

ABSTRACT

Yi, Bingming. Nonparametric, Parametric and Semiparametric Models for Screening and Decoding Pools of Chemical Compounds. (Under the direction of Dr. Jacqueline M. Hughes-Oliver and Dr. Sidney Stanley Young.)

During High Throughput Screening (HTS), large collections of chemical compounds are tested for potency with respect to one or more assays. In reality, only a very small fraction of the compounds in a collection will be potent enough to act as lead molecules in later drug discovery phases. Testing all compounds is neither cost-effective nor desirable. Based on the belief that chemical structure is highly related to potency of compounds, structure activity relationships (SARs) can be very helpful for selecting a handful of chemical compounds for testing.

This work investigates SARs using four different statistical methods. The first uses a latent class cell-based method. The second benefits from a fractional factorial design for optimizing the cell-based method to significantly increase hit rates. The third improves HTS efficiency by considering pooling experiments for chemical compounds in the presence of interaction and dilution. Rather than testing one compound at a time, chemical compounds are mixed together and tested by groups. Likelihood models are built and hit rates are shown to be higher than for traditional methods. The fourth solves the estimation problem in a pooling experiment by treating the pooling data as missing at random. Semiparametric models are implemented and estimators are shown to be more efficient than likelihood methods based on the same data.

**Nonparametric, Parametric and Semiparametric Models for Screening
and Decoding Pools of Chemical Compounds**

by

Bingming Yi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2002

APPROVED BY:

Jacqueline M. Hughes-Oliver
Co-Chair of Advisory Committee

Sidney Stanley Young
Co-Chair of Advisory Committee

Anastasios Tsiatis

Carla Mattos

To my family

Biography

Bingming Yi was born in Hubei, China on Feb. 22, 1968. He entered the Department of Probability and Statistics at Peking University in 1985 and earned his B.S. and M.S. in mathematical statistics in 1989 and 1994 respectively. He worked as an assistant professor at Beijing Broadcasting Institute before moving to Raleigh to attend North Carolina State University for the pursuit of Ph.D. degree. After graduation from NCSU, Bingming will join Merck & Co. to work on clinical trials as a biometrician.

Acknowledgements

I would like to express my deepest gratitude and biggest appreciation to my advisors Dr. Jacqueline M. Hughes-Oliver and Dr. Sidney Stanley Young. Not only has their guidance been immensely helpful in completing this research, but the constant encouragement and never-ending supply of patience have been equally indispensable.

My great thanks also go to Dr. Anastasios Tsiatis and Dr. Carla Mattos for serving on my committee. Their insights, suggestions and all other selfless help have contributed a great deal to my research and the final completion of this dissertation.

I would like to thank my research group members, especially Dr. Lei Zhu from GlaxoSmtihKline, Katja Remlinger and Ke Zhang from North Carolina State University. When I need help, they are always there.

Special thanks go to Dr. Sastry G. Pantula and Ms. Janice Gaddy for their constant help in providing my financial support and many other things. Their great job makes it possible for me to have a very pleasant and joyful time at North Carolina during the period from Aug. 1999 to Dec. 2002.

Finally and most importantly, I want to thank my wife, Lan Gao. Without her support and continuous encouragement, none of this work could have been accomplished.

Contents

List of Tables	vii
List of Figures	ix
1 Introduction	1
2 Latent Class Regression Analysis on the Potency of Chemical Compounds and Comparison to Recursive Partitioning	7
3 A Factorial Design to Optimize Cell-Based Drug Discovery Analysis	26
4 Statistical Models for Decoding Pools of Chemical Compounds in the Presence of Compound Interactions and Dilution	61
4.1 Literature Review	61
4.2 Atom Pairs as Covariate Class	67
4.3 ZHY Models and Assumptions	68
4.4 Model Extensions	70
4.5 Application to Real Data and Simulation Study	72
4.5.1 The Data	72
4.5.2 Application and Assessment of Models	75
4.5.3 Simulation Study	79
4.6 Threshold for Pool Potency	80
4.6.1 Hit Rates vs Various Threshold	82
4.6.2 Is It Possible to Find an Optimal Threshold?	84
4.6.3 A Close Look at Small Thresholds	85
4.6.4 Methodology Robust to Choice of Threshold for Pool Potency	85
4.7 A Big Pooling Data Set	88
5 Applications of Semiparametric Theory on Missing Data Problems for Decoding Pools of Chemical Compounds	95
5.1 Introduction	95
5.2 Literature Review	96

5.2.1	Influence Functions	97
5.2.2	Restricted Moment Models	99
5.2.3	Data Missing at Random	100
5.3	Semiparametric and Likelihood Models for Pooling Experiments . . .	103
5.3.1	Semiparametric Models	103
5.3.2	Likelihood Model	106
5.4	Real Data Applications	108
5.5	Another Likelihood Model	113
5.6	Future Work	118
	Bibliography	123

List of Tables

3.1	<i>Potency Proportion Quantiles of Non-Empty Cells</i>	55
3.2	<i>Factors for Creating Alternative Cell-Based Methods</i>	55
3.3	<i>Optimal Design Points for Creating Alternative Cell-Based Methods</i> .	56
3.4	<i>Summary of ANOVA Results for h_{10}, h_{20}, and h_{30}, the Hit Rate Responses at Tests of 10, 20, and 30 Compounds in the Training Set. (F is the value of F statistics)</i>	57
3.5	<i>Comparisons Between Design 7, LWY, and Design 17 for h_{10}, h_{20}, and h_{30}, the Hit Rate Responses at Tests of 10, 20, and 30 Compounds in the Training Set. (Experimentwise levels of significance are all set at 5%)</i>	57
3.6	<i>Mean Hit Rates, at All Levels of All Factors for h_{10}, h_{20}, and h_{30}, the Hit Rate Responses at Tests of 10, 20, 30 Compounds in the Training Set</i>	57
3.7	<i>Four Best Cell-Based Methods Predicted from the Statistical Analyses</i>	58
3.8	<i>Global and Initial Enhancement for Methods LWY and OPT</i>	58
4.1	<i>Distribution of Potent Compounds in the Covariate Classes</i>	73
4.2	<i>Maximum likelihood estimates of p from data DATA1 using four models, where important classes are labeled with an * beside the appropriate \hat{p}</i>	75
4.3	<i>Hit Rate Formulas for Four Methods</i>	78
4.4	<i>Hit rates for 7 methods, where ind. for individuals, ext. for extension</i>	78
4.5	<i>Means and standard deviations of hit rates in simulation</i>	80
4.6	<i>Using Pool Potency Threshold of 18.9</i>	84
4.7	<i>Using optimal threshold, hit rates results for 7 methods, where ind. for individuals, ext. for extension</i>	85
4.8	<i>Means and standard deviations of hit rates in simulation with small blocker rate</i>	85
4.9	<i>Frequency being identified as potent classes for each class</i>	87
4.10	<i>Correlation Coefficient Matrix of Duplicate Potencies for Compounds in Potent Pools</i>	90

5.1	<i>Distribution of Potent Compounds in Covariate Classes</i>	109
5.2	<i>Maximum Likelihood Estimates</i>	110
5.3	<i>Estimated Variance-Covariance Matrix of Maximum Likelihood Estimators</i>	110
5.4	<i>Estimates from the Semiparametric Model</i>	111
5.5	<i>Estimated Variance-Covariance Matrix of Semiparametric Estimators</i>	111
5.6	<i>Comparison of Estimates and Standard Deviations for Two Methods. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)</i>	111
5.7	<i>Average Results for Likelihood and Semiparametric Models. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)</i>	112
5.8	<i>Comparison of Estimations and Standard Deviations for Three Methods. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)</i>	118
5.9	<i>Average Results for Semiparametric and Two Likelihood Methods. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)</i>	118

List of Figures

2.1	Lift Charts for Four Methods, Averaged over 20 Data Sets	21
3.1	Histograms of six BCUT number descriptors.	32
3.2	Forming 1-D and 2-D cells. There are 64 1-D cells where the extreme cells each contain 1% of the compounds and the middle 62 cells have equal width. There are also 64 2-D cells each of which is formed by collapsing 8 1-D cells from each dimension.	35
3.3	Lift Charts for Three Methods, Averaged over 20 Training Data Sets.	45
3.4	Accumulation Curves for Three Methods, Averaged over 20 Training Data Sets.	46
3.5	Cumulative Recalls for Three Methods, Averaged over 20 Training Data Sets.	46
3.6	Lift Charts for LWY, Design 7 and OPT, Averaged over 20 Validation Data Sets.	49
4.1	A typical 8×12 plate used in pooling experiments.	62
4.2	Histograms of pool potency	74
4.3	Flow chart of classification using Model 1 or Model 2	76
4.4	Simulated hit rates with high blocker rate	81
4.5	Hit rates vs various thresholds for DATA1, DATA2, and DATA3	83
4.6	Simulated hit rates with smaller blocker rate	86
4.7	Scatter Plot of Duplicate Potencies for Compounds in Potent Pools	90

Chapter 1

Introduction

High Throughput Screening (HTS) technology is routinely used to identify lead molecules in the discovery of a new drug. During HTS, large collections of compounds are tested for potency with respect to one or more assays. In reality, only a very small fraction of the compounds in a collection will be potent enough to act as lead molecules in later drug discovery phases. Moreover, the expense of further investigation of leads places constraints on the number and quality of leads that will be pursued. The consequence is that out of a very large collection of compounds only a select handful will need to be identified for further study. Clearly, testing hundreds of thousands of compounds, one at a time, can be very wasteful, both in terms of time and money. Cost-effectiveness is critical for HTS programs.

For this reason, biologists, chemists, computer scientists, and statisticians have studied and searched for structure activity relationships (SARs) based on the belief that chemical structure is highly related to potency of compounds. However, the models that relate biological potency and chemical structure are usually unclear

and are often made more complex because molecular activity follows more than one mechanism. Nevertheless, it is well agreed that two molecules are more likely to have similar biological potency when they have fairly close chemical structure (McFarland and Gans, 1986). Based on this belief, many kinds of descriptors of chemical structure have been computed with the goal of improving HTS efficiency. Descriptors such as atom pairs, topological torsions, and fragments (Carhart, Smith, and Venkataraghavan, 1985; Nilakantan et al., 1987) have been useful in recursive partitioning analysis and/or pooling methods for HTS (Hung, 1993; Hawkins, Young, and Rusinko, 1997; Young and Hawkins, 1998; Rusinko et al., 1999; Zhu, Hughes-Oliver, and Young, 2001). Other researchers use the continuous BCUT descriptors given by Pearlman and Smith (1999) and derived from Burden (1989).

Using these chemical structural descriptors, our research goal is to develop sensible statistical models for improving the efficiency of HTS. From the point of view of distributional assumptions, our work can be divided into three parts: parametric, non-parametric, and semiparametric. From the point of view of ways for testing compounds, our work can be divided in two different ways: individual testing and pooled decoding. Chapter 2 investigates a parametric enhancement to a non-parametric approach to modeling data obtained by testing individual compounds. Chapter 3 extends a nonparametric approach to modeling data also obtained by testing individuals. Chapters 4 and 5 investigate models for data obtained primarily from pools or mixtures of compounds. Chapter 4 applies a parametric approach while Chapter 5 applies a semiparametric approach and more likelihood models.

A recently proposed cell-based method, due to Lam, Welch and Young (2002),

appears to be uniformly more efficient than the highly successful method of recursive partitioning. The cell-based method develops a relational model between biological potency of a compound and that compound's BCUT numbers. This model is built from a relatively small selection of compounds (called the training set) for which both chemical structure and biological potency are known or available. For the remainder of the compound collection (called the validation set), only the information on chemical structure of compounds is available for determining prediction of potency.

The cell-based method implements investigation of SARs by implicitly accounting for the effects of descriptors; no explicit functional forms are suggested. It is a non-parametric method. Metric spaces and subspaces of BCUT numbers are binned into many fine cells. By evaluating the proportion of potent compounds in each cell, good cells can be identified. Compounds in the validation set will then be ranked according to how often they lie in these good cells. Biological testing will proceed, based on ranks, until reaching a desired number of potent compounds or number of tests.

In Chapter 2, a Latent Class model using BCUT numbers as covariates is applied to enhance the cell-based method. After ranking compounds in the validation set, it is reasonable to assume that highly ranked compounds follow different distributions than compounds with lower ranks. Compounds with extremely high ranks are tested in the order suggested by the cell-based method. Compounds with medium rank are not expected to yield a large number of potent compounds, so extra effort is expended in refining the order of testing. We propose a Latent Class model for predicting the potencies of these compounds with medium ranks. Updated ranks, which in turn determine a new testing order, are assigned to these compounds and hit rates can be

improved beyond the cell-based method. The paper in Chapter 2 that describes this work appears in *Proceedings of the 2001 Joint Statistical Meetings*.

The research that led to Chapter 2 raised several questions concerning the cell-based method of Lam, et al. (2002). It was not clear how certain choices in the method affected the outcome or why the stated decisions were made. Many factors could impact performance of the cell-based method on potency prediction, and optimum levels of factors are not guaranteed to be the same for all applications. In Chapter 3, we propose using the training set to conduct an optimal fractional factorial experiment for determining the effect of seven factors on potency prediction from the cell-based method. The result is that careful selection of factor levels can dramatically improve potency prediction. The paper in Chapter 3 that describes this work has appeared in the *Journal of Chemical Information and Computer Sciences*, 42, 1221-1229, September 2002.

In the pharmaceutical industry, pooling experiments are conducted for the purpose of increasing compound throughput. Samples of different chemical compounds are combined together for a group test with respect to a particular biological assay. A positive pool test indicates that the pool contains at least one potent compound, while a negative pool test does not necessarily mean there are zero potent compounds in that pool. Due to the existence of interaction or dilution effects among compounds within a pool, potency of a potent compound may be “blocked” by another compound, thus leading to a negative pool test.

Based on only the pool potencies, can we make prediction for the potencies of

individual compounds and hence improve HTS efficiency? With the help of chemical structural descriptors, the answer is yes. This further investigation of individual potency based on pool results is called decoding of pools. In Dorfman classical decoding (Dorfman, 1943), all compounds in all potent pools are individually retested for potency. Potent compounds in negative pools are neglected. Considering the interaction effect within pools, Zhu, Hughes-Oliver, and Young (2001) suggested two parametric models (called ZHY models) to improve upon classical decoding. In Chapter 4, we propose extensions of these models to allow more practical assumptions regarding blocking effect.

Also in Chapter 4, we investigate the dilution effect by considering smaller thresholds for pool potency than that for individual potency. If information about the “correct” threshold for pool potency is available, some models present better behaviors than others. We actually investigate the performance of different models under various thresholds. Based on the performance of these models, we propose a method that is more effective for predicting potency and is robust to different thresholds. Thus, under the more practical situation that the “correct” threshold is unknown, the method is able to determine an appropriate threshold from the data.

The pooling models in Chapter 4 are parametric. Potency prediction is the only criterion used to evaluate whether or not a model is good, that is, more focus is placed on the HTS issue rather than on parameter estimations. In Chapter 5, we propose a semiparametric model that puts focus on the parameter estimation problem. Semiparametric theory for missing data can be successfully applied to pooling experiments because pooling data can be regarded as incomplete data. After obtaining pooled re-

sponses, an individual retesting procedure is usually implemented. For pools with no retesting, all individual potencies are missing. If pools are selected for retesting based on pool potencies (for example, pools with higher potencies are more likely to be individually retested), then the missing mechanism depends on observed data only. Consequently, data from pooling experiments can be regarded as data “missing at random,” and hence inverse probability estimating equations can be used to estimate potency parameters. The results in Chapter 5 show that semiparametric estimators are more efficient than traditional maximum likelihood estimators obtained from the incomplete data of pooling experiments.

Chapter 2

Latent Class Regression Analysis on the Potency of Chemical Compounds and Comparison to Recursive Partitioning

This chapter is published as an article in *Proceedings of the 2001 Joint Statistical Meetings*.

Latent Class Regression Analysis on the Potency of Chemical Compounds and Comparison to Recursive Partitioning

Bingming Yi¹, Jacqueline M. Hughes-Oliver¹,
S. Stanley Young², Lei Zhu²

¹Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

²Cheminformatics and Statistics Unit, GlaxoSmithKline. Research Triangle Park,
NC 27709

April, 25, 2001

Abstract

Drug discovery is dependent on finding a very small number of biologically active or potent compounds among hundreds of thousands of compounds stored in chemical libraries. Quantitative structure activity relationships suggest that potency of a compound is highly related to that compound's chemical makeup or structure. As such, a statistical model that predicts potency based on chemical structure can be very cost-effective since it would eliminate the need to individually test all the hundreds of thousands of compounds included in the library. We build a cell-based latent class regression model to accommodate differing classes of compounds and/or differing potency mechanisms. Comparisons are made between several cell-based methods and Helixtree recursive partitioning method for a real data set.

Introduction

Mixture models have been of great interest in recent years because of their successful application to many practical problems. Mixture distributions can be historically traced back to the work of Newcomb (1886) and Pearson (1894). Today, mixture models are often developed to explain the structure of multivariate categorical data. Finite mixture models assume that objects within the population come from a finite number (S) of underlying classes, G_1, G_2, \dots, G_S . The goals of analysis are to determine an appropriate value for S , to estimate the distribution and properties in each class, and to determine class membership.

Let $y_i = \{y_{ij}\}$ be the vector-valued data for the i th object, where $i = 1, \dots, N$ and $j = 1, \dots, J$. For each object, there are J response variables. Given y_i comes from class s , the conditional probability density of y_i is:

$$y_i \sim P(y_i, \theta_s) \quad s = 1, \dots, S,$$

where, the form of $P(\cdot, \cdot)$ is assumed known but θ_s is unknown. Let π_s represent the unknown proportion of objects in the population that come from class s . These proportions are restricted with condition:

$$\sum_{s=1}^S \pi_s = 1 \quad 0 < \pi_s \leq 1.$$

Hence, under finite mixture modeling, the unconditional density of the i th object is

obtained as:

$$f(y_i) = \sum_{s=1}^S \pi_s P(y_i, \theta_s).$$

As a special kind of finite mixture model, a latent class model assumes the concept of conditional independence (Goodman 1974). This means that the response variables of a particular object are assumed to be independent within each latent class. Therefore,

$$P(y_i, \theta_s) = \prod_{j=1}^J P(y_{ij}, \theta_s) \quad s = 1, \dots, S.$$

The theory underlying latent class models was presented by Lazarsfeld and Henry (1968), Goodman (1974), Haberman (1974, 1979), and Clogg and Goodman (1984, 1986).

A useful extension of latent class models is to introduce the relation between model parameters and independent covariates (Dayton and Macready, 1988; Bockenholt, 1993); this extension is called latent class regression analysis (LCRA). The covariates are fixed variables that are assumed to be known for all objects. In these models, some parameters, such as class proportions, are determined by the covariates. Consequently, the class proportions may be different for different individuals, while they are identical in the previous latent class modeling as described above.

Latent class models have proved to be useful and successful in many areas, including marketing, psychology, sociology, and education. However, it has not been applied to the problem of identifying chemical compounds that are potent in some sense and can thus be potentially developed into pharmaceuticals. This paper proposes such an analysis.

The problem of identifying potent compounds from a very large collection of com-

pounds is exacerbated by the fact that typically only a very small proportion, say 2%, will be potent. Many authors have investigated methods for improving the efficiency of identification (Hawkins, Young, and Rusinko, 1997; Langfeldt et al., 1997; Xie et al., 1998; Young and Hawkins, 1998; Zhu, Hughes-Oliver, and Young, 2001). In this paper, we will use continuous covariates that describe the chemical constructions and features of compounds to help in identifying and separating potent compounds from “junk” compounds (McFarland and Gans, 1986). A latent class model is proposed to separately identify potent and junk classes.

The remainder of this paper is organized as follows. Section 2 is a short introduction to the data used in the analysis. Section 3 describes the LCRA model and outlines the algorithm used for estimation. Section 4 introduces the cell-based method proposed by Lam, Welch, and Young (2001) (called LWY method) and Section 5 presents the cell-based LCRA. After a brief review of Helixtree recursive partitioning in Section 6, Section 7 makes comparisons between these methods and provides concluding remarks.

NCI Data, Augmented with BCUT Descriptors

The National Cancer Institute (NCI) maintains several databases for the purpose of accelerating research on treatment of HIV/AIDS. One such database is located at http://dtp.nci.nih.gov/docs/aids/aids_data.html. It provides screening results and chemical structure on about 32,000 compounds in response to a specific antiviral assay. The chemical structure has been provided electronically in the form of

a connection table giving the atoms and how they are bonded. These structures have been converted into quantitative features better suited for use in deriving quantitative structure activity relationships (QSARs). In this paper we use BCUTs to numerically characterize the chemical structure.

BCUT numbers are eigenvalues from connectivity matrices derived from the molecular graph (Burden, 1989; Pearlman and Smith, 1999). We were able to compute BCUTs for 29,812 compounds from the full NCI database, so our analysis is limited to this as the full NCI dataset. Only 608 (2.04%) of the 29,812 compounds were potent for this antiviral assay.

We select 20 pairs of training and validation sets. Each training set contains 4,096 compounds, so the associated validation set contains 25,716 compounds. These 20 training sets were selected by randomly sampling from the 29,812 compounds each with roughly 2.04% potent compounds.

Clearly, this is a case where biological potency is known for all compounds in the collection, so High Throughput Screening (HTS) is not needed. However, it provides an excellent test case for determining the effectiveness of an HTS method. For the purpose of using the relational model developed from the training set to predict potency, activities are assumed unknown for all compounds in the validation set. After the model is built using the training set, the potencies of the compounds in the validation set are “revealed” and used to determine the quality of the relational model.

LCRA: Model and Inference

For the NCI data, objects are compounds and their responses are univariate (i.e., $J = 1$), so that y_i is either 1 (potent) or 0 (not potent). This makes the latent class model quite simple in our problem:

$$f(y_i) = \pi_1 \Pr(y_i, p_1) + \pi_2 \Pr(y_i, p_2),$$

where $\Pr(y_i, p_s) = p_s^{y_i}(1 - p_s)^{1-y_i}$, $s = 1, 2$ and $\pi_1 + \pi_2 = 1$

Further more, we can augment information by introducing BCUT number covariates into the class proportions for each compound, thus forming an LCRA model. The density of the i th compound can be rewritten as:

$$f(y_i) = \pi_{1i} p_1^{y_i} (1 - p_1)^{1-y_i} + \pi_{2i} p_2^{y_i} (1 - p_2)^{1-y_i},$$

where $\pi_{1i} = \frac{e^{\mathbf{X}_i \beta}}{e^{\mathbf{X}_i \beta} + e^{-\mathbf{X}_i \beta}}$, $\pi_{2i} = 1 - \pi_{1i}$, \mathbf{X}_i is a 6-dimensional vector of BCUT covariates, and β is a 6-dimensional parameter-vector.

So, in this problem, there are 8 parameters to estimate: $p_1, p_2, \beta_1, \dots, \beta_6$. For this purpose of estimation, Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977; Bockenholt, 1993) is used.

The idealized “complete data” are the observed data $Y = (y_1, \dots, y_n)$ combined with the missing data $\mathbf{Z} = (Z_1, \dots, Z_n)$, where $Z_i = I(\text{compound } i \in \text{class 1})$ is the indicator of compound i being in class 1.

In the E -step of the algorithm, the expectation of missing data given observed

data is

$$E(Z_i|Y) = \frac{\hat{\pi}_{1i}\hat{p}_1^{y_i}(1-\hat{p}_1)^{1-y_i}}{\hat{\pi}_{1i}\hat{p}_1^{y_i}(1-\hat{p}_1)^{1-y_i} + \hat{\pi}_{2i}\hat{p}_2^{y_i}(1-\hat{p}_2)^{1-y_i}}.$$

In the M -step, the log-likelihood function of the “complete” data is maximized with respect to $p_1, p_2, \beta_1, \dots, \beta_6$. The log-likelihood function of the “complete” data is

$$\log \mathcal{L}(\beta, p_1, p_2) = \sum_{i=1}^N (A_i + B_i),$$

where $A_i = Z_i \log \pi_{1i} + (1 - Z_i) \log \pi_{2i}$ and $B_i = Z_i \log p_1^{y_i} (1 - p_1)^{1-y_i} + (1 - Z_i) \log p_2^{y_i} (1 - p_2)^{1-y_i}$.

This log-likelihood is easy to be maximized since A_i has information based only on $\beta = (\beta_1, \dots, \beta_6)$ and B_i has information on p_1, p_2 .

For a particular iteration, estimators of p_1 and p_2 are obtained as

$$\hat{p}_1 = \frac{\sum_{i=1}^N E(Z_i|Y)y_i}{\sum_{i=1}^N E(Z_i|Y)} \quad \hat{p}_2 = \frac{\sum_{i=1}^N (1 - E(Z_i|Y))y_i}{\sum_{i=1}^N (1 - E(Z_i|Y))}$$

The parameter β can be estimated by Newton-Raphson methods.

During iteration, new parameter estimators are obtained after evaluating $E(Z_i|Y)$ using previous estimators. Convergence is usually obtained after a few iterations. Because the EM-algorithm does only guarantee convergence to a local maximum, it is important to experiment with alternative initial values.

LWY Cell-Based Method

LWY cell-based method (Lam et al., 2001) is based on dividing the six-dimensional BCUT space into cells. These cells are expected to be formed “good” enough to capture active regions that are dense with potent compounds. Once active regions have been identified, a sensible approach is applied to test only compounds that fall in these active regions. In the following subsections we describe the components of method LWY. Section 4.1 describes how cells are formed, Section 4.2 discusses which cells can be called good, and Section 4.3 introduces how to apply the model developed from training set to the prediction set.

Forming Cells

The six-dimensional BCUT space may be viewed more simply as low-dimensional subspaces. By focusing on these lower dimensions, LWY is able to create a very fine resolution of cells in order to find active regions. Moreover, these cells can be shifted to effectively multiply the number of cells available.

In the most extreme case, the six-dimensional space can be marginalized to create six one-dimensional (1-D) subspaces, one for each of the six BCUTs. Any compound falling in the six-dimensional space will also fall in exactly one cell of each of the six 1-D subspaces. All 1-D subspaces are divided into 64 disjoint bins. In order to diminish

the impact of outliers, the first bin is formed by containing the lowest 1% of values for that BCUT and the last bin is formed by containing the highest 1% of values. The remaining 62 bins have equal width between the extreme bins, irrespective of content.

To form two-dimensional (2-D) subspaces, each dimension has eight bins, each of which is formed by amalgamating eight 1-D bins into a single one. This maintains the total number of cells in the 2-D subspace to be 64, as it was for the 1-D subspace. It is easy to see that there are $\binom{6}{2} = 15$ 2-D subspaces.

Similarly, three-dimensional (3-D) subspaces are formed by amalgamating four sets of 16 1-D bins to create four bins in each dimension. There are $\binom{6}{3} = 20$ 3-D subspaces. Four-, five-, and six-dimensional (sub)spaces are excluded from use.

Beyond considering 1-D, 2-D, and 3-D subspaces, shifted cells are also created. Each dimension of a subspace is shifted to the right by half the length of one bin. So for each 1-D subspace, another 1-D subspace with 64 shifted bins is formed after shifting. For each 2-D subspace, with two dimensions allowed to shift, a total of four subspaces are formed (including the original subspace without shifting). Similarly, for each 3-D subspace, a total of eight subspaces are formed after shifting.

Consequently, we have a total of $6 \times 2 + 15 \times 4 + 20 \times 8 = 232$ subspaces, each of which has 64 cells.

Identifying Good Cells

After all cells are formed, cells with two or more active compounds will be separated from others and called the preliminary good cells. From now on, we will work on only these preliminary good cells.

In order to further evaluate the preliminary good cells, a statistic is proposed to rank the cells. Now consider a cell with a total of n compounds and a of them are potent. If we assume the number of potent compounds follows the binomial distribution, we can find the 95% confidence limit for the potency proportion. (This proportion, the number of active compounds identified divided by the number of tests performed, is called the hit rate when testing compounds in this cell.)

The lower bound of the confidence interval can be used as a good criterion to know whether or not a cell is active. This lower 95% bound is labeled H_{L95} and defined as

$$H_{L95} = \min_p \{\Pr[\text{Bin}(n, p) \geq a] \geq 0.05\}.$$

After the H_{L95} are computed for all preliminary good cells, we also need a cutoff value to separate good cells from the others. We use a permutation test to find the cutoff.

Suppose we reorder the potencies of compounds in the training set but not the BCUT numbers associated with the compounds. This means that the potencies are randomly reassigned to the BCUT descriptors. Ideally, after this permutation no cells should be identified as good cells and large values of the H_{L95} of some cells are just the results of chance. So, we can perform the random permutation a thousand times and choose an H_{L95} value corresponding to the 95th-percentile after ordering the H_{L95} values obtained from all permutations. This 95th-percentile can be used as

the cutoff value H_0 . Returning to the true training set, all cells with H_{L95} larger than H_0 will be regarded as good cells.

