

**STATISTICAL METHODS FOR PATIENT
CHEMOSENSITIVITY PREDICTION BASED ON
IN VITRO DOSE-RESPONSE DATA AND A
MODIFIED EXPECTATION-MAXIMIZATION (EM)
ALGORITHM FOR REGRESSION ANALYSIS OF
DATA WITH NON-IGNORABLE NON-RESPONSE**

by

Yang Zhang

BS, Nanjing University, China, 2006

MS, Northeastern University, 2008

Submitted to the Graduate Faculty of
the Graduate School of Public Health in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yang Zhang

It was defended on

November 22, 2013

and approved by

Gong Tang, PhD, Associate Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Joseph P. Costantino, DrPH, Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Joyce Chang, PhD, Associate Professor

Division of General Internal Medicine, School of Medicine, University of Pittsburgh

Priya Rastogi, MD, Associate Professor

Division of Hematology/Oncology, School of Medicine, University of Pittsburgh

Shuguang Huang, PhD, Associate Vice President

Department of Statistics, Precision Therapeutics Inc., Pittsburgh, PA

Dissertation Director: Gong Tang, PhD, Associate Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Copyright © by Yang Zhang
2013

**STATISTICAL METHODS FOR PATIENT CHEMOSENSITIVITY
PREDICTION BASED ON IN VITRO DOSE-RESPONSE DATA AND A
MODIFIED EXPECTATION-MAXIMIZATION (EM) ALGORITHM FOR
REGRESSION ANALYSIS OF DATA WITH NON-IGNORABLE
NON-RESPONSE**

Yang Zhang, PhD

University of Pittsburgh, 2013

ABSTRACT

The first part of this dissertation concerns statistical analysis of *in vitro* assay data derived from tumor cells. An *in vitro* assay, ChemoFx[®], has been developed to predict patient's clinical chemosensitivity. We explore statistical methods that can more efficiently use the assay data and improve the prediction of clinical outcome. In typical analysis of assay dose-response data, summary statistics such as the area under the dose-response curve, and concentration at half inhibition (IC₅₀) are estimated from the assay data, then these statistics are dichotomized to predict the clinical outcome as sensitive or resistant. Considering the rigidness of the traditional models for dose-response curve fitting and the information loss in use of cell counts in the control wells, here we propose a mixture of exponential functions for fitting the dose-response curve and a branching process-based method to summarize the control well data. Simulation studies and analysis of clinical trial data show that the proposed method improves the prediction performance over some traditional methods. The second part concerns statistical analysis of regression data with nonresponses. Missing data are prevalent in clinical trials and public health studies. The often unknown mechanism for the missing data process may actually be associated with the underlying

values. Standard statistical methods, including likelihood-based methods and weighted estimating equations, require a model for the missing-data mechanism and incorporate it in the estimation and inference. Misspecification of the missing-data model often causes biased estimates and wrongful conclusions. The expectation-maximization (EM) algorithm is an iterative algorithm that is often used to find the maximum likelihood estimate for the likelihood-based methods. In the E-steps, given a current estimate and the missing-data mechanism, the conditional expectations of the sufficient statistics are calculated. Under the premise that the current estimate is consistent, we find that those conditional expectations could be approximated from the empirical data without the need for assuming or modeling the missing-data mechanism. Subsequently, we propose a modified EM algorithm regardless of the potential missing-data mechanism. Simulation studies show that the parameter estimates have negligible bias and are more efficient than the initial estimates obtained from external data.

Keywords: Dose-response curve; Curve fitting; Branching process; Analysis of missing data; Non-ignorable nonresponse; EM algorithm.

TABLE OF CONTENTS

PREFACE	xi
1.0 INTRODUCTION	1
1.1 Classification of assay dose-response data	1
1.2 Statistical analysis of missing data	3
2.0 STATISTICAL METHODS FOR PATIENT CHEMOSENSITIVITY PREDICTION BASED ON IN VITRO DOSE-RESPONSE DATA	5
2.1 Overview of analysis of assay dose-response data	5
2.1.1 Cancer biomarkers and ChemoFx [®] assay	5
2.1.2 Review of metrics for quantifying dose-response assay	7
2.2 The Proposed methods for classification based on dose-response data	13
2.2.1 Five-parameter mixture of exponential function (5ME) method	13
2.2.2 Branching process using dose-0 cell counts data	16
2.3 Simulation studies	20
2.3.1 Comparing 5ME, IC50, and AUC using simulated cell counts data	20
2.3.2 Comparing 5ME, IC50, and AUC using simulated dose-response curves	22
2.4 Application to clinical trial data	25
2.4.1 Analysis of B40 trial data	25
2.4.2 Analysis of PTI-206 ChemoFx [®] assay data	30
2.5 Discussions	33
3.0 A MODIFIED EM ALGORITHM FOR REGRESSION ANALYSIS OF DATA WITH NON-IGNORABLE NON-RESPONSE	35
3.1 Overview of analysis of missing data	35

3.1.1	Missing data in practice	35
3.1.2	Statistical methods for analysis of missing data	37
3.2	EM algorithm for regression analysis of data with non-responses when the missing-data mechanism is modeled	43
3.3	A modified EM algorithm for regression analysis of data with non-responses	45
3.3.1	EM algorithm for regression analysis of data with non-responses when missing-data mechanism is known	46
3.3.2	An modified EM algorithm	47
3.3.3	Implementation of proposed methods in simple linear regression	49
3.3.4	Improvement of modified EM algorithm by involving external data	51
3.3.5	The modified EM algorithm under discrete covariates	52
3.3.6	Smoothing methods and its implementation in the modified EM algorithm	52
3.4	Simulation studies	54
3.4.1	Simulation setup	54
3.4.2	Simulation results	56
3.5	Analysis of qualify-of-life data from a cancer clinical trial	59
3.6	Summary and discussion	60
BIBLIOGRAPHY		64

LIST OF TABLES

1	Comparison of 5ME, IC50 and AUC based on simulated cell counts	22
2	Comparing 5ME, IC50 and AUC using simulated dose-response data	23
3	Comparison of methods using B40 ChemoFx assay data	29
4	Comparison of methods by AUROC based on PTI-206 assay data	31
5	Simulation results when the predictor $X \sim BIN(5, 0.3)$	57
6	Simulation results when the predictor $X \sim N(0, 1)$	58
7	Data analysis: patients' QOL in R-04 trial.	62

LIST OF FIGURES

1	Layout of a typical 384 (16 × 24) - well plate in the ChemoFx [®] Assay	7
2	Dose response curves (DRCs) of the breast specimens treated with sunitinib. Responsive (R) specimens are denoted with a solid black line 3/39 (7.6%), intermediate responsive (IR) specimens are delineated with a dashed black line 8/39 (20.5%) and non-responsive (NR) specimens are indicated by a gray line 28/39 (71.7%)	8
3	Definition of IC50, relative IC50 and AUC on a DRC.	10
4	Four-parameter logistic function (4PL). 4PL with $(\beta_1, \beta_2, \beta_2, \beta_4) = (1, 0.2, 1, 5)$ is corresponding to the red curve, 4PL with $(\beta_1, \beta_2, \beta_2, \beta_4) = (1, 0.2, 3, 5)$ is corresponding to the green curve.	12
5	Mean of DRCs of the breast specimens treated with doxorubicin (A). Specimens from pathological complete response (pCR) groups are corresponding to the red line, specimens from non-pCR groups are corresponding to the black line.	14
6	Five-parameter mixture of exponential function (5ME). 5ME with $(a, b, c, d, \rho) = (0.01, 0.4, 0.05, 0.5, 0.8)$ is corresponding to the red curve, 5ME with $(a, b, c, d, \rho) = (0.01, 0.4, 0.05, 0.5, 0.2)$ is corresponding to the green curve.	15
7	Branching Process: free growth and drug effect	17
8	Mean DRCs of simulated cell counts data.	21
9	Mean of Simulated DRC following seven-parameter mixture of a 4PL function and an exponential function (7MEL) functions.	24
10	B-40 Schema	26

11	Sample dose-response curves in assay with drug A or T, grouped by pCR of patients in NSABP B-40 Arm 1A.	28
12	Mean dose-response curves(DRC's) in assay with drug A or T, grouped by pCR of patients in NSABP B-40 Arm 1A.	29
13	Comparison of methods by ROC analysis using PTI-206 assay data	32
14	Plots of simulated data (x,y) in different scenarios.	55
15	NSABP R-04 Trial: QOL before treatment vs QOL after treatment.	61

PREFACE

I would like to offer my sincere gratitude to Dr. Gong Tang, my advisor. In the past two and half years, he has supported me in research with his knowledge, enthusiasm, and commitment. Without his guidance and inspiration, the modified EM algorithm would not have been possible. Besides research support, he has also shared with me his wisdom and experience in life.

I would like to thank Dr. Shuguang Huang, my supervisor during my internship at Precision Therapeutics Inc. He has given me important support and advice in the study of dose-response-curve classification (Part I). It is the career experience he has shared with me that has guided me to industry.

I also thank Dr. Joyce Chang, my previous graduate student researcher (GSR) supervisor and Dr. Joseph Costantino, my current GSR supervisor. They have both provided invaluable guidance and have given me the flexibility to balance between my research and GSR work.

Last but not the least, I thank my parents, for giving me the opportunity of life, and their unconditional and consistent love.

1.0 INTRODUCTION

1.1 CLASSIFICATION OF ASSAY DOSE-RESPONSE DATA

ChemoFx[®] assay is an experimental tool that uses a patient's tumor chemoresponse *in vitro* to predict the patient's clinical response. This technology has been used in clinical practice for some time to help physicians in treatment decision [Ochs et al., 2005].

In vitro assays, e.g. ChemoFx, are usually conducted in a dose-response fashion, with tumor cells exposed to a series of a monotonically increasing dosages of the chemotherapeutic agent of interest[Brower et al., 2008]. The goal is to use the *in vitro* assay results to predict patients' clinical outcome such as clinical/pathological response, disease progression or survival. Typically, the analysis of assay data is involved with two stages: first, the dose-response data is reduced to some summary statistics, such as area under the dose-response curve (AUC), the drug concentration that achieves a target effect (e.g. IC50); these summary statistics are then used to predict the clinical outcome using some statistical methods (e.g., logistic regression model or Cox model).

The relationship between the *in vitro* and *in vivo* systems is usually complex. Traditional methods such as AUC (usually based on non-parametric approach, using trapezoidal rule) and IC50 (usually based on 4-parameter logistic regression (4PL)) sometimes can not well differentiate sensitive and resistant curves. Four parameter logistic model (4PL) -derived IC50 shows some relative advantage comparing with the summation of the last three responses on the dose-response curve [Huang and Pang, 2008], but the 4PL parametric curve fitting requires the compliance of the data. For instance, the observed data have to show

a sigmoidal shape as well as the lower and higher plateaus. In reality, though, the data do not always satisfy these requirements. Therefore, we propose a five-parameter mixture of exponential functional form (5ME) which is more flexible in modeling dose-response curves. Similar to the 4PL method, the fitted parameters of 5ME have explicit biological meanings and can be used as predictors in the second stage of patient classification.

Typically, the dose-response curve fitting (e.g. 4PL) for the purpose to estimate IC50 or the AUC analysis is based on the baseline-standardized data, where the data at each point (e.g. number of survival cell counts at each drug concentration) is standardized by the data at the baseline (e.g. cell count when the sample is not treated by drugs) so that the model is fitted on the relative activity. Obviously, the baseline information itself is not fully used in the curve fitting, and it is lost if only AUC or IC50 is used in the downstream analysis. Even if the baseline information is used, oftentimes people only consider the mean but ignore the variance of the cell counts from the different control wells. It has been proposed that the growth rate of tumor cells (without exposing to drug) may be reflective of some intrinsic biologic behaviors and therefore affects the patient prognosis. We propose a branching process method for possibly better utilizing the baseline data information (dose-0 cell counts). Here we simply assumed a cell splitting process following a Bernoulli distribution, in which the estimated Bernoulli probability that quantify the cell characters is then used for patient classification.

The performance of 5ME methods is compared with IC50 and AUC methods through simulation studies and the prediction of clinical outcome was evaluated based on the data from National Surgical Adjuvant Breast and Bowel Project (NSABP) B-40 trial and Precision Therapeutics Inc. (PTI) -206 trial.

1.2 STATISTICAL ANALYSIS OF MISSING DATA

There are many methods developed in the past few decades for statistical analysis of data with non-response. Based on how the joint distribution of the complete data and nonresponse indicator is factorized, these methods classified into selection models and pattern-mixture models. Selection models factor the likelihood into the distribution of the underlying complete data and the conditional distribution of the missing data indicator given the underlying complete data. Pattern-mixture models stratify the data by the patterns of missing values and model the distribution of data within each stratum.

For selection models, usually standard statistical methods such as maximum likelihood, weighted GEE and Bayesian methods require a correct model for the missing-data mechanism. Misspecification of the missing-data model often causes biased estimates and wrong conclusions. However, in practice, investigators do not have concrete knowledge about missing-data mechanism. With the selection model framework here we proposed a modified EM algorithm in maximum likelihood methods, assuming the distinction between the parameters for the complete data distribution and those for the missing data mechanism.

The expectation-maximization (EM) algorithm is an iterative algorithm that is often used to find the maximum likelihood estimate for the likelihood-based methods. In the E-steps, given a current estimate and a model for the missing-data mechanism, the conditional expectations of the sufficient statistics are calculated. Under the premise that the current estimate is consistent, we found that those conditional expectations could be approximated from the empirical data without the need for modeling the missing-data mechanism.

Therefore we can achieve inference of data distribution parameters using this modified EM algorithm regardless of the potential missing-data mechanism. The consistent initial value can be either obtained from an external complete dataset or complete recall on a subset.

The proposed modified EM algorithm method was illustrated in simulation studies and an analysis of quality of life data from a clinical trial. The simulation studies showed that the parameter estimates had negligible bias and were more efficient than the initial values obtained from external data.

2.0 STATISTICAL METHODS FOR PATIENT CHEMOSENSITIVITY PREDICTION BASED ON IN VITRO DOSE-RESPONSE DATA

2.1 OVERVIEW OF ANALYSIS OF ASSAY DOSE-RESPONSE DATA

2.1.1 Cancer biomarkers and ChemoFx[®] assay

Based on the definition provided by the National Cancer Institute (NCI) ¹:

Tumor is an abnormal mass of tissue that results when cells divide more than they should or do not die when they should.

A tumor marker is a substance found in tissue, blood, or other body fluids that may be a sign of cancer or certain benign (noncancerous) conditions. Most tumor markers are made by both normal cells and cancer cells, but they are made in larger amounts by cancer cells. A tumor marker may help to diagnose cancer, plan treatment, or find out how well treatment is working or if cancer has come back.

The classification of biomarkers does not have a general rule. From different aspect of knowledge [Mishra and Verma, 2010], we may have many choices of biomarker classification methods. We can classify biomarkers into prediction biomarkers, detection biomarkers, diagnosis biomarkers and prognosis biomarkers based on the disease state in which the biomarkers are utilized, or into DNA biomarkers, RNA biomarkers, protein biomarkers and carbohydrate biomarkers based on the biomolecules that have been used. Also for cancer biomarkers, it is

¹<http://www.cancer.gov/dictionary>

natural that we classify them by organ sites. For example, human fibroblast growth factor receptor 2 (HER2) is one of the recommended biomarkers for treatment planning for breast cancer patients. It is a predictive biomarker for breast cancer, and is usually measured by mRNA expression, so it is also an RNA biomarker. Other markers such as viral markers and imaging markers are also broadly used in cancer diagnosis and prognosis.[Yu et al., 1990, Abraham et al., 1996]

In vitro assay can be used as a predictive biomarker to assist clinicians in selecting the most suitable chemotherapeutic treatment for each cancer patient. Developed by Precision Therapeutics Inc. (PTI) ², the ChemoFx[®] assay is an *ex vivo* assay that measures the cancer patient's tumor responses to multiple chemotherapeutic agents simultaneously. The rationale of the ChemoFx[®] assay is based on the adherent character of monolayer-growing cells in culture that the cells die when they detach from the culture surface [Brower et al., 2008]. The anti-cancer agent's effect can be measured based on the survival fraction of tumor cells after treated for a fixed time - 72 hours. The assay results for different agents are compared within each patient and the most sensitive one(s) can be selected for individualized therapy.

The typical design of a ChemoFx[®] assay plate is shown in Figure 1. The assay plate consists of 384 (16 × 24) wells. In this assay experiment, the wells at the margin are not used to avoid edge effect. An assay is defined as a patient's cell treated by the standard of care (SOC) treatment options. Therefore, for each patient, the assay is consisted of a panel of treatments. Every two assays took three lines of wells with increasing dose from left to right. For example, the wells in line B-D, column 02-23 are used by two assays (one in gray color and one in white color), the dose or drug concentration is indexed numerically from 0-10. Note that for each assay, the same dose is applied repeatedly in three wells. As such, for each patient, there are three wells of cell counts for a drug concentration, except that for dose 0 there are $3 \times n$ of replicates, where n is the number of SOC treatment options applied to the cells.

²<http://www.precisiontherapeutics.com/>

At time 0, each well is placed with the same amount of cells, variation can occur due to technical variability, then different concentrations of chemo-drugs are applied to each well. After 72 hours of either free or prohibited growth, the cell counts in the wells are recorded. Then the mean of cell counts in the untreated wells (dose-0) for each patient is used as the baseline for that patient. The cell survival proportion (dose-response) from dose 1-10 for the patient is then derived by dividing each data point (mean cell counts at each concentration) by the baseline.

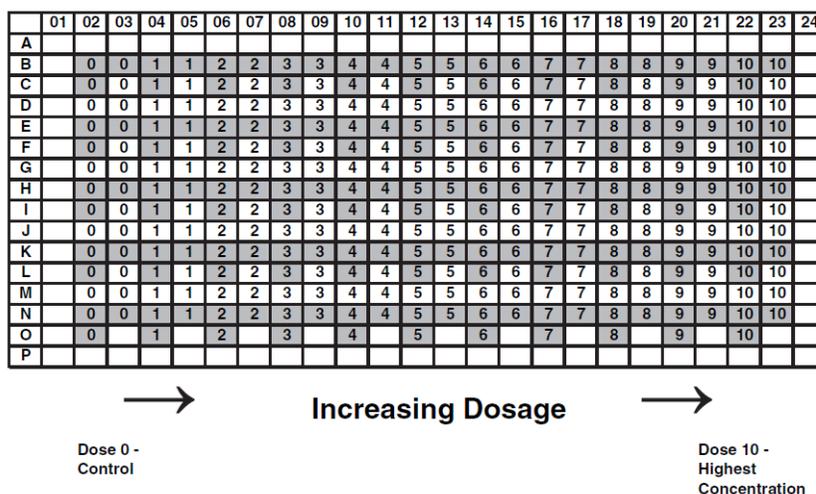


Figure 1: Layout of a typical 384 (16 × 24) - well plate in the ChemoFx[®] Assay

In actual practice, the tumor cells are acquired from fresh tissues collected at surgery. The cells are then cultured in the lab as mono-layers for several weeks. After the cell quality and quantity meet pre-defined standards, they will be transferred from flasks to the plates for the assay [Ochs et al., 2005.] to assess their chemo-sensitivity.

