Spring 5-11-2024

Statistical Approaches For The Early Detection Of Colorectal Cancer Using Longitudinal Biomarkers

Emily Berry
Southern Methodist University, eaberry@mail.smu.edu

STATISTICAL APPROACHES FOR
THE EARLY DETECTION OF COLORECTAL CANCER
USING LONGITUDINAL BIOMARKERS

Approved by:

Dr. Monnie McGee
Associate Professor

Dr. Daniel F. Heitjan
Professor

Dr. Stephen Robertson
Senior Lecturer

Dr. Steven Chiou
Assistant Professor

Dr. Amit Singal (External)
Professor - UT Southwestern Medical Center

STATISTICAL APPROACHES FOR THE EARLY DETECTION OF COLORECTAL CANCER USING LONGITUDINAL BIOMARKERS

A Dissertation Presented to the Graduate Faculty of the

Dedman College: School of Humanities and Sciences

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

Emily A. Berry

(B.S., Texas A&M University - College Station)
(M.S.P.H, Texas A&M University School of Public Health - College Station)

May 11, 2024

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Monnie McGee for shepherding me through the dissertation experience and without whom I would not have been able to complete the process. I am so incredibly grateful for the promise you saw in me, particularly when I was unable to see it myself. I would like to express my gratitude to Dr. Daniel F. Heitjan for continuously challenging me to rise to the next level and allowing me the space and opportunity to see what I was truly capable of achieving. I would also like to thank Dr. Steven Chiou for taking an interest in this project and providing his expertise to aid in addressing "real-world" data weirdness. I would like to thank Amit Singal for his support along the way and his commitment to ensuring I was not "building a bridge to nowhere" through these efforts. I also want to thank Dr. Stephen Robertson for his help during my time at SMU. Without your teaching and mentorship, I truly would never have survived in this program; your course notebook continues to be my best reference.

I am also very grateful to Dr. Keith E. Argenbright for his continued mentorship and support as I struggled through achievement. I am so appreciative of your patience and for providing the space necessary for me to pursue this goal.

Lastly, I dedicate this work to my husband, Jonathan, and my children, Keathley, Gus, and Ainsley. I hope to repay the sacrifices you have endured for me to pursue this dream with the many opportunities for joy and success in our future.

Berry , Emily A. B.S., Texas A&M University - College Station
M.S.P.H, Texas A&M University School of Public Health - College Station

Statistical approaches for
the early detection of colorectal cancer
using longitudinal biomarkers

Advisor: Monnie McGee

Doctor of Philosophy degree conferred May 11, 2024
Dissertation completed March 30, 2024

Colorectal cancer (CRC) is the third leading cause of cancer-related death in the United States [45]. CRC is believed to advance from adenomatous polyps creating a unique opportunity for both early detection and cancer prevention [4, 23]. Like other diseases, CRC screening reduces mortality by detecting cancer at earlier, more treatable stages; however, it can also reduce incidence through the removal of precancerous lesions [4]. As a result, screening is recommended for average-risk adults $\geq$ 45 years of age and includes a variety of tests [4, 12]. Despite alternate screening options, colonoscopy capacity is often cited as a barrier in colorectal cancer (CRC) screening [28, 39, 44]. In this dissertation, we address capacity as a statistical problem rather than a resource one.

In the first part, we apply methods developed to incorporate longitudinal biomarkers for ovarian cancer screening to the data accumulated through a large FIT-based CRC screening program. This requires us to consider multiple methods to accommodate the necessary data transformation given the range for quantitative fecal hemoglobin concentration.

The second part of the dissertation looks at a new diagnostic marker obtained by extracting information from the biomarker trajectories using functional data analysis. The approach addresses problems of missing data and verification bias. Performance however is hindered by data sparsity which can be attributed to the screening process.

The third part of the dissertation revisits the method highlighted in part one to derive and evaluate a decision threshold for clinical implementation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

To my husband, Jonathan, your encouragement and support of me goes beyond what words can adequately express. You are my person, and I am so incredibly grateful to be able to have done this with you by my side.

To my children, Keathley, Gus, and Ainsley, thank you for your understanding when I was distracted or not fully present for you during this project. I look forward to the many great years of adventures we will have together. I hope you will always remember we can do hard things!

Chapter 1

Introduction

Colorectal cancer (CRC) is the third leading cause of cancer-related death in the United States; 52,550 new deaths attributed to CRC are projected for 2023 [46]. CRC is believed to advance from adenomatous polyps and this slow process creates a unique opportunity for both early detection and cancer prevention [4, 23]. Like other diseases, screening can reduce mortality by detecting cancer at earlier stages when it is more treatable; however, CRC screening can also reduce incidence by detecting and removing precancerous lesions [4]. As a result, the US Preventative Services Task Force (USPSTF) recommends screening for colorectal cancer in all adults aged 50 to 75 years and has recently expanded that recommendation to include those adults aged 45 to 49 years [12]. This recommendation accommodates a variety of tests, including visual examinations (colonoscopy, computed tomography (CT) colonography, and flexible sigmoidoscopy) and high-sensitivity stool-based tests (Guiac-based fecal occult blood tests (gFOBT), fecal immunochemical tests (FIT), and stool DNA test)[4, 12]. Performance among different tests is comparable when completed at the appropriate time interval and with the recommended follow-up. The inclusion of options for screening is intended to boost adherence [4]. Despite the endorsement of alternate screening modalities, colonoscopy capacity is often cited as a barrier in colorectal cancer (CRC) screening [28, 39, 44].

In this dissertation, we address capacity as a statistical problem rather than a resource one. We apply methods designed to optimize the use of longitudinal biomarker measurements for the early detection of CRC. We also derive and evaluate a decision threshold for clinical implementation.

1.1. Longitudinal Biomarkers

A biomarker is a biological variable, genetic or phenotypic, used to signal the normal or abnormal process of a condition or disease. It is objectively measured through a biochemical, molecular, or imaging technique. It is often used to detect current disease or progression, predict the onset of disease, or evaluate the course of therapeutic intervention. In the absence of disease, biomarker levels are expected to remain stable over time. In the case of cancer, however, biomarker levels in an otherwise asymptomatic patient would increase on an exponential scale, reflecting the change in tumor volume [47]. As a result, biomarkers can be an inexpensive noninvasive approach to screening. Two examples are Cancer Antigen 125 (CA-125) used to screen for ovarian cancer and high-sensitivity stool-based tests for colorectal cancer.

1.1.1. Ovarian Cancer Screening

Ovarian cancer is the fifth leading cause of cancer-related death for women in the US [19, 55]. This can be attributed to nearly 80% of ovarian cancers being diagnosed with advanced staging (III/VI) at diagnosis [55]. Once the disease has spread within the pelvis and abdomen the 5-year survival rate is only 29.2% [55]. Early disease detection could help to prevent death from ovarian cancer, as most patients (70% - 90%) diagnosed with stage I or II ovarian cancer can be treated with conventional surgery and chemotherapy, increasing the 5-year survival rate to 93% [3, 55].

In the United States, the prevalence of ovarian cancer is 1 in 2500 for postmenopausal women age 50 and older [55]. As a result, the ideal screening test would require a sensitivity $\geq$ 75% and a specificity of at least 99.6% to achieve a PPV of 10% [9, 55]. CA-125 is a glycoprotein found on the cell surface of most ovarian cancers and is shed into the blood, where it can be detected using immunoassays [47, 55]. This continuous measurement is dichotomized using a population-based cutoff ($\geq$ 35 U/mL) to indicate a positive or abnormal test result [19, 47]. While 80% of ovarian cancers express CA-125, a single measurement lacks the sensitivity and specificity required for early detection [55]. Serial monitoring of

CA-125, however, has been shown to improve specificity, as CA-125 values are expected to rise exponentially in ovarian cancer patients but generally do not rise in those without disease [9, 47, 55]. This is illustrated in figure 1.1 in the CA-125 trajectories of 50 cases, women diagnosed with ovarian cancer, and 50 controls, women without ovarian cancer. Algorithms used to differentiate patients with ovarian cancer from those without the disease have achieved a sensitivity of 86%, a specificity of 99.7%, and a PPV of 16%, increasing the fraction of patients detected with ovarian cancer in early stage to 41.4% [47, 55].

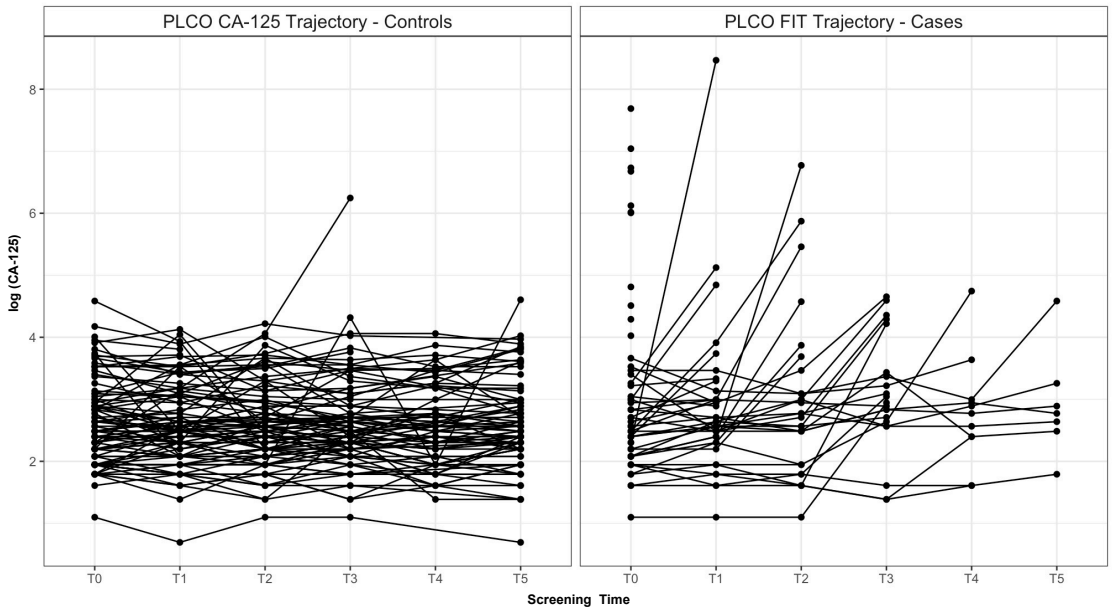


Figure 1.1: CA-125 trajectories of 100 cases and 100 controls that were randomly selected from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial data [11]

1.1.2. FIT-Based Screening

Both cancerous tumors and large adenomas bleed intermittently into the intestine, using antibodies FIT can detect this blood in the stool [4, 12]. Pooled FIT sensitivity is approximately 80% for cancer and 20% –30% for advanced neoplasia detection in a single application [40]. FIT effectiveness is maximized through repeated screening in participants with a normal screening result and follow-up colonoscopy when the test is positive [40]. This approach then yields similar reductions in mortality when compared with screening colonoscopy [4, 12].

FIT results are available in two formats. Qualitative tests report results as either positive or negative, while quantitative tests also include the estimated hemoglobin concentration in the result [16]. Quantitative FIT allows the cutoff concentration to be modified a priori when defining an abnormal result; this choice influences the performance characteristics of the test, in addition to the resources needed to support screening [16, 24]. Increasing the cutoff value decreases sensitivity while increasing specificity [24]. As a result, considerable attention has been paid to defining the optimal cutoff value in response to colonoscopy capacity concerns [24].

1.2. Modeling Biomarker Trajectories for Early Detection

Regardless of disease site, screening participants are likely to accumulate multiple measurements collected at regular intervals when participating in routine screening. While screening recommendations for both CA-125 and FIT suggest serial monitoring could improve sensitivity and specificity, the standard of care uses a single (most recent) measurement evaluated using a population-based cutoff to indicate a positive or negative result. This is because regression techniques commonly used to predict death or morbid events require each subject to have the same number of measurements, measured at the same time points [54]. In screening these criteria often cannot be met as patients with abnormal results are removed from the screening cycle to pursue more rigorous follow-up testing; instead, the most recent measurement is used.

Combining multiple biomarkers has also been investigated to improve diagnostic accuracy relative to a single marker, but again these methods described in the literature do not account for longitudinal measurements [30]. What is needed is a risk analysis method capable of using repeated measures, as they are accumulated, to allow for a more efficient use of these data. Addressing this gap would allow disease diagnosis to account for the trajectory in combination with the correlation structures for measurements collected repeatedly over time. The following approaches have been proposed to address these limitations and applied in the context of ovarian cancer screening. Each model used repeated CA-125 measurements

collected annually through The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial [19].

1.2.1. Shared Random Effects Model (SREM)

SREM was developed for risk prediction [1, 19, 30]. SREM jointly models binary disease outcomes and the longitudinal biomarker trajectories, assuming they share the same set of random effects [19]. Let $D_i = 0$ indicate that the i^{th} subject, $i = 1, \ldots, n$ is healthy and $D_i = 1$ indicate the i^{th} subject is diseased. The state of health of the subject is determined by the value of $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, a univariate biomarker measured at times $t_{i1}, \ldots, t_{in_i}$. The linear mixed model for those with (cases) and without disease (controls) is

$$\mathbf{Y}_i = \mathbf{X}_i\theta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i,$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices for the fixed and random effects, respectively; θ is a vector of fixed effects, $\mathbf{b}_i \sim \text{MVN}\ (0, \sigma_b^2)$ a vector of random effects, and $\varepsilon_i \sim \text{MVN}\ (0, \sigma_\varepsilon^2)$ are the random measurement errors. Then D_i is linked to the longitudinal process as

$$P(D_i = 1|b_i) = g\{\mathbf{W}_i\eta + \phi h(b_i)\},$$

where $g(\cdot)$ is a link function, $\mathbf{W}_i'$ is a vector of covariates specific to the i^{th} subject, $h(\mathbf{b}_i)$ is a known function of random effects, and ϕ represents the strength of the longitudinal association between the biomarker value and the outcome. [1, 30, 19]. The form of $h(\cdot)$ is application-specific. For example, in the fetal growth example described in Chapter 4, where SREM was originally outlined, $h(\cdot)$ was based on a quadratic growth curve for both fixed and random effects [1]. A simpler choice would be a linear combination of random effects [19].

Calculation of the diagnosis probability $P(D_i = 1|\mathbf{Y}_i)$ is greatly simplified when one assumes that the random effects are normally distributed and that the link function is a probit function [1, 30, 19]. Then, likelihood function is

$$L = L_1 \times L_2 = \prod_{i=1}^{N} f(\mathbf{Y}_i) \prod_{i=1}^{N} P(D_i|\mathbf{Y}_i).$$

L_1 involves only fixed effects, and L_2 has the explicit expression

$$P(D_i = 1|\mathbf{Y}_i) = \Phi\left\{\frac{\mathbf{W}_i'\eta + \phi h(\hat{b}_i)}{\sqrt{1 + \phi^2 \text{var}(h(\hat{b}_i - b_i))}}\right\}$$

where $\hat{b}_i = \mathbb{E}(b_i|\mathbf{Y}_i)$. The probability is obtained when the parameters are replaced with their maximum likelihood estimates (MLEs)[1, 19, 30]. These estimates are obtained through a two-step process, where L_1 is maximized first and its estimates are then used in maximizing L_2 to obtain the remaining parameters [1, 19, 30].