It is computationally expensive to implement large number of permutations. Fortunately for the cell-based analysis, one permutation is equivalent to reassigning the potency in all 14,848 cells. This is much like doing thousands of random permutations or re-ordering the potency for each of 232 subspaces. Thus, we can take the 95th-percentile from one permutation as the cutoff H_0 to determine whether or not there are any real good cells in these preliminary good cells.

Prediction and Validation

Until now we have been working in the world of the training set. It is time to make predictions. Once good cells have been identified, we will select from the validation set some compounds living in these good cells. We will test these compounds one at a time. Which compounds should be tested first? This is very important since the order has a big impact on hit rate results.

So some criterion is needed to determine the order of testing. Frequency of occurrence of living in good cells and the quality of the inhabited cell are important features. For compound C_i , the score is defined as

$$S_i = \sum_{cell_k \text{ is good}} I(C_i \in cell_k) H_{L95,k},$$

where $I(\cdot)$ is the indicator function and $H_{L95,k}$ is the H_{L95} value for $cell_k$. The higher the score the greater the chance the compound will be potent, so we can rank the compounds in the validation set by the order of scores from high to low. These

compounds will be tested in this order and the hit rates can be computed.

Cell-Based LCRA

In LWY method, scores are computed only for compounds in the validation set. Actually, we can do the same thing for compounds in the training set: rank the compounds by their scores and compute the hit rates.

Let T_1, T_2, \dots, T_n denote the compounds in the training set, h_1, h_2, \dots, h_n denote the corresponding hit rates and TS_1, TS_2, \dots, TS_n denote their scores ($n = 4096$). Suppose these compounds have been ordered by their scores from high to low. We can find n_0 such that h_1, h_2, \dots, h_{n_0} are equal to 100% but $h_{n_0+1} < 100\%$. However, when $h_1 < 100\%$, let $n_0 = 0$.

Let V_1, V_2, \dots, V_n denote the compounds from the validation set ($n = 25716$) and VS_k denote the score of compound V_k . For the compounds with scores no less than TS_{n_0} , say V_1, V_2, \dots, V_m , we will test them directly. For the other compounds of interest, say $V_{m+1}, V_{m+2}, \dots, V_{m+l}$, we will determine the testing order by LCRA. In this paper, $m + l$ is set to 200.

Returning to the training set, those compounds with scores between TS_{n_0} and VS_{m+l} are selected out for use. In the special case when $n_0 = 0$, those compounds with scores no less than VS_{m+l} are selected. The LCRA model proposed in Section 3 is built upon these compounds from the training set. Once the estimates of model parameters are obtained, the class proportions can be easily computed for the compounds $V_{m+1}, V_{m+2}, \dots, V_{m+l}$ in the validation set. Remember these proportions are

denoted as $\pi_{1,m+1}, \pi_{1,m+2}, \dots, \pi_{1,m+l}$ for class 1 (set as potent class). These proportions provide information on testing order. The higher the proportions, the bigger the chance of being tested earlier.

Thus, the cell-based LCRA is a two stage method. After applying LWY cell-based method, the compounds from the validation set with large scores will be tested directly by the LWY scheme. For the compounds with medium scores, we will use LCRA to determine the testing order.

Helixtree Recursive Partitioning Method

Recursive Partitioning (RP) is a tree-based method. Many implementations of this method obtain good results (Hawkins et al., 1997; Young and Hawkins, 1998; Rusinko et al., 1999).

Recursively, it splits a data set into disjoint subsets called nodes. A node may be split into two or more daughter nodes. The most significant of all possible splits is searched and implemented for each node. When there are no more significant splits or the stopping criterion is reached, terminal nodes are achieved. The tree obtained from the training set is called the training tree.

RP splits a node into two or more nodes using t -tests of observed potencies from two groups formed by separating compounds according to high or low values of a BCUT covariate. This is repeated for all BCUT covariates. The descriptors can be discrete or continuous. The Helixtree RP method here uses 6 continuous BCUT numbers as descriptors.

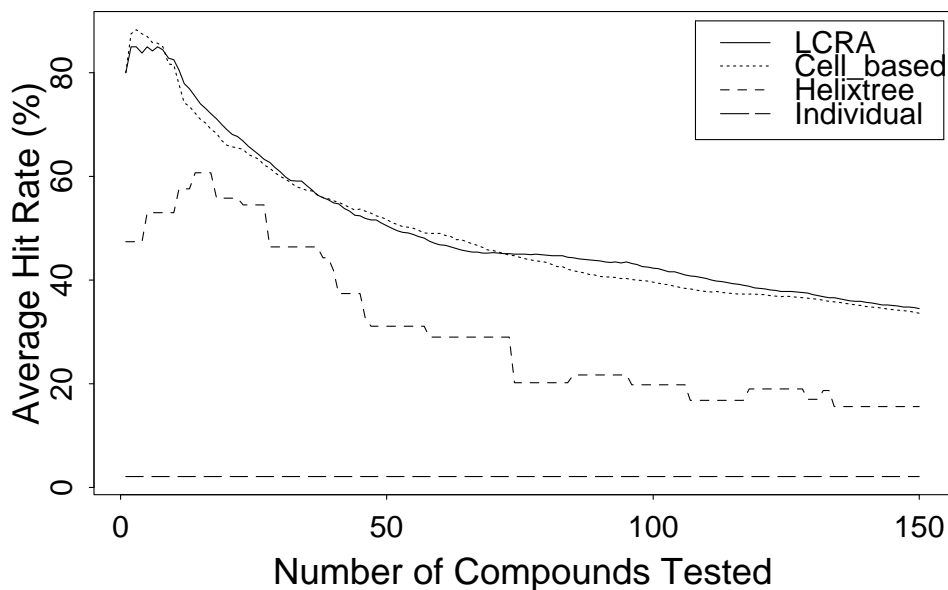


Figure 2.1: Lift Charts for Four Methods, Averaged over 20 Data Sets

Once the training tree is available, its partitioning criteria will be applied to the compounds in the validation set to build a corresponding prediction tree. The compounds in the prediction tree are then ranked by the hit rates of nodes in which they fall and will be tested by the order of these ranks from high to low. Testing results determine the lift chart as discussed in Section 7.

Software information for Helixtree is available on the website www.goldenhelix.com.

Lift Chart Results and Concluding Remarks

We make comparisons between LWY, cell-based LCRA and Helixtree recursive partitioning methods by plotting lift charts of their hit rates.

A lift chart is a convenient and informative graphic for viewing the results of an HTS method. The lift chart plots the number of compounds tested in the validation set versus the cumulative percent of potent compounds found (also known as the hit rate) from this testing. Obviously, the higher the hit rate, the better.

Figure 3.6 summarizes lift chart results from four different methods: cell-based LCRA, LWY, Helixtree RP method and random testing.

For each method and each validation set, a lift chart is obtained. Because 20 data sets (that is, 20 replicates) are used, we average the 20 lift charts and these are the plotted curves in Figure 3.6.

We can see that random testing, which has a constant hit rate of about 2%, is much worse than the other three methods and that the two cell-based methods are in turn better than recursive partitioning. RP selects only one descriptor at a time to split the node. However, when the mechanism of a split depends on two or more descriptors simultaneously, RP tends to give improper results.

The two cell-based methods give comparable results. There is evidence that class membership as derived from LCRA based on structural descriptors is related to compound potency. This suggests that LCRA on its own may be an effective HTS method. However, because LWY and LCRA apparently target similar relationships, applying LCRA after LWY provides little additional benefit.

Bibliography

- [1] Bockenholt, Ulf (1993). A Latent-Class Regression Approach for the Analysis of Recurrent Choice Data. *British Journal of Mathematical and Statistical Psychology*, **46**, pp.95-118.
- [2] Burden, F. R. (1989). Molecular Identification Number for Substructure Searches. *Journal of Chemical Information and Computer Sciences*, **29**, pp.225-227.
- [3] Clogg, C. C. and Goodman, L. A. (1984). Latent Structure Analysis of a Set of Multidimensional Contingency Tables. *Journal of the American Statistical Association*, **79**, pp.762-771.
- [4] Clogg, C. C. and Goodman, L. A. (1986). On Scaling Models Applied to Data From Several Groups. *Psychometrika*, **51**, pp.123-135.
- [5] Dayton, C. M. and Macready, G. B. (1988). Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association*, **83**, **401**, pp.173-178.
- [6] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM-algorithm. *Journal of the Royal*

Statistical Society, Series B **39**, pp.1-22.

- [7] Goodman, L. A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, **61**, pp.215-231.
- [8] Haberman, S. J. (1974). Log-Linear Model for Frequency Tables Derived by Indirect Observation: Maximum-Likelihood Equations. *The Annals of Statistics* **2**, pp.911-924.
- [9] Haberman, S. J. (1979). *Analysis of Qualitative Data, Volume 2: New Developments*. New York: Academic Press.
- [10] Hawkins, D. M., Young, S. S., and Rusinko A. (1997). Analysis of a Large Structure-activity Data Set Using Recursive Partitioning. *Quantitative Structure-Activity Relationship*, **16**, pp.296-302.
- [11] Lam, R. L. H., Welch, W. J., and Young, S. S. (2001). Cell-Based Analysis for Large Chemical Databases. *Technometrics* submitted.
- [12] Langfeldt, S. A., Hughes-Oliver, J. M., Ghosh, S., and Young, S. S. (1997). Optimal Group Testing in the Presence of Blockers. *Institute of Statistics Mimeograph Series No. 2297*.
- [13] Lazarsfeld, P. F. and Henry, N. W. (1968). Latent Structure Analysis. *Boston: Houghton Mifflin*.
- [14] McFarland, J. W. and Gans, D. J. (1986). On the Significance of Clusters in the Graphical Display of Structure-Activity Data. *Journal of Medicinal Chemistry* **29**, pp.505-514.
- [15] Newcomb, S. (1886). A Generalized Theory of the Combination of Observations

- so as to Obtain the Best Result. *American Journal of Mathematics* **8**, pp.343-366.
- [16] Pearlman, R. S. and Smith, K. M. (1999). Metric Validation and the Receptor-Relevant Subspace Concept. *Journal of Chemical Information and Computer Sciences* **39**, pp.28-35.
- [17] Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Trans. A* **185**, pp.71-110.
- [18] Rusinko, A., Farnen, M. W., Lambert, C. G., Brown, P. L., and Young, S. S. (1999). Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *Journal of Chemical Information and Computer Sciences* **38**, pp.1017-1026.
- [19] Xie, M., Tatsuoka, K., Sacks, J., and Young, S. S. (2001). Group Testing with Blockers and Synergism. *Journal of the American Statistical Association* **96**, pp.92-102.
- [20] Young, S. S. and Hawkins, D. M. (1998). Using Recursive Partitioning to Analyze a Large SAR Data Set. *SAR and QSAR in Environmental Research*, **8**, pp.183-193.
- [21] Zhu, L., Hughes-Oliver, J. M., and Young, S. S. (2001). Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics* **57**, pp.922-930.

Chapter 3

A Factorial Design to Optimize Cell-Based Drug Discovery Analysis

This chapter is published as an article in *Journal of Chemical Information and Computer Sciences*, 42, 1221-1229, Sept. 2002

A Factorial Design to Optimize Cell-Based Drug Discovery Analysis

Bingming Yi¹, Jacqueline M. Hughes-Oliver¹,
Lei Zhu², S. Stanley Young²

¹Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

²Cheminformatics and Statistics Unit, GlaxoSmithKline. Research Triangle Park,
NC 27709

March, 1, 2002

Abstract

Drug discovery is dependent on finding a very small number of biologically active or potent compounds among millions of compounds stored in chemical collections. Quantitative structure activity relationships suggest that potency of a compound is highly related to that compound's chemical makeup or structure. In order to improve the efficiency of cell-based analysis methods for high throughput screening, where information of a compound's structure is used to predict potency, we consider a number of potentially influential factors in the cell-based approach. A fractional factorial design is implemented to evaluate the effects of these factors and lift chart results show that the design scheme is able to find conditions that enhance hit rates.

INTRODUCTION

High Throughput Screening (HTS) technology is routinely used to identify lead molecules in the discovery of a new drug. During HTS, large collections of compounds are tested for potency with respect to one or more assays. In reality, only a very small fraction of the compounds in a collection will be potent enough to act as lead molecules in later drug discovery phases. Moreover, the expense of further investigation of leads places constraints on the number and quality of leads that will be pursued. The consequence is that out of a very large collection of compounds only a select handful will need to be identified for further study. Clearly, testing hundreds of thousands of compounds, one at a time, can be very wasteful, both in terms of time and money. Cost-effectiveness is critical for HTS programs.

For this reason, biologists, chemists, computer scientists, and statisticians have studied and searched for structure activity relationships (SARs), which are built on the belief that chemical structure is highly related to potency of compounds. However, the models that relate biological potency and molecular structure are usually unclear and are often made more complex because molecular activity follows more than one mechanism. Nevertheless, it is well agreed that two molecules with fairly close molecular structure will have similar biological potency^[1]. Based on this belief, many kinds of descriptors of molecular structure have been computed with the goal of improving HTS efficiency. Descriptors such as atom pairs, topological torsions, and fragments^{[2],[3]} have been useful in recursive partitioning analysis and/or pooling

methods for HTS^{[4]–[7]}. Other researchers use the continuous BCUT descriptors given by Pearlman and Smith^[8] and derived from Burden^[9]. Lam, Welch, and Young^{[10],[11]} propose a cell-based analysis (we call it the LWY method) using BCUT numbers.

The purpose of this paper is to improve the LWY methodology for screening large collections of compounds with respect to biological potency based on testing only a small fraction of the compounds. The approach develops a relational model between biological potency of a compound and its chemical structure. This relational model is built upon a relatively small selection of compounds (called the training set) for which both chemical structure and biological potency must be known or available. For the remainder of the compound collection (called the validation set), only information on chemical structure of compounds is needed. Compounds in the validation set will then be ranked according to predicted biological potency, using the relational model developed from the training set, and testing will proceed based on ranks until a desired number of potent compounds have been identified to serve as leads in the remainder of the drug discovery process.

Lam et al.^{[10],[11]} have already demonstrated that great gains can be achieved using the LWY method. We argue that even greater gains are possible by tuning the method as well as incorporating modifications that make the method less computationally intensive and hence more feasible. We illustrate these methods using a dataset from the National Cancer Institute (NCI). The second section contains a short description of the NCI data. The third section describes the cell-based method LWY in three subsections. The fourth section explores possible modifications to LWY using a fractional factorial experimental approach. The “best” cell-based methods are selected

in the fifth section and confirmatory results are shown in the sixth section. We close with a summary and discussion in the final section.

NCI DATA, AND BCUT DESCRIPTORS

The NCI maintains databases for the purpose of accelerating research on treatment of HIV/AIDS. One such database is located at http://dtp.nci.nih.gov/docs/aids/aids_data.html. It provides screening results and chemical structure on about 32,000 compounds for a specific antiviral assay. The chemical structures are provided electronically in the form of a connection table giving the atoms and how they are bonded. These structures have been converted into quantitative features better suited for use in deriving quantitative structure activity relationships (QSARs). In this paper we use BCUTs to numerically characterize the chemical structure.

BCUT numbers are eigenvalues from connectivity matrices derived from the molecular graph^{[8],[9]}. These numbers can be defined according to a variety of atomic properties such as size, atomic number, charge, etc. A different BCUT descriptor is created for each atomic property selected. More than 60 BCUT descriptors are in common use, but the high degree of multicollinearity existing between them has resulted in fairly common use of only six relatively uncorrelated BCUT descriptors. Given a particular atomic property of interest, a connectivity matrix from a molecular graph is constructed as a square matrix with dimension equal to the number of heavy (non-hydrogen) atoms. The property is placed along the diagonal for each

heavy atom, while off-diagonal elements measure the degree of connectivity between two heavy atoms. Because eigenvalues are matrix invariant, they measure properties of the molecular graph. Being functions of all the heavy atoms in the molecule, the eigenvalues are thought to represent the properties of the molecule as a whole.

We were able to compute BCUTs for 29,812 compounds from the full NCI database, so our analysis is limited to this as the full NCI data set. Only 608 (2.04%) of the 29,812 compounds were potent for this antiviral assay. The histograms shown in Figure 3.1 indicate that BCUTs 1–6 capture different features as evidenced by the differences in central tendency and dispersion. There exist relatively low correlations among these six BCUTs, and the correlation coefficients range from -0.49 to 0.46 , as illustrated in the following correlation matrix.

$$\begin{bmatrix} & \text{BCUT1} & \text{BCUT2} & \text{BCUT3} & \text{BCUT4} & \text{BCUT5} & \text{BCUT6} \\ \text{BCUT1} & 1.00 & -0.22 & -0.42 & -0.16 & 0.46 & 0.32 \\ \text{BCUT2} & & 1.00 & 0.20 & 0.32 & -0.25 & -0.49 \\ \text{BCUT3} & & & 1.00 & 0.08 & -0.25 & -0.17 \\ \text{BCUT4} & & & & 1.00 & -0.03 & -0.16 \\ \text{BCUT5} & & & & & 1.00 & 0.24 \\ \text{BCUT6} & & & & & & 1.00 \end{bmatrix}$$

As previously mentioned, we need to select a training set from which the relational model will be developed. However, to assess issues of sensitivity of the method to the particular choice of training set, we select 20 pairs of training and validation sets. Each training set contains 4,096 compounds (13.7% of the collection), so the associated validation set contains 25,716 compounds. These 20 training sets were selected by randomly sampling from the 29,812 compounds. They contain between 82 (2.00%) and 87 (2.12%) potent compounds.

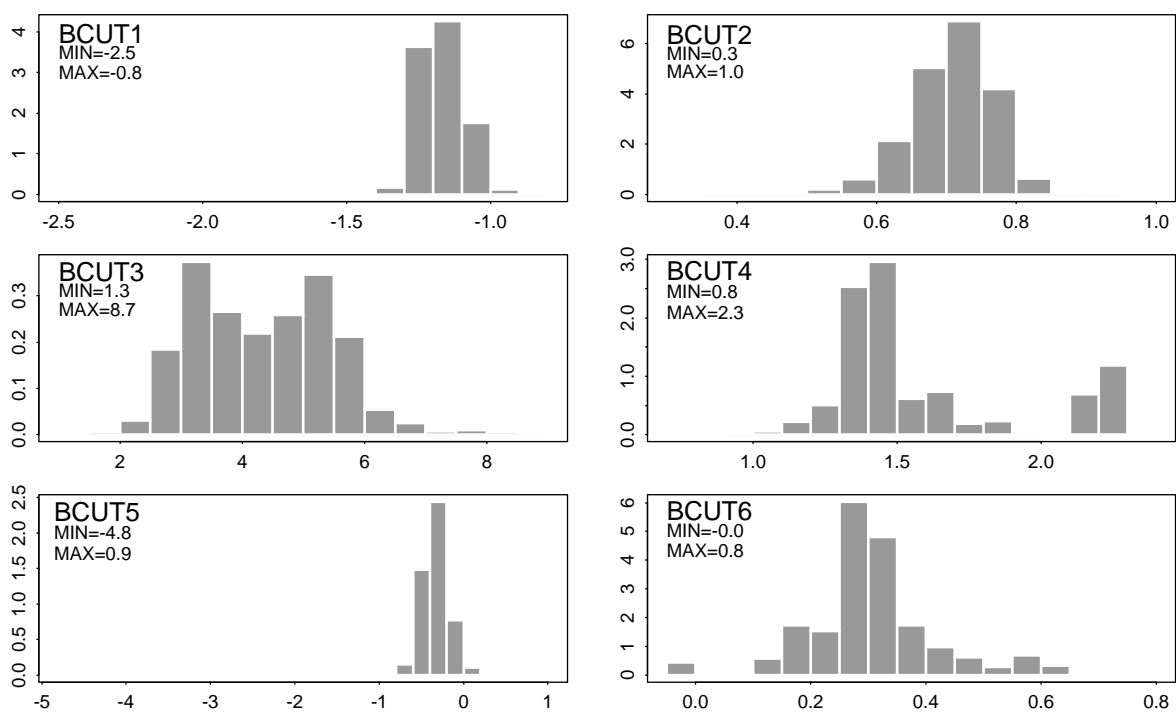


Figure 3.1: Histograms of six BCUT number descriptors.

Clearly, this is a case where biological potency is known for all compounds in the collection, so prediction is not needed. However, it provides an excellent test case for determining the effectiveness of a method. For the purpose of using the relational model developed from the training set to predict potency, activities are assumed unknown for all compounds in the validation set. After the model is built using the training set, the potencies of the compounds in the validation set are “revealed” and used to determine the quality of the relational model.

METHOD LWY, A CELL-BASED METHOD

Method LWY, as proposed by Lam et al.^{[10],[11]}, is based on dividing the six-dimensional BCUT space into cells. These cells are expected to be formed “good” enough to capture active regions that are dense with potent compounds. Once active regions have been identified, a sensible prediction approach is to test only compounds that fall in these active regions. This binning approach raises several questions, however. In the following subsections we describe the components of method LWY:

- How are cells formed? How do we divide a six-dimensional descriptor space into small enough regions to capture a high proportion of potent compounds without making the regions so small that they are too numerous to handle or contain too few compounds to be useful? This is discussed in the first subsection.
- How are cells classified as being good? Is a cell containing four potents among

five compounds as good as a cell containing two potents among two compounds?

This is discussed in the second subsection.

- The relational model developed in the two previous steps must now be applied to the validation set. How is this done? This is discussed in the third subsection.

Forming Cells

The six-dimensional BCUT space may be viewed more simply as low-dimensional subspaces, for example, all one-dimensional, two-dimensional and three-dimensional subspaces. By focusing on these lower dimensions, LWY is able to create a very fine resolution of cells in order to find active regions. Moreover, these cells can be shifted to effectively multiply the number of cells available.

In the most extreme case, the six-dimensional space can be marginalized to create six one-dimensional (1-D) subspaces, one for each of the six BCUTs. Any compound falling in the six-dimensional space will also fall in exactly one cell of each of the six 1-D subspaces. The disadvantage of 1-D subspaces is that information on the relationship with other BCUTs is lost; the advantage is that we have a much smaller space that will accommodate a finer resolution of cells. All 1-D subspaces are divided into 64 disjoint bins. In order to diminish the impact of outliers (common in chemistry data sets), the first bin is formed by containing the lowest 1% of values for that BCUT and the last bin is formed by containing the highest 1% of values. The remaining 62 bins have equal width between the extreme bins, irrespective of content.

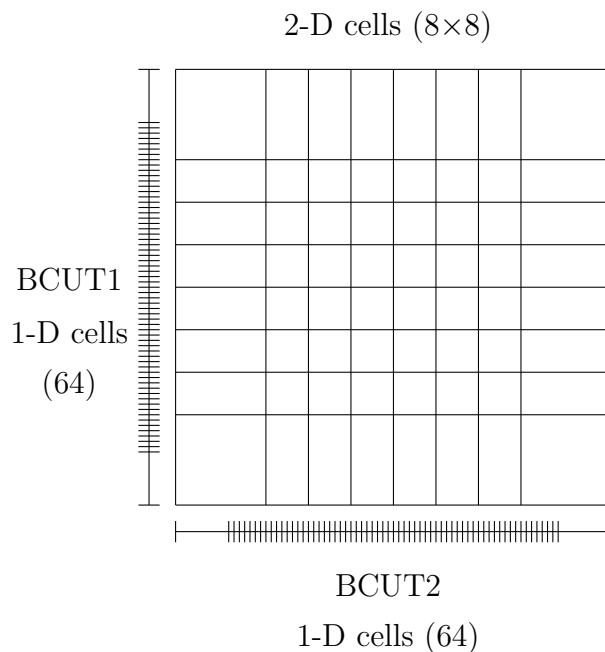


Figure 3.2: Forming 1-D and 2-D cells. There are 64 1-D cells where the extreme cells each contain 1% of the compounds and the middle 62 cells have equal width. There are also 64 2-D cells each of which is formed by collapsing 8 1-D cells from each dimension.

To form two-dimensional (2-D) subspaces, each dimension has eight bins that are formed by amalgamating eight 1-D bins into a single bin. This maintains the total number of cells in the 2-D subspace to be 64, as it was for the 1-D subspace. Figure 3.2 illustrates this amalgamating process. It is easy to see that there are $\binom{6}{2} = 15$ 2-D subspaces.

Similarly, three-dimensional (3-D) subspaces are formed by amalgamating 16 1-D bins to create four bins in each dimension. There are $\binom{6}{3} = 20$ 3-D subspaces. Four-, five-, and six-dimensional (sub)spaces are excluded from use.

Beyond considering 1-D, 2-D, and 3-D subspaces, shifted cells are also created.

Each dimension of a subspace is shifted to the right by half the length of one bin. So for each 1-D subspace, another 1-D subspace with 64 shifted bins is formed after shifting. For each 2-D subspace, with two dimensions allowed to shift, a total of four subspaces are formed (including the original subspace without shifting). Similarly, for each 3-D subspace, a total of eight subspaces are formed after shifting.

Consequently, we have a total of $6 \times 2 + 15 \times 4 + 20 \times 8 = 232$ subspaces. Each subspace has 64 cells, so that we have a total of $232 \times 64 = 14,848$ cells. It is important to remember that any given compound resides in every subspace. But in each subspace a compound resides in only one cell. Thus, we know each compound resides in 232 cells.

Identifying Good Cells

LWY method applied to a training set from the NCI data yields 13010 non-empty cells among total of 14848 cells. In these non-empty cells, about 43% contain at least one potent compound. Proportion quantiles of these “potent” cells are shown in Table 3.1. Clearly, a method is needed for identifying cells as being good or not. In LWY method, after all cells are formed, cells with two or more potent compounds will be separated from others and called the preliminary good cells. From now on, we will work on only these preliminary good cells.

In order to further evaluate the preliminary good cells, a statistic is proposed to rank the cells. Now consider a cell with a total of n compounds where a of them

are potent. If we assume the number of potent compounds follows the binomial distribution, we can find the 95% confidence limit for the potency proportion. (This proportion is called the hit rate when testing compounds in this cell: hit rate is the number of potent compounds identified divided by the number of compounds in the cell.)

The lower bound of the confidence interval can be used as a good criterion to know whether or not a cell is active. This lower 95% bound is labeled H_{L95} and defined as

$$H_{L95} = \min_p \{\Pr[Bin(n, p) \geq a] \geq 0.05\}.$$

After the H_{L95} are computed for all preliminary good cells, we also need a cutoff value to separate good cells from the others. We use a permutation test to find the cutoff.

The training data contains 4,096 compounds, each of which has a value of potency and six numerical BCUT descriptors. Suppose we randomly reorder the potencies of compounds, but keep the BCUT numbers in their initial positions. This means that the potencies are randomly reassigned to the BCUT descriptors. Ideally, after this permutation no cells should be identified as good cells and large values of the H_{L95} of some cells are just the results of chance. The 95th-percentile of all such H_{L95} values obtained from permutations is labeled H_0 . The value H_0 is used to distinguish between random variation and systematic variation. Returning to the true training set, all cells with H_{L95} larger than H_0 will be regarded as good cells.

It is computationally expensive to implement many permutations. Fortunately for the cell-based analysis, an equivalent way is to implement one permutation that reassigns the potencies in all 14,848 cells. This is much like doing thousands of random

permutations or re-ordering the potency for each of 232 subspaces. Thus, we can take the 95th-percentile from one permutation as the cutoff H_0 to determine whether or not there are any real good cells in these preliminary good cells.

This is also a good way to determine if BCUT descriptors are relevant to activity. (Although widely used, BCUT numbers are rather abstract and it is not clear if they contain structural information pertinent to biological activity.)

Prediction and Validation

Until now we have been working in the world of the training set. It is time to make predictions. Once good cells have been identified using the training set, we will select compounds residing in these good cells from the validation set. We will test these compounds one at a time. Which compounds should be tested first? This is very important since the order has considerable impact on hit rate results. For example, when we select 10 compounds to test sequentially, we have 10 chances to compute the hit rate after each test and obtain 10 hit rates. If all the 10 compounds selected happen to be inactive, the 10 hit rates are all zero. However, if the 10 compounds tested are all potent, the 10 hit rates are all 100%, which is the perfect result we desire.

So some criterion is needed to determine the order of testing. For this purpose, we will compute scores for the compounds in the validation set. We will determine the testing order by the scores. Frequency of occurrence of residing in good cells is a

reasonable way of defining scores. For compound C_i , the score is defined as

$$S_i = \sum_{\text{cell}_k \text{ is good}} I(C_i \in \text{cell}_k),$$

where $I(A)$ is the indicator function that takes value one when A is true and zero otherwise.

Alternatively, we can apply a weight to each good cell and compute a new score for the compounds:

$$S_i = \sum_{\text{cell}_k \text{ is good}} I(C_i \in \text{cell}_k) H_{L95,k},$$

where $H_{L95,k}$ is the H_{L95} value for cell_k .

For both scoring functions suggested above, we expect that the higher the score the greater the chance the compound will be potent, and thus it should have a higher priority to be among the first tested. So, we can rank the compounds in the validation set by the order of high score to low. These compounds will be tested by this order and the hit rate results can be computed. LWY method uses the latter as scores.

