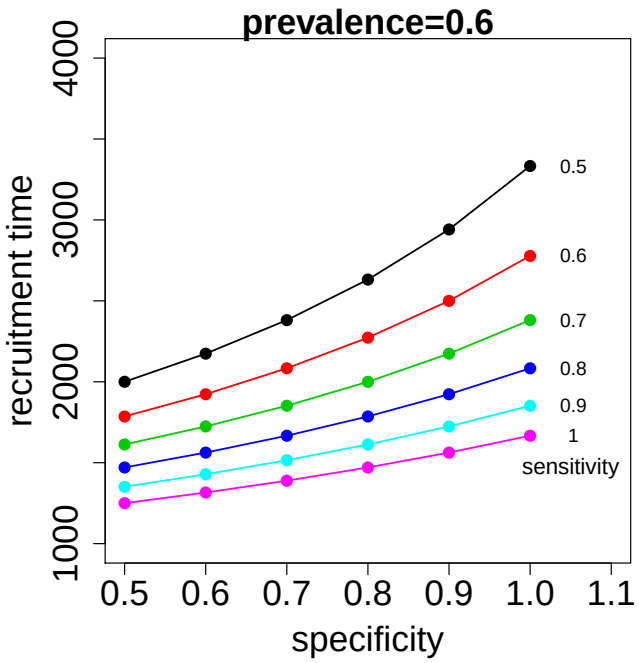

Figure 5: Mean of the recruitment time of 600 marker positive patients for different values of the sensitivity and specificity of the marker-test given a recruitment rate of 10 patients in the overall recruitment centers.

Figure 5 represents the variation of the recruitment time given one value of the sensitivity and different values of the specificity. The greater the sensitivity, the lower the recruitment time. However, the recruitment time increases when the specificity increases. Low sensitivity means some positive patients are erroneously tested as negative. This may increase the recruitment time. Similarly, for small specificity values, some false positive patients are recruited and the recruitment time decreases. Formula (5) helps in quantifying and visualizing the impact of the marker-test characteristics on the recruitment time if they are available. In practice, the entry probability is usually assumed to be

equal to θ instead of $Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)$, since the unknown misclassification errors are included in the estimate of θ.

Center Capacities

The importance of a given center can be given by the proportion of patients enrolled in each center compared to the total number of enrolled patients. The standardized weight of center k can be estimated as

$$\omega_k = [Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda_k / [Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda = \lambda_k / \lambda$$

as suggested by Anisimov and Fedorov [2007]. This is an estimate of the probability that a randomly selected patient comes from the center k. The probability that n_k^+ of the n^+ recruited patients come from center k is given by:

$$P(N_k^+ = n_k^+ | n^+) = \binom{n^+}{n_k^+} \omega_k^{n_k^+} (1 - \omega_k)^{n^+ - n_k^+}.$$

In other words, the vector $(N{+}_1, N_2^+, \cdots, N_K^+)$ follows a Multinomial distribution, where N_k^+ denotes a random variable representing the number of patients recruited in center k. The probability density of $(N_1^+, N_2^+, \cdots, N_K^+)$ is given by

$$P(N_1^+ = n_1^+, \cdots, N_k^+ = n_K^+) \frac{n^+!}{n_1^+! \cdots n_K^+!} \prod_{k=1}^{K} \omega_k^{n_k^+},$$

where $\sum_{i=1}^{K} N_i^+ = n^+$.

2.1.4 Models with Random Recruitment Rate

Poisson processes with Gamma distributed rates have been suggested by Anisimov and Fedorov [2007] for modeling of patient recruitment in unselected clinical trials. Mijoule et al. [2012] used instead the Pareto distribution for the recruitment rate. Pareto-Poisson may fit better than Gamma-Poisson when a large number of centers is of small capacity, as argued by the authors. However, the Pareto distribution is computationally harder to handle, which forces the use of Monte Carlo techniques to drop samples from target distributions. The authors recommend to use the simpler but very flexible Gamma-Poisson in most cases. The motivation of choosing the Gamma distribution is that the recruitment rate is a positive variable and this distribution offers flexibility in analysing

the model in a Bayesian framework. The process is defined as

$$N_t^+ \sim Pois\left(\sum_{k=1}^{K} \theta \lambda_t^{(k)}\right), \tag{7}$$
$$\lambda_t^{(k)} \sim \Gamma(a_k, b_k).$$

Without loss of generality, let us assume for simplicity that the scale parameters of the Gamma distributions are the same ($b_k = b,\ k = 1, 2, \cdots, K$). The overall intensity of the process sums to

$$\sum_{k=1}^{K} \theta \lambda_t^{(k)} \sim \Gamma(a, b/\theta),$$

where the shape parameter is given by $a = \sum_{k=1}^{K} a_k$. This property of the Gamma distribution shows that model (8) turns to a Gamma-Poisson model as investigated by Anisimov and Fedorov [2007]. The intensity of N_t^+ can also be written as $\lambda_t^{(+k)} \sim \Gamma(a_k, b/\theta)$. The restriction of equal scale parameters does not change the form of the distribution of the overall intensity, since the sum of independent random Gamma distributed variables remains Gamma distributed.

Recruitment Time

The recruitment time needed to enroll n^+ marker positive patients $T^+(n^+)$ is Erlang distributed with a Gamma distributed scale parameter: $T^+(n^+) \sim \Gamma[n^+, \Gamma(a, b/\theta)]$.
$T^+(n^+)$ is then Gamma-Gamma distributed:

$$T^+(n^+) \sim Gg(a, b/\theta, n^+).$$

This corresponds to a Pearson type VI distribution with a location parameter equal to zero Anisimov and Fedorov [2007]. The probability density of a Pearson type VI distribution $PearsonVI(a, b, s, \delta)$ is given by:

$$f(x) = \frac{1}{s\mathcal{B}(a, b)} \left(\frac{x - \delta}{s}\right)^{a-1} \left(1 + \frac{x - \delta}{s}\right)^{-a-b},$$

where $a > 0,\ b > 0,\ s > 0,\ \delta \geq 0$ and $\mathcal{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$. We deduce the density function of $T^+(n^+)$ by setting $\delta = 0$.

$$f(t) = \frac{1}{\frac{b}{\theta}\mathcal{B}(n^+,a)}\left(\frac{bt}{\theta}\right)^{n^+-1}\left(1+\frac{bt}{\theta}\right)^{-n^+-a},$$

$$E[T^+(n^+)] = \frac{n^+b}{\theta(a-1)} \quad \text{for } a > 1,$$

$$Var[T^+(n^+)] = \frac{b^2n^+(n^++a-1)}{\theta^2(a-1)^2(a-2)} \quad \text{for } a > 2.$$

We return to the above example 2.3, where the sample size of an enrichment study is supposed to be 600 marker positive patients. Now assume that the expected total number of unselected patients in all centers per time unit is Gamma distributed $\Gamma(100, 10)$ with mean $100/10 = 10$. Figure 6 presents the distribution of the recruitment time for different values of the prevalence. Figure 6 shows that the recruitment time increases when the

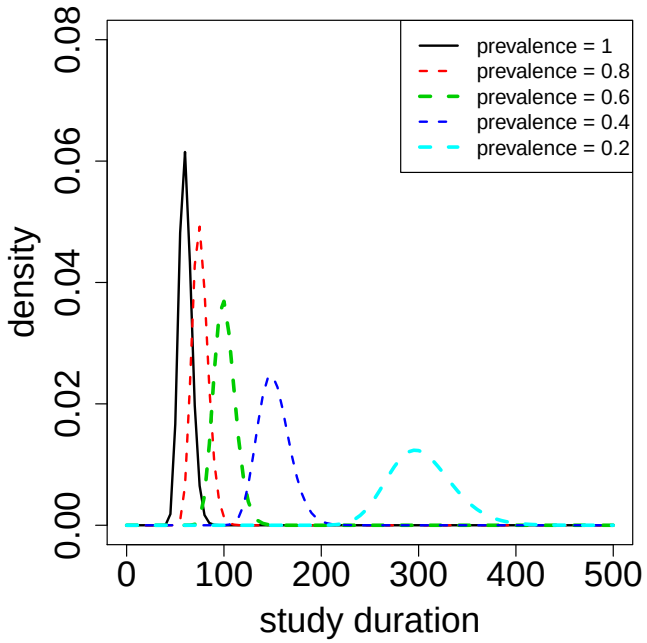

Figure 6: Density of the distribution of the recruitment time in a marker study of sample size $N^+ = 600$ given the overall recruitment rate $\lambda \sim \Gamma(100, 10)$.

marker prevalence decreases. It is also important to see that the variability increases in the distribution of the recruitment time when the prevalence decreases. Figure 6 compared to Figure 4 show a global increase of the variability when the the recruitment rate is a random variable instead of being a constant. As in the above section, the weight

of center k is given by : $\omega_k \sim \Gamma(a_k, b/\theta)/\Gamma(\sum_{i=1}^{K} a_i, b/\theta)$, which can also be written as $\omega_k \sim \Gamma(a_k, b/\theta)/[\Gamma(a_k, b/\theta) + \Gamma(a_k - \sum_{i=1}^{K} a_i, b/\theta)]$. ω_k is then Beta distributed $\omega_k \sim Beta(a_k, a_k - \sum_{i=1}^{K} a_i)$. Given n^+ recruited patients, the number of patients coming from center k is binomially distributed with parameter n^+ and ω_k, where $\omega_k \sim Beta(a_k, a_k - \sum_{i=1}^{K} a_i)$. By factorizing the scale parameter, we can see that the whole vector of the K weights is Dirichlet distributed:

$$\begin{aligned}
\omega_k &\sim \Gamma(a_k, b/\theta)/\Gamma(\sum_{k=1}^{K} a_k, b/\theta) \\
&\Rightarrow \omega_k \sim \Gamma(a_k, 1)/\Gamma(\sum_{k=1}^{K} a_k, 1) \\
&\Rightarrow (\omega_1, \omega_2, \cdots, \omega_K) \sim Dirich(a_1, a_2, \cdots, a_K)
\end{aligned}$$

The vector $(N_1^+, N_2^+, \cdots, N_K^+)$ is multinomially distributed with parameters n^+ and $(\omega_1, \omega_2, \cdots, \omega_K)$.

$$P(N_1^+ = n_1^+, \cdots, N_k^+ = n_K^+) = \frac{n^+!}{n_1^+! \cdots n_K^+!} \frac{\Gamma(\sum_{k=1}^{K} a_k) \prod_{k=1}^{K} \Gamma(a_k + n_k^+)}{\prod_{k=1}^{K} \Gamma(a_k) \Gamma(\sum_{k=1}^{K} a_k + n^+)},$$

where $\sum_{i=1}^{K} N_i^+ = n^+$.

Models with Random Recruitment Rate and Marker-test Characteristics

Here, only the probability of testing a patient as marker-positive changes in the above analysis by considering the sensitivity and specificity of the marker-test. The model is given by

$$\begin{aligned}
N_t^+ &\sim Pois\left(\sum_{k=1}^{K} [Sens \cdot \theta + (1 - Spec)(1 - \theta)] \lambda_t^{(k)}\right), \\
\lambda_t^{(k)} &\sim \Gamma(a_k, b_k).
\end{aligned}$$

Additional recruitment time graphics and formulas can obtained by replacing θ by $Sens \cdot \theta + (1 - Spec)(1 - \theta)$ in the formulas in Section 2.1.4. For Example 2.3, the recruitment time is represented by the following graphics for some combinations of the sensitivity and specificity. The impact of the test characteristics on the distribution of the recruitment time is illustrated by Figure 7. The two graphs on the top show that a decrease in the sensitivity for constant specificity leads to an increase in the recruitment time and its variability. A small specificity value leads to lower recruitment time as illustrated by the two graphs on the right side. This is due to the fact that many false positive are recruited.

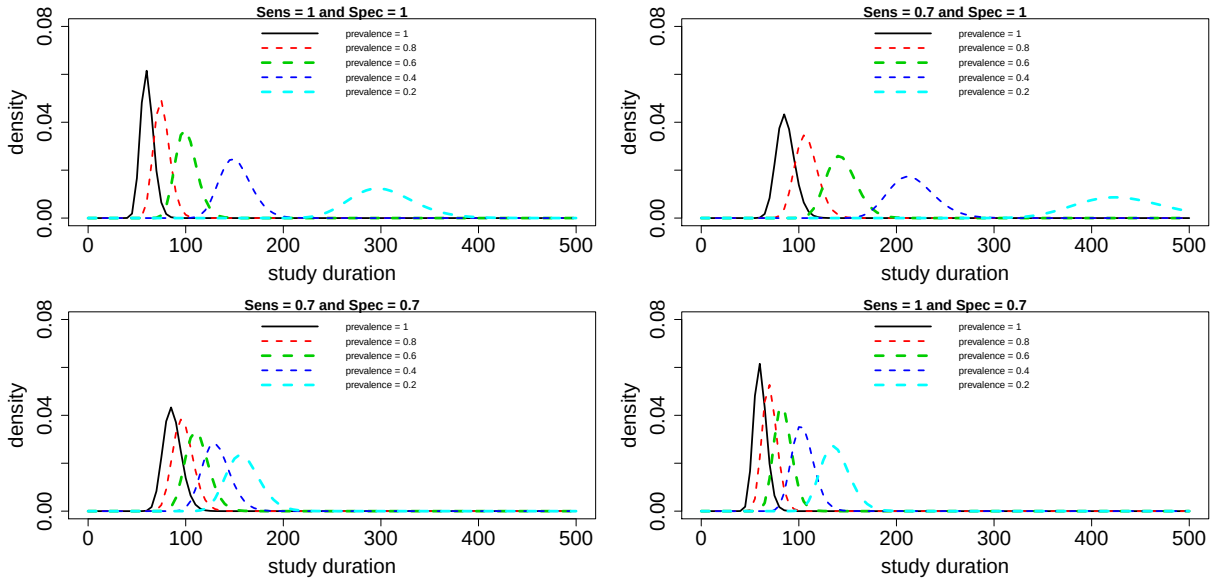

Figure 7: Density of the distribution of the recruitment time in the marker study given by Example 2.3 plotted for $Sens = 1$ $Spec = 1$; $Sens = 0.7$ $Spec = 1$; $Sens = 0.7$ $Spec = 0.7$; $Sens = 1$ $Spec = 0.7$, respectively and $\lambda_t \sim \Gamma(100, 10)$.

2.1.5 Autoregressive Models for Enrichment Studies

This section focuses on the investigation of integer GARCH models (see Appendix A.1.2) in the context of patient recruitment in enrichment studies. INGARCH models are special cases of Poisson autoregressive models. Inspired by the GARCH model (Bollerslev [1986]), Ferland et al. [2006] introduced INGARCH models in modeling the number of new infections from campylobacterosis in Canada as represented in Figure 8. At the time point 100, there is an outlier in the process. After this outlier, the process does not turn black immediately to its normal progression. In addition, there is a level change and then the rate does not vary completely at random. INGARCH models encourage a feedback mechanism so that the intensity of the process varies not only at random but also depends on past information. There is growing interest in INGARCH processes for the modeling of time series of counts. Fokianos et al. [2009] investigated Poisson linear and non-linear auto-regressive models. In these non-linear cases, the intensity of the process depends on its past values and the past observations through a non-linear function. Fokianos and Fried [2010] investigated outliers and level shift in such processes. However, these models in their original formulation are not designed for patient recruitment in enrichment studies. We redefine these models by considering the

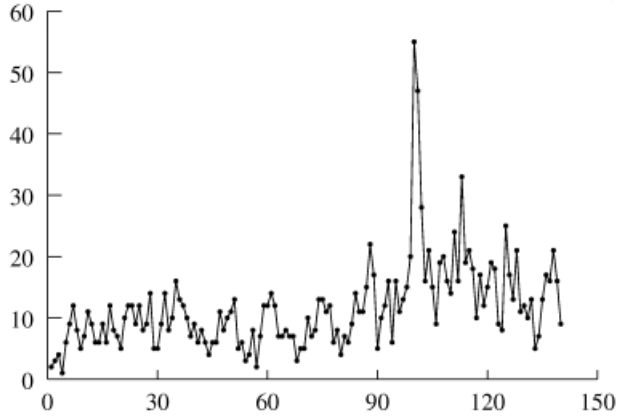

Figure 8: Number of cases of campylobacterosis infections from January 1990 to the end of October 2000. The series was reported every 28 days (13 times a year). *Source* Ferland et al. [2006]

marker prevalence and the marker-test characteristics.

$$N_t^+|\mathcal{F}_{t-1} \sim Pois\Big(\Big[Sens \cdot \theta + (1 - Spec)(1 - \theta)\Big]\lambda_t\Big),$$
$$\lambda_t = \beta_0 + \sum_{i=1}^{p} \beta_i B^i N_t + \sum_{j=1}^{q} \alpha_j B^j \lambda_t,$$

where B denotes the backward shift operator ($B^1 N_t = N_{t-1}$) and $\beta_0, \beta_1 \cdots \beta_p, \alpha_1, \cdots, \alpha_q \geq 0$. The stationarity condition can be deduced from that of the original model as stated by Ferland et al. [2006] (see Appendix A.1.2):

$$\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1/[Sens \cdot \theta + (1 - Spec)(1 - \theta)] \tag{8}$$

We focus on an INGARCH(1,1) process for $p = q = 1$.

$$N_t^+|\mathcal{F}_{t-1} \sim Pois\Big(\Big[Sens \cdot \theta + (1 - Spec)(1 - \theta)\Big]\lambda_t\Big),$$
$$\lambda_t = \beta_0 + \beta_1 N_{t-1}^+ + \alpha_1 \lambda_{t-1}$$

It is quite complicated to find an analytic distribution for the recruitment time by assuming an autoregressive processes. Thus, one approach is to simulate it by starting with fixed λ_0 and N_0^+.

Recruitment Time

Recall Example 2.3 and assume $\lambda_0 = 10$; $N_0^+ = 5$. We simulate the recruitment time for a marker prevalence of 60%, $\beta_0 = 2$, $\alpha_1 = 0.3$, $\beta_1 = 1$ and $Sens = Spec = 1$. Figure

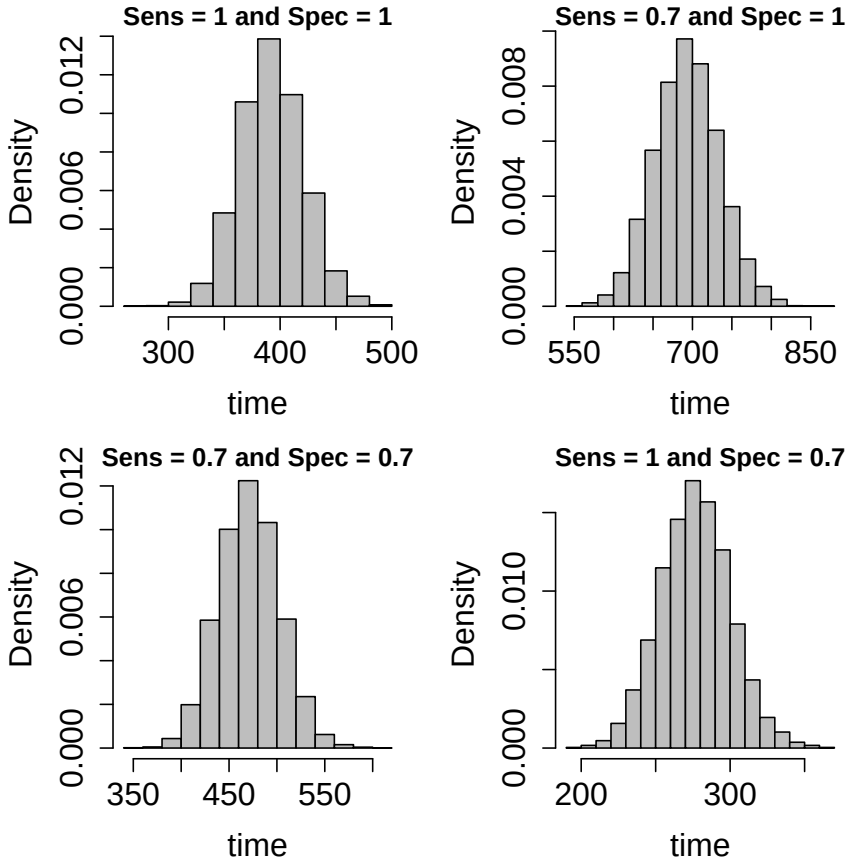

Figure 9: Simulated recruitment time in a marker study of sample size $N^+ = 600$ under an INGARCH(1,1) model with parameters $\beta_0 = 2$, $\alpha_1 = 0.3$ and $\beta_1 = 1$ and given an initial rate of $\lambda_0 = 10$ patients. The marker prevalence is chosen to 60%.

9 shows the simulated recruitment time under model (9). The impact of the sensitivity and specificity on the distribution of the recruitment time is illustrated by the different histograms corresponding to various combinations of the sensitivity and specificity of the marker-test.

2.1.6 Models with Change Points

The patient recruitment process in clinical trials may progress differently during the recruitment time (Tang et al. [2012]). Three main phases are often identifiable: The recruitment rate increases continuously at the beginning, since all centers do not start at the same time, but rather progressively. Another reason is an increase in information and advertisement. In this recruitment stage, recruitment rate depends on time. After all centers are initialized, the process reaches a relatively stationarity phase and varies about

a fixed mean. In the last phase, the recruitment rate may increase or decreases until the closure date. Figure (10) presents a simulated example of such a recruitment process. The assumptions on the recruitment rate as proposed Tang et al. [2012] present some

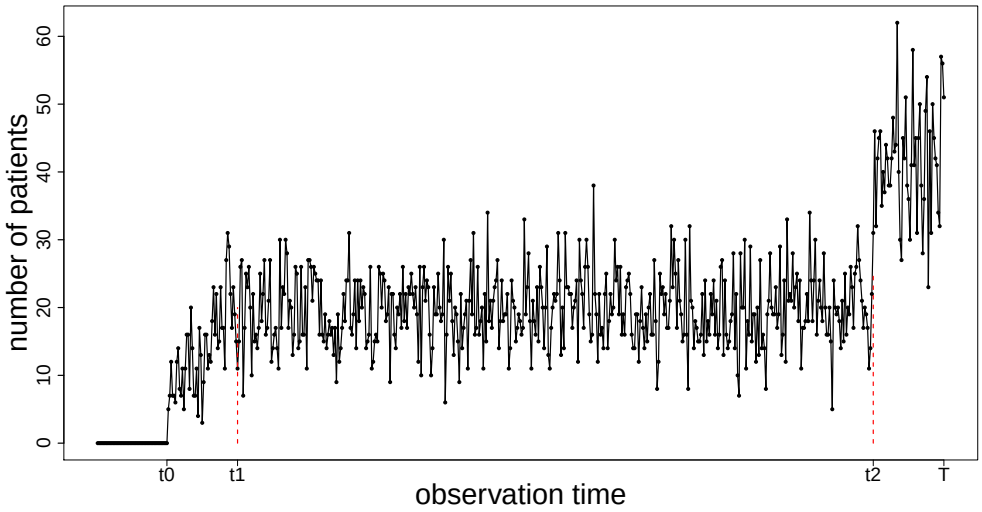

Figure 10: Overall number of patients arriving at recruitment centers. Three different phases are considered depending on the intensity of the process

weaknesses as mentioned by Anisimov [2012]: The recruitment rate at the first phase has not been explicitly proposed and a homogenous Poisson process has been assumed in the stable phase as suggested previously by Carter [2004], although patient recruitment process are often over dispersed. We follow the idea of assuming a multi-phase process an propose a Poisson process for patient recruitment in enrichment studies. The recruitment rate at the start phase is defined as a monotone increasing function of time. For the remaining two phases we investigate two cases: constant rate and Gamma distributed rate.

Let us assume that the first center is initialized and recruits the first patient at t_0. The overall recruitment rate increases from t_0 to t_1; from t_1 on, the process become stable for a substantial period. Suppose an eventual level shift may occur at $t_2 > T$, where T is the closure date. This phenomenon is illustrated as s simulated processes shown in Figure 10. t_1 and t_2 have been called change points, where t_2 represents the

point of level shift. The general definition of our model is given by:

$$N_t^+ \sim Pois\left(\left[Sens \cdot \theta + (1 - Spec)(1 - \theta)\right]\lambda_t\right) \tag{9}$$
$$\lambda_t = I_{(t_0 \le t < t_1)}\lambda_t^{(1)} + I_{(\tau_1 \le t < \tau_2)}\lambda_t^{(2)} + [\lambda_t^{(2)} + \delta]I_{(t \ge t_2)},$$

where $\lambda_t^{(1)}$ denotes the expected number of unselected patients arriving at all centers at the start phase, $\lambda_t^{(2)}$ denotes the expected number of unselected patients at stable phase and $\delta \in [0, \infty)$ is the level shift parameter representing the jump in the recruitment rate at last phase of the process. I is the indicator function taking the value 1 if the condition inside the brackets is satisfied and 0 otherwise. Model (10) is a general formulation of a patient recruitment model in an enrichment study under consideration of a level change in the process. It corresponds to model (1), when $t_0 = t_1$ and $\delta = 0$.

Model with Level Change and constant Rate The recruitment process with change points and constant recruitment rate at stable phase can be defined as:

$$N_t^+ \sim Pois\left(\left[Sens \cdot \theta + (1 - Spec)(1 - \theta)\right]\lambda_t\right)$$
$$\lambda_t = I_{(t_0 \le t < t_1)}(A + Bt) + I_{(t_1 \le t < t_2)}(A + Bt_1) + [A + Bt_1 + \delta]I_{(t \ge t_2)}$$

where $A \in [0, \infty)$ and $B \in [0, \infty)$. Here it is assumed that the recruitment rate at the end of the first $(A + Bt_1)$ remains constant in the second phase and increases in δ in the last phase.