1.2.2. Risk of Ovarian Cancer Algorithm (ROCA)

The Risk of Ovarian Cancer Algorithm (ROCA) was developed for the early detection of ovarian cancer [47]. Cases and controls are modeled separately with the expectation that CA-125 trajectories will remain flat in patients without the disease (controls) but increase rapidly in patients with the disease (cases) [47]. Controls are modeled using a constant mean model, such that

$$\mathbb{E}(Y_{ij}|D_i = 0) = \theta_i,$$

where Y_{ij} is the biomarker value for the ith person at time t_{ij} and D_i is the binary outcome, where $D_i = 0$ are controls [47, 19]. Cases, $D_i = 1$, are more complex because approximately 15% of ovarian cancer tumors do not produce additional CA-125 [47]. As a result, those without elevated CA-125 are modeled like controls, with mean

$$\mathbb{E}(Y_{ij}|D_i = 1, T_i) = \theta_i;$$

note T_i is the cancer diagnosis time and differs from t_{ij} which is the screening time [47, 19]. Those with elevated CA-125 are modeled using a piecewise linear model with a latent person-specific changepoint τ_i conditional on T_i [19, 47]. Skates et al (2001) assumed τ_i followed a known truncated normal distribution; however, when implemented by Han, et al (2020) the parameters for the change point distribution were estimated instead of being prespecified. The mean is

$$\mathbb{E}(Y_{ij}|D_i = 1, T_i) = \theta_i + \gamma_i(t_{ij} - \tau_i)^+,$$

where γ_i is the increase in slope after τ_i, θ_i is the subject-specific intercept, and $(x)^+ = x$ when $x > 0$ or 0 otherwise, [19, 47]. All parameters are estimated using a Bayesian framework [19, 47].

ROCA outperformed the standard of care approach in both a simulated setting and an independent trial [47]. Han, et al (2020) later compared it with two alternative risk prediction frameworks. In their implementation, the control model was modified to adjust for both screening time and baseline age, such that

$$Y_{ij} = \theta_0 + \theta_1 t_{ij} + \theta_2 \text{Age}_i + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij},$$

where random error $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ and as mentioned previously, the parameters for the changepoint distribution were estimated to improve performance [19].

Even with these adjustments, a loss in prediction accuracy is still to be expected when predicting cancer early detection of a new subject k. This is because ROCA can only obtain an approximation of $P(Y_k|D_k = 1)$ as it is modeling $Y_k|D_k = 1, T_k$ rather than $Y_k|D_k = 1$ [19]. ROCA then must marginalize over-diagnosis time T_k, which is often unknown for the new subject, to obtain this estimate [19]. Marginalization occurs by "borrowing" information about diagnosis time from known cases, but this reduces the prediction accuracy when the sample size for cases is small, which is to be expected given the incidence of ovarian cancer [19].

1.2.3. Pattern Mixture Model (PMM)

Like ROCA, the pattern mixture model (PMM) uses separate longitudinal formulations for those with (cases) and without disease (controls) [19]. They implement the same control model as with ROCA, but the case model uses a linear mixed model with natural cubic splines in place of the latent change-point structure to account for the nonlinear effects of screening time and baseline age

$$Y_{ij} = \theta_0 + \sum_{\ell=1}^{3} \theta_\ell B_\ell(\text{Age}_i, \lambda) + b_{0i} + \sum_{\ell=1}^{3} (\theta_{\ell+3} + b_{\ell i}) B_\ell(t_{ij}, \lambda) + \varepsilon_{ij},$$

where $B_\ell(x, \lambda)$ is the B-spline basis of order three for the natural cubic spline with knot λ defined as the minimum and maximum values for boundary knots and first and third quantiles for the internal knots. The fixed effects are θ_0, the intercept, $\theta_\ell,\ \ell = 1, 2, 3$ for the cubic splines for baseline age, and $\theta_{\ell+3},\ \ell = 1, 2, 3$ for the cubic splines for screening time. The random effects are b_{0i} and $b_{\ell i}$ and ε_{ij} is the random measurement error [19]. PMM also differs from ROCA in that it directly models $Y_i|D_i = 1$, allowing the diagnosis probability to be calculated without marginalization, avoiding the loss of prediction accuracy [19]. The probability $P(D_i = 1|Y_i)$ is obtained using Bayes' rule

$$\frac{P(D_i = 1|Y_i)}{P(D_i = 0|Y_i)} = \frac{P(Y_i|D_i = 1)}{P(Y_i|D_i = 0)} \times \frac{P(D_i = 1)}{P(D_i = 0)},$$

where $Y_i|D_i$ follows a multivariate normal distribution with normal random effects and error terms [19, 30]. If the prior disease information, $P(D_i = d)$, is known or can accurately be estimated, then the likelihood ratio under PMM, $P(D_i = 1|Y_i)/P(D_i = 0|Y_i)$, is then the optimal combination of the longitudinal biomarkers [19, 30].

While PMM was originally designed for risk prediction rather than early detection Han et al. (2020) demonstrated how a risk prediction framework could successfully be applied for the early detection of ovarian cancer. In the model comparison, PMM obtained a significantly better predictive performance relative to the other two approaches [19].

1.3. Model Comparison

ROCA, PMM, and SREM were evaluated using discrimination and calibration performance [19]. Diagnostic predictive accuracy was derived from time-dependent ROC curves and AUCs generated for 0.5 to 3 years post-screening. Calibration performance was defined by the calibration intercept and slope, as well as the ratio of observed (O) to expected (E) number of cases, where metrics closer to nominal values were considered better [19]. PMM had the highest time-dependent AUCs, ranging from 1.8 to 3.4% compared to ROCA and 1.6 to 4.8%, compared to SREM [19]. Using the bootstrapped replicates of the AUCs to compare ROCA and SREM there was no discernible difference in discrimination [19]. Discrimination performance for SREM is likely attributed to the simultaneous modeling of cases

and controls, which may not be reasonable given the demonstrated difference in the trajectories [19]. All three approaches were well calibrated, yet PMM and ROCA were better able to classify individuals as high and low risk, suggesting they could be more useful in clinical practice [19].

No one model is guaranteed to be uniformly better in practical application. However given the superior performance of PMM in the context of ovarian cancer screening combined with the similarities in FIT-based screening for CRC and ovarian cancer screening with CA-125, we will move forward using PMM to analyze FIT-based screening data.

1.4. Organization of the Chapters

In Chapter 2 we describe a large FIT-based CRC screening program and discuss how the distribution for quantitative FIT measurements may influence the implementation of the PMM approach. Chapter 3 begins with reproducing results from the PMM model in the context of ovarian cancer screening. We then consider several ad hoc substitution methods and complete-case analysis to address the need for data transformation in the presence of right-skewed data heavily concentrated with zeros. We find PMM to be sensitive to the substitution method. In chapter 4 we present an alternative approach, where we consider quantitative FIT to be left-censored. Here we provide examples of imputation using the non-parametric inverse Buckley-James approach, in both a "complete" dataset and the CRC screening data. Chapter 5 introduces functional data analysis and use of the underlying FIT trajectories to predict risk for CRC, rather than the repeated measurements themselves. Chapter 6 utilizes bootstrap sampling to identify and evaluate a decision threshold for the PMM risk score. We summarize the conclusions from this thesis in Chapter 7 and discuss issues that require further research and consideration.

Chapter 2

Longitudinal Biomarkers in FIT-Based Screening

Like ovarian cancer screening, FIT-based screening can produce a quantitative measurement. This result is then evaluated using a pre-specified, population-based cutoff to determine the screening outcome. Similarly, participants with a normal result are encouraged to repeat screening annually in both settings. Despite the accumulation of data, the test result is still determined by a single, most recent, measurement.

A single application of FIT-based screening has better test performance when compared to CA-125 using sensitivity and specificity [9, 40, 55]. However, the literature indicates serial monitoring of CA-125 levels can improve test performance, particularly in terms of specificity [9, 55]. This has been investigated further in a comparison of multiple models to incorporate longitudinal CA-125 measurements for the early detection of ovarian cancer, where the PMM was shown to have superior performance [19]. Given the similarities in ovarian cancer screening and FIT-based screening for CRC, we anticipate comparable improvements when incorporating longitudinal biomarkers to screen for the early detection of CRC. Here we introduce the FIT-based screening data, highlighting differences from the CA-125 data to provide context for the approaches considered with implementation in Chapter 3.

2.1. Colorectal Cancer Screening and Patient Navigation

The Colorectal Cancer Screening and Patient Navigation (CSPAN) program was established in November 2013. It began as a randomized comparative effectiveness trial to evaluate the impact of financial incentives on participation when paired with mailed outreach to offer FIT-based CRC screening through John Peter Smith Health Network (JPS). JPS includes a large public hospital and a network of more than 60 community clinics to provide primary and tertiary care services in Tarrant County, Texas, including Fort Worth. JPS offers a med-

ical assistance program, JPS Connection, for uninsured individuals with insufficient financial resources who need medical care. Qualifying individuals must reside in Tarrant County, be a US citizen or a legal permanent resident, and meet the required income guidelines of less than 250% of the federal poverty level (FPL). FIT invitations were mailed to all uninsured individuals enrolled in JPS Connection, who were aged 50–64 years, not up-to-date with CRC screening at baseline. Screening was defined as guaiac fecal occult blood test or FIT in the past year, sigmoidoscopy in the past 5 years, or colonoscopy in the last 10 years. All invitees had one or more visits to a primary care clinic within the year before program initiation. Individuals with a prior history of CRC or colonic resection, missing address or phone number, or who were incarcerated at baseline were excluded. Individuals in the resulting cohort were then randomly assigned to receive the following: (1) mailed outreach; (2) mailed outreach plus a $5 incentive for FIT completion; or (3) mailed outreach plus a $10 incentive for FIT completion.

Mailed FIT outreach consisted of: (1) an invitation in English and Spanish to return a FIT, along with pictorial instructions on how to complete; (2) an enclosed 1-sample FIT test (OC-Micro/Sensor, Polymedco); (3) two automated telephone reminders in English and Spanish to encourage test completion, delivered at the time of invitation and 1 week later; (4) up to two "live" telephone reminders attempted within 4 weeks post-invitation if screening was not completed and the patient was not reached on the initial call attempt. Returned FIT kits were processed by the health system clinical laboratory per manufacturer instructions. When defining an abnormal result, the cutoff was chosen a priori as ≥ 10 μg/g to evaluate the impact of threshold on test performance and colonoscopy demand. All patients with an abnormal test result were referred for diagnostic colonoscopy. Colonoscopy completion was encouraged through: (1) telephone-based navigation for appointment scheduling; (2) provision of bowel prep (via mail or clinic pickup); and (3) appointment reminders and review of preparation instructions 5 and 2 days before the colonoscopy appointment. Certified letters with the test results were also sent to patients with an abnormal FIT result, as well as their primary care provider, and included the recommendation to schedule a colonoscopy.

Patients with normal test results were notified by mail and reminded to repeat screening in one year. Gift cards were included with the result letter for both incentive groups. All clinical services, including FIT tests, pre- and post-operative visits, and diagnostic colonoscopy after an abnormal FIT were provided at no cost to participants, using funding from the Cancer Prevention Research Institute of Texas (CPRIT; PP120229).

In year one, invitations were distributed over five mail-out "rounds." In years 2 and 3, patients were re-invited with the same intervention assignment (outreach only, $5, or $10) only if they completed a screening test in the year prior, with a normal result. Patients who did not complete a FIT were not re-invited. New patients invited in years 2 and 3, including those newly enrolled in JPS connections or those newly age-eligible for screening, were not randomized but rather assigned to the outreach-only group. In year 4, financial incentives were discontinued. All patients with a normal result in year 3 were invited to complete FIT using outreach only, regardless of their original assignment. Before data collection, this trial was preregistered on ClinicalTrials.gov (identifier: NCT01946282). The Colorectal Cancer Screening and Navigation (C-SPAN) is funded through the Cancer Prevention & Research Institute of Texas (CPRIT PP120229; PI Argenbright). This program was reviewed by the University of Texas Southwestern Medical Center Institutional Review Board and determined as a non-research activity.

In February 2016, the program transitioned from a closed system, JPS, to an open system. Specifically, rural and medically underserved Texans from 35 north Texas counties could then be referred to the program by local healthcare networks, community clinics, safety-net systems, Federally Qualified Health Centers (FQHCs), and primary care clinics. Individuals could also self-enroll in screening either during community outreach events and health fairs or through a dedicated 800 number, provided they met the inclusion criteria established previously. FIT kits continued to be mailed to all FIT-negative patients from the original program, along with the self-enrolled and those referred by a new partnering provider. The outreach methods outlined previously were again used to encourage screening completion. Completed FIT kits were processed at UT Southwestern Medical Center under standing

orders from the Medical Director at Moncrief Cancer Institute (UTSW-MCI). The program reverted to the manufacturer-specified cutoff of $\geq$ 20 μg/g to define an abnormal positive FIT. This was in response to the results from the initial program implementation, where it was found that reducing the abnormal FIT result cut-off value ($\geq 10\mu g$/g) might increase advanced neoplasia detection, but doubled the proportion of patients requiring a diagnostic colonoscopy [6]. Patients with a normal screening result were reminded to repeat screening in one year and then re-invited for continued participation in the program, while patients with an abnormal screening result were navigated to diagnostic colonoscopy. Diagnostic services were scheduled with a partnering provider within the patient's county of residence, when possible, or in a neighboring county. To reduce the impact of transportation on colonoscopy completion, travel time to care was no more than one hour, and transportation assistance was also available. All clinical services were provided at no cost to participants through continued support from CPRIT (PP150061).

The current phase of the program began in March 2020, expanding coverage to serve an additional 22 counties. Again, FIT-negative participants from the previous two phases were invited to continue annual screening, along with new referrals from community partners and those who self-enrolled, and all clinical services are provided at no cost to participants through continued support from CPRIT (PP200009).

The resulting CSPAN dataset contains 19,796 patients who completed at least one FIT between November 1, 2013, and December 31, 2021, as part of either one or more phases of the CSPAN program. Participants with a baseline screening age younger than 50 (n = 199) or older than 74 years of age (n = 62) were excluded to align with screening guidelines for the majority of the study period. Note, the USPSTF only modified screening eligibility in May of 2021 to include those aged 45 - 49 [12]. The sample available for analysis then includes 29,398 completed FIT entries for 16,637 participants. Among them, we define FIT-negative patients (n = 15,336) as those who had only negative qualitative results, regardless of the number of rounds completed, and FIT-positive patients (n = 1,301) as those who had a positive qualitative result, which may be preceded by one or more negative results

depending on the number of rounds completed. This translates to a FIT positive rate of 7.82%, which is consistent with what has been reported in the literature in recent studies, where 7% to 8% of FIT were positive [2]. Demographically the positive and negative FIT groups are similar, as shown in the descriptive statistics included in Table:2.1.

Characteristic	Group	CSPAN Participants (N = 16,637)	
		FIT Negative(n = 15,336)	FIT Positive (n = 1,301)
Baseline Age	50 - 54	5,019 (32.7%)	399 (30.7%)
	55 - 59	4,400 (28.6%)	427 (32.8%)
	60 - 64	3,817 (24.8%)	350 (26.9%)
	65 - 69	1,414 (9.2%)	93 (7.1%)
	70 - 74	716 (4.7%)	32 (2.5%)
Sex	Male	4,878 (31.7%)	448 (34.4%)
	Female	10,483 (68.2%)	853 (65.6%)
Race	White	7,522 (49.0%)	656 (50.4%)
	Black	2,718 (17.7%)	278 (21.4%)
	Asian	259 (1.7%)	16 (1.2%)
	AI/AN	55 (0.4%)	3 (0.2%)
	NH/PI	17 (0.1%)	3 (0.2%)
	Other	2,742 (17.8%)	232 (17.8%)
	Unknown	2,053 (13.4%)	113 (8.7%)
Ethnicity	Hispanic	6,052 (39.4%)	390 (30.0%)
	Not Hispanic	8,186 (53.3%)	834 (64.1%)
	Unknown	1,128 (7.3%)	77 (5.9%)
Number of Observations	1	8,934 (58.1%)	866 (66.6%)
	2	3,136 (20.4%)	258 (19.8%)
	3	1,923 (12.5%)	113 (8.7%)
	4	828 (5.4%)	43 (3.3%)
	5	263 (1.7%)	13 (1.0%)
	6	195 (1.3%)	5 (0.4%)
	7	79 (0.5%)	3 (0.2%)
	8	8 (0.1%)	0 (0.0%)

Table 2.1: Descriptive statistics for CSPAN participants November 2013 — December 2021. AI/AN is American Indian/Alaska Native and NH/PI is Native Hawaiian/Pacific Islander.