2.1.2 Review of metrics for quantifying dose-response assay

A dose-response curve (DRC) is an X-Y plot relating assay response (e.g., cell survival rate for ChemoFx) and corresponding dose [Huang and Pang, 2008]. Figure 2 provides an example of dose-response curves from ChemoFx assay [Suchy et al., 2011].

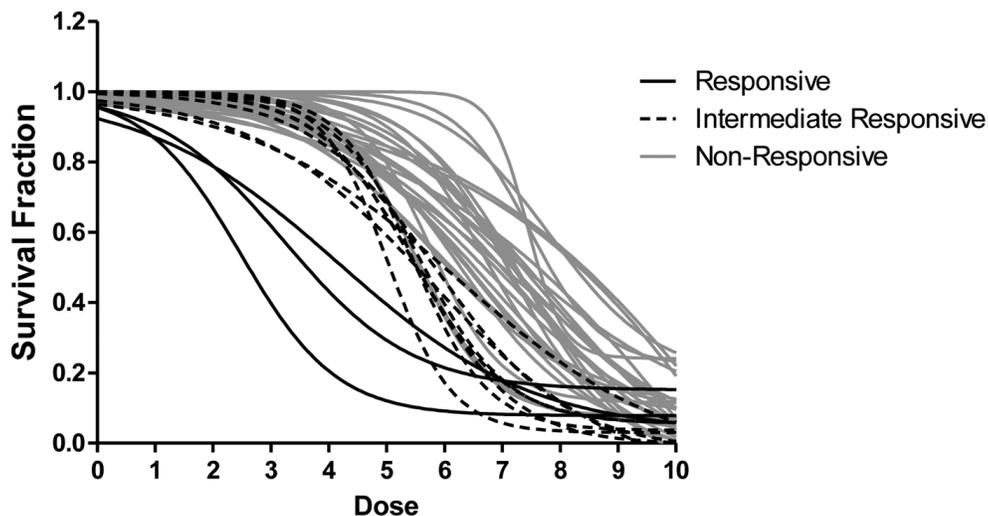


Figure 2: Dose response curves (DRCs) of the breast specimens treated with sunitinib. Responsive (R) specimens are denoted with a solid black line 3/39 (7.6%), intermediate responsive (IR) specimens are delineated with a dashed black line 8/39 (20.5%) and non-responsive (NR) specimens are indicated by a gray line 28/39 (71.7%)

For the baseline adjustment of the DRCs, the control well mean (CWM) for a patient is defined as the mean cell counts in the untreated wells:

$$CWM_i = \frac{1}{M} \sum_{m=1}^M X_{i,m}^0 \quad (2.1)$$

Where M represents number of dose-0 wells for each patient, and $X_{i,m}^0$ represents the cell counts from each dose-0 well. Since nine drugs are usually tested on one plate, with each measurements repeated three times, the maximum of M is 27 based on the assay design. But in practice M varies because of the possible failure of assays or insufficient cells for all of the treatment.

Dose response (DR) at dose k is then defined as the cell survival fraction using the mean of cell counts from different wells at dose- k divided by the CWM for a patient i :

$$DR_{i,j,k} = \frac{Y_i^k}{CWM_i}, k = 1, 2, \dots, 10. \quad (2.2)$$

where mean cell counts at dose- k is the average of the three reps, as $Y_i^k = \frac{1}{3} \sum_{p=1}^3 Y_{i,p}^k$.

A DRC describes the change of cell-killing percentages with the increasing drug concentrations for each treatment option. The goal is to assess the tumor cells in vitro sensitivity to each of the clinical treatment options, this is then used to predict the patient’s clinical response to the treatments and thus help physicians in making treatment decisions.

In practice, investigators usually use a two-stage approach for classification based on DRCs [Huang and Pang, 2008]. In the first stage, summary statistics are extracted from each DRC, with choices of methods including half-maximal inhibitory concentration (IC50), area under dose-response curves (AUC). In the second stage, the summary statistics extracted from the first stage and other baseline variables are used to predict clinical response.

The half-maximal inhibitory concentration (IC50) and area under dose response curves (AUC) are both common choices of traditional methods for extracting summary statistics from the dose-response data from assay experiments [Huang and Pang, 2008].

Absolute IC50 is defined as the concentration which generates half of the maximal effect - mid-point between the max effect and zero [Cheng, 2002]. Relative IC50 is similar to absolute IC50 except that the maximal effect is adjusted by a baseline (minimal effect)- the mid-point between the max effect and the min effect. The difference between absolute IC50 and relative IC50 can be demonstrated in Figure 3.

In *in vitro* dose-response assays, IC50 is the most widely used metric to analyze DRCs, and it is usually estimated based on a parametric approach. For a given drug, assays with a larger IC50 suggests that the tumor cells are more resistant to the drug.

The area under the DRC, a.k.a. AUC, is another common method for quantifying dose-response curves and used for curve comparisons [Mi et al., 2008]. It is defined as the sum of

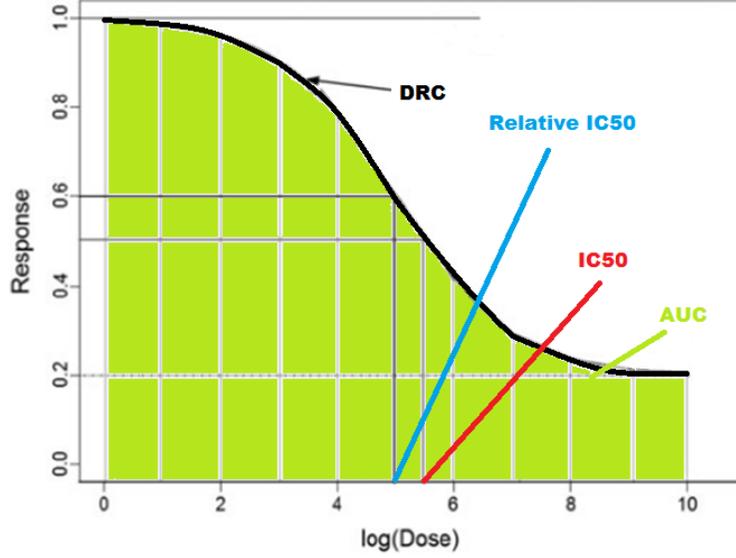


Figure 3: Definition of IC50, relative IC50 and AUC on a DRC.

DR at dose from 1-10:

$$AUC = \sum_{k=1}^{10} DR_k \quad (2.3)$$

AUC method is used to quantify the relative height of the DRCs. When comparing the DRCs of two patients treated with the same drug, the lower DRC suggests that the tumor cells of that patient is be more sensitive to the drug. The convenience of interpretation of AUC in predicting clinical response is a great advantage of the method.

Truncated or partial AUC is a modification of AUC method. The motivation for this approach is that during the model training it is found that not all of the concentrations are informative (correlated with) the clinical outcome. AUC7, for example, computes the AUC using DRs at dose from 1-7, considering that the growth of tumor cells under high concentration (dose 8-10) may not be distinguished for different drugs.

AUC and IC50 both reduce the dimension of the DR data from 10 to one (assuming that the triplicate measurements at each concentration are averaged first). This dimension

reduction may simplify the prediction model in the second stage. AUC and IC50 methods also have clear and intuitive interpretation.

When the CWMs change, IC50s will stay the same, because the calculation of IC50 depends on the relative change of dose-response. AUCs can be affected in that situation, and thus prediction of clinical response is also impacted. The information about the free growth of tumor cells itself is indicative of how aggressive the tumor cells are, and it should be in the prediction model.

As usual, there could be loss of information in the dimension reduction process when only IC50 and AUC are used to summarize the dose-response data. For example, as shown in Figure 4, two dose-response curves have the same AUCs and IC50s, but the chemo-sensitive kinetics might be totally different, because part of the information on the cell killing mechanism is not captured by AUC or IC50.

To address the above concern, the DRCs should be fitted with a function that is required to not only reduce the dimension of the data, but also describe the curve kinetics. Four-parameter logistic function (4PL) is widely used for fitting DRCs in industry. The 4PL function is defined as in [Huang and Pang, 2008]:

$$y = \beta_2 + \frac{\beta_1 - \beta_2}{1 + (\frac{\beta_4}{x})^{\beta_3}} \quad (2.4)$$

Here, β_2 denotes upper plateau, β_1 denotes lower plateau, β_3 denotes slope/speed of response to dosage change, and β_4 denotes IC50.

An alternative version of 4PL function could be used[Ritz and Streibig, 2005]

$$y = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{\beta_3(\beta_4 - x)}} \quad (2.5)$$

In this form of four parameter logistic model, β_1, β_2 and β_3 will be similar to that in formula (2.4); β_4 and x here will be similar to $\log(\beta_4)$ and $\log(x)$ in formula (2.4), considering the fact that $e^{\beta_3\beta_4}$ should be very large. In assay dose-response data analysis, we usually use

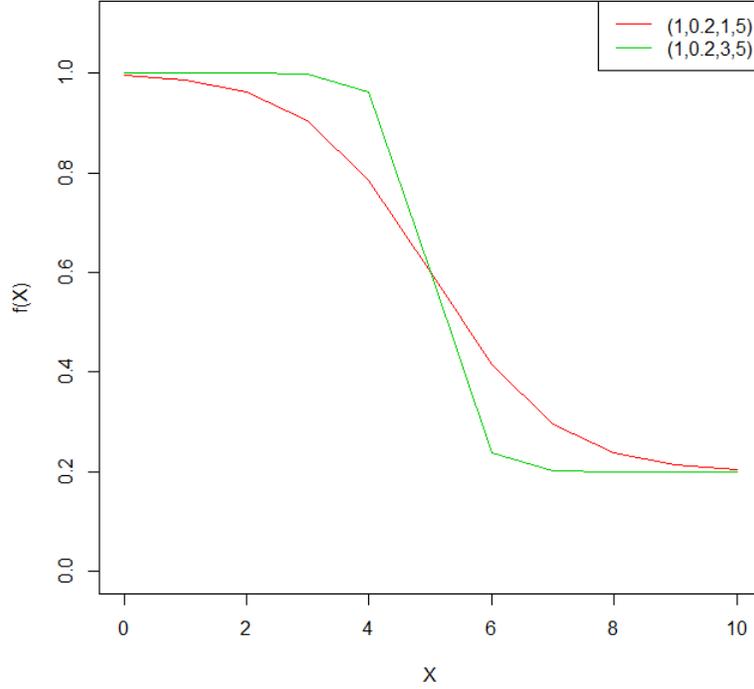


Figure 4: Four-parameter logistic function (4PL). 4PL with $(\beta_1, \beta_2, \beta_3, \beta_4) = (1, 0.2, 1, 5)$ is corresponding to the red curve, 4PL with $(\beta_1, \beta_2, \beta_3, \beta_4) = (1, 0.2, 3, 5)$ is corresponding to the green curve.

(2.5) taking into account that the dose k from 1-10 represent the drug concentrations exponentially increasing with base 10.

In Figure 4, the 4PL function of index dosage x from 0 to 10 are plotted under two setups of parameters. $\beta_4 = 5$ was set for both curves, this means that both curves have $IC_{50} = 5$ (i.e., the dose x where derivative of $f(x)$ reaches maximum). By setting $\beta_1 = 1$ and $\beta_2 = 0.2$, both curves start from cell survival proportion 1 and end with 0.2. By setting $\beta_3 = 3$ (red) or 5 (green), they have different slopes at dose $x = 5$.

As interpreted with the plot of DRCs, each parameter in the 4PL function has explicit meaning in the interpretation. Once the DRC is fitted with 4PL function, the parameters may be used as inputs in the prediction of patient's chemotherapy sensitivity.

4PL model reduces the dimension of the DR data from 10 to 4, it preserves more information regarding dose-response mechanism than AUC or IC50 alone. Generally, when the drug concentrations are well designed and the noise level is well controlled (so that the data complies with the 4PL model fitting assumptions), the 4PL model performs better than using only the highest 3 dose-responses [Huang and Pang, 2008].

2.2 THE PROPOSED METHODS FOR CLASSIFICATION BASED ON DOSE-RESPONSE DATA

2.2.1 Five-parameter mixture of exponential function (5ME) method

The 4PL model is a useful functional form that describes DRC's with flat top and bottom (low and high dosage). However, in reality such low/high-dose plateaus do not always exist, it is then a mathematical challenge to estimate the IC50 if the DRC does not comply with the model fitting assumptions. Illustrated in Figure 5, specimens from National Surgical Adjuvant Breast and Bowel Project (NSABP) B-40 trial with patients treated by T+AC without Bev. In ChemoFx assay experiment, the DRCs using doxorubicin (drug A) usually does not have flat beginning and ending. Biologically, such two-phased exponentially decreasing trend in DRCs is not uncommon and quite often the cause is unknown; in this case, it could be caused by the mixture of tumor cells and fibroblast cells in the specimens. Therefore, another functional form is needed to fit this type of unusual DRCs with steep ends.

Here we propose a five-parameter mixture of exponential function (5ME) method for dose response curve fitting. The 5ME function consists of two different exponential decreasing components:

$$f(x) = \rho\{1 + a[1 - e^{bx}]\} + (1 - \rho)\{c + (1 - c)e^{-dx}\}, x = 0, 1, \dots, 10. \quad (2.6)$$

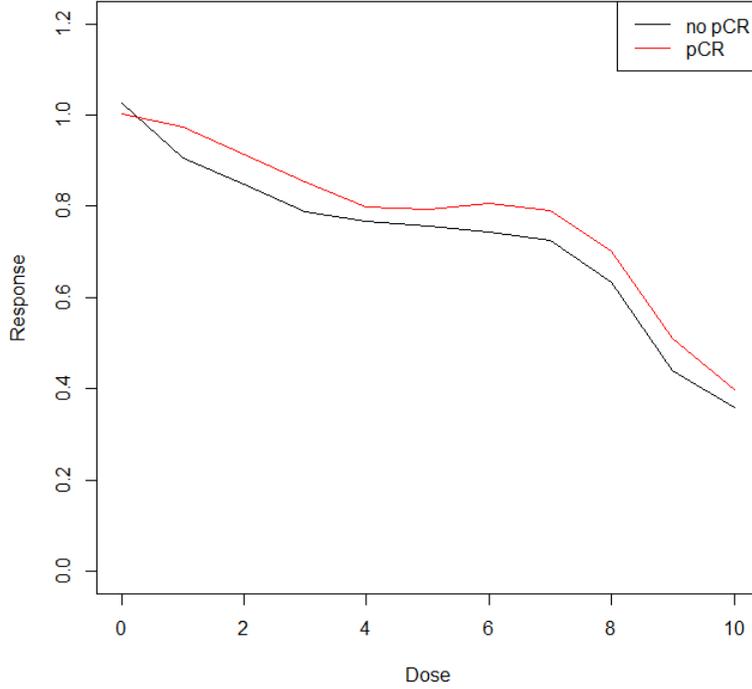


Figure 5: Mean of DRCs of the breast specimens treated with doxorubicin (A). Specimens from pathological complete response (pCR) groups are corresponding to the red line, specimens from non-pCR groups are corresponding to the black line.

where x denotes the dosage, (a, b, c, d) are parameters describing the dose-response mechanism of the DRC: Under lower dosage, we assume that the DRCs start with a fast drop and slow down as $f(x) = e^{-dx}$; Under higher dosage, we assume that the DRCs start with a slow drop and accelerate the decreasing as $f(x) = 1 - e^{bx}$. ρ denotes a weight parameter for how much proportion of the DRC is shared by each component of exponential function.

In Figure 6, we present plots of 5ME functions under two setups of parameters. Parameters (a, b, c, d) are set as the same for both curves, which means that the 2 components of exponentially decreasing function are exactly the same for both curves. We set the weight

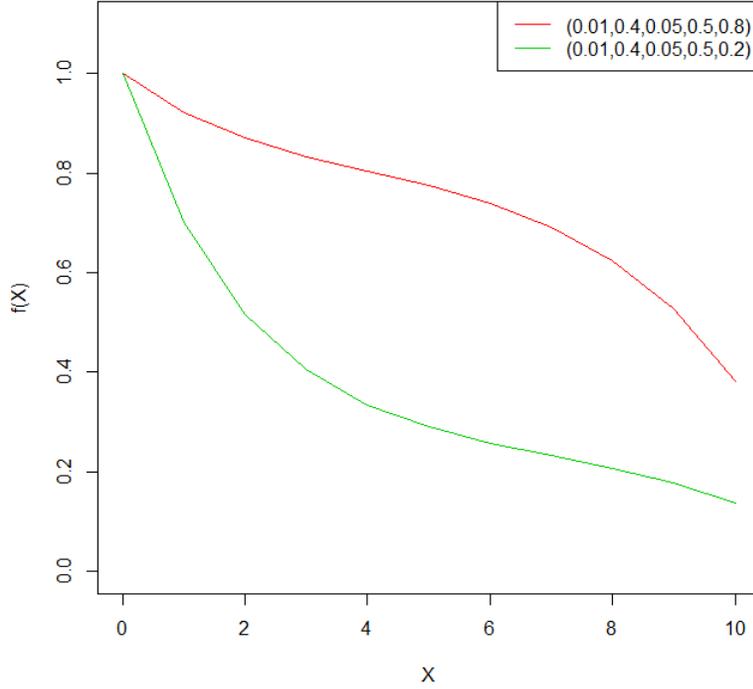


Figure 6: Five-parameter mixture of exponential function (5ME). 5ME with $(a, b, c, d, \rho) = (0.01, 0.4, 0.05, 0.5, 0.8)$ is corresponding to the red curve, 5ME with $(a, b, c, d, \rho) = (0.01, 0.4, 0.05, 0.5, 0.2)$ is corresponding to the green curve.

parameter to be $\rho = 0.8$ or 0.2 , such that the two curves have different proportions of the two exponential functions. The red curve consists of 80% of $f(x) = 1 - e^{bx}$ and 20% of $f(x) = e^{-dx}$, whereas the blue curve consists of 20% of $f(x) = 1 - e^{bx}$ and 80% of $f(x) = e^{-dx}$.