OTHER CELL-BASED METHODS

In the cell-based method proposed by Lam et al.^{[10],[11]}, many tuning factors may be investigated. For example, we can consider only 3-D subspaces rather than all 1-D, 2-D, and 3-D subspaces. If 3-D subspaces are enough to produce comparable hit rate results, then the computational burden could be reduced to consider only 160 subspaces instead of 232. So subspace is a factor to be considered and we examine two

levels: 3-D subspaces only and all 1-D, 2-D, and 3-D subspaces. In fact, we decide on seven factors that could be further investigated to optimize LWY’s cell-based method. These are summarized in Table 3.2 and fully discussed below.

Factor A considers the number of cells into which a subspace is divided. Method LWY uses 64 cells always, for 1-D, 2-D, and 3-D subspaces. We consider two higher numbers of cells, 216 and 729. Is it beneficial to have a finer resolution of cells and is there a point beyond which no gains are realized? For 729 cells, 1-D subspaces are divided into 729 bins, 2-D subspaces are divided into 27 bins on each dimension ($27^2 = 729$), and 3-D subspaces are divided into 9 bins on each dimension ($9^3 = 729$). For 216 cells, things do not work out as cleanly. Notice that 216 cells can be obtained for 1-D and 3-D subspaces ($6^3 = 216$), but not for 2-D subspaces. We make a slight modification and use $15^2 = 225$ for 2-D subspaces. So Factor A, called Cell, has three levels: low (coded as -1), represents 64 cells in each subspace; center (coded as 0), represents 216 cells in each subspace; and high (coded as 1), represents 729 cells in each subspace.

Factor B, called amalgamating, considers whether or not amalgamating will be used in creating 2-D and 3-D cells from 1-D cells; see the subsection on Forming Cells. Factor B has three levels, where the high level (coded as 1) means amalgamating will be performed, as in method LWY. The other levels, low (coded -1) and center (coded 0), do not use amalgamating. For the low level, denoted “amalgamating=No1,” the first and last bins in each dimension contain the maximum of 1% of all compounds or the inverse of the number of bins in that dimension. In other words, the number of compounds in the extreme bins is $\max(1\%, (\# \text{ of bins in that dimension})^{-1} \times 100\%)$.

For example, for 3-D subspaces with 64 cells, each dimension has four bins and the first and fourth bins will be formed to contain 25% of the compounds in the training set. But for 2-D subspaces with 64 cells, the first and eighth bins will be formed to contain 12.5% of compounds in the training set. For the center level, denoted “amalgamating=No2,” the first and last bins in each dimension contain 1% of all compounds. For each dimension of 3-D subspaces with 64 cells, the second and third bin together will contain 98% of all compounds. It seems to gather too much data in the central cells and does not seem to be a very sensible approach, but we nevertheless compare it to the other levels.

Factor C considers whether or not all of the 1-D, 2-D, and 3-D subspaces will be used, as in LWY, or only the 3-D subspace. The latter level is considered low (coded -1) and the former is considered high (coded 1) in this two-level factor.

Factor D has two levels to represent the number of permutation runs used to determine the cut point for good cells: low (coded as -1) represents one permutation and high (coded as 1) represents five permutations. As discussed in subsection Identifying Good Cells, LWY uses one permutation with the justification that the large number of cells makes it reasonable. We expect no difference between the two levels.

Factor E has two levels to indicate whether or not preliminary selection of good cells is performed. As discussed in subsection Identifying Good Cells, preliminary good cells with two or more potent compounds are first identified and the really good cells are selected from among these cells. We doubt that preliminary selection is particularly effective. Low level (coded as -1) is no preliminary selection, and the

high level (coded as 1) indicates preliminary selection is performed.

Factor F has two levels concerning what to use as the cutoff from the permutation test. Method LWY requires that in order to find a cutoff value to separate good cells from the others, we need to implement the permutation process and find the 95th-percentile H_{L95} value. We propose to increase the cutoff value by replacing the 95th-percentile with the maximum value after permutation. This will lead to a more strict selection for good cells. Low level (coded as -1) indicates use of the 95th-percentile and the high level (coded as 1) indicates use of the maximum of 100th-percentile.

Factor G concerns the weight used in scoring compounds for predicted potency. In method LWY, weight H_{L95} is used to compute scores for compounds in the validation set. We propose using either $\sqrt{H_{L95}}$ or H_{L95}/\sqrt{n} to replace H_{L95} . Low level (coded as -1) uses H_{L95}/\sqrt{n} , center level (coded as 0) uses H_{L95} , and high level (coded as 1) uses $\sqrt{H_{L95}}$.

We are not willing to consider all $3^3 2^4 = 432$ combinations of alternative cell-based methods, but we very much want to know what factors might be important in obtaining an effective method. In order to evaluate the main effects of these factors, we use the design procedure ADX in SAS to find an optimal main effects design. It gives the fractional factorial design with 21 design points shown in Table 3.3. Method LWY is not in this set of 21 but can be expressed and added as $A = -1$, $B = 1$, $C = 1$, $D = -1$, $E = 1$, $F = -1$, $G = 0$.

The resulting 22 cell-based methods (including method LWY) were applied to

20 training sets. Lift charts were obtained and analyzed using standard analysis of variance procedures.

SELECT THE “BEST” CELL-BASED METHOD

Using only 22 observed cell-based methods, our goal is to determine which of the 432 possible cell-based methods will be most effective for our assay-library combination. Realizing that no method will be uniformly optimal for all assay-library combinations, we propose: (a) using all 22 cell-based methods to screen the training sets; (b) assessing the effectiveness of each screening method on each training set; (c) using analysis of variance to determine the best cell-based method for the training sets; (d) if the best cell-based method has not previously been observed, then apply it to the training sets to obtain confirmation of its effectiveness and to prepare for application to the validation sets; then (e) apply this best method to the validation set.

Our assessment of the effectiveness of screening is done by the lift chart. Given input data, cell-based methods create relational models that output a ranked ordering of compounds, where ranking is according to predicted potency. Testing of compounds is done according to this ordering until “enough” potents have been identified. The relational model developed from the cell-based method, if successful, will assign high ranks to potent compounds, thus making it likely that the cumulative percent of potent compounds actually found, relative to the number of compounds tested, will

be very high in early testing and will decrease to approach the average potency in the full data set. A lift chart plots the cumulative percent (hit rate) of potents found, relative to the number of compounds tested, as a function of the number of compounds tested. Obviously, the higher the hit rate, the better. Lift charts are directly related to accumulation curves^[13] and cumulative recalls^[14], as discussed below, but are more amenable to the types of statistical analyses conducted here. Specifically, assumptions of normality and homogeneity of variances are very reasonable for hit rates.

Kearsley et al.^[13] propose measures based on the accumulation curve. The accumulation curve plots the total number of potent compounds found ($A@n$) versus the total number of compounds tested (n). For example, $A@20$ is the number of potents found from testing the first 20 compounds. The lift chart at 20 tested compounds is simply $h_{20} = \frac{A@20}{20} \times 100\%$. In other words, the lift chart plots $(\frac{A@n}{n} \times 100)\%$ versus n . The ideal case for the accumulation curve is the line “ $A@n = n$,” which corresponds to a flat line “ $h_n = 100\%$ ” in the lift chart. The cumulative recall^[14] plots the cumulative percent of potents found, relative to the total number of potents in the entire set, versus the total number of compounds tested. In other words, the cumulative recall plots $(\frac{A@n}{C} \times 100\%)$ versus n , where C is a constant that represents the total number of potent compounds in the entire set. We see that the three types of plots are actually equivalent.

Figure 3.3 summarizes lift chart results from three cell-based methods, namely, LWY, Design 7, and Design 17. For each method and each training set, a lift chart is obtained. Because 20 training sets (that is, 20 replicates) are used, we average the 20 lift charts, and these are the plotted curves in Figure 3.3. It appears that

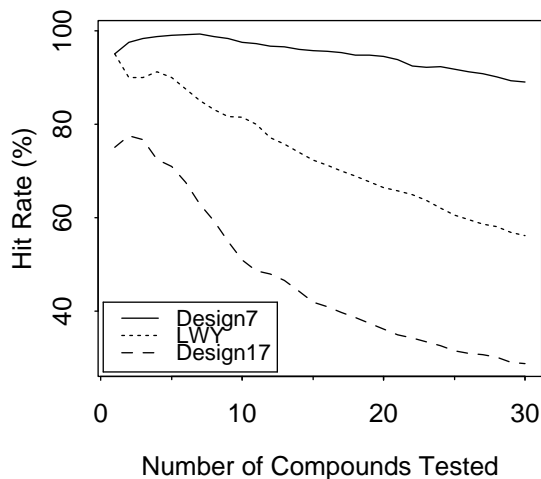


Figure 3.3: Lift Charts for Three Methods, Averaged over 20 Training Data Sets.

Design 7 is considerably better than LWY, while Design 17 is considerably worse. Figures 3.4 and 3.5 show the same results as Figure 3.3, but in the forms of accumulation curve and cumulative recall, respectively. Of course, our designed experiment allows us to test for differences in light of uncertainty. But we must first determine what is the “response” of interest.

Kearsley et al.^[13] consider a global enhancement based on A_{50} , the number of compounds that must be tested until half the potents are found. Initial enhancement, another proposal based on $A@N$ from a large database consisting of M compounds ($N=M/100$), is argued to be a more appropriate measure than A_{50} when only a small percent of a large database will be tested.

In obtaining our lift charts, we are interested in testing only a small percent of our training sets. Because the training sets contain only 4096 compounds, we consider

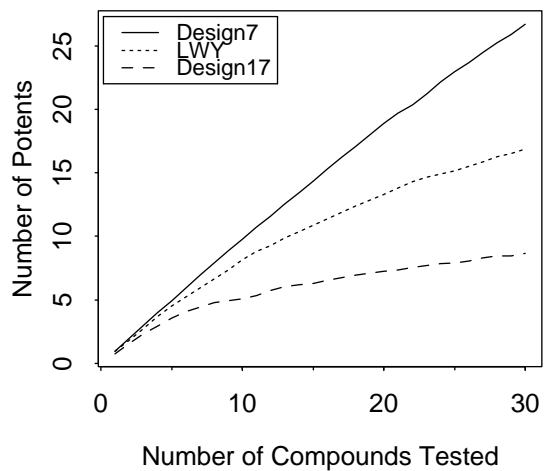


Figure 3.4: Accumulation Curves for Three Methods, Averaged over 20 Training Data Sets.

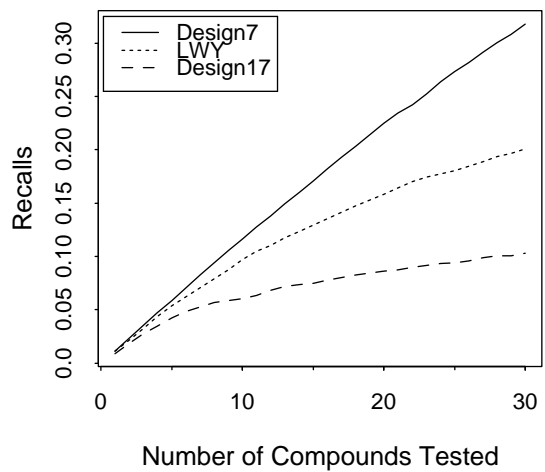


Figure 3.5: Cumulative Recalls for Three Methods, Averaged over 20 Training Data Sets.

h_{10} , h_{20} , and h_{30} (equivalent to $A@10$, $A@20$, and $A@30$) as measures of assessment. Analysis of variance on the three responses h_{10} , h_{20} , and h_{30} show slightly different results, as summarized in Table 3.4. Possible correlation between responses is not addressed.

Factor E is not important in creating cell-based methods for these responses and training sets. Especially when testing fewer compounds, Factors D and F are not as important as Factors A, B, C, and G. Multiple comparison testing suggests that hit rates for Design 7 are significantly higher than for LWY, which in turn are significantly higher than for Design 17 (see Table 3.5). Of the 22 cell-based methods listed in Table 3.3, the best method is Design 7, which is profiled with 729 cells, amalgamating No1, using 3-D subspaces only, a single permutation, no preliminary selection, permutation percentile 95%, and weight H_{L95}/\sqrt{n} . The worst method is Design 17, which has 64 cells, amalgamating No2, uses 3-D subspaces only, five permutations, no preliminary selection, permutation percentile 95%, and weight $\sqrt{H_{L95}}$.

Moreover, based partially on details presented in Table 3.6, it appears that level 1 of Factor A (729 cells), level -1 of Factor B (amalgamating No1), level 1 of Factor C (1-, 2-, 3-D subspaces), level -1 of Factor F (permutation percentile 95%), and level -1 of Factor G (weight H_{L95}/\sqrt{n}) are associated with better methods. In fact, four cell-based methods, as predicted best from the statistical analyses, are given in Table 3.7. The best among these 4 designs is denoted OPT and coded as (A,B,C,D,E,F,G) = (1,-1,1,-1,1,-1,-1). All four methods share the key features listed above, and they are not significantly different at the $\alpha = .05$ level. In fact, they differ only in levels for the mostly unimportant Factors D and E. The ignorable impact of

Factors D and E is evidenced by the largely overlapping 95% confidence intervals. None of these four methods are included in the set of 22 methods listed in Table 3.3, and any of them could be selected as the best cell-based method. However, these methods have also overlapped 95% confidence interval of Design 7 (see Table 3.7, with one exception) which means that they are not significantly better than Design 7. In our case, Design 7 is selected to make prediction as the best cell-based method. In the circumstances when the methods predicted best are significantly better than the observed best, they should be chosen to make prediction.

CONFIRMATORY RESULTS

The design of experiments approach to selecting a best cell-based method for this assay-library combination allows us to conclude that Design 7 can be used to give near-optimum performance. Applying Design 7 to the 20 training sets, the observed mean responses are 97.5%, 94.5% and 89.0% for h_{10} , h_{20} and h_{30} , respectively. We are now ready to investigate the enhancements offered by this cell-based method compared to random testing and method LWY. Enhancements are measured by applying the methods to the 20 large validation sets, each consisting of 25,716 compounds.

Following the approach of Kearsley et al.^[13], we use the measure of initial enhancement. Initial enhancement is the ratio of the actual A@250 for the method over the A@250 expected for random testing ($250 \times 520/25,716$ in our case), that is, $IE = A@250 \cdot \frac{25,716}{250 \times 520}$. These numbers are displayed in Table 3.8. Again, the benefits

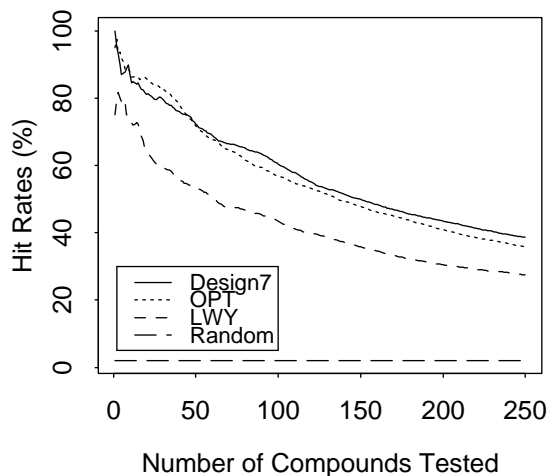


Figure 3.6: Lift Charts for LWY, Design 7 and OPT, Averaged over 20 Validation Data Sets.

of method Design 7 and OPT are clear, with much better performance relative to both random testing and method LWY.

Figure 3.6 summarizes the lift chart results of LWY, Design 7 and OPT, averaged over the 20 validation sets. It also shows the expected lift chart from random testing. The message is quite clear, both Design 7 and OPT offer substantial improvements and they make comparable predictions.

SUMMARY AND DISCUSSION

Cell-based methodology is successful at improving HTS efficiency. It is much more efficient than random testing and it has also been asserted to be more favorable than

recursive partitioning^{[10],[11]}. This cell-based approach combines the power of multiple good cells with the incorporation of molecular structural descriptors. The novel point of this method is the investigation of SARs by implicitly accounting for the effects of descriptors; no explicit functional forms are suggested. These results re-enforce the idea that compounds with similar structure will have comparable potency.

Gains in efficiency of cell-based methods compared to random testing are, however, sensitive to many factors of the process. It has been useful to implement an appropriate design of experiments for investigating the effects of these factors. From the results discussed above, we can see that careful selection of these factors can dramatically improve the screening efficiency. According to these results, the best method among the 22 methods in the main-effects design (including LWY) comes from Design 7.

In Design 7, only 3-D subspaces are required. This can lead to savings in computational time during the practical screening process. The gain can be quite substantial if the number of descriptors is increased. In this paper, only six of the more than 60 BCUT descriptors are used, and so the savings by considering only 3-D subspaces is that 72 fewer subspaces are needed; see the third and fourth section. On the other hand, if 10 BCUT descriptors are used to develop the relational model, then 200 fewer subspaces are needed if only 3-D subspaces are used.

These benefits can, and should, be investigated for a variety of applications. We make no claims that method Design 7 will be best for all assay-library combinations. In fact, we strongly believe that the best cell-based method for one application may

be horribly inefficient for another application and so it is imperative that the selection process be application-specific. Indeed, this is what we consider as our major contribution: the ability to determine, for each assay-library combination, the best cell-based method for screening the bulk of the library. The best method is entirely determined from a relatively small training set and no additional testing is required to assess many possible methods; the only additional cost is computing time.

Random selection of 20 pairs of training and validation sets offers realization of extreme behaviors, both positive and negative. For example, the lowest initial enhancement of Design 7 over random testing is 16.3 and comes from training-validation pair 12. On the other hand, the highest initial enhancement of Design 7 over random testing is 23.5 and comes from training-validation pair 4. Enhancements quoted in Table 3.8 are averages over all 20 training-validation pairs, just as the enhancements quoted on page 124 of Kearsley et al.^[13] are averages over 10 different probes within the same library. While our 20 training-validation pairs are not exactly equivalent to considering 20 separate assays, they do provide some information on the variability of the technique across applications.

There are, of course, areas in which our suggestions can be further investigated and possibly improved. It would be useful to have some sense of the relative importance of the BCUT numbers in predicting potency, but our current work makes no such attempts, with good reason. Cell-based methods isolate small regions in 1-, 2- or 3-D subspaces where potent compounds reside. Examination of the BCUTs that index the important dimensions tells us which BCUTs are important and the cell coordinates tell us the important ranges of those BCUT numbers. Examination of

the potent compounds in these good cells and the nature of the particular BCUTs can point to specific atoms and substructures. Our experience is that this is complex and the simple examination of the compounds generally illuminates the class of potent compounds in the sub-region. The nature of the BCUTs, atomic, lipophilic, etc. offers some information on the type of interatomic interactions involved. Having said all of this, the analysis here is largely aimed at getting high hit rates and interpretation of the molecular features is somewhat secondary. We should note that use of BCUTs for interpretable QSAR is considered problematic^[15], although recent work ^[16] is more positive.

Lam et al.^{[10],[11],[12]} spend considerable effort in examination of various ways to evaluate the importance of a cell. Two things are important, the level of potency or hit rate, and some measure of the reliability of the estimated hit rate. There are obviously many ways to weight these factors. One of the ways recommended by Lam et al.^{[10],[11],[12]} is to use a statistical lower bound, H_{L95} , on the proportion of potent compounds in the cell. Our work also uses H_{L95} in scoring cells, but others may find a different metric to be more desirable. Whatever the measure adopted, the cells can be ranked and unscreened compounds can be tested either in the order of their occurrence or a predetermined number of compounds can be used to determine how many of the good cells will be tested.

Cell-based methods can be used as the statistical method for selecting compounds in a sequential screening scheme^[17]. In sequential screening, an initial set of compounds is screened and the results are used to determine a predictive model. A cell-based method could be used to make these predictions and eventually a block of

compounds are selected and tested. So either a pre-selected number of compounds would be assayed or all the compounds that fell into good cells. If the logistics of compound handling and screening are very good and if the assay is very expensive, it would make sense to screen very small, incremental sets of compounds and stop when the screening objectives are met, namely, when either a fixed number of potent compounds or compound classes are identified. This sequential screening paradigm might be even more effective if the scoring statistic H_{L95} is replaced by a more dynamic index.

In this work, we consider seven factors and specific levels for each in order to determine the best cell-based method. Actually, factors could be expanded to include additional levels or even to introduce new factors. An additional factor could be used to investigate the use of equal-width bins versus equal-frequency bins versus hybrid (a combination of equal-width and equal-frequency) binning. Lam et al.^{[10],[11],[12]} report that they conducted extensive investigations of the effects of these various methods of binning and concluded that the hybrid approach was most effective. This finding could, however, be assay dependent and it may be a good idea to include the factor in the analysis. Also, one could consider adding more levels for some factors. For example, 4-D and 5-D subspaces could be included as levels of Factor C and Factor G might also include other choices for the weights.

Selection of a main-effects design is somewhat limiting because two-factor interactions are only estimable when some main effects are negligible. If one is willing to entertain more runs, that is, observe more cell-based methods, larger designs that accommodate estimation of interaction effects could also be pursued. For example,

assuming that two of the factors in the NCI data application were negligible (which is true of Factors E and D), a design that allows estimation of the five remaining main effects and all 10 two-factor interactions contains 44 design points.

The design of experiments approach presented here is general enough to be applicable to many assay-library combinations, yet specific enough to result in significant improvements for a particular assay-library application.

Acknowledgment

This work was supported in part by a grant from the National Science Foundation, award number DMS-0072809. We also gratefully acknowledge Lap-Hing Raymond Lam at GlaxoSmithKline for providing the NCI data, early access to his dissertation, and other help.

Table 3.1: *Potency Proportion Quantiles of Non-Empty Cells*

Quantile	Min.	1%	5%	25%	50%	75%	95%	99%	Max.
Proportion	0.01	0.01	0.01	0.02	0.03	0.05	0.18	0.50	1.00

Table 3.2: *Factors for Creating Alternative Cell-Based Methods*

Factor	Low	Center	High	Description			
A	-1	0	1	Cell	64	216	729
B	-1	0	1	Amalgamating	No1	No2	Yes
C	-1		1	Subspaces	3-D		1-, 2-, 3-D
D	-1		1	Permutation	1		5
E	-1		1	Pre. Selection	No		Yes
F	-1		1	Percentile	95%		100%
G	-1	0	1	Weights	H_{L95}/\sqrt{n}	H_{L95}	$\sqrt{H_{L95}}$

Table 3.3: *Optimal Design Points for Creating Alternative Cell-Based Methods*

Design	A	B	C	D	E	F	G
1	1	1	-1	1	1	-1	1
2	1	1	-1	-1	1	1	1
3	1	1	1	-1	-1	-1	0
4	1	0	1	1	1	1	0
5	1	0	-1	1	-1	1	0
6	1	0	1	1	1	-1	-1
7	1	-1	-1	-1	-1	-1	-1
8	0	1	1	-1	-1	-1	0
9	0	1	1	1	1	1	-1
10	0	0	-1	-1	1	-1	1
11	0	0	-1	-1	-1	1	-1
12	0	-1	1	1	-1	1	1
13	0	-1	-1	1	1	1	0
14	-1	1	-1	1	-1	-1	0
15	-1	1	-1	1	-1	1	-1
16	-1	1	-1	-1	1	1	-1
17	-1	0	-1	1	-1	-1	1
18	-1	0	1	-1	1	1	0
19	-1	-1	1	-1	-1	1	1
20	-1	-1	-1	-1	1	-1	0
21	-1	-1	1	1	1	-1	-1
LWY	-1	1	1	-1	1	-1	0

Table 3.4: Summary of ANOVA Results for h_{10} , h_{20} , and h_{30} , the Hit Rate Responses at Tests of 10, 20, and 30 Compounds in the Training Set. (F is the value of F statistics)

Source	h_{10}		h_{20}		h_{30}	
	F	p -value	F	p -value	F	p -value
Model Effect	37.3	$\leq .0001$	47.7	$\leq .0001$	62.6	$\leq .0001$
Main Effect A	141.1	$\leq .0001$	178.2	$\leq .0001$	229.1	$\leq .0001$
Main Effect B	26.9	$\leq .0001$	22.8	$\leq .0001$	33.1	$\leq .0001$
Main Effect C	36.0	$\leq .0001$	22.6	$\leq .0001$	21.2	$\leq .0001$
Main Effect D	0.5	0.4981	3.1	0.0767	14.0	0.0002
Main Effect E	0.4	0.5060	0.5	0.4757	0.0	0.9668
Main Effect F	2.4	0.1215	35.0	$\leq .0001$	65.5	$\leq .0001$
Main Effect G	11.0	$\leq .0001$	19.6	$\leq .0001$	18.7	$\leq .0001$

Table 3.5: Comparisons Between Design 7, LWY, and Design 17 for h_{10} , h_{20} , and h_{30} , the Hit Rate Responses at Tests of 10, 20, and 30 Compounds in the Training Set. (Experimentwise levels of significance are all set at 5%)

	h_{10}	h_{20}	h_{30}
Difference between Design 7 and LWY	16.00	28.00	32.83
Difference between LWY and Design 17	30.50	30.25	27.34
LSD Min. Significant difference	7.58	7.59	6.53
HSD Min. Significant difference	13.95	13.97	12.03
Bonferroni Min. Significant difference	14.39	14.41	12.41

Table 3.6: Mean Hit Rates, at All Levels of All Factors for h_{10} , h_{20} , and h_{30} , the Hit Rate Responses at Tests of 10, 20, 30 Compounds in the Training Set

Factor	h_{10}			h_{20}			h_{30}		
	-1	0	1	-1	0	1	-1	0	1
A	70.4	86.3	94.1	53.8	67.7	81.0	45.0	56.9	70.9
B	89.0	77.3	84.4	72.5	61.6	68.5	63.1	51.9	57.8
C	79.9		87.3	64.6		70.5	55.2		60.0
D	83.2		84.0	68.6		66.4	59.6		55.6
E	83.2		84.0	67.1		68.0	57.6		57.6
F	84.5		82.6	71.2		63.9	61.8		53.4
G	86.9	84.4	79.5	72.8	67.3	62.5	62.1	57.3	53.4

Table 3.7: *Four Best Cell-Based Methods Predicted from the Statistical Analyses*

Factors							95% Confidence Intervals					
A	B	C	D	E	F	G	h_{10}		h_{20}		h_{30}	
1	-1	1	-1	1	-1	-1	[103.2,	111.8]	[95.1,	103.6]	[85.9,	93.1]
1	-1	1	-1	-1	-1	-1	[102.2,	111.1]	[94.1,	102.9]	[85.9,	93.2]
1	-1	1	1	1	-1	-1	[104.2,	112.5]	[93.1,	101.3]	[82.1,	89.0]
1	-1	1	1	-1	-1	-1	[103.2,	111.8]	[92.1,	100.5]	[82.1,	89.1]
Design 7							[94.9,	103.6]	[88.3,	97.0]	[81.1,	88.3]

Table 3.8: *Global and Initial Enhancement for Methods LWY and OPT*

Method	A@250	IE
Design 7	90	17.7
OPT	90	17.6
LWY	68	13.3

Bibliography

- [1] McFarland, J. W.; Gans, D. J. On the Significance of Clusters in the Graphical Display of Structure-Activity Data. *Journal of Medicinal Chemistry* **1986**, *29*, 505-514.
- [2] Carhart, R. E.; Smith, D. H.; Venkataraghavan R. Atom Pairs as Molecular Features in Structure-activity studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
- [3] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.
- [4] Hawkins, D. M.; Young, S. S.; Rusinko A. Analysis of a large structure-activity data set using recursive partitioning. *Quantitative Structure-Activity Relationship* **1997**, *16*, 296-302.
- [5] Young, S. S.; Hawkins, D. M. Using Recursive Partitioning to Analyze a Large SAR Data Set. *SAR and QSAR in Environmental Research* **1998**, *8*, 183-193.
- [6] Rusinko A.; Farmen M. W.; Lambert C. G.; Brown P. L.; Young S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *38*, 1017-1026.
- [7] Zhu, L.; Hughes-Oliver, J. M.; Young, S. S. Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics* **2001**, *57*, 922-930.
- [8] Pearlman R. S.; Smith K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28-35.

- [9] Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225-227.
- [10] Lam, R. L. H. Design and Analysis of Large Chemical Databases for Drug Discovery *University of Waterloo* Ph.D. Dissertation **2001**.
- [11] Lam, R. L. H.; Welch W. J.; Young, S. S. Cell-Based Analysis for Large Chemical Databases. *Technometrics* **2002**, submitted.
- [12] Lam, R. L. H.; Welch W. J.; Young, S. S. Uniform Coverage Designs for Molecule Selection. *Technometrics*, **2002**, *44*, 99-109.
- [13] Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118-127.
- [14] Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *Journal of Molecular Graphics and Modelling* **2000**, *18*, 343-357.
- [15] Pearlman R. S.; Smith K. M. Novel Software tools for chemical diversity. *Perspectives in Drug Discovery and Design* **1998**, *9-11*, 339-353.
- [16] Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39(1)*, 11-20.
- [17] Engels M. F.; Venkatarangan P. Smart screening: Approaches to efficient HTS. *Current Opinion in Drug Discovery & Development* **2001**, *4(3)*, 275-283.