The FIT-positive cohort can further be subdivided by colonoscopy outcome to include CRC, advanced adenoma, adenoma, and normal results. Advanced adenomas are $\geq$ 1 cm,

with villous or tubulovillous features, and/or high-grade dysplasia, while adenomas are < 1 cm, without villous or tubulovillous features, and/or without high-grade dysplasia [38]. Of these subgroups, cases are a combination of CRC and advanced adenoma patients, also known as advanced neoplasia (AN) patients (n = 225). The quantitative FIT score trajectories for 100 randomly chosen cases (right panel) are shown along 100 randomly selected controls (left panel) from the FIT negative cohort in Figure 2.1. Each black line in the figure represents FIT scores for one subject.

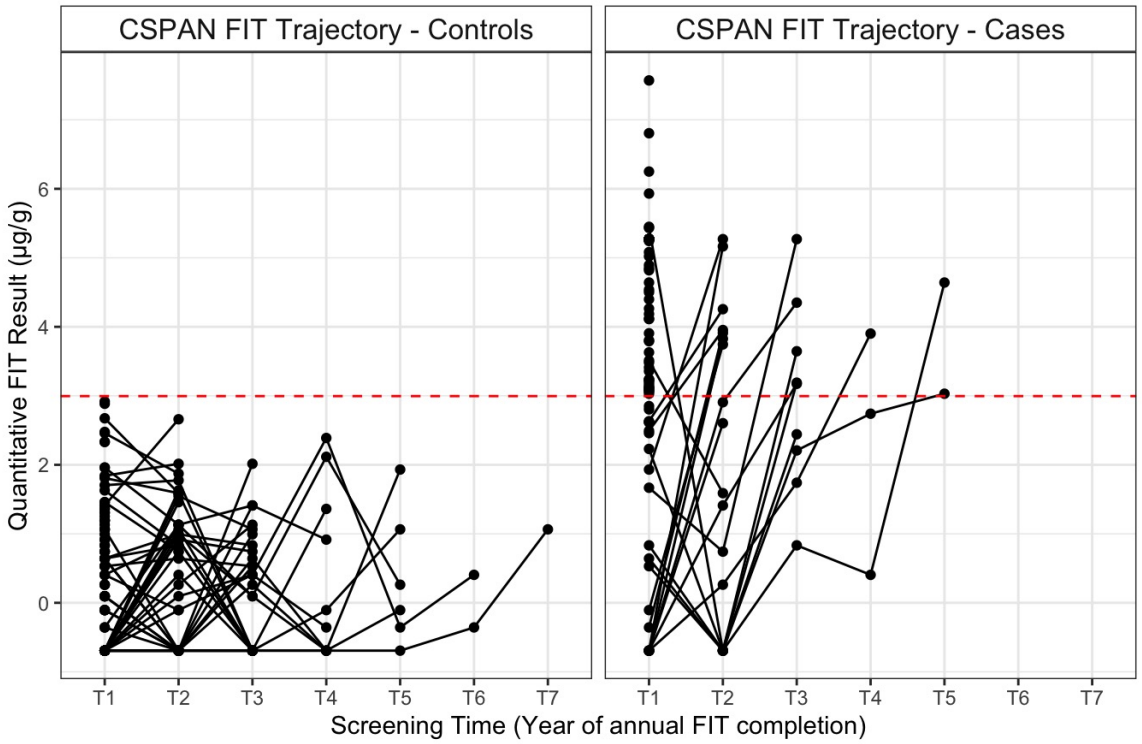


Figure 2.1: Quantitative FIT Trajectories for 100 Cases (right panel) and 100 Controls (left panel) randomly selected from the CSPAN program. Each black line shows the FIT levels for one subject in the program. The horizontal red line indicates the 20 μg of hemoglobin per gram of feces cutoff used to define an abnormal FIT

The FIT trajectories remain nearly flat for the controls over multiple rounds of screening, while the case trajectories increase rapidly at some point despite having begun as flat in some patients. The patterns shown in Figure 2.1 are similar to those published for CA-125 from the

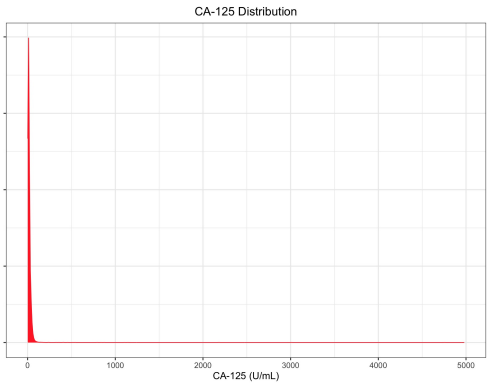
PLCO dataset and shown below in Figure 1.1. Given these similarities, the PMM approach used with the PLCO data for ovarian cancer screening should also perform well with the quantitative FIT data for CRC.

2.2. Limitations for PMM Implementation with Quantitative FIT

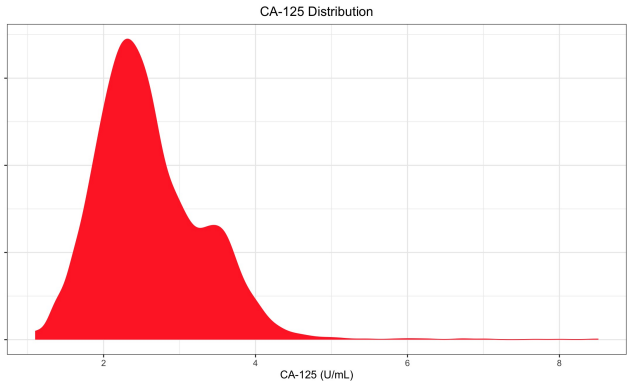
PMM directly makes the assumption of a linear mixed model on the biomarker trajectories conditional on disease status, where $Y_i|D_i = d$, $d = 0, 1$ [19]. Recall $Y_i|D_i$ follows a multivariate normal distribution

$$Y_i|D_i \sim \text{MVN}(\mu_d, \Sigma_d),$$

with normal random effects and error terms [30]. Often real data are extremely skewed leading to invalid results in a standard statistical analysis; the log transformation is a widely used method to address skewness [15]. In this context, CA-125 spans orders of magnitude, ranging from 2 - 10,000 U/ml as shown in Figure 2.2a. Taking the logarithm transforms these levels to an arithmetic scale better suited for modeling with linear additive components [19, 47]. The use of the log transformation is further grounded in the biological understanding of the tumor process, where marker levels are thought to be proportional to the tumor volume; this approach results in a linear increase in biomarker levels during the preclinical phase of the disease [47]. As a result, the authors apply log transformation to obtain a distribution much closer to normal than the original right-skewed biomarker data with a long right tail (Figure 2.2b)[15, 19, 47].



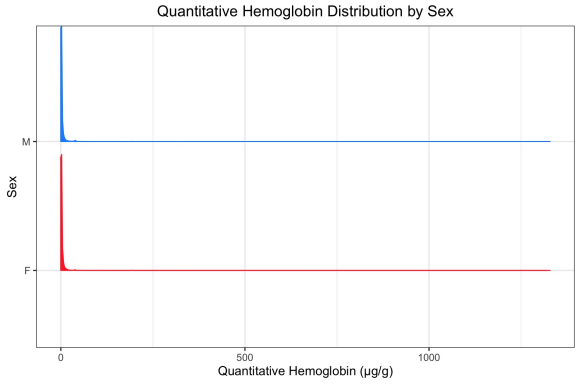
(a) Distribution of CA-125 measurements — original scale.

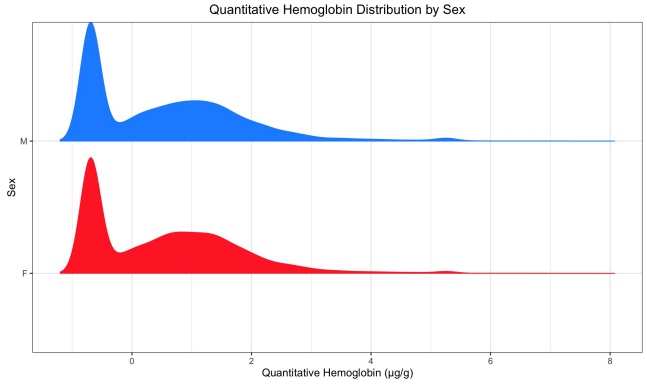


(b) Distribution of CA-125 measurements — Log scale.

Figure 2.2: Distribution of CA-125 measurements for participants in the PLCO trial (n = 980). [11]

The quantitative FIT values for the CSPAN cohort range from 0 - 388 μg/g. Like CA-125, fecal hemoglobin concentrations are positive with a long right tail (Figure 2.3a). The minimum value however is now zero which can cause difficulty with transformations (Figure 2.3b).





(a) Distribution of quantitative FIT — original scale. (b) Distribution of quantitative FIT — Log scale.

Figure 2.3: Distribution of quantitative FIT results by sex for participants CSPAN program (n = 16,667).

Specifically, the log is undefined at zero, such that $\log(x) \to -\infty$ as $x \to 0$ [15, 53]. When only a few zero values are present for a variable ($< 2\%$), a shift parameter (M) can be added to all values without issue for subsequent analyses, [15, 53]. M is typically a small positive constant, like 0.01, to minimize the difference from the true value while allowing for a valid

transformation [53]. It is worth noting, however, that guidelines for selecting this value are limited in the literature. It is instead left to the analyst to decide, making any conclusions subjective to investigator bias [15].

As an example, if we say $g(M) = \exp(\mathbb{E}\log(X+M)) - M$, then

$$g''(M) = \exp[\mathbb{E}\log(x+M)]\left\{\left(\mathbb{E}\frac{1}{X+M}\right)^2 - \mathbb{E}\frac{1}{(X+M)^2}\right\} < 0,$$

unless X is constant [15]. When $M > 0$,

$$\exp(\mathbb{E}\log(X+M) \neq M + \exp(\mathbb{E}\log(X))$$

indicating $g(M)$ is dependant upon M [15]. Including a nonzero M adds another layer of complexity when interpreting model estimates from the translated data, dependent on how M was chosen [15].

We divided our CSPAN population into quintiles for the male and female subgroups in Table 2.2 to compare with data from the National Scottish Bowel Screening Programme in Table 2.2 [31]. The distributions are largely similar; any differences can likely be attributed to differences in breakdown for sex, as men typically have higher hemoglobin concentrations than women [31].

Table 2.2: Percentiles (with 95% CI) of fecal hemoglobin concentration by program

CSPAN						
Sex	n (%)	25.0%	50.0%	75.0%	90.0%	95.0%
Female	10,483 (68.2%)	0 (0, 0)	1 (1, 1.2)	3.2 (3.2, 3.4)	8.0 (7.6, 8.4)	16.0 (14.8, 17.6)
Male	4,878 (31.7%)	0 (0, 0)	1.2 (1, 1.2)	3.6 (3.4, 3.8)	9.4 (8.6, 10.2)	19.6 (17.2, 23.6)
Scottish Bowel Programme						
Sex	n (%)	25.0%	50.0%	75.0%	90.0%	95.0%
Female	20,662 (53.4%)	0 (0, 0)	0 (0, 0)	1.8 (1.6, 1.8)	7.6 (7, 8.2)	22.8 (21.2, 24.2)
Male	18,058 (46.6%)	0 (0, 0)	0.2 (0.2, 0.2)	2.4 (2.4, 2.6)	13.4 (12.2, 14.6)	36.8 (35, 42.6)

McDonald, et al. (2012) used the D'Agostino-Pearson test to confirm the quantitative FIT distribution is not Gaussian ($p < 0.0001$). The D'Agostino-Pearson test is the sum of

these skewness and kurtosis tests, such that when data are normally distributed the test statistic $z_k^2 + z_s^2$ has a chi-square distribution with 2 degrees of freedom, i.e.

$$z_k^2 + z_s^2 \sim \chi_2^2,$$

where z_s is the test statistic for skewness and z_k is the test statistic for kurtosis. The null hypothesis is that the data is a realization of independent, identically distributed Gaussian random variables, a significant p-value would then indicate the FIT data is not, specifically the coefficients of skewness and kurtosis were > 1 ($p < 0.0001$) [31]. While it is widely accepted that everyone has some blood in their feces, normal fecal blood loss can approach 1.5 mL of blood per day. FIT can detect as little as 0.3 mL of blood added to the stool; therefore, Table 2.2 suggests as much as half of the population may have no detectable hemoglobin in their feces or a quantitative FIT score of zero [31]. Within the CSPAN dataset, 9,327 of the 111,176 quantitative FIT scores, or 29.8%, measure 0 μg/g, which well exceeds the $< 2\%$ recommended in the literature [53].

Incorporating longitudinal biomarkers has been shown to improve performance when screening for the early detection of ovarian cancer. Given the similarities to FIT-based screening, we anticipate incorporating longitudinal measurements could have a similar effect on the early detection of CRC. To do so, we will apply PMM as it was shown to have superior performance in a model comparison but this will require log-transforming the data [19]. While this method could readily be applied to CA-125 without a shift parameter because they are positive, nonzero measurements, we have shown that not to be the case for quantitive hemoglobin measurements. In the next chapter, we consider multiple choices for the shift parameter, M, to facilitate the necessary data transformation in our implementation of PMM.

Chapter 3

Analyzing Longitudinal Biomarkers with PMM

In this chapter, we implement the PMM approach for early detection. We first recreate results from Han, et al. (2020) using the PLCO Ovarian Phase III Validation Study dataset from LABCAS Public Collections to validate the process. Then we analyze the CSPAN dataset using PMM, applying various approaches to accommodate the necessary data transformation. These same methods are used within the standard-of-care approach, using univariate logistic regression to predict risk based on the most recent FIT. Performance is evaluated based on diagnostic predictive accuracy as defined by the area under the ROC curve (AUC).

3.1. PMM Example: PLCO ovarian cancer data

Our analytic sample is the publicly available subset (N =976) of the PLCO Cancer Screening Trial used in the original paper (N = 30,402). The sample includes 71 ovarian cancer cases and 905 controls, compared to 132 cases and 30,269 controls in the original data set [19]. CA-125 measurements ranged from 2 - 4,755 U/mL before a log transformation. The median number of measurements in cases was three, and six in controls.

Recall the PMM approach models cases and controls separately. We adjust for both screening time and baseline age in the control model. For consistency with Han, et al. (2020), we label it Model 3 (CN3):

$$Y_{ij} = \theta_0 + \theta_1 t_{ij} + \theta_2 \text{Age}_i + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij},$$

where t_{ij} is screening time, Age_i is baseline age, θ_0, θ_1 and θ_2 are fixed effects, b_{0i} and b_{1i} are random effects, and ε_{ij} is the random measurement error. Similarly, we label the case model

as PMM:

$$Y_{ij} = \theta_0 + \sum_{\ell=1}^{3} \theta_\ell B_\ell(\text{Age}_i, \lambda) + b_{0i} + \sum_{\ell=1}^{3} (\theta_{\ell+3} + b_{\ell i} B_\ell(t_{ij}, \lambda) + \varepsilon_{ij},$$

where $B_\ell(x, \lambda)$ is the B-spline basis of order 3 for the natural cubic spline with knot λ decided based on x, θ_0, θ_ℓ and $\theta_{\ell+3}$ are fixed effects, b_{0i} and $b_{\ell i}$ are random effects, and ε_{ij} is the random measurement error.