Model with Level Change and Gamma distributed Rate Processes with Gamma distributed recruitment rate at the stable phase are defined as:

$$N_t^+ \sim Pois\left(\left[Sens \cdot \theta + (1 - Spec)(1 - \theta)\right]\lambda_t\right)$$
$$\lambda_t = I_{(t_0 \le t < t_1)}(A + Bt) + I_{(t_1 \le t < t_2)}\lambda_t^{(2)} + [\lambda_t^{(2)} + \delta]I_{(t \ge t_2)} \tag{10}$$
$$\lambda_t^{(2)} \sim \Gamma(a, b).$$

If estimates of t_1 and t_2 are available at the initial stage of the recruitment, then only the remaining time from t_2 to the end of the recruitment process is to be estimated. A practical use of these last two models would be to assume the same intensity in the different time intervals at the initial stage as presented in the previous sections to access the prediction of the recruitment time. After observing some data t_1 and t_2 can be detected and the model adjusted to improve the prediction of the closure time. We propose a Bayesian approach to detect the change points t_1 and t_2.

2.2 Parameters Update in Bayesian Framework

Bayesian methods have been getting more popular for statistical inference in pharmaceutical industry and particulary for the statistical analysis of clinical trials. Classical methods in evaluating the treatment effect of a new drug essentially test the hypothesis that the treatment effect difference is equal to zero and provide its point and interval estimates. Bayesian approaches supplement this by focusing on how to change our opinion about the treatment effect (Berry [2006]). Our final opinion (posterior information) is a summary of the initial information before carrying out the trial (prior information) and information obtained from the trial results (likelihood). Simon [1999] suggested a Bayesian approach in designing and analyzing active control clinical trials and estimated the posteriori probability of the new drug to be superior to the placebo. Bayesian methods have been used in constructing adaptive and sequential designs (Wathen and Thall [2008], Chen et al. [2010], Thall and Wathen [2007]). They may offer more flexibility if information from earlier drug development phases are to be considered in the later phases.

The drawback of Bayesian methods, to require a huge computational platform has been remedied with the new generation of high performance computers. Bayesian approaches offer the possibility of incorporating the expert information, information from past studies and experience in the form of priors in the inference. This section presents a way of updating the parameters of the models presented in the previous sections. Bayesian methods provide a very flexible way of progressively updating the parameters and the prediction of the recruitment time using the observations up to the evaluation time. The prior distribution is combined with the distribution of the observed data given the parameters (likelihood) to obtain the posterior distribution of the parameters and the whole distribution's parameters are derived. The literature on Bayesian statistics is very large (see Hoff [2009] for an introduction to Bayesian methods).

Background

Let $\mathcal{N}$ be the set of all possible observations (sample space) and $n \in \mathcal{N}$ be a samples vector representing the observed process until τ: $n = (n_1, n_2, \cdots, n_\tau)$. Consider a sample model with parameter vector η which is from the set Θ, the set of all possible parameters (parameter space). Let the prior distribution of η be denoted by $p(\eta)$ and the sample distribution be given by $p(n|\eta)$ (likelihood). The posterior distribution $p(\eta|n)$ is given

by the Bayes's formula

$$p(\eta|n) = \frac{p(n|\eta)p(\eta)}{\int_\Theta p(n|\widetilde{\eta})p(\widetilde{\eta})d\widetilde{\eta}} \propto p(n|\eta)p(\eta).$$

In some cases, the posterior can be explicitly derived and leads to a known probability distribution of the same class as the prior distribution. In such cases, the prior is called a conjugate prior to the sample distribution. The Gamma distribution is a conjugate prior for a Poisson sample distribution, similar to the beta prior in Binomial model as well as the Dirichlet prior in the Multinomial model, among others. In the absence of conjugacy, the posterior distribution cannot be explicitly derived and Markov chain Monte Carlo methods are used to generate a sequence $\eta^{(1)}, \eta^{(2)}, \cdots, \eta^{(T)}$ from the posterior. The moments of the posterior are estimated using the Monte Carlo approximation defined by

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} g(\eta^{(t)}) = \int g(\eta)p(\eta|n)d\eta,$$

for many functions g. The Metropolis-Hastings algorithm introduced by Metropolis et al. [2004] can be used to generate such a sequence. It is often combined with the Gibbs sampler introduced by Geman and Geman [1993], when η is a vector of many parameters. The Gibbs sampler generates a vector $\eta^{(t)}$ sequentially componentwise or a block after another block. Each complete cycle through the components of the vector constitutes one step in a Markov chain whose stationary distribution is, under suitable conditions, the distribution to be simulated see Casella and George [1992] for more details. Raftery and Lewis [1992] derived the number of iterations required for accurate estimates based on simulated samples from the posterior distribution.

2.2.1 Analysis of the Model with constant Recruitment Rate

The test of new patients does not provide information that can be used to update the sensitivity and specificity of the marker-test. Only if the real status of patients to be tested is known, can a confusion matrix be constructed after the test. That is not the case in selecting the study population for enrichment studies. Under the model with constant recruitment rate (see model (5)), if the marker prevalence θ is considered as constant during the recruitment process, then the posterior distribution of λ is a Gamma distribution given a Gamma prior. The prevalence may be reasonably kept as constant if its present estimation is based on a very large sample size compared to the number patients to be tested in the new study. Let $(n_1^+, n_2^+, \cdots, n_\tau^+)$ be the recruited

biomarker-test-positive patients until τ. We update the recruitment model (5) in a Bayesian framework as follows:

$$N_t^+ \sim Pois\Big([Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda\Big)$$
$$\text{a priori} \quad \lambda \sim \Gamma(a_\lambda, b_\lambda)$$

Since $Sens, Spec$ and θ are assumed to be constant, the posterior distribution of λ is given by $(\lambda|n_1^+, n_2^+, \cdots, n_\tau^+) \sim \Gamma(a_\lambda + n^+(\tau), (b_\lambda + \tau)[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)])$. Although considering the prevalence as constant simplifies the model, it leads to loss of new information about the model since the marker-test is designed to reflect the composition (marker-positive and marker-negative) of the unselected patient population as long as the test characteristics are large enough. We also study the case where θ is considered as a model parameter following a Beta distributed a priori. Note that the total number of marker positive patients recruited until τ sums to $n^+(\tau)$ and follows a Binomial distribution $n^+(\tau) \sim Bin(n(\tau), [Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)])$ given $n(\tau)$ (number of tested patients) and θ. The priors of both parameters of model (5) are given by

$$\text{a priori} \quad \lambda \sim \Gamma(a_\lambda, b_\lambda)$$
$$\text{a priori} \quad \theta \sim Beta(a_\theta, b_\theta).$$

The likelihood function $p(n_1^+, n_2^+, \cdots, n_\tau^+|\theta, \lambda)$ is proportional to

$$\prod_{i=1}^{\tau} \Big([Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda\Big)^{n_i^+} e^{(-[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda)}$$
$$= \Big([Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda\Big)^{n^+(\tau)} e^{(-\tau[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda)}.$$

The two model parameters (θ, λ) can be updated componentwise according to algorithm 2.4. N samples of $\eta = (\theta, \lambda)$ are generated using Gibbs sampling as briefly described in the following paragraph:

Algorithm 2.4

1- Choose $\eta^{(0)}$. *Giving* $\eta^{(i)} = (\theta^{(i)}, \lambda^{(i)})$, *repeat*

2- Generate $\lambda^{(i+1)} \sim p(\lambda|\theta^{(i)}, n_1^+, n_2^+, \cdots, n_\tau^+)$ *direct sampling*

3- Generate $\theta^{(i+1)} \sim p(\theta|\lambda^{(i+1)}, n_1^+, n_2^+, \cdots, n_\tau^+)$ *Metropolis-Hastings algorithm (MH)*

4- Set $\eta^{(i+1)} = (\theta^{(i+1)}, \lambda^{(i+1)})$

4- $i = i + 1$

Until $i = N$.

Remark 2.5

The full conditional distribution of θ $p(\theta|\lambda, n_1^+, n_2^+, \cdots, n_\tau^+)$ *is proportional to*

$$\left([Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]\lambda\right)^{n^+(\tau)} e^{(-\tau[Sens \cdot \theta + (1-Spec)\cdot(1-\theta)]\lambda)} \theta^{a_\theta} (1 - \theta)^{b_\theta - 1},$$

which does not match with a known distribution. Samples can be drooped from this distribution by using the MH algorithm. The full conditional distribution $p(\lambda|\theta, n_1^+, n_2^+, \cdots, n_\tau^+)$ *of* λ *is proportional to the Gamma distribution* $\Gamma(n^+(\tau) + a_\tau, \tau[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)] + b_\lambda)$.

For visibility of our equations, we concentrate on the analysis of the models without test characteristics ($Sens = Spec = 1$), since the sensitivity and specificity remain constant in the models and do not fundamentally change the analysis methodology. In this case, full conditional distributions derived in the situations above take the same form if the $Sens$ and $Spec$ were set to 1.

2.2.2 Gamma-Poisson Model

This section presents a Bayesian analysis of model (8). A Gamma prior is assumed for a and b and a beta prior for θ. The model with the different priors can be rewritten as follows:

$$\begin{aligned} n_t^+ &\sim Pois(\theta\lambda_t), \\ \lambda_t &\sim \Gamma(a, b), \\ \text{a priori} \quad \theta &\sim Beta(a_\theta, b_\theta), \\ \text{a priori} \quad a &\sim \Gamma(\alpha_a, \beta_a), \\ \text{a priori} \quad b &\sim \Gamma(\alpha_b, \beta_b), \end{aligned}$$

where $a_\theta, b_\theta, \alpha_a, \beta_a, \alpha_b, \beta_b$ are constant in the model. The posteriors for θ, a, b and $\lambda_t, i = 1, \cdots, \tau$ are to be estimated.

Remark 2.6

The full conditional distribution of λ_t *is*

$$\begin{aligned} p(\lambda_t|n_1^+, n_2^+, \cdots, n_\tau^+, \lambda_{-t}, a, b, \theta) &\propto \prod_{t=1}^{\tau} (\theta\lambda_t)^{n_t^+} \exp(-\theta\lambda_t)\lambda_t^{a-1} \exp(-\lambda_t b) \\ &\propto \lambda_t^{n_t^+ + a - 1} \exp\left[-\lambda_t(\theta + b)\right] \\ &= \Gamma(n_t^+ + a, \theta + b), \end{aligned}$$

where $\lambda_{-t} = (\lambda_1, \cdots \lambda_{t-1}, \lambda_{t+1}, \cdots \lambda_\tau)$.

Full conditional distribution of a *can be derived using the sample distribution of* λs, $p(\lambda_1, \cdots, \lambda_\tau|a, b) \propto \prod_{t=1}^{\tau} \lambda^{a_\lambda - 1} \exp(-\lambda_t b_\lambda)$

$$\begin{aligned} p(a|\lambda_1, \lambda_2, \cdots, \lambda_\tau, b, \theta) &\propto \prod_{t=1}^{\tau} \lambda_t^{a-1} \exp(-\lambda_t b) a^{\alpha_a - 1} \exp(-a\beta_a) \\ &\propto \prod_{t=1}^{\tau} \lambda_t^{\alpha_a - 1} \exp(-a\beta_a). \end{aligned}$$

The full conditional distribution of b *is*

$$\begin{aligned} p(b|\lambda_1, \lambda_2, \cdots, \lambda_\tau, a, \theta) &\propto \prod_{t=1}^{\tau} \exp(-\lambda_t b) b^{\alpha_b - 1} \exp(-b\beta_b) \\ &\propto \prod_{t=1}^{\tau} b^{\alpha_b - 1} \exp -(\lambda_t + \beta_b) \\ &= \Gamma(\alpha_b, \sum_{t=1}^{\tau} \lambda_t + \beta_b). \end{aligned}$$

The full conditional distribution of θ *is*

$$\begin{aligned} p(\theta|n_1^+, n_2^+, \cdots, n_\tau^+, \lambda_1, \lambda_2, \cdots, \lambda_\tau, a, b) &\propto \prod_{t=1}^{\tau} (\theta\lambda_i)^{n_i^+} \exp(-\theta\lambda_i) \theta^{a_\theta} (1-\theta)^{b_\theta - 1} \\ &\theta^{a_\theta + \sum_{i=1}^{\tau} n_i^+} \exp(-\theta \sum_{i=1}^{\tau} \lambda_i)(1-\theta)^{b_\theta - 1}. \end{aligned}$$

We could not find known distributions which match the above distributions. The MH algorithm is used to sample from these distributions.

2.2.3 Integer GARCH Model

In the previous section, the INGARCH(1,1) model has been defined for patient recruitment in enrichment studies,

$$\begin{aligned} N_t^+|\mathcal{F}_{t-1} &\sim Pois(\theta\lambda_t) \\ \lambda_t &= \beta_0 + \beta_1 N_{t-1}^+ + \alpha_1 \lambda_{t-1}. \end{aligned}$$

The traditional INGARCH(1,1) corresponding to $\theta = 1$ has been analyzed in the Bayesian context by Fried et al. [2013]. We update the model parameters $(\beta_0, \beta_1, \alpha_1, \theta)$ componentwise by assuming a Gamma prior for β_0, and a beta prior for θ. (β_1, α_1) should be updated together to better control the stationarity condition $\beta_1 + \alpha_1 < 1/\theta$. We use a three dimensional Dirichlet (Dirich) distribution for (β_1, α_1) in which the third component α_2 is a control parameter which ensures that the three components sum up to 1. That means $(\beta_1 + \alpha_1) < 1$ as first two components and then $\beta_1 + \alpha_1 < 1/\theta$, since $1/\theta \geq 1$. The full model can be rewritten as follows

$$\begin{aligned} N_t^+ | \mathcal{F}_{t-1} &\sim Pois(\theta\lambda_t) \\ \lambda_t &= \beta_0 + \beta_1 N_t^+ + \alpha_1 \lambda_{t-1} \\ \text{a priori} \quad \beta_0 &\sim \Gamma(a_{\beta_0}, b_{\beta_0}) \\ \text{a priori} \quad (\beta_1, \alpha_1, \alpha_2) &\sim Dirich(a_{\beta_1}, a_{\alpha_1}, a_{\alpha_2}) \\ \text{a priori} \quad \theta &\sim Beta(a_\theta, b_\theta). \end{aligned}$$

The full conditional distribution of β_0 is

$$\begin{aligned} p(\beta_0 | n_1^+, \cdots, n_\tau^+, \theta, \beta_1, \alpha_1) &\sim \beta_0^{a_{\beta_0}-1} e^{-\beta_0 b_{\beta_0}} e^{-\theta \sum_{t=1}^{\tau} \lambda_t} \prod_{t=1}^{\tau} (\theta\lambda_t)^{n_t^+} \\ &\propto \beta_0^{a_{\beta_0}-1} e^{(-\beta_0 b_{\beta_0} - \theta \sum_{t=1}^{\tau} \lambda_t)} \prod_{t=1}^{\tau} \lambda_t^{n_t^+}. \end{aligned}$$

The full conditional distribution of β_1, α_1 is

$$\begin{aligned} p(\beta_1, \alpha_1 | n_1^+, \cdots, n_\tau^+, \theta, \beta_0) &\sim \beta_1^{a_{\beta_1}-1} \alpha_1^{a_{\alpha_1}-1} \alpha_2^{a_{\alpha_2}-1} e^{-\theta \sum_{t=1}^{\tau} \lambda_t} \prod_{t=1}^{\tau} (\theta\lambda_t)^{n_t^+} \\ &\beta_1^{a_{\beta_1}-1} \alpha_1^{a_{\alpha_1}-1} e^{-\theta \sum_{t=1}^{\tau} \lambda_t} \prod_{t=1}^{\tau} (\lambda_t)^{n_t^+}. \end{aligned}$$

The full conditional distribution of θ is

$$\begin{aligned} p(\theta | n_1^+, \cdots, n_\tau^+, \theta, \beta_0, \beta_1, \alpha_1) &\propto \theta^{a_\theta} (1-\theta)^{b_\theta - 1} e^{-\theta \sum_{t=1}^{\tau} \lambda_t} \prod_{t=1}^{\tau} (\theta\lambda_t)^{n_t^+} \\ &\propto \theta^{a_\theta + n^+(\tau)} (1-\theta)^{b_\theta - 1} e^{-\theta \sum_{t=1}^{\tau} \lambda_t}. \end{aligned}$$

2.2.4 Change Point Detection in Bayesian Framework

This section presents A Bayesian analysis of the two models that explicitly consider the different phases of the recruitment process in an enrichment study:

$$\begin{aligned} N_t^+ &\sim Pois(\theta\lambda_t) \\ \lambda_t &= I_{(t_0\le t<t_1)}(A+Bt) + I_{(t_1\le t<t_2)}(A+Bt_1) + [A+Bt_1+\delta]I_{(t\ge t_2)}, \end{aligned}$$

and

$$\begin{aligned} N_t^+ &\sim Pois(\theta\lambda_t) \\ \lambda_t &= I_{(t_0\le t<t_1)}(A+Bt) + I_{(t_1\le t<t_2)}\lambda_t^{(2)} + [\lambda_t^{(2)}+\delta]I_{(t\ge t_2)} \\ \lambda_t^{(2)} &\sim \Gamma(a,b). \end{aligned}$$

Intuitively, informative priors for t_1 and t_2 can be derived as follows: Estimate the recruitment time T in study planning stage and assume $t_1 \sim U(t_0, T/4)$ and $t_2 \sim U(\frac{3}{4}T, T)$ a priori. The idea behind these assumptions is that all centers will likely initialize in the first time interval $[t_0, T/4]$, which may then contain t_1. Similarly, an eventual level shift would occur around the closure date. We assume the following priors for the other parameters: $A \sim \Gamma(\alpha_A, \beta_A), B \sim \Gamma(\alpha_B, \beta_B), a \sim \Gamma(\alpha_a, \beta_a), b \sim \Gamma(\alpha_b, \beta_b)$, $\theta \sim Beta(a_\theta, b_\theta)$ and $\delta \sim \Gamma(a_\delta, b_\delta)$. These models have been implemented in OpenBUGS 3.2.2 (see http://www.openbugs.info/w/). The following example illustrates the change point detection.

The process with constant rate is simulated by using the parameter $t_1 = 15,\ t_2 = 90, A = B = 1, \theta = 0.6$ and $\delta = 10$. Figure 11 represents 100 simulated observations with changes points at $t_1 = 15$ and $t_2 = 90$. The posterior means of these parameters are given in Table 2.

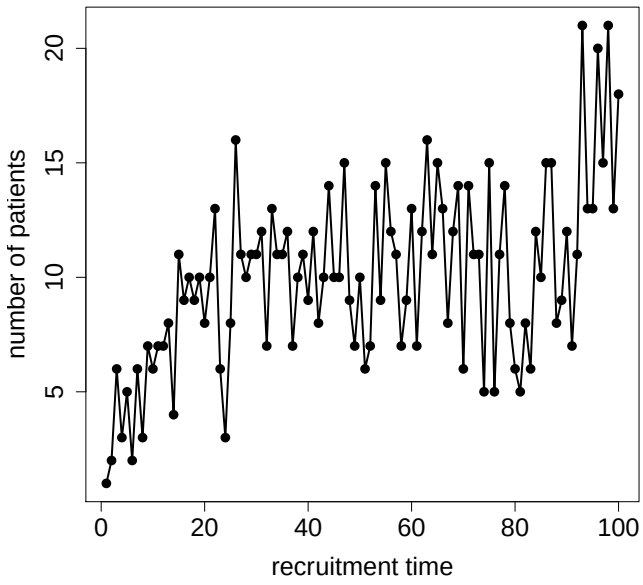

Figure 11: Simulated process with level shift and constant rate at the stable phase.

	True parameters	mean	sd	MC error	val2.5pc	median	val97.5pc
t_1	15	14.27	2.18	0.0541	10.34	14.24	19.9
t_2	90	89.17	3.545	0.09873	79.35	89.59	97.9
θ	0.6	0.6425	0.2386	0.006523	0.196	0.6523	0.985
δ	10	4.771	3.637	0.09387	8.943E-5	4.451	14.39
A	1	0.3515	0.862	0.02261	5.701E-7	0.01237	2.632
B	1	1.305	0.7838	0.02232	0.5848	1.078	3.415

Table 2: Change points detection and parameter estimates in Bayesian framework in a Poisson multi-phases model with constant rate at the stable phase (time in a time unit such as day).

As an additional example, let us consider the process with level changes and a Gamma distributed rate at stable phase. We simulate this process by using the parameter $t_1 = 15,\ t_2 = 90,\ A = B = 1,\ a = 16,\ b = 5,\ \theta = 0.6$ and $\delta = 10$. Figure 12 represents 100 simulated observations. The predicted posterior means are given in Table 3. A larger variability can be observed, compared to the results on Figure 11. The change points were set to $t_1 = 15$ and $t_2 = 90$ and the detected values are 17.25 and 87.27, respectively.

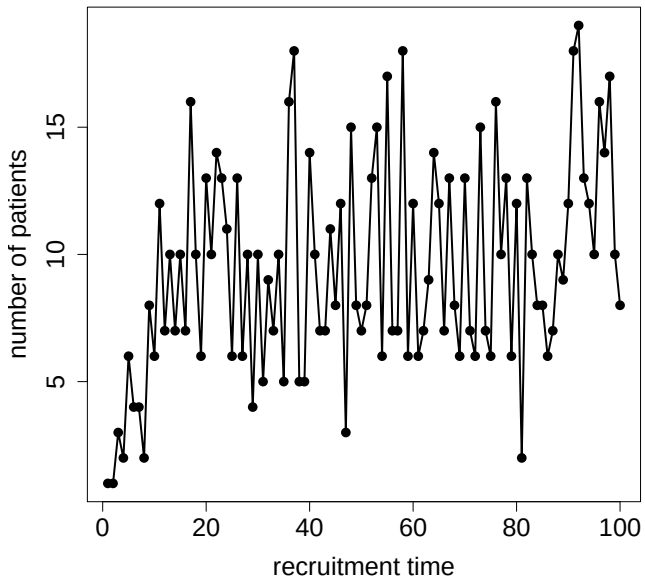

Figure 12: Simulated process with level shift and Gamma distributed rate at the stable phase.

	True parameters	mean	sd	MC error	val2.5pc	median	val97.5pc
t_1	15	17.25	5.269	0.1405	8.357	17.31	24.69
t_2	90	87.27	5.138	0.136	76.48	88.26	98.23
θ	0.6	0.5919	0.2314	0.006973	0.174	0.5931	0.9867
δ	10	3.34	3.65	0.1011	3.582E-7	2.346	12.3
A	1	0.5025	1.036	0.02813	2.961E-7	0.06586	3.657
B	1	1.331	0.7908	0.02405	0.4846	1.107	3.522
a	16	12.73	2.87	0.03372	7.939	12.44	19.14
b	5	0.7781	0.3642	0.009654	0.2058	0.736	1.581

Table 3: Change points detection and parameter estimates in a Bayesian framework in a Poisson multi-phases model with a Gamma distributed rate at the stable phase.

2.3 Costs Management in Enrichment Studies

Drug development costs must be kept as low as possible to guarantee the accessibility of new drugs for all patients in need. This requires techniques for time and cost planning and optimization at both the project and study level. In this context, Kramer and Schul-

man [2012] suggested strategies such as an optimal use of new technology to improve data exchange between researchers and for easier access to data including patient data. New information and communication technology should improve communication and information flux between researchers to reduce waiting time. Waiting time implies additional personnel, infrastructures and materials costs. In addition, capitalization costs increase depending on time. At the trial level, the impact of the study sample size on the costs is often underestimated (Claxton and Posnett [1996])

Enrichment studies are thought to accelerate the drug development process and reduce the development costs while improving the chance of positive study results (Temple [2005], Temple and Becker [2012], Wassmer [2013]). However, several factors may considerably affect the study costs in enrichment studies. Small marker prevalence may raise the patient recruitment time with impact on the costs. The prevalence may significantly increase the number of patients that must be screened and tested until achieving the study sample size. Another factor affecting the screening and marker-test process and the recruitment time is the marker-test characteristics. Since enrichment studies have been getting more and more popular in drug development, it is relevant to explicitly investigate the impact of specific factors of enrichment studies on the study costs for better planning and decision making.

The marker-test and screening costs may seriously increase the total study costs, depending on the screening conditions to be met by patients and the type of the marker-test as well as the number of patients to be screened and tested. The number of patients passing through the screening and marker-test procedures until achieving the required study sample size depends on the study sample size, but also on the marker-prevalence. To the best of our knowledge, there is no work which explicitly investigates the cost of enrichment studies. DiMasi et al. [2003] and DiMasi et al. [1991] focused on the estimation of the average costs of drug development. Cline et al. [1998] studied the impact of patient management and education on the care costs using an example study on patients with heart failure.

In this section, we suggest a simple technique for cost prediction in enrichment studies. The studies costs are summarized in four main categories: The screening costs, the marker-test costs, the care costs and the duration costs. The care costs represent all expenditures after patients have entered into the trial (drug supply, visits, indemnity,

personnel costs, ...). The care costs depend essentially on the number of patients in the study and are thus expected to be lower than in unselected trials, since the sample size of enrichment studies tend to be smaller. The duration costs represent the total expenditure related to the overall duration of the trial. This depends on several factors. An increasing number of centers activated for the trial decreases the recruitment time but may increase the costs per time unit. More recruitment centers implies more infrastructure and more personnel. The screening costs and marker-test costs per patient and the cost per time unit depend on the trial and should be provided by researchers.

2.3.1 Costs Estimates without Marker-test Characteristics

Let us consider an enrichment trial and an unselected trial with sample size N^+ and N_0 respectively. The total costs of the unselected study consists of screening costs, the care costs and the costs due to the recruitment time. They can be defined as

$$Z = \kappa_c N_0 + \kappa_s M_0 + \kappa_t T(N_0),$$

where κ_s, κ_c and κ_t represent the screening and care costs per patient and the expenditure per time unit, respectively. The number of patients to be screened until achieving N_0 unselected patients is modelled by a Negative Binomial distributed random variable $M_0 \sim NBin(N_0, \mu)$, where μ represents the probability of screening an unselected patient as unselected. $T(N_0)$ denotes the time required to accomplish the recruitment process. It follows an Erlang or Pearson type VI distribution depending on the underlying recruitment process. More generally, the distribution of $T(N_0)$ can be gathered through the strategies proposed in the above section and included in the costs prediction. For example $T(N_0) \sim Erlang(N_0, \lambda)$ if a total number of patients equal to λ arrive at the recruitment centers per time unit under a Poisson recruitment process with constant intensity.