Chapter 4

Statistical Models for Decoding Pools of Chemical Compounds in the Presence of Compound Interactions and Dilution

4.1 Literature Review

As described in Chapter 1, modern discovery of a new drug begins with testing hundreds of thousands of compounds with respect to one or more biological assays. As the proportion of activity is usually very small, testing all compounds is not cost-effective or desirable. For example, the assay used to test a single compound may cost from a few cents to several dollars depending upon the complexity of the assay. It means screening one million samples for one assay could easily go up to \$1 million. For this reason, pooling experiments are now often conducted in many pharmaceutical companies with the purpose of time and cost effectiveness.

Typically, robotic liquid handling systems are used to form pools. After solid

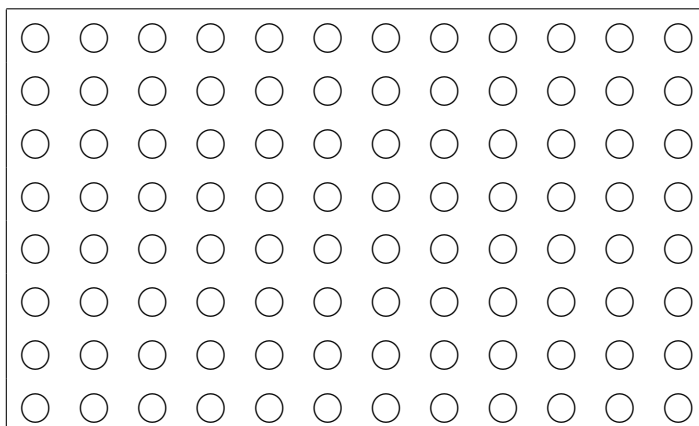


Figure 4.1: A typical 8×12 plate used in pooling experiments.

samples of chemical compounds are dissolved into solvent, the handling system will be programmed to place samples on a plate; a typical plate of 8×12 is shown in Figure 4.1. Compounds are pooled by combining all samples in a given row or a column and pools are stored in separate plates to prepare for testing.

Pooling along the rows or along the columns, but not both, is called one-way pooling. Obviously, each individual compound appears in only one pool in one-way pooling and independence could be assumed among potencies of different pools. Two-way pooling requires that pools are formed in both directions of rows and columns, thus leading to dependence between pool potencies; each compound appears in two different pools. Three way pooling or even multi-way pooling can also be considered, depending on the particular circumstance of the assay, but it should be noted that in multi-way pooling the total number of tests on pools can increase to the extent that the pooling experiment is no longer cost-effective.

Potency values with respect to a particular biological assay are usually given on a continuous scale. Dichotomizing compounds as “potent” or “junk” using a threshold value on this continuous scale is often easily done by a chemist or biologist. This same threshold for individual potencies is not, however, often appropriate for pools. Even if compounds are placed in pools at the same concentration at which they

are individually tested, dilution can occur to make the potency of a highly potent compound that is pooled with many junk compounds yield a pool potency well below the maximum individual potency. This effect is exacerbated if individual compounds are placed in pools at weaker concentrations to allow the pool to have concentration of organic materials comparable to the level of individual testing. The net effect is that specification of threshold for labeling pools as potent or not is rather speculative.

Using a selected threshold, pools with potencies greater than this threshold will be labeled as potent and given discrete potency value 1; pools with potencies less than this threshold will be labeled as not potent and given discrete potency value 0. After completion of this classification process, pooling experiments for compound screening will follow the pattern of group testing. The following is a brief literature review on group testing and pooling experiments for chemical compounds.

The group testing problem originated with Dorfman (1943). In order to detect blood samples infected by syphilis for donors called for induction in World War Two, Dorfman proposed this group testing idea to improve screening efficiency. From then on, group testing has been attracting statistical attention and has also been obtaining successful application in various areas.

An important issue in group testing is estimation of the proportion of individuals that possess a trait of interest. Group testing estimators can reduce the mean square error (MSE) of estimates of proportions in the cases where the true proportion does not exceed $2/3$, although benefits depend on using an appropriate group size (Swallow 1985 and 1987; Sobel and Elashoff 1975; Thompson 1962). Because properties of group testing estimators are very sensitive to group size (Hughes-Oliver and Swallow 1992; Swallow 1985; Thompson 1962), Hughes-Oliver and Swallow (1994) suggest several methods to overcome this deficiency. These methods include using different group sizes; two stage estimation; adaptively adjusting group size from time to time; and performing Bayesian analysis.

Recently, group testing has been applied successfully in the estimation of disease risk, especially HIV risk (Gastwirth and Hammick 1989; Tu, Litvak, and Pagano,

1994 and 1995; Hung and Swallow, 1999 and 2000). While considering imperfect testing with False Positive Rate (FPR) and False Negative Rate (FNR), Tu et al. (1994, 1995) also show that fewer tests with bigger pool size achieve greater efficiency on the same number of samples. Brookmeyer (1999) estimated HIV incidence rates using a multistage pooling estimator that is a generalization of Thompson's (1962) estimator.

Group testing can be a very good approach for cost effectiveness as well as for estimation of proportions. Along with the estimation of FPR and FNR using pooled data on disease prevalence, Kline et al. (1989) and Cahoon-Young et al. (1989) cut the original testing cost by 80%.

The literature usually assumes a small prevalence proportion for the trait of interest. However, Hammick and Gastwirth (1994) investigate group testing strategies to handle circumstances with much higher prevalence rate. Hughes-Oliver and Rosenberger (2002) also demonstrate the effectiveness of a pooling strategy for estimating a 0.25 prevalence of chlamydia. Moreover, Hughes-Oliver and Rosenberger extend the adaptive or multistage approach to simultaneously estimate the prevalence of the three inter-related diseases HIV, chlamydia and hepatitis B.

Applying the ideas of group testing to pools of chemical compounds for assessing biological potency introduces many difficulties. Biological activity of chemical compounds is usually measured on a continuous scale, rather than as presence or absence of a trait, as is typical in regular group testing. A threshold is needed for determining which pools are potent, but this is not an easy task given the possible existence of multiple mechanisms within a pool.

Dilution effect within pools is inevitable. The concentration of a compound is changed after pooling and this can impact the pool's potency. In the presence of strong dilution effects, a good pool with at least one potent compound in it will be hard to discriminate from a bad pool with no potent compounds in it. There is an early paper that discussed the optimal pool size under dilution effect (Hwang, 1976), where the dilution effect is included in the model as the probability that a defective

group is correctly identified by the test. Hung and Swallow (1999) proposed two types of models for dilution effect and consequences robustness of group testing for proportion estimating problems.

Another possible mechanism in a pool is synergistic effect. Synergism occurs when several individually non-potent compounds yield a potent pool result when put together in a single pool. Xie et al. (2001) investigate group testing problems under presence of synergism.

A major concern about pooling chemical compounds is the blocker effect. A pool containing at least one potent compound is called a good pool. A pool is labeled as potent after a threshold is used to assign a discrete potency value. Unfortunately, a good pool may be not potent due to interactions between the compounds in this pool. Specifically, a negative pool may actually be a good pool containing a potent compound that is altered by a blocker effect. This is more common with pooling chemical compounds than with group testing of blood samples. Phatarfod and Sudbury (1994) give a blocker effect example that one sample might mask another. They also reported that the two-way pooling is less susceptible to the false negative problem caused by blocker effect.

Langfeldt et al. (1997) and Xie et al. (2001) also discuss optimal pooling strategies with the presence of blocker effect. Langfeldt et al. (1997) try to minimize the test cost and reduce the number of missed positive individuals by assuming a known common blocker rate. Xie et al. (2001) assume an unknown common blocker rate to derive probabilities of potency for individual compounds.

Zhu, Hughes-Oliver, and Young (2001) also consider blocker effect in their probabilistic models for pooling decoding. But, unlike many previous researchers, they use chemical structural descriptors to further classify compounds into several classes. Compounds in each class are assumed to follow a trinomial distribution to be potent, blocker (the whole pool will not be potent once this pool contains at least one blocker), or neither. After class probabilities of potency are estimated, potency predictions of individual compounds can be used to assess the performance of various models and

methods.

Group testing with explanatory variables is also discussed by Farrington (1992), Hung and Swallow (2000), and Vansteelandt, Goetghebeur, and Verstraeten (2000). Farrington (1992) and Vansteelandt et al. (2000) use link functions to relate probability that a trait is present in a pool with covariates of all individuals in that pool. Generalized linear models are proposed to give inference on individual risk based on the pooled data. In special cases, standard software can be used to implement this analysis. Hung and Swallow (2000) explore the potential advantages of group testing to some hypothesis-testing problems by investigating the relationship between individual risk and a covariate.

In this Chapter, we propose two models based on pooled data of chemical compounds with the consideration of blockers and dilution effect. This is an extension of Zhu, Hughes-Oliver, and Young (2001) in the sense that a more realistic blocker effect model is used. The blocker effect happens as an interaction of two compounds and hence is determined by both of them. The probability of having a blocker effect in a specific pool depends on what kinds of compound pairs are found in the pool. Considering covariates by using chemical structural classes (see Section 4.2), this probability is determined by what kinds of pairs of classes are in a pool. A pool containing compounds all from one class is unlikely to suffer from blocker effects, but may appear potent due to additive effects. This is an assumption more reasonable than that in ZHY models where blocker effect is considered as an intrinsic character of a single compound and a single compound can sometimes act as a blocker which will mask the potencies of all other potent compounds (see Section 4.3). The new models are introduced in detail in Section 4.4 and applied to real data in Section 4.5. Application to a large pooling experiment was planned but, for several reasons explained in Section 4.7, this plan was not executed.

In Section 4.6, we also investigate the effect of various thresholds for pool potency on the performance of different models. Consequently, we propose a method that is successful yet does not need a user-supplied specific pool threshold. This method is

robust in the sense that it can identify potent classes with high precision even when nothing is known about the “correct” threshold. It is also a good solution when we have no information about the dilution effect.

4.2 Atom Pairs as Covariate Class

As mentioned in Section 4.1, we build models using chemical structural descriptors. Based on the belief that chemical structure is highly related to potency of compounds, many kinds of descriptors have been computed from chemical structural features. Among such structural features, atom pairs, topological torsions, and fragments (Carhart, Smith, and Venkataraghavan, 1985; Nilakantan et al., 1987) have been useful in recursive partitioning analysis and/or pooling methods for the purpose of improving HTS efficiency (Hawkins, Young, and Rusinko, 1997; Young and Hawkins, 1998; Rusinko et al., 1999; Zhu, Hughes-Oliver, and Young, 2001). We consider atom pairs as covariates in our pooling models. Atom pairs account for all pairs of non-hydrogen atoms and the minimal topological distances between the pairs in a compound (Carhart, Smith, and Venkataraghavan, 1985). The presence and absence of atom pairs are represented as 1’s and 0’s, so that a compound can be described by a bit string where each element of the string corresponds to an atom pair. For a compound with n non-hydrogen atoms, the number of atom pairs of that compound is on the order of n^2 . A collection of compounds usually has many thousands of atom pairs, thus making it impractical to individually consider all atom pairs appearing in the data. We need some preliminary selection of atom pairs. Suppose we have already determined a relatively small subset of potentially important atom pairs with the belief that compounds containing a subset of these atom pairs will share common properties. Using combinations from these atom pairs, we form L atom pair covariate classes.

Preliminary selection of atom pairs and rules for other covariates under different circumstances can be identified by subject-matter experts or through an analysis of

preliminary data. For example, Dorfman (1943) noted differing incidence of syphilis depending on the home state of the blood donor. Thus, in the case of blood testing, demographic variables can be reasonably used as covariates. In our case of testing compounds for biological activity, atom pairs are selected to capture the structural features of chemical compounds, and recursive partitioning analysis can play the role as a preliminary selector.

Hawkins, Young, and Rusinko (1997) define chemical classes by the terminal nodes from a recursive partitioning tree. Each node will split into two nodes by presence and absence of a certain atom pair which makes the most significant split. Rules come from the recursive partitioning tree and can then be used to define chemical classes. Each compound belongs to a unique class by following a specific rule. Each active terminal node is used to define a chemical class and all inactive nodes are combined to define a single inactive class. The strategy would be to screen a few thousand compounds as a small fraction of a large data set and analyze this small data by recursive partitioning.

4.3 ZHY Models and Assumptions

Zhu, Hughes-Oliver, and Young (2001) propose two models (called ZHY models) for decoding pools of chemical compounds. ZHY models consider blocker effects by assuming the existence of a special compound called “blocker” that inhibits the potency of other compounds in the same pool. It is assumed that each compound is classified in one of three ways: potent, blocker, or neither. The probabilistic model assumes that each individual compound follows independent trinomial distribution:

$$(W_{ij}, V_{ij}, 1 - W_{ij} - V_{ij}) \sim \text{Multinomial}(1, p_{ij}, f_{ij}, 1 - p_{ij} - f_{ij}), \quad i = 1, \dots, n; \quad j = 1, \dots, k,$$

where n is the number of pools, k is the number of compounds within each pool, W_{ij} and V_{ij} are indicator functions denoting the j th compound in the i th pool being potent or a blocker, respectively, and $p_{ij} = \Pr(W_{ij} = 1)$ and $f_{ij} = \Pr(V_{ij} = 1)$. A

blocker cannot be identified unless it is tested with one or more active compounds, so V_{ij} is not directly observable.

It is also assumed that a pool is potent if there are no blockers and at least one potent individual compounds within it; a pool is negative either when there is at least one blocker, or when there are no potent individuals in this pool. Therefore,

$$Y_i = I(\text{pool } i \text{ is potent}) = I\left(\sum_{j=1}^k W_{ij} > 0, \sum_{j=1}^k V_{ij} = 0\right).$$

As we have already had L atom pair classes as covariates, each compound will belong to one of these classes and it is assumed that all compounds in the same class will have the same probability to be potent or a blocker. Consequently, we will focus on the probability that a given covariate class is a potent or blocker class. In other words, we consider $p = (p_{*1}, \dots, p_{*L})$ and $f = (f_{*1}, \dots, f_{*L})$, where $p_{*l} = \Pr(\text{a compound in covariate class } l \text{ is potent})$ and $f_{*l} = \Pr(\text{a compound in covariate class } l \text{ is a blocker})$, rather than p_{ij} and f_{ij} for all $i = 1, \dots, n$. and $j = 1, \dots, k$. We also know the information of counts, s_{il} , the number of compounds in pool i that come from covariate class l .

Two separate models are proposed for the pooled data, depending on the level of retesting of potent pools. ZHY Model 1 considers the case where limited resources do not allow retesting of individuals in potent pools. Only $Y = (Y_1, Y_2, \dots, Y_n)'$, the observed pool potency, is available for use in Model 1. ZHY Model 2 is proposed for the case where all compounds in potent pools are retested. In other words, ZHY Model 2 is based on Dorfman Classical Decoding (see Chapter 1). In that case, both pool potencies Y and the individual potencies in potent pools, a portion of $W = (W_{11}, \dots, W_{1k}, \dots, W_{n1}, \dots, W_{nk})'$, are observed and available for use in Model 2.

Using the corresponding data in different models, we can find maximum likelihood estimators for these parameters. These estimates may then be used to determine “important covariate classes,” according to whether the corresponding estimate of p_{*l} is large relative to other estimates. All individual compounds within these important

covariate classes will then be selected for individual testing to determine potency. The hit rate can be calculated as the number of potent compounds found divided by the total number of tests performed. The numerator is total of the number of tests to obtain the pooled data and that to complete the individual testing.

In ZHY Model 1, pool potencies Y and the covariate class counts s_{il} are the observed data. It can be shown that Y_i , the potency of the i th pool, is distributed as Bernoulli $\{\psi_i = \prod_{l=1}^L (1 - f_{*l})^{s_{il}} - \prod_{l=1}^L (1 - p_{*l} - f_{*l})^{s_{il}}\}$. Let $\theta = (p, f)$ be the vector of all parameters. The likelihood of the n independent pools is $L(\theta) = \prod_{i=1}^n \psi_i^{y_i} (1 - \psi_i)^{(1-y_i)}$. The s_{il} 's satisfy the relationship $\sum_{l=1}^L s_{il} = k$.

For ZHY Model 2, the observed data is now: $Y = (Y_1, Y_2, \dots, Y_n)'$, $W = (W_{ij} : W_{ij} \text{ when } Y_i = 1)$. Let $W_{*i} = (W_{*i1}, \dots, W_{*iL})$, where W_{*il} is the number of compounds tested potent in pool i and covariate class l . The joint density of the observed data (Y_i, W_{*i}) is

$$f(y_i, w_{*i}) = \begin{cases} \phi_{1i} = \prod_{l=1}^L p_{*l}^{w_{*il}} (1 - p_{*l} - f_{*l})^{(s_{il} - w_{*il})} & \text{if } y_i = 1, \sum_{l=1}^L w_{*il} > 0 \\ 1 - \psi_i & \text{if } y_i = 0, \end{cases}$$

for $i = 1, \dots, n$. The likelihood will then be $L(\theta) = \prod_{i=1}^n \phi_{1i}^{y_i} (1 - \psi_i)^{(1-y_i)}$.

4.4 Model Extensions

For our statistical models, assumptions about blocker effects are different from these in ZHY models. Blocking in a given pool happens when two compounds in the pool interact with each other. One compound may alter the binding site of another compound, thus causing the potent compound to lose its original potency. From this point of view, the probability that a pool experiences blocking depends on compound pairs rather than on a single compound. We propose an extension of ZHY models by considering this new kind of blocker effect.

Suppose compounds follow independent binomial distributions for indicating po-

tency, that is,

$$W_{ij} \sim \text{Binomial}(1, p_{ij}), \quad i = 1, \dots, n, j = 1, \dots, k,$$

where W_{ij} , p_{ij} , n , and k are as defined in Section 4.3. Let f_{ij} denote the probability that blocking happens between the i th and j th compounds, where $i, j = 1, 2, \dots, nk$. Under these assumptions, a pool is potent if this pool contains at least one potent compound (these pools are called good pools) and no blocker interaction happens between any pairs of two compounds in this pool; a pool is negative either if this pool is not good or it is a good pool that contains two compounds with a blocking interaction.

For L atom pair covariate classes, the potency parameter is $p = (p_{*1}, \dots, p_{*L})$, the same as in ZHY models, but the blocker parameter is different. Let $f = (f_{*1}, \dots, f_{*B})$, where $B = \binom{L}{2}$ is the total number of pairs of different classes and $f_{*b} = \Pr(\text{Blocker effect happens for pair } b)$. Hence, if two compounds come from different covariate classes where this class pair is number b among all class pairs, then the probability that blocking happens between the two compounds will be f_{*b} ; if two compounds come from the same class, then no blocker effect is assumed.

For extension of ZHY Model 1, which assumes no retesting of individuals, the pool potency Y_i is distributed as Bernoulli $\{\psi_i\}$, where

$$\psi_i = \{1 - B_i(f)\} \left\{ 1 - \prod_{l=1}^L (1 - p_{*l})^{s_{il}} \right\},$$

and $B_i(f) = 1 - \prod_{k=k_1}^{k_{n_i}} (1 - f_{*k})$ is the probability that pool i has a blocker effect. Class pairs k_1, \dots, k_{n_i} are all the class pairs present in pool i . The likelihood of the n independent pools is

$$L(\theta) = \prod_{i=1}^n \psi_i^{y_i} (1 - \psi_i)^{(1-y_i)}.$$

Let C be a subset of classes in which all classes appear in all pools at the same time and $D = \{1, 2, \dots, L\}/C$ which is all the remaining classes other than classes in C . Then for any $d \in D$, the blocking effect parameters between d and $c_1 \in C$, between d and $c_2 \in C, \dots$, and between d and $c_u \in C$ are non-identifiable, where u is total number of classes in C .

For extension of ZHY Model 2, which assumes retesting all individuals in potent pools, the observed data are $Y = (Y_1, Y_2, \dots, Y_n)'$, and $W = (W_{ij} : W_{ij} \text{ when } Y_i = 1)$. The likelihood function is:

$$L(\theta) = \prod_{i=1}^n \phi_{1i}^{y_i} (1 - \psi_i)^{(1-y_i)},$$

where

$$\phi_{1i} = (1 - B_i(f)) \prod_{l=1}^L p_{*l}^{w_{*il}} (1 - p_{*l})^{(s_{il} - w_{*il})}$$

4.5 Application to Real Data and Simulation Study

4.5.1 The Data

In a pooling experiment at GlaxoSmithKline, 1000 chemical compounds are selected from storage where selection is based on Burden numbers (Burden, 1989) in order to reduce the additive effect in pools. Compounds with similar Burden numbers are more likely to be topologically similar, so assignment of compounds to pools is to achieve diversity of Burden numbers within a pool; this will avoid testing at twice the concentration of that covariate class.

Compounds are grouped in pools of ten, resulting in 100 pools. Pools are formed by three-way pooling, thus leading to a total of 300 tests on pools. Imagine 1000 points representing 1000 compounds to form a cube of $10 \times 10 \times 10$. We can produce 100 pools from top to bottom, another 100 pools from left to right, and another 100 pools again from back to front. Obviously, the 300 pools are not independent in potency, but independence is a reasonable assumption within each set of 100 potencies. We will separately analyze the three different data sets each obtained from one way of the three-way pooling. Individual potencies, covariate class information and all 100 pool potencies are available for each set of pools (sets of pools are called DATA1, DATA2, and DATA3, respectively) and will be used for model evaluation.

Biological potency of a compound is measured as the inhibition ability to the assay.

Table 4.1: *Distribution of Potent Compounds in the Covariate Classes*

Covariate classes	1	2	3	4	5	6	7	Sum
Potent compounds	2	1	2	16	6	4	9	40
Total compounds	509	5	11	412	27	5	31	1000
Proportion	.0039	.0200	.1818	.0388	.2222	.8000	.2903	.0400

Inhibition is calculated by an equation relating the target compound to a reference compound; the result is a measurement on a continuous scale. Theoretically, percent inhibition should be between 0 and 100, but in practice, under some special binding situations leading to unusual results, the computation can give potencies less than 0 or above 100. A single compound is labeled as potent or not depending on its inhibition percentage being greater or less than a given threshold. In our case, for the assay we use, the potency threshold for individual compounds is 60. There are 40 compounds with potency value greater than 60 and will be said to be potent. The percentage of potent compounds is a low 4%, and is typical for screening experiments of chemical compounds.

By a recursive partitioning analysis, we select seven covariate classes based on eight atom pairs (see Section 4.2). The number of potent compounds in these classes is shown in Table 4.1.

Pools are also tested using the same assay as for the individual compounds and potencies are given in the same scale for percentage inhibition. Figure 4.2 shows the distribution of pool potencies for each 100 pools. When applying potency threshold of 60 for individual compounds to pool potency, there are only 4 potent pools out of 31 good pools for data set DATA1, 4 out of 33 for DATA2 and 5 out of 31 for DATA3. Recall that a good pool contains at least one potent compound irrespective of whether or not a blocker effect occurs in that pool.

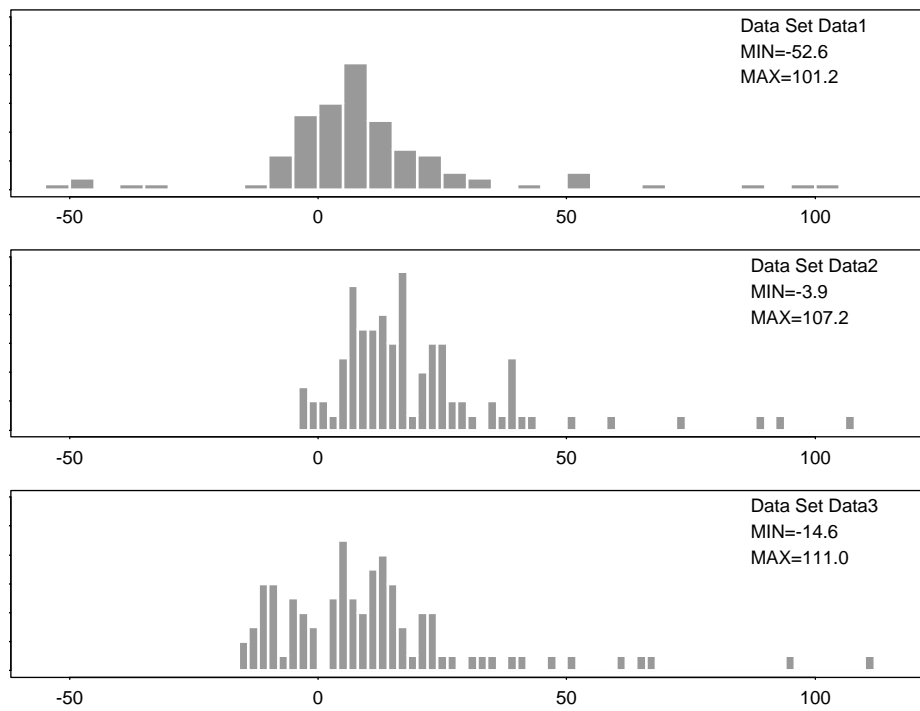


Figure 4.2: Histograms of pool potency

Table 4.2: *Maximum likelihood estimates of p from data DATA1 using four models, where important classes are labeled with an * beside the appropriate \hat{p}*

Classes	Model 1	Model 1 ext.	Model 2	Model 2 ext.
1	0.00	0.00	0.00	0.00
2	0.77*	0.56*	1.00*	1.00*
3	0.00	0.22*	0.00	0.00
4	0.00	0.00	0.00	0.00
5	0.27*	0.10*	0.63*	0.63*
6	0.16*	0.17*	1.00*	1.00*
7	0.00	0.00	0.22*	0.22*

4.5.2 Application and Assessment of Models

We are now in the position to estimate all parameters in models in order to obtain hit rates. Note that ZHY Model 1 and Extension Model 1 use only the pool potencies as input data and assume no retesting of the individual compounds. ZHY Model 2 and Extension Model 2 use both pool potencies and the individual potencies in potent pools, which means that before the model estimation process begins, the compounds in potent pools have already been individually tested and their potencies are available for direct use. Calculation of hit rates for the two types of models occur in different ways.

The classification methods using Model 1 and Model 2 are shown in flow charts in Figure 4.3 (Zhu, Hughes-Oliver, and Young, 2001).

The MLEs of parameters p and f for all four models are obtained using SAS IML's nonlinear optimization function NLPCG. Estimates of p for all four models from data DATA1 are given in Table 4.2. In order to find global rather than local maximum likelihood, many starting points are chosen to cover the entire parameter space.

After determining \hat{p} , we evaluate covariate classes in terms of relative importance. First, we rank \hat{p}_i from high to low, say $\hat{p}_{i_1} \geq \hat{p}_{i_2} \geq \dots \geq \hat{p}_{i_7}$, where i_1, i_2, \dots, i_7 are a permutation of $1, 2, \dots, 7$. Potency probabilities are compared to all other potency probabilities for determining important classes. Specifically, we find m such that $\frac{\sum_{j=1}^m \hat{p}_{(j)}}{\sum_{j=1}^7 \hat{p}_j} \geq 95\%$ but $\frac{\sum_{j=1}^{m-1} \hat{p}_{(j)}}{\sum_{j=1}^7 \hat{p}_j} < 95\%$, where $\hat{p}_{(j)}$ is the j th ordered value within \hat{p} .