As described in chapter 1, the PMM model accounts for the nonlinear effects of screening time and baseline age using natural cubic splines. To build the corresponding design matrices, we first calculate screening time as the difference, in years, from the date the patient entered the study to the date of serum collection for each measurement. Baseline age is defined as age at study entry [11]. Using the splines package, we generated the B-spline basis of order 3 for the natural cubic splines. The knots (λ) were determined in the same way and using all cases for both screening time and baseline age. Specifically, the boundary knots were set as the maximum and minimum values for each variable, while the internal knots were set as the first and third quantiles. Stratifying data by disease status, we then pass screening time and baseline age to X_0, the design matrix for fixed effects in the control model, and screening time to Z_0, the design matrix for random effects in the control model. Similarly, X_1 is comprised of the spline values for baseline age and the spline values for screening time, while Z_1 contains the spline values for screening time. Lastly, the CA-125 measurements were log-transformed.

We estimated the case (PMM) and control (CN3) models using the lme4 package. The vector of fixed effects, β^d, along with the variance components from the random effects ($b_i^d \sim \text{MVN}(0, \Delta^d)$) and random errors ($\varepsilon_i^d \sim \text{MVN}(0, \Sigma_i^d)$) were extracted. Estimating individual disease risk score first requires us to calculate $\mu_d \equiv X_i \beta_d$ and $\Gamma_d \equiv Z_i \Delta^d Z_i' + \Sigma_i^d$ for cases ($d = 1$) and controls ($d = 0$). Suppose we regard disease classification as a hypothesis-testing problem. In that case, we can consider ϕ to be the combination rule mapping the J-dimensional longitudinal marker Y to a univariate marker $\phi(Y)$, where a larger value of $\phi(Y)$ is more characteristic of disease [30]. Based on the Neyman-Pearson Lemma, the

likelihood ratio

$$\Lambda(Y) = \frac{P(Y|D=1)}{P(Y|D=0)}$$

can be considered the most powerful test. As stated in chapter 1, this gives us the optimal combination of the longitudinal biomarkers [30]. We obtain $\Lambda(Y)$ under the PMM setting as

$$\log \Lambda(Y) \propto (y_i - \mu_0)' \, \Gamma_0^{-1} \, (y_i - \mu_0) - (y_i - \mu_1)' \, \Gamma_1^{-1} \, (y_i - \mu_1),$$

where $\log \Lambda(Y) \equiv \lambda(Y)$ [30]. The disease risk score , $P(D=1|Y)$ can then be written as a function of $\lambda(Y)$

$$\log \frac{(P(D=1|Y)}{(P(D=0|Y)} = \log \frac{(P(D=1,Y)}{(P(D=0,Y)} = \log \frac{f_{Y|D=1}(Y)P(D=1)}{f_{Y|D=0}(Y)P(D=0)} = \log \frac{(P(D=1)}{(P(D=0)} + \frac{1}{2} \log \frac{|\Gamma_0|}{|\Gamma_0|} + \frac{1}{2}\lambda(Y)$$

using Bayes' theorem [30]. As a result, the risk score is merely a byproduct of computing the likelihood ratio combination.

Concurrently, we are also estimating variance (Ω) for a disease risk score (ρ_0) using

$$\Omega = \frac{\partial \rho_0}{\partial \gamma^1} \Pi_1 \left(\frac{\partial \rho_0}{\partial \gamma^1} \right)' + \frac{\partial \rho_0}{\partial \gamma^0} \Pi_0 \left(\frac{\partial \rho_0}{\partial \gamma^0} \right)' + \frac{1}{p_D(1-p_D)}, \tag{3.1}$$

from which we can calculate a corresponding confidence interval. To do so, we first obtain Π_d, the information matrix of the stratified likelihood function $L(Y_i|D_i = d; \gamma^d)$, where $\gamma^d = (\beta^d, \eta^d)$, such that β^d is the vector of fixed effects and η^d is the vector of parameters in the variance components for the PMM case and control models [30]. Second, we compute the derivatives $\frac{\partial \rho_0}{\partial \gamma^d} = \left(\frac{\partial \rho_0}{\partial \beta^d}, \frac{\partial \rho_0}{\partial \eta^d} \right)$ for cases ($d = 1$) and controls ($d = 0$) [30]. Lastly, we evaluate the variance contribution from estimating disease prevalence. If $p_D \equiv P(D = 1)$, we can approximate p_D as $\frac{1}{N} \sum_{i=1}^{N} D_i$ [30]. Taking the derivative yields $\frac{1}{p_D(1-p_D)}$. The contributions of p_D, γ^1, and γ^0 to the variance, Ω, are additive given their mutual independence, resulting in Equation 3.1 [30].

Out-of-sample risk predictions were obtained for each participant (N = 976) using leave-one-out-cross-validation (LOOCV) to minimize over-fitting [19]. Specifically, the CA-125 measurements were removed for a given participant, the models were estimated using the remaining participant data, and then the risk score was calculated for the excluded participant. Time-dependent ROC curves and AUCs generated with survivalROC for 0.5 to 3 years

post CA-125 screening were used to compare prediction accuracy; the corresponding 95% confidence intervals were calculated for each time-dependent AUC using 2,000 bootstrapping replicates. Analysis was conducted in R version 4.3.1 [35].

3.1.1. Results

Table 3.1 includes the fitted parameters we obtained for the control model (CN3), along with those published by Han et al. (2020). Our results align in both direction and magnitude with those reported originally. While Han, et al. (2020) reported baseline age and screening time had small but significant effects on the CA-125 trajectories (0.019 (0.018, 0.020) and 0.002 (0.001, 0.003), respectively). However, we found evidence for an effect of screening time ($p = 0.002$) but not baseline age ($p = 0.224$). Similarly, table 3.2 includes the fitted results for PMM from both our implementation and those published originally. Our estimates largely fall within the confidence intervals reported by Han et al. (2020), except for those effects (fixed and random) related to screening time, where we see a greater variance in magnitude.

	Reproduced CN3	Original CN3
Parameter	Estimate (95% CI)	Estimate (95% CI)
intercept: θ_0	2.261 (1.734, 2.788)	2.158 (2.097, 2.219)
screening time: θ_1	0.009 (0.003, 0.016)	0.019 (0.018, 0.20)
baseline age: θ_2	0.005 (-0.004, 0.013)	0.002 (0.001, 0.003)
SD of random intercept:σ_{b_0}	0.607 (0.575, 0.641)	0.451 (0.442, 0.460)
SD of random slope:σ_{b_1}	0.058 (0.050, 0.066)	0.039 (0.008, 0.069)
random effect correlation:$\rho_{b_0 b_1}$	-0.108 (-0.220, -0.012)	-0.147 (-0.191, -0.103)
SD of random error: σ_ξ	0.276 (0.268, 0.283)	0.215(0.211, 0.220)

Table 3.1: Parameter estimates and 95% confidence intervals (CI) for CN3 reproduced from the 905 controls in the PLCO trial subset and as published originally using 30,269 controls. SD denotes standard deviation.

	Reproduced PMM	Original PMM
Parameter	Estimate (95% CI)	Estimate (95% CI)
Fixed Effects		
intercept: θ_0	2.313 (1.096, 3.536)	2.604 (2.079, 3.129)
AgeSpline1: θ_1	1.440 (-0.017, 2.820)	0.251 (-0.629, 1.130)
AgeSpline2: θ_2	1.066 (-1.903, 4.036)	0.466 (-0.930, 1.862)
AgeSpline3: θ_3	-0.062 (-0.500, 0.383)	0.480 (-0.234, 1.192)
ScrTimeSpline1:θ_4	7.219 (2.811, 11.574)	1.649 (0.945, 2.353)
ScrTimeSpline2:θ_5	9.540 (3.915, 15.126)	0.701 (0.135, 1.267)
ScrTimeSpline3:θ_6	1.393 (1.097, 1.620)	0.834 (0.343, 1.326)
Random Effects		
intercept: σ_{b_0}	1.393 (1.097, 1.620)	1.223 (1.034, 1.446)
ScrTimeSpline1:σ_{b_1}	0.604 (0.315, 1.124)	3.194 (2.293, 4.450)
ScrTimeSpline2:σ_{b_2}	14.463 (11.087, 17.414)	2.317 (1.080, 4.971)
ScrTimeSpline3:σ_{b_3}	18.505 (14.122, 22.267)	1.5663 (0.824, 2.967)
correlation: $\rho_{b_0 b_1}$	-0.115 (-0.951, 0.597)	-0.306 (-0.704, 0.238)
correlation: $\rho_{b_0 b_2}$	-0.327 (-0.602, -0.014)	-0.675 (-0.859, -0.337)
correlation: $\rho_{b_0 b_3}$	-0.378 (-0.642, -0.071)	-0.431 (-0.884, 0.440)
correlation: $\rho_{b_1 b_2}$	0.450 (-0.116, 0.983)	0.735 (0.232, 0.928)
correlation: $\rho_{b_1 b_3}$	0.453 (-0.128, 0.983)	0.181 (-0.907, 0.954)
correlation: $\rho_{b_2 b_3}$	0.998 (0.993, 0.999)	0.506 (-0.737, 0.970)
random error: σ_ξ	0.321 (0.246, 0.353)	0.437 (0.356, 0.536)

Table 3.2: Parameter estimates for PMM reproduced from the 71 in the PLCO trial subset and as published originally using 132 cases: estimate and the 95% confidence interval (CI) reported.

We further evaluate our implementation using diagnostic predictive accuracy. As in the original paper, we report time-dependent AUCs generated at six cutoff times ranging from 0.5 to 3 years post-screening in Table 3.3. We can interpret these values as the probability of

a randomly selected "case" with a cancer diagnosis before time t having a larger predicted risk than a randomly selected "control" with a diagnosis time after time t [19]. Predictive accuracy is considered better at time t at greater values of AUCs [19]. Our results yield a similar pattern to the outcomes reported by Han et al. (2020), where AUCs are greatest early in the follow-up period but decrease as time t increases. While our values exceed those reported originally, our comparison using the bootstrapped replicates of the AUCs indicates no difference between the two results at each of the cutoff times, excluding $AUC_{1.0}$ These results suggest our reproduction of the model is satisfactory and our implementation approach sufficiently mirrors the one presented by Han et al. (2020). We will now apply the PMM approach to the CSPAN dataset.

	Reproduced	Original
Metric	PMM-CN3	PMM-CN3
$AUC_{0.5}$	0.971 (0.952, 0.986)	0.946 (0.937, 0.954)
$AUC_{1.0}$	0.947 (0.922, 0.968)	0.894 (0.886, 0.902)
$AUC_{1.5}$	0.889 (0.835, 0.938)	0.865 (0.858, 0.872)
$AUC_{2.0}$	0.867 (0.806, 0.921)	0.842 (0.832, 0.851)
$AUC_{2.5}$	0.867 (0.806, 0.918)	0.819 (0.810, 0.828)
$AUC_{3.0}$	0.859 (0.799, 0.913)	0.801 (0.791, 0.809)

Table 3.3: Time-dependent AUCs and associated 95% bootstrapped confidence intervals (shown in parentheses) for the original and reproduced PMM-CN3 method

3.2. PMM implementation and evaluation: CSPAN

In chapter 2 we introduced the need for a shift parameter to log-transform quantitative FIT from the CPSAN dataset. Specifically, the measurements are heavily right-skewed, ranging in value from 0 - 388 μg/g with a high concentration of zeros. We selected values commonly employed to address this scenario, in addition to a complete case analysis. These include:

- M1: Complete case (CC) analysis, removing participants with a quantitative FIT measurement equal to zero (N = 9,393);
- M2: Add a small positive number (0.01) to all values before taking the log transform to avoid taking the log of 0. [53];
- M3: Substitute each recorded 0 value with $\min(Y_{ij})$, where Y_{ij} is the quantitative FIT measured indexed by patient and time, indicating all samples;
- M4: substitution of zero by $\min(Y_{ij}) \pm X \sim \mathcal{N}(0, 0.00005)$, where Y_{ij} is the quantitative FIT measured indexed by patient and time, indicating all samples;
- M5: substitution of zero by $\frac{1}{2}\min(Y_{ij})$, where Y_{ij} is the quantitative FIT measured indexed by patient and time, indicating all samples;
- M6: substitution of zero by $\frac{1}{2}\min(Y_{ij}) \pm X \sim \mathcal{N}(0, 0.00005)$, where Y_{ij} is the quantitative FIT measured indexed by patient and time, indicating all samples.

We made one exception before applying the PMM approach outlined in the previous section. Rather than calculate risk scores using LOOCV, we used K-fold cross-validation, where K = 10, to minimize overfitting while still being computationally efficient. Before making this modification, we tested the approach on the CA-125 data and compared those results to the LOOCV results from our reproduction and those published by Han et al. (2020). Performance using 10-fold cross-validation differed by less than 1% and are included in Table A.1. Similarly, these results are comparable to the AUCs published originally for each of the cutoff times, suggesting any influence of the proposed modification was negligible. The 10-fold validation also greatly decreases computational time.

The data corresponding to each method was analyzed with the PMM-CN3 approach and univariate logistic regression. This second approach includes only the most recent FIT to evaluate the probability of advanced neoplasia, mirroring current screening and [51]. Table 3.4 includes the estimated AUC and standard error obtained from these models for each method. We compare model performance using the DeLong test with pROC [41]. The

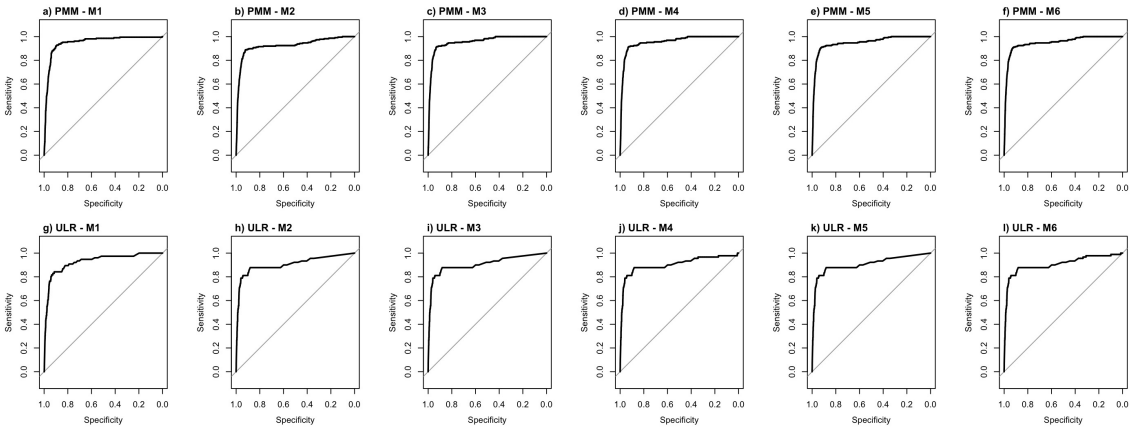
DeLong test incorporates AUC estimates and their corresponding variances to assess whether there is a statistically significant difference in the predictive ability of the models [13]. The "Difference of AUC" is calculated as the estimated AUC for PMM-CN3 minus the estimated AUC for the univariate logistic regression model for each method.

Method	PMM-CN3	Logistic Regression	P-Value	Difference of AUC
M1	0.950 (0.007)	0.927 (0.017)	0.226	0.022
M2	0.925 (0.012)	0.906 (0.022)	0.437	0.019
M3	0.956 (0.007)	0.907 (0.022)	0.029	0.049
M4	0.956 (0.007)	0.905 (0.022)	0.030	0.051
M5	0.949 (0.008)	0.906 (0.022)	0.067	0.043
M6	0.949 (0.008)	0.906 (0.022)	0.071	0.057

Table 3.4: AUC estimates and standard errors (shown in parentheses) of various data methods for the PMM-CN3 and univariate logistic regression approaches

The estimated AUC for the PMM-CN3 approach is consistently greater than that of the logistic regression approach, regardless of the method employed(Figure 3.1). In the case of M3 and M4 there is a notable difference (p = 0.029 and 0.030, respectively).