In an enrichment study, the cost per patient due to both the screening and test procedure is given by $\kappa_s + \kappa_m$, where κ_m represents the marker-test costs for one patient. The probability of an unselected patient to successfully pass through the screening and test procedure and being classified as marker-positive is equal to $\theta\mu$, where θ denotes the marker prevalence. We use this probability to derive the number X of patients to be screened and tested until achieving the enrichment study sample size N^+. This number follows the distribution $X \sim NBin(N^+, \theta\mu)$. We define the total expenditures in an

enrichment study as

$$Y = (\kappa_s + \kappa_m)X + \kappa_c N^+ + \kappa_t T^+(N^+). \tag{11}$$

This definition of the costs explicitly considers factors affecting enrichment trials such as the marker prevalence, through X and $T^+(N^+)$. The impact of the probability of screening a patient as unselected is not of great interest here as the screening procedure is common to any clinical trial. We keep it constant and focus on the impact of the marker prevalence and marker-test characteristics on the total costs

$$\begin{aligned} E(Y) &= (\kappa_s + \kappa_m)N^+/\theta\mu + \kappa_c N^+ + \kappa_t E(T^+(N^+)) \\ Var(Y) &= (\kappa_s + \kappa_m)^2 N^+(1-\theta\mu)/\mu^2\theta^2 + \kappa_t^2 Var(T^+(N^+)). \end{aligned}$$

Example 2.2 *Let the sample size of an unselected study be $N_0 = 800$ patients. Assume there is a marker group of patients in which a higher effect size is expected. The sample size is reassessed to conduct the study in this marker population, and the new sample size is estimated as $N^+ = 600$ to achieve the same power. The eligibility rate in the overall patients population is $\mu = 70\%$. Finally it is assumed, that the care costs $\kappa_c = 20\kappa_s$ and $\kappa_s = \kappa_m = \kappa_t$.*

We investigate example 2.2 by estimating the total cost of the unselected study and particulary the total cost of the enrichment study for different values of the marker prevalence. Figure 13 shows the variation of the number of patients to screen and test until recruiting 600 marker-test positive patients. For the results represented in Figure 14, we assume a Poisson recruitment process with constant intensity $\lambda = 10$. For Figure 15, $\lambda \sim \Gamma(10, 1)$ is assumed with the mean equal to 10 patients and variance equal to 10. Figure 14 and Figure 15 show that the total enrichment study costs decrease when the marker prevalence increases regardless of the assumed recruitment process. That can be easily explained by the fact that the number of patients to screen and test decreases when the prevalence increases and the study time varies in the same direction. The prevalence is a crucial factor in planning the enrichment study costs. For Example (2.2) and under the Gamma-Poisson recruitment process, the enrichment study costs are equal to 18038 for a marker prevalence of 60% and increase to 30092 for the prevalence of 20%. In addition, the present example shows how fast the marker study costs can exceed the costs of an unselected trial.

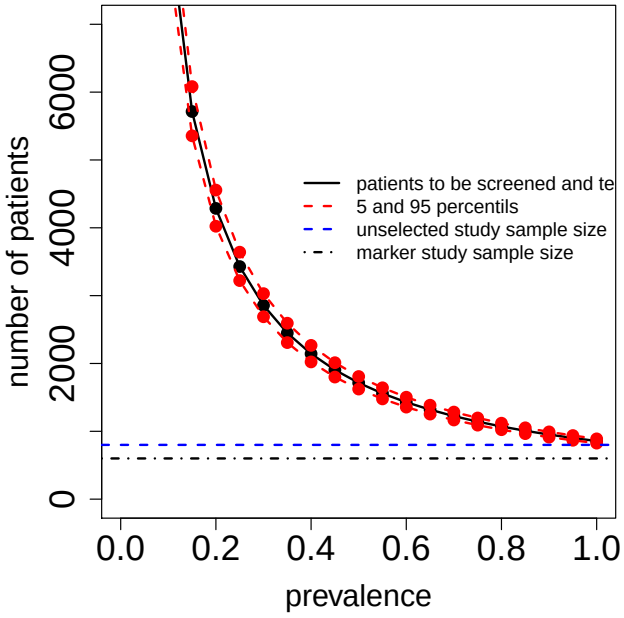

Figure 13: Patients to be tested until $N^+ = 600$ marker positive patients are recruited as a function of the prevalence.

2.3.2 Studies Costs and Marker-test Characteristics

The sensitivity and specificity of the marker-test may significantly affect the probability of testing an unselected patient as marker-positive. For $Sens \leq 1$ and $Spec \leq 1$, this probability is given by $Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)$ instead of simply θ. The probability of an unselected patient to be screened and tested as marker-positive is given by $\mu[Sens \cdot \theta + (1 - Spec) \cdot (1 - \theta)]$, where μ represents the probability of screening an unselected patients as unselected. Analogously to Section 2.3.1 we derive analytically the impact of the sensitivity and specificity of the marker-test as well as the marker prevalence on the marker study costs. Then the figures in that section correspond to $Sens = Spec = 1$. The costs of an enrichment study under consideration of the marker-test characteristics are given by

$$\begin{aligned} \widetilde{Y} &= (\kappa_s + \kappa_m)X + \kappa_c N^+ + \kappa_t T^+(N^+) \\ \widetilde{X} &\sim NBin(N^+, \mu[Sens \cdot \theta + (1 - Spec)(1 - \theta)]), \end{aligned} \tag{12}$$

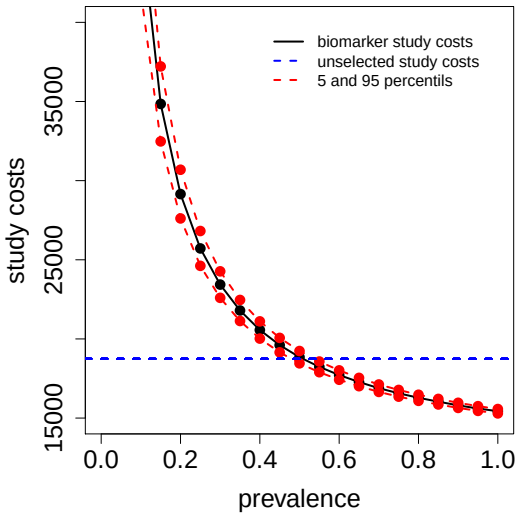

Figure 14: Variation of enrichment study costs depending on the marker prevalence. The recruitment is assumed to be a Poisson process with constant intensity $\lambda = 10$ and the study sample is $N^+ = 600$.

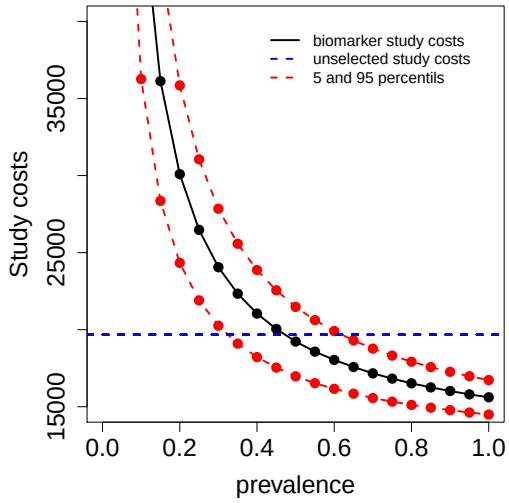

Figure 15: Variation of enrichment study costs depending on the marker prevalence. The recruitment process is assumed to be a Poisson process with intensity $\lambda \sim \Gamma(10, 1)$ and the study sample size is $N^+ = 600$.

where

$$E(\widetilde{Y}) = (\kappa_s + \kappa_m)N^+/[Sens \cdot \theta + (1 - Spec)(1 - \theta)]\mu + \kappa_c N^+ + \kappa_t E(T^+(N^+))$$
$$Var(\widetilde{Y}) = (\kappa_s + \kappa_m)^2 N^+(1 - [Sens \cdot \theta + (1 - Spec)(1 - \theta)]\mu)/$$
$$\mu^2[Sens \cdot \theta + (1 - Spec)(1 - \theta)]^2 + \kappa_t^2 Var(T^+(N^+)).$$

To illustrate this relationship between the enrichment study costs and the marker-test characteristics, we consider Example 2.2 and keep all other parameters constant except the test characteristics. Figure 16 shows the change in the study costs as the sensitivity and specificity of the test vary. The marker prevalence is assumed to be equal to 60%. In addition, we assume a recruitment process with arrival rate of $\lambda \sim \Gamma(10, 1)$, so that the recruitment time needed to enroll the 600 marker positive patients follows a Pearson type VI distribution $Gg(600, 10, 1/(0.6 * 0.7))$. Figure 16 shows that the costs vary depending on the sensitivity, specificity and the marker prevalence and range from about $16000k_m$ to $24000k_m$ for the same value of the marker prevalence (60%). The black square represents the cost of $18000k_m$, when the $Sens = Spec = 1$. It corresponds to the value of the cost in Figure 15 for the marker prevalence of 60%. Each line represents

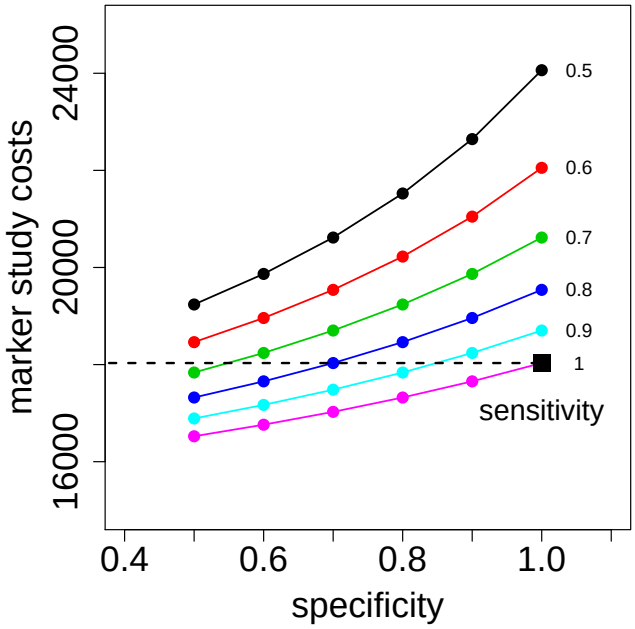

Figure 16: Costs of an enrichment study as a function of the sensitivity and specificity of the marker-test. Here, $\kappa_c = 20\kappa_m = \kappa_t$ and $\kappa_m = \kappa_s$ and the costs are in the unit of κ_m.

different values of the study costs for one value of the sensitivity and different values of the specificity. The larger the specificity, the higher the costs. However, the specificity must be large enough to allow for a reasonable number of false positive patients to enter into the trial. Such patients would receive the therapy although they are not supposed to benefit from it and this may decrease the power of the study. When the sensitivity increases, the chance of testing a marker-positive patient as positive increases and thus the costs decreases.

Let us now consider the population of patients tested as positive. The distribution of the true-positive patients in a sample of N^+ selected patients is given by

$$TP \sim Bin(N^+, PPV),$$

where TP represents the number of the true-positive patients and PPV represents the positive predictive value of the marker-test (see Appendix A.2). Then the expected number of true positive patients is given by $E(TP) = N^+PPV \leq N^+$. That means, the selected study population consists in expectation of N^+PPV marker-positive and $N^+(1-PPV)$ marker-negative patients. To ensure, for example, that the study population

contains an expected number of marker-positive patients equal to the study sample size N^+, more patients must be recruited. By enrolling N^+/PPV test-positive patients instead of N^+, we obtain a study population with $E(TP) = N^+PPV/PPV = N^+$ marker-positive patients. In this case, the number of patients to be screened and tested is given by $X_a \sim NBin(N^+PPV, \mu[Sens\cdot\theta+(1-Spec)(1-\theta)])$ and the adjusted marker study costs is defined as

$$\begin{aligned}
Y_a &= (\kappa_s + \kappa_m)X_a + \kappa_c N^+/PPV + \kappa_t T^+(N^+/PPV) \\
X_a &\sim NBin(N^+/PPV, \mu[Sens\cdot\theta + (1-Spec)(1-\theta)]) \\
E(Y_a) &= \frac{N^+(\kappa_s+\kappa_m)}{PPV[Sens\cdot\theta+(1-Spec)(1-\theta)]} + N^+\kappa_c/PPV + \kappa_t E(T^+(N^+/PPV)) \\
Var(Y_a) &= (\kappa_s+\kappa_m)^2(N^+/PPV)(1-[Sens\cdot\theta+(1-Spec)(1-\theta)]\mu)/ \\
&\quad \mu^2[Sens\cdot\theta+(1-Spec)(1-\theta)]^2 + \kappa_t^2 Var(T^+(N^+/PPV)).
\end{aligned}$$

The sample size adjustment to increase the number of true positive patients may considerably increase the study costs. The expected additional costs are given by $E(Y_a)-E(\widetilde{Y})$.

2.3.3 Costs by random Marker-Prevalence

If the information about the marker-prevalence is in form of a beta distribution, then the distribution of the number of patients passing through the screening and marker-test until the study sample size is achieved follows a Negative binomial distribution as follows

$$\begin{aligned}
X &\sim NegB(N^+, \mu\theta) \\
\theta &\sim Beta(\omega_1, \omega_2).
\end{aligned}$$

Note that $\mu\theta$ also follows a Beta distribution with mean and variance $E(\mu\theta) = \mu\omega_1/(\omega_1+\omega_2)$, $Var(\mu\theta) = \mu^2\omega_1\omega_2/(\omega_1+\omega_2)^2(\omega_1+\omega_2+1)$, respectively. This implies X is Beta Negative Binomial distributed $X \sim BNB(N^+, \omega_1, \omega_2)$ and thus the distribution of the total expenditure at the screening and marker-test stage can be derived as $((\kappa_s+\kappa_m)X)$. The care costs $\kappa_c N^+$ do not depend on the marker prevalence. The time component $\kappa_t T(N^+)$ of the total costs must be simulated from the following distribution:

$$\begin{aligned}
T(N^+) &\sim Erlang(N^+, \mu\theta\lambda) \\
\theta &\sim Beta(\omega_1, \omega_2),
\end{aligned}$$

when λ is constant and

$$T(N^+) \sim Erlang(N^+, \mu\theta\lambda)$$
$$\theta \sim Beta(\omega_1, \omega_2)$$
$$\lambda \sim \Gamma(a, b)$$

for a Gamma distributed recruitment rate.

2.3.4 Power and Marker-test Characteristics

It is difficult if not impossible the find a perfect marker-test. The study population (test-positive) is a mixed population of marker-positive patients (true positive) and some marker-negative patients (false positive). This mixed population is randomized in the treatment groups as illustrated by Figure 17. The power of the marker study may de-

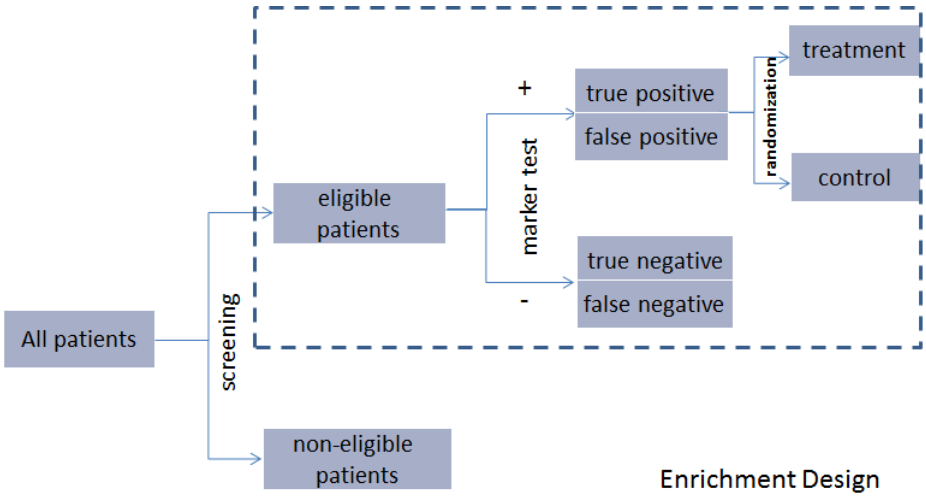

Figure 17: Enrichment Design considering the Marker-test Characteristics

crease, when a considerable number of marker-negative patients are erroneously enrolled. This may be due to the fact that the effect size provided by the study population is less than that used in estimating the study sample size. The relationship between the power $(1-\beta)$ of a marker study, its sample size (N^+) and the effect size in the marker positive population δ^+ is well known for several types of study designs (non-inferiority, equivalence,...). This relationship provides the answer to the following questions in planning a trial with a type I error α: What sample size is required to ensure a power of $1-\beta$ in detecting an effect difference δ^+? What is the power of the trial in detecting an effect difference δ^+, when N^+ patients are recruited for the study? What effect size can be

detected with power $1-\beta$ if the study sample size is N^+? The relationship between power, sample size and effect size has been described in Lachin [1981] for many study designs. The sample size should be high enough, the samples representatively selected and randomized in the study groups to maintain the power.

In this section we investigate the impact of marker-test characteristics and marker prevalence on the power of an enrichment study. To derive the relationship between the power and these parameters, we consider a trial in which the decision is based on the hypothesis $H_0 : \delta^{(+)} > 0$ vs $H_1 : \delta^{(+)} \leq 0$. For simplicity, we assume balanced study groups (new drug and control) and equal effect variability in both study groups. Let $\delta^{(+)}$ and $\delta^{(-)}$ be the standardized effect difference in the marker positive population and the marker negative population respectively, where $\delta^{(+)} > 0$. Then, the number of marker-positive patients needed can be written as

$$N^+ = \frac{2(Z_{1-\alpha} + Z_{1-\beta})^2}{\delta^{(+)2}}, \tag{13}$$

where $Z_{1-\alpha}$ denotes the $100\alpha\%$ percentile of the standard normal distribution (Lachin [1981]). N^+ is the amount of test-positive patients consisting of TP true positive and $N^+ - TP$ false positive patients. $TP \sim Bin(N^+, PPV)$, where PPV denotes the proportion of positive patients in the population of test-positive ones. The power $1-\beta$ is a random variable given by the number of true positive patients in the sample size N^+. Its mean and variance can be approximated using the Δ-Rule (Lehmann and Casella [1998]) as follows

$$2(Z_{1-\alpha} + Z_{1-\beta})^2 = TP\delta^{(+)2} + (N^+ - TP)\delta^{(-)2},$$

and then

$$\begin{aligned} E(Z_{1-\beta}) &= \sqrt{\frac{\delta^{(-)2}(1-PPV)N^+ + \delta^{(+)2}PPV \cdot N^+}{2}} - Z_{1-\alpha} \quad \text{and} \\ Var(Z_{1-\beta}) &\approx \frac{[(\delta^{(+)2} - \delta^{(-)2}) \cdot N^+ PPV(1-PPV)]^2}{2[\delta^{(+)2}PPVN^+ + \delta^{(-)2}(1-PPV)N^+]}. \end{aligned} \tag{14}$$

An estimate of the expected power can thus be derived from $E(Z_{1-\beta})$. Figure 16 shows the variation of the power of a marker study for different values of the sensitivity and specificity of the marker-test. We assume an initial power of 90% used in estimating the study sample size. By a specificity of 100%, which means no negative patient is expected to be tested as positive, a sensitivity different from zero is enough to maintain the desired power. Of course a reasonable sensitivity is required to ensure that positive samples are

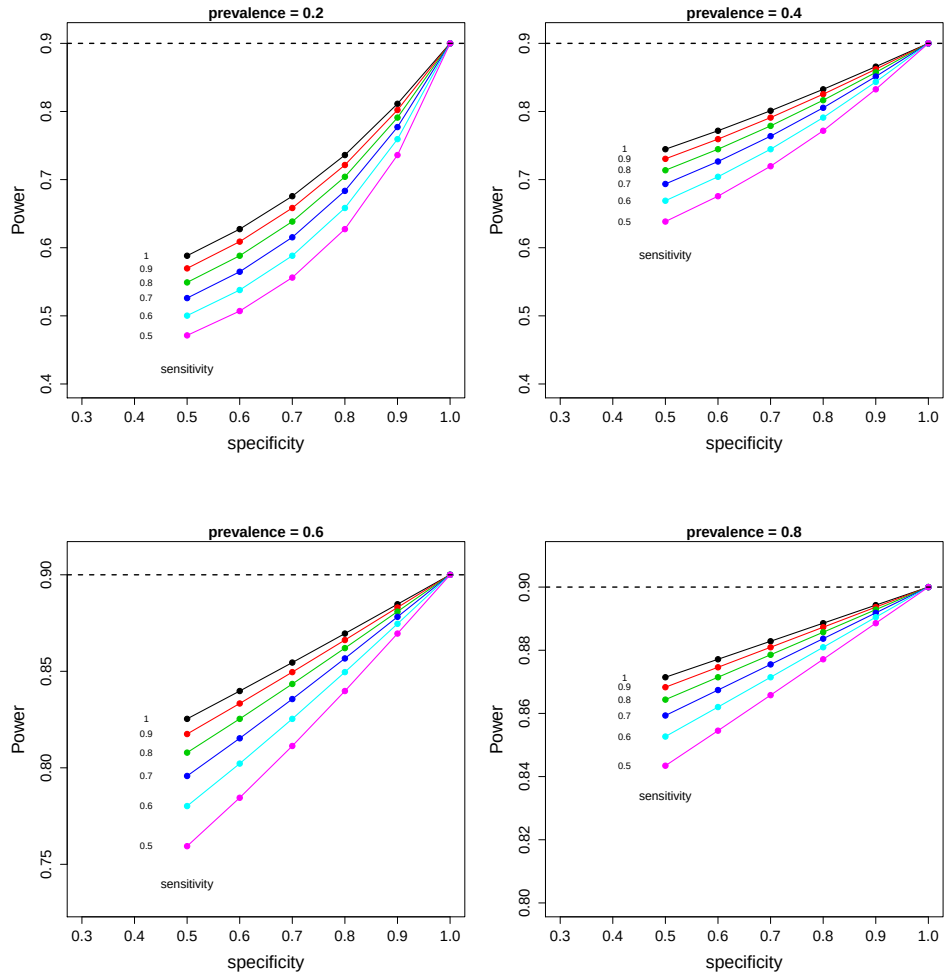

Figure 18: Power versus sensitivity and specificity of the marker-test

not rejected all the time. The second factor affecting the power is the marker prevalence (see Figure 18). The smaller the marker prevalence the lower the power. Using the decomposition of the sample size in TP and $N^+ - TP$, we derive the adjusted sample size expected to maintain the initial power. To avoid a decrease in power, the sample size can be defined as follows:

$$N_c = \frac{\delta^{(+)2} N^+}{\delta^{(+)2} PPV + (1 - PPV)\delta^{(-)2}}$$

where $N_c \geq N^+$ is the number of test-positive patients to be enrolled instead of N^+ patients.

Example 2.3 *Let us assume that the effect size in the marker positive population is twice than that of the negative group. Let the marker prevalence be equal to* 40% *and the sensitivity and the specificity of the marker-test* 80%*. For an initial estimated sample*

size of $N^+ = 600$ *marker positive patients, the compensated sample size is* $N_c = 672$*. That means 72 more patients must be recruited to avoid a loss in the power due to the enrollment of false positive patients.*

This way of reassessing the sample size depends on the study design. In the previous section, we suggested recruiting patients so that expected number of true positives is equal to the given sample size. That means for the above example, that N^+/PPV patients equal to $N_c = 700$ should be enrolled. This is slightly inaccurate compared to the present strategy, but seems to be faster and does not depend on the study design.

2.4 Outlook

The recruitment model should be as simple possible to reduce the variability in the predictions. However, it may be relevant to add some additional levels of randomness to the model to better reflect the process. The following models may be used in modeling the recruitment process in enrichment studies.

2.4.1 Models with Beta distributed Marker Prevalence

An estimate of the distribution of the marker prevalence may be available in the form of a Beta distribution. In this case, the recruitment model is given by

$$\begin{aligned} N_t^+ &\sim Pois\left(\sum_{k=1}^{K} \theta \lambda_t^{(k)}\right) \\ \theta &\sim Beta(\omega_1, \omega_2), \end{aligned} \tag{15}$$

where the recruitment rate $\lambda_t = \sum_{k=1}^{K} \lambda_t^{(k)}$ may take different forms as investigated in the previous sections. If λ_t is constant, the recruitment time needed to enroll N^+ patients follows the $Erlang(N^+, \theta\lambda_t)$ distribution, where $\theta \sim Beta(\omega_1, \omega_2)$. In addition, if λ_t is Gamma distributed then the shape parameter in the distribution of the recruitment time is the product of Gamma and Beta distributions. The difficulty remains to connect the distribution of the recruitment in these cases to a known distribution. However, an estimate of the recruitment time can be obtained through simulation and the models can be updated in OpenBUGS by assuming a Gamma prior for ω_1 and ω_2.

2.4.2 Hierarchical Modeling

Patient recruitment models in enrichment trials can be defined hierarchically by assuming that the recruitment rate of test-positive patients given the arrival rate of unselected patients is a binomially distributed random variable:
$\lambda_t^+|\lambda_t \sim Bin\Big(\lambda_t, \theta \cdot Sens + (1-\theta)(1-Spec)\Big)$. This leads to the following hierarchical model of patient recruitment in enrichment trials:

$$\begin{aligned} N_t^+ &\sim Pois(\lambda_t^+) \\ \lambda_t^+ &\sim Bin(\lambda_t, \theta \cdot Sens + (1-\theta)(1-Spec)) \\ \theta &\sim Beta(\omega_1, \omega_2), \end{aligned} \tag{16}$$

where λ_t follows, for example, a Gamma distribution.

Remark 2.7 *The above model is equal to the model (2) investigated in the previous section, when the mean of the distribution* $Bin(\lambda_t, \theta \cdot Sens + (1-\theta)(1-Spec))$ *of* λ_t^+ *is considered instead of the distribution itself. That means, it is assumed that* $\lambda_t^+ = [\theta \cdot Sens + (1-\theta)(1-Spec)]\lambda_t$ *instead of* $\lambda_t^+ \sim Bin(\lambda_t, \theta \cdot Sens + (1-\theta)(1-Spec))$.