Similar as 4PL method, after the nonlinear curve fitting, the estimated parameters can be used as inputs of the prediction model for patients' chemotherapy sensitivity.

Compared with non-parametric methods, parametric curve fitting methods such as 4PL and 5ME focus more on the dose-response mechanism when extracting summary statistics from DRCs. If the mechanisms agree with the model assumptions, parametric curve fit-

ting should outperform the non-parametric methods. On the other hand, non-parametric models such as AUC and non-parametric IC50 are less sensitive to the violation of the model .

2.2.2 Branching process using dose-0 cell counts data

As discussed earlier, the sample size of dose-0 cell counts in ChemoFx assays usually varies from 6 (2 drugs) to 27 (9 drugs). The mean of the dose-0 cell counts, CWM, is used for the standardization of DRCs. The dose-0 cell counts data are usually not used in predicting patients' chemotherapy sensitivity. Even if the baseline information is used, oftentimes people only consider the mean but ignore the variance. Simply adding mean and variance of dose-0 cell counts may improve the prediction of patient chemo-sensitivity, but extracting cell growth characteristics from the dose-0 cell counts data could benefit us more in prediction and model interpretation. Here we propose a branching process method for extracting information on the uninhibited growth of tumor cells from dose-0 cell counts data. By setting up a proper stochastic process model for cell growth in an assay, we can compute patient specific characters such as initial cell count in the well, cell division probability, and the number of cell cycles with a fixed time window.

A branching process models the reproduction of particles such as cell proliferation. A branching process assumes each particle follows the same probability/rule to produce offspring, and there is no interaction between any particles [[Lange, 2010](#)]. Time is measured discretely as generations in a branching process. Here we assume a cell division process following a Bernoulli distribution.

At time zero of ChemoFx assay, tumor cells are placed in assay plates and are allowed to grow with or without inhibition for 72 hours. The initial cell counts in wells are unknown. The assay protocol requires that 320 cells are placed in each well, but this number varies due to technical variability. Since a typical cell cycle takes 24 hours, we assume that all cells have experienced three cell cycles. We denote:

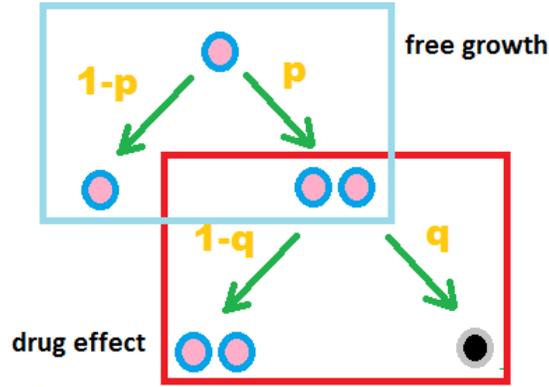


Figure 7: Branching Process: free growth and drug effect

W_0 (TBD): the number of cells initially placed in a well at time zero;

W_t^0 (measured): the number of cells in the well 72 hours after time zero, no treatment was applied;

W_t^k (measured): the number of cells in the well 72 hours after time zero, treatment of dose- k was applied;

p (TBD): the probability of division in each cell cycle;

$n = 3$: the number of cell cycles experienced in time duration $[0,t]$, here $t = 72$ hours.

For each patient, the original assay cell counts data are composed of, (W_t^0, W_t^k) , measured in triplicates, for each treatment. Usually W_t^0 has more than 20 replicates and W_t^k has 3 replicates for each k . We are interested in the patient specific parameters (W_0, p) that describes the free growth character of patients' tumor cells. For a given patient, W_0 reflects how healthy the tumor cells are because the true starting number of cells for each patient depends on the surviving proportion of cells when they are plated; p reflects how aggressively the tumor cells grow.

As shown in Figure 7, for cell growth without inhibition, the branching process is assumed to follow a Bernoulli distribution with probability of division p . The probability of killing q is only applicable to drug-treated wells. Based on the branching process [Lange, 2010], for the free growth starting from one cell, we have:

$$\begin{aligned}\mu &= E(Y_t^0) = (1 + p)^n \\ \sigma^2 &= Var(Y_t^0) = (1 - p)(1 + p)^{n-1}[(1 + p)^n - 1]\end{aligned}$$

Starting from W_0 cells:

$$\begin{aligned}W_0 &= \sum_{j=1}^{W_0} Y_{0j} = \sum_{j=1}^{W_0} 1 \\ W_t^0 &= \sum_{j=1}^{W_0} Y_{tj}^0\end{aligned}$$

Suppose that W_0 is a random variable with mean μ_W and variance σ_W^2 , then conditioning on p and n we have the mean of W_t^0 :

$$\begin{aligned}E[W_t^0|p, n] &= E_{W_0}[E[\sum_{j=1}^{W_0} Y_{tj}^0|W_0, p, n]] \\ &= E[(W_0|p, n)\mu] \\ &= \mu_W \mu\end{aligned}$$

$$E[W_t^0|p, n] = \mu_W(1 + p)^n \tag{2.7}$$

Similarly we have conditional variance of W_t^0 :

$$\begin{aligned}Var[W_t^0|p, n] &= Var_{W_0}[E[\sum_{j=1}^{W_0} Y_{tj}^0|W_0, p, n]] + E_{W_0}[Var[\sum_{j=1}^{W_0} Y_{tj}^0|W_0, p, n]] \\ &= Var[(W_0|p, n)\mu] + E[(W_0|p, n)\sigma^2] \\ &= \sigma_W^2 \mu^2 + \mu_W \sigma^2\end{aligned}$$

$$Var[W_t^0|p, n] = \sigma_W^2(1+p)^{2n} + \mu_W\{(1-p)(1+p)^{n-1}[(1+p)^n - 1]\} \quad (2.8)$$

Note that our data is only consisted of the dose response well counts after-treatment : $\{W_t^{01}, W_t^{02}, W_t^{03}, \dots, W_t^{0M}\}$ for dose-0 cell counts (M is the number of replication of dose-0 wells) and $\{W_t^{k1}, W_t^{k2}, W_t^{k3}\}$ for dose- k cell counts ($k=1,2,\dots,10$). For patient i , we have:

$$E(W_{t_i}^0) = \mu_{W_i}(1 + p_{0_i})^{n_i}$$

$$E(W_{t_i}^k) = \mu_{W_i}(1 + p_{0_i}(1 - 2q_{k_i}))^{n_i}$$

Because the variance of $W_{t_i}^k$ can not be reliably estimated due to the small sample size (at most three), here we apply the branching process model only to dose-0 cell counts data to solve for free growth characters (μ_W, σ_W^2, p) for each patient.

By setting $n = 3$, here we have three unknown parameters (μ_W, σ_W^2, p) and two (mean and variance) equations (2.7)(2.8), therefore we must add assumptions/constraints on $(\mu_W, \sigma_W^2, p$ and/or $n)$. Here are some tentative choices:

If the variance is assumed to be σ_W^2 for the initial cell counts W_0 , we may compute the mean well counts at initial time μ_W^i and cell division probability p_i for each patient i .

Alternatively, if a Poisson distribution of W_0 can be assumed that $W_{0_i} \sim POI(\lambda_i)$, we have

$$E[W_{t_i}^0|p, n] = \lambda_i(1 + p_i)^{n_i}$$

$$Var[W_{t_i}^0|p, n] = \lambda_i(1 + p_i)^{2n_i} + \lambda_i\{(1 - p_i)(1 + p_i)^{n_i-1}[(1 + p_i)^{n_i} - 1]\}$$

then we will have cell division probability and initial cell counts (p_i, λ_i) solved for each patient i .

Similar to 4PL and 5ME, the parameters derived from modeling dose-0 cell counts data using branching process have explicit meanings, and they are then used as inputs to the prediction model for patient clinical response. For example, if we assume Poisson distribution of W_0 , we may have (p_i, λ_i) used as inputs in the prediction of patients' clinical response.

2.3 SIMULATION STUDIES

2.3.1 Comparing 5ME, IC50, and AUC using simulated cell counts data

In this simulation study we generated cell counts data from a mixture of two branching process to simulate the growth of the mixture of fibroblast cells and tumor cells. In each simulation we generated data for 300 patients, 100 labeled with pathological complete response (PCR) and the remaining 200 labeled with non-PCR. For a patient i , 10 dose-0 cell counts and 3 dose- k ($k=1, \dots, 10$) cell counts were generated. Initial cell counts W_0 were generated from $W_0 \sim POI(Z)$, where $Z \sim N(300, 50^2)$, with the ratio of tumor cells $r \sim UNIF(0.7, 0.9)$. Fibroblast cell killing probability $q_F(k)$ and tumor cell killing probability $q_T(k)$ (q_T^+ for pCR group and q_T^- for non-pCR group) follows:

$$q_F(k) = 1 - \exp\{-0.5k\}$$

$$q_T^+(k) = 0.4 + \frac{0.5}{1 + \exp\{0.7(7 - k)\}}$$

$$q_T^-(k) = \frac{1}{1 + \exp\{1.2(6.5 - k)\}}$$

The mean DRC are presented in Figure 8. Here we chose not to generate the dose-response data following an exact 5ME function (in fact a 4PL+ exponential function form was used) because 5ME model is expected to be flexible in fitting DRCs and show its advantage in prediction. After each simulation, with the 300 patients' dose-0 and dose- k well counts, CWM for each patient was computed and used to standardize the DRCs, then we applied the proposed 5ME method and existing methods such as AUC and IC50 to extract

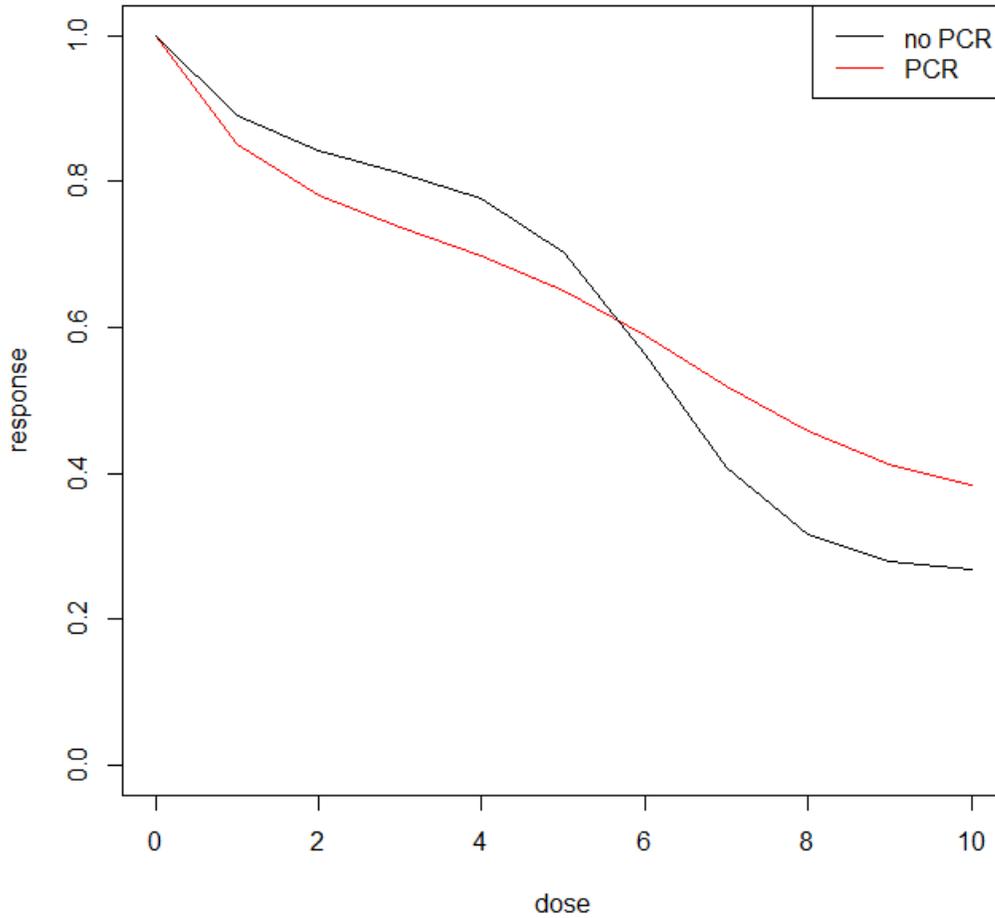


Figure 8: Mean DRCs of simulated cell counts data.

summary statistics from the DRCs. We separated the inputs into training set and testing set (ratio as 4:1) and used a logistic model to predict pCR. The process was repeated for 1000 times. Mean and empirical standard deviation of sensitivity, specificity, PPV, NPV, and prediction accuracy were computed and compared for each method.

The simulation results are shown in Table 1. The *NA's* in PPV column for CWM, IC50 and AUC methods are caused by some extreme predictions of pCR by using these methods

when all patients in testing set are predicted resistant to chemodrugs. Under the given setup of parameters, CWM, IC50 and AUC do not differ between two groups. Therefore the logistic models using CWM, IC50 or AUC did not provide good prediction of chemosensitivity. The 5ME curve fitting method in this case can transform the data into the parameter space and tell the difference between the two groups. Although we choose not to use 5ME function to generate the data, the method still show its advantage. The prediction accuracy using 5ME method is significantly better than those using CWM, IC50 and AUC methods.

Table 1: Comparison of 5ME, IC50 and AUC based on simulated cell counts

	Entries	sensitivity	specificity	PPV	NPV	PA
1	CWM	0.001	0.99	NA	0.667	0.666
	Emp.SD	0.01	0.01	NA	0.002	0.004
2	IC50 + CWM	0.007	0.963	NA	0.659	0.644
	Emp.SD	0.03	0.05	NA	0.013	0.034
3	AUC + CWM	0.02	0.99	NA	0.67	0.67
	Emp.SD	0.04	0.02	NA	0.01	0.02
4	5ME + CWM	0.50	0.89	0.71	0.78	0.76
	Emp.SD	0.12	0.06	0.12	0.04	0.05

2.3.2 Comparing 5ME, IC50, and AUC using simulated dose-response curves

In this simulation study, instead of generating the cell counts from mixture of branching processes, here we repeatedly ($K=1000$) simulated the proportion of cell surviving from some given functions and artificially adding some errors. Here we generated the dose response curves from a seven-parameter mixture of an exponential function and a logistic function (7MEL):

$$f(k) = \rho \cdot \left\{ a + \frac{b - a}{1 + \exp(c(d - k))} \right\} + (1 - \rho) \cdot \{ A + (1 - A) \cdot \exp(-Bk) \} + \epsilon \quad (2.9)$$

where $\epsilon \sim UNIF(-0.1, 0.1)$.

As presented in Figure 9, in each simulation we generate dose-response data following 7MEL function as in (2.9), and eight different setups of parameters $\theta = (a, b, c, d, A, B, \rho)$ as

$$\theta_1^+ = (1, 0.2, 3, 5, 0.5, 0.5, 0.4)$$

$$\theta_1^- = (1, 0.2, 1, 5, 0.5, 0.5, 0.4)$$

$$\theta_2^+ = (1, 0.2, 3, 4.5, 0.5, 0.15, 0.4)$$

$$\theta_2^- = (1, 0.18, 1, 4.7, 0.4, 0.1, 0.44)$$

$$\theta_3^+ = (1, 0.2, 3, 5.6, 0.4, 0.15, 0.3)$$

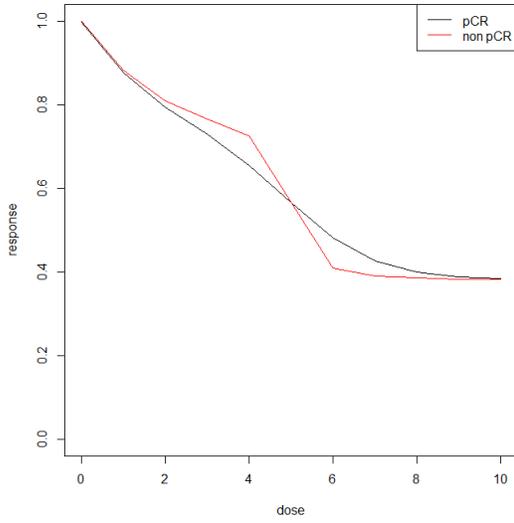
$$\theta_3^- = (1, 0.3, 1, 6, 0.5, 0.3, 0.44)$$

$$\theta_4^+ = (1, 0.1, 3, 8, 0.4, 0.15, 0.3)$$

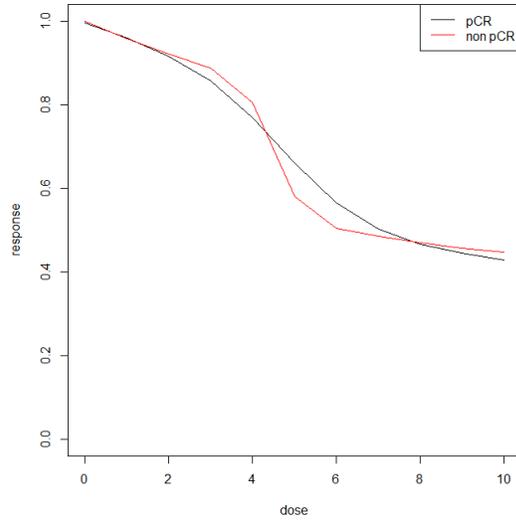
$$\theta_4^- = (1, 0.2, 1, 8, 0.5, 0.3, 0.44)$$

Table 2: Comparing 5ME, IC50 and AUC using simulated dose-response data

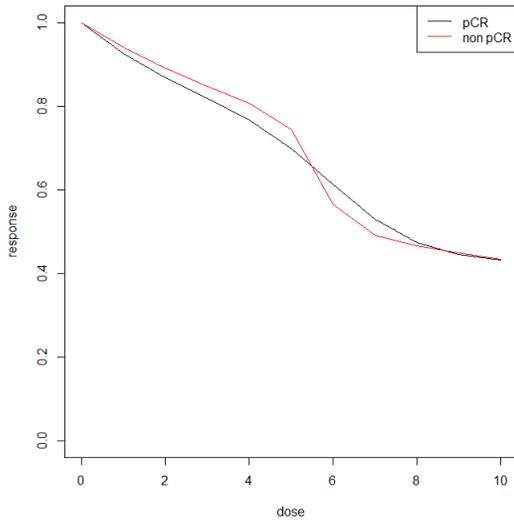
	Entries	sensitivity	specificity	PPV	NPV	PA
1	IC50	0.61	0.49	0.56	0.57	0.55
	Emp.SD	0.19	0.20	0.05	0.05	0.02
2	AUC	0.71	0.55	0.61	0.65	0.63
	Emp.SD	0.06	0.07	0.03	0.05	0.04
3	5ME	0.79	0.56	0.64	0.73	0.67
	Emp.SD	0.06	0.06	0.03	0.05	0.03



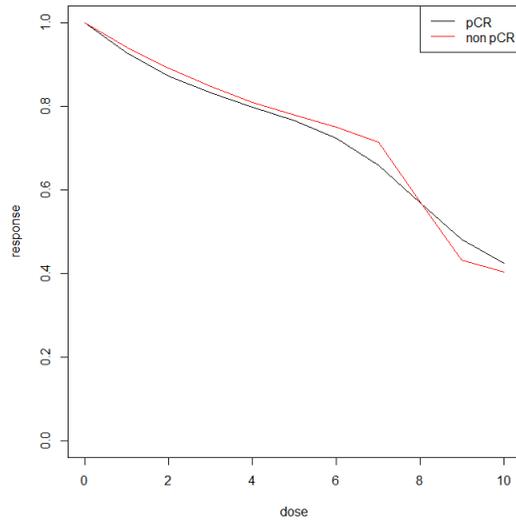
(a) 7MEL functions with θ_1^+ and θ_1^-



(b) 7MEL functions with θ_2^+ and θ_2^-



(c) 7MEL functions with θ_3^+ and θ_3^-



(d) 7MEL functions with θ_4^+ and θ_4^-

Figure 9: Mean of Simulated DRC following seven-parameter mixture of a 4PL function and an exponential function (7MEL) functions.