Figure 3.1: Receiver operating characteristic (ROC) curves for model performance by transformation accommodation method.



The standard error for the estimated AUC is also smaller for PMM-CN3 compared to the univariate logistic regression model for each implementation. Logistic regression performance seems more robust to variation in the ad hoc substitution methods used, yielding consistent AUC estimates and corresponding standard errors. The only exception was the complete case (M1) dataset, which yielded the largest AUC for the logistic regression approach. It is important to note, however, the improvement came at great expense, forfeiting nearly half (43.64%) of the available patient data.

Unlike logistic regression, PMM-CN3 appears to be more sensitive to the method used to accommodate the log transformation, linking us back to the original concern outlined in Chapter 2. Specifically, we see the fluctuation in estimated AUC and corresponding SE based on M, the small positive constant added before transformation. AUC is lowest when M is set to 0.01, the recommendation in West (2022). AUC increases 2.3% when M is 50 times greater at one-half the observed non-zero minimum, with and without variation. Similarly, when M is 100 times greater at the observed non-zero minimum, with and without variation, AUC increases by 3.1%. These results align with the caution issued by Feng, et al. (2012) that the model estimates from the translated data depend on how M is selected. In the next chapter, we will introduce and implement an alternative approach to accommodate log transformation in data with a high concentration of zeros before analyzing the CSPAN data using PMM.

Chapter 4

Quantitative FIT Scores as a Left-Censored Outcome

In Chapter 2, we compared CA-125 and hemoglobin concentration distributions as shown in figures 2.2 and 2.3, respectively. Both are right-skewed, but the range of possible values differs in that CA-125 is always greater than zero but quantitative hemoglobin is often not (30%), impacting the feasibility of a logarithmic transformation. Using biomarkers requires some consideration for instrument precision and random measurement error. Even when levels are sufficient, biomarker quantification could be compromised because the instrumentation could not detect low values, resulting in missing data [43]. Suppose the quantitative hemoglobin measurements equal to zero are considered below the detection limit and evaluated as missing, rather than truly zero. The commonly chosen shift parameters would then be referred to as ad hoc substitution while exclusion of the nonzero measurements is truly a complete case analysis. The approaches implemented to facilitate data transformation are still appropriate in the context of missing data but not without concern. Complete-case analysis is inefficient, as was illustrated with the CSPAN dataset [10, 20, 43]. Ad hoc substitutions, often chosen as the detection limit (M3) or one-half the detection limit (M5), can lead to biased estimates for parameters and standard errors [10, 43, 50]. This approach can also understate variability, which we attempted to address through coerced variation M4 and M6 [20].

In this chapter, we consider measurements of zero as being below the detection limit and therefore missing. Imputation-based approaches allow missing data to be replaced by one or more plausible set of values to be analyzed in conjunction with the observed data as a completed data set [20, 36]. The "complete" dataset can then be considered a valid sample from the population under certain assumptions and given the other information available for those subjects, resulting in inferences within the realm of statistical plausibility obtained

had there been no missing data [20, 36]. While bias is still a concern, these approaches have been shown to outperform ad hoc substitution with a constant [50]. As a result, we used an imputation-based approach to replace the zeros and accommodate data transformation before analyzing the "completed" dataset with PMM and univariate logistic regression.

4.1. Left-Censoring

The limit of detection (LOD) is the smallest value at which an instrument can delineate between the presence and absence of substance [10, 43]. When the level of the substance is below the LOD, the value recorded is censored but does not always indicate the absence of the biomarker. When the assay has a lower detection limit, the result is considered to be censored from below, or left-censored.

While no minimum detection limit has been reported for quantitative hemoglobin concentrations obtained by the automated semi-quantitative OC-Sensor, data and literature provided by the manufacturer document the test reliability only for the range of 10 - 400 μg/g [51]. It is important to note this cutoff is well below the 20μg/g threshold recommended by the manufacturer and commonly used in screening. Any variability in this range would not be expected to influence performance when used as a qualitative test. Because no data is available for test reliability at cutoffs below 10 μg/g, it is possible these results become gradually less reliable below this threshold [25]. As a result, we intend to use the minimum observed non-zero FIT score as the censoring threshold. Any measurement less than the censoring threshold will be treated as missing or left-censored and replaced with an imputed value before applying PMM.

4.2. Non-Parametric Inverse Buckley-James

Many approaches applied to left-censored data analysis are based on methods developed for right-censored survival data. The accelerated failure time model is frequently used as a regression model for right-censored survival data, as it directly links the expected response to the predictors [50]. The Buckley-James estimator is a popular method for fitting the ac-

celerated failure time model allowing the dependent variable to be censored and the residual distribution unspecified [7]. It was originally used in conjunction with data from the Stanford Heart Transplantation Program to illustrate its application but has also been applied more recently to left-censored human immunodeficiency virus (HIV) viral load data [7, 50].

This approach imputes censored values by their estimated conditional mean to provide censoring and predictor values [50]. Even though it was developed under right censoring, it can be readily applied to left-censored data by reversing the scale. Once the data has been inverted from left-censored to right-censored, the algorithm can be directly applied using the bujar package in R [35, 52]. The process is as follows [50]:

1. Let Y_i be the response variable and LOD be fixed and known. Z_i is the observed response defined as:
$$Z_i = \begin{cases} Y_i & \text{if} \quad Y_i > \text{LOD} \\ \text{LOD} & \text{if} \quad Y_i \leq \text{LOD} \end{cases}$$
2. Define an arbitrary constant $M \geq \max(Y_i)$
3. Reverse the order of the data: $M - Y_i$
   - ▷ *Note:* Left-censored $\mathbf{Z}$ is replaced by $(M - \mathbf{Z})$ which is now right-censored at $M - \text{LOD}$.
4. Impute $(M - Z_i)^*$ as:
$$\delta_i(M - Y_i) + (1 - \delta_i)\mathbb{E}(M - Y_i | M - Y_i \geq M - \text{LOD}, \mathbf{X_i}),$$
   where $\mathbf{X_i}$ is a p-vector of fixed predictors
5. Calculate the conditional expectation by:
$$\mathbb{E}(M - Y_i | M - Y_i \geq M - \text{LOD}, \mathbf{X_i}) = \int_{M-\text{LOD}}^{M} \frac{u f(u, \mathbf{X_i}, \boldsymbol{\beta}) du}{1 - F(M - \text{LOD}, \mathbf{X_i}, \boldsymbol{\beta})}, \qquad (4.1)$$
   where $\mathbf{X_i}$ is a p-vector of fixed predictors, $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters, and $F(u, \mathbf{X_i}, \boldsymbol{\beta})$ is the (unknown) cumulative density function for $M - Y_i$ with mean $M - \mathbf{X_i}\boldsymbol{\beta}$ evaluated at u. $F(u, \mathbf{X_i}, \boldsymbol{\beta})$ can be estimated using Kaplan-Meier [50].

6. Compute the Buckley-James estimate using a semiparametric iterative algorithm:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} ||(M - \mathbf{Z})^* - (M - \mathbf{X}\beta)||_2^2 + \lambda||\beta||_1,$$

alternating between the imputation of censored values using Equation 4.1 and the Lasso, where $\lambda = 0$ corresponds to the unpenalized ordinary least-squares estimate.

Soret, et al. (2018) defined censoring rates as moderate (20% censoring), high (50%), and severe (70%). Figure 4.1 indicates similar performance between the nonparametric Buckley-James approach with moderate censoring and the gold standard with uncensored data [50]. The error increases when censoring is high, but the change is most notable when censoring is severe [50]. Recall nearly 30% of the quantitative FIT measurements are zero in the CSPAN dataset. If we consider those measurements below the LOD, the censoring rate would fall midway between moderate and high based on the classification used by Soret, et al. (2018). The simulation results presented in figure 4.1 then suggest non-parametric Buckley-James approach should still perform well when applied to the CSPAN data.

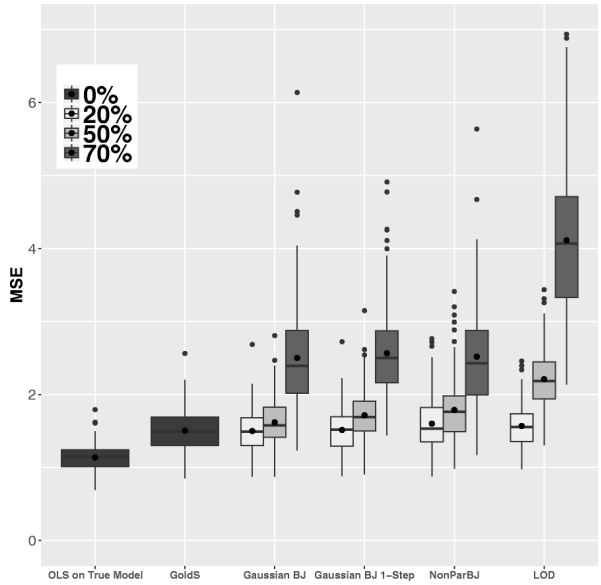


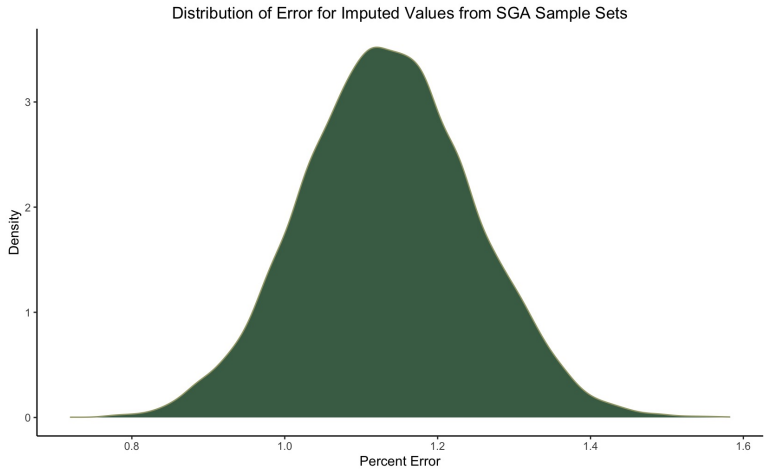
Figure 4.1: Simulation Results made available through the Creative Commons Attribution 4.0 International License (`https://tinyurl.com/ReprintsPermission`) [50]. Mean Square Error (MSE) from gold standard method (GoldS), Gaussian Buckley-James (Gaussian BJ), 1-Step version (Gaussian BJ 1-Step), non-Parametric Buckley-James (NonParBJ), and simple imputation by LOD (LOD). Censoring rates: 0% (uncensored), 20% (moderate), 50% (high), and 70% (severe).

4.3. Simulated Data: Fetal Growth Example

Before applying the non-parametric Buckley-James approach to the CSPAN dataset, we tested it on a complete data set with imposed censoring. We used data from the Scandinavian portion of the NICHD (National Institute of Child Health and Human Development) Study of Successive Small-for-Gestational Age (SGA) Births [5]. This dataset was used by Liu and Albert (2014) to illustrate the application of the PMM approach compared with SREM. Of the 5,722 eligible women expecting a second or third child between January 1986 and March 1988, 1,945 women and their births were selected for follow-up at four prenatal visits, delivery, and during the first year of life [5]. Excluding those women who failed to complete more than one of the four pregnancy examinations, the final dataset contained 1,116 subjects. Data elements included maternal pre-pregnancy risk factors associated with SGA birth, including 1) a prior low birth weight (LBW) birth, 2) maternal cigarette smoking at conception, 3) low pre-pregnancy weight (< 50 kg), 4) a previous perinatal death, or 5) the presence of chronic maternal disease (namely, chronic renal disease, essential hypertension, or heart disease) [5]. The longitudinal marker is the mean abdominal diameter measured at approximately 17, 25, 33, and 37 weeks of gestation and reported on the log scale [5, 30]. We chose this dataset because it has been analyzed previously with PMM, has no detection limit, and unlike the CA-125 dataset, the complete dataset used in the publication is available.

We used the quantile function in R to identify the 30^{th} percentile in the fetal growth data. Any observations less than the 30^{th} percentile were considered censored. The LOD was set at 2.262 to maintain a 30% censoring rate. Sampling with replacement, we generated 10,000 samples of size N = 1,555. We found values imputed for measurements less than the imposed LOD were very close to the actual measurement recorded for mean abdominal diameter. We calculated the percent error for each of the imputed measurements in each of the sample datasets using $\frac{(\text{Imputed}-\text{Original})}{\text{Original}} \times 100$. When averaged over all sample sets, the mean was 1.14% with a standard deviation of 2.28%. The distribution of percent error for imputed measurements in the sample sets is shown in figure 4.2. While some bias is present, as the percent error is greater than zero, it is small.

Figure 4.2: Distribution of percent error for imputed values in 10,000 samples (N= 1,555) with 30% censoring



Using the imputed value in place of the "censored" value, each "completed" dataset was analyzed using PMM. For fetal growth, the longitudinal profile for ultrasound anthropomorphic measurements is characterized by

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + b_{i0} + b_{i1} t_{ij} + b_{i2} t_{ij}^2 + \epsilon_{ij}.$$

As a result, the case and control models used in the PMM were the same [14, 30, 48]. Table 4.1 includes the parameter estimates and AUC generated by the original dataset and the mean for the estimates and AUC obtained by each sample set (N = 10,000).

	Original Dataset		Imputed Datasets	
	Controls	Cases	Controls	Cases
β_0	1.5001 (0.0045)	1.5117 (0.0107)	1.5230 (0.0045)	1.5300 (0.0078)
β_1	0.8803 (0.0046)	0.8736 (0.0122)	0.8578 (0.0046)	0.8532(0.0080)
β_2	-0.0882 (0.0010)	-0.0927 (0.0028)	-0.0837 (0.0010)	-0.0831(0.0017)
σ_{b_0}	0.0798	0.0203	0.1286	0.0910
σ_{b_1}	0.0769	0.0381	0.1396	0.0934
σ_{b_2}	0.0151	0.0087	0.0315	0.0200
ρ_{01}	-0.9109	-0.8294	-0.9523	-0.9467
ρ_{02}	0.8391	0.6110	0.9002	0.8911
ρ_{12}	-0.9789	-0.9490	-0.9844	-0.9810
σ_ϵ	0.0227	0.0241	0.0190	0.0286
AUC	0.847		0.856	

Table 4.1: Estimated coefficients and standard errors for the PMM models based on the original and imputed datasets for the fetal growth study.

The imputed datasets produced similar estimates to the original dataset. While the random effects showed greater variability in magnitude compared to the fixed effects, directionality is consistent between the two sets of outcomes. Similarly, trends in magnitude are also maintained when comparing the case and control models in each scenario. Lastly, model performance is consistent even when the imputed data is used. The percent error was 1.05%, suggesting this approach can be successfully implemented even in the presence of heavier censoring.