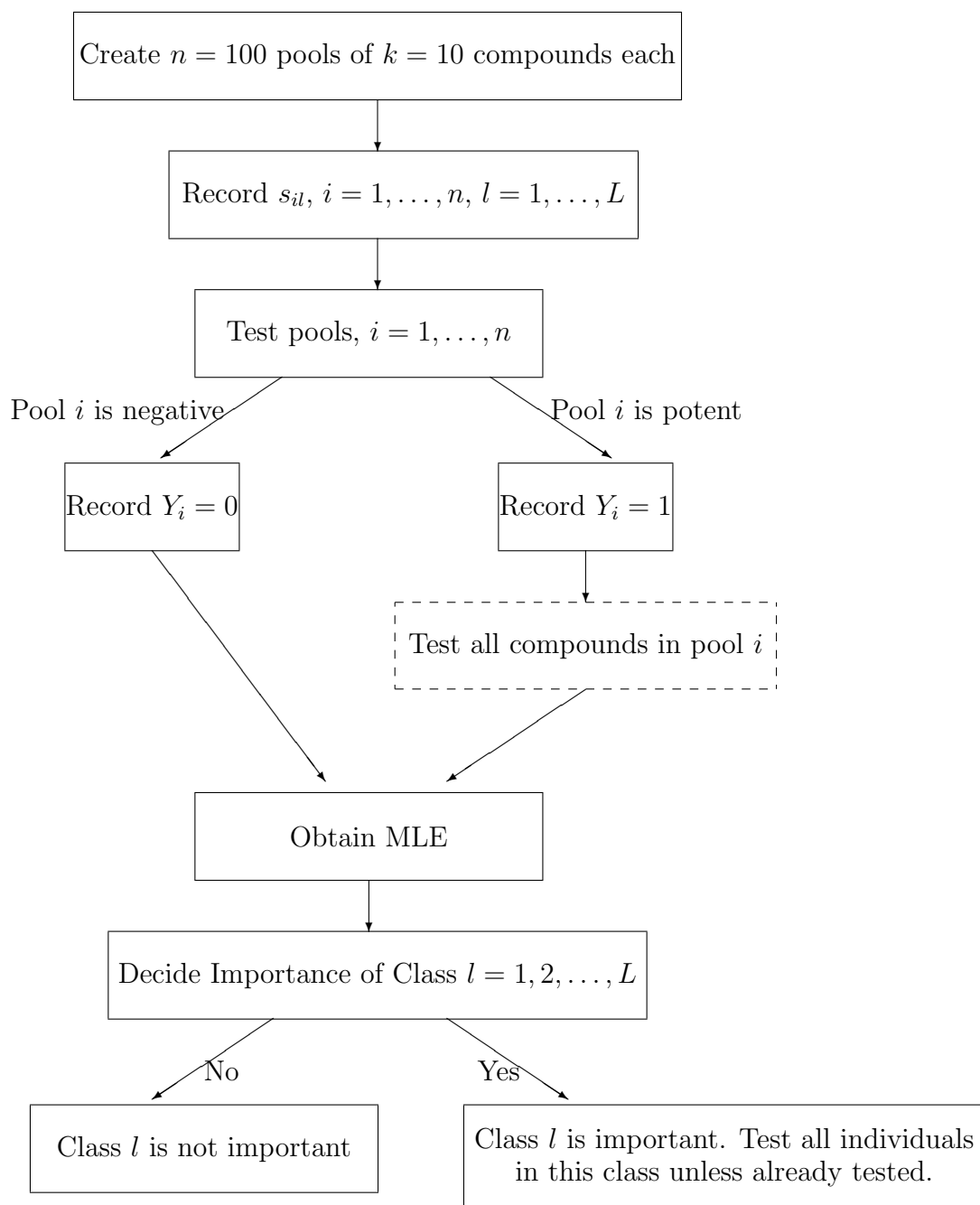


Figure 4.3: Flow chart of classification using Model 1 or Model 2

In other words, class i is important if \hat{p}_i is within the top 95% of the sum of all \hat{p}_i s. For data DATA1, ZHY Model 1 identifies covariate classes 2, 5, 6, as important or potent classes. Extension Model 1 identifies 2, 3, 5, 6 as potent classes, while both ZHY Model 2 and Extension Model 2 identify 2, 5, 6, 7 as potent classes.

How does one evaluate model performance? Because our research goal is to find as many potent compounds as possible at the least cost, we use hit rate as the criterion to assess model performance. Hit rate is the number of potent compounds identified divided by the total number of tests performed. The higher the hit rate, the more potent compounds are identified using the same number of tests, or on the other hand, fewer of tests are needed to identify a pre-specified number of potent compounds or hits. After important potent classes are determined, all compounds in these classes are predicted to be potent and should undergo individual screening. Our approach is actually a classification method.

Because Model 1's and Model 2's use different data resources, hit rate calculations have different forms. Including individual testing (which tests compounds one at a time without pooling) and Dorfman Classical Decoding method, we have 6 different methods. We will also consider a seventh method. If we use only the individual data from potent pools, we can also estimate p , except for cases where there are some classes from which no compounds appear in the potent pools at all. For these special classes, we will estimate their \hat{p}_i s with 0. Using the estimates given, potent classes are identified and hit rates can be computed. This method is called Simplified method. Formulas for computing hit rates are given in Table 4.3, where $Z_l = I(\text{class } l \text{ is important})$ (Zhu, Hughes-Oliver, and Young, 2001). ZHY Model 1 and Extension Model 1 have the same formula for computing hit rates, which are given under label Model 1. ZHY Model 2, Extension Model 2 and Simplified method also have the same formula, given under label Model 2.

Hit rate results for data sets DATA1, DATA2, and DATA3 are given in Table 4.4. For DATA1, individual testing identifies 40 compounds using 1000 tests, so the hit rate is 0.04. Classical Decoding needs 100 tests for pool potencies plus 40 for individuals

Table 4.3: *Hit Rate Formulas for Four Methods*

Method	Hit Rate
Test Individuals	$\frac{\sum_{i=1}^n \sum_{l=1}^L W_{*il}}{nk}$
Classical Decoding	$\frac{\sum_{i=1}^n \sum_{l=1}^L Y_i W_{*il}}{n+k \sum_{i=1}^n Y_i}$
Model 1	$\frac{\sum_{i=1}^n \sum_{l=1}^L Z_l W_{*il}}{n + \sum_{i=1}^n \sum_{l=1}^L Z_l s_{il}}$
Model 2	$\frac{\sum_{i=1}^n \sum_{l=1}^L (Y_i + Z_l - Y_i Z_l) W_{*il}}{n+k \sum_{i=1}^n Y_i + \sum_{i=1}^n \sum_{l=1}^L Z_l s_{il} - \sum_{i=1}^n \sum_{l=1}^L Y_i s_{il} Z_l}$

Table 4.4: *Hit rates for 7 methods, where ind. for individuals, ext. for extension*

DATA	Test ind.	Classical	Model 1	Model 1 ext.	Model 2	Model 2 ext.	Simplified
1	$\frac{40}{1000}=0.040$	$\frac{6}{140}=0.043$	$\frac{11}{137}=0.080$	$\frac{13}{148}=0.088$	$\frac{20}{200}=0.100$	$\frac{20}{200}=0.100$	$\frac{20}{200}=0.100$
2	$\frac{40}{1000}=0.040$	$\frac{140}{9}=0.036$	$\frac{137}{11}=0.080$	$\frac{148}{13}=0.088$	$\frac{182}{207}=0.071$	$\frac{182}{207}=0.071$	$\frac{182}{207}=0.071$
3	$\frac{40}{1000}=0.040$	$\frac{9}{150}=0.060$	$\frac{10}{132}=0.076$	$\frac{11}{137}=0.080$	$\frac{23}{217}=0.106$	$\frac{23}{217}=0.106$	$\frac{23}{217}=0.106$

in 4 potent pools and it identifies 6 potent compounds. The hit rate is 0.043. ZHY Model 1 deems classes 2, 5, and 6 as potent classes. Individual tests are obtained only for compounds in these three classes, thus identifying 11 potent compounds while needing 100 tests for pools plus 37 tests for the three classes. Hit rate for ZHY Model 1 is 0.08. ZHY Model 2 needs 200 tests in total (100 for pools, 40 for retesting individuals within potent pools, 68 for individuals in classes 2, 5, 6 and 7, minus 8 individuals in classes 2, 5, 6, 7 that are already individually tested because they are in the potent pools) to identify 20 potent compounds for a hit rate of 0.10. Hit rates for Extension Model 1 and Extension Model 2 can be computed in the same way as ZHY Model 1 and ZHY Model 2, respectively.

It is not surprising to see that model-based approaches perform better than individual testing and Dorfman Classical Decoding. One reason is that the model-based methods all assume the existence of blocker effects. As we know, for data set DATA1, there are 31 good pools and only 4 potent pools when using threshold=60 for pool potency. For this case, the apparent blocker rate is extremely high, resulting in the fact that no consideration of blocker effect is a disaster. This is also the reason that in DATA2 Classical Decoding is not better than individual testing although we expect the former should behave better. Another reason for improved results of model-based

methods is the use of chemical structural descriptors. Classical Decoding can only identify as potent these compounds that have been individually tested. On the other hand, the model-based approaches can predict compounds to be potent solely based on structural information, even if they are tested in pools labeled not potent or possibly never tested, neither in pools or individually.

Extension Model 1 is doing better than ZHY Model 1 with respect to both the hit rates and the number of potent compounds identified. However, Extension Model 2 is performing the same as ZHY Model 2. We obtain insight after checking the data set in detail. We believe that Model 2’s estimation results appear to be dominated by the information from the individual data in the potent pools since this part of the data is used in model estimations. We can see that, for each of the three data sets, Simplified method identifies the same potent classes as ZHY Model 2 and Extension Model 2, and obtain the same hit rates (see Table 4.4). However, in the simulation study later, we will find Extension Model 2 is slightly better than ZHY Model 2 although they are mostly comparable.

4.5.3 Simulation Study

We perform a small simulation study to compare models. Chemical structural descriptors is as observed in data DATA1. The vector of potency probabilities is $p=(0.004, 0.200, 0.180, 0.040, 0.220, 0.800, 0.290)$, as observed in DATA1. Blocker effect is simulated at a relatively high level that sets blocking probabilities to be $f = 27/31$. This value is determined from data DATA1, where 31 good pools are observed but only 4 are labeled potent using a pool threshold of 60.

Sixty simulation replicates resulted in mean hit rates as given in Table 4.5. Variability of these hit rates is also given in Table 4.5. Actual hit rates are given in Figure 4.4. From this plot, it is clear that individual testing and Classical Decoding give comparable results in the presence of high blocker effect. All model-based approaches are more variable but also have higher mean hit rates than individual testing

Table 4.5: Means and standard deviations of hit rates in simulation

	Test ind.	Classical	Model 1	Model 1 ext.	Model 2	Model 2 ext.	Simplified
Mean	0.0308	0.0276	0.0743	0.0978	0.0816	0.0902	0.0749
Std. Dev.	0.0045	0.0100	0.0341	0.0295	0.0268	0.0248	0.0305

and Classical Decoding. Extension Model 1 is better than ZHY Model 1 and Extension Model 2 is slightly better than ZHY Model 2. Comparing Extension Model 1 and Extension Model 2, the former behaves much better many times, but it is highly variable. Overall, they two extension models have comparable hit rates results.

4.6 Threshold for Pool Potency

In the model assessment discussed in Section 4.5, we use 60 as the threshold for pool potency. In other words, when a pool has potency value greater than 60, it is classified as potent and its discrete potency is assigned as 1; otherwise, it is classified as a junk pool and given discrete potency 0. Using these binary pool “potencies” and chemical structural information for all individual compounds, model estimates are obtained, important covariate classes are determined, and hit rates are calculated in order to evaluate model performance.

However, as mentioned in Section 4.5, a pool threshold of 60 leads to identifying only 4 potent pools and incorrectly labeling 27 good pools as “junk.” Considering the effect of dilution, this threshold of 60 is not appropriate for pool potency, since concentration is reduced when a compound mixes with other compounds in the pool. For example, suppose a compound C has potency value of 60 and there are 9 other compounds with potency 0 in the same pool. For the pool potency, we reasonably expect it will be much lower than 60. So, this pool turns out to be not potent when using a threshold of 60, although it is actually a good pool. We propose to address this dilution effect by lowering the threshold for pool potency.

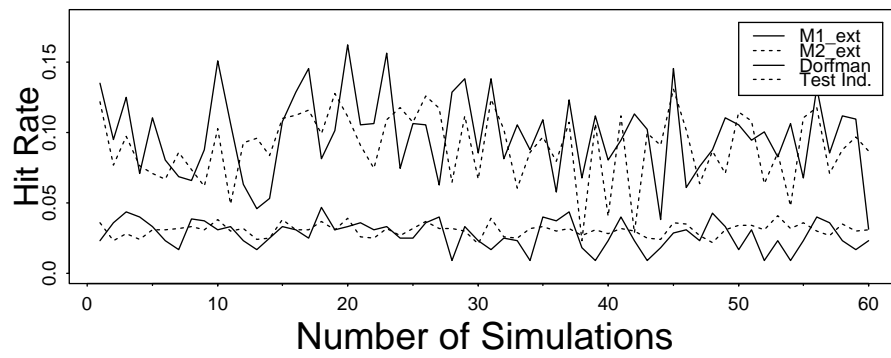
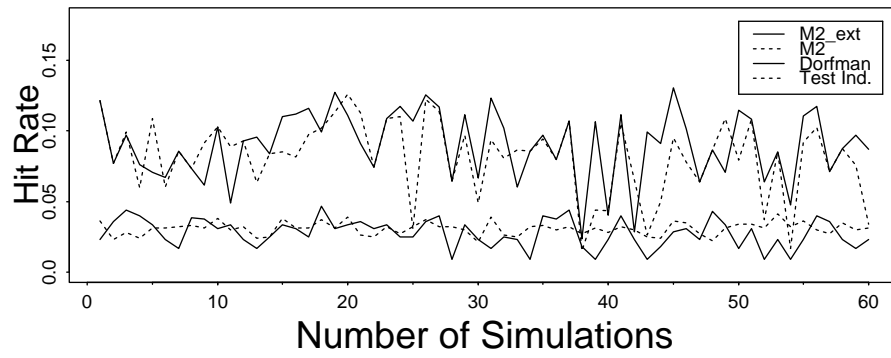
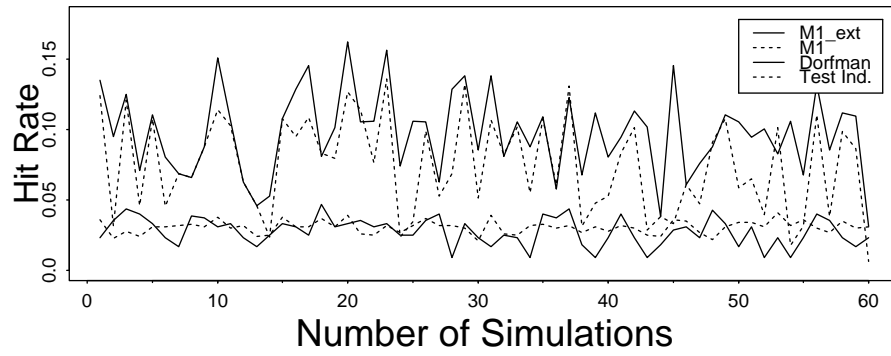


Figure 4.4: Simulated hit rates with high blocker rate

4.6.1 Hit Rates vs Various Threshold

In order to investigate model performance when using different pool thresholds, we ran Extension Model 1 and Extension Model 2 using many pool potency thresholds for data sets DATA1, DATA2, DATA3. In other words, we repeat the process in Section 4.5 for many different thresholds other than 60, but using only Extension Model 1 and Extension Model 2. We do not consider ZHY Model 1 and ZHY Model 2 since they did not perform as well as the extension models in Section 4.5. We also compute hit rates for the Classical Decoding method.

Specifically, we consider thresholds in the range of 10%-100% of the individual potency threshold. Considering the heaviest dilution effect, we expect that the potency for a good pool should not be less than this minimum value. Also, there is nothing to suggest that good pools should have potency greater than the individual threshold in order to be potent. Thus, for our example, all thresholds greater than 6 and less than 60 are chosen to be considered. Resulting hit rates are plotted as a function of different thresholds in Figure 4.5.

We can see that at high thresholds, Extension Model 2 behaves better than Extension Model 1 and Dorfman Classical Decoding, and is also less variable than Extension Model 1. But when thresholds become smaller, Extension Model 1 performs the best and becomes more stable than at high thresholds. However, when the thresholds become too small, Extension Model 1 becomes highly variable again. For most threshold values, Extension Model 1 has the best properties.

These plots may not be feasible in practice, since individual potencies are needed to determine hit rates after model estimation for each threshold. Suppose we have a total of m different thresholds denoted by C_i , $i = 1, \dots, m$. For each threshold C_i , let PC_i be the set of covariate classes deemed important. When these PC_i 's are not nested sequentially, individual testing to evaluate hit rate for threshold C_i may not be useful for assessing hit rate for C_{i+1} , thus leading to a waste of valuable resources. This places limits on direct practical application. However, these plots

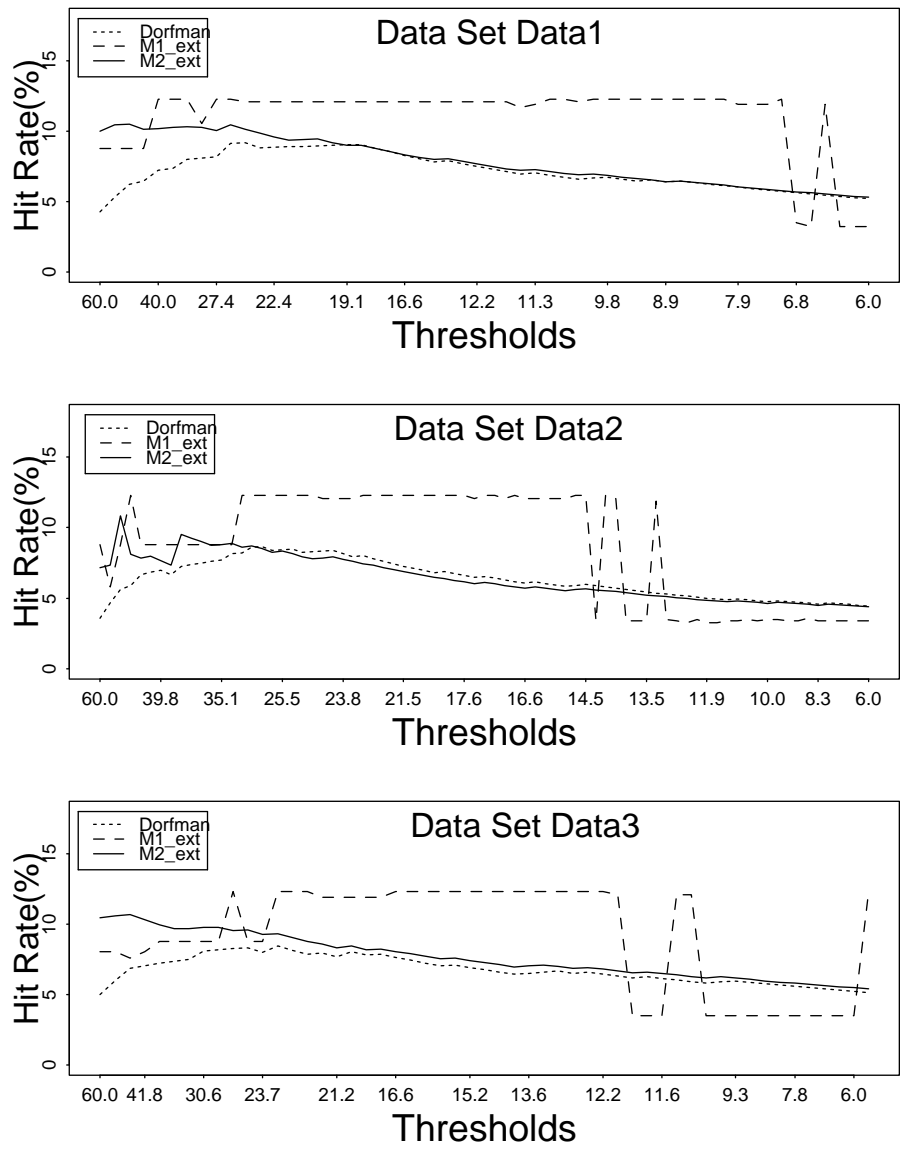


Figure 4.5: Hit rates vs various thresholds for DATA1, DATA2, and DATA3

Table 4.6: *Using Pool Potency Threshold of 18.9*

		Labeled Status of Pool		
		potent	not potent	
Actual Status of Pool	good	21/6.8	10/24.2	31
	junk	1/15.2	68/53.8	69
		22	78	100

provide an intriguing big-picture view of the impact of pool potency thresholds on model performance.

4.6.2 Is It Possible to Find an Optimal Threshold?

A clear pattern is observed in Figure 4.5 for Classical Decoding hit rate curves: increasing, reaching maximum, and then decreasing. We can use this feature to identify an optimal threshold for pool potency, which is desirable for future use. In order to do this, we do not have to experiment with all possible thresholds. When we see the trend of decreasing hit rates, we stop. Actually, we make the rule as stopping after we meet two continuous times of going down in hit rates. Using this rule, we find that the optimal threshold is 18.9 for DATA1, 24.0 for DATA2 and 23.7 for DATA3.

Are these thresholds good? Suppose all individual potencies are available. We can then construct a χ^2 test to evaluate these thresholds. For data set DATA1, using threshold 18.9, 22 pools have potencies greater 18.9 and then are labeled as potent. Among these labeled potent pools, 21 are good pools and only one is a junk pool. The contingency table given in Table 4.6 allows us to test the relation between actually being a good pool and being labeled potent.

Entries in Table 4.6 are in the form observed/expected, where expected=(row total) \times (column total)/100. For this contingency table, we can compute the p -value for a test of independence. In fact, we can compute p -values for all possible thresholds C_i , $i = 1, \dots, m$. The threshold with the minimum p -value is the ideal threshold we desire, which is 18.9 for DATA1, 24 for DATA2, 24 for DATA3. It shows that the above threshold obtained by Classical Decoding method is very precise.

Table 4.7: *Using optimal threshold, hit rates results for 7 methods, where ind. for individuals, ext. for extension*

	Test ind.	Classical	Model 1	Model 1 ext.	Model 2	Model 2 ext.	Simplified
DATA1	$\frac{40}{1000}=0.040$	$\frac{29}{320}=0.091$	$\frac{21}{174}=0.121$	$\frac{21}{174}=0.121$	$\frac{33}{366}=0.090$	$\frac{33}{366}=0.090$	$\frac{33}{366}=0.090$
DATA2	$\frac{40}{1000}=0.040$	$\frac{31}{370}=0.084$	$\frac{21}{174}=0.121$	$\frac{22}{179}=0.123$	$\frac{33}{366}=0.079$	$\frac{33}{366}=0.079$	$\frac{33}{366}=0.079$
DATA3	$\frac{40}{1000}=0.040$	$\frac{20}{240}=0.083$	$\frac{21}{174}=0.121$	$\frac{22}{179}=0.123$	$\frac{29}{311}=0.093$	$\frac{29}{311}=0.093$	$\frac{29}{311}=0.093$

Table 4.8: *Means and standard deviations of hit rates in simulation with small blocker rate*

	Test ind.	Classical	Model 1	Model 1 ext.	Model 2	Model 2 ext.	Simplified
Mean	0.0403	0.0876	0.1221	0.1241	0.0867	0.0874	0.0814
Std. Dev.	0.0059	0.0071	0.0276	0.0246	0.0106	0.0087	0.0150

4.6.3 A Close Look at Small Thresholds

We take a close look at the performance of several models using the optimal thresholds: 18.9 for DATA1, 24 for DATA2, and 24 for DATA3. The results are shown in Table 4.7.

These real data results show that ZHY models and model extensions are quite comparable when using optimal threshold. We also conduct a small simulation experiment similar to the one described in Section 4.5, except that the blocker rate of $27/31$ is replaced by the rate of $6/31$. Extension models have higher hit rates and are less variable than ZHY models (see Table 4.8). On the other hand, Extension Model 1 is obviously preferable to Extension Model 2 even though the former is more variable; see Figure 4.6.

4.6.4 Methodology Robust to Choice of Threshold for Pool Potency

From the discussion above, it is clear that model performance is very much dependent on threshold. In some cases, we may know the optimal threshold very well, either by previous experiments, or just by experience. Then, Extension Model 1 is the best choice. It has the highest hit rate among all these methods. But, in many other situations, we may have no idea regarding the appropriate threshold. In these cases, we need some new methods.

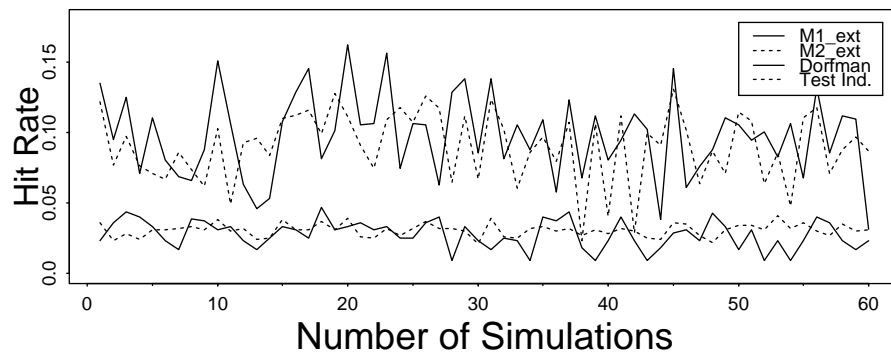
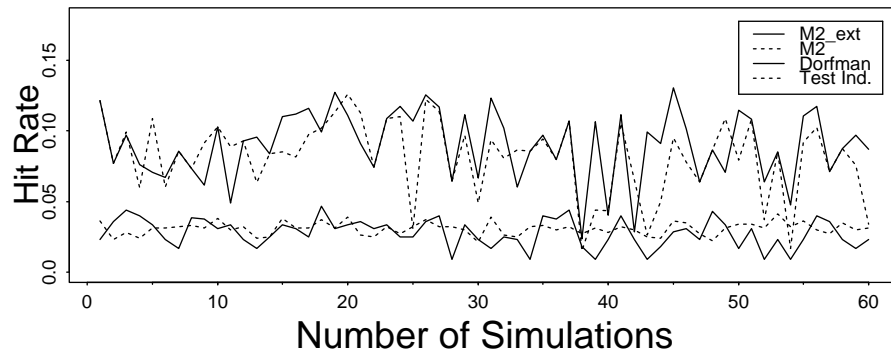
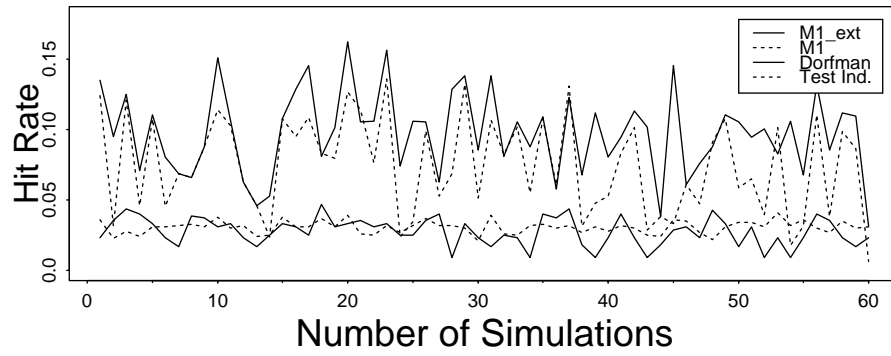


Figure 4.6: Simulated hit rates with smaller blocker rate

Table 4.9: *Frequency being identified as potent classes for each class*

Class	1	2	3	4	5	6	7	total
DATA1	0	33	44	5	53	54	50	54
DATA2	3	48	71	25	75	77	64	77
DATA3	0	49	44	15	53	53	42	53

We propose a method to benefit from Model 1’s property: the only data needed is the set of pool potencies and covariate class information for all compounds. No individual potency data is required. So, we run Extension Model 1 for all possible thresholds (see Subsection 4.6.1). This will lead to identification of potent classes for each threshold. We then count the frequency of being identified as a potent class for each class, and rank these frequencies from high to low. Ranked frequencies provide an order of individually testing compounds based entirely on chemical structural descriptors of the compounds. Individual testing terminates after obtaining a pre-specified number of potent compounds or observing a sharp down-turn in hit rates. Frequencies for the real data are given in Table 4.9. Fifty-four thresholds were investigated for data set DATA1, leading to a testing order of covariate classes 6, 5, 7, 3, 2; classes 4 and 1 were almost never selected. The testing hit rate is 0.123. Similar results are obtained for sets DATA2 and DATA3.

In practice, many factors may be at work in a pool, such as dilution or blocking. Various complicated situations will make it very hard to determine a good threshold for pool potency. The dangers of using an improper pool threshold are two-fold. On one hand, a different model may be more appropriate for that threshold, and on the other hand, the optimal threshold may yield a much higher hit rate. The methodology presented in this section is robust to the choice of threshold and is thus useful for practical decoding of pooling experiments.

4.7 A Big Pooling Data Set

At GlaxoSmithKline, a pooling experiment grouped half million compounds into about 50,000 pools of size 11. Potency measures the inhibition ability of a compound, and is calculated by an equation relating the target compound to a reference compound. GlaxoSmithKline (GSK) scientists specify 75 as the threshold for labeling a compound to be potent, that is, only compounds with potency greater than 75 are regarded as potent. Using this threshold, about 0.4% of these half million compounds are potent.

Because of missing values and other reasons, there are 43,659 pools with pool size 11. Using pool threshold 84.36 suggested by GSK scientists, 1,254 pools with pool size 11 are labeled as potent. The proportion of potent pools is about 2.87%.

If the compounds within these 1,254 pools are individually retested, it is found that there are 180 containing at least one potent compound. Hence, only 14.35% of pools labeled as potent are really potent. This percentage is low, indicating that many pools are potent by effect of either additivity or synergism. Actually, in the 43,659 pools with pool size 11, 42,352 contain no potent compounds, but $1,254 - 180 = 1,074$ of these pools are labeled as potent. When potencies are known for all individual compounds, the additivity or synergism effect can be computed to be about 2.54%. Because of the existence of this non-negligible additivity or synergism effect, the strategy of Dorfman testing (Dorfman, 1943), which retests all individual compounds in potent pools, is not successful in this data set as in other cases (hit rate of Dorfman testing is about 0.42%, compared to 0.4% for random testing).