Among the eight setups of parameters, we use θ_1^+ , θ_2^+ , θ_3^+ , and θ_4^+ for generating DRCs for sensitive patients, and θ_1^- , θ_2^- , θ_3^- , and θ_4^- for generating DRCs for resistant patients. We generate 9600 DRCs (1200 for each θ), split them into training set and testing set in a 5:1 ratio, and apply the proposed 5ME methods and traditional methods such as IC50 and AUC. We repeated the process for $K = 1000$ times. Mean and empirical standard deviation of sensitivity, specificity, PPV, NPV, and prediction accuracy were computed and compared for each method.

Table 2 summarizes the performance of these methods using dose-response data simulated from 7MEL functions. Here since we set the parameters θ 's in pairs such that the AUC and IC50 are similar between the sensitive and the resistant groups, the prediction accuracy is relatively low as we expected. Compared to IC50 and AUC, the proposed 5ME method shows better performance predicting responses in such a mixture pattern situation.

2.4 APPLICATION TO CLINICAL TRIAL DATA

2.4.1 Analysis of B40 trial data

National Surgical Adjuvant Breast and Bowel Project (NSABP) Protocol B-40 was designed to determine if adding capecitabine(X) or gemcitabine(G) to neo-adjuvant docetaxel followed by doxorubicin/cyclophosphamide(T+AC) and if adding bevacizumab(Bev) to chemotherapy would improve pathologic complete response(pCR) rate of breast cancer patients.

Pathologic complete response(pCR) is defined as no invasive and no in situ residuals in breast and nodes for breast cancer patients [Minckwitz et al., 2012]. pCR can be associated with patient disease free survival (DFS) after treatment. For some subgroup of patients, pCR can be used as a suitable surrogate end point.

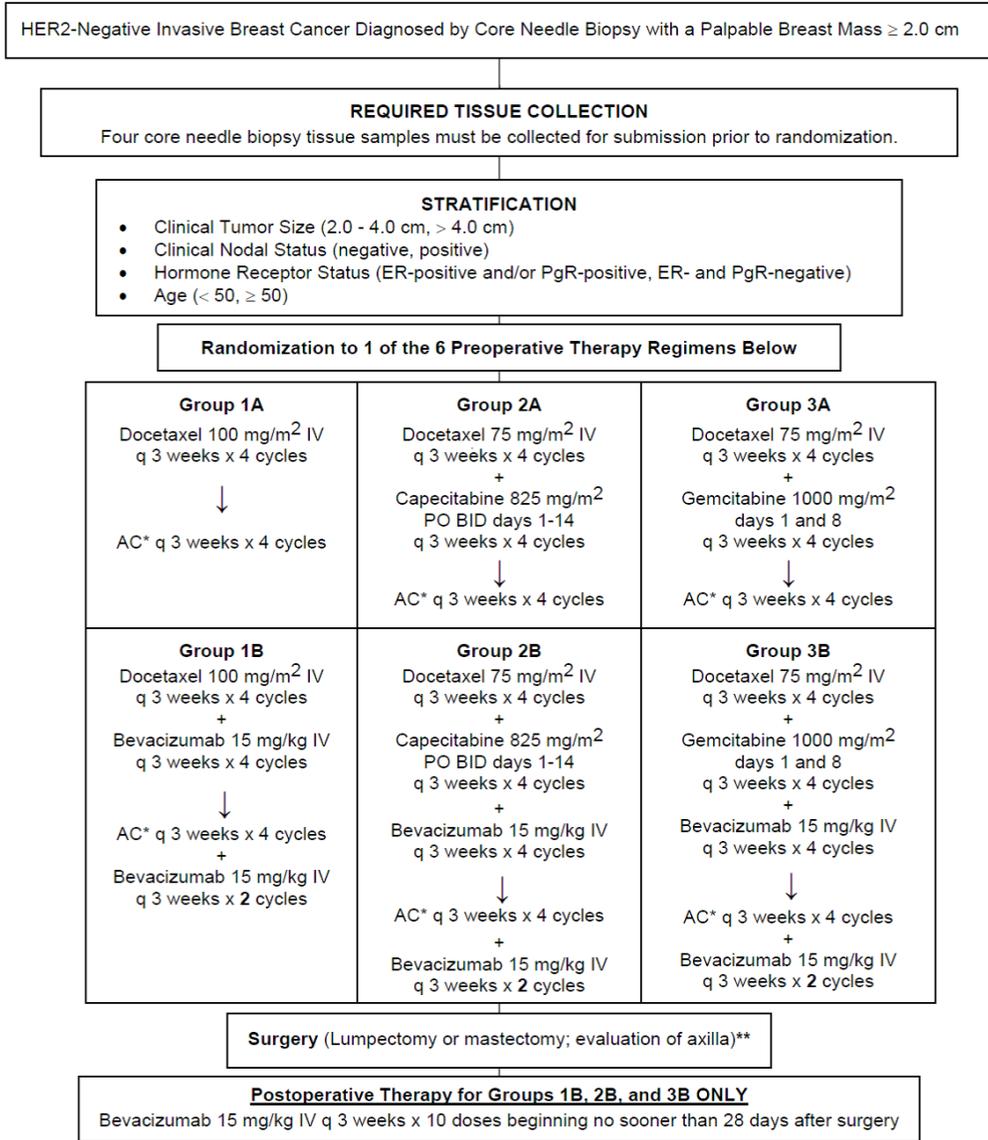


Figure 10: B-40 Schema

It was learned from another neoadjuvant trial conducted and reported by NSABP, protocol B-27, that adding docetaxel(T) after doxorubicin/cyclophosphamide (AC) significantly increased clinical and pathological response rates for operable breast cancer [Bear et al., 2003, 2006, Fisher et al., 1997]. Recent research showed that bevacizumab(Bev) and capecitabine (X) / gemcitabine (G) could possibly improve response rate for some metastatic-disease [Robert et al., 2011, O' Shaughnessy et al., 2002, Albain et al., 2008]. Therefore NSABP protocol B-40 trial followed such factorial design that eligible patients were randomly assigned to 3 arms T+AC, TX+AC, and TG+AC, then patients in each arm were randomized to receive Bev or not(see Figure 10).

At the end of B-40 trial, 1186 of the 1206 patients had available pCR data. The conclusion from analysis of B-40 clinical data was that addition of X or G did not significantly improve pCR rate compared with T alone (pCR rate in Arm 2 and 3 was not significantly higher than that in Arm 1); addition of Bev to chemotherapy significantly increased pCR rate (pCR rate in Arm B's was significantly higher than that in Arm A's)[Bear et al., 2012]. For patients with positive hormone receptor(HR), the effect of adding Bev was significant, while for patients with negative HR, the effect was minimal.

In B40 trial, the biopsies from breast cancer patients were delivered to ChemoFx[®] assay lab in Precision Therapeutics Inc. for ex vivo experiments. Tumors cells were cultured for several weeks until they reached the requirements for assays in both quality and quantity. Then each patient's cells were placed by robots into wells on plates to receive all single drugs and combinations of those drugs (11 assays for B40 in total: T,A,C,X,G, TX,TG,AC,TAC,TXAC,TGAC).

A total of 473 patients in B40 trial had successful ChemoFx assay experiments using drug A and T, and 223 of them were involved in arm 1A, 2A, 3A (treated without Bev) and the remaining 250 were involved in arm 1B, 2B, 3B (treated with Bev). Our validation was only conducted on 223 patients from arm A and using the assay results for drug A and T. Patients from arm B were not used because Bev had been proved to significantly

improve pCR rate, which could confound the results. Drug C was also not used for analysis because there is evidence from the lab that the assay results for this drug were not reliable.

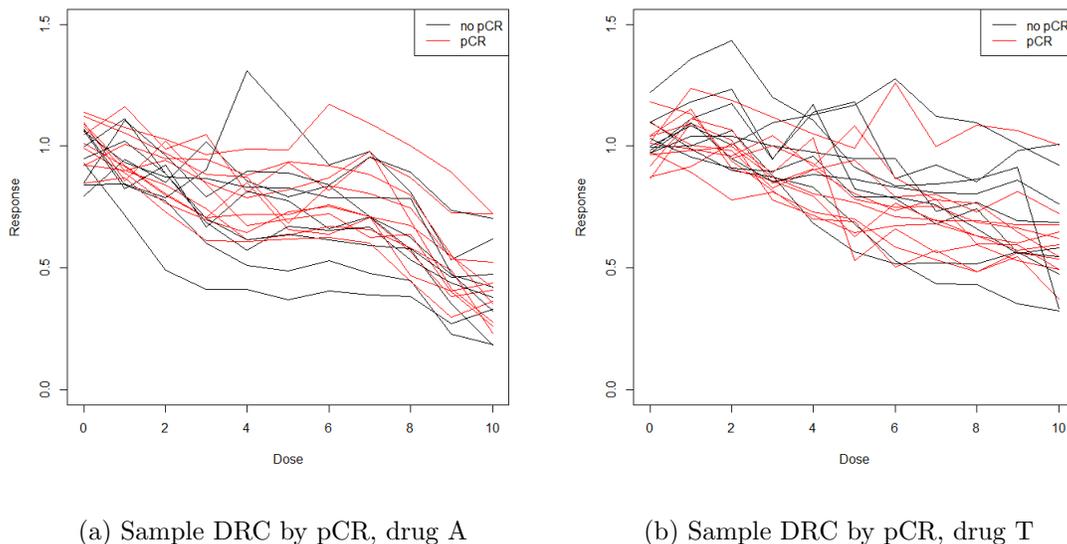
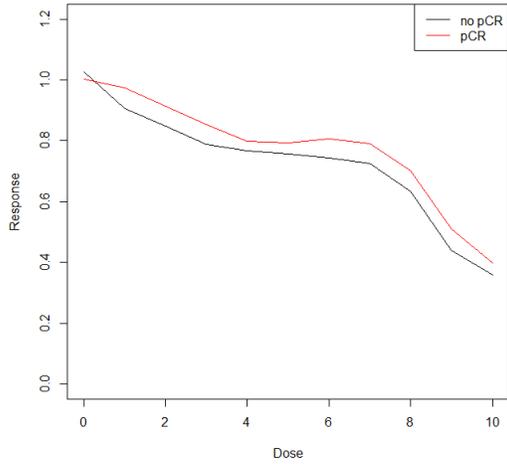


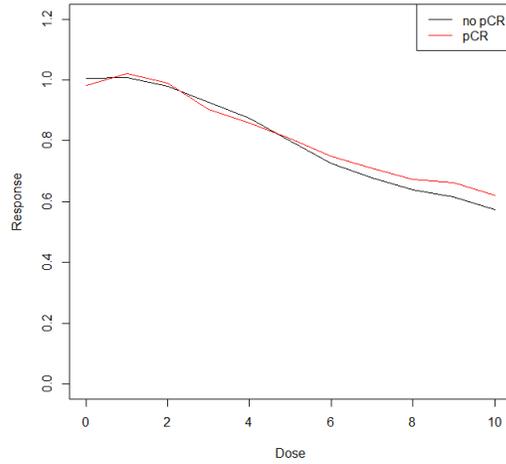
Figure 11: Sample dose-response curves in assay with drug A or T, grouped by pCR of patients in NSABP B-40 Arm 1A.

Figure shows some sample individual DRCs in assay with drug A and T for patients from arm A1 where T+AC was applied. Unfortunately the mean DRC of patients with pCR was even higher than that of patients without pCR based on assay results of drug A, which was the opposite of what we expected. The mean DRCs of patients with pCR and non-pCR were also not separable based on assay results of drug T, adding into the difficulties of predictions.

The DRCs were standardized by the CWMs. For extracting summary statistics, we applied AUC7, AUC10, IC50, BP and parametric curve fitting (4PL method for assay using drug T and 5ME for assay using drug A). We fitted drug A and drug T assays with different models, because drug T showed DRCs with high and low plateaus, while drug A assays could only be fitted by 5ME model. The summary statistics from drug A assays and drug T assays are combined and used as inputs for prediction. We applied 5-fold cross-validation



(a) Mean DRC by pCR, drug A



(b) Mean DRC by pCR, drug T

Figure 12: Mean dose-response curves(DRC's) in assay with drug A or T, grouped by pCR of patients in NSABP B-40 Arm 1A.

for $K = 1000$ times using logistic model to predict patient pCR. The prediction accuracy and corresponding resampling standard error of these methods are summarized in Table 3.

Table 3: Comparison of methods using B40 ChemoFx assay data

Method	NO.pts	Prediction Accuracy	Resampling SE
1 DRC + CWM	223	0.67	0.09
2 AUC10 + CWM	223	0.71	0.14
3 AUC7 + CWM	223	0.65	0.09
4 IC_{50} + CWM	223	0.69	0.07
5 4PL/5ME+CWM	222	0.70	0.13
6 4PL/5ME+BP	198	0.68	0.07

To compare the methods 1-5, with logistic models using different inputs such as dose-response data, AUC10, AUC7, nonparametric IC50, or 4PL/5ME, IC50 appeared to have a smaller resampling SE, and a slightly better prediction accuracy. The 4PL/5ME method showed a higher prediction accuracy, but the resampling SE was also larger. This was probably caused by the large noise of the dose-response data, and because parametric models have more parameters, the variation of prediction performance could be more affected compared to simple model such as non-parametric IC50. AUC10 also showed a large SE in prediction, this could be caused by the instability of the responses at higher doses.

To compare methods 5 and 6, with 4PL/5ME approach with or without character variables from branching process in the model, branching process did not provide any improvement in predicting pCR, this method was also unable to achieve a solution of p for 24 patients.

2.4.2 Analysis of PTI-206 ChemoFx[®] assay data

To further validate the performance of the proposed branching process (BP) method, herein we conducted the analysis using PTI-206 assay data. PTI-206 was a study for advanced ovarian cancer. In the study, patients were enrolled in a pre-defined protocol. Tumor samples from 54 institutions were submitted for chemo-response testing between 2006 and 2010. Women with FIGO stage III-IV epithelial ovarian (EOC), fallopian tube (FTC) and peritoneal (PPC) cancer treated with carboplatin (C)/ paclitaxel (P)-based chemotherapy following initial cytoreductive surgery were included in the study. Clinical response (CR) was defined as positive if a patient experienced more than six months from the end of primary chemotherapy. If recurrence was reported within six months, then the patient had no CR.

262 patients in the PTI-206 trial had successful ChemoFx assay results. After filtering out missing data, 230 patients remained with clinical response (CR) and Carboplatin assay

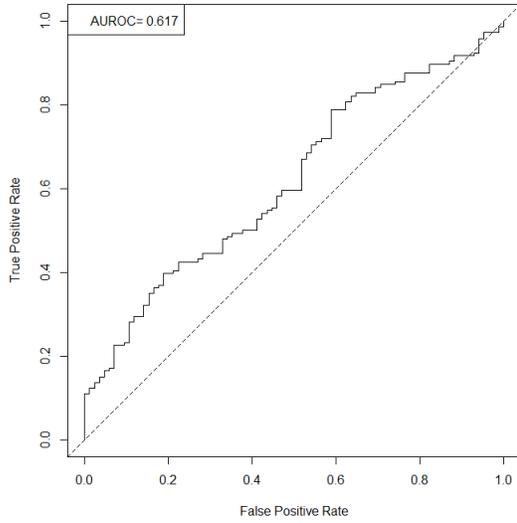
result (Carboplatin is the drug applied to patients in this trial). The goal of this application is to compare the BP with traditional baseline adjustment methods in predicting CR. AUC7 is used for the prediction in these Chemofx assays, considering the instability of assay responses at high doses.

For each patient, the data to be used includes AUC7, CWM and dose-0 cell counts. To apply the BP method, we first computed the mean and variance of dose-0 cell count. Then assuming all cells experiencing three cycles in 72 hours in the assay experiment ($n = 3$) and Poisson distribution for initial well counts ($W_0 \sim POI(\lambda)$), we used (2.7,2.8) to compute the initial mean cell counts λ and cell division probability p for each patient.

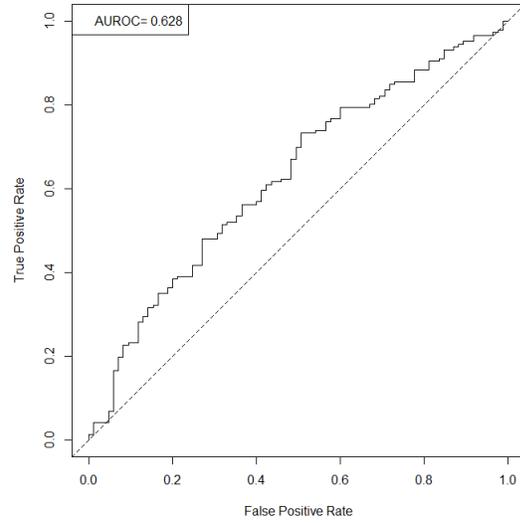
To predict CR, logistic models are performed with the following inputs: CWM + variance of dose-0 counts, $\lambda + p$, AUC7 + CWM + variance of dose-0 counts, AUC7 + $\lambda + p$. The performance of prediction was compared by receiver operating characteristic (ROC) analysis. The ROC plots are shown in Figure 2.4.2. The Area under the ROC curve (AUROC) of each method are summarized in Table 4. AUROC for CWM + variance of dose-0 counts model was 0.617, AUROC for $\lambda + p$ model was 0.628, AUROC for AUC7 + CWM + variance of dose-0 counts model was 0.632, AUROC for AUC7 + $\lambda + p$ model was 0.647.