4.4. CSPAN Data

We then applied the inverse nonparametric Buckley-James approach to the CSPAN data and analyzed the "completed" dataset with PMM. To do so, we first split the dataset to exclude entries where values are truly missing in subsequent rounds of screening, because our intent is not to fill in the missing data, but rather only to update those measurements below the imposed LOD. The quantitative FIT measurement becomes Y_i within the Buckley-James

algorithm, and the LOD is set to one, the minimum observed value. We identify M as the maximum observed value plus one, before subtracting all observed values to reverse order the data. We then built the model matrix, $\mathbf{X}$, including baseline age, race, ethnicity, sex, and screening time as variables. We let $\lambda = 0$ in step 7 from section 4.2, corresponding to the unpenalized ordinary least-squares estimate. Again using the bujar function in R we obtain the imputed values, subtracting from M to reorder the data. It is important to note that with the CSPAN data, the algorithm produced negative values for the imputed quantitative FIT score; this was not an issue in the fetal growth dataset. We replaced the zeros with a small positive constant less than the smallest observed non-zero value and log-transform the measurements. This was considered acceptable because the MLE for LOD is the minimum of the observed values. Also, the small positive constant is only added to those measurements below the LOD, which are then replaced through imputation. After running the rest of the algorithm, the imputed response variables are now non-negative after reordering and exponentiating back to the original scale. We then combine those records with newly imputed values or observations exceeding the LOD with those missing data to create the "completed" dataset.

These data are analyzed with the PMM approach described in chapter 3, using 10-fold cross-validation when calculating risk scores. The data was also analyzed using univariate logistic regression, where only the most recent FIT result is considered, mirroring current screening practice. As seen in Chapter 3, the PMM approach led to moderate improvement in the out-of-sample AUC (0.951) compared with univariate logistic regression (0.921). The standard error for PMM-CN3 is also approximately one-third that of the univariate logistic regression model (0.008 vs. 0.023). While neither result exceeds the maximum AUCs reported in table 3.4, both are consistent with the results achieved through M1, the complete case analysis, without sacrificing any data. Our results indicate the inverse non-parametric Buckley-James approach is a strong alternative when modeling with a moderate to high concentration of zeros in the response variable.

Chapter 5

Modeling Biomarkers with Functional Data Analysis

Biomarkers were defined in Chapter 1, highlighting their role as an inexpensive and noninvasive screening tool. We included CA-125 and fecal hemoglobin as specific examples, where the first is used for ovarian cancer screening and the latter to screen for CRC. We also provided case and control trajectories for CA-125 in Figure 1.1 and for hemoglobin in Figure 2.1. Both biomarkers show distinctive patterns for those with the disease (cases) compared to those without disease (controls). Until now, our attempts to harness the additional information available through repeated measurements have treated these data as sequential discrete observations. Here we propose an alternate strategy, extracting information from the entire biomarker trajectories using functional data analysis (FDA) in place of the individual biomarker measurements.

5.1. Functional Data Analysis (FDA)

In functional data analysis (FDA) we consider longitudinal data as sets of discrete observations on smooth underlying curves [37]. The basic unit of information is then the entire observed function rather than a string of individual values. The goals of functional data analysis include: 1) representing and transforming data in ways to aid further analysis, 2) displaying data in a way that highlights various characteristics, 3) studying important sources of pattern and variation in the data, and 4) explaining variation in an outcome or dependent variable by using input or independent variable information [37]. FDA enables us to bring the biomarker trajectory information into the model using a countable linear combination of the basis functions. The functional principal component (FPC) scores are one choice for the set of basis functions [26, 56]. The goal is to obtain the orthogonal functions that most efficiently describe the variation in the data [26, 56]. FPC analysis is commonly

used for dimension reduction, condensing the trajectories from longitudinal data to a set of FPC scores [26, 56]. Yao et al. (2005) developed a method for FPC analysis in which the FPC scores are framed as conditional expectations. Their nonparametric approach, principal component analysis through conditional expectation (PACE) for longitudinal data, was designed for sparse, irregularly spaced longitudinal data [56]. PACE allows us to represent the biomarker trajectories through the Karhunen-Loéve expansion and is computed by determining the eigenfunctions from the data [26, 56]. The number of FPC scores needed is identified based on a combination of the Akaike information criterion (AIC) and the scree plot, i.e. where AIC has approached its minimum and the plot has plateaued[26]. We implement this approach in R using the fdapace package [57]. We can then approximate a generalized functional linear regression model, by including the FPC scores obtained through this approach as covariates in the model.

5.2. Modeling with Biomarker Trajectories

Patients in most cancer screening programs are selectively chosen for disease verification with the gold standard based on the results from preliminary testing. The patient's true disease status is confirmed only when the first test is abnormal. Otherwise, the patient's disease status is missing, introducing the potential for verification bias. We can address verification bias as a missing data problem by considering true disease status to be either missing at random (MAR) or missing not at random (MNAR) [26]. As we describe how to incorporate FDA using PACE to obtain a new composite diagnostic marker for disease risk, we include models for MAR and MNAR.

5.2.1. Notation

PACE allows for irregularly spaced longitudinal data. To align the data we discretize visit time t_{ik} into the kth visit, where $k = 1, 2, 3, ...$, for each individual [26]. For each patient, at

each visit t_k we collect the following: disease status

$$d_{t_k} = \begin{cases} 1 & \text{if verified and diseased} \\ 0 & \text{if verified and nondiseased} \\ missing & \text{if not verified,} \end{cases} \tag{5.1}$$

a p-dimensional biomarker measurement $M_{t_k} = (M_{t_{k_1}}, M_{t_{k_2}}, ..., M_{t_{k_p}})^\mathsf{T}$, a $t_k \times p$ biomarker measurement $Q_{t_k} = (M_{t_1}^\mathsf{T}, M_{t_2}^\mathsf{T}, ..., M_{t_k}^\mathsf{T})^\mathsf{T}$, and the missing data indicator

$$r_{t_k} = \begin{cases} 1 & \text{if } d_{t_k} \text{ is observed} \\ 0 & \text{otherwise,} \end{cases} \tag{5.2}$$

and a test result variable to be defined later [26]. We use the missing disease status indicator to model as MAR or MNAR. The next two subsections give the model details for each mechanism.

5.2.2. MAR

MAR indicates the missing data mechanism depends only on the observed data [29]. In FIT-based cancer screening, test positivity is defined by the manufacturer's recommended cut-off value of 20μg/g. When the fecal hemoglobin concentration exceeds the cut-off value, the test is positive and the patient is referred to colonoscopy, otherwise the test is negative and the patient is re-invited to FIT in the subsequent year. Only patients completing colonoscopy will have true disease status verified as advanced neoplasia or "normal." The disease status will be missing in all other patients. Under the MAR assumption, verification status is obtained directly from the missing data indicator r_{t_k} for any patient at a given visit time t_k because both the biomarker value M_{t_k} and missing data indicator r_{t_k} are observed. Disease status d_{t_k} is assumed to be observed only for those patients with a positive test result [26]. As a result, we can also assume that given both the biomarker measurements and disease status, the verification process depends only on the biomarker measurements [26].

For a set of n patients, we will have data $(Q_{it_k}, d_{it_k}, r_{it_k})$ at time t_k. We are interested in modeling the risk of disease given the biomarker measurements, $P(d_{it_k} = 1|Q_{it_k})$. This probability is the new composite diagnostic marker for disease risk to be used in place of a single biomarker value. Based on the MAR assumptions outlined previously, this can be done using only the data from those patients with verified true disease status at time t_k [26].

We first compute a finite number of FPC scores for the data at measurement time t_k with fdapace. We then approximate the logistic functional model for risk using a logistic regression model where true disease status is defined as a binary variable and the extracted FPC scores are covariates. Specifically,

$$\begin{aligned} g(P(d_{it_k} = 1|f_{M_{i1}}(t), \ldots, f_{M_{ip}}(t))) &\approx g(P(d_{it_k} = 1|\xi_{it_k})) \\ &\approx \alpha_0 + \sum_{l=1}^{L_{1k}} \alpha_{1l}\xi_{i1l}(t_k) + ... + \sum_{l=1}^{L_{pk}} \alpha_{pl}\xi_{ipl}(t_k), \end{aligned}$$

where $L_{1k}, L_{2k}, \ldots, L_{pk}$ are the finite number of principal components for biomarker trajectories 1,2,...,p, and $\xi_{ijl}(t_k)$ is the lth FPC score obtained from the measurements of biomarker j for patient i up to visit t_k [26]. The likelihood function for the observed d_{it_k} is then proportional to $\prod_{i=1}^{\nu_k} p_{it_k}^{d_{it_k}}(1 - p_{it_k})^{1-d_{it_k}}$, where ν_k is the total number of patients whose true disease status is observed at time t_k, and

$$p_{it_k} = g^{-1}\left(\alpha_0 + \sum_{l=1}^{L_{1k}} \alpha_{1l}\xi_{i1l}(t_k) + ... + \sum_{l=1}^{L_{pk}} \alpha_{pl}\xi_{ipl}(t_k)\right).$$

We then combine this result with the optimal classification rule developed by McIntosh and Pepe (2002) to define patients as testing positive or negative at time t_k. If $\mathbf{Y}$ is the vector of test results (i.e. markers or other screening tests), the Neyman-Pearson lemma states that for any f_0, where f_0 is defined as 1 - specificity, the screening rule with the highest true positive rate (TPR) based on $\mathbf{Y}$ among all possible rules based on $\mathbf{Y}$ is the likelihood ratio rule [32]. McIntosh and Pepe (2002) then show rules based on LR($\mathbf{Y}$) are equivalent to rules based on the risk score $p(\mathbf{Y}) = P(D = 1|\mathbf{Y})$, which can be approximated with binomial regression as $\text{logit}(p(\mathbf{Y})) = \beta_0 + h(\beta, \mathbf{Y})$. Having established $p(\mathbf{Y}) = P(D = 1|\mathbf{Y})$, this can

be rewritten using Bayes' rules as

$$\frac{P(\mathbf{Y}|D=1)P(D=1)}{\{P(\mathbf{Y}|D=1)P(D=1)+P(\mathbf{Y}|D=0)P(D=0)\}}$$
$$=\frac{LR(\mathbf{Y})q}{\{LR(\mathbf{Y})q+1\}},$$

where $q = P(D=1)/P(D=0)$ is the odds of disease in the population [32]. Given the risk score is a monotone increasing function of the likelihood ratio, the likelihood ratio rule can be rewritten as $p(\mathbf{Y}) > c^*(f_0)$, where $c^*(f_0)$ is chosen where f_0 = 1-specificity [32]. Replacing $\mathbf{Y}$ with Q_{t_k} to reflect our notation, we then define risk, $p(Q_t)$, as the new diagnostic marker [26].

5.2.3. MNAR

Ideally, true disease status would be verified for all patients with an abnormal result in the initial screening test; however, this is not feasible in many cases because the patient is either too sick or refuses to be verified [22]. In this scenario, the verification process is considered to be MNAR, where the missing data mechanism may depend on both the observed data (i.e. test results) and unobserved data (i.e. true disease status or other risk factor(s) directly related to true disease status) [29]. As we cannot verify the missing data mechanism, we also present the model derived under the less restrictive MNAR setting.

Unlike under MAR we cannot reduce the likelihood function to include only the verified patients. Instead the joint likelihood function is

$$\prod_{i=1}^{n} P(d_{it_k}, r_{it_k}|Q_{it}) = \prod_{i=1}^{n} P(r_{it_k}|Q_{it})P(d_{it_k}|r_{it_k}, Q_{it})$$

, where n is the total number of patients [26]. Using the same notation for biomarker measurements and corresponding smooth functions, we use a pattern mixture model for missing data [34]. Using the same notation as before, the missing data indicator model is

$$g(P(r_{it_k}=1|f_{M_{i1}}(t),\ldots,f_{M_{ip}}(t))) \approx g(P(r_{it_k}=1|\xi_{it_k}))$$
$$\approx \beta_0 + \sum_{l=1}^{L_{1k}} \beta_{1l}\xi_{i1l}(t_k) + \ldots + \sum_{l=1}^{L_{pk}} \beta_{pl}\xi_{ipl}(t_k),$$

and functional logistic regression is again approximated using the extracted FPC scores as covariates [26]. We use the same approach to model risk given the missing indicator and biomarker measurements, where

$$g(P(d_{it_k}=1|r_{it_k},f_{M_{i1}}(t),\ldots,f_{M_{ip}}(t))) \approx g(P(d_{it_k}=1|r_{it_k},\xi_{it_k})) \\ \approx \gamma_0+\gamma_1 r_{it_k}+\sum_{l=1}^{L_{1k}}\gamma_{21l}\xi_{i1l}(t_k)+...+\sum_{l=1}^{L_{pk}}\gamma_{2pl}\xi_{ipl}(t_k)$$

[26]. Thus at time t_k, $r_{it_k}|\xi(t_k) \sim \text{Bernoulli}(\theta_{it_k})$ and $d_{it_k}=1|r_{it_k},\xi(t_k) \sim \text{Bernoulli}(q_i t_k)$, where θ_{it_k} and q_{it_k} are the inverse logit for the corresponding models.

Given d_{t_k} can either be missing or observed, we have two missing data patterns requiring two separate models. In pattern 1, $r_{it_k}=1$ and d_{it_k} has been verified in all patients. This can be modeled as

$$g(P(d_{it_k}=1|r_{it_k}=1,f_{M_{i1}}(t),\ldots,f_{M_{ip}}(t))) \approx \gamma_0^{(1)}+\sum_{l=1}^{L_{1k}}\gamma_{21l}^{(1)}\xi_{i1l}(t_k)+...+\sum_{l=1}^{L_{pk}}\gamma_{2pl}^{(1)}\xi_{ipl}(t_k),$$

where all parameters are identifiable [26]. While in pattern 2, $r_{it_k}=0$ and d_{it_k} is missing. This is modeled similarly as

$$g(P(d_{it_k}=1|r_{it_k}=0,f_{M_{i1}}(t),\ldots,f_{M_{ip}}(t))) \approx \gamma_0^{(0)}+\sum_{l=1}^{L_{1k}}\gamma_{21l}^{(0)}\xi_{i1l}(t_k)+...+\sum_{l=1}^{L_{pk}}\gamma_{2pl}^{(0)}\xi_{ipl}(t_k),$$

the exception being the parameters here are not identifiable [26]. We can then estimate disease risk using

$$P(d_{it_k}=1|\xi(t_k))=P(d_{it_k}=1,r_{it_k}=1|\xi(t_k))+P(d_{it_k}=1,r_{it_k}=0|\xi(t_k))$$
$$=P(r_{it_k}=1|\xi(t_k))P(d_{it_k}=1|r_{it_k}=1,\xi(t_k))+P(r_{it_k}=0|\xi(t_k))P(d_{it_k}=1|r_{it_k}=0,\xi(t_k)),$$

where $\gamma_0^{(0)}=\gamma_0^{(1)},\gamma_{21l}^{(0)}=\gamma_{21l}^{(1)},\ldots\gamma_{2pl}^{(0)}=\gamma_{2pl}^{(1)}$ are the identifying restrictions established under the MAR assumptions [26]. This can then be rewritten as $\gamma_0^{(0)}=\gamma_0^{(1)}+\delta_0,\gamma_{21l}^{(0)}=\gamma_{21l}^{(1)}+\delta_1,\ldots\gamma_{2pl}^{(0)}=\gamma_{2pl}^{(1)}+\delta_p$ once we reparametrize the model to embed the MAR constraints using the sensitivity parameter δ [26]. The missing data mechanism can thus be classified as MAR when $\delta_0=\delta_1=\ldots=\delta_p=0$ and MNAR otherwise.

5.3. Model Implementation

We first apply this approach to analyze the PLCO ovarian cancer data also described in chapter 1. Women aged 55 and older completed a CA-125 blood test annually and those with an abnormal result received a follow-up biopsy. In our subset, participants completed between 1 and 6 blood tests, where the median number of completed tests was 3 for cases and 6 for controls. These tests were completed on average at 1, 12, 24, 37, 48, and 61 months.