Based on details provided by the screeners, the problems is more likely due to additivity rather than synergism. Compounds are placed in storage in the order in which they are created, thus inducing a strong degree of compound similarity over time and storage location. Unlike the pooling experiment from Section 4.5, no attempt was made to avoid additivity in this pooling design. Pools were not even randomly created. In fact, pools were created from the 10×12 plates pulled from

storage. The consequence is that highly similar compounds were pooled, thus leading to many potent pools that do not contain potent individuals. The strain from the large number of labeled potent pools affected the quality of the pooling experiment in yet another way. The specified threshold for pool potency (84.36) was determined entirely from testing capacity; no science was used.

On the other hand, it can also be seen that in the pools with size 11, there are a total of 1,307 pools containing at least one potent compounds. However, in these good pools, only 180 are labeled as potent. The blocking effect can be computed to be 13.8%. The blocking rate is higher than is usually expected, suggesting that the pool threshold is likely to be inappropriate. This data set may have the same problem of determining a good threshold as the small GSK data in Section 4.5. The methods proposed in Section 4.6 could be applied here to find a better threshold. Unfortunately, because the models developed in this chapter do not consider synergism, analysis of this data, as well as the pool threshold issue, will be left as future work.

However, caution should be taken when analyzing this data set because the measurement variability for potency is relatively high. In this experiment, compounds within the potent pools are individually retested twice to decide potency. Including the original individual measurements (remember that all half million compounds have been tested once for potency before being put into pools — this part of data is called “individual” in Table 4.10 and Figure 4.7), all compounds in potent pools are measured in triplicate. Results of correlation analysis for the duplicate potencies and individual measurements are given in Table 4.10. The corresponding scatter plots are shown in Figure 4.7. Correlations of these multiple measurements are relatively low, suggesting that assay variability is relatively high.

Table 4.10: *Correlation Coefficient Matrix of Duplicate Potencies for Compounds in Potent Pools*

	Duplicate 1	Duplicate 2	Individual
Duplicate 1	1.000	0.458	0.209
Duplicate 2	0.458	1.000	0.198
Individual	0.209	0.198	1.000

Scatter Plots of Duplicate Measurements

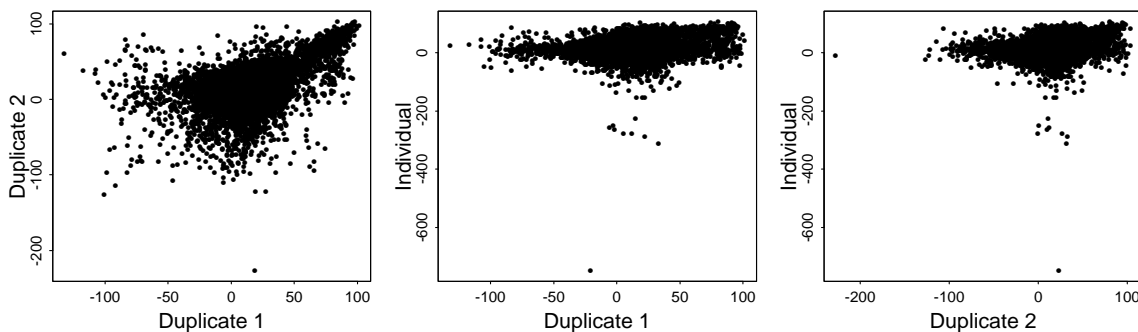


Figure 4.7: Scatter Plot of Duplicate Potencies for Compounds in Potent Pools

Bibliography

- [1] Brookmeyer, R. Analysis of Multistage Pooling Studies of Biological Specimens for Estimating Disease Incidence and Prevalence. *Biometrics* **1999**, *55*, 608-612.
- [2] Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225-227.
- [3] Cahoon-Young, B.; Chandler, A.; Livermore, T.; Gaudino, J.; Benjamin, R. Sensitivity and Specificity of Pooled versus Individual Sera in a Human Immunodeficiency Virus Antibody Prevalence Study. *Journal of Clinical Microbiology* **1989**, *27*, 1893-1895.
- [4] Carhart, R. E.; Smith, D. H.; Venkataraghavan R. Atom Pairs as Molecular Features in Structure-activity studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
- [5] R. Dorfman The Detection of Defective Members of Large Populations. *Annals of Mathematical Statistics* **1943**, *14*, 436-440.
- [6] Farrington, C. Estimating Prevalence by Group Testing Using Generalized Linear Models. *Statistics in Medicine* **1992**, *11*, 1591-1597.
- [7] Gastwirth, J. L.; Hammick, P. A. Estimation of the Prevalence of a Rare Disease, Preserving the Anonymity of the Subjects by Group Testing: Application to Estimating the Prevalence of Aids Antibodies in Blood Donors. *Journal of Statistical Planning and Inference* **1989**, *22*, 15-27.
- [8] Hammick, P. A.; Gastwirth, J. L. Group Testing for Sensitive Characteristics:

- Extension to Higher Prevalence Levels. *Internat. Statist. Assoc.* **1994**, *62*, 319-331.
- [9] Hawkins, D. M.; Young, S. S.; Rusinko A. Analysis of a Large Structure-activity Data Set Using Recursive Partitioning. *Quantitative Structure-Activity Relationship* **1997**, *16*, 296-302.
- [10] Hughes-Oliver, J. M.; Swallow, W. H. Choosing the Group Size for Group Testing to Estimate a Proportion. *Institute of Statistics Mimeograph Series No. 2209, North Carolina State University* **1992**.
- [11] Hughes-Oliver, J. M.; Swallow, W. H. A Two-Stage Adaptive Group-Testing Procedure for Estimating Small Proportions. *Journal of the American Statistical Association* **1994**, *89*, 982-993.
- [12] Hughes-Oliver, J. M.; Rosenberger, W. F. Efficient Estimation of the Prevalence of Multiply Rare Traits. *Biometrika* **2000**, *87*, 315-327.
- [13] Hung, M. Advantages of Group Testing with Stratified or Regression Data, and Its Robustness in Estimation of Proportions. *Ph.D dissertation*, **1993**.
- [14] Hung, M.; Swallow, W. H. Robustness of Group Testing in the Estimation of Proportions. *Boimetrics* **1999**, *55*, 231-237.
- [15] Hung, M.; Swallow, W. H. Use of Binomial Group Testing in Tests of Hypotheses for Classification or Quantitative Covariables. *Boimetrics* **2000**, *56*, 204-212.
- [16] Hwang, F. K. Group Tesing with a Dilution Effect. *Biometrika* **1976**, *63*, 671-673.
- [17] Kline, R. L.; Brothers, T. A.; Brookmeyer, R.; Zeger, S.; Quinn, T. C. Evaluation of Human Immunodeficiency Virus Seroprevalence in Population Surveys Using Pooled Data. *Journal of Clinical Microbiology* **1989**, *27*, 1449-1452.
- [18] Lam, R. L. H.; Welch W. J.; Young, S. S. Cell-Based Analysis for Large Chemical Databases. *Technometrics* **2002**, submitted.
- [19] Lam, R. L. H.; Welch W. J.; Young, S. S. Uniform Coverage Designs for Molecule Selection. *Technometrics* **2002**, accepted.

- [20] Langfeldt, S. A.; Hughes-Oliver, H. M.; Ghosh, S.; Young, S. S. Optimal Group Testing in the Presence of Blockers. *Institute of Statistics Mimeograph Series No. 2297*, **1997**
- [21] McFarland, J. W.; Gans, D. J. On the Significance of Clusters in the Graphical Display of Structure-Activity Data. *Journal of Medicinal Chemistry* **1986**, *29*, 505-514.
- [22] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.
- [23] Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28-35.
- [24] Phatarfod, R. M.; Sudbury, A. The Use of a Square-Array Scheme in Blood Testing. *Statistics in Medicine*, **1994**, *13*, 2337-2343.
- [25] Rusinko, A.; Farnen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *38*, 1017-1026.
- [26] Sobel, M.; Elashoff, R. M. Group Testing With a New Goal, Estimation. *Biometrika*, **1975**, *62*, 181-193.
- [27] Swallow, W. H. Group Testing for Estimating Infection Rates and Probability of Disease Transmission. *Phytopathology* **1985**, *75*, 882-889.
- [28] Swallow, W. H. Relative Mean Squared Error and Cost Considerations in Choosing Group Size for Group Testing to Estimate Infection Rates and Probabilities of Disease Transmission. *Phytopathology* **1987**, *77*, 1376-1381.
- [29] Thompson, K. H. Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics*, **1962**, *18*, 568-578.
- [30] Tu, X. M.; Litvak, E.; Pagano, M. Studies of AIDS and HIV surveillance. Screening tests: Can we get more by doing less? *Statistics in Medicine*, **1994**, *13*, 1905-1919.

- [31] Tu, X. M.; Litvak, E.; Pagano, M. On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika*, **1995**, *82*, 287-298.
- [32] Xie, M; Tatsuoka, K.; Sacks, J.; Young S. S., Group Testing Scheme in Cases of Blockers and Synergism. *Journal of the American Statistical Association*, **2001**, *96*, 92-102.
- [33] Young, S. S.; Hawkins, D. M. Using Recursive Partitioning to Analyze a Large SAR Data Set. *SAR and QSAR in Environmental Research* **1998**, *8*, 183-193.
- [34] Zhu, L.; Hughes-Oliver, J. M.; Young, S. S. Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics* **2001**, *57*, 922-930.
- [35] Vansteelandt, S.; Goetghebeur, E.; Verstraeten, T. Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples. *Biometrics* **2000**, *56*, 1126-1133.

Chapter 5

Applications of Semiparametric Theory on Missing Data Problems for Decoding Pools of Chemical Compounds

5.1 Introduction

In Chapter 4, likelihood models are used to estimate potency and blocking probabilities. Asymptotic properties of the resulting estimators are not obvious because the information matrix is singular in some cases (Zhu, Hughes-Oliver and Young, 2001). In this chapter, new likelihood models are developed for additionally estimating synergism. These models also allow variance estimation for the estimators. Further, semiparametric theory on missing data is successfully applied to the pooling experiment of chemical compounds and results in increased efficiency of estimation.

Pooling experiments can be viewed as a special missing data problem. Individual responses within a pool are missing when only the pooled response is obtained and

no retesting is done on individuals within this pool. For each pool, pooled responses and all structural information for all compounds within the pool are always available. The strategy for determining which pools will be retested is based on observed data only, thus the pooling data can be regarded as missing at random.

Semiparametric theory has been successfully applied in missing data problems to get efficient parameter estimation. Application to pooling experiments has never been done, at least to our knowledge. Here we work through the details and ultimately plan to compare the accumulation curves for semiparametric and likelihood-model-based approaches.

Section 5.2 gives a short literature review and introduces the theory of semiparametric models, as well as the application in restricted moment models and data missing at random. Section 5.3 specifies the semiparametric and corresponding likelihood models for pooling experiments in drug discovery. Section 5.4 applies the two models to a real pooling data set and compares the estimation results. Section 5.5 proposes another likelihood model and also compares it to the previous models. The last section proposes future work.

5.2 Literature Review

The literature on semiparametric models and estimators is extensive. Bickel (1982), Newey (1990) and Bickel et al. (1993) give ideas for finding consistent estimators. Ritov (1987), Newey, (1990), Bickel et al. (1993) discuss the limiting distribution of a semiparametric m -estimator. A very important issue that concerns many researchers is the bound for semiparametric efficiency. Newey (1990) gives a good introduction to this problem and provides a discussion of the research methods for determining and calculating bounds.

In the circumstance of missing data, especially data missing at random in the sense of Rubin (1976), Robins, Rotnitzky, and Zhao (1994) give a new class of semiparametric estimators, based on inverse probability weighted estimating equations.

Robins, Rotnitzky, and Zhao (1994) also compare estimators in their class with other estimators and show that each previous estimator is asymptotically equivalent to some, usually inefficient, estimator in their class.

This section contains three subsections summarizing relevant results of efficient semiparametric estimation, especially on data missing at random. The first subsection introduces the influence function and identifies the efficient score in semiparametric models. The second subsection gives the influence function and efficient score for restricted moment models. Following the one work of Robins, Rotnitzky, and Zhao (1994), the third subsection describes the process of finding the influence function and the efficient score for data missing at random in the sense of Rubin (1976). Efficient semiparametric estimators can be obtained by solving the corresponding efficient estimating equations.

This section is essentially a review of portions from lecture notes of Dr. Anastasios Tsiatis (Tsiatis, 2001). The lecture named *Semiparametric Theory and Missing Data Problem* was given in 2001 at Department of Statistics, North Carolina State University.

5.2.1 Influence Functions

Let Z_1, \dots, Z_n be i.i.d. random variables where Z_i has density assumed to belong to the class $\{P_Z(z; \theta), \theta \in \Omega\}$. The parameter θ can be written as $(\beta^T, \eta^T)^T$, where $\beta^{q \times 1}$ is the parameter of interest, and η , the nuisance parameter, may be finite- or infinite-dimensional.

We first consider the parametric model where the nuisance parameter is r -dimensional (see Pfanzagl and Wefelmeier, 1982; Begun et al., 1983; Newey, 1990; Bickel et al., 1993); extension can be made to the case where the nuisance parameter is infinite-dimensional. The score vector of a single observation Z in a parametric model is denoted by $S_\theta(Z, \theta_0) = \frac{\partial \log P_Z(Z, \theta)}{\partial \theta} \Big|_{\theta = \theta_0}$, which is of dimension $q + r$ and θ_0 is the true

value creating the data we observe. Let

$$S_\beta(Z, \theta_0) = \frac{\partial \log P_Z(Z, \theta)}{\partial \beta} \Big|_{\theta=\theta_0}, \quad q \times 1 \text{ vector},$$

$$S_\eta(Z, \theta_0) = \frac{\partial \log P_Z(Z, \theta)}{\partial \eta} \Big|_{\theta=\theta_0}, \quad r \times 1 \text{ vector}.$$

For simplicity, let S_β , and S_η represent $S_\beta(Z, \theta_0)$ and $S_\eta(Z, \theta_0)$ respectively. Then $S_\theta(Z, \theta_0) = (S_\beta^T, S_\eta^T)^T$.

The following theorem about the influence function is due to Newey (1990) and Bickel et al. (1993).

Theorem: Suppose $\hat{\beta}_n$ is an asymptotically linear estimator with influence function $\varphi(Z)$ such that $E_\theta(\varphi^T \varphi)$ exists and is continuous in θ in a neighborhood of θ_0 . If $\hat{\beta}_n$ is regular, then the following holds:

$$(i) E\{\varphi(Z)S_\beta\} = I^{q \times q}$$

$$(ii) E\{\varphi(Z)S_\eta\} = 0^{q \times r}.$$

Influence functions are uniquely determined by their estimators, that is, different $\hat{\beta}$ s have different influence functions. Thus, the previous theorem could be used to identify the efficient estimator. Specifically, we could first find all influence functions that satisfy the above two conditions, then determine the most efficient one. The resulting influence function uniquely determines the efficient $\hat{\beta}$.

Consider a space \mathcal{H} consisting of all q -dimensional random functions $h : \mathcal{Z} \rightarrow \mathcal{R}^q$, where h is measurable and satisfies $E(h) = 0$, $E(h^T h) < \infty$. \mathcal{Z} is the sample space and \mathcal{R}^q is the q -dimensional real number space. If the space \mathcal{H} is accommodated with an inner product $\langle h_1, h_2 \rangle = E(h_1^T h_2)$ for any h_1 and h_2 in this space, then this space is linear and complete, and is thus a Hilbert space. A subspace within \mathcal{H} , called a nuisance tangent space, is defined as

$$\Lambda = \{BS_\eta(Z, \theta_0) : \forall B^{q \times p}\}.$$

By this definition, the influence functions $\varphi(Z)$ must satisfy: (1) $E\{\varphi(Z)S_\beta^T\} = I^{q \times q}$; and (2) $\varphi(Z) \in \Lambda^\perp$. So, in order to find the influence functions, we need to identify the subspace perpendicular to Λ . The elements in subspace Λ^\perp are called scores.

It is shown by Newey (1990) that the efficient score is

$$S^{eff}(Z, \theta_0) = S_\beta - \Pi(S_\beta|\Lambda),$$

where $\Pi(S_\beta|\Lambda)$ is the projection of S_β on the nuisance tangent space.

It is easy to verify (Newey, 1990) that the efficient influence function is

$$\varphi^{eff}(Z) = E(S^{eff}S^{effT})^{-1}S^{eff}(Z, \theta_0).$$

5.2.2 Restricted Moment Models

Consider the model

$$Y = M(X, \beta) + \varepsilon,$$

where $E(\varepsilon|X) = 0$ and the functional form of $M(\cdot, \cdot)$ is known. It is assumed that data are the realization of the iid random vectors Z_1, \dots, Z_n , where $Z_i = (Y_i, X_i)$.

The density is characterized as

$$p(z, \beta, \eta_1(\cdot), \eta_2(\cdot)) = \eta_1(y - M(x, \beta), x) \cdot \eta_2(x).$$

where $\eta_1(\varepsilon, x) = p_{\varepsilon|x}(\varepsilon|x)$, and $\eta_1(\varepsilon, x)$, $\eta_2(x)$ are two densities. It could be shown (Tsiatis, 2001) that the score function, that is, the element perpendicular to the nuisance tangent space, is given by

$$A^{q \times d}(x)(y - M(x, \beta)),$$

and the corresponding influence function is

$$E[A(x)D(x)]^{-1}A(x)(y - M(x, \beta)).$$

Here $A^{q \times d}(x)$ is an arbitrary function of x , and $D(x) = \frac{\partial M(x, \beta)}{\partial \beta^T}$.

The efficient score is given by

$$D^T(x)V^{-1}(x)(y - M(x, \beta)).$$

where $V(x) = E(\varepsilon\varepsilon^T | x)$ (Tsiatis, 2001). Therefore, the most efficient estimator is the solution to the equation

$$\sum_{i=1}^n D^T(x_i)V^{-1}(x_i)(y_i - M(x_i, \beta)) = 0.$$

5.2.3 Data Missing at Random

When no missing data occurs, we observe the full data Z_1, \dots, Z_n . Let R_i be the indicator variable of observing complete data for the i th observation. When $R_i = 1$, we observe the complete data Z_i ; when $R_i = 0$, there is missing data for the i th sample and Z_i is not completely observed. Let $G_r(Z_i)$ be a function of Z_i and r and $G_r(Z_i)$ are partial data of Z_i . So when $R_i = 0$, we observe $G_0(Z_i)$, rather than Z_i , and when $R_i = 1$, we observe $G_1(Z_i)$ which is actually Z_i .

If it holds that $P(R = r | Z) = \pi(r, G_r(Z))$, which implies that the probability of observing complete data depends on Z only through a function of the observed data, then this situation is called missing at random (Rubin, 1976).

Given data missing at random, it can be shown that the corresponding score function is in the form of

$$\frac{I(R = 1)\varphi^F(Z)}{\pi(1, Z)} + L(R, G_R(Z)),$$

where $L(R, G_R(Z))$ is an arbitrary function from the subspace $\Lambda_2 = \{h(R, G_R(Z)) \in \mathcal{H} : E(h(R, G_R(Z)) | Z) = 0\}$ and $\varphi^F(Z)$ is the full data score function (Robins, Rotnitzky, and Zhao, 1994).

Provided that the missing variable R is binary, we can express any function $h(R, G_R(Z)) \in \mathcal{H}$ as

$$I(R = 1)L(1, G_1(Z)) + I(R = 0)L(0, G_0(Z)).$$

Because any function $L(R, G_R(Z)) \in \Lambda_2 \subset \mathcal{H}$ must satisfy

$$E\{L(R, G_R(Z)) | Z\} = 0,$$

that is

$$E\{[I(R = 1)L(1, G_1(Z)) + I(R = 0)L(0, G_0(Z))]\} = 0,$$

it can be solved to obtain

$$L(1, G_1(Z)) = -\frac{1 - \pi(1, Z)}{\pi(1, Z)}L(0, G_0(Z)).$$

Thus, a typical element of Λ_2 can be written as

$$-\frac{I(R = 1)(1 - \pi(1, Z))}{\pi(1, Z)}L(0, G_0(Z)) + I(R = 0)L(0, G_0(Z)).$$

R is either 1 or 0, so R can be used to replace $I(R = 1)$. Consequently, a typical element of Λ_2 can be rewritten as

$$\frac{R - \pi(1, Z)}{\pi(1, Z)}L(G_0(Z)),$$

where $L(G_0(Z))$ is an arbitrary function of $G_0(Z)$ satisfying $E[L(G_0(Z)) | Z] = 0$, for example, $L(G_0(Z)) = -L(0, G_0(Z))$

Therefore, the observed data score function is in the form of

$$\frac{R \varphi^F(Z)}{\pi(1, Z)} + \frac{R - \pi(1, Z)}{\pi(1, Z)}L(G_0(Z)).$$

The full data score function for restricted moment model is $\varphi^F(Z) = A(x)(y - M(x, \beta))$. We now need to find optimal $\varphi^F(Z)$ and $L(G_0(Z))$, or optimal $A(x)$ and $L(G_0(Z))$ for the restricted moment model correspondingly.

Suppose we fix $\varphi^F(Z)$ as some arbitrary full-data influence function. Then, among the class of observed-data influence functions

$$\left\{ \frac{R \varphi^F(Z)}{\pi(1, Z)} + L_2(R, G_R(Z)) : L_2(R, G_R(Z)) \in \Lambda_2 \right\},$$

the element with the smallest variance is obtained by using the residual

$$\frac{R \varphi^F(Z)}{\pi(1, Z)} - \Pi \left(\frac{R \varphi^F(Z)}{\pi(1, Z)} \mid \Lambda_2 \right),$$

where $\Pi\left(\frac{R\varphi^F(Z)}{\pi(1,Z)} \mid \Lambda_2\right)$ is the projection of $\frac{R\varphi^F(Z)}{\pi(1,Z)}$ onto the subspace Λ_2 (Tsiatis, 2001).

In order to find the most efficient influence function, it is enough to consider influence functions in the form of

$$\frac{R\varphi^F(Z)}{\pi(1,Z)} - \Pi\left(\frac{R\varphi^F(Z)}{\pi(1,Z)} \mid \Lambda_2\right),$$

where $\varphi^F(Z)$ is an arbitrary full data influence function.

We first compute the projection of $\frac{R\varphi^F(Z)}{\pi(1,Z)}$ onto Λ_2 . As above, a typical element in space Λ_2 can be written as $\frac{R-\pi(1,Z)}{\pi(1,Z)}L(G_0(Z))$. So in order to find $\Pi\left(\frac{R\varphi^F(Z)}{\pi(1,Z)} \mid \Lambda_2\right)$, we need to find an element $L_0(G_0(Z))$ in Λ_2 such that for all $L(G_0(Z)) \in \Lambda_2$, the following holds:

$$E\left\{\left[\frac{R\varphi^F(Z)}{\pi(1,Z)} - \frac{R-\pi(1,Z)}{\pi(1,Z)}L_0(G_0(Z))\right]\frac{R-\pi(1,Z)}{\pi(1,Z)}L(G_0(Z))\right\} = 0.$$

Solving this equation, we obtain

$$L_0(G_0(Z)) = E(\varphi^F(Z) \mid G_0(Z)).$$

So, the projection of $\frac{R\varphi^F(Z)}{\pi(1,Z)}$ onto Λ_2 is obtained as

$$\frac{R-\pi(1,Z)}{\pi(1,Z)}E(\varphi^F(Z) \mid G_0(Z)).$$

Therefore, for a fixed arbitrary function $\varphi^F(Z)$, the most efficient score function can be written as

$$\frac{R\varphi^F(Z)}{\pi(1,Z)} - \frac{R-\pi(1,Z)}{\pi(1,Z)}E(\varphi^F(Z) \mid G_0(Z)).$$

For the restricted moment model, after replacing $\varphi^F(Z)$ by $A(x)(y - M(x, \beta))$, the corresponding efficient score can be obtained.

The most efficient score function for the restricted moment models, shown by Robins, Rotnitzky, and Zhao (1994), is associated with $A_{eff}(x)$ satisfying the following iteration relationship

$$A_{eff}(x) = D^T(x)T^*(x) + E\left\{\frac{1-\pi(1,Z)}{\pi(1,Z)}E[A_{eff}(x)\varepsilon \mid G_0(Z)]\varepsilon^T \mid x\right\}T^*(x),$$

where $\varepsilon = Y - M(x, \beta)$, $D(x) = \frac{\partial M(x, \beta)}{\partial \beta^T}$ and $T^*(x) = \left\{E\left[\frac{\varepsilon\varepsilon^T}{\pi(1,Z)} \mid x\right]\right\}^{-1}$.

5.3 Semiparametric and Likelihood Models for Pooling Experiments

In pooling experiments for chemical compounds, retesting is usually limited to the potent pools. Consequently, some pools have both observed pool responses and individual potencies. However, within pools, retested or not, the structural information of each compound is always available. The information is called covariate and denoted as x_{ij} for the j th individual compound in the i th pool. x_{ij} is an L dimensional row vector of 0's except the l th element is 1, provided this compound belongs to the l th class. The determination of the classes is described in Chapter 4. In this section, we are going to estimate the probability for each class potency, $p_l = \Pr(\text{compound from class } l \text{ is potent})$.

In order to achieve efficient estimation for the parameters, some information on individual potencies, as well as the pool responses, is needed. This calls for a procedure to determine which pools will be retested in order to obtain individual potencies within these pools.

It is reasonable that potent pools should have a big chance of being selected for retesting. Inactive pools will also have a small chance to be retested; this is consistent with the existence of blocking effects. Let R_i be the indicator that pool i is retested. The retesting model is assumed as

$$\Pr(R_i = 1|y_i) = \pi(y_i) = \begin{cases} \pi_1 & \text{if } y_i = 1 \\ \pi_0 & \text{if } y_i = 0 \end{cases},$$

where π_1 is going to be set close to 1 and π_0 close to 0.

5.3.1 Semiparametric Models

It is easily seen that the full data for one pool is $Z_i = (Y_i, x_i, Y_i^*)$, where Y_i is the potency of the i th pool; $Y_i^* = (Y_{i1}, Y_{i2}, \dots, Y_{ik})^T$; Y_{ij} is the potency of the j th compound within the i th pool; $x_i = (x_{i1}^T, x_{i2}^T, \dots, x_{ik}^T)^T$ is the $k \times q$ covariate matrix for

the i th pool; and x_{ij} is a q -dimensional row vector as covariate for the j th compound within the i th pool. For a retested pool, we observe Z_i and hence $R_i = 1$; otherwise, we observe (Y_i, x_i) and $R_i = 0$.

As it is true that

$$\Pr(R_i = 1 \mid y_i, x_i, y_i^*) = \Pr(R_i = 1 \mid y_i) = \pi(y_i),$$

this is a Missing at Random (MAR) problem. The full data model is assumed to be

$$Y_i^* = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{pmatrix} = \begin{pmatrix} M(x_{i1}, \beta) + \varepsilon_{i1} \\ M(x_{i2}, \beta) + \varepsilon_{i2} \\ \vdots \\ M(x_{ik}, \beta) + \varepsilon_{ik} \end{pmatrix} = \begin{pmatrix} M(x_{i1}, \beta) \\ M(x_{i2}, \beta) \\ \vdots \\ M(x_{ik}, \beta) \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{ik} \end{pmatrix},$$

where $E(\varepsilon_{ij} \mid x_{ij}) = 0$, $(x_{ij}, \varepsilon_{ij})$ are i.i.d. and $M(x_{ij}, \beta) = \exp(x_{ij}\beta)/(1 + \exp(x_{ij}\beta))$. As part of the full data model, it is also assumed that $\Pr(Y_i = 1 \mid \sum_{j=1}^k Y_{ij} > 0) = 1 - B$ and $\Pr(Y_i = 1 \mid \sum_{j=1}^k Y_{ij} = 0) = S$. B is the probability that a potent pool has blocking effect so that it is observed to be not potent and S is the probability that a truly inactive pool has synergism effect so that it is observed to be potent even though none of its individuals are potent. For simplicity, both B and S are assumed known in this subsection.

Let $M^*(x_i, \beta)$ denote $(M(x_{i1}, \beta), M(x_{i2}, \beta), \dots, M(x_{ik}, \beta))^T$, and ε_i^* denote $(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{ik})^T$. As given in Section 5.2.3, the most efficient score is

$$\frac{R \varphi^F(Z)}{\pi(1, Z)} - \frac{R - \pi(1, Z)}{\pi(1, Z)} E(\varphi^F(Z) \mid G_0(Z)).$$

For our case, $\varphi^F(Z_i) = A_{eff}(x_i)(Y_i^* - M^*(x_i, \beta))$, $\pi(1, Z_i) = \pi(Y_i)$, and $G_0(Z_i) = (Y_i, x_i)$. It is easily seen that $E(\varphi^F(Z_i) \mid G_0(Z_i)) = A_{eff}(x_i)(E(Y_i^* \mid Y_i, x_i) - M^*(x_i, \beta))$.