Table 4: Comparison of methods by AUROC based on PTI-206 assay data

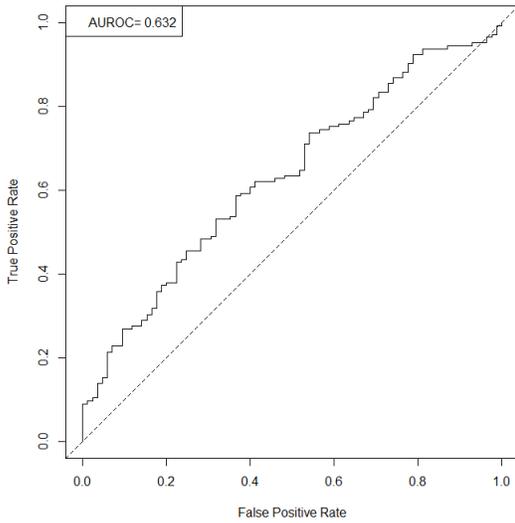
Predictors	AUROC
1 mean and variance of dose-0 cell counts	0.617
2 (λ, p) from BP	0.628
3 AUC7 + mean and variance of dose-0 cell counts	0.632
4 AUC7 + (λ, p) from BP	0.647



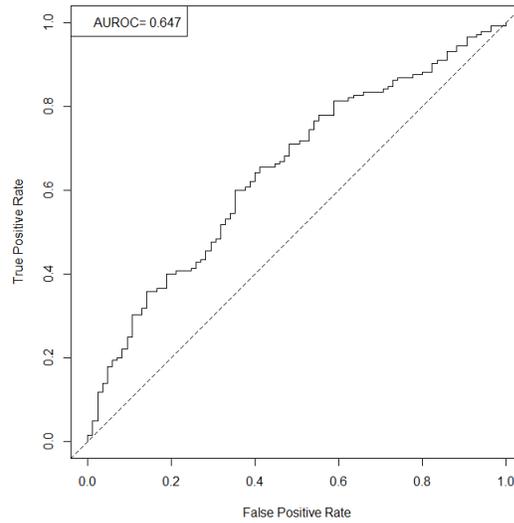
(a) Inputs: CWM + variance of dose-0 counts



(b) Inputs: $\lambda + p$



(c) Inputs: AUC7 + CWM + variance of dose-0 counts



(d) Inputs: AUC7 + $\lambda + p$

Figure 13: Comparison of methods by ROC analysis using PTI-206 assay data

First, we compared model 1 and model 2 (CWM + variance of dose-0 counts model vs $\lambda + p$ model). The inputs in these two methods were equivalent in information, because that $\lambda + p$ was computed from CWM + variance of dose-0 counts by using BP method. In this comparison, the BP method improved the AUROC by 1.8%.

We then compared model 3 and model 4 (AUC7 + CWM + variance of dose-0 counts model vs AUC7 + $\lambda + p$ model). The inputs in these two methods were also equivalent. In this comparison, the BP method improved the AUROC by 2.4%.

2.5 DISCUSSIONS

Here we proposed a five-parameter mixture of exponential function (5ME) method for dose-response curve (DRC) fitting and a branching process (BP) method for dose-response baseline characterization.

The 5ME model is a parametric curve fitting method similar to the four parameter logistic model (4PL). The difference between these two models is the type of dose-response curves they describe. The 5ME model can be fitted into DRCs without upper and bottom plateau, and an improved performance is seen in simulation study by comparing with other traditional methods such as IC50 and AUC.

When noise is large as that in the B40 assays, the fitting of 5ME function to DRC is not precise, and the prediction performance using the curve parameters is than ideal.

The BP method is based on the cell division modeling in an *in vitro* assay. This method was illustrated in the analysis of PTI-206 assay data. When applying this model, because the variance of dose-0 cell counts were utilized, it was expected to provide more information on the character of patient's tumor cells. Compared to using mean and variance of dose-0 cell counts, using λ and p from the BP method improved the prediction of clinical response.

The improvement by the branching process method, relies on the available biological knowledge of cell division process in assay experiment. Here we simply assumed that the number of cell cycles was fixed and the probability of cell division followed a Bernoulli probability. These assumptions can be refined and the prediction should be improved if we have more understanding of the experiment and the behavior of tumor cells in the assays, such as the distribution of initial cells in each well and the cell cycle variation.

3.0 A MODIFIED EM ALGORITHM FOR REGRESSION ANALYSIS OF DATA WITH NON-IGNORABLE NON-RESPONSE

3.1 OVERVIEW OF ANALYSIS OF MISSING DATA

3.1.1 Missing data in practice

Missing data are prevalent in biomedical studies especially in large clinical trials and longitudinal studies where values of some variables are missing from many subjects. In general, missing values may occur due to design, loss to follow-up or inability to record [Little and Rubin, 2002] [Little et al., 2012].

For example, in some case-control studies, the assessment of a bio-marker is too expensive that the investigators can only afford to perform the assay on a subset, or values of certain exposure variables may not be available from some subjects based on their existing records. In many longitudinal studies, an outcome variable is repeatedly measured over time to provide information on its trend overtime in individuals. The values of the outcome variable may be missing during the study because of missed visits or dropouts.

In most statistical software packages, subjects with missing values are deleted automatically and analysis is solely based on the subsets with complete records on involved variables. Simply applying standard statistical methods using complete cases often leads to biased estimates. The bias can be substantial when the proportion of missing values is high or the missingness does not occur randomly. Missingness in some variables may cause imbalance in treatment among the complete cases. For example, in a randomized clinical trial, patients

are randomized into two treatment arms, to balance the two groups for both known and unknown factors. If patients' dropout is related with less improvement of the symptoms than expected, then the comparison between the treatment groups will be biased in favor of higher rate of dropout.

It is important to distinguish missing-data patterns and missing-data mechanism because of their implication in appropriate statistical analysis[[Little and Rubin, 2002](#)]. Missing-value indicators, with value at 1 or 0, are often used to indicate the missing status of a variable for a subject. Then we have a vector for each subject and a matrix of missing indicators for all subjects. Such missing indicator matrix forms the missing-data pattern. Missing-data patterns are used to describe which values are observed and which values are missing. For example, monotone missing data are related with longitudinal dropouts, when all future observations are missing after the dropout. Non-monotone missing data are more complicated and require modeling of the dropout mechanism.

Missing-data mechanism is used to describe how missingness is related with the hypothetical complete data, including missing values. Rubin [[Rubin, 1976](#)] defined three types of missing-data mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). When missingness is not related with the values of any variables in the study/data, the data are called MCAR. When missingness is related with the values of observed variables only, the data are called MAR. When missingness is related with the values of unobserved variables after conditioning on the observed variables, the data are called NMAR.

For example, in a randomized clinical trial with two treatment arms, suppose that patients in treatment group will receive IV injection, whereas those in control group will take pills. If a patient discontinues the treatment because he does not like the flavor of the pills, the missing mechanism here is MAR. This is because that the group assignment is observed. If loss of follow-up is caused by heart attack which is only related to real-time blood pressure, then the missing-data mechanism is NMAR. This is because that the values which

missingness depend on may not be observed. If loss of follow-up is caused by relocation for job change, the data are in general MCAR.

3.1.2 Statistical methods for analysis of missing data

In the following context, we denote $Y = (Y_1, \dots, Y_K)$ the vector of variables and $R = (R_1, \dots, R_K)$ the vector of missing data indicators. $Y = (Y_{obs}, Y_{mis})$, Y_{obs} and Y_{mis} denote the observed and missing components of Y . $R_k = 0$ if Y_k is missing; $R_k = 1$ if Y_k is observed. Missing-data mechanism is generally described via conditional distribution $f(R|Y, \psi)$, where ψ denotes parameters for the missing-data mechanism model. Based on how the missingness is related to the hypothetical complete-data vector Y , Rubin classified three missing-data mechanisms [Rubin, 1976]:

1. MCAR

Data are missing completely at random(MCAR) if the missingness does not depend on the underlying complete data at all.

$$f(R|Y; \psi) = f(R; \psi) \text{ for all } Y.$$

2. MAR

Data are missing at random(MAR) if the missingness depends on observed values only.

$$f(R|Y; \psi) = f(R|Y_{obs}, \psi) \text{ for all } Y_{mis}, \psi.$$

3. NMAR

When missingness still depends on Y_{mis} after conditioning on Y_{obs} , the data are called not missing at random (NMAR).

For regression analysis of a data set, missingness could happen in response, covariates, or both. The impacts of missing values are different for response and covariate, and yield different analytical methods.

In the following context, we consider regression analysis of data with nonresponse where the covariates are fully observed and the response variable is subject to missing values. Suppose the observed data for (X, Y) are $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i 's are fully observed, and the response variable Y is only observed for the first r subjects, i.e., $(y_r, y_{r+1}, \dots, y_n)$ are missing. In practice, we are usually concerned with the following regression models:

$$[Y|X] \sim g(y|x; \theta)$$

where $g(\cdot|x; \theta)$ is a parametric or semi-parametric model. Here the interest is to make inference on θ . There are generally two types of methods in missing-data analyses [Ibrahim et al., 2012]: ad-hoc methods such as complete case analysis and imputations, and standard statistical methods such as inverse-probability weighted estimating equations (IPWEE), maximum likelihood, Bayesian / data augmentation methods.

Complete-case analysis is a simple methods that applies on the complete cases. In complete-case analysis, the statistical inference on θ is based on the complete cases $\{(x_1, y_1), (x_2, y_2), \dots, (x_r, y_r)\}$. Complete-case analysis is based on a well-defined subset and simple to perform. However, it does not use information collected from incomplete cases. It is usually inefficient and may be biased when data are not MCAR [Ibrahim et al., 2012].

Imputation is also a common ad-hoc method for analysis of missing data. In general, the imputation methods are conducted as following: at first we identify missing values from the data, then we obtain a predictive model for the distribution of missing values given observed data. Values are drawn from the predictive distribution. After imputations, the usual estimation can be performed on the imputed dataset

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_r, y_r), (x_1, \tilde{y}_{r+1}), (x_2, \tilde{y}_{r+2}), \dots, (x_r, \tilde{y}_n)\}.$$

where \tilde{y}'_j 's are the imputed values. Imputation is a straightforward and flexible method. However, appropriate imputation methods usually require a reasonable (estimated) imputation model. The subsequent inference has to take into account of the variations due to imputation models and the intrinsic variation due to the draws [Little and Rubin, 2002].

Multiple imputation are used to justify variation to improve single imputation. In multiple imputation, we repeatedly generate D (usually $2 \leq D \leq 10$) complete data sets, then we can achieve inference on θ by combining the estimations from D data sets [Rubin, 1987] as follows:

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$$

$$(\theta - \bar{\theta}_D) T_D^{-\frac{1}{2}} \sim t_\nu$$

where T_D is the variance of $\bar{\theta}_D$ computed from within-imputation variance and between imputation variance, ν is the degrees of freedom of the t distribution computed from D and the variances.

The advantage of multiple imputation is the computation efficiency. Compared with re-sampling methods such as bootstrap and jackknife, multiple imputation required fewer times of repeating computation to obtain variance estimates of imputation estimates.

In summary, ad-hoc methods usually require data missing completely at random (MCAR) or missing at random (MAR), which limited the usage of such methods.

A generalized estimating equation (GEE) is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes [Liang and Zeger, 1986]. The inference on θ could be done by solving the follow equation:

$$\sum_{i=1}^n D_i(x_i, \theta) \{y_i - \mu(x_i, \theta)\} = 0$$

where $D_i(x), \theta$ is a $d \times K$ matrix chosen according to the distribution of y_i , d is the dimension of parameter θ , $\mu(x, \theta) = E[Y|x]$.

Rubin and others presented the GEE method for the data missing at random (MAR) using inverse probability weight (IPW) [Robins et al., 1994]. When we consider the missing in data, the GEE based on the complete cases is:

$$\sum_{i=1}^r D_i(x_i, \theta) \{y_i - g(x_i, \theta)\} = 0$$

where the first r y_i 's are observed. Supposed there is a fully observed auxiliary variable Z . Assuming data are MAR

$$pr(R_i = 1|x_i, y_i, z_i) = pr(R_i = 1|x_i, z_i; \psi) = f(x_i, z_i; \psi)$$

Using inverse-probability weight [Fitzmaurice et al., 1995], the inference on θ can be achieved by solving the weighted GEE:

$$\sum_{i=1}^r f(x_i, z_i; \hat{\psi})^{-1} D_i(x_i, \theta) \{y_i - g(x_i, \theta)\} = 0$$

where $\hat{\psi}$ is usually estimated by regression of R on X and Z .

GEE is a well-established method, flexible for correlation structure adjustment. [Troxel et al., 1997] extended GEE method for analysis of data with non-ignorable nonresponse. However in general GEE methods are limited to data MAR or MCAR and the missing-data mechanism is required in the likelihood function.

Likelihood-based methods can be classified into selection models and pattern-mixture models by the way of likelihood partition. Selection models [Diggle and Kenward, 1994] factor the joint distribution into underlying complete data distribution and missing data mechanism:

$$p(Y, R; \theta, \psi) = p(Y; \theta)p(R|Y; \psi)$$

Pattern mixture model [Little, 1994] stratify the data based on missing-data pattern:

$$p(Y, R; \theta, \psi) = p(Y|R; \theta)p(R; \psi)$$

Selection model is a natural way of factoring the model considering that the relationships between Y and X in the full population is usually of interest. Also it will be shown later in this section that when data are MAR, modeling missing-data mechanism is not necessary for likelihood-based inference for selection models.

Pattern-mixture models, on the other hand, can be more suitable if the subpopulation distribution is of interest. Also in some special situations, when data are NMAR, pattern mixture models could be more convenient to perform than selection model.

Rubin showed that when data are missing at random (MAR) and missing-data mechanism is distinct from the regression model, the missing-data mechanism can be ignored in the likelihood-based inference [Rubin, 1976]:

$$\begin{aligned} L_{full}(\theta, \phi; Y_{obs}, R) &= \int f(Y_{obs}, Y_{mis}; \theta) f(R|Y_{obs}, Y_{mis}; \phi) dY_{mis} \\ &= \int f(Y_{obs}, Y_{mis}; \theta) f(R|Y_{obs}; \phi) dY_{mis} \\ &= f(Y_{obs}; \theta) f(R|Y_{obs}; \phi) \\ &= L_{ign}(\theta) f(R|Y_{obs}; \phi) \end{aligned}$$

where $L_{ign}(\theta) = f(Y_{obs}; \theta)$ is the ignorable likelihood function. For data MAR, since the second term in full likelihood becomes irrelevant to θ , inference based on $L_{full}(\theta, \phi; Y_{obs}, R)$ is equivalent to inference based on $L_{ign}(\theta)$, and the missing-data mechanism does not need to be modeled here.

When the MAR condition is not met, missing-data mechanism is required in constructing full likelihood function for estimation and inference of regression parameters. Misspecification of missing-data model often yields biased estimates and wrong conclusions.

In practice, it is often the case that we can not identify the missing-data mechanism. Therefore, when applying likelihood based method for data missing not at random (NMAR), people usually make some assumptions about the missing-data model. This is the motivation of the article to develop a general (missing-data mechanism assumption free) method for regression analysis when missing-data mechanism is unknown.

Tanner and Wong showed that the posterior distribution of regression parameter given observed data can be achieved by an iterative data augmentation procedure [Tanner and Wong, 1987]. In the imputation step (I-step), the missing values are drawn from the distribution of missing values given observed values and given regression parameters. In the posterior step (P-step), the distribution of regression parameters are updated by the posterior distribution of model parameters derived from the imputed dataset. This method is called data augmentation.

For example, when data are MAR, Bayesian inference is based on the posterior distribution

$$p(\theta|Y_{obs}) \propto p(\theta)f(Y_{obs};\theta)$$

where $p(\theta)$ is the prior distribution. A random sample are drawn from the posterior distribution of θ and its sample property can be used to make inference on θ . However, the posterior distribution $p(\theta|Y_{obs})$ could be complicated. By data augmentation we iteratively simulates random samples of missing values and models parameter given observed data, with 2 steps in each iteration:

I Step (imputation): draw $Y_{mis}^{(t+1)}$ with density $p(Y_{mis}|Y_{obs}, \theta^{(t+1)})$;

P Step (posterior): Draw $\theta_{mis}^{(t+1)}$ with density $p(\theta|Y_{obs}, Y_{mis}^{(t+1)})$ as if data were complete.

The iterations start with initial draw $\theta^{(0)}$ from an approximation to the posterior distribution of θ . As $t \rightarrow \infty$, the iterative procedure will eventually yield a draw from the joint distribution of $p(\theta|Y_{obs})$. Data augmentation methods are usually computationally intensive.

Here we focus on maximum likelihood method, for the interest of dealing with more general missing-data mechanisms. In the following context we will develop a modified EM algorithm for regression analysis of data with non-responses when missing-data mechanism unknown.

3.2 EM ALGORITHM FOR REGRESSION ANALYSIS OF DATA WITH NON-RESPONSES WHEN THE MISSING-DATA MECHANISM IS MODELED

Assume observed data $\mathcal{D}_{obs,i} = (x_i, R_i y_i, R_i)$ for subject $i = 1, 2, \dots, n$. Suppose x_i 's be the fully observed covariate; y_i 's be partially observed responses, observed for $i=1, 2, \dots, m$; R_i 's be the missing data indicator where $R_i = 1$ stand for observed y_i and $R_i = 0$ for missing y_i .

Without loss of generality, we assume for the complete data, the conditional distribution with parameter θ follows an exponential family and canonical link.

$$y_i|x_i \sim g(y_i|x_i; \theta) \propto \exp\{\theta S(x_i, y_i) + h(x_i, y_i) + a(\theta)\} \quad (3.1)$$

Consider the following missing-data model

$$pr[R_i = 1|x_i, y_i] = w(x_i, y_i; \psi)$$

The interest is the inference of the distribution parameters θ .