We used R version 4.3.1 for our analysis[35]. We used fdapace to apply the PACE method and obtain FPC scores for the CA-125 trajectory. Based on the AIC and scree plot, we identified $L_1 = 1$ as the number of FPC scores needed in the model. We randomly selected two-thirds of the data as training data, leaving the remaining one-third as test data; this was the ratio used by Li and Gatsonis (2017). We fit the logistic regression model with the FPC scores as covariates to approximate the logistic functional regression model and obtain an estimate for the composite diagnostic marker. Following Li and Gatsonis (2017), we used R^2, really the McFadden pseudo-R^2, and the Hosmer and Lemeshow goodness-of-fit test to evaluate fit for the logistic regression model, while the C statistic was used to check the prediction error. We also used the AUC of the empirical ROC curve to evaluate diagnostic accuracy.

Analyzing the data under the MAR assumption, we reduced the dataset to include only those participants with verified disease status (n = 284). Our R^2 was 0.204, the Hosmer and Lemeshow goodness-of-fit test suggests the evidence may be insufficient to indicate poor model fit (p = 0.286), and our C statistic was 0.693. AUC was estimated at 0.717, which suggests discrimination for this approach is acceptable but falls short of that achieved by PMM as reported in table 3.3.

We repeated the analysis under the MNAR assumption. Rather than add the sensitivity parameter δ to every parameter, where we may not know its real effect, our sensitivity analysis only includes the intercept term [26]. This was chosen because δ_0 is the log odds difference when comparing the odds $d_{it_k} = 1$ between observed true disease status and missing true disease status, where the log odds of $d_{it_k} = 1$ for missing true disease status

is δ_0 units(s) larger than the log odds of $d_{it_k} = 1$ for observing true disease status relative to the MAR. As a result, the intercept for the regression model was $\gamma_0^{(0)} = \gamma_0^{(1)} + \delta_0$ under the MNAR assumption [26]. We varied δ from -1.6 to 0.8, which accounts for a range in prevalence from 9% to 54% among patients whose true disease status is missing. We used the same number of principal components for the CA-125 biomarker, $L_1 = 1$. The diagnostic results are included in table 5.1. Our results indicate discrimination is below an acceptable threshold and still far lower than what was achieved using PMM.

Table 5.1: AUC estimates and standard errors (shown in parentheses) for sensitivity analysis from the FDA approach under the MNAR assumption for CA-125 biomarker

δ_0	Estimated AUC (se)
-1.6	0.584 (0.052)
-1	0.595 (0.094)
-0.4	0.612 (0.074)
0.4	0.610 (0.052)
0.8	0.617 (0.072)

We also used this approach to analyze the data from the CSPAN program. Recall in Chapter 2, patients aged 50 and older were invited annually to complete FIT-based screening. Participants with an abnormal FIT result were navigated to colonoscopy, while those with a normal FIT result were re-invited to FIT screening. Participants completed between 1 and 8 FIT during the study period (Table 2.1), with a median of 1 and interquartile range of 1. These FIT were completed on average at 2, 17, 29, 42, 48, 61, 80, and 87 months.

We first obtained FPC scores for the FIT trajectory. Based on the AIC and scree plot, we identified $L_1 = 4$ as the number of FPC scores needed in the model. We then analyzed the data under the MAR assumption, reducing the dataset to include only those participants with verified disease status (n = 1,301). We used the same ratio to split the data into testing and training sets. We fit the model using the training data and then applied the model-fitting results to the test data Our R^2 was 0.018, the Hosmer and Lemeshow goodness-of-fit test

was statistically significant (p = 0.021), and our C statistic was 0.511. These results suggest the logistic regression model did not fit well and our prediction rate was poor. Our AUC estimate was 0.540, suggesting the FDA approach has insufficient discrimination compared to PMM or even the standard screening approach, both of which had AUCs ranging from 0.905 to 0.956 as shown in Table 3.4.

We repeated the analysis under the MNAR assumption, applying sensitivity analysis only to the intercept term and maintaining the same number of principal components, L_1 = 4. We allowed δ to vary from -1.6 to 0.8, accounting for the same range in prevalence. The diagnostic results are included in table 5.2. Like Li and Gatsonis (2017), we see small variations in the estimated AUCs when the δ_0 varies, but our results again suggest poor discrimination when applied in this screening setting.

δ_0	Estimated AUC (se)
-1.6	0.515 (0.022)
-1	0.514 (0.021)
-0.4	0.514 (0.022)
0.4	0.513 (0.021)
0.8	0.513 (0.021)

Table 5.2: AUC estimates and standard errors (shown in parentheses) for sensitivity analysis from the FDA approach under the MNAR assumption for FIT

5.4. Limitations to the FDA Approach

While the results obtained by Li and Gatsonis (2017) suggest the richer information available through biomarker trajectories should improve the precision with which we predict disease, our implementation showed otherwise. To understand the difference in performance, it is important to highlight key aspects of divergence between our settings.

The FDA approach was presented as a means of improving prediction for disease recurrence by combining information from multiple biomarker trajectories [26]. They similarly

showed this approach performed well when applied to a single biomarker using simulated CA-125 data [26]. While we may attribute some loss in performance to the constraint of a single biomarker in both the CRC and ovarian cancer screening scenarios, we anticipate two other factors to be greater contributors to the change.

The first is sparseness of measurements. PACE was designed to used in settings where the number of repeated measurements available per subject is small, yet the working assumption is a dataset is treated as sparse if it has on average ≤ 20, potentially irregularly spaced, measurements per subject [56]. The examples included by Yao et al. (2005) range from 1 to 14 measurements per patient, with a median of 6, and 1 to 6 with a median of 3. While the predictions based on a single measure work reasonably well, it is not feasible to apply the method when there is only one observation available per subject for all subjects as we need to be able to consistently estimate the covariance structure. Similarly, the simulated CA-125 dataset would have 1 to 12 measurements per patient, and the applied datasets range from 1 to 7 and 1 to 5 measurements per patient, but no median value is provided [26]. When we consider the CSPAN data, however, we have 1 to 8 measurements per patient with a median of 1, which may be too sparse for PACE to glean enough meaningful information from the FIT trajectories. This notion is further supported by our application to the PLCO ovarian cancer screening data, where again the number of measurements ranged from 1 to 6, but had a median of 3 for cases and 6 for controls. While our results were still not as favorable as those published by Li and Gatsonis (2017), they were markedly improved compared to those obtained using the CSPAN dataset.

The other factor is the sparseness of the disease. Li and Gatsonis (2017) focused on predicting disease recurrence, which often occurs at a higher rate than initial disease onset. This is evidenced by the specifications for the simulated data and in the description of the first applied dataset; both have outcome occurrence of 80% [26]. Our biomarker data, however, originates from cancer screening, where incidence is much lower. As an example, a woman's risk for getting ovarian cancer is approximately 1 in 87, and the risk for CRC is nearly 1 in 23 for men and 1 in 25 for women [4, 46]. While confirmed incidence in the PLCO

dataset is higher at 7.24% and confirmed incidence in the CSPAN dataset is only 1.5%, both are much lower than recurrence rates in the examples published and may contribute to some loss in performance.

Chapter 6

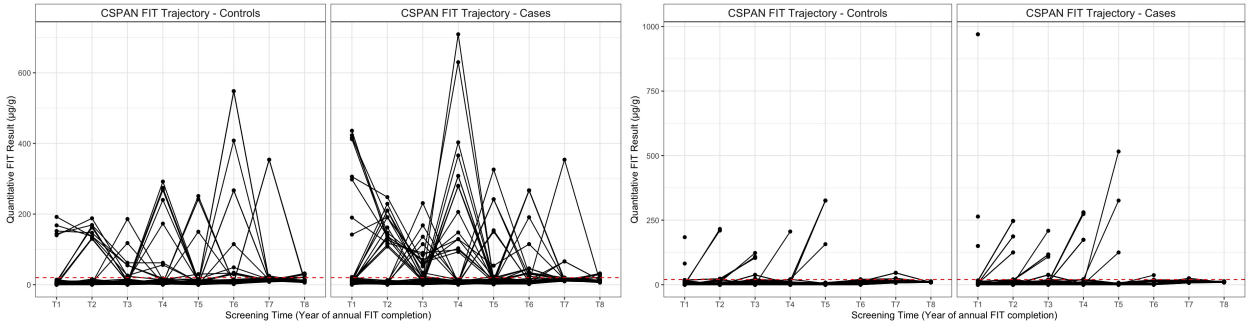
Establishing a New Decision Threshold

In prior chapters, we presented approaches for estimating risk scores using longitudinal biomarker measurements. We need a decision threshold to translate these findings into medical practice and support clinical diagnosis. This value dichotomizes the risk score into a binary decision, classifying values greater than or equal to the threshold as positive, and those below as negative. We focus on the PMM approach paired with inverse non-parametric Buckley-James based on its superior performance using the AUC as a measure.

6.1. Test Data

We first considered two imputation-based approaches to generate sample sets for testing. Using the microsimulation model, CRC-SPIN, we simulated colorectal cancer disease trajectories paired with outcomes from screening tests [42]. However, this approach only yields qualitative screening outcomes. Using the additional data available for the simulated cohort along with the data available through the CSPAN program, we attempted to impute the missing qualitative FIT measurements using Multivariate Imputation by Chained Equations (mice) and donor-based imputation, specifically k-Nearest Neighbour (KNN). Variables considered in the imputation included sex, age, screening time, qualitative screening outcome, and advanced neoplasia status. We did not include race or ethnicity, as CRC-SPIN does not account for either in its modeling [42]. We used the mice and VIM packages in R to implement imputation, but neither produced useable results. As indicated previously, PMM capitalizes on the differing biomarker trajectories for those with disease (cases) and those without disease (controls) illustrated in Figures 1.1 and 2.1, something imputation fails to capture through either approach. In Figure 6.1 we show the trajectories for a random sample of cases and controls by imputation method. While KNN (6.1b) more closely resembles what

we would expect given the true FIT trajectories from CSPAN, we see a regression to the mean-like behavior. Some values in the control group appear to be greater than expected and some values in the case group appear lower than expected, with some nonsensical values in both groups. The mice trajectories (6.1a) fail to capture any semblance of the expected trends. While the control peaks appear less severe than for cases, both contain implausible values.



(a) Imputation using mice

(b) Imputation using knn in VIM

Figure 6.1: Quantitative FIT Trajectories for 100 Cases (right panel) and 100 Controls (left panel) randomly selected from the imputed CRC-SPIN datasets. Each black line shows the FIT levels for one subject in the program. The horizontal red line indicates the 20 μg of hemoglobin per gram of feces cutoff used to define an abnormal FIT

Instead, we produced 500 bootstrap samples of size N = 10,000 from the original CSPAN dataset. While FIT-positivity for CSPAN was 7.82%, which aligns with values reported by other programs in the literature, cost-effectiveness models use FIT positivity parameters $\leq$10% [18]. We sampled FIT-positive and FIT-negative patients separately to allow FIT-positivity to vary while staying at or below 10% but greater than 0.

We first applied the inverse non-parametric Buckley-James approach described in Chapter 4 to impute the non-zero quantitative FIT measurements necessary to accommodate logarithmic transformation. The case and control models (PMM-CN3) described in Chapter 3 were then fit with the transformed data and out-of-sample risk predictions were calculated using 10-fold cross-validation. The original CSPAN data yielded an AUC of 0.951 with a standard error of 0.008, similarly, the mean AUC for the sample sets was 0.936 with a stan-

dard error of 0.010. We anticipate varying the FIT-positive rate is likely responsible for the small difference in AUC and slightly larger standard error. Nonetheless, these results indicate this approach is consistent and well able to discriminate between two clinically relevant subclasses of participants in a CRC screening program, those with and without advanced neoplasia.

6.2. Selecting a Decision Threshold

Incorporating a test into patient management requires a decision threshold and each threshold corresponds to a specific sensitivity/specificity pair [58]. As we consider all possible sensitivity/specificity pairs in pursuit of the optimal decision threshold, we also account for the relative burden of errors (i.e. false-positive and false-negative classifications) and the relative proportions of the two subclasses.

Burden can be defined in the context of financial or health-specific cost and from varying perspectives (i.e. patient, provider, insurer, community, etc.), but including some judgment for the relative burden of a false result is necessary [58]. We define burden using the clinical costs associated with a false result. A false positive result is approximately $1,300 based on the 2022 Medicare rate for unnecessary diagnostic testing, including pre-and post-operative consults, bowel prep, and the average cost of colonoscopy with and without polypectomy and pathology [8]. We define the burden of a false negative result by the increase in treatment cost associated with a later stage at diagnosis. We use the average difference in treatment cost associated with a single delay in cancer staging, which is approximately $35,000 [17]. In this setting, the relative cost for a false-negative result is much greater than the cost for a false-positive result, indicating our pair will favor sensitivity rather than specificity [58].

Prevalence interacts with the sensitivity and specificity to determine the actual probabilities of false-positive and false-negative results in the population and as such must also be included to identify the optimal sensitivity/specificity pair [58]. While the prevalence of advanced neoplasia increases with age, we conservatively used 6.2%, which is the prevalence in 50–54 year-olds [27]. We then combine these two elements to calculate a slope (m) for the

ROC curve, using:

$$\mathrm{m} = \left(\frac{\text{false-positive cost}}{\text{false-negative cost}}\right) \times \left(\frac{1-\mathrm{P}}{\mathrm{P}}\right),$$

where P = disease prevalence and [58]. The decision threshold corresponds to the operating point where a line with slope (m) intersects the ROC curve, and can also be obtained from the sensitivity and specificity pair which maximizes [sensitivity − m(1 − specificity)] [58]. The resulting threshold is the value that will yield the optimal mix of false-positive and false-negative results [58].

6.3. Results

To maximize the above function, we extracted sensitivity and specificity pairs from the AUC results obtained previously at varying thresholds for each sample set. We combine our false-positive/false-negative cost ratio of 0.036 and prevalence estimate of 6.2% to calculate m = 0.5505. We then iterate through the pairs until we obtain the maximum.

Based on this approach, the decision threshold is 0.02528; the corresponding sensitivity/specificity pair is 0.8997 and 0.9018, respectively. Specificity is slightly lower in comparison to values reported in the literature for single-sample FIT (0.91 (0.87 - 0.94)) and colonoscopy (0.93 (0.91 - 0.95)), but sensitivity is slightly higher relative to single-sample FIT (0.86 (0.68 - 0.95)) and colonoscopy (0.71 (0.58 - 0.81)) [24]. This is likely a result of the choices made when defining the elements used to calculate the slope (m). Particularly, we attributed a greater burden to false-negative classification and thus would expect the pair to favor sensitivity over specificity, as is shown here. These results further suggest the PMM risk score, at this threshold, can correctly distinguish between participants with and without AN.

We also consider efficiency and predictive value in our evaluation. Specifically, efficiency is the fraction of correct results (true-positive and true-negative) among all results, while predictive results are either positive (PPV), the percentage of correct positive results, or negative (NPV), the percentage of correct negative results. We include these measurements separately as they are to be viewed more as aid in interpreting the test result, rather than a

measure of performance. This is because both efficiency and predictive value also incorporate prevalence, meaning they cannot be considered as inherent properties of the test alone, like sensitivity and specificity, but rather are the results from applying the test in a particular context (decision threshold and disease prevalence) [58]. As a result, it is also important to note this indicates multiple values are possible for each measurement, i.e. holding prevalence constant, a value then exists for each possible decision threshold [58]. Given a prevalence estimate of 6.2% and a decision threshold of 0.02528, the mean value for efficiency was 0.933 (0.002), while PPV was 0.141 (0.009) and NPV was 0.998 (0.0004).