We consider a deterministic screening procedure: A fixed proportion of patients arriving at the recruitment centers, such as 90%, are screened as unselected. It may be relevant to also consider the variability in the number of patients screened as positive. An alternative could be to consider the screening procedure as a Bernoulli experiment with parameter 0.9 assumed for the marker-test procedure. That would lead to the following model:

$$\begin{aligned} N^+ &\sim Pois(\lambda_t^+) \\ \lambda_t^+ &\sim Bin(\lambda_t, \theta \cdot Sens + (1-\theta)(1-Spec)) \\ \lambda_t &\sim Bin(\lambda_t^0, \mu) \\ \theta &\sim Beta(\omega_1, \omega_2), \end{aligned}$$

where λ_t^0 represents the total arrival rate of actually unselected patients that are screened and eventually tested. μ denotes the probability of screening a patient as unselected. This hierarchical modeling of patient recruitment processes in enrichment studies will be looked at closer. The impact of this additional level of randomness on the accuracy of predictions should be studied. Note that each additional random level in the model usually increases the variability of the estimates.

3 Sensitivity-preferred Classification Rules

A marker-test procedure used in selecting the study population for an enrichment trial is a binary classification problem. In conducting an enrichment trial for target development of a new drug for a life-threatening disease without standard treatment, it is more important not to misclassify a diseased patient as healthy, so that he can benefit from the new therapy. In building binary classification rules, one tries to identify a rule that best assigns new observations to the correct classes. Usually, the quality label "best" refers to the maximization of the total rate of correctly classified observations which is equivalent to the minimization of the total misclassification errors. However, this definition of the quality "best" is not always appropriate because the impact of misclassification costs may differ between both classes. An incorrect assignment to one class can be more harmful in comparison to a misclassification in the other class.

A medical diagnosis (diseased versus healthy) is a typical example of the binary classification problem where one class is usually more important than the other. For example, in classifying moles into malignant and benign, the impact of removing a harmless mole can be less severe than overlooking a malignant mole that can develop into a melanoma. In such cases, the researcher is primarily interested in minimizing the error rate in classifying malignant moles so as not to misclassify them as benign. In some situations the error rate will not be allowed to exceed a certain level in that class. Throughout this thesis, the class of diseased patients will be assumed as more important.

Classification techniques that guarantee a minimal pre-specified true classification rate into the most important class are required. Such techniques should favor the more important class while building classifiers and thus lead to an acceptable true classification rate in that class. A minimal true classification rate in the important class could be fixed by researchers according to the consequences of misclassification on the patient's health and economic situation. Often, patient's diagnosis is based on biomarkers. Thus classification or diagnosis methods must have the ability to simultaneously pinpoint the relevant markers from a list of potential candidates (gene expression for example).

Our contribution in this part of the thesis is the introduction of a new approach for building binary classification rules that meet the above requirements. We suggest a new approach for entirely controlling the true classification rate in the most important

class with respect to the sensitivity while at the same time maximizing the specificity. The new classification strategy is designed to guarantee a pre-determined sensitivity (e.g. sensitivity $\geq 90\%$), to provide the highest specificity and to select the relevant predictors for this purpose.

For this purpose, we suggest building classification rules by optimizing loss or utility functions of binary classification such as the log-likelihood function, subject to the constraint that the sensitivity belongs to a pre-determined interval (of high values). In other words, the optimization algorithm runs only in the admissible region. To deal with the high-dimensionality of the data, we add a L_1-norm penalty to the model parameters for simultaneous selection of relevant predictors. L_1-norm penalized regression was introduced by Tibshirani [1996] for variable selection in regression models. Here, the predictor selection procedure is connected to the model and optimization procedure through the L_1-norm penalization of the model parameters, unlike in filter methods such as "p-values screening". Liu et al. [2007] has suggested the use of the L_q-norm $0 < q < 1$ for more sparsity which presents more optimization burdens (non convexity of the L_q-norm). The new approach is also applicable to unpenalized problems by optimizing the loss-function of classification subject to the constraint on the sensitivity.

In the following sections, we describe two cases of optimization of loss-functions under constrained sensitivity. The first one is the optimization of the likelihood (log-likelihood) function of a binary logistic regression subject to a constraint on the sensitivity and the L_1-norm penalization of the model parameters. Penalized logistic regression for simultaneous feature selection in computing a logistic regression model has been investigated by many authors (Meier et al. [2008], Shevade and Keerthi [2003],Koh et al. [2007]). The second illustration is the optimization of specificity, again subject to a constraint on the sensitivity. The L_1-norm penalization of the model parameters will be used for the feature selection. The sensitivity and specificity have been used by Liu and Tan [2008] as loss functions of binary classification. They suggested to optimize the objective function of the weighted sum of the sensitivity and specificity in dealing with classification problems with different class importance.

A very simplistic classification rule which guarantees high sensitivity can be formulated as follows: Assigning all patients with symptoms into the diseased class may lead to 100% sensitivity. However, such a trivial classification rule is unreasonable, since a healthy patient would receive therapy with the disadvantages of experiencing adverse

effects of therapy, which may include psychological and physiological stress as well as incurring treatment costs.

A reasonable and commonly used strategy which guarantees a certain sensitivity has been investigated by Jung et al. [2010]: First compute the decision score (such as disease probability) and then find the cut-off that leads to the desired sensitivity instead of trying to optimize the balance between the sensitivity and specificity. However, a small gain in the sensitivity may lead to a large loss in specificity. All information pertaining to the research question should be included to the optimization equation and procedure in order to guarantee an optimal solution.

Simple solutions as described above do not meet real world requirements and classification by different class importance remains a challenging issue in biomarker research. Other methods have been suggested for dealing with this issue which try to increase the sensitivity without guaranteeing a specific value. The improvement of the sensitivity leads to a loss in specificity. One of these methods, for example, is arrived at by weighting the observations (patients) in the training procedure (Yi [2005] and Liu and Tan [2008]): higher weights are assigned to the observations from the most important class. Thus they introduced the different weights early by the definition of the loss-function of classification.

Weighting has also been used in the later stages of classification rules construction. Elkan [2001] suggested computing the cut-off value of the estimated decision scores by considering different weights for the different classes. In practice, it is difficult to find interpretable weights. Affecting any number as weight only for mathematical and optimization convenience may be unacceptable. Misclassification costs represent ethically and economically interpretable weights. Sheng and Ling [2006] and Ling and Sheng [2008] used the misclassification costs to weight the classes differently in building classification rules. Some authors tried to increase the number of observations of the most important class in the training procedure to favor it (Japkowicz and al. [2000],cha). However, these approaches are not designed to guarantee a pre-determined value of the sensitivity while providing the largest specificity. Our strategy is to perform the optimization in the subregion of ethically and economically acceptable sensitivity values fixed by experts and then no classification rule should be better under the given constraint.

3.1 Data Material and Background

We use endometriosis data as examples in implementing our approach. Endometriosis, the presence of endometrial-like tissue outside the uterus, is a disease associated with pelvic pain and infertility. It essentially affects women at childbearing age with a prevalence of 10% (May et al. [2010]). A higher prevalence of endometriosis is reported in the population of women with subfertility. One of the most common symptoms related to endometriosis is pelvic pain, but in most cases it progresses asymptomatically. The etiology of endometriosis as well as the relationship between the disease stages and the severity of its symptoms remains unclear. It can take several years or decades for symptoms to present themselves after the disease appears. Up to now, endometriosis has only been effectively diagnosed through invasive procedures such as a laparoscopy (May et al. [2011]). Laparoscopy is a surgical technique in which small incisions are done in the abdomen to access the organ on which the surgical operation is performed under assistance of cameras and monitors (Arai [2012]).

Several studies have been conducted on the identification of markers that can be used in diagnosis and monitoring of the treatment of endometriosis to prevent women having to undergo unnecessary surgical procedures. May et al. [2010] reviewed 189 publications for endometriosis in urine, serum and plasma. These papers presented a large number (more than 100) of markers that are elevated in endometriosis patients (e.g. Interleukin 6 and 8, Interferon gamma, Cancer antigen CA125 and CA19-9). None of these markers were demonstrated as unequivocally useful in the clinical practice (May et al. [2010]). Other markers such as endometrial biopsy, endometrial fluid aspirates and menstrual effluent have been investigated (May et al. [2011]). Finding a panel of markers (list of genes or proteins) with clear clinical benefit for diagnosis and monitoring of endometriosis remains of great interest. We apply the suggested technique on data from a research project conducted by Bayer Schering Pharma in finding diagnostic markers for endometriosis screening and monitoring. Women aged from 18 to 45 years, scheduled for therapeutic laparoscopy, either for confirmation of endometriosis or diagnostic laparoscopy because of subfertility or tubal ligation, were enrolled for the trial (Walzer et al.). The coordinating investigator for this study was Prof. Dr. med. Dr. phil. A. D. Ebert, Director of the Clinic for Gynaecology and maternity at Vivantes Humboldt-Clinic.

3.1.1 Endometriosis Gene Expression Data

Messenger RNA expression profiling of blood samples was collected from 113 women with symptoms of endometriosis (pain) and women without symptoms undergoing laparoscopy due to subfertility, or tubal ligation. From these two categories of patients, four different groups could be identified through laparoscopy: women having pain without having endometrial lesions, women having pain and endometrial lesions, women without pain and without endometrial lesions and women without pain but with endometrial lesions. The data are be summarized in Table 4 and does not have any missing values. We are interested in different data sets that can be used to construct binary classifiers.

Blood RNA endometriosis data (45861 predictors)

Study group	No pain & no endo.	No pain & endo.	pain & endo.	pain& no endo.
Sample size	10	14	63	26

Table 4: Study groups and sample sizes of a RNA data for the identification of diagnosis and monitoring markers of endometriosis.

Each data set has binary outcomes (two groups). The four subgroups presented in Table (4) coded by patients with pain and confirmed endometriosis lesions (G_{11}), patients with pain but without lesions (G_{12}), patients without pain but with lesions (G_{21}) and finally patients without pain and no lesions (G_{22}) represent possible classes. These classes can also be combined and analyzed according to the research question. For example, a classification rule built by using only the classes G_{11} and G_{12} provides diagnostic biomarkers and a rule for the identification of patients who have endometriosis lesions, from the population of patients having pain. We build five such data sets and five different classifiers using our new approach. There are no missing values in the data.

3.1.2 Protein Data

For earlier diagnosis and monitoring of endometriosis, a protein study has been conducted (Henze et al. [2013]). Here, protein levels of 191 proteins in the peritoneum fluid were measured from a population of 34 women aged between 21 and 47 years suffering from endometriosis in three different stages and 16 women aged between 21 and 48 without endometriosis. There are no missing values in this data set but 35 proteins have the same level of protein in all patients regardless of their disease status. Such proteins, viewed as constant, were discarded from the analysis. Two endometriosis patients

Peritoneum fluid endometriosis proteins (191)				
Study group	No endo.	endometriosis in grad 1	endo. in grad 2	endo. in grad 3
Sample size	16	5	5	12

Table 5: Study groups and sample sizes of a protein data for the identification of diagnosis and monitoring markers of endometriosis

of grade 3 presented extremely high levels of the proteins CA-125 (20000 and 40000), which was about 38 and 76 times higher than the highest level without considering these two patients. There was one patient with an outlier level of CA-19-9. Henze et al. [2013] focused particularly on Trefoil Factor 3 (TFF3) and found out that TFF3 is significantly higher in women with endometriosis and correlated with known biomarkers related to endometriosis such as CA-125, CA-19-9, inflammatory mediators (IL-8, MCP-1) and Matrix Metalloproteinase 7 (MMP-7, Matrilysin). Other proteins investigated for endometriosis diagnosis and monitoring have been listed in May et al. [2010] and May et al. [2011]. Our approach will be used in building a multi-dimensional classification rule with given constraints based on these protein data.

3.2 Classification

A two-class classification problem is considered (diseased versus healthy). Classification methods (e.g. logistic regression, Support Vector Machine, classification tree) try to find a rule for correct assignment of observations to the right class using predictors. These predictors are used to compute a single decision score such as a disease probability. The computed decision score vector is used as a surrogate marker similar to the classification procedure in one dimensional settings.

The distributions of the decision score in both classes usually overlap. It is then relevant to select a cut-off that leads to the minimal misclassification. A cut-off is a value where observations whose decision score exceeds this value are assigned to one class and to the other class if lower than this value. Figure 19 presents an example of the distributions of a decision score in the diseased and healthy class, respectively. This example has been simulated under the assumption of higher regulation in the diseased class. A reasonable cut-off is 2 on the x-axis. This cut-off leads to four possible results as usual: True Positive (TP) representing the number of diseased patients predicted as diseased, True Negative (TN) representing the amount of healthy patients classified as

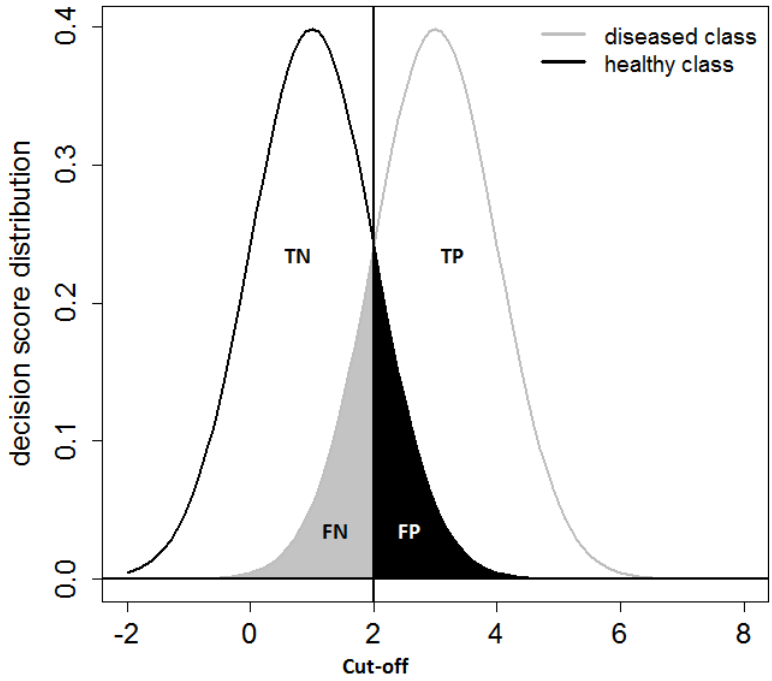

Figure 19: Distributions of a decision score in the disease (grey) and healthy (black) class. The cut-off is assumed to be equal to 2.

healthy. FP and FN denote, respectively, the False Positive and False Negative patients.

3.2.1 Logistic Regression

We implement our strategy of optimization under constrained sensitivity in the context of logistic regression. The logistic regression is one of the most used classification methods, where the outcome variable is the probability of the observations to be from the reference class (here the disease class). Here, the logistic regression is presented in the context of binary classification. Consider a training set denoted by (X, y), where X represents a $n \times (k+1)$ predictor matrix and y the $n-$ dimensional binary outcome vector usually coded as $1 \hat{=}$ diseased and $0 \hat{=}$ healthy. Each component of y is viewed as a realization of a Bernoulli random variable with mean $P(Y = 1|X) = p$, which is defined as a function of the model parameters as

$$logit(p) = \log(\frac{p}{1-p}) = X\beta,$$

where β represents the $(k+1)$-dimensional parameter vector including the intercept. The logistic regression belongs to the generalized linear models introduced by McCullagh and Nelder [1989] with a *logit* link function. The likelihood function can be derived as

$$\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, i \in \{1, \cdots, n\}$$

and then the model parameters are estimated by solving the optimization equation

$$\hat{\beta} = \underset{\beta}{argmin} \sum_{i=1}^{n} \Big\{ -y_i \log(p_i) - (1-y_i)\log(1-p_i) \Big\}.$$

This optimization of the $\log-$ likelihood function does not have an explicit solution and is usually solved using numeric algorithms such as Newton-Raphson (Ben-Israel [1966]).

3.2.2 Penalized Logistic Regression

The analysis of high-dimensional data such as gene expression data requires statistical methods that enable feature selection. Sparse models are easier to compute and to interpret and have many ethical and economic advantages. For example, a patient diagnosis based on the evaluation of the level of only one protein may be easier to perform and to interpret than a diagnosis based on a multidimensional evaluation of ten proteins. Most high-dimensional data contains only a few relevant predictors and the others are just noise. Furthermore, when the number of predictors is far larger than the observations, the logistic regression may lead to over-fitting: The model performs well on the present data, but poorly on new observations.

Classification methods are often combined with filter methods. The predictors are evaluated separately in their capacity for distinguishing between the two classes (p-value of two-tailed t-test, AUC). However, these methods do not consider the properties of the model and do not take into account the group effect that an ostensibly irrelevant predictor could show a relevant effect, when it is combined with other predictors. Penalized logistic regression includes feature selection in the optimization algorithm.

LASSO for Logistic Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) in the context of logistic regression consists of the optimization of the $\log-$likelihood function subject to a constrained L_1-norm of the model parameters. It was originally introduced by Tibshirani [1996] to simultaneously perform feature selection in linear regression models. This

strategy belongs to one of the most frequently used methods for the analysis of high-dimensional data such as gene expression data (Shevade and Keerthi [2003], Liu et al. [2007]). The LASSO penalization of model parameters has been extended to generalized linear models (Park and Hastie [2007]).

The L_1-norm penalization shrinks irrelevant predictors to zero and provides sparse classifiers. The logistic regression model with LASSO penalty is computed by solving the optimization problem

$$[\hat{\beta}, \hat{\beta}_0] = \underset{\beta_0,\beta}{argmin} \sum_{i=1}^{n} \Big\{ -y_i \log(p_i) - (1-y_i)\log(1-p_i) \Big\} + \lambda \|\beta\|_1, \tag{17}$$

where $\|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|$ denotes the L_1-norm of the parameter vector β and $\lambda > 0$ is a penalty parameter. The optimal λ is specified using a cross-validation methodology.

Equation 17 is a convex optimization problem since the log-likelihood is concave and $\|\beta\|_1$ convex. However $\|\beta\|_1$ is non-differentiable at 0. Generic methods for non-differentiable convex problems can be used, such as the ellipsoid method or sub-gradient methods Shor et al. [1985]. These methods are usually very slow in practice (Koh et al. [2007]). The SmoothL1 uses a smooth approximation of the L_1-norm such as $\sum_{i=1}^{p} |\beta_i| \approx \sum_{i=1}^{p} (\beta_i^2 + \epsilon)^{1/2}$ (Whittle [1971]) and solves the problem with traditional gradient based methods. The LASSO estimates can be interpreted as posterior estimates, when the β_is have independent identical Laplace priors (Tibshirani [1996]). The parameters are updated using the prior distribution

$$P(\beta|\lambda) = \prod_{i=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\beta_i|\right)$$

and likelihood as defined above (Genkin et al. [2007]).

3.3 Classification under Constrained Sensitivity

In this section, we investigate a new approach in building binary classification rules, while entirely controlling the sensitivity. The main idea is to incorporate into the optimization equation the information that only classifiers with at least a pre-determined sensitivity are acceptable. Specifically, the objective function of classification is optimized subject to the constraint that the sensitivity belongs to the pre-determined interval of admissible values. This results in explicitly searching for the largest specificity in the interval of ethically and economically acceptable values of the sensitivity value. The L_1-norm penalization of the model parameters is added to the optimization equation for simultaneous selection of relevant predictors which encourage large sensitivity values. This technique is applied to the optimization of two objective functions of classification: The log-likelihood function of the logistic regression with LASSO penalty and a Youden-index based objective function of the sum of the sensitivity and specificity. We start this section with some basic terminologies about the appraisal of classification rules.

3.3.1 Classifiers Evaluation

Consider a complete data set of n observations (patients with known status). Assume these observations consist of n^+ positive (diseased) and n^- negative (healthy) patients. The standard procedure in evaluating classifier performance is cross-validation: The data set is randomly divided into M equal sub-data sets. $M-1$ parts are used to construct a classifier and the remaining data set is used to test that classifier. In the test stage, the known disease status is assumed as unknown and is predicted by using the built classification rule. This procedure is repeated M times, so that each observation is used once in the training and once in the test stages. This procedure is repeated M times, so that each observation is used once in the training and test stages. This is called $M-$fold cross-validation, where M is usually set to 10. The true labels and the corresponding predictions constitute raw materials for the classifier evaluation.

Often, classification methods do not directly estimate the class label but predict a decision score such as disease probabilities. To turn back to class labels, a cut-off is required for the dichotomization of such decision scores. Figure 20 is an example of a cross-validated data set of patients with known diseased status. A cut-off (δ) is selected between the minimum and the maximum of the decision score. Here, $\delta \in [0, 1]$ and patients with disease probability greater than δ are labelled as diseased and

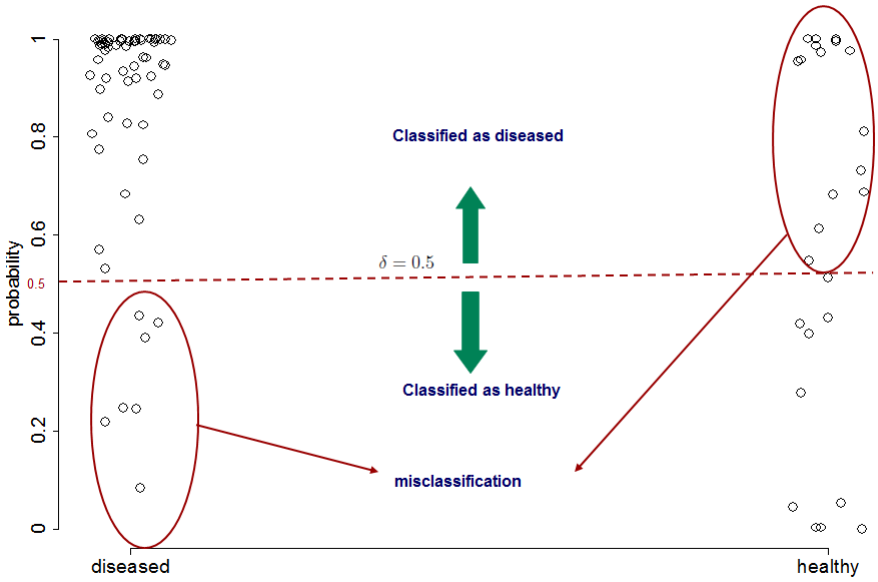

Figure 20: Graphical 2 x 2 confusion matrix representing diseased patients (left) and healthy patients (right) with their predicted disease probability.

healthy otherwise. In the example presented in Figure 20, δ is set to 0.5 corresponding to the dotted line leading to the four possible results (TP, TN, FP and FN). The success of classification in each class as well as the overall accuracy of the classifier depends on the cut-off used. A wide range of performance measurements as defined in Appendix (A.2) have been used for the evaluation of the goodness of fit of classifiers. This includes: the Sensitivity and Specificity, the Accuracy, Positive predictive value (PPV) and Negative predictive values (NPV). The use of sensitivity and specificity instead of accuracy facilitates the selection of the best threshold between the misclassification in the two classes.

Receiver Operating Characteristic

The Receiver Operating Characteristic (ROC) curve is the most commonly used graphical tool for the visualization and comparison of classifiers. Each cut-off value leads to one value of sensitivity and specificity. A ROC curve is plotting the sensitivity versus 1-specificity. For example, a ROC curve will be obtained by moving the dotted line from the deepest position to the highest (Figure 20) and by representing the results graphically, since each position corresponds the one value of the sensitivity and specificity.

Figure 21 represents an example of a ROC curve. The expected curve for a random

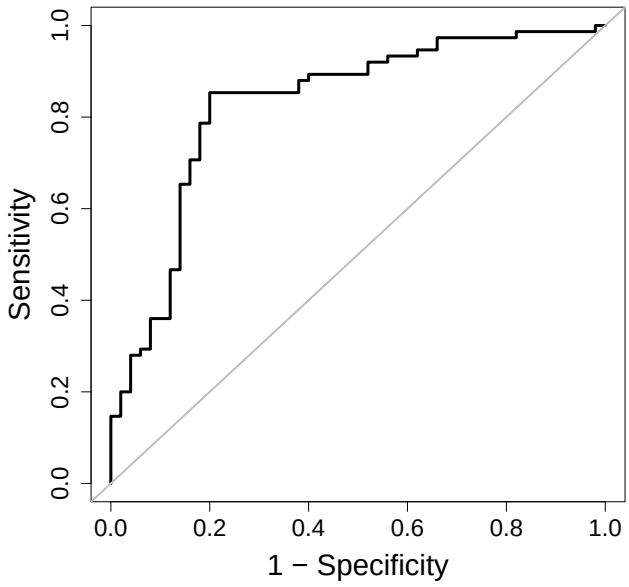

Figure 21: Example of ROC curve.

assignments of classes corresponds to the first bisector (grey line).

Classifiers can be compared through their ROC curves. The best will provide the ROC with the largest AUC (Area Under the Curve). In one-dimensional settings, where the predictor is used as the decision score, the greater the AUC is is, the better that predictor is as a diagnostic marker. The AUC has been used to rank predictors in high-dimensional data analysis. It corresponds to the probability that the predicted disease probability of a randomly selected diseased patient is higher than that of a randomly selected healthy patient. It is computed by

$$AUC = \frac{\sum_{i \in n^+} \sum_{j \in n^-} I(f(x_i^+) > f(x_j^-))}{n^+ n^-},$$

where I denotes the indicator function and f the score function in ranking positive (x_i^+) and negative samples (x_i^-).

Optimal Cut-off Selection

Each cut-off value leads to one classification result. To select the best cut-off that can be used for new patients, many techniques have been used. The cut-off that maximizes

the Youden-index given by

$$\frac{TP}{n^+} + \frac{TN}{n^-} - 1 = \text{sensitivity} + \text{specificity - 1}$$

is the most popular optimal cut-off. It leads to the point on the ROC curve most distant from the first bisector. Another technique is to select the cut-off leading to the point on the ROC curve closest to the point with coordinates $(0, 1)$. It minimizes the distance to $(0, 1)$ defined by

$$D = \sqrt{\left(\frac{TP}{n^+} - 1\right)^2 + \left(\frac{TN}{n^-} - 1\right)^2}.$$

In the best situation, the ROC curve passes through the upper left corner (point with coordinates (0,1)) and leads to 100% sensitivity and 100% specificity. All these approaches seek to find the balance between the true classification rate in the different classes while keeping the accuracy as high as possible.