When the parameters of the missing-data mechanism is distinct from the parameters of the regression model, the likelihood function for (θ, ψ) given observed data follows

$$\begin{aligned} L(\theta, \psi; \mathcal{D}_{obs}) &= \prod_{i=1}^n p(R_i, R_i y_i|x_i; \theta, \psi) \\ &= \prod_{i=1}^m g(y_i|x_i; \theta) w(x_i, y_i; \psi) \cdot \prod_{i=m+1}^n \int g(y_i|x_i; \theta) \{1 - w(x_i, y_i; \psi)\} dy_i \end{aligned}$$

The likelihood function could also be written as

$$\begin{aligned}
L(\theta, \psi; \mathcal{D}_{obs}) &= \prod_{i=1}^n \frac{p(R_i, R_i y_i, (1 - R_i) y_i | x_i; \theta, \psi)}{p((1 - R_i) y_i | R_i, R_i y_i, x_i; \theta, \psi)} \\
&= \prod_{i=1}^n \frac{p(R_i, y_i | x_i; \theta, \psi)}{p((1 - R_i) y_i | R_i, R_i y_i, x_i; \theta, \psi)} \\
&= \prod_{i=1}^n \frac{g(y_i | x_i; \theta) p(R_i | x_i, y_i; \psi)}{p((1 - R_i) y_i | R_i, R_i y_i, x_i; \theta, \psi)}
\end{aligned}$$

The corresponding log-likelihood function is

$$\begin{aligned}
l(\theta, \psi; \mathcal{D}_{obs}) &= \sum_{i=1}^n \{ \log g(y_i | x_i; \theta) + \log p(R_i | x_i, y_i; \psi) - \log p((1 - R_i) y_i | R_i, R_i y_i, x_i; \theta, \psi) \} \quad (3.2)
\end{aligned}$$

The MLE is

$$(\hat{\theta}, \hat{\psi}) = \arg \max_{\theta, \psi} l(\theta, \psi; \mathcal{D}_{obs})$$

The expectation-maximization (EM) algorithm is an iterative algorithm that is often used to find the maximum likelihood estimate [Dempster et al., 1977]. Little and Rubin showed that [Little and Rubin, 2002]

$$l(\theta, \psi | \mathcal{D}_{obs}) = l(\theta, \psi | \mathcal{D}) - \ln f(\mathcal{D}_{mis} | \mathcal{D}_{obs}, \theta, \psi)$$

where \mathcal{D}_{mis} is the missing data and $\mathcal{D} = (\mathcal{D}_{obs}, \mathcal{D}_{mis})$ is the hypothetical complete data . The expectation of log-likelihood function given a current estimate of θ , say $\tilde{\theta}$, is

$$l(\theta, \psi | \mathcal{D}_{obs}) = E[l(\theta, \psi | \mathcal{D}) | \tilde{\theta}, \tilde{\psi}] - E[\ln f(\mathcal{D}_{mis} | \mathcal{D}_{obs}, \theta, \psi) | \tilde{\theta}, \tilde{\psi}] \quad (3.3)$$

Denote $Q(\theta, \psi | \tilde{\theta}, \tilde{\psi}) = E[l(\theta, \psi | \mathcal{D}) | \tilde{\theta}, \tilde{\psi}]$ and $H(\theta, \psi | \tilde{\theta}, \tilde{\psi}) = E[\ln f(\mathcal{D}_{mis} | \mathcal{D}_{obs}, \theta, \psi) | \tilde{\theta}, \tilde{\psi}]$, which yields $l(\theta, \psi | \mathcal{D}_{obs}) = Q(\theta, \psi | \tilde{\theta}, \tilde{\psi}) - H(\theta, \psi | \tilde{\theta}, \tilde{\psi})$.

Each iteration of the EM algorithm include an E-step and an M-step. In the E-step, given the current estimate $(\theta^{(t)}, \psi^{(t)})$, we need to calculate

$$Q(\theta, \psi | \theta^{(t)}, \psi^{(t)}) = \sum_{i=1}^n E[\{ \log g(y_i | x_i; \theta) + \log p(R_i | x_i, y_i; \psi) \} | R_i, R_i y_i, x_i; \theta^{(t)}, \psi^{(t)}]$$

In the M-step, we update the parameters by finding the maximizer of the Q-function giving the current estimate of parameters

$$(\theta^{(t+1)}, \psi^{(t+1)}) = \arg \max_{(\theta, \psi)} Q(\theta, \psi; \theta^{(t)}, \psi^{(t)})$$

By Jansen's Inequality, we have

$$H(\theta, \psi; \theta^{(t)}, \psi^{(t)}) \leq H(\theta^{(t)}, \psi^{(t)}; \theta^{(t)}, \psi^{(t)})$$

this yield that $l(\theta^{(t+1)}, \psi^{(t+1)}; \mathcal{D}_{obs}) \geq l(\theta^{(t)}, \psi^{(t)}; \mathcal{D}_{obs})$. Or in another way, maximizing log-likelihood $l(\theta, \psi | \mathcal{D}_{obs})$ with respect to (θ, ψ) is equivalent as maximizing $Q(\theta, \psi; \theta^{(t)}, \psi^{(t)})$ with respect to (θ, ψ) iteratively. Therefore the EM algorithm ensures increasing of the log-likelihood, and the estimate of parameters converges at the MLE.

EM algorithm is relatively slow in convergence. Sometimes in the M-step, we do not have explicit form of solution even the underlying complete data are from exponential family. EM algorithm is stable that the likelihood function is always increasing.

However, in such maximum-likelihood method, since (θ, ψ) have to be estimated/updated together, misspecification of the missing data model could lead to biased estimate of θ , and usually researchers do not have information on the missing-data mechanism. This is the motivation of our modified EM algorithm.

3.3 A MODIFIED EM ALGORITHM FOR REGRESSION ANALYSIS OF DATA WITH NON-RESPONSES

We propose the modified EM algorithm using the following reasoning: when data are missing not at random (NMAR), and $w(x, y; \psi)$ is unknown, if we can update $\theta^{(t)}$ without using information on ψ , the method will be useful in practice by ignoring the missing-data mechanism.

3.3.1 EM algorithm for regression analysis of data with non-responses when missing-data mechanism is known

Consider when $\psi = \psi_0$ is known, the MLE is $\hat{\theta} = \arg \max l(\theta, \psi_0; \mathcal{D}_{obs})$, where

$$l(\theta, \psi_0; \mathcal{D}_{obs}) = \sum_{i=1}^n \{[\log g(y_i|x_i; \theta) + \log p(R_i|x_i, y_i; \psi_0) - \log p((1 - R_i)y_i)|R_i, R_i y_i, x_i; \theta, \psi_0)]\}$$

Given current estimate $\theta^{(t)}$, the corresponding Q-function becomes

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= \sum_{i=1}^n E[\{\log g(y_i|x_i; \theta) + \log p(R_i|x_i, y_i; \psi_0)\}|R_i, R_i y_i, x_i; \theta^{(t)}, \psi_0] \\ &\propto \sum_{i=1}^n E[\log g(y_i|x_i; \theta)|R_i, R_i y_i, x_i; \theta^{(t)}, \psi_0] \end{aligned}$$

Here the second term of Q-function is ignored because it is not related to θ . In the M-step

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta; \theta^{(t)})$$

The computation of Q-function in E-step is also simplified to

$$Q(\theta; \theta^{(t)}) \propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n \int \log g(y_i|x_i; \theta) g(y_i|x_i; \theta^{(t)}) (1 - w(x_i, y_i; \psi_0)) dy$$

where the first m subjects are observed, and the first term in the expectation can be directly computed from the data; the rest $(n - m)$ subjects have missing responses, and the computation of the second term depends on the missing-data mechanism $w(x, y; \psi_0)$.

Without loss of generality, consider the exponential family with canonical link as in (3.1), we have

$$\begin{aligned} Q(\theta; \theta^{(t)}) &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n \int \log g(y_i|x_i; \theta) g(y_i|x_i; \theta^{(t)}) (1 - w(x_i, y_i; \psi_0)) dy \\ &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n \int \{\theta S(x_i, y_i) + a(\theta)\} g(y_i|x_i; \theta^{(t)}) (1 - w(x_i, y_i; \psi_0)) dy \\ &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \theta \sum_{i=1}^m E[S(x_i, y_i)|x_i, R_i = 0, \theta^{(t)}, \psi_0] + (n - m)a(\theta) \end{aligned}$$

Calculating $Q(\theta; \theta^{(t)})$ is equivalent as updating $E[S(x_i, y_i)|x_i, R_i = 0, \theta^{(t)}, \psi_0]$:

$$E[S(x_i, y_i)|x_i, R_i = 0, \tilde{\theta}, \psi_0] = \frac{\int S(x_i, y_i)w(x_i, y, \psi_0)g(y|x_i; \tilde{\theta})dy}{pr[R_i = 0|x_i, \tilde{\theta}, \psi_0]}$$

Therefore, with known missing-data mechanism, we can updating the Q-function by iteratively applying the EM algorithm till convergence of θ .

3.3.2 An modified EM algorithm

For updating $E[S(x_i, y_i)|x_i, R_i = 0, \theta^{(t)}, \psi_0]$, alternatively we have

$$\begin{aligned} E[S(x_i, y_i)|x_i, \theta^{(t)}, \psi_0] &= E[S(x_i, y_i)|x_i, R_i = 1, \theta^{(t)}, \psi_0] \cdot pr[R_i = 1|x_i, \theta^{(t)}, \psi_0] \\ &\quad + E[S(x_i, y_i)|x_i, R_i = 0, \theta^{(t)}, \psi_0] \cdot pr[R_i = 0|x_i, \theta^{(t)}, \psi_0] \end{aligned}$$

Then we have

$$\begin{aligned} &E[S(x_i, y_i)|x_i, R_i = 0, \theta^{(t)}, \psi_0] \\ &= \frac{E[S(x_i, y_i)|x_i, \theta^{(t)}, \psi_0] - E[S(x_i, y_i)|x_i, R_i = 1, \theta^{(t)}, \psi_0] \cdot pr[R_i = 1|x_i, \theta^{(t)}, \psi_0]}{pr[R_i = 0|x_i, \theta^{(t)}, \psi_0]} \quad (3.4) \end{aligned}$$

where $E[S(x_i, y_i)|x_i, R_i = 1, \theta^{(t)}, \psi_0]$, $pr[R_i = 1|x_i, \theta^{(t)}, \psi_0]$ and $pr[R_i = 0|x_i, \theta^{(t)}, \psi_0]$ are calculated as:

$$\begin{aligned} E[S(x_i, y_i)|x_i, R_i = 1, \theta^{(t)}, \psi_0] &= \frac{\int S(x_i, y_i)w(x_i, y, \psi_0)g(y|x_i; \theta^{(t)})dy}{pr[R_i = 1|x_i, \theta^{(t)}, \psi_0]} \\ pr[R_i = 1|x_i, \theta^{(t)}, \psi_0] &= \int w(x_i, y, \psi_0)g(y|x_i; \theta^{(t)})dy \\ pr[R_i = 0|x_i, \theta^{(t)}, \psi_0] &= 1 - pr[R_i = 1|x_i, \theta^{(t)}, \psi_0] \end{aligned}$$

To compute the expectation of sufficient statistics giving missing, missing-data mechanism is required. However, alternatively the terms $E[S(x_i, y_i)|x_i, R_i = 1, \theta^{(t)}, \psi_0]$, $pr[R_i = 1|x_i, \theta^{(t)}, \psi_0]$ and $pr[R_i = 0|x_i, \theta^{(t)}, \psi_0]$ in (3.4) can be replaced with empirical estimates $\hat{E}[S(x_i, y_i)|x_i, R_i = 1]$, $\hat{pr}[R_i = 1|x_i]$, and $\hat{pr}[R_i = 0|x_i]$ via nonparametric regression methods using the observed data (x_i, y_i, R_i) . The missing-data mechanism $w(x_i, y, \psi_0)$ is not

required in such nonparametric smoothing process. The expectation of sufficient statistics in the missing pattern can be approximated as following:

$$\hat{E}[S(x_i, y_i)|x_i, R_i = 0, \theta^{(t)}] = \frac{E[S(x_i, y_i)|x_i, \theta^{(t)}] - \hat{E}[S(x_i, y_i)|x_i, R_i = 1] \cdot \hat{pr}[R_i = 1|x_i]}{\hat{pr}[R_i = 0|x_i]} \quad (3.5)$$

After the expectation of sufficient statistics given θ is obtained, the approximated Q-function becomes

$$\begin{aligned} Q(\theta; \tilde{\theta}) &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n E[\log g(y_i|x_i; \theta)|x_i, R_i = 0; \tilde{\theta}, \psi_0] \\ &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n \{\theta \hat{E}[S(x_i, y_i)|x_i, R_i = 0; \tilde{\theta}] + a(\theta)\} \quad (3.6) \end{aligned}$$

In order to obtain valid approximation in (3.6), $\tilde{\theta}$ has to be a consistent estimate of θ , such that empirical estimates in (3.5) be consistent with the original terms. Therefore, a reasonable initial value of θ is very important for starting the proposed modified EM algorithm. In general, we would need an external dataset without missing values for estimate the parameters for the start of EM algorithm in the main data, and this is the trade-off of the empirical estimation replacement.

In practice, we can obtain the initial value from an external data set with no missing values, or from the recall results of a random subset of original sample \mathcal{D}_{obs} . For simplicity, all these complete small set are named as external set \mathcal{E} for subsequent context.

Therefore, we have the proposed modified EM algorithm outlined as follows:

I. The preparation of the modified EM algorithm.

- (i) We obtain empirical estimates of $E[S(x_i, y_i)|x_i, R_i = 1]$ and $pr[R_i = 1|x_i]$ via nonparametric smoothing methods using the observed data (x_i, y_i, R_i) .
- (ii) We obtain an estimate of θ via an external data set with no missing values, and use it as the initial value $\theta^{(0)}$ to start the EM algorithm.

- II. In the E-step of the t^{th} iteration, given the current estimate $\theta^{(t)}$, we update the Q-function as in (3.5) and (3.6) using empirical estimates of $E[S(x_i, y_i)|x_i, R_i = 1]$ and $pr[R_i = 1|x_i]$ from the preparation step.
- III. In the M-step, we update the values of the parameters as $\theta^{(t+1)}$ by maximizing $Q(\theta; \theta^{(t)})$ with respect to θ .
- IV. We repeat the E-step and the M-step iteratively till convergence of θ .

The variance of estimates $\hat{\theta}$ from the modified EM algorithm is not directly accessible because the full likelihood function does not have explicit form and the observed information matrix can not be computed. Therefore, in implementation, we use bootstrap re-sampling method to compute an empirical variance for the estimates from modified EM algorithm.

3.3.3 Implementation of proposed methods in simple linear regression

As an example, consider simple linear regression as

$$y_i = \beta_0 + \beta_1 x_i + N(0, \sigma^2) \tag{3.7}$$

when there are no missing values, the MLE of $\theta = (\beta_0, \beta_1, \sigma^2)$ can be calculated as

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \tag{3.8}$$

where $\bar{x} = \sum_{i=1}^n x_i$ and $\bar{y} = \sum_{i=1}^n y_i$. Let the sufficient statistics from complete-data be $SS = (S_1, S_2, S_3)$, where

$$\begin{aligned} S_1 &= \sum_{i=1}^n y_i \\ S_2 &= \sum_{i=1}^n x_i y_i \\ S_3 &= \sum_{i=1}^n y_i^2 \end{aligned}$$

Suppose the responses Y 's have missing values (observed for $i = 1, \dots, m$) and covariate X 's fully observed. In E-step of the $(t + 1)$ th iteration of the modified EM algorithm, We compute $E[S_k|X, Y_{obs}; \theta^{(t)}], k = 1, 2, 3$, expectation of sufficient statistics given observed data and current parameter estimates $\theta^{(t)}$.

$$\begin{aligned}
S_1^{(t)} &= E[S_1|\mathcal{D}_{obs}; \theta^{(t)}] \\
&= E\left[\sum_{i=1}^n y_i|\mathcal{D}_{obs}; \theta^{(t)}\right] \\
&= \sum_{i=1}^m y_i + E\left[\sum_{i=m+1}^n y_i|x_i, R_i = 0, \theta^{(t)}\right] \\
&= \sum_{i=1}^m y_i + \sum_{i=m+1}^n \frac{\beta_0^{(t)} + \beta_1^{(t)}x_i - \hat{E}[y|x_i, R = 1]\hat{p}r(R = 1|x_i)}{\hat{p}r(R = 0|x_i)}
\end{aligned}$$

In the above formula, $\hat{E}[y|x_i, R = 1]$ and $\hat{p}r(R = 1|x_i)$ are empirical estimates from observed data via non-parametric regression.