Therefore, the efficient inverse probability weighted estimating equation (IPWEE) is as follows:

$$\sum_{i=1}^n \left[\frac{r_i}{\pi(y_i)} A_{eff}(x_i)(y_i^* - M^*(x_i, \beta)) - \frac{r_i - \pi(y_i)}{\pi(y_i)} A_{eff}(x_i)(E(Y_i^* \mid y_i, x_i) - M^*(x_i, \beta)) \right] = 0,$$

where $A_{eff}(x_i)$ is determined by Robins' iteration equation (Robins, Rotnitzky, and Zhao, 1994). For this application, Robins' iteration equation is equivalent to the following

$$A_{eff}(x_i) = D^T(x_i)T^*(x_i) + A_{eff}(x_i)E\left\{\frac{1 - \pi(Y_i)}{\pi(Y_i)}E[\varepsilon_i^*|y_i, x_i]\varepsilon_i^{*T} \mid x_i\right\}T^*(x_i),$$

where $D(x_i) = \frac{\partial M(x_i, \beta)}{\partial \beta^T}$ and $T^*(x_i) = \left\{E\left[\frac{\varepsilon\varepsilon^T}{\pi(Y_i)} \mid x_i\right]\right\}^{-1}$.

From the equation above, $A_{eff}(x_i)$ can be directly solved for each pool i . Given the value of β , it can be computed that

$$\begin{aligned} & E\left\{\frac{1 - \pi(Y_i)}{\pi(Y_i)}E[\varepsilon_i^* \mid y_i, x_i]\varepsilon_i^{*T} \mid x_i\right\} \\ &= (1 - B - H_i)\left[\left(\frac{1}{\pi(1)} - 1\right)\frac{1-B}{H_i} - \frac{1}{\pi(1)} + \frac{1}{\pi(0)} - \left(\frac{1}{\pi(0)} - 1\right)\frac{B}{1-H_i}\right]M_i^*M_i^{*T}, \\ D(x_i) &= (M(x_{i1}, \beta)(1 - M(x_{i1}, \beta))x_{i1}^T, \dots, M(x_{ik}, \beta)(1 - M(x_{ik}, \beta))x_{ik}^T)^T, \\ T^*(x_i) &= \left\{\left(\frac{1}{\pi(1)} - \frac{1}{\pi(0)}\right)(H_i - 1 + B)M_i^*M_i^{*T} + \left(\frac{1-B}{\pi(1)} + \frac{B}{\pi(0)}\right)diag(M_i^*)\right. \\ & \quad \left. \cdot diag(1_k - M_i^*)\right\}^{-1}, \end{aligned}$$

where $M_i^* = M^*(x_i, \beta)$, $H_i = 1 - B - (1 - B - S)\prod_{j=1}^k \frac{1}{1 + \exp(x_{ij}\beta)}$, and 1_k is the k -dimension vector with all elements equal to 1.

In order to solve the efficient estimating equation

$$\sum_{i=1}^n \left[\frac{r_i}{\pi(y_i)}A_{eff}(x_i)(y_i^* - M^*(x_i, \beta)) - \frac{r_i - \pi(y_i)}{\pi(y_i)}A_{eff}(x_i)(E(Y_i^*|y_i, x_i) - M^*(x_i, \beta)) \right] = 0,$$

Taylor's expansion is applied to $M^*(x_i, \beta)$ in the neighborhood of β_0 :

$$M^*(x_i, \beta) \approx M^*(x_i, \beta_0) + D(x_i, \beta_0)(\beta - \beta_0).$$

Therefore, we can rewrite the estimating equation as

$$\beta \approx \beta_0 + \left[\sum_{i=1}^n A_{eff}(x_i)D(x_i) \right]^{-1} \sum_{i=1}^n A_{eff}(x_i)[f(x_i, \beta_0) - M^*(x_i, \beta_0)],$$

where $f(x_i, \beta_0) = \frac{r_i}{\pi(y_i)}y_i^* - \frac{r_i - \pi(y_i)}{\pi(y_i)}E(Y_i^*|y_i, x_i)$ and $A_{eff}(x_i)$ and $D(x_i)$ are evaluated based on β_0 . In order to obtain the solution to the efficient estimating equation, we

iteratively update β_0 with β computed from the last step. Upon convergence, when the difference between β and β_0 is very small, we obtain our final estimator $\hat{\beta}$ of β .

Let A , D , and f be abbreviated forms for $A_{eff}(x_i)$, $D(x_i)$, and $f(x_i, \beta)$, respectively. It can be shown that the influence function $\varphi(x_i)$ of $\hat{\beta}$ for the i th pool is given as

$$[E(AD)]^{-1}A(f - M^*(x_i, \beta_0)).$$

Therefore,

$$var(\hat{\beta}) = \frac{1}{n}var(\varphi(x_i)) = \frac{1}{n}[E(AD)]^{-1}var[A(f - M^*(x_i, \beta_0))][E(D^T A^T)]^{-1},$$

where

$$var[A(f - M^*(x_i, \beta_0))] = A \text{diag}(M^*(x_i, \beta_0))\text{diag}(1_k - M^*(x_i, \beta_0))A^T$$

is derived in the Appendix. Thus,

$$var(\hat{\beta}) = \frac{1}{n}[E(AD)]^{-1}A \text{diag}(M^*(x_i, \beta_0))\text{diag}(1_k - M^*(x_i, \beta_0))A^T[E(D^T A^T)]^{-1}.$$

Consequently, we obtain an estimator of $var(\hat{\beta})$ that is given by

$$\widehat{var}(\hat{\beta}) = \left[\sum_{i=1}^n (AD) \right]^{-1} \sum_{i=1}^n \left\{ A \text{diag}(M^*(x_i, \hat{\beta}))\text{diag}(1_k - M^*(x_i, \hat{\beta}))A^T \right\} \left[\sum_{i=1}^n (D^T A^T) \right]^{-1}.$$

5.3.2 Likelihood Model

The likelihood function for the pooling data as described in Section 5.3.1 on semi-parametric modeling is now obtained. The model is hierarchical, consisting of at most three layers: one layer for individual potencies, another layer for the effect of pooling, and a final layer for the retesting procedure.

When a pool is not retested ($R_i = 0$), we do not observe the first layer, so the contribution to the likelihood is obtained as

$$\Pr(Y_i = y_i) \Pr(R_i = 0 | Y_i = y_i) = \Pr(Y_i = y_i) (1 - \pi(y_i)),$$

where Y_i follows a Bernoulli distribution with probability

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(Y_i = 1 \mid \sum_{j=1}^k Y_{ij} > 0) \Pr(\sum_{j=1}^k Y_{ij} > 0) \\ &\quad + \Pr(Y_i = 1 \mid \sum_{j=1}^k Y_{ij} = 0) \Pr(\sum_{j=1}^k Y_{ij} = 0) \\ &= (1 - B) [1 - \prod_{j=1}^k (1 - p_{ij})] + S \prod_{j=1}^k (1 - p_{ij}) \\ &= (1 - B) - (1 - B - S) \prod_{j=1}^k (1 - p_{ij}),\end{aligned}$$

$p_{ij} = \Pr(Y_{ij} = 1) = E(Y_{ij}) = M(x_{ij}, \beta) = \exp(x_{ij}\beta)/(1 + \exp(x_{ij}\beta))$, and B , S , x_{ij} are defined the same as in Section 5.3.1, that is, $B = \Pr(Y_i = 0 \mid \sum_{i=1}^k Y_{ij} > 0)$, $S = \Pr(Y_i = 1 \mid \sum_{i=1}^k Y_{ij} = 0)$, and x_{ij} is a vector of 0's except the l th element of this vector is 1 when the compound x_{ij} belongs to the class l .

When a pool is retested ($R_i = 1$), we observe all layers and the contribution to the likelihood is obtained as

$$\Pr(Y_i^* = y_i^*) \Pr(Y_i = y_i \mid Y_i^* = y_i^*) \Pr(R_i = 1 \mid Y_i = y_i, Y_i^* = y_i^*).$$

Recall that $Y_i^* = (Y_{i1}, \dots, Y_{ik})^T$ is the vector of individual potencies within pool i , so that

$$\Pr(Y_i^* = y_i^*) = \prod_{j=1}^k p_{ij}^{y_{ij}^*} (1 - p_{ij})^{1 - y_{ij}^*}.$$

Moreover,

$$\begin{aligned}\Pr(Y_i = y_i \mid Y_i^* = y_i^*) &= \begin{cases} \Pr(Y_i = y_i \mid \sum_{i=1}^k Y_{ij} > 0) = (1 - B)^{y_i} B^{1 - y_i} \\ \Pr(Y_i = y_i \mid \sum_{i=1}^k Y_{ij} = 0) = S^{y_i} (1 - S)^{1 - y_i} \end{cases} \\ &= [(1 - B)^{y_i} B^{1 - y_i}]^{y_i^0} [S^{y_i} (1 - S)^{1 - y_i}]^{1 - y_i^0},\end{aligned}$$

where $y_i^0 = I(\sum_{j=1}^k y_{ij} > 0)$, and

$$\Pr(R_i = 1 \mid Y_i = y_i, Y_i^* = y_i^*) = \Pr(R_i = 1 \mid Y_i = y_i) = \pi(y_i),$$

as established for missing-at-random in Section 5.3.1.

Hence the likelihood, for $\theta = (p_{11}, \dots, p_{nk})^T$, is obtained as

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^n \left\{ [\pi(y_i)]^{r_i} [1 - \pi(y_i)]^{(1 - r_i)} \right\} \\ &\quad \cdot \left\{ [1 - B - (1 - B - S) \prod_{j=1}^k (1 - p_{ij})]^{y_i} [B + (1 - B - S) \prod_{j=1}^k (1 - p_{ij})]^{1 - y_i} \right\}^{1 - r_i} \\ &\quad \cdot \left\{ [\prod_{j=1}^k p_{ij}^{y_{ij}^*} (1 - p_{ij})^{1 - y_{ij}^*}] [(1 - B)^{y_i} B^{(1 - y_i)}]^{y_i^0} [S^{y_i} (1 - S)^{(1 - y_i)}]^{1 - y_i^0} \right\}^{r_i}\end{aligned}$$

This likelihood function can be simplified. Compounds fall in only one class out of a total of L classes. Therefore, there are only L different p_{ij} s, which are denoted as $p_l = \exp(\beta_l)/(1 + \exp(\beta_l))$, $l = 1, 2, \dots, L$ with vector $p = (p_1, \dots, p_L)^T$. Now, denote $S_{il} = \sum_{j=1}^k I(\text{compound } x_{ij} \text{ is from class } l)$, $W_{il} = \sum_{j=1}^k y_{ij} I(\text{compound } x_{ij} \text{ is from class } l)$, and $H_i = 1 - B - (1 - B - S) \prod_{l=1}^L (1 - p_l)^{S_{il}}$. Then the likelihood function, for $p = (p_1, \dots, p_L)^T$, can be written as:

$$\mathcal{L}(p) = \prod_{i=1}^n \left\{ [\pi(y_i)]^{r_i} [1 - \pi(y_i)]^{(1-r_i)} \right\} \left\{ H_i^{y_i} (1 - H_i)^{(1-y_i)} \right\}^{1-r_i} \cdot \left\{ \left[\prod_{l=1}^L p_l^{W_{il}} (1 - p_l)^{S_{il} - W_{il}} \right] [(1 - B)^{y_i} B^{(1-y_i)}]^{y_i} [(1 - S)^{(1-y_i)} S^{y_i}]^{1-y_i} \right\}^{r_i}$$

The Fisher score is $(\frac{\partial \ln \mathcal{L}(p)}{\partial p_1}, \frac{\partial \ln \mathcal{L}(p)}{\partial p_2}, \dots, \frac{\partial \ln \mathcal{L}(p)}{\partial p_L})$, where $\frac{\partial \ln \mathcal{L}(p)}{\partial p_m}$ is equal to

$$\sum_{i=1}^n \left\{ r_i \left(\frac{W_{im}}{p_m} - \frac{S_{im} - W_{im}}{1 - p_m} \right) + (1 - r_i) y_i \frac{S_{im}(1 - B - H_i)}{(1 - p_m) H_i} I(S_{im} \geq 1) + (1 - r_i)(1 - y_i) \frac{S_{im}(B + H_i - 1)}{(1 - p_m)(1 - H_i)} I(S_{im} \geq 1) \right\},$$

for $m = 1, 2, \dots, L$.

Let $I(p)$ denote the $L \times L$ observed information matrix. Then the element of the m th row and u th column ($m \neq u$) is in the form of

$$\sum_{i=1}^n \left\{ (1 - r_i) y_i \frac{S_{im} S_{iu} (1 - B)(1 - B - H_i)}{(1 - p_m)(1 - p_u) H_i^2} I(S_{im} \geq 1) I(S_{iu} \geq 1) + (1 - r_i)(1 - y_i) \frac{S_{im} S_{iu} B(B + H_i - 1)}{(1 - p_m)(1 - p_u)(1 - H_i)^2} I(S_{im} \geq 1) I(S_{iu} \geq 1) \right\},$$

for $m = 1, 2, \dots, L$, $u = 1, 2, \dots, L$, and $m \neq u$. The element of the m th row and m th column is in the form of

$$\sum_{i=1}^n \left\{ r_i \left(\frac{W_{im}}{p_m^2} + \frac{S_{im} - W_{im}}{(1 - p_m)^2} \right) + (1 - r_i) y_i \frac{S_{im}(H_i + B - 1)(H_i - S_{im}(1 - B))}{(1 - p_m)^2 H_i^2} I(S_{im} \geq 2) + (1 - r_i)(1 - y_i) \frac{S_{im}(H_i + B - 1)(H_i + S_{im}B - 1)}{(1 - p_m)^2 (1 - H_i)^2} I(S_{im} \geq 2) \right\},$$

for $m = 1, 2, \dots, L$.

5.4 Real Data Applications

The application in this section uses the same data as in Chapter 4. Here is a brief summary for that data set. In a pooling experiment at GlaxoSmithKline, 1000

Table 5.1: *Distribution of Potent Compounds in Covariate Classes*

Covariate classes	1	2	3	4	5	6	7	Sum
Potent compounds	2	1	2	16	6	4	9	40
Total compounds	509	5	11	412	27	5	31	1000
Observed Proportion	.0039	.2000	.1818	.0388	.2222	.8000	.2903	.0400

chemical compounds are tested and percentage of inhibition is given as potencies. Compounds are pooled in groups of 10, resulting in 100 groups. Individual covariate class information and 100 pool potencies are available for three types of pools (called DATA1, DATA2, and DATA3 in Chapter 4). In this section, only DATA1 is used.

As in Chapter 4, the threshold for binary potency of individuals is 60. There are 40 compounds with percent inhibition greater than 60 and labeled as potent. So, the percentage of potent compounds is 4%. By using the optimal threshold of 18.9 for defining binary pool potent (see Section 4.6 in Chapter 4), 22 pools out of 100 are labeled as potent.

Also, we are able to classify these 1000 compounds into seven covariate classes based on eight atom pairs (see Section 4.2 in Chapter 4). The number of potent compounds in each class is shown in Table 5.1.

We model the existence of both blocking effect B and synergism effect S in pools. These two effects can be treated either as unknown (need to be estimated) or are known (set to specific values). From the real data, the effects of blocking and synergism are estimated as $B = 0.32$, $S = 0.01$, when supposing all individual potencies are known and using potency thresholds of 60 for individuals and 18.9 for pools. In this section, $B = 0.32$ and $S = 0.01$ are assumed known for the purpose of fitting the semiparametric and likelihood models proposed in Section 5.3. However, B and S are assumed unknown and are estimated as part of the likelihood model proposed in Section 5.5.

Table 5.2: *Maximum Likelihood Estimates*

Parameter	p_1	p_2	p_3	p_4	p_5	p_6	p_7
Estimates	0.003	0.337	0.237	0.036	0.256	0.795	0.281
Observed	0.004	0.200	0.182	0.039	0.222	0.800	0.290

Table 5.3: *Estimated Variance-Covariance Matrix of Maximum Likelihood Estimators*

9.00E-6	4.23E-6	2.38E-6	3.07E-7	1.73E-6	-2.89E-6	-1.52E-6
4.23E-6	0.091	2.00E-5	1.00E-4	5.49E-4	-1.10E-5	5.90E-4
2.38E-6	2.00E-5	0.023	1.74E-5	5.74E-6	-2.40E-6	1.24E-4
3.07E-7	1.00E-4	1.74E-5	1.34E-4	8.71E-7	-1.30E-5	1.97E-5
1.73E-6	5.49E-4	5.74E-6	8.71E-7	0.011	-6.34E-7	5.01E-4
-2.89E-6	-1.10E-5	-2.40E-6	-1.30E-5	-6.34E-7	0.034	-1.37E-6
-1.52E-6	5.90E-4	1.24E-4	1.97E-5	5.01E-4	-1.37E-6	0.012

Selection of pools for retesting is accomplished using

$$\Pr(R_i = 1|y_i) = \pi(y_i) = \begin{cases} \pi_1 = 0.9 & \text{if } y_i = 1 \\ \pi_0 = 0.1 & \text{if } y_i = 0 \end{cases}.$$

When a pool is active, it has 90% chance to be retested; otherwise, an inactive pool has 10% chance to be retested. (It is not necessary that $\pi_1 + \pi_0 = 1$.) For example, one randomly selected retesting data set retests 19 out of 22 active pools and 8 out of 78 inactive pools. For this data set, the maximum likelihood estimator \hat{p} is given in Table 5.2. The corresponding estimated variance-covariance matrix of these MLEs, obtained from inverting the observed information matrix, is shown in Table 5.3.

For this same data set, estimates obtained from the semiparametric approach are given in Table 5.4. The corresponding estimated variance-covariance matrix is given in Table 5.5.

In order to compare the two estimators, Table 5.6 lists both sets of estimates and standard errors for the likelihood and semiparametric methods. We can see that, except for \hat{p}_6 , the semiparametric estimator is more efficient than the MLE.

Table 5.4: *Estimates from the Semiparametric Model*

Parameter	p_1	p_2	p_3	p_4	p_5	p_6	p_7
Estimates	0.002	0.318	0.163	0.026	0.245	0.733	0.299
Observed	0.004	0.200	0.182	0.039	0.222	0.800	0.290

Table 5.5: *Estimated Variance-Covariance Matrix of Semiparametric Estimators*

4.13E-6	-5.28E-8	3.60E-8	7.80E-9	2.95E-9	1.21E-7	1.28E-8
-5.28E-8	0.044	-4.59E-6	-2.65E-7	-6.58E-6	3.98E-6	-1.00E-5
3.60E-8	-4.59E-6	0.012	-2.40E-7	-4.19E-6	3.92E-6	-6.16E-6
7.80E-9	-2.65E-7	-2.40E-7	6.06E-5	-6.79E-8	1.45E-6	-2.88E-7
2.95E-9	-6.58E-6	-4.19E-6	-6.79E-8	0.007	2.35E-5	-7.04E-6
1.21E-7	3.98E-6	3.92E-6	1.45E-6	2.35E-5	0.039	1.73E-5
1.28E-8	-1.00E-5	-6.16E-6	-2.88E-7	-7.04E-6	1.73E-5	0.007

Table 5.6: *Comparison of Estimates and Standard Deviations for Two Methods. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)*

Parameter	Observed	Counts	MLE	Semiparametric
p_1	.004	2/509	.003 (.003)	.002 (.002)
p_2	.200	1/5	.337 (.313)	.318 (.209)
p_3	.182	2/11	.237 (.151)	.163 (.112)
p_4	.039	16/412	.036 (.012)	.026 (.008)
p_5	.222	6/27	.256 (.107)	.245 (.083)
p_6	.800	4/5	.795 (.184)	.733 (.198)
p_7	.290	9/31	.281 (.109)	.299 (.083)

Table 5.7: *Average Results for Likelihood and Semiparametric Models. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)*

Parameter	Observed	Counts	MLE	Semiparametric
p_1	.004	2/509	.003 (.003)	.019 (.005)
p_2	.200	1/5	.291 (.351)	.305 (.198)
p_3	.182	2/11	.243 (.163)	.213 (.122)
p_4	.039	16/412	.035 (.011)	.046 (.010)
p_5	.222	6/27	.254 (.102)	.270 (.081)
p_6	.800	4/5	.811 (.217)	.748 (.191)
p_7	.290	9/31	.292 (.108)	.305 (.078)

The above conclusion is based on one randomly selected retesting data set. Now, 100 randomly selected retesting data sets are chosen to make such evaluations. Table 5.7 shows the average estimates and standard deviations for the two methods. On average, standard deviations from the semiparametric method are smaller than those from MLE, except for \hat{p}_1 . Some insight is obtained from comparing estimating equations for the two methods.

The semiparametric method has estimating equation

$$\sum_{i=1}^n \left[\frac{r_i}{\pi(y_i)} A_{eff}(x_i)(y_i^* - M^*(x_i, \beta)) - \frac{r_i - \pi(y_i)}{\pi(y_i)} A_{eff}(x_i)(E(y_i^* | y_i, x_i) - M^*(x_i, \beta)) \right] = 0,$$

where $A_{eff}(x_i)$ is determined by Robins' iteration equation (Robins, Rotnitzky, and Zhao, 1994). Under the special case of no missing data, $A_{eff}(x_i)$ is denoted as $A_{eff}^F(x_i)$, which is equal to $D^T(x_i)V^{-1}(x_i)$. In our case,

$$D(x_i) = (M(x_{i1}, \beta)(1 - M(x_{i1}, \beta))x_{i1}^T, \dots, M(x_{ik}, \beta)(1 - M(x_{ik}, \beta))x_{ik}^T)^T,$$

and $V(x_i) = \text{diag}(M_i^*)\text{diag}(1_k - M_i^*)$, so that $A_{eff}^F(x_i) = (x_{i1}^T, \dots, x_{ik}^T)$. Recalling that $r_i = 1$ and $\pi(y_i) = 1$ for all i s when there is no missing data, the estimating equation becomes

$$\sum_{i=1}^n A_{eff}^F(x_i)(y_i^* - M^*(x_i, \beta)) = 0.$$

For the likelihood model, the estimating equation is obtained simply by setting the Fisher score equal to 0. For our application, it is

$$\sum_{i=1}^n \left[r_i \sum_{j=1}^k x_{ij}^T (y_{ij} - M(x_{ij}, \beta)) + (1 - r_i) y_i \sum_{j=1}^k x_{ij}^T M(x_{ij}, \beta) \frac{(1-B-S)U_i}{1-B-(1-B-S)U_i} - (1 - r_i)(1 - y_i) \sum_{j=1}^k x_{ij}^T M(x_{ij}, \beta) \frac{(1-B-S)U_i}{B+(1-B-S)U_i} \right] = 0,$$

which can be rewritten as

$$\sum_{i=1}^n \left[r_i A_{eff}^F(x_i)(y_i^* - M^*(x_i, \beta)) + (1 - r_i) y_i A_{eff}^F(x_i) M^*(x_i, \beta) \frac{(1-B-S)U_i}{1-B-(1-B-S)U_i} - (1 - r_i)(1 - y_i) A_{eff}^F(x_i) M^*(x_i, \beta) \frac{(1-B-S)U_i}{B+(1-B-S)U_i} \right] = 0,$$

where $U_i = \prod_{j=1}^k (1 + \exp(x_{ij}\beta))^{-1}$.

Comparing the two estimating equations, we can see that when there is no missing data, that is, $r_i = 1$ for all i s, the two equations are equivalent. However, when some pools have missing data (so that the individual potencies are missing due to no retesting for this pool), the two estimating equations are very different. The likelihood equation still uses the $A_{eff}^F(x_i)$ that is associated with the full data, while the semiparametric equation newly determines $A_{eff}(x_i)$. By the semiparametric theory on data missing at random, this estimating equation obtains estimators locally efficient in the sense of Robins, Rotnitzky, and Zhao (1994). Therefore, in our application, we observe more efficient estimators than MLEs.

Unfortunately, the semiparametric estimators are more biased in this application. That may be because the sample, which has only 100 pools, is too small to allow realization of the benefits of the asymptotic property of consistency for semiparametric estimators.

5.5 Another Likelihood Model

In the likelihood and semiparametric models proposed in Section 5.3, no measurement errors are assumed, and blocking effect B and synergism effect S are assumed to be known. In this section, B and S are unknown and needed to be estimated from

the observed data. Measurement errors can exist and are included in this model, but they are assumed to be known.

When measurement errors are considered in the pooling experiment, they exist for measurements of both individual and pool responses. In this section, common sensitivity and specificity rates are assumed for both individuals and pools. They are denoted as S_e and S_p .

As for the likelihood model proposed in Section 5.3, the model considered here is also hierarchical, consisting of at most five layers. The first layer is the true individual potencies. The second is the observed individual potencies, that is, the true potencies plus the measurement error. The third layer is the pooling effect, where the true pool potency is obtained after possible adjustment by blocking or synergism effect on the sum of true individual potencies of compounds within the pool. The fourth layer is the observed pool potency, based on the third layer plus the measurement error. The fifth is the retesting procedure.

Let Y_i^0 be the true pool potency after either blocking or synergism (if there is no measurement error, Y_i^0 will be the observed pool potency) and Y_{ij}^0 be the true potency of individual. When a pool is not retested ($R_i = 0$), we do not observe the first or second layers. The contribution to the likelihood is obtained as

$$\Pr(Y_i = y_i) \Pr(R_i = 0 | Y_i = y_i) = \Pr(Y_i = y_i) (1 - \pi(y_i)),$$

where Y_i follows a Bernoulli distribution with probability

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(Y_i = 1 | Y_i^0 = 1) \Pr(Y_i^0 = 1) + \Pr(Y_i = 1 | Y_i^0 = 0) \Pr(Y_i^0 = 0) \\ &= S_e \Pr(Y_i^0 = 1) + (1 - S_p) \Pr(Y_i^0 = 0). \end{aligned}$$

Exactly the same as in Section 5.3.2, $\Pr(Y_i^0 = 1) = 1 - B - C_{BS}V_i$, where $V_i = \prod_{j=1}^k (1 - p_{ij})$ and $C_{BS} = 1 - B - S$. Therefore,

$$\begin{aligned} \Pr(Y_i = 1) &= S_e(1 - B - C_{BS}V_i) + (1 - S_p)(B + C_{BS}V_i) \\ &= S_e + C_{ep}(B + C_{BS}V_i), \end{aligned}$$

where $C_{ep} = 1 - S_e - S_p$.

Recall that $Y_i^* = (Y_{i1}, \dots, Y_{ik})^T$ is the vector of individual potencies within pool i , and let

$$Q_i = \Pr(Y_i^* = y_i^*) = \prod_{j=1}^k \left[p_{ij} S_e^{y_{ij}} (1 - S_e)^{1-y_{ij}} + (1 - p_{ij})(1 - S_p)^{y_{ij}} S_p^{1-y_{ij}} \right],$$

and

$$G_i = \prod_{j=1}^k \Pr(Y_{ij} = y_{ij}, Y_{ij}^0 = 0) = \prod_{j=1}^k (1 - p_{ij})(1 - S_p)^{y_{ij}} S_p^{1-y_{ij}}.$$

When a pool is retested ($R_i = 1$), we observe all layers and the contribution to the likelihood is obtained as

$$\Pr(Y_i = y_i, Y_i^* = y_i^*) \Pr(R_i = 1 | Y_i = y_i, Y_i^* = y_i^*).$$

Moreover,

$$\begin{aligned} \Pr(Y_i = 1, Y_i^* = y_i^*) &= Q_i \Pr(Y_i = 1 | Y_i^* = y_i^*) \\ &= Q_i (S_e + C_{ep} \Pr(Y_i^0 = 0 | Y_i^* = y_i^*)) \\ &= Q_i (S_e + C_{ep} (B + C_{BS} \Pr(\sum_{j=1}^k Y_{ij}^0 = 0 | Y_i^* = y_i^*))) \\ &= Q_i S_e + C_{ep} (Q_i B + C_{BS} G_i), \end{aligned}$$

and

$$\Pr(R_i = 1 | Y_i = y_i, Y_i^* = y_i^*) = \Pr(R_i = 1 | Y_i = y_i) = \pi(y_i),$$

as established for Missing at Random (MAR) in Section 5.3.1.

Hence, the likelihood can be written for $\theta = (B, S, p_{11}, \dots, p_{nk})$ as follows

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^n \left\{ [\pi(y_i)]^{r_i} [1 - \pi(y_i)]^{(1-r_i)} \right\} \\ &\quad \{ [S_e + C_{ep} (B + C_{BS} V_i)]^{y_i} [1 - S_e - C_{ep} (B + C_{BS} V_i)]^{1-y_i} \}^{1-r_i} \\ &\quad \{ [Q_i S_e + C_{ep} (B Q_i + C_{BS} G_i)]^{y_i} [Q_i - Q_i S_e - C_{ep} (B Q_i + C_{BS} G_i)]^{1-y_i} \}^{r_i}. \end{aligned}$$

Compounds exist in a total of at most L classes, so the likelihood can be simplified and rewritten as a function of $\eta = (B, S, p_1, \dots, p_L)$ through rewriting V_i , Q_i , and G_i as $V_i = \prod_{l=1}^L (1 - p_l)^{S_{il}}$, $Q_i = \prod_{l=1}^L [p_l S_e + (1 - p_l)(1 - S_p)]^{W_{il}} [p_l (1 - S_e) + (1 - p_l) S_p]^{S_{il} - W_{il}}$, and $G_i = \prod_{l=1}^L (1 - p_l)^{S_{il}} (1 - S_p)^{W_{il}} S_p^{S_{il} - W_{il}}$, respectively. Recall that $S_{il} = \sum_{j=1}^k I(\text{compound } x_{ij} \text{ is from class } l)$, $W_{il} = \sum_{j=1}^k y_{ij} I(\text{compound } x_{ij} \text{ is from class } l)$.