3.3.2 Sensitivity and Specificity Approximation

The following approximation of the sensitivity and specificity are important for the next section. Let us consider a training data set with n samples (n^+ positive and n^- negative samples).

Definition 3.1 *The sensitivity is defined as*

$$Sens(\beta_0, \beta) = \frac{1}{n^+} \sum_{i \in n^+} I(p_i > \delta),$$

where I represents the indicator function taking the value one, if $p_i > \delta$ and zero otherwise and δ denotes a probability cut-off. The sensitivity as defined above is a function of the model parameters through $p_i = 1/[1 + \exp(-\beta_0 - X\beta)]$. Analogously, the specificity is given by

$$Spec(\beta_0, \beta) = \frac{1}{n^-} \sum_{i \in n^-} I(p_i \leq \delta).$$

The sensitivity and specificity as defined above are non-differentiable functions of the model parameters due to the indicator function. Liu and Tan [2008] suggested to use the generalized logistic function to approximate the indicator function.

$$I(z > 0) \approx g(\eta z) = \frac{1}{1 + \exp(-\eta z)},$$

where $\eta \geq 1$ is a constant that controls the trade-off between the smoothness of the approximation and the convergence to the indicator function. The function $\frac{1}{1+\exp(-\eta z)}$

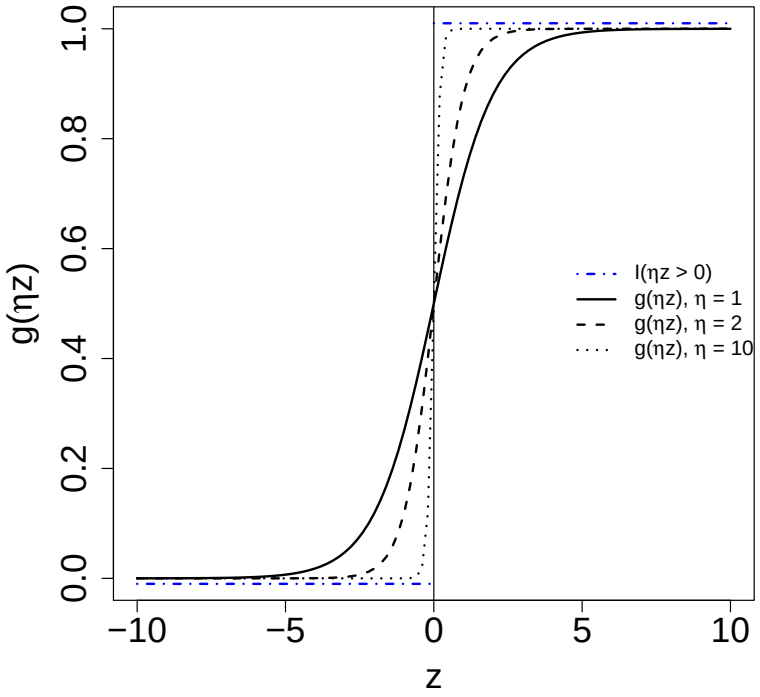

Figure 22: Indicator function approximation through the generalized logistic function.

converges to $I(z > 0)$ as η tends to infinity (see Figure 22 for different values of η). The indicator function is represented by the blue lines. A differentiable approximation of the sensitivity can be derived as

$$Sens(\beta_0, \beta) \approx \frac{1}{n^+} \sum_{i \in n^+} g[\eta(p_i - \delta)], \tag{18}$$

where $\lim_{\eta \longrightarrow \infty} g[\eta(p_i - \delta)] = I(p_i - \delta > 0)$. This is equivalent in the linear predictor space to using the cut-off $\log\left(\frac{\delta}{1-\delta}\right) \in \mathbb{R}$ instead of δ in the probability space. In other words, the sensitivity can be defined in terms of the linear predictor by

$$Sens(\beta_0, \beta) \approx \frac{1}{n^+} \sum_{i \in n^+} g[\eta(\beta_0 + X_i\beta - \log\left(\frac{\delta}{1-\delta}\right))]. \tag{19}$$

This last definition of the sensitivity corresponds to the approximation used by Liu and Tan [2008] for $\delta = 0.5$.

Similarly, the specificity is approximated by

$$Spec(\beta_0, \beta) \approx \frac{1}{n^-} \sum_{i \in n^-} g[-\eta(p_i - \delta)]$$
$$= 1 - \frac{1}{n^-} \sum_{i \in n^-} g[\eta(\beta_0 + \beta X_i - \log(\frac{\delta}{1-\delta}))]. \quad (20)$$

3.3.3 Sensitivity-preferred Logistic Regression with LASSO Penalty

The log-likelihood function of the logistic regression model can be denoted by

$$L(\beta_0, \beta) = y_i \log(p_i) + (1 - y_i) \log(1 - p_i). \quad (21)$$

The new strategy is to solve the logistic regression with LASSO penalty subject to the constraint that the sensitivity belongs to the interval of admissible values. The optimization equation is given by

$$\underset{\beta_0, \beta}{argmin} - L(\beta_0, \beta) + \lambda \|\beta\|_1 \quad \text{subject to}$$
$$Sens(\beta_0, \beta) \geq \theta, \quad (22)$$

where θ denotes the lower boundary of the interval of acceptable sensitivity values for example, 90% if the sensitivity must be greater than 90%, $\lambda > 0$ denotes the LASSO parameter and $Sens(\beta_0, \beta)$ the sensitivity. The L_1-norm of the parameter vector is given by $\|\beta\|_1 = \sum_{j=1}^{k} |\beta|_j$. $L(\beta_0, \beta)$ denotes the likelihood function as given by equation 21.

The functions $\|\beta\|_1$ and $Sens(\beta_0, \beta)$ are non-differentiable functions of the models parameters. The optimization of the problem 22 using gradient based algorithms requires differentiable approximation of these functions. The L_1-norm is approximated by

$$\|\beta\|_1 \approx \sum_{j=1}^{k} (\beta^2 + \epsilon)^{1/2},$$

where $\epsilon > 0$ controls the trade-off between the differentiability and the convergence to the L_1-norm. The sensitivity is approximated as given in equation 19. A differentiable optimization of the problem 22 is given by:

$$\underset{\beta_0, \beta}{argmin} - L(\beta_0, \beta) + \lambda \sum_{j=1}^{k} (\beta^2 + \epsilon)^{1/2} \quad \text{subject to}$$
$$\frac{1}{n^+} \sum_{i \in n^+} g[\eta(\beta_0 + X_i \beta - \log(\frac{\delta}{1-\delta}))] \geq \theta. \quad (23)$$

The problem (23) is a non-linear constrained optimization problem. The gradient of the objective function is given by

$$\frac{\partial -L(\beta_0,\beta)+\lambda\sum_{j=1}^{k}(\beta^2+\epsilon)^{1/2}}{\partial\beta_j} = \begin{cases} -\sum_{i\in n} X_{ij}(y_i - p) & \text{if} \quad j = 0 \\ -\sum_{i\in n} X_{ij}(y_i - p) + \lambda\beta_j(\beta_j^2+\epsilon)^{-1/2} & \text{if} \quad j > 0. \end{cases}$$

and the Jacobian matrix of the vector of inequality functions is the one-row matrix defined as

$$\frac{\partial Sens(\beta_0,\beta)}{\partial\beta} = \frac{1}{n^+}\left(X^T p(1-p)(1-g[\eta(p-\delta)])g[\eta(p-\delta)]\right)^T.$$

We use the logarithmic-barrier (log-barrier) methods optimization method.

3.3.4 Log-barrier for the Sensitivity-preferred Problem

The log-barrier method is an interior point method that forces the solution path of the optimization algorithm to be in the feasible region (where constraints are satisfied) (see Wright [1992]). The logarithmic barrier will force $S(\beta_0,\beta) - \theta$ to always be positive through the log-function and thus the inequality $Sens(\beta_0,\beta) > \theta$ will be satisfied.
The log-barrier function is defined as

$$B(\beta_0,\beta) = -\log\left[Sens(\beta_0,\beta) - \theta\right].$$

This function is used as a penalty term in the optimization equation 22. The barrier function $B(\beta_0,\beta)$ will grow very fast when the solution approaches the boundary of the feasible region. The unconstrained version of the optimization problem 22 is given by

$$\underset{\beta_0,\beta}{argmin} - L(\beta_0,\beta) + \lambda\sum_{j=1}^{k}(\beta^2+\epsilon)^{1/2} + \nu B(\beta_0,\beta), \tag{24}$$

where $\nu > 0$ denotes the barrier parameter and $\lambda > 0$ the LASSO parameter. The solution of this problem $(\beta,\beta_0)(\nu)$ is a function of the barrier parameter and converges to the solution of the problem 22 for $\nu \to 0$ (Whittle [1971]).

Theorem 3.2 *The optimization problem (24) is convex for all cut-off δ, so that $p_i \geq \delta$ for $i \in n^+$.*

Proof 3.3 *It is sufficient to prove that each of the three parts of the equation is convex since a linear combination with non-negative coefficients of convex functions is convex.*

- The log-likelihood function L of the logistic regression is known to be a concave function. It can be verified that the Hessian matrix is negative semi-definite. This

matrix is given by

$$L_{ij} = \frac{\partial^2 L(\beta_0, \beta)}{\partial\beta_i \partial\beta_j} = -\sum_{m=1}^{n} X_{mi} X_{mk} \frac{\exp(-\beta_0 - X_m\beta)}{(1+\exp(-\beta_0 - X_m\beta))^2}$$
$$= -\sum_{m=1}^{n} X_{mi} X_{mk} p_m (1-p_m),$$

where $p_m = \frac{1}{1+exp(-\beta_0 - X_m\beta)}$ and X_m denotes the vector of predictor values of the m^{th} observation and X_{mj} the j^{th} component of the vector X_m. The matrix $L \in \mathbb{R}^{(k+1)\mathrm{x}(k+1)}$ is negative semi definite, since for each vector $a \in \mathbb{R}^{(k+1)}$,

$$\begin{aligned} aLa &= -\sum_{m=1}^{n}\sum_{j=0}^{k}\sum_{i=0}^{k} a_i a_j X_{mi} X_{mj} p_m (1-p_m) \\ &= -\sum_{m=1}^{n}\sum_{j=0}^{k}\sum_{i=0}^{k} a_i X_{mi}\sqrt{p_m(1-p_m)}\, a_j X_{mj}\sqrt{p_m(1-p_m)} \\ &= -\sum_{m=1}^{n} a^T \pi_m \pi_m^T a \\ &= -\sum_{m=1}^{n} (a^T \pi_m)^2 \leq 0, \end{aligned}$$

where $\pi_m = (p_m(1-p_m))^{1/2} X_m$.
Hence, $-L(\beta_0, \beta)$ is convex.

- The function $\sum_{j=1}^{k}(\beta_j^2 + \epsilon)^{1/2}$ is also convex. The Hessian matrix is given by

$$N_{ij} = \frac{\partial^2}{\partial\beta_i\partial\beta_j}\left[\sum_{m=1}^{k}(\beta_m^2+\epsilon)^{1/2}\right] = \begin{cases} 0 & \text{if} \quad i \neq j \\ [(\beta_j^2+\epsilon)^2 - \beta_j^2](\beta_j^2+\epsilon)^{-3/2} & \text{if} \quad i = j \end{cases}$$

and

$$aNa = \sum_{j=1}^{k} a_j^2 [(\beta_j^2+\epsilon)^2 - \beta_j^2](\beta_j^2+\epsilon)^{-3/2} \geq 0.$$

Thus, this differentiable approximation of the L_1-norm of the model parameters is convex.

- Also $B(\beta_0, \beta) = -\log\left(\frac{1}{n^+}\sum_{i\in n^+} g[\eta(\beta_0 + X_i\beta - \log(\frac{\delta}{1-\delta}))] - \theta\right)$ is convex. Since the log is monotone non-decreasing, it remains to show that the function $G(\beta_0, \beta) = \left(\frac{1}{n^+}\sum_{i\in n^+} g[\eta(\beta_0 + X_i\beta - \log(\frac{\delta}{1-\delta}))] - \theta\right)$ is concave. The Hessian matrix is defined as

$$G_{ij} = \frac{\partial^2 G(\beta_0, \beta)}{\partial\beta_j\partial\beta_j} = \frac{\eta^2}{n^+}\sum_{m\in n^+} X_{mi} X_{mj} g_m (1-g_m)(1-2g_m)$$

where $g_m = g[\eta(\beta_0 + X_m\beta - \log\frac{\delta}{1-\delta})]$. Similar to the matrix L, the sign of aGa is determined as follows

$$\begin{aligned} aGa &= \frac{\eta^2}{n^+}\sum_{m\in n^+}\sum_{i=0}^{k}\sum_{j=0}^{k} a_i a_j X_{mi} X_{mj} g_m(1-g_m)(1-2g_m), \\ &= -\frac{\eta^2}{n^+}\sum_{m\in n^+}(a^T\pi_m)^2, \end{aligned}$$

where $a \in \mathbb{R}^{(k+1)}$ and $\pi_{ij} = X_{ij}\sqrt{g_m(1-g_m)(1-2g_m)}$. aGa is positive for $1 - 2g_m \geq 0$. This is the case when $g_m \geq 0.5$ which is equivalent to $p_i \geq \delta$. In this case, $G(\beta_0, \beta)$ is concave and since $-\log$ function is convex, $B(\beta_0, \beta)$ is convex $\square$

The convexity of equation (24) is not guaranteed, when there are some $p_i < \delta$ for $i \in n^+$. However, aGa consists of a true positive and a false negative part and can be written as

$$\begin{aligned} aGa = \frac{\eta^2}{n^+} &\underbrace{\sum_{m\in n^+\cap g_m\geq 0.5}\sum_{i=0}^{k}\sum_{j=0}^{k} a_i a_j X_{mi} X_{mj} g_m(1-g_m)(1-2g_m)}_{\text{True positive}} \\ + \frac{\eta^2}{n^+} &\underbrace{\sum_{m\in n^+\cap g_m< 0.5}\sum_{i=0}^{k}\sum_{j=0}^{k} a_i a_j X_{mi} X_{mj} g_m(1-g_m)(1-2g_m).}_{\text{False positive}}, \end{aligned}$$

The convexity will be obtained when the true positive part exceeds the false negative one. This could easily be achieved by selecting a small cut-off value and when the two classes are relatively well separable in the training data.

The optimization problem (24) requires a very large computational capacity in high-dimensional settings such as gene expression data. Tibshirani et al. [2012] and Ghaoui et al. [2012] proposed the way to derive conditions for pre-selection of relevant predictors in LASSO type problems. We derive rules for the identification of irrelevant predictors, in our model, that should be ignored in the optimization procedure to save computation time based on the *strongrule* as suggested by Tibshirani et al. [2012]. Let us consider the following model with the original definition of the L_1-norm to fully exploit its property of setting some predictors to zero.

$$\underset{\beta_0,\beta}{argmin} \; -f(\beta_0, \beta) + \lambda\|\beta\|_1,$$

where $L(\beta_0, \beta) + \nu \log\left(\frac{1}{n^+}\sum_{i\in n^+} g[\eta(p_i - \delta)] - \theta\right) = f(\beta_0, \beta)$. The solution to this optimization problem $\hat{\beta}(\lambda) = (\hat{\beta}_0(\lambda), \hat{\beta}_1(\lambda), \cdots, \hat{\beta}_k(\lambda))$ depends on the tuning parameter λ. The

optimal λ is provided by a cross validation in a grid search in the interval $[\lambda_{min}, \lambda_{max}]$. λ_{max} denotes the smallest λ for which all model parameters are set to zero except the intercept. That means, the solution at λ_{max} has the form $\hat{\beta}(\lambda_{max}) = (\hat{\beta}_0, 0, \cdots, 0)$ and the corresponding estimated probability vector is given by $p(\hat{\beta}(\lambda_{max})) = \bar{p} = 1_n\bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. On the other hand, λ_{min} is the largest positive number for which no predictor is set to zero. Let us consider the sequence of λ values $\lambda_1, \lambda_2 \cdots \lambda_m$.

Theorem 3.4 *Given the solution at λ_{k-1} ($\hat{\beta}(\lambda_{k-1})$), the irrelevant predictors to be discarded at λ_k must satisfy the inequality*

$$\left|\frac{\partial f}{\partial \beta_j}(\hat{\beta}(\lambda_{k-1}))\right| < 2\lambda_{k-1} - \lambda_k.$$

This provides the solution path in varying λ from λ_1 to λ_m. $\frac{\partial}{\partial \beta_j} f[\hat{\beta}(\lambda_{k-1})]$ denotes the partial derivative of f with respect to β_j evaluated at $\hat{\beta}(\lambda_{k-1})$. Theorem 3.4 was stated by Tibshirani et al. [2012], who called it sequential strong rule for discarding predictors.

Proof 3.5 *See Tibshirani et al. [2012].*

Note that the Karush-Kuhn-Tucker (KKT) first order conditions for the above optimization problem are given by

$$\begin{aligned} &\frac{\partial[-f(\beta_0, \beta) + \lambda\|\beta\|_1]}{\partial \beta_j} = 0 \\ \implies &\frac{\partial f(\beta_0, \beta_j)}{\partial \beta_j} = \lambda \frac{\partial |\beta_j|}{\partial \beta_j}, \end{aligned} \tag{25}$$

where

$$\frac{\partial(|\beta_j|)}{\partial \beta_j} \in \begin{cases} \{+1\} & \text{if} \quad \beta_j > 0 \\ \{-1\} & \text{if} \quad \beta_j < 0 \\ [-1, +1] & \text{if} \quad \beta_j = 0. \end{cases}$$

This condition must be satisfied by each optimal solution. We derive the condition given in Theorem 3.4 for the optimization problem 24.

$$\frac{\partial}{\partial \beta_j} \log\left[Sens(\beta_0, \beta) - \theta\right] = \frac{\frac{\eta}{n^+} X_j^T p(1-p)(1 - g[\eta(p-\delta)])g[\eta(p-\delta)]}{\frac{1}{n^+}\sum_{i \in n^+} g[\eta(p_i - \delta)] - \theta}$$

$$\frac{\partial}{\partial \beta_j} L(\beta_0, \beta) = X_j^T(y - p).$$

The partial derivative of $f(\beta_0, \beta)$ as defined above is given by

$$\frac{\partial}{\partial \beta_j} f(\beta_0, \beta) = \frac{\partial}{\partial \beta_j} L(\beta_0, \beta) + \nu \frac{\partial}{\partial \beta_j} \log \Big[Sens(\beta_0, \beta) - \theta \Big]$$

and then

$$\left| \frac{\partial}{\partial \beta_j} f(\beta_0, \beta) \right| = \left| X_j^T (y - p) + \nu \frac{\frac{\eta}{n^+} X_j^T p(1-p)(1 - g[\eta(p-\delta)])g[\eta(p-\delta)]}{\frac{1}{n^+} \sum_{i \in n^+} g[\eta(p_i - \delta)] - \theta} \right|.$$

This last equation provides the sequence strong rule for the optimization problem (24) as given by Theorem 3.4. In the special case where $k-1$ is set to a constant equal to λ_{max}, at which all parameters are set to zero, the conditions for discarding irrelevant predictors are given by

$$\left| X_j^T (y - \bar{p}) + \nu \frac{\frac{\eta}{n^+} X_j^T \bar{p}(1-\bar{p})(1 - g[\eta(\bar{p}-\delta)])g[\eta(\bar{p}-\delta)]}{\frac{1}{n^+} \sum_{i \in n^+} g[\eta(\bar{p}_i - \delta)] - \theta} \right| < 2\lambda - \lambda_{max}, \qquad (26)$$

where

$$\lambda_{max} = \max_j \left(\left| X_j^T (y - \bar{p}) + \nu \frac{\frac{\eta}{n^+} X_j^T \bar{p}(1-\bar{p})(1 - g[\eta(\bar{p}-\delta)])g[\eta(\bar{p}-\delta)]}{\frac{1}{n^+} \sum_{i \in n^+} g[\eta(\bar{p}_i - \delta)] - \theta} \right| \right).$$

Note that the first part $X_j^T(y - \bar{p})$ is computed using the complete data matrix, while the second part comes from the definition of the sensitivity and should be computed using only the positive samples. Inequality 26 represents the condition satisfied by irrelevant predictors under model 22. The feature selection stage is connected to the model. The selected predictors lead to classifiers with sensitivity larger than θ, while maximizing the likelihood function.

3.3.5 Specificity Maximization under Constrained Sensitivity

In this section, we investigate a new classification method based on the optimization of the specificity subject to the constraint that the sensitivity belongs to a pre-determined interval of large values such as $[90\%, 100\%]$. The loss-function of $sens + spec$ has been suggested by Liu and Tan [2008]. The consideration of the sensitivity and specificity in defining loss-functions of classification is motivated by the fact, that the goal in building binary classifiers is to achieve the largest Youden index ($sens + Spec + 1$). Thus this can be directly considered as a loss-function to be optimized instead of considering the likelihood function. Here, we optimize only the specificity since the sensitivity is already constrained to the admissible interval. This optimization problem provides the largest specificity regardless of the value of the sensitivity belonging to the pre-specified

interval. The L_1-norm penalization of the model parameters is used to perform the feature selection simultaneously. The model is defined as

$$\underset{\beta_0,\beta}{argmin}\left\{-Spec(\beta_0,\beta)+\lambda\|\beta\|_1\right\} \quad \text{subject to}$$
$$Sens(\beta_0,\beta)\geq\theta, \tag{27}$$

where θ denotes the lower bound of the sensitivity, $\beta\in\mathbb{R}^k, \beta_0\in\mathbb{R}$ the model parameters and λ the LASSO parameter.

The gradient of the objective function is given by

$$\frac{\partial -Spec(\beta_0,\beta)+\lambda\|\beta\|_1}{\partial\beta_j} =$$
$$\begin{cases} \frac{1}{n^-}\eta X_j^T\bar{p}_{n^-}(1-\bar{p}_{n^-})g[-\eta(\bar{p}_{n^-}-\delta)](1-g[-\eta(\bar{p}_{n^-}-\delta)]) & \text{if} \quad j=0 \\ \frac{1}{n^-}\eta X_j^T\bar{p}_{n^-}(1-\bar{p}_{n^-})g[-\eta(\bar{p}_{n^-}-\delta)](1-g[-\eta(\bar{p}_{n^-}-\delta)])+\lambda\beta_j(\beta_j^2+\epsilon)^{-1/2} & \text{if} \quad j>0. \end{cases}$$

The sensitivity and specificity are functions of the model parameters as given by equation 19 and equation 20 respectively. Maximizing the specificity subject to the constraint that the sensitivity is large is comparable to the maximization of the sum of sensitivity and specificity subject to the same constraint. This last optimization problem under LASSO penalty is given by

$$\underset{\beta_0,\beta}{argmin}\left\{-Sens(\beta_0,\beta)-Spec(\beta_0,\beta)+\lambda\|\beta\|_1\right\} \quad \text{subject to}$$
$$S(\beta_0,\beta)\geq\theta.$$

The above equations are optimization problems with a non-linear constraint. These can be solved similarly to the optimization of the likelihood as presented in the previous section. We use the log-barrier method as described in the previous section. Note that these are non-convex problems and that the algorithm should be started at many different points to increase the likelihood of convergence. In solving problem 27, irrelevant predictors satisfy the conditions

$$\left|\frac{1}{n^-}\eta X_j^T\bar{p}_{n^-}(1-\bar{p}_{n^-})g[-\eta(\bar{p}_{n^-}-\delta)](1-g[-\eta(\bar{p}_{n^-}-\delta)])\right.$$
$$\left.+\frac{\frac{\nu}{n^+}\eta X_j^T\bar{p}_{n^+}(1-\bar{p}_{n^+})g[\eta(\bar{p}_{n^+}-\delta)](1-g[\eta(\bar{p}_{n^+}-\delta)])}{\frac{1}{n^+}\sum_{i\in n^+}g[\eta(\bar{p}_{n^+}-\delta)]-\theta}\right|\leq 2\lambda-\lambda_{max},$$

where $\bar{p}_{n^-}=1_{n^-}\bar{y}_{n^-}$, $\bar{p}_{n^+}=1_{n^+}\bar{y}_{n^+}$, $\bar{y}_{n^+}=\frac{1}{n^+}\sum_{i\in n^+}y_i$ and $\bar{y}_{n^-}=\frac{1}{n^-}\sum_{i\in n^-}y_i$,

$$\lambda_{max}=\max_j\left|\frac{1}{n^-}\eta X_j^T\bar{p}_{n^-}(1-\bar{p}_{n^-})g[-\eta(\bar{p}_{n^-}-\delta)](1-g[-\eta(\bar{p}_{n^-}-\delta)])\right.$$
$$\left.+\frac{\frac{\nu}{n^+}\eta X_j^T\bar{p}_{n^+}(1-\bar{p}_{n^+})g[\eta(\bar{p}_{n^+}-\delta)](1-g[\eta(\bar{p}_{n^+}-\delta)])}{\frac{1}{n^+}\sum_{i\in n^+}g[\eta(\bar{p}_{n^+}-\delta)]-\theta}\right|.$$

The standard approach in finding classifiers with a pre-specified sensitivity value is to move the ROC curve up to the point that leads to that sensitivity. That means, the cut-off is varied until the pre-determined sensitivity is achieved regardless of the specificity. If there are many cut-offs that provide the same sensitivity, the optimal one leads to the highest specificity. An illustration of this approach is given by the following example. We simulate 30 values from the Gamma distributions $\Gamma(10, 2)$ and $\Gamma(15, 2)$, respectively. These values can be viewed as protein levels of a specific protein measured on 30 healthy patients ($\Gamma(10, 2)$) and 30 diseased patients ($\Gamma(15, 2)$). To evaluate how well this protein can discriminate between the two classes, the following ROC-curve is plotted by varying the cut-off on the whole values. Assume that a 95% sensitivity is required for the results

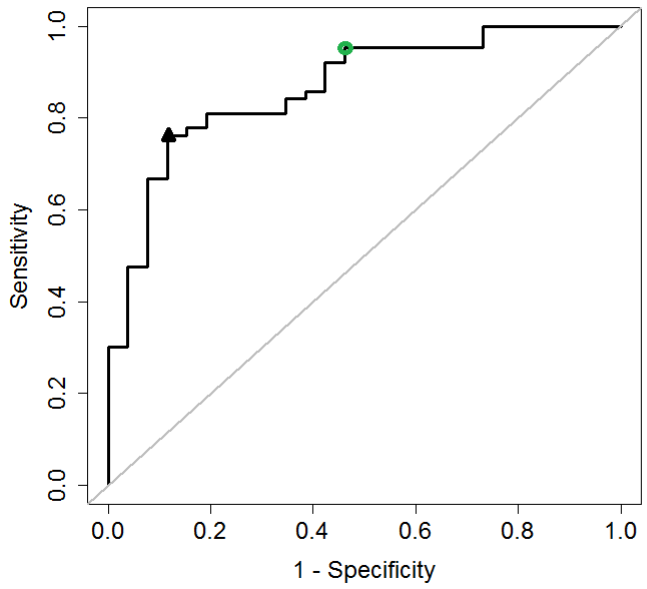

Figure 23: Sensitivity-preferred cut-off represented by the green dot and the cut-off maximizing the Youden-index represented by the black triangle.

presented in Figure 23. To achieve this objective, the cut-off leading to the green point is selected and corresponds to 95% sensitivity and 58% specificity, while the black point corresponding to the maximization of the Youden-index provides 74% sensitivity and 87% specificity. The interpretation of these results depends on the class importance. The gain of 21 percentage points in sensitivity provided by this technique may be more beneficial than the 29 percentage points decline in the specificity.