Similarly, we have

$$\begin{aligned}
S_2^{(t)} &= E[S_2|\mathcal{D}_{obs}; \theta^{(t)}] \\
&= E\left[\sum_{i=1}^n x_i y_i|\mathcal{D}_{obs}; \theta^{(t)}\right] \\
&= \sum_{i=1}^m x_i y_i + E\left[\sum_{i=m+1}^n x_i y_i|x_i, R_i = 0, \theta^{(t)}\right] \\
&= \sum_{i=1}^m x_i y_i + \sum_{i=m+1}^n x_i \left\{ \frac{\beta_0^{(t)} + \beta_1^{(t)}x_i - \hat{E}[y|x_i, R = 1]\hat{p}r(R = 1|x_i)}{\hat{p}r(R = 0|x_i)} \right\}
\end{aligned}$$

and

$$\begin{aligned}
S_3^{(t)} &= E[S_3|\mathcal{D}_{obs}; \theta^{(t)}] \\
&= E\left[\sum_{i=1}^n y_i^2|\mathcal{D}_{obs}; \theta^{(t)}\right] \\
&= \sum_{i=1}^m y_i^2 + E\left[\sum_{i=m+1}^n y_i^2|x_i, R_i = 0, \theta^{(t)}\right] \\
&= \sum_{i=1}^m y_i^2 + \sum_{i=m+1}^n \frac{(\beta_0^{(t)} + \beta_1^{(t)}x_i)^2 + \sigma^2 - \hat{E}[y^2|x_i, R = 1]\hat{p}r(R = 1|x_i)}{\hat{p}r(R = 0|x_i)}
\end{aligned}$$

In the M-step, by replacing the data part in (3.8) with the updated expectation of sufficient statistics, we have θ updated as

$$\beta_1^{(t+1)} = \frac{S_2^{(t)} - \bar{x}S_1^{(t)}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_2^{(t)} - (\frac{1}{n} \sum_{i=1}^n x_i)S_1^{(t)}}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \quad (3.9)$$

$$\beta_0^{(t+1)} = \frac{1}{n} S_1^{(t)} - \beta_1^{(t+1)} \bar{x} = \frac{1}{n} S_1^{(t)} - \beta_1^{(t+1)} \frac{1}{n} \sum_{i=1}^n x_i \quad (3.10)$$

$$\begin{aligned} \sigma^{2(t+1)} &= \frac{1}{n} \{S_3^{(t)} - \frac{1}{n} (S_1^{(t)})^2 - \beta_1^{(t+1)2} \sum_{i=1}^n (x_i - \bar{x})^2\} \\ &= \frac{1}{n} \{S_3^{(t)} - \frac{1}{n} (S_1^{(t)})^2 - \beta_1^{(t+1)2} [\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2]\} \end{aligned} \quad (3.11)$$

3.3.4 Improvement of modified EM algorithm by involving external data

Let $\mathcal{E}_i = (x_i, y_i)$, $i = n+1, \dots, n+n^*$ be the external data. In this complete external data \mathcal{E} with sample size n^* , (x, y) are all observed. Adding the information of \mathcal{E} into the former log-likelihood (3.2) gives the new log-likelihood:

$$\begin{aligned} l(\theta, \psi_0; \mathcal{D}_{obs}, \mathcal{E}) &= \sum_{i=1}^n \{[\log g(y_i|x_i; \theta) + \log p(R_i|x_i, y_i; \psi_0)] \\ &\quad - \log p((1 - R_i)y_i|R_i, R_i y_i, x_i; \theta, \psi_0)\} + \sum_{i=n+1}^{n+n^*} \log g(y_i|x_i; \theta) \end{aligned}$$

and compared with (3.6), the new Q-function is

$$\begin{aligned} Q(\theta; \tilde{\theta}) &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=n+1}^{n+n^*} \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n E[\log p(y_i|x_i; \theta)|x_i, R_i = 0; \tilde{\theta}, \psi_0] \\ &\propto \sum_{i=1}^m \log g(y_i|x_i; \theta) + \sum_{i=n+1}^{n+n^*} \log g(y_i|x_i; \theta) + \sum_{i=m+1}^n \{\theta \hat{E}[S(x_i, y_i)|x_i, R_i = 0; \tilde{\theta}] + a(\theta)\} \end{aligned}$$

The replacement of expectation of sufficient statistics terms with empirical estimate remains the same as shown before. When the external complete data \mathcal{E} with sample size n^* are added in the log-likelihood function, we benefit more in the efficiency of the estimates from the modified EM algorithm.

3.3.5 The modified EM algorithm under discrete covariates

When X 's are discrete, calculation of the empirical estimate terms in (3.6) can be simplified as computing the mean in the group with the same given value of X . For example, when $x_i \in (1, 2, \dots, L)$

$$\begin{aligned}\hat{E}[S(x_i, y_i)|x_i = l, R_i = 1, \theta^{(t)}] &= \frac{\sum_{j=1}^m S(l, y_j; \theta^{(t)}) \cdot I\{x_j = l\}}{\sum_{j=1}^m I\{x_j = l\}} \\ \hat{p}r[R_i = 1|x_i = l] &= \frac{\sum_{j=1}^n R_j \cdot I\{x_j = l\}}{\sum_{j=1}^n I\{x_j = l\}} \\ \hat{p}r[R_i = 0|x_i = l] &= 1 - \hat{p}r[R_i = 1|x_i = l]\end{aligned}$$

3.3.6 Smoothing methods and its implementation in the modified EM algorithm

When X 's are continuous, the calculation of the empirical estimates $\hat{E}[y|x_i, R = 1]$ and $\hat{p}r(R = 1|x_i)$ requires smoothing methods.

Given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$, where $X_i, Y_i \in \mathbb{R}$, non-parametric regression or smoothing is concerned with estimating the regression function $m(x) = E(Y|X = x)$, predicting Y given X from joint distribution of X and Y . We can write

$$Y = m(X) + \epsilon$$

A one-dimensional smoothing Kernel is any smooth function K such that $K(x) \geq 0$ and

$$\int K(x) = 1, \int xK(x) = 0 \text{ and } \sigma_K^2 \equiv \int x^2 K(x) > 0$$

Let $h > 0$ be bandwidth. *Localized Square Error* is defined by

$$\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) (\hat{m}(x) - Y_i)^2$$

By minimizing Localized Square Error, we have The Nadaraya-Watson kernel estimator

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)} = \sum_{i=1}^n Y_i l_i(x)$$

where $l_i(x) = \frac{K(\frac{x-X_i}{h})}{\sum_{i=1}^n K(\frac{x-X_i}{h})}$. Thus $\hat{m}_h(x)$ is a local average of Y_i 's. Here is the functional form of Epanechnikov kernel

$$K(u) = \frac{3}{4}(1 - u^2)I_{|u| \leq 1}$$

Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$$

Box kernel

$$K(u) = \frac{1}{2}I_{|u| \leq 1}$$

As described in section 3.3.2, in E-step of the proposed modified EM algorithm for regression analysis of data with non-ignorable non-responses, we need to replace $E[S(x_i, y)|x_i, R_i = 1, \tilde{\theta}, \psi_0]$, $pr[R_i = 1|x_i, \tilde{\theta}, \psi_0]$ and $pr[R_i = 0|x_i, \tilde{\theta}, \psi_0]$ with empirical estimates $\hat{E}[S(x_i, y_i)|x_i, R_i = 1]$, $\hat{pr}[R_i = 1|x_i]$ and $\hat{pr}[R_i = 0|x_i]$. With the Nadaraya-Watson kernel estimator we have

$$\hat{E}[S(x_i, y_i)|x_i, R_i = 1, \tilde{\theta}] = \hat{m}_{1h}(x_i) = \frac{\sum_{j=1}^m S(X_j, Y_j)K(\frac{x_i-X_j}{h})}{\sum_{j=1}^m K(\frac{x_i-X_j}{h})} = \sum_{j=1}^m S(X_j, Y_j)l_j(x_i)$$

$$\hat{pr}[R_i = 1|x_i] = \hat{m}_{2h}(x_i) = \frac{\sum_{j=1}^n R_j K(\frac{x_i-X_j}{h})}{\sum_{j=1}^n K(\frac{x_i-X_j}{h})} = \sum_{j=1}^n R_j l_j(x_i)$$

Bandwidth selection is important in kernel regression. To simplify, we use fixed bandwidth $h = 0.1$ after normalizing the data.

We use R function (ksmooth)¹ for the nonparametric regression in calculating the empirical estimates $\hat{E}[S(x_i, y_i)|x_i, R_i = 1]$ and $\hat{pr}[R_i = 1|x_i]$. To improve the computing performance of subroutines, we also develop our own R packages using C language for N-W kernel smoothing with Epanechnikov, Gaussian and box kernel. The subroutines of modified EM algorithms are written in R language.

¹<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/ksmooth.html>

3.4 SIMULATION STUDIES

3.4.1 Simulation setup

In each simulation, data $\mathcal{D}_{obs} = (x_i, R_i y_i), i = 1, \dots, n$ and $\mathcal{E} = (x_i, y_i), i = n + 1, \dots, n + n^*$ are generated following a simple linear function as in (3.7). The regression parameters are $\theta = (\beta_0, \beta_1, \sigma^2)$. Covariate x_i 's are fully observed and response y_i 's are observed for $i = 1, \dots, m$. The missing-data indicator R_i 's are generated following a missing data mechanism as

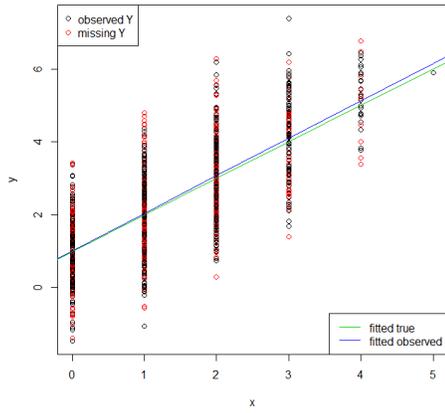
$$pr(R_i = 1|x_i, y_i; \psi) = \Phi(\psi_0 + \psi_1 x_i + \psi_2 y_i) \quad (3.12)$$

where $\psi = (\psi_0, \psi_1, \psi_2)$ are the missing-data mechanism parameters.

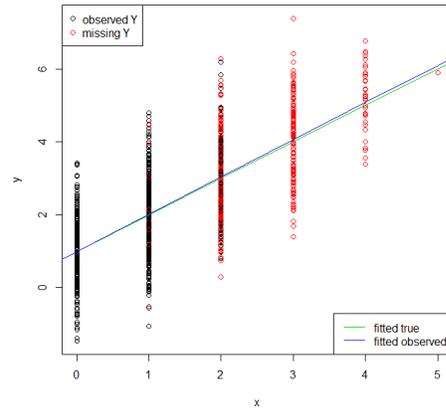
In each simulation, we generate an external data with complete records and sample size $n^* = 200$ and another dataset with the response y is subject to missing values with sample size $n = 1000$. Covariate X was either generated from $X \sim N(0, 1)$ or $X \sim BIN(5, 0.3)$ for the consideration of continuous X and discrete X . Response Y s were generated following (3.7) with $\theta = (1, 1, 1)$. When X s are discrete, missingness of response Y s were generated following (3.12) with $\psi = (-0.3, 0, 0)$ for MCAR, $\psi = (-3.6, 2, 0)$ for MAR and $\psi = (-6.3, 2, 1)$ for NMAR. When X s are continuous, missingness of response Y s were generated following (3.12) $\psi = (-0.3, 0, 0)$ for MCAR, $\psi = (-0.7, 2, 0)$ for MAR and $\psi = (-2, 2, 1)$ for NMAR. The response missing rates are around 38% for all scenarios.

Plots of simulated data are shown in Figure 14. The red points are data with missing responses Y , and black points are complete cases. Green lines are the underlying true linear functions, and blue lines are fitted regression line using complete cases only. We can tell that for data with responses missing not at random, the estimates could be severely biased, while for data with response missing completely at random or missing at random, the estimates are very close to the true regression line in the plots.

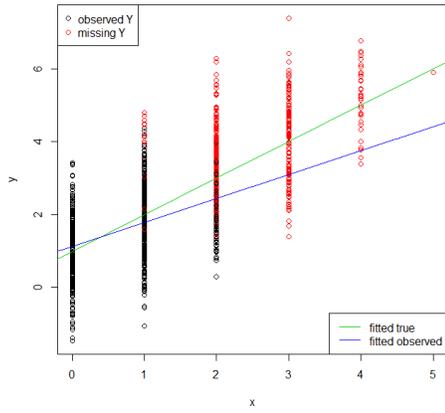
The following methods were considered for estimation of regression parameters in the simulation study: the MLE from the analysis of the external data \mathcal{E} , complete-case analysis



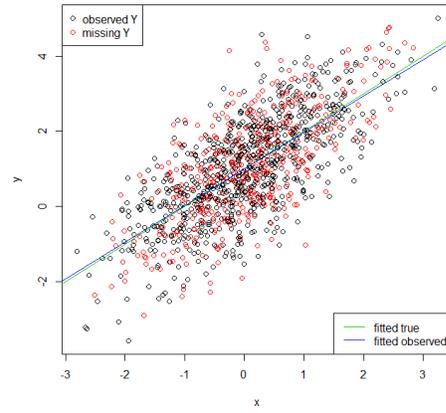
(a) Y is MCAR & $X \sim \text{BIN}(5, 0.3)$



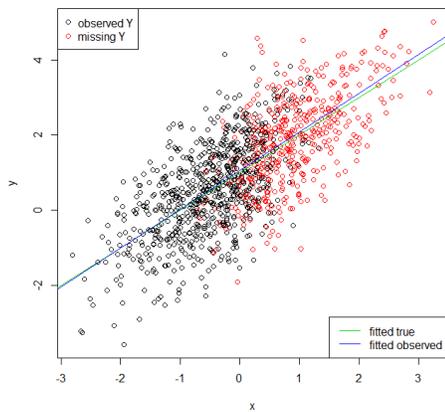
(b) Y is MAR & $X \sim \text{BIN}(5, 0.3)$



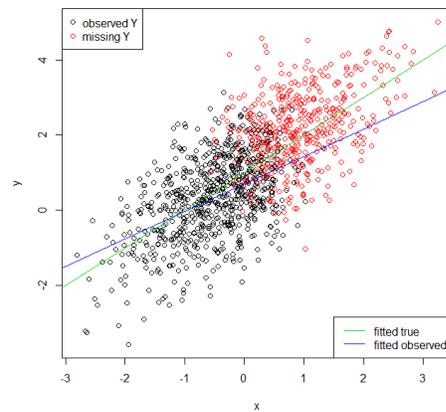
(c) Y is NMAR & $X \sim \text{BIN}(5, 0.3)$



(d) Y is MCAR & $X \sim N(0, 1)$



(e) Y is MAR & $X \sim N(0, 1)$



(f) Y is NMAR & $X \sim N(0, 1)$

Figure 14: Plots of simulated data (x, y) in different scenarios.

based on \mathcal{E} and \mathcal{D}_{obs} , the MLE from the analysis of the underlying complete data (CD), the proposed modified EM algorithm using \mathcal{D}_{obs} and \mathcal{E} (proposed methods). In Modified EM algorithm using \mathcal{D}_{obs} and \mathcal{E} method, when X is continuous, we used box kernel, Gaussian kernel and Epanechnikov kernel in empirical estimation of the expectation of sufficient statistics given missing.

Simulations were repeated for $K=1000$ times. For each simulated dataset, each method will be applied in estimating parameters. The empirical bias and empirical standard deviation of estimates from the above-mentioned methods were computed and compared. Bootstrap re-sampling for the estimation of regression parameters was performed for $B=200$ times in each simulation, and the 95%CI coverage from each methods are computed and compared.

3.4.2 Simulation results

Table 5 summarizes the performance of these methods on inference of distribution parameter $\theta = (\beta_0, \beta_1, \sigma^2)$ for regression of simulated data with discrete predictor X and response Y MCAR, MAR and NMAR. Predictor X 's were generated following a binomial distribution $X \sim BIN(5, 0.3)$.

Table 6 summarizes the methods' performance on inference of distribution parameter $\theta = (\beta_0, \beta_1, \sigma^2)$ for regression of simulated data with continuous predictor X and response Y missing completely at random (MCAR), missing at random (MAR) and missing not at random (NMAR). Predictor X 's were generated following a normal distribution $X \sim N(0, 1)$.

From Table 5 and Table 6, when response Y is missing completely at random (MCAR) or missing at random (MAR), the results suggest that the modified EM algorithm (except for Gaussian kernel when predictor X is continuous) yielded negligible bias, which is similar to the MLE from the external data set and the MLE using complete cases from the combined data set. The 95% Bootstrap confidence intervals from the modified EM algorithm (except

Table 5: Simulation results when the predictor $X \sim BIN(5, 0.3)$

Mechanism	Method	Emp. Bias			Emp. SD ^a			95%CI Coverage ^b		
		β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
MCAR	MLE for \mathcal{E}	-12	25	-111	1245	688	976	94.7%	94.2%	94.3%
	CC Analysis	-26	14	-46	618	341	490	94.6%	94.4%	94.3%
	CD MLE	-20	8	-34	526	297	404	94.1%	93.5%	94.8%
	Modified EM	-9	23	-107	1237	677	975	94.3%	95.0%	92.9%
MAR	MLE for \mathcal{E}	-12	25	-111	1245	688	976	94.7%	94.2%	94.3%
	CC Analysis	-32	27	-37	571	431	491	95.1%	94.8%	94.8%
	CD MLE	-20	8	-34	526	297	404	94.1%	93.5%	94.8%
	Modified EM	-17	27	-54	671	495	635	94.6%	94.2%	93.5%
NMAR	MLE for \mathcal{E}	-12	25	-111	1245	688	976	94.7%	94.2%	94.3%
	CC Analysis	156	-1412	-354	574	446	477	94.1%	10.4%	88.4%
	CD MLE	-20	8	-34	526	297	404	94.1%	93.5%	94.8%
	Modified EM	-16	27	-50	662	492	731	94.7%	93.9%	93.6%

^aAll values in the table $\times 10,000$

^bAsymptotic CI used for MLE using external data, complete-case analysis, and MLE using underlying complete data; Bootstrap re-sampling for $B = 200$ times used for modified EM.

Table 6: Simulation results when the predictor $X \sim N(0, 1)$

Mechanism	Method ^c	Emp. Bias			Emp. SD ^a			95%CI Coverage ^b		
		β_0	β_1	σ^2	β_0	β_1	σ^2	β_0	β_1	σ^2
MCAR	MLE for \mathcal{E}	-26	9	-103	693	743	991	95.7%	93.6%	95.2%
	Complete cases	-7	7	-26	350	350	494	94.0%	94.4%	95.4%
	CD MLE	-5	10	-15	289	289	408	95.0%	94.8%	96.0%
	MEM-Box	-22	28	-135	764	659	1052	97.2%	94.5%	95.7%
	MEM-Gaussian	-25	231	-381	650	577	897	95.1%	92.0%	89.5%
	MEM-Epan.	-23	48	-148	643	561	890	95.6%	93.2%	92.3%
MAR	MLE for \mathcal{E}	-26	9	-103	693	743	991	95.7%	93.6%	95.2%
	Complete cases	-12	1	-24	386	411	488	94.2%	95.6%	93.6%
	CD MLE	-5	10	-15	289	289	408	95.0%	94.8%	96.0%
	MEM-Box	-0.1	-1	-63	596	481	678	95.0%	94.8%	95.6%
	MEM-Gaussian	113	831	-151	558	476	640	93.5%	94.6%	92.8%
	MEM-Epan.	11	11	-73	556	472	641	93.9%	94.5%	93.7%
NMAR	MLE for \mathcal{E}	-26	9	-103	693	743	991	95.7%	93.6%	95.2%
	Complete cases	-2009	-1388	-486	380	422	468	0.0%	7.6%	81.4%
	CD MLE	-5	10	-15	289	289	408	95.0%	94.8%	96.0%
	MEM-Box	-69	2	-156	592	484	652	95.9%	95.2%	93.4%
	MEM-Gaussian	25	73	-248	552	476	624	94.0%	94.8%	91.1%
	MEM-Epan.	-97	-11	-161	551	472	624	95.0%	95.2%	93.4%

^aAll values in the table $\times 10,000$.