Let $K_i = S_e + C_{ep}(B + C_{BS}V_i)$ and $M_i = Q_iS_e + C_{ep}(BQ_i + C_{BS}G_i)$, then the Fisher scores for B and S are

$$\begin{aligned} \sum_{i=1}^n & \frac{y_i(1-R_i)C_{ep}(1-V_i)}{K_i} - \frac{(1-y_i)(1-R_i)C_{ep}(1-V_i)}{1-K_i} + \frac{y_iR_iC_{ep}(Q_i-G_i)}{M_i} - \frac{(1-y_i)R_iC_{ep}(Q_i-G_i)}{Q_i-M_i}, \\ \sum_{i=1}^n & -\frac{y_i(1-R_i)C_{ep}V_i}{K_i} + \frac{(1-y_i)(1-R_i)C_{ep}V_i}{1-K_i} - \frac{y_iR_iC_{ep}G_i}{M_i} + \frac{(1-y_i)R_iC_{ep}G_i}{Q_i-M_i}. \end{aligned}$$

For $m = 1, 2, \dots, L$, let $\delta = p_mS_e + (1-p_m)(1-S_p)$, $T_m = C_{ep}(S_{im}\delta - W_{im})/(\delta(1-\delta))$, and $T_{gm} = C_{ep}C_{BS}S_{im}G_i/(1-p_m)$. Then the score for p_m is

$$\begin{aligned} \sum_{i=1}^n & -\frac{y_i(1-R_i)C_{ep}C_{BS}S_{im}V_i}{(1-p_m)K_i} + \frac{(1-y_i)(1-R_i)C_{ep}C_{BS}S_{im}V_i}{(1-p_m)(1-K_i)} \\ & + \frac{y_iR_i(Q_iT_m(S_e+C_{ep}B)-T_{gm})}{M_i} + \frac{(1-y_i)R_i(Q_iT_m(1-S_e-C_{ep}B)+T_{gm})}{Q_i-M_i}. \end{aligned}$$

The information matrix $I(\theta)$ can also be obtained directly from the Fisher scores.

The element for B and B is

$$\begin{aligned} \sum_{i=1}^n & \frac{y_i(1-R_i)C_{ep}^2(1-V_i)^2}{K_i^2} + \frac{(1-y_i)(1-R_i)C_{ep}^2(1-V_i)^2}{(1-K_i)^2} \\ & + \frac{y_iR_iC_{ep}^2(Q_i-G_i)^2}{M_i^2} + \frac{(1-y_i)R_iC_{ep}^2(Q_i-G_i)^2}{(Q_i-M_i)^2}. \end{aligned}$$

The element for S and S is

$$\begin{aligned} \sum_{i=1}^n & \frac{y_i(1-R_i)C_{ep}^2V_i^2}{K_i^2} + \frac{(1-y_i)(1-R_i)C_{ep}^2V_i^2}{(1-K_i)^2} \\ & + \frac{y_iR_iC_{ep}^2G_i^2}{M_i^2} + \frac{(1-y_i)R_iC_{ep}^2G_i^2}{(Q_i-M_i)^2}. \end{aligned}$$

The element for B and S is

$$\begin{aligned} \sum_{i=1}^n & -\frac{y_i(1-R_i)C_{ep}^2V_i(1-V_i)}{K_i^2} - \frac{(1-y_i)(1-R_i)C_{ep}^2V_i(1-V_i)}{(1-K_i)^2} \\ & -\frac{y_iR_iC_{ep}^2G_i(Q_i-G_i)}{M_i^2} - \frac{(1-y_i)R_iC_{ep}^2G_i(Q_i-G_i)}{(Q_i-M_i)^2}. \end{aligned}$$

The element for B and p_m , $m = 1, 2, \dots, L$ is

$$\begin{aligned} \sum_{i=1}^n & -\frac{y_i(1-R_i)C_{ep}^2C_{BS}S_{im}V_i(1-V_i)}{(1-p_m)K_i^2} - \frac{y_i(1-R_i)C_{ep}S_{im}V_i}{(1-p_m)K_i} \\ & -\frac{(1-y_i)(1-R_i)C_{ep}^2C_{BS}S_{im}V_i(1-V_i)}{(1-p_m)(1-K_i)^2} + \frac{(1-y_i)(1-R_i)C_{ep}S_{im}V_i}{(1-p_m)(1-K_i)} \\ & + \frac{y_iR_i(Q_iT_m(S_e+C_{ep}B)-T_{gm})C_{ep}(Q_i-G_i)}{M_i^2} - \frac{y_iR_i(Q_iC_{ep}T_m+T_{gm}/C_{BS})}{M_i} \\ & -\frac{(1-y_i)R_i(Q_iT_m(1-S_e-C_{ep}B)+T_{gm})C_{ep}(Q_i-G_i)}{(Q_i-M_i)^2} + \frac{(1-y_i)R_i(Q_iC_{ep}T_m+T_{gm}/C_{BS})}{Q_i-M_i}. \end{aligned}$$

The element for S and p_m , $m = 1, 2, \dots, L$ is

$$\begin{aligned} \sum_{i=1}^n & \frac{y_i(1-R_i)C_{ep}^2C_{BS}S_{im}V_i^2}{(1-p_m)K_i^2} - \frac{y_i(1-R_i)C_{ep}S_{im}V_i}{(1-p_m)K_i} \\ & + \frac{(1-y_i)(1-R_i)C_{ep}^2C_{BS}S_{im}V_i^2}{(1-p_m)(1-K_i)^2} + \frac{(1-y_i)(1-R_i)C_{ep}S_{im}V_i}{(1-p_m)(1-K_i)} \\ & -\frac{y_iR_i(Q_iT_m(S_e+C_{ep}B)-T_{gm})C_{ep}G_i}{M_i^2} - \frac{y_iR_iT_g}{C_{BS}M_i} \\ & + \frac{(1-y_i)R_i(Q_iT_m(1-S_e-C_{ep}B)+T_{gm})C_{ep}G_i}{(Q_i-M_i)^2} + \frac{(1-y_i)R_iT_g}{C_{BS}(Q_i-M_i)}. \end{aligned}$$

The element for p_m and p_m , $m = 1, 2, \dots, L$ is

$$\begin{aligned} \sum_{i=1}^n & \frac{y_i(1-R_i)C_{ep}^2 C_{BS}^2 S_{im}^2 V_i^2}{(1-p_m)^2 K_i^2} - \frac{y_i(1-R_i)C_{ep} C_{BS} S_{im} (S_{im}-1) V_i}{(1-p_m)^2 K_i} \\ & + \frac{(1-y_i)(1-R_i)C_{ep}^2 C_{BS}^2 S_{im}^2 V_i^2}{(1-p_m)^2 (1-K_i)^2} + \frac{(1-y_i)(1-R_i)C_{ep} C_{BS} S_{im} (S_{im}-1) V_i}{(1-p_m)^2 (1-K_i)} \\ & + \frac{y_i R_i (Q_i T_m (S_e + C_{ep} B) - T_{gm})^2}{M_i^2} + \frac{(1-y_i) R_i (Q_i T_m (1-S_e - C_{ep} B) + T_{gm})^2}{(Q_i - M_i)^2} \\ & - \frac{y_i R_i ((Q_i T_m^2 - Q_i C_{ep}^2 (W_{im}/\delta^2 + (S_{im} - W_{im})/(1-\delta)^2)) (S_e + C_{ep} B) + T_{gm} (S_{im}-1)/(1-p_m))}{M_i} \\ & - \frac{(1-y_i) R_i ((Q_i T_m^2 - Q_i C_{ep}^2 (W_{im}/\delta^2 + (S_{im} - W_{im})/(1-\delta)^2)) (1-S_e - C_{ep} B) - T_{gm} (S_{im}-1)/(1-p_m))}{Q_i - M_i}. \end{aligned}$$

The element for p_m and p_u ($m = 1, 2, \dots, L$, $u = 1, 2, \dots, L$, $m \neq u$) is

$$\begin{aligned} \sum_{i=1}^n & \frac{y_i(1-R_i)C_{ep}^2 C_{BS}^2 S_{im} S_{iu} V_i^2}{(1-p_m)(1-p_u) K_i^2} - \frac{y_i(1-R_i)C_{ep} C_{BS} S_{im} S_{iu} V_i}{(1-p_m)(1-p_u) K_i} \\ & + \frac{(1-y_i)(1-R_i)C_{ep}^2 C_{BS}^2 S_{im} S_{iu} V_i^2}{(1-p_m)(1-p_u)(1-K_i)^2} + \frac{(1-y_i)(1-R_i)C_{ep} C_{BS} S_{im} S_{iu} V_i}{(1-p_m)(1-p_u)(1-K_i)} \\ & + \frac{y_i R_i (Q_i T_m (S_e + C_{ep} B) - T_{gm})(Q_i T_u (S_e + C_{ep} B) - T_{gu})}{M_i^2} \\ & - \frac{y_i R_i (Q_i T_m T_u (S_e + C_{ep} B) + T_{gm} S_{iu}/(1-p_u))}{M_i} \\ & + \frac{(1-y_i) R_i (Q_i T_m (1-S_e - C_{ep} B) + T_{gm})(Q_i T_u (1-S_e - C_{ep} B) + T_{gu})}{(Q_i - M_i)^2} \\ & - \frac{(1-y_i) R_i (Q_i T_m T_u (1-S_e - C_{ep} B) - T_{gm} S_{iu}/(1-p_u))}{(Q_i - M_i)}. \end{aligned}$$

Applying this likelihood model to the same data set as described in Section 5.4, we obtain estimates $\hat{B} = 0.144$ and $\hat{S} = 0.014$ with standard errors 0.376 and 0.014 respectively. The estimated \hat{p} s of this likelihood model are listed in Table 5.8 and compared to the semiparametric and likelihood models from Section 5.3. The results for the two previous models are copied from Table 5.6. It can be seen that for this specific data set, the variance from both MLEs are comparable and are worse than for the semiparametric estimators except for \hat{p}_6 .

As in Section 5.4, the same 100 randomly selected retesting data sets are used to make comparisons between the semiparametric and likelihood estimators. Table 5.9 shows the average estimates and standard deviations for the three methods. The results for the semiparametric and the first likelihood model are copied from Table 5.7. It can be seen that, except for \hat{p}_1 , the semiparametric method gives more efficient estimators than both types of maximum likelihood estimators. As mentioned above, semiparametric estimators are more biased, we think because of the small sample size. The results also tell us that both likelihood methods are very comparable.

Table 5.8: *Comparison of Estimations and Standard Deviations for Three Methods. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)*

Parameter	Observed	Counts	MLE	Semiparametric	MLE2
p_1	.004	2/509	.003 (.003)	.002 (.002)	.003 (.003)
p_2	.200	1/5	.337 (.313)	.318 (.209)	.245 (.263)
p_3	.182	2/11	.237 (.151)	.163 (.112)	.204 (.145)
p_4	.039	16/412	.036 (.012)	.026 (.008)	.030 (.015)
p_5	.222	6/27	.256 (.107)	.245 (.083)	.214 (.119)
p_6	.800	4/5	.795 (.184)	.733 (.198)	.796 (.183)
p_7	.290	9/31	.281 (.109)	.299 (.083)	.220 (.136)

Table 5.9: *Average Results for Semiparametric and Two Likelihood Methods. (Counts are given in the form of potent/total for each class; Std. dev.s are given in brackets after estimates)*

Parameter	Observed	Counts	MLE	Semiparametric	MLE2
p_1	.004	2/509	.003 (.003)	.019 (.005)	.003 (.003)
p_2	.200	1/5	.291 (.351)	.305 (.198)	.271 (.392)
p_3	.182	2/11	.243 (.163)	.213 (.122)	.229 (.169)
p_4	.039	16/412	.035 (.011)	.046 (.010)	.033 (.015)
p_5	.222	6/27	.254 (.102)	.270 (.081)	.240 (.113)
p_6	.800	4/5	.811 (.217)	.748 (.191)	.811 (.219)
p_7	.290	9/31	.292 (.108)	.305 (.078)	.275 (.133)

Estimating the effect of blocking and synergism does not necessarily sacrifice the efficiency of estimating p in the likelihood models. The averaged estimates for B and S are $\hat{B} = 0.228$ and $\hat{S} = 0.013$, with standard deviations 0.320 and 0.073 respectively.

5.6 Future Work

Semiparametric models can be developed assuming the blocking effect B and synergism effect S are unknown and the measurement errors are known, in a manner similar to the way the likelihood model was developed in Section 5.5.

Potencies are typically measured on a continuous scale, but in this dissertation we dichotomize these continuous responses into binary indicators using either individual or pool thresholds. In order to recover information lost by dichotomization, continuous responses can be handled using semiparametric models. Benefitting from the double robustness of semiparametric estimators for data missing at random, estimators that are consistent and asymptotically normal could be constructed even when the first moment model is incorrect, given that the missing data model depends on the observed pool responses. Specifically, consider the model

$$Y = M(X, \beta) + \varepsilon,$$

where $E(\varepsilon | x) = 0$. It is very easy to fit a wrong model $M(X, \beta)$ in the investigation of the structure activity relationship (SAR) for chemical compounds in drug discovery. In addition, the distribution of the error ε makes the problem more complicated. However, semiparametric models allow greater freedom to deal with ε 's distribution.

On the continuous scale, it is important to model the pool potency depending on the individual potencies within that pool. For example, before considering the effect of blocking and synergism or measurement error, we could assume that $Y_i = \frac{1}{k} \sum_{j=1}^k Y_{ij}$ or $Y_i = \max(Y_{i1}, Y_{i2}, \dots, Y_{ik})$, and many other possible connections. However, it is not clear how to correctly determine this dependence. More research is needed in this area.

Bibliography

- [1] Begun, J.; Hall, W.; Huang, W.; Wellner, J. Information and Asymptotic Efficiency in Parametric-nonparametric Models. *Annals of Statistics* **1983**, *11*, 432-452.
- [2] Bickel, P. On Adaptive Estimation. *Annals of Statistics* **1982**, *10*, 647-671.
- [3] Bickel, P.; Klaassen, C. A. J.; Ritov, Y.; Wellner, J. A. Efficient and Adaptive Inference in Semiparametric Models. *Baltimore: Johns Hopkins University Press* **1993**.
- [4] Newey, W. K. Semiparametric Efficiency Bounds. *Journal of Applied Econometrics* **1990**, *5*, 99-135.
- [5] Pfanzagl, J.; Wefelmeyer, W. Contributions to a General Asymptotic Statistical Theory *Springer-Verlag, New York* **1982**.
- [6] Ritov, Y.; Bickel, P. J. Achieving Information Bounds in Non and Semiparametric Models. *Technical Report no. 116, Department of Statistics, University of California, Berkeley* **1987**.
- [7] Robins, J. M.; Rotnitzky, A.; Zhao, L. P. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* **1994**, *89*, 846-866.
- [8] Rubin, D. B. Inference and Missing Data. *Biometrika* **1976**, *63*, 581-592.
- [9] Tsiatis, A. Semiparametric Theory and Missing Data Problem *Lecture notes, Department of Statistics, North Carolina State University* **2001**.

- [10] Zhu, L.; Hughes-Oliver, J. M.; Young, S. S. Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics* **2001**, *57*, 922-930.

APPENDIXES

In this appendix, it will be shown that

$$\text{var}[A(f(x_i, \beta_0) - M^*(x_i, \beta_0))] = A \text{diag}(M^*(x_i, \beta_0)) \text{diag}(1_k - M^*(x_i, \beta_0)) A^T,$$

where $f(x_i, \beta_0) = \frac{R_i}{\pi(Y_i)} Y_i^* - \frac{R_i - \pi(Y_i)}{\pi(Y_i)} E(Y_i^* | y_i, x_i)$ and 1_k is a k -dimensional vector with all elements equal to 1.

Proof: Let $U = f(x_i, \beta_0) - M^*(x_i, \beta_0)$, and $M = M^*(x_i, \beta_0)$ then

$$\text{var}[A(f(x_i, \beta_0) - M)] = \text{var}[E(AU | y_i, x_i)] + E[\text{var}(AU | y_i, x_i)].$$

The above two terms are computed in the following two steps, respectively.

(1) As

$$\begin{aligned} E(AU | y_i, x_i) &= A[E(Y_i^* | y_i, x_i) - M] \\ &= A(W_i - 1)M, \end{aligned}$$

where $W_i = \frac{S_e^{y_i}(1-S_e)^{1-y_i}}{H_i^{y_i}(1-H_i)^{1-y_i}}$ and $H_i = S_e + (1 - S_e - S_p) \prod_{j=1}^k (1 + \exp(x_{ij}\beta))^{-1}$, it thus holds (x_i is not considered as random variables)

$$\begin{aligned} \text{var}[E(AU | y_i, x_i)] &= AMM^T A^T \text{var}(W_i - 1) \\ &= \frac{(S_e - H_i)^2}{H_i(1-H_i)} AMM^T A^T. \end{aligned}$$

(2) In order to evaluate $E[\text{var}(AU | y_i, x_i)]$, let's compute $\text{var}(U | y_i, x_i)$ first.

It is already known in the first step that

$$E(U | y_i, x_i) = (W_i - 1)M,$$

and moreover, through direct computation, we have

$$E(UU^T | y_i, x_i) = W_i \text{diag}(M) \text{diag}(1_k - M) - (W_i - 1)MM^T.$$

Diagonal matrix $diag(M)$'s elements are all zero except the diagonal elements equal to vector M . Therefore,

$$\begin{aligned} var(U | y_i, x_i) &= E(UU^T | y_i, x_i) - E(U | y_i, x_i)E(U^T | y_i, x_i) \\ &= W_i diag(M)diag(1_k - M) - (W_i - 1)MM^T - (W_i - 1)^2MM^T. \end{aligned}$$

It could also be verified that $E(W_i) = 1$, and $var(W_i) = \frac{(S_e - H_i)^2}{H_i(1 - H_i)}$. So, we obtain

$$E[var(U | y_i, x_i)] = diag(M)diag(1_k - M) - \frac{(S_e - H_i)^2}{H_i(1 - H_i)}MM^T,$$

and,

$$E[var(AU | y_i, x_i)] = A \left[diag(M)diag(1_k - M) - \frac{(S_e - H_i)^2}{H_i(1 - H_i)}MM^T \right] A^T.$$

Consequently, adding up the results from the above two steps, it holds

$$var[A(f - M)] = A diag(M)diag(1_k - M)A^T.$$

Bibliography

- [1] Begun, J.; Hall, W.; Huang, W.; Wellner, J. Information and Asymptotic Efficiency in Parametric-nonparametric Models. *Annals of Statistics* **1983**, *11*, 432-452.
- [2] Bickel, P. On Adaptive Estimation. *Annals of Statistics* **1982**, *10*, 647-671.
- [3] Bickel, P.; Klaassen, C. A. J.; Ritov, Y.; Wellner, J. A. Efficient and Adaptive Inference in Semiparametric Models. *Baltimore: Johns Hopkins University Press* **1993**.
- [4] Bockenholt, Ulf A Latent-Class Regression Approach for the Analysis of Recurrent Choice Data. *British Journal of Mathematical and Statistical Psychology*, **1993**, *46*, 95-118.
- [5] Brookmeyer, R. Analysis of Multistage Pooling Studies of Biological Specimens for Estimating Disease Incidence and Prevalence. *Biometrics* **1999**, *55*, 608-612.
- [6] Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225-227.
- [7] Cahoon-Young, B.; Chandler, A.; Livermore, T.; Gaudino, J.; Benjamin, R. Sensitivity and Specificity of Pooled versus Individual Sera in a Human Immunodeficiency Virus Antibody Prevalence Study. *Journal of Clinical Microbiology* **1989**, *27*, 1893-1895.
- [8] Carhart, R. E.; Smith, D. H.; Venkataraghavan R. Atom Pairs as Molecular Features in Structure-activity studies: Definition and Applications. *J. Chem.*

- Inf. Comput. Sci.* **1985**, *25*, 64-73.
- [9] Clogg, C. C.; Goodman, L. A. Latent Structure Analysis of a Set of Multidimensional Contingency Tables. *Journal of the American Statistical Association*, **1984**, *79*, 762-771.
- [10] Clogg, C. C.; Goodman, L. A. On Scaling Models Applied to Data From Several Groups. *Psychometrika*, **1986**, *51*, 123-135.
- [11] Dayton, C. M.; Macready, G. B. Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association*, **1988**, *83*, 401, 173-178.
- [12] Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood Estimation from Incomplete Data via the EM-algorithm. *Journal of the Royal Statistical Society, Series B* **1977**, *39*, 1-22.
- [13] Dorfman, R. The Detection of Defective Members of Large Populations. *Annals of Mathematical Statistics* **1943**, *14*, 436-440.
- [14] Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *Journal of Molecular Graphics and Modelling* **2000**, *18*, 343-357.
- [15] Engels M. F.; Venkatarangan P. Smart screening: Approaches to efficient HTS. *Current Opinion in Drug Discovery & Development* **2001**, *4(3)*, 275-283.
- [16] Farrington, C. Estimating Prevalence by Group Testing Using Generalized Linear Models. *Statistics in Medicine* **1992**, *11*, 1591-1597.
- [17] Gastwirth, J. L.; Hammick, P. A. Estimation of the Prevalence of a Rare Disease, Preserving the Anonymity of the Subjects by Group Testing: Application to Estimating the Prevalence of Aids Antibodies in Blood Donors. *Journal of Statistical Planning and Inference* **1989**, *22*, 15-27.
- [18] Goodman, L. A. Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, **1974**, *61*, 215-231.
- [19] Haberman, S. J. Log-Linear Model for Frequency Tables Derived by Indirect Observation: Maximum-Likelihood Equations. *The Annals of Statistics* **1974**,

2, 911-924.

- [20] Haberman, S. J. Analysis of Qualitative Data, Volume 2: New Developments. *New York: Academic Press* **1979**.
- [21] Hammick, P. A.; Gastwirth, J. L. Group Testing for Sensitive Characteristics: Extension to Higher Prevalence Levels. *Internat. Statist. Assoc.* **1994**, *62*, 319-331.
- [22] Hawkins, D. M.; Young, S. S.; Rusinko A. Analysis of a Large Structure-activity Data Set Using Recursive Partitioning. *Quantitative Structure-Activity Relationship* **1997**, *16*, 296-302.
- [23] Hughes-Oliver, J. M.; Swallow, W. H. Choosing the Group Size for Group Testing to Estimate a Proportion. *Institute of Statistics Mimeograph Series No. 2209, North Carolina State University* **1992**.
- [24] Hughes-Oliver, J. M.; Swallow, W. H. A Two-Stage Adaptive Group-Testing Procedure for Estimating Small Proportions. *Journal of the American Statistical Association* **1994**, *89*, 982-993.
- [25] Hughes-Oliver, J. M.; Rosenberger, W. F. Efficient Estimation of the Prevalence of Multiply Rare Traits. *Biometrika* **2000**, *87*, 315-327.
- [26] Hung, M. Advantages of Group Testing with Stratified or Regression Data, and Its Robustness in Estimation of Proportions. *Ph.D dissertation*, **1993**.
- [27] Hung, M.; Swallow, W. H. Robustness of Group Testing in the Estimation of Proportions. *Boimetrics* **1999**, *55*, 231-237.
- [28] Hung, M.; Swallow, W. H. Use of Binomial Group Testing in Tests of Hypotheses for Classification or Quantitative Covariables. *Boimetrics* **2000**, *56*, 204-212.
- [29] Hwang, F. K. Group Tesing with a Dilution Effect. *Biometrika* **1976**, *63*, 671-673.
- [30] Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118-127.

- [31] Kline, R. L.; Brothers, T. A.; Brookmeyer, R.; Zeger, S.; Quinn, T. C. Evaluation of Human Immunodeficiency Virus Seroprevalence in Population Surveys Using Pooled Data. *Journal of Clinical Microbiology* **1989**, *27*, 1449-1452.
- [32] Lam, R. L. H. Design and Analysis of Large Chemical Databases for Drug Discovery *University of Waterloo* Ph.D. Dissertation **2001**.
- [33] Lam, R. L. H.; Welch W. J.; Young, S. S. Cell-Based Analysis for Large Chemical Databases. *Technometrics* **2002**, submitted.
- [34] Lam, R. L. H.; Welch W. J.; Young, S. S. Uniform Coverage Designs for Molecule Selection. *Technometrics*, **2002**, *44*, 99-109.
- [35] Langfeldt, S. A.; Hughes-Oliver, H. M.; Ghosh, S.; Young, S. S. Optimal Group Testing in the Presence of Blockers. *Institute of Statistics Mimeograph Series No. 2297*, **1997**
- [36] Lazarsfeld, P. F.; Henry, N. W. Latent Structure Analysis. *Boston: Houghton Mifflin* **1968**.
- [37] McFarland, J. W.; Gans, D. J. On the Significance of Clusters in the Graphical Display of Structure-Activity Data. *Journal of Medicinal Chemistry* **1986**, *29*, 505-514.
- [38] Newcomb, S. A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. *American Journal of Mathematics* **1886**, *8*, 343-366.
- [39] Newey, W. K. Semiparametric Efficiency Bounds. *Journal of Applied Econometrics* **1990**, *5*, 99-135.
- [40] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.
- [41] Pearlman R. S.; Smith K. M. Novel Software tools for chemical diversity. *Perspectives in Drug Discovery and Design* **1998**, *9-11*, 339-353.
- [42] Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28-35.

- [43] Pearson, K. Contributions to the Mathematical Theory of Evolution. *Philosophical Trans.* **1894**, A 185, 71-110.
- [44] Pfanzagl, J.; Wefelmeyer, W. Contributions to a General Asymptotic Statistical Theory *Springer-Verlag, New York* **1982**.
- [45] Phatarfod, R. M.; Sudbury, A. The Use of a Square-Array Scheme in Blood Testing. *Statistics in Medicine*, **1994**, 13, 2337-2343.
- [46] Ritov, Y.; Bickel, P. J. Achieving Information Bounds in Non and Semiparametric Models. *Technical Report no. 116, Department of Statistics, University of California, Berkeley* **1987**.
- [47] Robins, J. M.; Rotnitzky, A.; Zhao, L. P. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* **1994**, 89, 846-866.
- [48] Rubin, D. B. Inference and Missing Data. *Biometrika* **1976**, 63, 581-592.
- [49] Rusinko A.; Farnen M. W.; Lambert C. G.; Brown P. L.; Young S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, 38, 1017-1026.
- [50] Sobel, M.; Elashoff, R. M. Group Testing With a New Goal, Estimation. *Biometrika*, **1975**, 62, 181-193.
- [51] Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, 39(1), 11-20.
- [52] Swallow, W. H. Group Testing for Estimating Infection Rates and Probability of Disease Transmission. *Phytopathology* **1985**, 75, 882-889.
- [53] Swallow, W. H. Relative Mean Squared Error and Cost Considerations in Choosing Group Size for Group Testing to Estimate Infection Rates and Probabilities of Disease Transmission. *Phytopathology* **1987**, 77, 1376-1381.
- [54] Thompson, K. H. Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics*, **1962**, 18, 568-578.

- [55] Tsiatis, A. Semiparametric Theory and Missing Data Problem. *Lecture notes, Department of Statistics, North Carolina State University* **2001**.
- [56] Tu, X. M.; Litvak, E.; Pagano, M. Studies of AIDS and HIV surveillance. Screening tests: Can we get more by doing less? *Statistics in Medicine*, **1994**, *13*, 1905-1919.
- [57] Tu, X. M.; Litvak, E.; Pagano, M. On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare Disease: Application to HIV Screening. *Biometrika*, **1995**, *82*, 287-298.
- [58] Vansteelandt, S.; Goetghebeur, E.; Verstraeten, T. Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples. *Biometrics* **2000**, *56*, 1126-1133.
- [59] Xie, M; Tatsuoka, K.; Sacks, J.; Young S. S., Group Testing Scheme in Cases of Blockers and Synergism. *Journal of the American Statistical Association*, **2001**, *96*, 92-102.
- [60] Young, S. S.; Hawkins, D. M. Using Recursive Partitioning to Analyze a Large SAR Data Set. *SAR and QSAR in Environmental Research* **1998**, *8*, 183-193.
- [61] Zhu, L.; Hughes-Oliver, J. M.; Young, S. S. Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics* **2001**, *57*, 922-930.