3.4 Overview of Classification Strategies by Different Class Importance

This section presents a literature overview of some techniques that have been used to improve the true classification rate in the most important class in building binary classifiers. These techniques favor the sensitivity, which leads to a decrease in specificity. Unlike the strategy of constrained optimization presented above, these methods cannot guarantee the achievement of a pre-determined sensitivity, while guaranteeing the optimal specificity.

Cost-sensitive methods take in to account the different misclassification costs in building classifiers. They can be applied to find the best cut-off on the already computed decision score vector or earlier during the definition of the objective function of classification. Consider the following cost matrix: $c(i|j) = c_{ij}$ represents the cost of classifying

Prediction	Really positive = 1	Really negative = 0
Positive =1	$c(1\|1) = c_{11}$	$c(1\|0) = c_{10}$
Negative =0	$c(0\|1) = c_{01}$	$c(0\|0) = c_{00}$

Table 6: Cost matrix of binary classification.

a sample in class i given its true class is j. The expected costs of classifying a sample x in the positive and negative class are given respectively by

$$\begin{aligned} E(X,1) =& p(y=1|X=x)c_{11} + p(y=0|X=x)c_{10} \\ E(X,0) =& p(y=0|X=x)c_{01} + p(y=1|X=x)c_{00}, \end{aligned}$$

where $p(y=i|X=x)$ denotes the probability of class i to be the true class of the sample x. Elkan [2001] suggested selecting the optimal cut-off by minimizing the expected costs. A sample x is classified as positive if $E(X,1) \leq E(X,0)$. By solving this inequality in $p(y=1|X=x)$ and by replacing $p(y=0|X=x)$ with $1 - p(y=1|X=x)$, the cost-sensitive cut-off δ_{sc} is obtained and takes the form:

$$\delta_{sc} = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}}.$$

Usually, $c_{00} = c_{11} = 0$ and then $\delta = \frac{c_{10}}{c_{10}+c_{01}}$. The probability cut-off for equal misclassification costs is 0.5 regardless of the class-frequencies. This may be problematic in the

case of strongly unbalanced data and could severely disadvantage the smaller of the two classes.

Misclassification costs can also be considered by using the thresholding strategy as investigated by Sheng and Ling [2006]. It consists of minimizing the total misclassification costs instead of the expected costs. Each cut-off results in $TP, TN, FP,\ FN$ and a total misclassification cost denoted by $M_c = c_{11}TP + c_{00}TN + c_{01}FN + c_{10}FP$. The optimal cut-off which yields to the minimal value of M_c is determined empirically. However, misclassification costs are rarely available and a cost sensitive cut-off cannot guarantee the achievement of a pre-determined sensitivity value.

Weighting of the likelihood function belongs to the so called direct cost-sensitive learning procedures. Each sample is weighted proportionally to the misclassification costs in its class before being introduced into the optimization algorithm. Often, numbers are chosen arbitrarily to weight samples. However, the misclassification costs remain the most interpretable weights in this context. Weighting samples before introducing them to the learning algorithm has been applied to many classification and regression methods. The weighted likelihood function in the case of the logistic regression is given by

$$L_\omega(\beta|X) = \prod_{i=1}^{n} p_{\omega_i}^{y_i}(1 - p_{\omega_i})^{1-y_i}, \tag{28}$$

where

$$p_{\omega_i}^{y_i} = \frac{1}{1 + \exp\left(-\beta_0 - \omega_i X_i \beta\right)}$$

and $\omega_i > 0$ denotes the weight assigned to the i^{th} sample. Weighting does not affect the convexity of the original problem (when $\omega_i = 1$, for $i = 1 \cdots n$). This technique can also be used to favor the minority class if training sets are strongly imbalanced.

Weighting the Youden-index utility function of classification offers the possibility of directly weighting the sensitivity and specificity as suggested by Liu and Tan [2008]. The weighted objective function of the Youden-index is defined by

$$\begin{aligned} T(\beta) &= \omega_1 S(\beta) + \omega_0 Sp(\beta) \\ &\quad \frac{1}{n^+}\sum_{i\in n^+} \omega_i^+ I(p_i > \delta) + \frac{1}{n^-}\sum_{i\in n^-} \omega_i^- I(p_i \leq \delta) \\ &\quad \approx \frac{1}{n^+}\sum_{i\in n^+} \omega_i^+ g[\eta(p_i - \delta)] + \frac{1}{n^-}\sum_{i\in n^-} \omega_i^- (1 - g[\eta(p_i - \delta)]). \end{aligned}$$

This optimization problem is convex only under the condition that the data are perfectly separable through δ. That means $p_i > \delta$ for $i \in n^+$ and $p_i \leq \delta$ for $i \in n^-$ (Liu and Tan [2008]).

Using misclassification costs in any way to weight samples during building of classifiers is a reasonable approach for the purpose of considering different class importance. However, an objective and accurate evaluation of misclassification costs in the medical context is a very difficult task. For example, misclassification cost evaluation in the context of patient diagnostic involves the quantification of the impact of the classification results on patient health, life quality, and life expectancy, in addition to economical considerations. In the same context of classification with different class importance, Japkowicz and al. [2000] and Seiffert et al. [2008] suggested to adjust the class frequencies by changing the baseline frequencies of the training set to favor the most important class (Sampling methods). However, none of these methods explicitly considers the relevant information that a pre-determined lower bound of the sensitivity value must hold and thus cannot guarantee the achievement of such a pre-determined acceptable sensitivity. The cut-off variation to find the acceptable sensitivity remains the unique technique that guarantees pre-specified sensitivity values and will be compared to the new strategy of constrained optimization of loss functions of binary classification as presented above.

3.5 Results

We implement the optimization equations (22) and (27) in the freeware software R (R Core Team [2013]). The package alabama (Augmented Lagrangian and Adaptive Barrier Minimization Algorithm) consists of functions for optimizing smooth non- linear objective functions with constraints. Here, linear and non-linear constraints are allowed (Ravi [2011]). The sensitivity is a non-linear function of the model parameters. We optimize the likelihood function and the specificity with LASSO penalty, subject to the constraint that the sensitivity is larger than 90%. The achieved specificity is compared to the specificity provided by the method described in Jung et al. [2010] which is to compute the disease probability, then search for the cut-off leading to 90% sensitivity and finally to derive the corresponding specificity (specificity L_1). Here, the disease probabilities are computed by using a 10 fold cross-validation of the logistic regression model with LASSO penalty. To the best of our knowledge, this method is the unique method that guarantees a sensitivity value in a pre-determined interval, aside from the trivial method of assigning all samples to the disease class. Functions in alabama require as input the objective function and its gradient as well as the inequality function and its Jacobian matrix as derived by the definition of the optimization equation. We use a 10-fold cross-validation to predict the class-probabilities and focus on the classification performance of sensitivity and specificity.

Gene expression data

We build 5 different classifiers from the gene expression data of endometriosis patients as described in the data material section by using the same number of predictors as the corresponding LASSO. The new technique is designed to return the largest specificity given that the sensitivity is larger than 90%, since the optimization is performed in that interval. The results presented in Table 7 meet this requirement. The new strategy, compared to the ROC based decision making, shows a clear gain in specificity in almost all cases (see columns specificity and specificity L_1). For example, in classifying the group of patients with pain ($G_{11} \cup G_{21}$) versus the patient population without pain ($G_{12} \cup G_{22}$), the specificity improved from 16.2% to 40.4%. There is no significant difference between the results of the likelihood optimization and the specificity optimization as evident by comparing Table 7 and Table 8. Both methods did not outperform the traditional method in the data set with G_{11} and G_{22}. This can be explained by the non-convexity of the constraint in some data situations. In this case, the algorithm must be started

Likelihood Optimization under constrained Sensitivity					
Data	sensitivity	specificity	predictors	specificity L_1	δ
$G_{11} vs. G_{12}$	0.904	0.407	20	0.216	0.68
$(G_{11} \cup G_{21}) vs. (G_{12} \cup G_{22})$	0.922	0.324	20	0.162	0.66
$G_{11} vs. (G_{12} \cup G_{21} \cup G_{22})$	0.920	0.372	20	0.117	0.62
$(G_{11} \cup G_{12}) vs. (G_{21} \cup G_{22})$	0.922	0.333	25	0.0833	0.58
$G_{11} vs. G_{22}$	0.921	0.5	10	0.571	0.7

Table 7: This table presents the specificity of the logistic regression with LASSO penalty at 90% sensitivity denoted by specificity L_1 and the value of the sensitivity and specificity provided by the new approach of constrained optimization of the likelihood function subject to the constraint that the sensitivity is larger than 90%. δ denotes the probability cut-off used in defining the sensitivity.

Specificity Optimization under constrained Sensitivity					
Data	sensitivity	specificity	predictors	specificity L_1	δ
$G_{11} vs. G_{12}$	0.904	0.370	20	0.216	0.65
$(G_{11} \cup G_{21}) vs. (G_{12} \cup G_{22})$	0.900	0.351	25	0.162	0.45
$G_{11} vs. (G_{12} \cup G_{21} \cup G_{22})$	0.904	0.372	25	0.117	0.55
$(G_{11} \cup G_{12}) vs. (G_{21} \cup G_{22})$	0.933	0.333	30	0.0833	0.5
$G_{11} vs. G_{22}$	0.900	0.5	20	0.571	0.7

Table 8: This table presents the specificity of the logistic regression with LASSO penalty at 90% sensitivity denoted by specificity L_1 and the new value of the sensitivity and specificity when the specificity function is optimized subject to the constraint that the sensitivity is greater than 90%.

at multiple points. No classification rule should outperform the new classifiers based on the way they are designed.

Protein data

We apply the new strategy to the protein data described in the data material section. To evaluate the results of the new method, the traditional LASSO is also applied and the corresponding ROC curve is plotted. We compute the specificity at 90% sensitivity and compare this with the specificity provided by the new method.

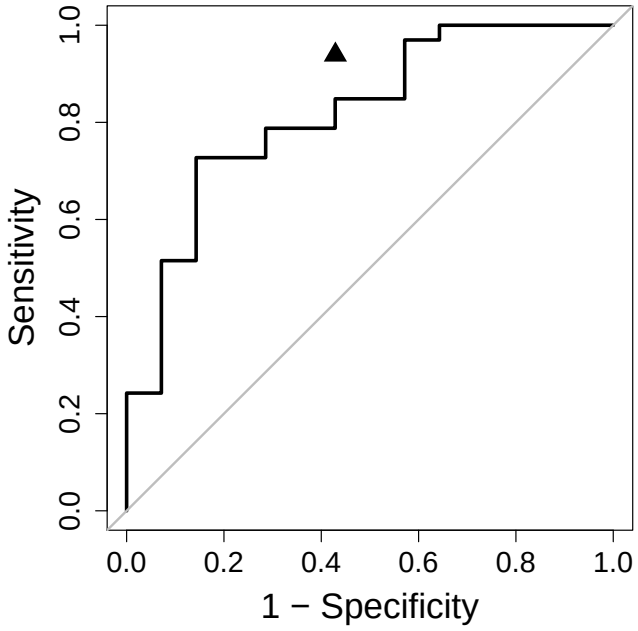

Figure 24: ROC curve representing the results of the optimization of the likelihood function subject to the constraint that the sensitivity is larger than 90%. The curve represents the results provided by LASSO and is obtained by varying the cut-off value. The new classifier is represented by the triangle.

The new strategy leads to 93.9% sensitivity and 57.1% specificity while the traditional method leads to 42.8% specificity at 90% sensitivity.
The specificity optimization provides lower classification performance compared to that of the likelihood optimization. The specificity at 90% is 50% which is larger than the 42% specificity provided by LASSO and cut-off variation.

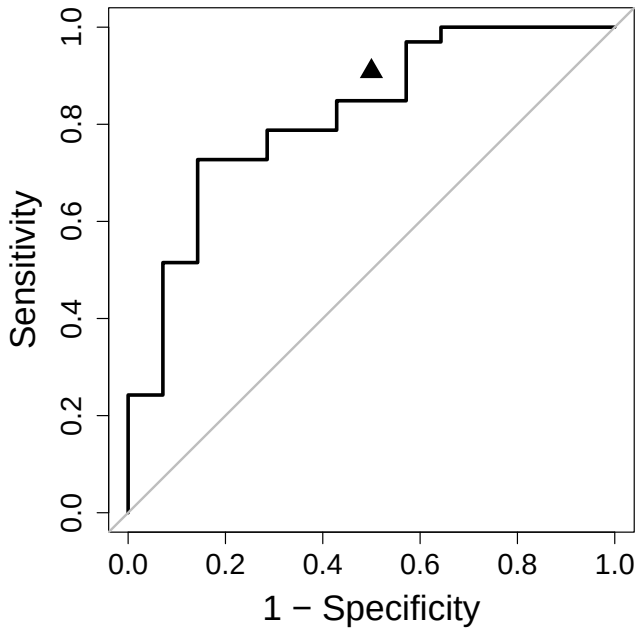

Figure 25: ROC curve representing the results of the optimization of the specificity function subject to the constraint that the sensitivity is larger than 90%. The curve represents the results provided by LASSO and is obtained by varying the cut-off value. The new classifier is represented by the triangle.

3.5.1 Outlook

The general formulation of the classification rules bases on constrained optimization of the objective function is giving by

$$\hat{\beta}_{Bridge} = \underset{\beta_0,\beta}{argmin} f(\beta_0, \beta) + \lambda \|\beta\|_\nu \qquad \text{subject to} \qquad (29)$$
$$Sens(\beta_0, \beta) \geq \theta,$$

where $\|\beta\|_\nu$ corresponds to the L_ν-norm of the model parameters with $\nu > 0$, $f(\beta_0, \beta)$ is a smooth objective function of classification, $S(\beta_0, \beta)$ denotes the sensitivity as a function of the model parameters and θ represents the lower bound of acceptable sensitivity values. We are interested in finding a convex approximation of $S(\beta_0, \beta)$ in the context of logistic regression and will investigate this strategy for other objective functions of classification.

The *Bridge* regression introduced by Frank and Friedman [1993] in the context of linear regression is the generalization of the LASSO regression involving the L_μ-norm

penalty of the model's parameters, where $\mu > 0$. We have investigated the Bridge estimates in the case of logistic regression subject to a constraint on the sensitivity. The optimization equation is defined as

$$\hat{\beta}_{Bridge} = \underset{\beta_0,\beta}{argmin} \sum_{i=1}^{n} \Big\{ -y_i \log(p_i) - (1-y_i)\log(1-p_i) \Big\} + \lambda \|\beta\|_\mu, \quad \text{subject to} \quad (30)$$
$$Sens(\beta_0, \beta) \geq \theta,$$

where $\|\beta\|_\mu = \sum_{j=1}^{k} |\beta_j|^\mu$. It contains the Ridge estimates ($\mu = 2$) which are usually used to deal with correlation between predictors (Hoerl and Kennard [1970]) as well as the more sparse logistic regression model that has been studied by Liu et al. [2007] for $0 < \mu < 1$.

4 Conclusions

The main contribution of this thesis starts in section 2, where we investigate strategies to predict patient recruitment time, study costs and power variation in enrichment studies. We assume that only patients who are classified as marker-test positive after passing the screening and marker-test procedures are enrolled for the trial. In section 2.1, we develop Poisson processes for modeling patient recruitment in enrichment trials. These processes consider the prevalence of biomarker-positive patients in the unselected patient population, the number of active clinical trial centers and their capacities, as well as the marker-test characteristics. Based on the suggested models, we derive the distributions of recruitment time analytically. Estimates of the recruitment time are relevant in determining the feasibility of a study and in planning deadlines. We examine the suggested Poisson processes under different assumptions about the recruitment rate (constant, Gamma distributed,...). Marker prevalence is one of the most relevant factors in evaluating recruitment time. The larger the marker prevalence, the shorter the recruitment time.

In section 2.2, we propose a Bayesian approach for progressively updating study components like marker prevalence and overall recruitment rate. Updates of model parameters in ongoing stage allow a more accurate prediction of the remaining time until the end of the recruitment process. Updated estimates of the remaining time may help in deciding whether to add more centers to accelerate the recruitment process. We propose appropriate priors for the change points, which can be detected by subsequently computing their posterior distributions. In section 2.3, we suggest a model for study costs evaluation that considers the screening and marker-test costs, the care costs after recruitment, and a time component. Time is a crucial cost factor which includes personnel, infrastructure, etc. We derive the distribution of patients passing through the screening and test phases, then the corresponding total expenditures as a function of the marker prevalence. The lower the marker prevalence, the longer the study time, thus the higher the costs. Some of our cost and time models consider the sensitivity and specificity of the marker-test to be set to one if they are not available. We find the impact of enrolling false positive patients may decrease the study's power.

Section 3 investigates another problem met not only in conducting enrichment studies, but more generally in dealing with binary classification. In this section, we introduce

a new approach to building binary classifiers while controlling the sensitivity as the true classification rate in the most important class. To assess binary classification rules, one must consider the difference in misclassification costs between the two classes. In some, if not most, diagnostic situations, it is crucial to include the control of the sensitivity in the classification building stage and to reject the classifiers which lead to a sensitivity under a pre-determined threshold (for example 90%). Our new strategy is based on optimizing the objective function of classification, under the constraint that the sensitivity belongs to an interval of admissible values such as $[90, 100]\%$. A traditional issue when using high-dimensional data, such as microarray gene expression data, is to reduce the number of predictors and provide classifiers based on a reasonable number of predictors. We add a LASSO penalty to the optimization equation to select relevant predictors.

Our new strategy is illustrated in section 3.3 within the context of logistic regression. Here we investigate the optimization of two different objective functions of binary classification: the likelihood function and an objective function based on the Youden-index (sensitivity+specificity-1). The sensitivity and specificity in their original form are discrete functions of the model parameters, which are then approximated to a differentiable function through the generalized logistic function. The two objective functions are optimized subject to both the constraint that the sensitivity is greater than 90% and a constraint on the L_1-norm of the model parameters. This two-constraint-optimization problem is resolved using an internal point method (log-barrier), which searches for the optimal solution exclusively in the feasible region.

Optimizing objective functions with constrained sensitivity when building classification rules in the context of disparate class importance has its advantages. It does not, for example, require the misclassification costs, which are rarely available, although it is often clear that they are different between classes. In addition, classification methods that consider the misclassification costs as an indicator of the difference in importance between the classes do not guarantee the achievement of pre-determined sensitivity values. Our approach finds classifiers with the largest specificity and the desired clinical acceptable sensitivity. This information is relevant and is therefore considered earlier in defining the optimization equation rather than later.

A traditional way of selecting classifiers with pre-determined sensitivity is to vary the cut-off on the predicted decision score vector and select the one that leads to the desired

sensitivity. In other words, the optimization is performed in classes thought to equal importance, although they do not. The optimization equations in the new approach are designed to provide the best classifier with a clinically admissible sensitivity value. This objective is supported by the real world examples detailed in the results section.

References

V. V. Anisimov. Using mixed Poisson models in patient recruitment in multicenter clinical trials. In *Proceedings of the World Congress on Engineering*, pages 1046–1049, 2008.

V. V. Anisimov. Predictive modelling of recruitment and drug supply in multicenter clinical trials. In *Proceedings of the Joint Statistical Meeting*, pages 1248–1259, 2009.

V. V. Anisimov. Discussion on the paper: Prediction of accrual closure date in multi-center clinical trials with discrete-time Poisson process models, by Gong Tang, Yuan Kong, Chung-Chou Ho Chang, Lan Kong, and Joseph P. Costantino. *Pharmaceutical Statistics*, 11(5):357–358, 2012.

V. V. Anisimov and V. V. Fedorov. Modelling, prediction and adaptive adjustment of recruitment in multicenter trials. *Statistics in Medicine*, 26(27):4958–4975, 2007.

K. Arai. Method for 3D object reconstruction using several portions of 2D images from the different aspects acquired with image scopes included in the fiber retractor. *International Journal of Advanced Research in Artificial Intelligence*, 1(9), 2012.

A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, and al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95, 2001.

A. Ben-Israel. A Newton-Raphson method for the solution of systems of equations. *Journal of Mathematical Analysis and Applications*, 15(243):1, 1966.

D. A. Berry. Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36, 2006.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: Forecasting and control*. New Jersey: John Wiley & Sons, fourth edition, 2013.

N. Brünner. What is the difference between predictive and prognostic biomarkers? Can you give some examples. *Connection*, 13:18, 2009.

R. E. Carter. Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials*, 25(5):429–436, 2004.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

M.-H. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye. Bayesian clinical trials. In *Frontiers of Statistical Decision Making and* B*ayesian Analysis*, pages 257–284. Philadelphia: Springer Science+Business Media LLC, 2010.

R. I. Chien. *Issues in pharmaceutical economics*. New York: Free Press, 1979.

K. Claxton and J. Posnett. An economic approach to clinical trial design and research priority-setting. *Health Economics*, 5(6):513–524, 1996.

C. M. J. Cline, B. Israelsson, R. B. Willenheimer, K. Broms, and L. R. Erhardt. Cost effective management programme for heart failure reduces hospitalisation. *Heart*, 80 (5):442–446, 1998.

R. Cont and P. Tankov. *Financial modelling with jump processes*, volume 2. Florida: CRC Press, 2004.

J. A. DiMasi, R. W. Hansen, H. G. Grabowski, and L. Lasagna. Cost of innovation in the pharmaceutical industry. *Journal of Health Economics*, 10(2):107–142, 1991.

J. A. DiMasi, R. W. Hansen, H. G. Grabowski, and al. The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, 22(2):151–186, 2003.

C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

R. Ferland, A. Latour, and D. Oraichi. Integer-valued GARCH process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

K. Fokianos and R. Fried. Interventions in INGARCH processes. *Journal of Time Series Analysis*, 31(3):210–225, 2010.

K. Fokianos and R. Fried. Interventions in log-linear Poisson autoregression. *Statistical Modelling*, 12(4):299–322, 2012.

K. Fokianos, A. Rahbek, and D. Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

R. Fried, I. Agueusop, B. Bornkamp, K. Fokianos, J. Fruth, and K. Ickstadt. Retrospective Bayesian outlier detection in INGARCH series. *Statistics and Computing*, 23(6):1–10, 2013.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics*, 20(5-6):25–62, 1993.

A. Genkin, D. D. Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

L. E. Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the Lasso and sparse supervised learning problems. *Pacific Journal of Optimization*, 8(4):667–698, 2012.

D. Henze, I. Gashaw, D. Hornung, I. Agueusop, K. Machens, A. Schmitz, T. M. Zollner, and W. D. Döcke. Trefoil factor 3 a new player in clinical and experimental endometriosis? *Nature Reviews Drug Discovery (Manuscript submitted for publication)*, 2013.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

P. D. Hoff. *A first course in Bayesian statistical methods.* Philadelphia: Springer Science+Business Media LLC, 2009.

N. Japkowicz and al. Learning from imbalanced data sets: A comparison of various strategies. In *Proceedings of the AAAI Workshop on Learning from Imbalanced Data Sets*, 2000.

K. Jung, M. Grade, J. Gaedcke, P. Jo, L. Opitz, H. Becker, B. M. Ghadimi, and T. Beißbarth. A new sensitivity-preferred strategy to build prediction rules for therapy response of cancer patients using gene expression data. *Computer Methods and Programs in Biomedicine*, 100(2):132–139, 2010.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2): 137–163, 2001.

K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale L1-regularized logistic regression. *Journal of Machine Learning Research*, 8(7), 2007.

J. Kramer and K. A. Schulman. Transforming the economics of clinical trials. In *Proceedings of the Institute of Medicine*, pages 7–8, 2012.

J. M. Lachin. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2(2):93–113, 1981.

E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. New York: Springer-Verlag, 2nd edition, 1998.

C. X. Ling and V. S. Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, pages 231–5, 2008.

Z. Liu and M. Tan. ROC-based utility function maximization for feature selection and classification with applications to high-dimensional protease data. *Biometrics*, 64(4): 1155–1161, 2008.

Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S. J. Meltzer, and M. Tan. Sparse logistic regression with Lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 6(1):6, 2007.

K. E. May, S. A. Conduit-Hulbert, J. Villar, S. Kirtley, S. H. Kennedy, and C. M. Becker. Peripheral biomarkers of endometriosis: A systematic review. *Human Reproduction Update*, 16(6):651–674, 2010.

K. E. May, J. Villar, S. Kirtley, S. H. Kennedy, and C. M. Becker. Endometrial alterations in endometriosis: A systematic review of putative biomarkers. *Human Reproduction Update*, 17(5):637–653, 2011.