^bAsymptotic CI used for MLE using external data, complete-case analysis, and MLE using underlying complete data; Bootstrap re-sampling for $B = 200$ times used for modified EM.

^cMEM kernel smoothing bandwidth is set fixed at 0.1.

for Gaussian kernel when predictor X is continuous) had nominal coverage of the true values of regression parameters. The estimates from modified EM algorithm were more efficient than the ones from the external data but less efficient than the MLE using complete cases from the combined data set.

When response Y is NMAR, the modified EM algorithm yielded negligible bias, which is similar to the MLE from the external data set. The 95% Bootstrap confidence intervals from the modified EM algorithm (except for Gaussian kernel when predictor X is continuous) had nominal coverage of the true values of regression parameters. Based on MSE, our proposed modified EM algorithm gives the best performance considering the consistency and efficiency.

3.5 ANALYSIS OF QUALIFY-OF-LIFE DATA FROM A CANCER CLINICAL TRIAL

R-04 is a clinical trial conducted by NSABP² comparing preoperative radiation therapy and Capecitabine with or without Oxaliplatin with preoperative radiation therapy and continuous intravenous infusion of 5-Fluorouracil with or without Oxaliplatin in the treatment of patients with operable carcinoma of the rectum [Monga and O'Connell, 2006].

The quality-of-life score (QOL) in R-04 trial were measured for the patients before treatment (baseline) and 1 year after treatment. We use QOL at baseline as predictor X and QOL after treatment as response Y . There were no missingness in the data. The sample size is 1266. For the purpose of illustration, 200 patients were selected as external complete data set. In the rest, we artificially generated missing values in Y following (3.12) where $\psi = (-1.2, 2, 1)$. Therefore the responses are not missing at random (NMAR) and the proportion of non-response was about 38%. Figure 15 shows the data with generated missingness.

²The National Surgical Adjuvant Breast and Bowel Project. <http://www.nsabp.pitt.edu>

Table 7 summarizes the performance of the MLE from the analysis of the external data \mathcal{E} , Complete case analysis based on \mathcal{E} and \mathcal{D}_{obs} , and proposed modified EM algorithm method on inference of distribution parameter $\theta = (\beta_0, \beta_1, \sigma^2)$ for regression analysis of quality of life (QOL) data with continuous X (QOL at baseline) and responses Y (QOL measured 1 year after treatment) missing not at random (NMAR). The missing indicators are generated following a NMAR mechanism. The response missing rate is 38.4%. The regression estimate of θ from original data is $\hat{\theta} = (0.00, 0.57, 0.68)$.

The data analysis results suggest that when response Y is NMAR and predictor X is continuous, the modified EM algorithm yielded negligible bias, which is similar to the MLE from the external data set; The 95% Bootstrap confidence intervals from the modified EM algorithm covered the regression estimates from original data; The estimates from modified EM algorithm are more efficient than the ones from the external data but less efficient than the MLE using complete cases from the combined data set. However, the complete-case analysis based on the combined data set is severely biased and the confidence intervals did not cover the regression estimates from original data. The conclusions agree with our simulation studies.

3.6 SUMMARY AND DISCUSSION

We proposed a modified EM algorithm in maximum likelihood approach for missing-data analysis, when the modeling of the missing-data mechanism is not necessary.

Simulation studies and data analysis were performed to compare the proposed modified EM algorithm method with the MLE using external data and complete case analysis.

The simulation studies suggest that the modified EM algorithm yields estimates with negligible bias regardless of the nature of the missing-data mechanism. For data with dis-

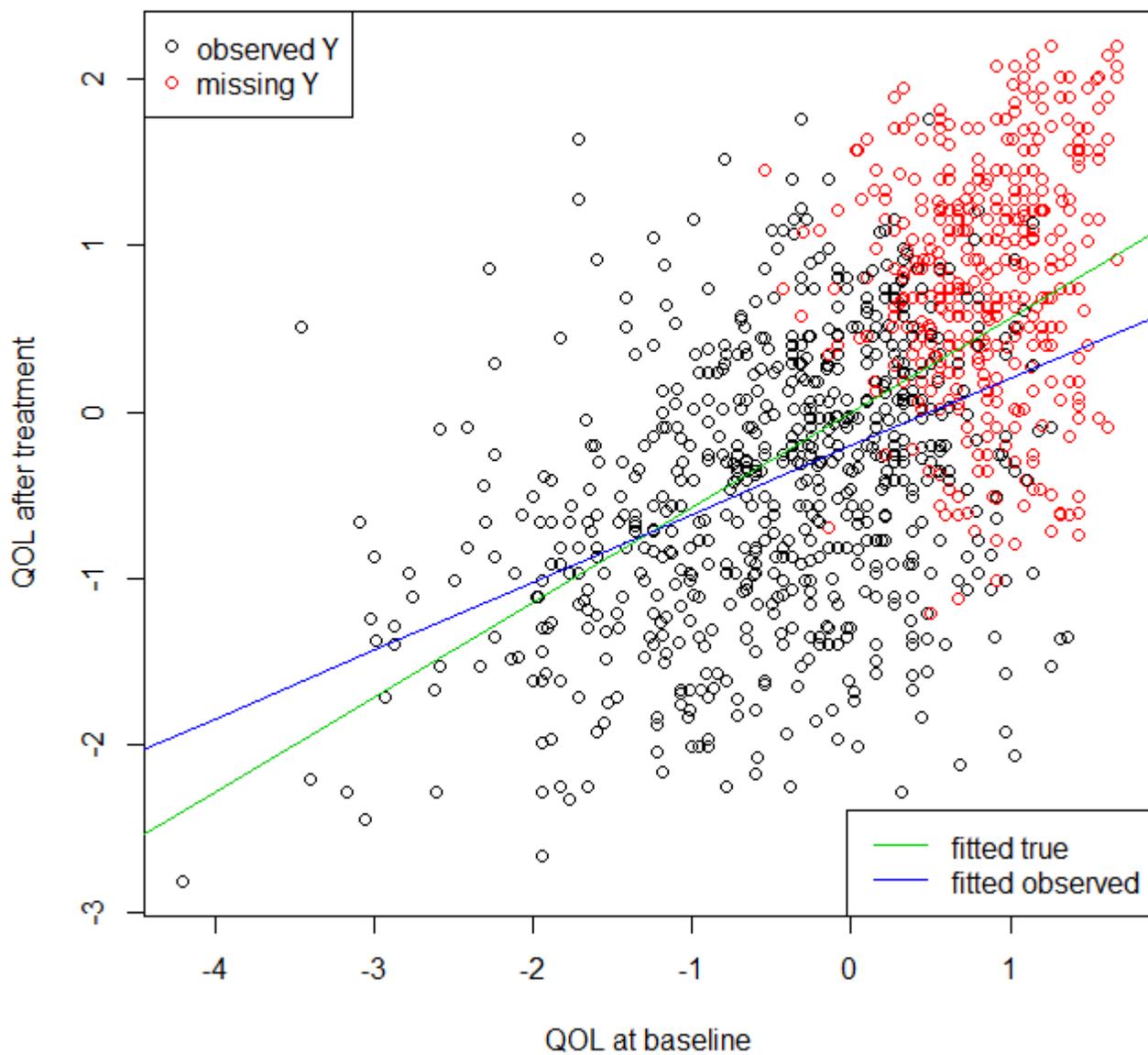


Figure 15: NSABP R-04 Trial: QOL before treatment vs QOL after treatment.

Table 7: Data analysis: patients' QOL in R-04 trial.

Method ^c	Smoothing ^d	Estimate Bias ^a			95% CI and width ^b		
		β_0	β_1	σ^2	β_0	β_1	σ^2
MLE \mathcal{E}		-0.02	-0.02	0.05	(-0.14,0.09)	(0.44,0.66)	(0.56,0.89)
					0.24	0.22	0.33
CC analysis		-0.20	-0.15	-0.02	(-0.26,-0.14)	(0.36,0.47)	(0.58,0.74)
					0.12	0.11	0.15
CD MLE		0.00	0.00	0.00	(-0.05,-0.05)	(0.52,0.61)	(0.61,0.74)
					0.09	0.09	0.13
Modified EM	Box	-0.07	-0.06	-0.02	(-0.18,0.04)	(0.43,0.59)	(0.57,0.75)
					0.22	0.16	0.17
	Gaussian	-0.05	-0.05	-0.02	(-0.15,0.05)	(0.44,0.59)	(0.57,0.75)
					0.20	0.16	0.18
	Epan.	-0.07	-0.06	-0.01	(-0.17,0.02)	(0.43,0.58)	(0.58,0.75)
					0.19	0.15	0.17

^aregression estimates from original data: $\theta = (0.00, 0.57, 0.68)$

^bAsymptotic CI's used for external data MLE, complete-case analysis and MLE using underlying complete data; Bootstrap for B=200 times used for modified EM methods

^cSample size 1266, external data $n^* = 200$, incomplete data $n = 1066$, response missing rate 38.4%.

^dMEM kernel smoothing bandwidth is set fixed $bw = 0.1$.

crete covariate X , the empirical bias of estimates is always smaller than 0.015, the empirical bias/ empirical standard deviation ratio is always smaller than $\frac{1}{9}$. For data with continuous covariate X , the empirical bias of estimates from modified EM method is slightly larger because of non-parametric regression boundary issue, but still less than 0.02 for Box and Epanechnikov kernel. The empirical bias/SD ratio is less than or around $\frac{1}{4}$ for Box and Epanechnikov kernel.

Second, the method does not require specification or modeling of the missing-data mechanism. This is the most important advantage of the proposed algorithm. The simulation results showed that no matter the data is MCAR, MAR or NMAR, the bias were mostly negligible and the coverage of the 95% CI were around the nominal level. Complete case analysis is biased when data are NMAR.

The estimates derived from the modified EM algorithm are more efficient than the MLE from the external complete data. Modified EM algorithm always have similar bias as the MLE from external data, and estimates from modified EM algorithm achieved an empirical SD (about 30%) smaller than the asymptotic SE of MLE using external data when data are MAR and MNAR.

In case where predictor X is discrete, the proposed modified EM algorithm outperform the MLE using external complete data, because the empirical estimate terms can be simplified as marginal means. When predictor X is continuous, the performance of the proposed methods is related with the kernel choice and bandwidth selection. From simulation results, Box and Epanechnikov kernel are appropriate choices for this algorithm.

In future research , we plan to improve the modified EM algorithm in magnitude of bias and efficiency when predictor X is continuous by the improving the approximation of the Q-function. We plan to work on bandwidth selection in the smoothing process and improving the estimation at boundary.

BIBLIOGRAPHY

- Dana C. Abraham, Ronald C. Jones, J. Harold Jones, Stephen E. and Cheek, George N. Peters, Sally M. Knox, Michael D. Grant, David W. Hampe, Daniel A. Savino, and Steven E. Harms. Evaluation of neoadjuvant chemotherapeutic response of locally advanced breast cancer by magnetic resonance imaging. *Cancer*, 78(1):91C100, 1996.
- K.S. Albain, S.M. Nag, G. Calderillo-Ruiz, J.P. Jordaan, A.C. Llombart, A. Pluzanska, J. Rolski, A.S. Melemed, J.M. Reyes-Vidal, J.S. Sekhon, L. Simms, and J. O'Shaughnessy. Gemcitabine plus paclitaxel versus paclitaxel monotherapy in patients with metastatic breast cancer and prior anthracycline treatment. *J. Clin. Oncol.*, 26:3950–7, 2008.
- H.D. Bear, S. Anderson, A. Brown, R. Smith, E.P. Mamounas, B. Fisher, R. Margolese, H. Theoret, A. Soran, D.L. Wickerham, N. Wolmark, National Surgical Adjuvant Breast, and Bowel Project Protocol B-27. The effect on tumor response of adding sequential preoperative docetaxel to preoperative doxorubicin and cyclophosphamide: preliminary results from national surgical adjuvant breast and bowel project protocol b-27. *J. Clin. Oncol.*, 21:4165–74, 2003.
- H.D. Bear, S. Anderson, R.E. Smith, C.E. Jr Geyer, E.P. Mamounas, B. Fisher, A.M. Brown, A. Robidoux, R. Margolese, M.S. Kahlenberg, S. Paik, A. Soran, D.L. Wickerham, and N. Wolmark. Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National surgical adjuvant breast and bowel project protocol b-27. *J. Clin. Oncol.*, 24:2019–27, 2006.
- H.D. Bear, G. Tang, P. Rastogi, C.E. Jr Geyer, A. Robidoux, J.N. Atkins, L. Baez-Diaz, A.M. Brufsky, R.S. Mehta, L. Fehrenbacher, J.A. Young, F.M. Senecal, R. Gaur, R.G. Margolese, P.T. Adams, H.M. Gross, J.P. Costantino, S.M. Swain, E.P. Mamounas, and N. Wolmark. Bevacizumab added to neoadjuvant chemotherapy for breast cancer. *N. Engl. J. Med.*, 366:310–20, 2012.
- S.L. Brower, J.E Fensterer, and J.E. Bush. *Methods in Molecular Biology, vol. 414: Apoptosis and Cancer*, chapter 6, pages 57–78. Humana Press, 2008.
- H.C. Cheng. The power issue: determination of kb or ki from ic50 - a closer look at the chengprusoff equation, the schild plot and related power equations. *Journal of Pharmaceutical and Toxicological Methods*, 46:61–71, 2002.

- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- P. J. Diggle and M. G. Kenward. Informative dropout in longitudinal data analysis. *Applied Statistics*, 43:49–94, 1994.
- B. Fisher, A. Brown, E. Mamounas, S. Wieand, A. Robidoux, R.G. Margolese, A.B. Jr Cruz, E.R. Fisher, D.L. Wickerham, N. Wolmark, A. DeCillis, J.L. Hoehn, A.W. Lees, and N.V. Dimitrov. Effect of preoperative chemotherapy on local-regional disease in women with operable breast cancer: findings from national surgical adjuvant breast and bowel project b-18. *J. Clin. Oncol.*, 15:2483–93, 1997.
- G. Fitzmaurice, G. Molenberghs, and S. Lipsitz. Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 4:691–704, 1995.
- S. Huang and L. Pang. Comparing statistical methods for quantifying drug sensitivity based on in vitro dose-response assays. *Anticancer Research*, 28:1733–1740, 2008.
- J.G. Ibrahim, H. Chu, and M.H. Chen. Missing data in clinical studies: issues and methods. *J. Clin. Oncol.*, 30(26):3297–303, 2012.
- K. Lange. *Applied Probability, 2nd Edition*. Springer, 2010.
- K.-Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1):13–22, 1986.
- R. A. Little. A class of pattern-mixture models for multivariate incomplete data. *Biometrika*, 81:471–483, 1994.
- R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002.
- R.J.A. Little, R. D’Agostino, K. Dickersin, S.S. Emerson, J.T. Farrar, C. Frangakis, J.W. Hogan, G. Molenberghs, S.A. Murphy, J.D. Neaton, A. Rotnitzky, D. Scharfstein, W.J. Shih, J.P. Siegel, and H. Stern. The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.*, 367:1355–1360, 2012.
- Z. Mi, F.A. Holmes, B. Hellerstedt, J. Pippen, R. Collea, A. Backner, J.E. Bush, H.H. Gallion, A. Wells, and J.A. O’Shaughnessy. Feasibility assessment of a chemoresponse assay to predict pathologic response in neoadjuvant chemotherapy for breast cancer patients. *Anticancer Research*, 28:1733–1740, 2008.
- Gunter von Minckwitz, Michael Untch, Jens-Uwe Blohmer, Serban D. Costa, Holger Eidtmann, Peter A. Fasching, Bernd Gerber, Wolfgang Eiermann, Jorn Hilfrich, Jens Huober, Christian Jackisch, Manfred Kaufmann, Gottfried E. Konecny, Carsten Denkert, Valentina Nekljudova, Keyur Mehta, and Sibylle Loibl. Definition and impact of pathologic complete

- response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J. Clin. Oncol.*, 30(15):1796–804, 2012.
- Alok Mishra and Mukesh Verma. Cancer biomarkers: Are we ready for the prime time? *Cancers*, 2:190–208, 2010.
- D.K. Monga and M.J. O’Connell. Surgical adjuvant therapy for colorectal cancer: Current approaches and future directions. *Ann. Surg. Oncol.*, 13(8):1021–34, 2006.
- J. O’ Shaughnessy, D. Miles, S. Vukelja, V. Moiseyenko, J.P. Ayoub, G. Cervantes, P. Fumoleau, S. Jones, W.Y. Lui, L. Mauriac, C. Twelves, G. Van Hazel, S. Verma, and R. Leonard. Superior survival with capecitabine plus docetaxel combination therapy in anthracycline-pretreated patients with advanced breast cancer: phase iii trial results. *J. Clin. Oncol.*, 20:2812–23, 2002.
- R.L. Ochs, D. Burholt, and P. Kornblith. *Methods in Molecular Medicine, vol. 110: Chemosensitivity: Vol. 1: In Vitro Assays*, chapter 13, pages 155–172. Humana Press, 2005.
- Christian Ritz and Jens C. Streibig. Bioassay analysis using r. *Journal of Statistical Software*, 12(5), 2005.
- N.J. Robert, V. Dieras, J. Glaspy, A.M. Brufsky, I. Bondarenko, O.N. Lipatov, E.A. Perez, D.A. Yardley, S.Y. Chan, X. Zhou, S.C. Phan, and J. O’Shaughnessy. Ribbon-1: randomized, double-blind, placebo-controlled, phase iii trial of chemotherapy with or without bevacizumab for first-line treatment of human epidermal growth factor receptor 2-negative, locally recurrent or metastatic breast cancer. *J. Clin. Oncol.*, 29:1252–60, 2011.
- J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.*, 89:846–866, 1994.
- D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
- D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.
- Sarah L. Suchy, Lauren M. Hancher, Dakun Wang, Paul R. Ervin Jr., and Stacey L. Brower. Chemoresponse assay for evaluating response to sunitinib in primary cultures of breast cancer. *Cancer Biology & Therapy*, 11(12):1059–1064, 2011.
- M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, Volume 82,:Issue 398, 1987.
- A.B. Troxel, S.R. Lipsitz, and T.A. Brennan. Weighted estimating equations with nonignorably missing response data. *Biometrics*, 53 (3):857–869, 1997.
- Mimi C. Yu, Myron J. Tong, Pierre Coursaget, Ronald K. Ross, Sugantha Govindarajan, and Brian E. Henderson. Prevalence of hepatitis b and c viral markers in black and white

patients with hepatocellular carcinoma in the united states. *J. Natl. Cancer Inst.*, 82 (12): 1038–1041, 1990.