P. McCullagh and J. A. Nelder. *Generalized linear models*. London: Chapman & Hall/CRC, 1989.

L. Meier, S. Van De Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71, 2008.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 2004.

G. Mijoule, S. Savy, and N. Savy. Models for patients' recruitment in clinical trials and sensitivity analysis. *Statistics in Medicine*, 31(16):1655–1674, 2012.

M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

R. M. Parker, S. C. Ratzan, and N. Lurie. Health literacy: A policy challenge for advancing high-quality health care. *Health Affairs*, 22(4):147–153, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

A. E. Raftery and S. Lewis. How many iterations in the Gibbs sampler. *Bayesian Statistics*, 4(2):763–773, 1992.

V. Ravi. *Alabama: Constrained nonlinear optimization*, 2011. URL `http://CRAN.R-project.org/package=alabama`. R Package Version 2011.9-1.

C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. A comparative study of data sampling and cost sensitive learning. In *Proceedings of the Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference*, pages 46–52. IEEE, 2008.

V. S. Sheng and C. X. Ling. Thresholding for making classifiers cost-sensitive. In *Proceedings of the National Conference on Artificial Intelligence*, page 476. AAAI Press, 2006.

S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

N. Z. Shor, K. C. Kiwiel, and A. Ruszcaynski. *Minimization methods for non-differentiable functions*. New York: Springer, 1985.

R. Simon. Bayesian design and analysis of active control clinical trials. *Biometrics*, 55 (2):484–487, 1999.

J. P. Swann. Summary of NDA approvals & receipts, 1938 to the present. `http://fda.gov/AboutFDA/WhatWeDo/History/ProductRegulation/SummaryofNDAApprovalsReceipts1938tothepresent/default.htm`, 2013. [Online; accessed 02-September-2013].

G. Tang, Y. Kong, C. C. H. Chang, L. Kong, and J. P. Costantino. Prediction of accrual closure date in multi-center clinical trials with discrete-time Poisson process models. *Pharmaceutical statistics*, 11(5):351–356, 2012.

R. J. Temple. Enrichment designs: Efficiency in development of cancer treatments. *Journal of Clinical Oncology*, 23(22):4838–4839, 2005.

R. J. Temple and R. L. Becker. Guidance for industry enrichment strategies for clinical trials to support approval of human drugs and biological products. *U.S. Department of Health and Human Services Food and Drug Administration*, 2012.

P. F. Thall and J. K. Wathen. Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 43(5):859–866, 2007.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

A. Walzer, J. Hilpert, K. Machens, S. Lehr, E. M. Von Hübbenet, and H. Schmitz. Identification of possible biomarkers for monitoring and diagnosis of endometriosis. *Bayer Schering Pharma (Research Report), 2007*.

G. Wassmer. Enrichment Designs: Methoden und Anwendungen. *ADDPLAN, Institutskolloquium Wien*, 2013.

J. K. Wathen and P. F. Thall. Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine*, 27(27):5586–5604, 2008.

P. Whittle. *Optimization under constraints: Theory and applications of nonlinear programming*. New York: Wiley-Interscience, 1971.

M. H. Wright. Interior methods for constrained optimization. *Acta Numerica*, 1:341–407, 1992.

S. Yi. *Loss function for binary classification and class probability estimation*. PhD thesis, University of Pennsylvania, 2005.

A Poisson Processes

In deterministic processes the number of observations changes over time t according to a known function (e.g., $3t^2 - 1$) and not a probability distribution. This is rarely the case in real life, where almost all processes have a certain degree of randomness.

Definition A.1 *Stochastic process*
$\{N(t); t \in [0, \infty)\}$ *is a family of random variables describing an empirical process whose development is governed by probabilistic laws. The parameter* t *often describes the time and may be either discrete or continuous.* $N(t)$ *represents the cumulated observations from 0 to* t *and is a random variable, which can be real valued or complex* ($t \in \mathcal{C}$) *and may take the form of a vector.*

This thesis focuses on discrete processes such as the process of patient recruitment in clinical trials which is observed at discrete times. Discrete process have observations in the set of non-negative integers. It assumed that the observation time space is divided in equispaced intervals of time (day, week, month,...). The observations at the t^{th} time unit are given by $N_t = N(t) - N(t-1)$. The main task in dealing with stochastic processes is to use available data points $N_1, N_2, \cdots, N_\tau$ to forecast the future observations of the process $(N_{\tau+1}, N_{\tau+2}, \cdots)$. Many classes of stochastic processes have been developed and describe well a wide range of real life processes. Empirical processes with a given stability called stationarity usually require less parameters by the modeling than non-stationary processes whose expectation and variability change over time.

Definition A.2 *Stationary processes*
A stochastic process is stationary if it varies around a fixed mean. It is called strictly stationary if the joint distribution of $\tau = 1, 2, 3, \cdots$ *successive observations does not depend on the time interval of selection. For example, for* $k = 1, 2, \cdots$ *the joint distribution of* $N_1, N_2, \cdots, N_\tau$ *is equal to that of* $N_{1+k}, N_{2+k}, \cdots, N_{\tau+k}$ *(Box et al. [2013]).*

If the probability distribution of N_t is the same for all t, the samples $N_1, N_2, \cdots, N_\tau$ can be used to estimate the mean and variance of the process
$\bar{\mu}_N = \frac{1}{\tau}\sum_{t=1}^{\tau} N_t$ and $\hat{\sigma}_N^2 = \frac{1}{\tau}\sum_{t=1}^{\tau}(N_t - \bar{N})^2$. The covariance of N_t and N_{t+k} defined as
$\Gamma_k = cov(N_t, N_{t-k}) = E[(N_t - \mu)(N_{t+k} - \mu)]$

is called the *autocovariance* at lag k and then the *autocorrelation* at lag k is given by:

$$\begin{aligned} corr(N_t, N_{t+k}) &= \frac{E[(N_t - \mu_N)(N_{t+k} - \mu_N)]}{\sqrt{E[(N_t - \mu_N)^2]E[(N_{t+k} - \mu_N)^2]}} \\ &= \frac{E[(N_t - \mu_N)(N_{t+k} - \mu_N)]}{\sigma_N^2}. \end{aligned}$$

The representation of the autocorrelation as a function of lag is called the autocorrelation function. For any stationary process, the autocovariance $cov(N_i, N_j)_{i,j\in\{1,2,\cdots\tau\}}$ and autocorrelation matrix $corr(N_i, N_j)_{i,j\in\{1,2,\cdots\tau\}}$ of τ successive observations are positive-definite (Box et al. [2013]).

A.1 Definitions

Let $\{N(t); t \in [0, \infty)\}$ be a stochastic process so that the observations in two different time intervals are independent and the probability of observing more than one event in a very small time interval tends to zero: $P[N(t+h) - N(t) \geq 2] = P[N(h) \geq 2] = o(h)$ for $h \to 0$, where $o(h)$ denotes a function of h that tends to zero faster than h itself. Such a process is called a Poisson process. Poisson processes have been used in modeling count data which take non-negative integer values (For example, the birth process or the patient recruitment process in a clinical trials). It is assumed that $N(0) = 0$ and $N(t)$ is nondecreasing in t, where $N(t)$ denotes the total number of events observed from 0 to t (King and Zeng [2001]).

Definition A.3 *Homogeneous Poisson process*
A Poisson process is called homogeneous if there is a constant $\lambda > 0$ called intensity, so that the probability to observe k events in a time interval $(0, t)$ is given by:

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \qquad k = 0, 1, \cdots .$$

In homogeneous Poisson processes the mean and the variance are equal: $E(N(t)) = Var(N(t)) = \lambda t$.

The intensity λ may be a continuous function of the observation time. In this case, the mean between two positive numbers a and b is the integral of the intensity in that interval $\lambda(t) = \int_a^b \lambda(\tau)$, and the probability density $P(N(t) = k)$ is given by:

$$P(N(t) = k) = \frac{\exp\left\{-\int_0^t \lambda(\tau)d\tau\right\}\left[\int_0^t \lambda(\tau)d\tau\right]^k}{k!}.$$

$E[N(t)] = \int_0^t \lambda(\tau)d\tau$ and the variance is defined by:

$$\begin{aligned} Var[N(t)] &= \sum_{k=0}^{\infty} k^2 \frac{\exp\left\{-\int_0^t \lambda(\tau)d\tau\right\}\left[\int_0^t \lambda(\tau)d\tau\right]^k}{k!} \\ &= E[N(t)]\int_0^t \lambda(\tau)d\tau + \int_0^t \lambda(\tau)d\tau. \end{aligned}$$

Remark A.4 *The generalization of λ to $\lambda(t)$ does not affect the form of the distribution, but the mean and variance are now different ($E[N(t)] \neq Var[N(t)]$).*

The following property of Poisson processes is important when a unique process is considered that consists of the sum of a given number of parallel Poisson processes.

Proposition A.5 *Sum of independent Poisson processes*
Let $\{N_1(t); t \in [0,\infty)\}$ and $\{N_2(t); t \in [0,\infty)\}$ be two independent Poisson processes with expectations $\lambda_1(t)$ and $\lambda_2(t)$, respectively. The process $\{N_1(t) + N_2(t); t \in [0,\infty)\}$ is also a Poisson process with intensity $\lambda_1(t) + \lambda_2(t)$ (King and Zeng [2001]).

This property states that a multi-center recruitment process can be investigated under weak conditions as a unique Poisson process. This is particulary relevant for a large number of centers, where some centers are likely to recruit zero patients (Anisimov and Fedorov [2007],Mijoule et al. [2012]). It can be extended to any finite sum using induction. The resulting process for an infinite sum of Poisson processes is given by the countable additive theorem.

Theorem A.6 *Countable additive theorem*
Let N_t, $t = 0, 1, 2, \cdots$ be independent random variables, and assume N_t is Poisson distributed with density $p(\lambda_t)$ for each t.
If $\Lambda = \sum_{t=1}^{\infty} \lambda_t$ converges, then $N = \sum_{t=1}^{\infty} N_t$ converges with probability 1, and N is Poisson distributed with intensity Λ and density $p(\Lambda)$. If, on the other hand, Λ diverges, then N diverges with probability 1.

Proof A.7 *See King and Zeng [2001].*

A.1.1 Some Connections to other Distributions

Let T_i be the waiting time between the $(i-1)^{st}$ and the i^{th} events of a count process and $T(n)$ be the waiting time until observing the n^{th} event. That means $T(n) = \sum_{i=1}^{n} T_i$. If the event times T_i, $i = 1, 2, \cdots$ are independent exponentially distributed with parameter

$\lambda > 0$ constant, then $\{N(t); t \in [0, \infty)\}$ is a homogenous Poisson process with parameter λ. Conversely, if $\{N(t); t \in [0, \infty)\}$ is a homogenous Poisson process with intensity λ, the waiting time between events are independent exponentially distributed with parameter λ. Note that the probability density of an exponentially distributed random variable T is defined as

$$P(T = t) = \lambda e^{-\lambda t},$$

for $t > 0$ and $\lambda > 0$.
The waiting time until n events occur $T(n)$ is a sum of independent exponentially distributed variables, which corresponds to the Erlang distribution. The Erlang distribution is a Gamma distribution where the shape parameter is a non-negative integer: $T(n) \sim Erlang(n, \lambda)$. Homogenous Poisson processes are relative simple to manage, but may lead to poor predictions when they are used for the modeling of an over-dispersed process. Over-dispersion occurs when the sample variance is greater than the theoretical variance.

Definition A.8 *Poisson process with Gamma distributed intensity*
The Poisson process with Gamma distributed intensity has been investigated by Anisimov and Fedorov [2007] for modeling of the patient recruitment process in clinical trials. $\{N(t); t \in [0, \infty)\}$ *is a Poisson process with parameter* λ *that varies randomly according to a Gamma distribution. The probability density* $f(\lambda)$ *of a Gamma distributed variable* $\lambda \sim Gamma(\alpha, \beta)$ *is given by:*

$$f(\lambda) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} & \text{if} \quad \lambda > 0, \alpha > 0, \beta > 0 \\ 0 & \text{if} \quad \lambda \leq 0, \end{cases}$$

where $\Gamma(\alpha)$ *denotes the Gamma function defined as*
$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ *(*$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$*).*

Remark A.9 *Let* $\{N(t); t \in [0, \infty)\}$ *be a Poisson process with Gamma distributed intensity* λ*. The conditional distribution of* $N(t)$ *given a value of* λ *is a Poisson distribution with density defined as*

$$\begin{aligned} P(N(t) = k|\lambda) &= \frac{e^{-\lambda t}(\lambda t)^k}{k!}, k = 1, 2, \cdots \\ &= \binom{k + \alpha - 1}{k} \left(\frac{t}{\beta + t}\right)^k \left(\frac{\beta}{\beta + t}\right)^\alpha. \end{aligned}$$

$N(t)$ is Negative Binomial distributed with parameters

$$r = \alpha, \qquad p = \frac{\beta}{\beta + t}$$

The mean and the variance are given by

$$E[N(t)] = \alpha \frac{t}{\beta},$$

and

$$Var(N(t)) = \alpha \frac{t}{\beta}\left(1 + \frac{t}{\beta}\right).$$

The variance is greater than the mean, since

$$\begin{aligned} Var(N(t)) &= \alpha \frac{t}{\beta} + \alpha \frac{t^2}{\beta^2} \\ &= E[N(t)] + \alpha \frac{t^2}{\beta^2} \geq E[N(t)]. \end{aligned}$$

The Negative Binomial distribution is over-dispersed compared to the Poisson distribution. It provides an additional parameter that can be useful in modeling the variance.

A.1.2 Poisson Autoregressive Process

Poisson autoregressive processes are Poisson processes with intensity defined as a function its past observations and the past values of the intensity. Such processes have been investigated by Ferland et al. [2006] and Fokianos et al. [2009] and have the advantage of taking into account the past information about process progression. The mean and variance not only depend on the randomness, but there is also a feedback mechanism.

Definition A.10 *Poisson linear autoregressive models*
Let $\mathcal{F}_{t-1}$ be a σ-field generated from $\{\lambda_0, N_0, N_1, \cdots, N_{t-1}\}$. A Poisson linear autoregressive model of order p, q, where $p = 1, 2 \cdots$ and $q = 1, 2, \cdots$ is defined as:

$$\begin{aligned} N_t | \mathcal{F}_{t-1} &\sim Pois(\lambda_t) \\ \lambda_t &= \beta_0 + \sum_{i=1}^{p} \beta_i B^i N_t + \sum_{j=1}^{q} \alpha_j B^j \lambda_t, \end{aligned} \tag{31}$$

where B denotes the backward shift operator ($B^1 N_t = N_{t-1}$) and λ_0, N_0 are fixed and positive constants (Ferland et al. [2006]). The parameters of the models are assumed to be positive, i.e. $\beta_0, \beta_1 \cdots \beta_p, \alpha_1, \cdots, \alpha_q \geq 0$. A necessary condition that the model parameters must satisfy for stationary has been derived by Ferland et al. [2006]:

$$\sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1 \tag{32}$$

Stationarity means the expectation does not change over time.

Remark A.11

This model is a discrete version of GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models investigated by Bollerslev [1986] and has been called INGARCH for integer-GARCH.

The sparsest INGARCH process corresponds to $p = q = 1$ and is defined by

$$\begin{aligned} N_t|\mathcal{F}_{t-1} &\sim Pois(\lambda_t) \\ \lambda_t &= \beta_0 + \beta_1 N_{t-1} + \alpha_1 \lambda_{t-1}. \end{aligned} \tag{33}$$

The mean of a stationary process is obtained by solving the equation $\lambda = \beta_0+\beta_1\lambda+\alpha_1\lambda \Rightarrow \lambda = \beta_0/(1-\beta_1-\alpha_1)$.

The auto-covariance to lag k and the variance of this process have been derived in Ferland et al. [2006]:

$$Cov(N_t, N_{t+k}) = \begin{cases} \frac{[1-(\beta_1+\alpha_1)^2+\beta_1^2]\lambda}{1-(\beta_1+\alpha_1)^2} & \text{if} \quad k = 0 \\ \frac{\beta_1[1-\alpha_1(\beta_1+\alpha_1)](\beta_1+\alpha_1)^{k-1}\lambda}{1-(\beta_1+\alpha_1)^2} & \text{if} \quad k \geq 1. \end{cases}$$

The variance is given by

$$\sigma^2 = \lambda\left(1 + \frac{\beta_1^2}{1-(\beta_1+\alpha_1)^2}\right).$$

Similar to the Poisson process with Gamma distributed intensity, the variance of the INGARCH process is greater than its mean. The general form of Poisson autoregressive models is obtained by $\lambda_t = f(\lambda_0, \lambda_1, \cdots, \lambda_{t-1}) + g(N_0, N_1, \cdots, N_{t-1})$, where f and g are given functions with values in $\mathbb{R}^+$ (Fokianos et al. [2009]).

A.2 Other Definitions

Let us consider a population of n unselected patients. Let us assume that n^+ of them are marker-positive and $n^- = n - n^+$ are marker-negative. Consider a marker-test that classifies $TP \leq n^+$ of the marker-positive patients as positive and TN of the marker-negative patients as negative.

The **prevalence** of marker positive patients in the unselected patient population is estimated by $\frac{N^+}{N} = prevalence$.

Accuracy: Represents the proportion of the overall true classified samples ($\frac{TP+TN}{n}$). Here, the best classifier leads to maximal accuracy. The accuracy depends on the class frequencies and its interpretation may be misleading if the classes are strongly imbalanced. For example, if $n^+ = 95$ and $n^- = 5$, a trivial classifier that assigns all patients

to the diseased class will have 95% accuracy.

Sensitivity: Is defined as $\frac{TP}{n^+}$ and denotes the proportion of diseased patients correctly classified as diseased.

Specificity: Is given by $\frac{TN}{n^-}$ and represents the proportion of healthy patients correctly classified as healthy.

The **positive predictive value (PPV)** is defined as the proportion of positive samples in the subpopulation of **test positive** samples. The PPV depends on the sensitivity, the specificity and the prevalence

$$PPV = \frac{Sens\ Prevalence}{Sens\ Prevalence + (1 - Spec)(1 - Prevalence)}.$$

The **negative predictive value (NPV)** is defined as the proportion of negative samples in the subpopulation of **test negative** samples. The NPV depends on the sensitivity, the specificity and the prevalence

$$NPV = \frac{Spec\ (1 - Prevalence)}{Spec\ (1 - Prevalence) + (1 - Sens)\ Prevalence}.$$

A.3 R-Code to the constrained Optimization of the Likelihood

```
require(alabama)
#*************************************************************************
# Generalized logistic function
#*************************************************************************
glog <- function(x){
  return(1/(1+exp(-10*x)))
}
#*************************************************************************
# Data standardazation
#*************************************************************************
standardata <- function(dat){
  .standard <- function(column){
    column <- (column-mean(column))/sd(column)
  }
  daten <- dat
  newdata<- cbind(apply(daten[,-ncol(daten)],2,.standard),
  daten[,ncol(daten)])
  return(newdata)
}
#*************************************************************************
# T Test screening for verification of the strong rules
#*************************************************************************
ttest <- function(dat){
  te <- numeric()
  for(i in 1:(ncol(dat)-1)){
    te[i] <- t.test(dat[,i][dat[,ncol(dat)]==1],
    dat[,i][dat[,ncol(dat)]==0])$p.value
  }
  orders <- order(te)
  signifdat <- cbind(dat[,which(te < 0.05)],dat[,ncol(dat)])
  return(signifdat)
}
```

```
#*************************************************************************
# Max lambda  for which all predictors are setting to zero
#*************************************************************************
maxlambda <- function(cutof,minsen=0.9,data,eta=10,bpar=0.01){
  nc <- ncol(data); nr <- nrow(data); dat <- data
  y <- dat[,nc]; pbar <- rep(mean(y),nr)
  lambdas <- numeric()
  matdata <- as.matrix(dat[,-nc])
  datone <- dat[dat[,ncol(dat)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(datcond)
  pbarsens <- pbar[1:nrow(datone)]
  a <- glog(eta*(pbarsens-cutof))
  inter <- t(matdata)%*%(y-pbar)
  grads <- eta*bpar*t(datenf)%*%(pbarsens*(1-pbarsens)*
  (1-a)*a)/((mean(a)-minsen)*sum(y))
  return(max(drop(abs(-inter-grads))))
}
#*************************************************************************
# Strong rules (See Tibshirani 2001)
#*************************************************************************
strulesign <- function(lambda, cutof,minsen,data,eta, bpar=0.01){
  nc <- ncol(data); nr <- nrow(data); dat <- data
  y <- dat[,nc]; pbar <- rep(mean(y),nr); lambdas <- numeric()
  matdata <- as.matrix(dat[,-nc])
  datone <- dat[dat[,ncol(dat)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(datcond)
  pbarsens <- rep(mean(y),nrow(datone))
  a <- glog(eta*(pbarsens-cutof))
  inter <- t(matdata)%*%(y-pbar)
  grads <- eta*bpar*t(datenf)%*%(pbarsens*(1-pbarsens)*
           (1-a)*a)/((mean(a)-minsen)*sum(y))
  lambdas <- drop(abs(-inter-grads))
  signdata <- dat[,which(lambdas >= 2*lambda - max(lambdas))]
```

```
  datasign <- cbind(signdata,study_group=dat[,nc])
  return(datasign)
}
#*********************************************************************
# This function compute the crossvalidtion of the likelihood function
# optimization subject to the constraint that the sensitivity is larger
# than theta. ********************************************************

cv_lik<-function(data, barrier=0.001, theta=0.9, sigmoid=10, cutoff=0.5,
                 fold=10, it=10, np=10, standard=TRUE, method="auglag"){

  if(standard==TRUE){
    data <- standardata(data)
  } else{
    data <- data
  }
  sdata <- data

  set.seed(6)
  n <- trunc(nrow(data)/fold); index <- sample(nrow(data))

  t.labels <- as.numeric(data[,ncol(data)][index])

  prediction <- numeric(); parameters <- numeric()
  # CV
  for(k in 1:fold){
    train <- data[-index[(n*(k-1)+1):(k*n)],]
    if (ncol(train) < 3){
      train <- train
      test <- sdata[index[(n*(k-1)+1):(k*n)],]
      colnames(test)[ncol(test)] <- "study_group"
      test <- test[,colnames(train)]
      test <- as.matrix(cbind(rep(1,n),test[,-ncol(test)]))
    }else{
    maxl <- maxlambda(cutof=cutoff,minsen=theta,data=train,eta=sigmoid,
```

```
  bpar=barrier)# maxlambda
  lamb <- seq(from = maxl, to = maxl/2, length.out=50)
  pred <- 1; j <- 1
  while(pred < np){
    lambda <- lamb[j]
    tr <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta,
          data=train, eta=sigmoid, bpar=barrier)
    pred <- ncol(tr)
    j <- j + 1
  }
  lambda <- lamb[j-1]
  train <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta,
           data=train, eta=sigmoid, bpar=barrier)
  test <- sdata[index[(n*(k-1)+1):(k*n)],]
  colnames(test)[ncol(test)] <- "study_group"
  test <- test[,colnames(train)]
  test <- as.matrix(cbind(rep(1,n),test[,-ncol(test)]))
  }

if(ncol(train)==1){
    probapred <- rep(1,nrow(test))
}else{

star <- c(1,rep(0.001,(ncol(train)-1)))
# likelihood function
loglik <- function(para){
  lambda <- lambda
  daten <- train
  nr <- nrow(daten)
  nc <- ncol(daten)
  datenf <- as.matrix(cbind(rep(1,nr),daten[,-nc]))
  y <- daten[,nc]
  x <- datenf%*%para
  lih <- t(x)%*%(y-1) - sum(log(1+exp(-x)))
return(-lih + lambda*sum((para[-1]^2 + 0.001)^0.5))
```

```
}
# gradient of the likelihood function
gradien <- function(para){
  lambda <- lambda
  daten <- train
  nr <- nrow(daten)
  nc <- ncol(daten)
  datenf <- as.matrix(cbind(rep(1,nr),daten[,-nc]))
  y <- daten[,nc]
  p <- 1/(1 + exp(-datenf%*%para))
  return(as.numeric(-t(datenf)%*%(y-p) + lambda*
        c(0, para[-1]/sqrt(para^2[-1] + 0.0001))))
}
# Constraint on the sensitivity
constr <- function(para){
  daten <- train; sigmoid <-  sigmoid; cutoff <- cutoff
  lambda1 <- lambda; minsens <- theta; h <- rep(NA, 1)
  datone <- daten[daten[,ncol(daten)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(cbind(rep(1,nrow(datone)),datcond))
  predictor <- drop(datenf%*%para)
  sens <- sum(1/1+exp(-sigmoid*(predictor-log(cutoff/(1-cutoff)))))
  sens <- sens/nrow(datone)
  h[1] <- sens - minsens
  return(h)
}
# Jacobien matrix of the constraint
jacob <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
  lambda <- lambda
  datone <- daten[daten[,ncol(daten)]==1,]
  ncl <- ncol(datone)
  datcond <- datone[,-ncl]
```

```
  datenf <- as.matrix(cbind(rep(1,nrow(datcond)),datcond))
  jac <- matrix(nrow=1,ncol=ncol(datenf))
  x <- drop(datenf%*%para)
  ge <-  1/(1 + exp(-sigmoid*(x - cutoff)))
  jac[1,] <- sigmoid*(1/nrow(datcond))*t(datenf)%*%(ge*(1-ge))
  return(jac)
}
inn <- auglag(par=star, fn=loglik, gr = gradien,hin = constr,
       hin.jac = jacob)
#inn <- constrOptim.nl(par=star, fn=loglik, gr = gradien,
        #hin = constr, hin.jac = jacob)
probapred <- 1/(1+exp(-test%*%inn$par))
}
prediction <- c(prediction,as.numeric(probapred))
}
if(n*fold==nrow(data)){
  sensitivity <- sum(ifelse(prediction >= cutoff &
                 t.labels == 1,1,0))/sum(t.labels)
  specificity <- sum(ifelse(prediction < cutoff &
                 t.labels == 0,1,0))/sum((1-t.labels))
} else{
  train <- data[index[1:(n*fold)],]
  if (ncol(train) < 3){
    train <- train
    test <- sdata[index[(n*(k-1)+1):(k*n)],]
    colnames(test)[ncol(test)] <- "study_group"
    test <- test[,colnames(train)]
    test <- as.matrix(cbind(rep(1,n),test[,-ncol(test)]))
  }else{
  maxl <- maxlambda(cutof=cutoff,minsen=theta,data=train,
          eta=sigmoid,bpar=barrier)
  lamb <- seq(from = maxl, to = maxl/2, length.out=50)
  pred <- 1; j <- 1
while(pred < np){
  lambda <- lamb[j]
```

```
  tr <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta, data=train,
                   eta=sigmoid, bpar=barrier)
  pred <- ncol(tr)
  j <- j + 1
}
lambda <- lamb[j-1]
train <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta, data=train,
         eta=sigmoid, bpar=barrier)
test <- sdata[index[(n*fold+1):nrow(data)],]
colnames(test)[ncol(test)] <- "study_group"
test <- test[,colnames(train)]
test <- as.matrix(cbind(rep(1,nrow(test)),test[,-ncol(test)]))
}
if(ncol(train)==1){
probapred <- rep(1,nrow(test))
}else{
star <- c(0,rep(0.001,(ncol(train)-1)))

loglik <- function(para){
  lambda <- lambda; daten <- train
  nr <- nrow(daten); nc <- ncol(daten)
  datenf <- as.matrix(cbind(rep(1,nr),daten[,-nc]))
  y <- daten[,nc]; x <- datenf%*%para
  lih <- t(x)%*%(y-1) - sum(log(1+exp(-x)))
return(-lih + lambda*sum((para[-1]^2 + 0.001)^0.5))
}

gradien <- function(para){
  lambda <- lambda; daten <- train
  nr <- nrow(daten); nc <- ncol(daten)
  datenf <- as.matrix(cbind(rep(1,nr),daten[,-nc]))
  y <- daten[,nc]
  p <- 1/(1 + exp(-datenf%*%para))
  return(as.numeric(-t(datenf)%*%(y-p) +
        lambda*c(0, para[-1]/sqrt(para^2[-1] + 0.0001))))
```

```
}

constr <- function(para){
  daten <- train; sigmoid <-  sigmoid
  cutoff <- cutoff; lambda1 <- lambda
  minsens <- theta; h <- rep(NA, 1)
  datone <- daten[daten[,ncol(daten)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(cbind(rep(1,nrow(datone)),datcond))
  predictor <- drop(datenf%*%para)
  sens <- sum(1/1+exp(-sigmoid*(predictor-log(cutoff/(1-cutoff)))))
  sens <- sens/nrow(datone)
  h[1] <- sens - minsens
  return(h)
}

jacob <- function(para){
  daten <- train; sigmoid <-  sigmoid
  cutoff <- cutoff; lambda <- lambda
  datone <- daten[daten[,ncol(daten)]==1,]
  ncl <- ncol(datone); datcond <- datone[,-ncl]
  datenf <- as.matrix(cbind(rep(1,nrow(datcond)),datcond))
  jac <- matrix(nrow=1,ncol=ncol(datenf))
  x <- drop(datenf%*%para)
  ge <-  1/(1 + exp(-sigmoid*(x - cutoff)))
  jac[1,] <- sigmoid*(1/nrow(datcond))*t(datenf)%*%(ge*(1-ge))
  return(jac)
}

inn <- auglag(par=star, fn=loglik, gr = gradien, hin = constr,
       hin.jac = jacob)
#inn <- constrOptim.nl(par=star, fn=loglik, gr = gradien, hin = constr,
        #hin.jac = jacob)
probapred <- 1/(1+exp(-test%*%inn$par))
```

```
  }
  prediction <- c(prediction,as.numeric(probapred))
  sensitivity <- sum(ifelse(prediction >=cutoff &
                 t.labels == 1,1,0))/sum(t.labels)
  specificity <- sum(ifelse(prediction < cutoff &
                 t.labels == 0,1,0))/sum((1-t.labels))
  }
  allpred <- list(sensitivity=sensitivity,specificity=specificity,
             prediction=prediction,t.labels=t.labels,parameters=parameters)
  return(allpred)
}

#result1 <- cv_lik(data=daten,theta=0.9,sigmoid = 15,
#cutoff=0.62,standard=TRUE)
# Find optimal cut-off and evt. other parameters such as sigmoid parameter

optimr <- function(data = daten, theta = 0.9, sig = 10, cutf,
                   npreds = c(10,20,30)){
    res <- numeric()
    for(i in 1:length(sig))
    {
    a <- sig[i]
    for(j in 1:length(npreds))
    {
      b <- npreds[j]
      for(k in 1:length(theta))
      {
        d <- theta[k]
        for(l in 1:length(cutf))
        {
          cu <- cutf[l]
          result1 <- cv_lik(data = daten,theta = d, sigmoid = a,
          cutoff = cu, standard = TRUE, np = b)
          if(result1$sensitivity > 0.89 & result1$specificity > 0.1)
          {
```

```
          vec <- c(result1$sensitivity, result1$specificity, a, b, d,cu)
          res <- rbind(res, vec)
          true_labels <- result1$t.labels
          predicted_prob <- result1$prediction
          }
          else
          {
          vec <- c(1, 0, a, b, d,cu)
            res <- rbind(res, vec)
            predicted_prob <- rep(1,nrow(data))
            true_labels <- daten[,ncol(data)]
          }
        }
      }
    }
  }
  best1 <- res[which(res[,2] == max(res[,2]))[1],]
  return(list(SensSpec=best1,true_labels=true_labels,
         predicted_prob=predicted_prob))
  }

npreds <- c(15,20,24)
cutf <- seq(0.4,0.7,by=0.02)
#*************************************************************
# Load the full data set and save it in daten as data frame.
#*************************************************************
bestr <- optimr(data = daten, theta = 0.9, sig = 10,
          cutf=cutf, npreds = npreds)
```

A.4 R-Code to constrained Optimization of the Specificity

```
#*****************************************************************
# This programs computes the specificity optimization subject to the
# constraint that the sensitivity is larger than theta = 90% *******
# The main package is alabama ************************************
```

```
require(alabama)
# generalized logistic function ***********************************
glog <- function(x){
  return(1/(1+exp(-x)))
}
# Function for data standardization, similar to scale **************
standardata <- function(dat){
  .standard <- function(column){
    column <- (column-mean(column))/sd(column)
  }
  daten <- dat
  newdata<- cbind(apply(daten[,-ncol(daten)],2,.standard),
                  daten[,ncol(daten)])
  colnames(newdata)[ncol(newdata)] <- "study_group"
  return(newdata)
}
# This function selects only significant variables (t-test level of 0.05)
ttest <- function(dat){
  te <- numeric()
  for(i in 1:(ncol(dat)-1)){
    te[i] <- t.test(dat[,i][dat[,ncol(dat)]==1],dat[,i]
                    [dat[,ncol(dat)]==0])$p.value
  }
  orders <- order(te)
  signifdat <- cbind(dat[,which(te<0.05)],dat[,ncol(dat)])
  return(signifdat)
}

# This function computes the minimal value of lambda from
# which all predictor are set to ****
# zero in lasso type problem***************************************

maxlambda <- function(cutof,minsen=0.9,data,eta=10,bpar=0.01){
  #data <- as.matrix(cbind(rep(1,nrow(data)),data))
  dat <- data
```

```
  y <- dat[,ncol(data)]
  pbar <- rep(mean(y),nrow(data))
  lambdas <- numeric()
  matdata <- as.matrix(dat[,-ncol(data)])
  datone <- dat[dat[,ncol(dat)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(datcond)
  pbarsens <- pbar[1:nrow(datone)]

  datzero <- data[data[,ncol(data)]==0,]
  datcondzero <- datzero[,-ncol(datzero)]
  datenfzero <- as.matrix(datcondzero)
  pbarspe <- pbar[1:nrow(datzero)]
  a <- glog(eta*(pbarsens-cutof))
  #predictorzero <- drop(datenfzero%*%para)
  p <- pbarspe
  g <- 1/(1 + exp(eta*(p - cutof)))
  grad0 <- (1/nrow(datzero))*t(datenfzero)%*%(eta*p*(1-p)*g*(1-g))
  gradbar <- - eta*bpar*t(datenf)%*%(pbarsens*(1-pbarsens)*(1-a)*a)/
               ((mean(a)-minsen)*sum(y))
  #grad1 <- - eta*t(datenf)%*%(pbarsens*(1-pbarsens)*(1-a)*a)/sum(y)
  return(max(abs(grad0 + gradbar)))
}
# ************************************************************
# This function selects relevant predictors using the strong rule *
strulesign <- function(lambda, cutof,minsen,data,eta, bpar=0.01){
  #data <-standardata(data)
  #data <- as.matrix(cbind(rep(1,nrow(data)),data))
  nc <- ncol(data); nr <- nrow(data)
  dat <- data
  y <- dat[,ncol(data)]
  pbar <- rep(mean(y),nr)
  lambdas <- numeric()
  matdata <- as.matrix(data[,-ncol(data)])
  datone <- dat[dat[,ncol(dat)]==1,]
```

```
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(datcond)
  pbarsens <- pbar[1:nrow(datone)]

  datzero <- data[data[,ncol(data)]==0,]
  datcondzero <- datzero[,-ncol(datzero)]
  datenfzero <- as.matrix(datcondzero)
  pbarspe <- pbar[1:nrow(datzero)]
  a <- glog(eta*(pbarsens-cutof))
  #predictorzero <- drop(datenfzero%*%para)
  p <- pbarspe
  g <- 1/(1 + exp(eta*(p - cutof)))
  grad0 <- (1/nrow(datzero))*t(datenfzero)%*%(eta*p*(1-p)*g*(1-g))
  gradbar <- - eta*bpar*t(datenf)%*%(pbarsens*(1-pbarsens)*(1-a)*a)/
              ((mean(a)-minsen)*sum(y))
 # grad1 <- - eta*t(datenf)%*%(pbarsens*(1-pbarsens)*(1-a)*a)/sum(y)
  lambdas <- abs(grad0 + gradbar)
  signdata <- dat[,which(lambdas >= 2*lambda - max(lambdas))]
  datasign <- cbind(signdata,study_group=dat[,ncol(data)])
  return(datasign)
}
# This function provides a feasible start value, when constrOptim.nl
  is used for the optimization
start <- function(data,sigmoid=10,cutoff=0.5,lambda=10,theta=0.9,
         barrier=0.001){
  data <- cbind(rep(1,nrow(data)), data)
  para <- runif((ncol(data)-1),min=-lambda/ncol(data),
          max=lambda/ncol(data))
  datone <- data[data[,ncol(data)]==1,]
  #datone <- as.matrix(cbind(rep(1,nrow(datone)),datone[,-ncol(datone)]))
  datone <- as.matrix(datone[,-ncol(datone)])
  x <- drop(datone%*%para)
  dif <- x - log(cutoff/(1-cutoff))
  se <- sum(1/(1 + exp(-sigmoid*dif)))/nrow(datone)
  c1 <- se - theta
```

```
  n <- 1
  #c2 <- lambda - sum((para^2 + 0.0001)^0.5)
  while(c1 < 0 & n < 200){
    para <- runif((ncol(data)-1),min=-lambda/ncol(data),
            max=lambda/ncol(data))
    x <- drop(datone%*%para)
    dif <- 10*(x - log(cutoff/(1-cutoff+0.0001)))
    se <- sum(1/(1 + exp(-sigmoid*dif)))/nrow(datone)
    c1 <- se - theta
    n <- n + 1
    #c2 <- lambda - sum((para^2 + 0.0001)^0.5)
  }
  return(para)
}

# Constrained optimization ************************************
cv_sp <- function(data,barrier=0.001,theta=0.9,sigmoid=10,cutoff=0.5,
          fold=10,it=10, np=15,standard=TRUE,method="auglag"){

  if(standard==TRUE){
    data <- standardata(data)
  } else{
    data <- data
  }
  sdata <- data
  set.seed(3)
  n <- trunc(nrow(data)/fold);index <- sample(nrow(data))
  t.labels <- as.numeric(data[,ncol(data)][index])
  prediction <- numeric(); parameters <- numeric()

  for(k in 1:fold){
    train <- data[-index[(n*(k-1)+1):(k*n)],]
    if (ncol(train) < 3){
      train <- train
      test <- sdata[index[(n*(k-1)+1):(k*n)],]
```

```
    colnames(test)[ncol(test)] <- "study_group"
    test <- test[,colnames(train)]
    test <- as.matrix(cbind(rep(1,n),test[,-ncol(test)]))
  }else{
  maxl <- maxlambda(cutof=cutoff,minsen=theta,data=train,eta=sigmoid,
          bpar=barrier)
  lamb <- seq(from = maxl, to = maxl/2, length.out=50)
  pred <- 1; j <- 1
  while(pred < np){
    lambda <- lamb[j]
    tr <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta,
    data=train, eta=sigmoid, bpar=barrier)
    pred <- ncol(tr)
    j <- j + 1
  }
  lambda <- lamb[j-1]
  train <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta,
  data=train, eta=sigmoid, bpar=barrier)
  test <- sdata[index[(n*(k-1)+1):(k*n)],]
  colnames(test)[ncol(test)] <- "study_group"
  test <- test[,colnames(train)]
  test <- as.matrix(cbind(rep(1,n),test[,-ncol(test)]))
  #test <- as.matrix(test[,-ncol(test)])
  }
if(ncol(train)==1){
    probapred <- rep(1,nrow(test))
}else{

  #star <- start(data=train,sigmoid=sigmoid,cutoff=cutoff,
  #lambda=lambda,theta=theta,barrier=barrier)
  star <- c(0,rep(0.001,(ncol(train)-1)))
fn <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
```

```
  lambda <- lambda
  datzero <- daten[daten[,ncol(daten)]==0,]
  datcondzero <- datzero[,-ncol(datzero)]
  datenfzero <- as.matrix(cbind(rep(1,nrow(datzero)),datcondzero))
  #datenfzero <- as.matrix(datcondzero)
  predictorzero <- drop(datenfzero%*%para)
  probazero <- 1/(1 + exp(-predictorzero))
  senszero <- sum(1/(1 + exp(sigmoid*(probazero - cutoff))))
  senszero <- senszero/nrow(datzero)
  return(-senszero)
}

gradien <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
  lambda1 <- lambda
  datzero <- daten[daten[,ncol(daten)]==0,]
  datcondzero <- datzero[,-ncol(datzero)]
  datenfzero <- as.matrix(cbind(rep(1,nrow(datzero)),datcondzero))
  #datenfzero <- as.matrix(datcondzero)
  predictorzero <- drop(datenfzero%*%para)
  p <- 1/(1 + exp(-predictorzero))
  g <- 1/(1 + exp(sigmoid*(p - cutoff)))
  grad<- (1/nrow(datzero))*t(datenfzero)%*%(sigmoid*p*(1-p)*g*(1-g))
  return(as.numeric(t(grad)))
}

constr <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
  lambda1 <- lambda
  minsens <- theta
  h <- rep(NA, 2)
```

```
  datone <- daten[daten[,ncol(daten)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(cbind(rep(1,nrow(datone)),datcond))
  #datenf <- as.matrix(datcond)
  predictor <- drop(datenf%*%para)
  proba <- 1/(1 + exp(-predictor))
  sens <- sum(1/(1 + exp(-sigmoid*(proba - cutoff))))
  sens <- sens/nrow(datone)
  h[1] <- sens - minsens
  h[2] <- lambda - sum((para^2[-1] + 0.0001)^0.5)
  return(h)
}
jacob <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
  lambda <- lambda
  datone <- daten[daten[,ncol(daten)]==1,]
  ncl <- ncol(datone)
  datcond <- datone[,-ncl]
  datenf <- as.matrix(cbind(rep(1,nrow(datcond)),datcond))
  #datenf <- as.matrix(datcond)
  jac <- matrix(nrow=2,ncol=ncol(datenf))
  x <- drop(datenf%*%para)
  p <- 1/(1 + exp(-x))
  ge <-  1/(1 + exp(-sigmoid*(p - cutoff)))
  jac[1,] <- (1/nrow(datcond))*t(datenf)%*%(sigmoid*p*(1-p)*ge*(1-ge))
  norm <- - c(0, para[-1]/sqrt(para^2[-1] + 0.0001))
  jac[2,] <- norm
  return(jac)
}
 inn <- auglag(par=star, fn=fn, gr = gradien,hin = constr,
        hin.jac = jacob)
 #inn <- constrOptim.nl(par=star, fn=fn, gr = gradien,hin = constr,
         #hin.jac = jacob)
```

```
  probapred <- 1/(1+exp(-test%*%inn$par))
}
prediction <- c(prediction,as.numeric(probapred))
parameters <- c(parameters,inn$par)
}
if(n*fold==nrow(data)){
  sensitivity <- sum(ifelse(prediction >= cutoff &
  t.labels == 1,1,0))/sum(t.labels)
  specificity <- sum(ifelse(prediction < cutoff &
  t.labels == 0,1,0))/sum((1-t.labels))
} else{
  train <- data[index[1:(n*fold)],]
  if (ncol(train) < 3){
    train <- train
    test <- sdata[index[(n*(k-1)+1):(k*n)],]
    colnames(test)[ncol(test)] <- "study_group"
    test <- test[,colnames(train)]
    test <- as.matrix(cbind(rep(1,n),test[,-ncol(test)]))
  }else{
  maxl <- maxlambda(cutof=cutoff,minsen=theta,data=train,eta=sigmoid,
          bpar=barrier)
  lamb <- seq(from = maxl, to = maxl/2, length.out=50)
  pred <- 1; j <- 1
while(pred < np){
  lambda <- lamb[j]
  tr <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta,
        data=train, eta=sigmoid, bpar=barrier)
  pred <- ncol(tr)
  j <- j + 1
}
lambda <- lamb[j-1]
train <- strulesign(lambda=lambda, cutof=cutoff, minsen=theta,
data=train, eta=sigmoid, bpar=barrier)
test <- sdata[index[(n*fold+1):nrow(data)],]
colnames(test)[ncol(test)] <- "study_group"
```

```
 test <- test[,colnames(train)]

 test <- as.matrix(cbind(rep(1,nrow(test)),test[,-ncol(test)]))
 #test <- as.matrix(test[,-ncol(test)])
 }
 if(ncol(train)==1){
 probapred <- rep(1,nrow(test))
 }else{
 star <- c(0,rep(0.001,(ncol(train)-1)))
# star <- start(data=train,sigmoid=sigmoid,cutoff=cutoff,lambda=lambda,
  #theta=theta,barrier=barrier)

 fn <- function(para){
   daten <- train
   sigmoid <-  sigmoid
   cutoff <- cutoff
   lambda <- lambda
   datzero <- daten[daten[,ncol(daten)]==0,]
   datcondzero <- datzero[,-ncol(datzero)]
   datenfzero <- as.matrix(cbind(rep(1,nrow(datzero)),datcondzero))
   #datenfzero <- as.matrix(datcondzero)
   predictorzero <- drop(datenfzero%*%para)
   probazero <- 1/(1 + exp(-predictorzero))
   senszero <- sum(1/(1 + exp(sigmoid*(probazero - cutoff))))
   senszero <- senszero/nrow(datzero)
   return(-senszero)
 }

 gradien <- function(para){
   daten <- train
   sigmoid <-  sigmoid
   cutoff <- cutoff
   lambda1 <- lambda
   datzero <- daten[daten[,ncol(daten)]==0,]
   datcondzero <- datzero[,-ncol(datzero)]
```

```
  datenfzero <- as.matrix(cbind(rep(1,nrow(datzero)),datcondzero))
  #datenfzero <- as.matrix(datcondzero)
  predictorzero <- drop(datenfzero%*%para)
  p <- 1/(1 + exp(-predictorzero))
  g <- 1/(1 + exp(sigmoid*(p - cutoff)))
  grad<- (1/nrow(datzero))*t(datenfzero)%*%(sigmoid*p*(1-p)*g*(1-g))
  return(as.numeric(t(grad)))
}

constr <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
  lambda1 <- lambda
  minsens <- theta
  h <- rep(NA, 2)
  datone <- daten[daten[,ncol(daten)]==1,]
  datcond <- datone[,-ncol(datone)]
  datenf <- as.matrix(cbind(rep(1,nrow(datone)),datcond))
  #datenf <- as.matrix(datcond)
  predictor <- drop(datenf%*%para)
  proba <- 1/(1 + exp(-predictor))
  sens <- sum(1/(1 + exp(-sigmoid*(proba - cutoff))))
  sens <- sens/nrow(datone)
  h[1] <- sens - minsens
  h[2] <- lambda - sum((para^2[-1] + 0.0001)^0.5)
  return(h)
}
jacob <- function(para){
  daten <- train
  sigmoid <-  sigmoid
  cutoff <- cutoff
  lambda <- lambda
  datone <- daten[daten[,ncol(daten)]==1,]
  ncl <- ncol(datone)
```

```
    datcond <- datone[,-ncl]
    datenf <- as.matrix(cbind(rep(1,nrow(datcond)),datcond))
    #datenf <- as.matrix(datcond)
    jac <- matrix(nrow=2,ncol=ncol(datenf))
    x <- drop(datenf%*%para)
    p <- 1/(1 + exp(-x))
    ge <-  1/(1 + exp(-sigmoid*(p - cutoff)))
    jac[1,] <- (1/nrow(datcond))*t(datenf)%*%(sigmoid*p*(1-p)*ge*(1-ge))
    norm <- - c(0, para[-1]/sqrt(para^2[-1] + 0.0001))
    jac[2,] <- norm
    return(jac)
  }
  inn <- auglag(par=star, fn=fn, gr = gradien, hin = constr, hin.jac = jacob)
  #inn <- constrOptim.nl(par=star, fn=fn, gr = gradien,hin = constr,
  #       hin.jac = jacob)
  probapred <- 1/(1+exp(-test%*%inn$par))
  }
  prediction <- c(prediction,as.numeric(probapred))
  parameters <- c(parameters,inn$par)
  sensitivity <- sum(ifelse(prediction >=cutoff & t.labels == 1,1,0))/
                 sum(t.labels)
  specificity <- sum(ifelse(prediction < cutoff & t.labels == 0,1,0))/
                 sum((1-t.labels))
  }
  allpred <- list(sensitivity=sensitivity,specificity=specificity,
             prediction=prediction,t.labels=t.labels,parameters=parameters)
  return(allpred)
  }

#result1 <- cv_sp(data=daten,theta=0.9,sigmoid = 15,cutoff=0.55,
#standard=TRUE,np=15)

#specoptimgenes <- result1
```

```
bestr <- optimr(data = daten, theta = 0.9, sig = 10,
         cutf=cutf, npreds = npreds)
 optimr <- function(data = daten, theta = 0.9, sig = 10, cutf,
                    npreds = c(10,20,30)){
    res <- numeric()
    for(i in 1:length(sig))
    {
    a <- sig[i]
    for(j in 1:length(npreds))
    {
      b <- npreds[j]
      for(k in 1:length(theta))
      {
        d <- theta[k]
        for(l in 1:length(cutf))
        {
          cu <- cutf[l]
          result1 <-  cv_sp(data = daten,theta = d, sigmoid = a,
          cutoff = cu, standard = TRUE, np = b)
          if(result1$sensitivity > 0.89 & result1$specificity > 0.1)
          {
          vec <- c(result1$sensitivity, result1$specificity, a, b, d,cu)
          res <- rbind(res, vec)
          true_labels <- result1$t.labels
          predicted_prob <- result1$prediction
          }
          else
          {
          vec <- c(1, 0, a, b, d,cu)
            res <- rbind(res, vec)
            predicted_prob <- rep(1,nrow(data))
            true_labels <- daten[,ncol(data)]
          }
        }
      }
```

```
    }
  }
  best1 <- res[which(res[,2] == max(res[,2]))[1],]
  return(list(SensSpec=best1,true_labels=true_labels,
         predicted_prob=predicted_prob))
  }

# Load data and save it as dataframe
cutf <- seq(0.3,0.7,by=0.02)
npreds <- c(5,10,15,20,25,30)
result <- optimr(data = daten,cutf = cutf, npreds = npreds)
```

Acknowledgement

I would like to express very great appreciation to my supervisor Prof. Dr. Katja Ickstadt. Your inducements and advices during this thesis were of great help; you are a great scientist I have never met before. Thank you for enabling the cooperation with Bayer Pharma AG. Without your supervision and constant help this dissertation would not have been possible. I am grateful to Bayer Pharma AG for the three-years financial support and for supplying equipments such as my office, computer and data. Many thanks to my second supervisor Prof. Dr. Jörg Rahnenführer for co-supervising this thesis. Special thank to my research supervisor Dr. Richardus Vonk for accepting me in his department at Bayer. Deepest thanks for all our fruitful discussions and your stimulating questions during our meetings.

Many thanks for motivating research questions on enrichment trials. Special thanks to Dr. Stephan Lehr for motivating sensitivity-preferred classification in biomarker research. I wish to thank my colleague Dr. Tina Müller for her support throughout this thesis. I am forever grateful for our very useful discussions, constructive criticism and for chairing your experience with me. You are a wonderful person and a great scientist. I would never forget my colleagues of *Research and Clinical Science Statistics* (RCSS) for the wonderful working atmosphere. RCSS is the best department I ever seen before with excellent and talented people. I enjoyed working there and I will truly miss my colleagues. Particular thanks to Dr. Bernd-Wolfgang Igl for reading this thesis and Dr. Hannes-Friedrich Ulbrich for sharing their experience with me.

I would like to thank Gwendolyn Reid, Rebecca Gachago, David Isreal and Dr. Arsene Ntiwa for reading this thesis. I am grateful for your comments and corrections. I am thankful to the faculty of statistics of the TU Dortmund university for offering me the opportunity to study statistics and graduate to master and doctor. Finally, I would like to thank my family and friends, particularly my wife Nina Nicole Agueusop and my brother Jean-Marie Dongo for their love, encouragements and patience throughout this thesis.

Declaration

I declare that this thesis is written by myself and that I exclusively used the indicated literature and resources. The thoughts taken directly or indirectly from external sources are proper marked as such.