Negative predictive value is not typically reported for FIT-based screening programs as colonoscopy is not completed for participants with a normal FIT result so true negative status is unknown. Similarly, as disease prevalence is low, indicating the number of affected participants is small, we should expect NPV to be high as the risk of being classified as a false negative is low. We focus instead on PPV. While values reported in the literature vary for FIT-based screening, we compare our results to those reported from a retrospective, longitudinal study performed in a fixed cohort of Kaiser Permanente health plan members over four rounds of annual screening; PPV for AN was summarized at 0.114 for the entire study period [21]. Our results suggest a modest improvement in our ability to correctly classify patients with AN based on their PMM risk score. Lastly, we consider efficiency as it allows us to consider all results together rather than only one class at a time, as with predictive value. Of the 6.7% results still misclassified, on average only 0.174% of participants would receive a false-negative result, while 6.526% would still be considered a false-positive result. Even so, this approach would still also yield a 19.87% reduction, on average, in false-positive results relative to the current screening approach. Using the costs defined previously for false-positive and false-negative results, this approach could save approximately $110 per patient screened. These results suggest the PMM approach could be a favorable solution to the colonoscopy capacity issue by reducing the number of unnecessary procedures, in addition to yielding cost savings for the healthcare system.

Chapter 7

Conclusion

This dissertation focuses on statistical methods using longitudinal biomarkers for the early detection of disease. We first apply a risk prediction framework (PMM) to identify cancer among asymptomatic patients. In the context of colorectal cancer screening, this strategy required additional consideration before the modeling approach could be implemented. Specifically, the distribution and range of values for fecal hemoglobin required we test various shift parameters to accommodate data transformation. In doing so, we found the PMM approach to be sensitive to the value used. We then transitioned to an imputation-based approach to accommodate data transformation. PMM continuously outperformed univariate logistic regression, implemented to represent the standard of care. In addition to the longitudinal measures, we implemented a functional data approach to instead leverage the information from the underlying trajectories, adjusting for verification bias under both MAR and MNAR assumptions. However, this approach did not perform as well as PMM or even the standard of care. We anticipate this approach was unsuccessful due to the sparsity of measurements associated with screening. In hindsight, this is unsurprising as one of the original models considered (ROCA), which incorporated the estimation of a changepoint distribution as part of its evaluation of risk, was thought to be unsuccessful because the intervals between two biomarker measurements were too wide and thus would not contain enough measurements around the changepoint. Lastly, we identified and evaluated a decision threshold for clinical implementation. This approach leveraged Receiver-Operating Characteristic (ROC) Plots in combination with the relative cost/undesirability of errors to identify the decision threshold corresponding to the optimal mix of false-positive and false-negative results. We applied these methods to data collected within a large FIT-based CRC screening program, with the motivation of improving early detection success while reducing

the number of colonoscopies needed to identify patients with AN.

Several topics warrant further research. First, a natural next step is to apply the PMM approach used in this thesis within a FIT-based screening program and evaluate clinical outcomes in a randomized control setting in comparison with the standard of care. We will also investigate different methods for determining the optimum binary cut-off threshold to pair with the PMM risk score and how they might be applied in a clinical setting. As an example, information theory offers multiple concepts thought to be useful when evaluating a clinical test. Both mutual information (MI) and information gain (IG) have been proposed as metrics for clinical test value, where MI quantifies the degree to which performing the test can be expected to reduce uncertainty regarding the underlying disease and IG (relative entropy) quantifies the expectation that a specific test result (i.e. positive or negative) will reduce diagnostic uncertainty [33, 49]. IG is of particular interest as the clinical considerations for cancer screening disproportionately weigh certain test results. Lastly, we continue to pursue methods for incorporating longitudinal biomarker measurements for the early detection of disease while accounting for the sparseness of data typically associated with cancer screening.

Appendix A

APPENDIX

x

Table A.1: Time-dependent AUCs and associated 95% bootstrapped confidence intervals (shown in parentheses) for the reproduced PMM-CN3 using the LOOCV and 10-Fold CV method

	LOOCV	10-Fold CV
Metric	PMM-CN3	PMM-CN3
$AUC_{0.5}$	0.971 (0.952, 0.986)	0.969 (0.947, 0.985)
$AUC_{1.0}$	0.947 (0.922, 0.968)	0.947 (0.922, 0.967)
$AUC_{1.5}$	0.889 (0.835, 0.938)	0.895 (0.845, 0.937)
$AUC_{2.0}$	0.867 (0.806, 0.921)	0.872 (0.816, 0.920)
$AUC_{2.5}$	0.867 (0.806, 0.918)	0.871 (0.818, 0.919)
$AUC_{3.0}$	0.859 (0.799, 0.913)	0.863 (0.802, 0.913)

REFERENCES

[1] Albert, P. S. A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statistics in medicine 31*, 2 (2012), 143–154.

[2] Allan, B. What is the positivity rate for colorectal cancer screening by fit? *NEJM Journal Watch* (2014).

[3] American Cancer Society. *Cancer Facts & Figures 2020 - 2022*. Atlanta: American Cancer Society, 2022.

[4] American Cancer Society. *Colorectal Cancer Facts & Figures 2020 - 2022*. Atlanta: American Cancer Society, 2022.

[5] Bakketeig, L. S., Jacobsen, G., Hoffman, H. J., Lindmark, G., Bergsjø, P., Molne, K., and Rødsten, J. Pre-pregnancy risk factors of small-for-gestational age births among parous women in scandinavia. *Acta obstetricia et gynecologica Scandinavica 72*, 4 (1993), 273–279.

[6] Berry, E., Miller, S., Koch, M., Balasubramanian, B., Argenbright, K., and Gupta, S. Lower abnormal fecal immunochemical test cut-off values improve detection of colorectal cancer in system-level screens. *Clinical Gastroenterology and Hepatology 18*, 3 (2020), 647–653.

[7] Buckley, J., and James, I. Linear regression with censored data. *Biometrika 66*, 3 (1979), 429–436.

[8] Centers for Medicare & Medicaid Services. `https://www.cms.gov/medicare/physician-fee-schedule/search`, 04/02/2024.

[9] Charkhchi, P., Cybulski, C., Gronwald, J., Wong, F. O., Narod, S. A., and Akbari, M. R. Ca125 and ovarian cancer: a comprehensive review. *Cancers 12*, 12 (2020), 3730.

[10] Cole, S. R., Chu, H., Nie, L., and Schisterman, E. F. Estimating the odds ratio when exposure has a limit of detection. *International journal of epidemiology 38*, 6 (2009), 1674–1680.

[11] Cramer, D. PLCO Phase III Analysis Files. `https://edrn-labcas.jpl.nasa.gov/labcas-ui/c/index.html?collection_id=PLCO_Phase_III_Dataset` , 08/01/2021.

[12] Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., Donahue, K. E., Doubeni, C. A., Krist, A. H., Kubik, M., et al. Screening for colorectal cancer: Us preventive services task force recommendation statement. *JAMA 325*, 19 (2021), 1965–1977.

[13] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* (1988), 837–845.

[14] Deter, R. L. Individualized growth assessment: evaluation of growth using each fetus as its own control. In *Seminars in perinatology* (2004), vol. 28, Elsevier, pp. 23–32.

[15] Feng, C., Wang, H., Lu, N., and Tu, X. M. Log transformation: application and interpretation in biomedical research. *Statistics in medicine 32*, 2 (2013), 230–239.

[16] Fraser, C. G., Allison, J. E., Halloran, S. P., Young, G. P., and Expert Working Group on Fecal Immunochemical Tests for Hemoglobin, Colorectal Cancer Screening Committee, W. E. O. A proposal to standardize reporting units for fecal immunochemical tests for hemoglobin. *Journal of the National Cancer Institute 104*, 11 (2012), 810–814.

[17] Goede, S. L., Kuntz, K. M., Van Ballegooijen, M., Knudsen, A. B., Lansdorp-Vogelaar, I., Tangka, F. K., Howard, D. H., Chin, J., Zauber, A. G., and Seeff, L. C. Cost-savings to medicare from pre-medicare colorectal cancer screening. *Medical care 53*, 7 (2015), 630–638.

[18] Goede, S. L., Rabeneck, L., van Ballegooijen, M., Zauber, A. G., Paszat, L. F., Hoch, J. S., Yong, J. H., Kroep, S., Tinmouth, J., and Lansdorp-Vogelaar, I. Harms, benefits and costs of fecal immunochemical testing versus guaiac fecal occult blood testing for colorectal cancer screening. *PloS one 12*, 3 (2017), e0172864.

[19] Han, Y., Albert, P. S., Berg, C. D., Wentzensen, N., Katki, H. A., and Liu, D. Statistical approaches using longitudinal biomarkers for disease early detection: A comparison of methodologies. *Statistics in medicine 39*, 29 (2020), 4405–4420.

[20] Horton, N. J., and Kleinman, K. P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician 61*, 1 (2007), 79–90.

[21] Jensen, C. D., Corley, D. A., Quinn, V. P., Doubeni, C. A., Zauber, A. G., Lee, J. K., Zhao, W. K., Marks, A. R., Schottinger, J. E., Ghai, N. R., et al. Fecal immunochemical test program performance over 4 rounds of annual screening: a retrospective cohort study. *Annals of internal medicine 164*, 7 (2016), 456–463.

[22] Jetelina, K. K., Yudkin, J. S., Miller, S., Berry, E., Lieberman, A., Gupta, S., and Balasubramanian, B. A. Patient-reported barriers to completing a diagnostic colonoscopy following abnormal fecal immunochemical test among uninsured patients. *Journal of general internal medicine 34* (2019), 1730–1736.

[23] Knudsen, A. B., Rutter, C. M., Peterse, E. F., Lietz, A. P., Seguin, C. L., Meester, R. G., Perdue, L. A., Lin, J. S., Siegel, R. L., Doria-Rose, V. P., et al. Colorectal cancer screening: an updated modeling study for the us preventive services task force. *JAMA 325*, 19 (2021), 1998–2011.

[24] Lee, J. K., Liles, E. G., Bent, S., Levin, T. R., and Corley, D. A. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. *Annals of internal medicine 160*, 3 (2014), 171–181.

[25] Levi, Z., Rozen, P., Hazazi, R., Vilkin, A., Waked, A., Maoz, E., Birkenfeld, S., Leshno, M., and Niv, Y. A quantitative immunochemical fecal occult blood test for colorectal neoplasia. *Annals of internal medicine 146*, 4 (2007), 244–255.

[26] Li, H., and Gatsonis, C. Combining biomarker trajectories to improve diagnostic accuracy in prospective cohort studies with verification bias. *Statistics in medicine 38*, 11 (2019), 1968–1990.

[27] Liang, P. S., Williams, J. L., Dominitz, J. A., Corley, D. A., and Zauber, A. G. Age-stratified prevalence and predictors of neoplasia among us adults undergoing screening colonoscopy in a national endoscopy registry. *Gastroenterology 163*, 3 (2022), 742–753.

[28] Lieberman, D. A., De Garmo, P. L., Fleischer, D. E., Eisen, G. M., and Helfand, M. Patterns of endoscopy use in the united states. *Gastroenterology 118*, 3 (2000), 619–624.

[29] Little, R., Rubin, D., and Safari, a. O. M. C. *Statistical Analysis with Missing Data., 3rd Edition*. Wiley, 2019.

[30] Liu, D., and Albert, P. S. Combination of longitudinal biomarkers in predicting binary events. *Biostatistics 15*, 4 (2014), 706–718.

[31] McDonald, P. J., Strachan, J. A., Digby, J., Steele, R. J., and Fraser, C. G. Faecal haemoglobin concentrations by gender and age: implications for population-based screening for colorectal cancer. *Clinical chemistry and laboratory medicine 50*, 5 (2012), 935–940.

[32] McIntosh, M. W., and Pepe, M. S. Combining several screening tests: optimality of the risk score. *Biometrics 58*, 3 (2002), 657–664.

[33] Metz, C. E., Goodenough, D. J., and Rossmann, K. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology 109*, 2 (1973), 297–303.

[34] Molenberghs, G., and Kenward, M. *Missing data in clinical studies*. John Wiley & Sons, 2007.

[35] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[36] Raghunathan, T. E. What do we do with missing data? some options for analysis of incomplete data. *Annu. Rev. Public Health 25* (2004), 99–117.

[37] Ramsay, J. O., and Silverman, B. W. *Functional Data Analysis*. Springer, 2005.

[38] Regula, J., Rupinski, M., Kraszewska, E., Polkowski, M., Pachlewski, J., Orlowska, J., Nowacki, M. P., and Butruk, E. Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. *New England Journal of Medicine 355*, 18 (2006), 1863–1872. PMID: 17079760.

[39] Rex, D. K. Colonoscopy: a review of its yield for cancers and adenomas by indication. *American Journal of Gastroenterology (Springer Nature) 90*, 3 (1995).

[40] Robertson, D. J., Lee, J. K., Boland, C. R., Dominitz, J. A., Giardiello, F. M., Johnson, D. A., Kaltenbach, T., Lieberman, D., Levin, T. R., and Rex, D. K. Recommendations on fecal immunochemical testing to screen for colorectal neoplasia: a consensus statement by the us multi-society task force on colorectal cancer. *Gastroenterology 152*, 5 (2017), 1217–1237.

[41] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics 12* (2011), 77.

[42] Rutter, C. Rand corporation (crc-spin), 2018. visited 22-August-2022.

[43] Schisterman, E. F., Vexler, A., Whitcomb, B. W., and Liu, A. The limitations due to exposure detection limits for regression models. *American journal of epidemiology 163*, 4 (2006), 374–383.

[44] Seeff, L. C., Manninen, D. L., Dong, F. B., Chattopadhyay, S. K., Nadel, M. R., Tangka, F. K., and Molinari, N.-A. M. Is there endoscopic capacity to provide colorectal cancer screening to the unscreened population in the united states? *Gastroenterology 127*, 6 (2004), 1661–1669.

[45] Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians 72*, 1 (2022), 7–33.

[46] Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. Cancer statistics, 2023. *Ca Cancer J Clin 73*, 1 (2023), 17–48.

[47] Skates, S. J., Pauler, D. K., and Jacobs, I. J. Screening based on the risk of cancer calculation from bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association 96*, 454 (2001), 429–439.

[48] Slaughter, J. C., Herring, A. H., and Thorp, J. M. A bayesian latent variable mixture model for longitudinal fetal growth. *Biometrics 65*, 4 (2009), 1233–1242.

[49] Somoza, E., and Mossman, D. Comparing and optimizing diagnostic tests: an information-theoretical approach. *Medical Decision Making 12*, 3 (1992), 179–188.

[50] Soret, P., Avalos, M., Wittkop, L., Commenges, D., and Thiébaut, R. Lasso regularization for left-censored gaussian outcome and high-dimensional predictors. *BMC Medical Research Methodology 18* (2018), 1–13.

[51] Van Rossum, L., Van Rijn, A., Laheij, R., Van Oijen, M., Fockens, P., Jansen, J., Verbeek, A., and Dekker, E. Cutoff value determines the performance of a semi-quantitative immunochemical faecal occult blood test in a colorectal cancer screening programme. *British journal of cancer 101*, 8 (2009), 1274–1281.

[52] Wang, Z., and Wang, C.-Y. Buckley-James Boosting for Survival Analysis with High-Dimensional Biomarker Data. *Statistical Applications in Genetics and Molecular Biology 9* (2010), Article 24.

[53] West, R. M. Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry 59*, 3 (2022), 162–165.

[54] Wu, M., and Ware, J. H. On the use of repeated measurements in regression analysis with dichotomous responses. *Biometrics* (1979), 513–521.

[55] Yang, W.-L., Lu, Z., and Bast Jr, R. C. The role of biomarkers in the management of epithelial ovarian cancer. *Expert review of molecular diagnostics 17*, 6 (2017), 577–591.

[56] Yao, F., Müller, H.-G., and Wang, J.-L. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association 100*, 470 (2005), 577–590.

[57] Zhou, Y., Bhattacharjee, S., Carroll, C., Chen, Y., Dai, X., Fan, J., Gajardo, A., Hadjipantelis, P. Z., Han, K., Ji, H., Zhu, C., Müller, H.-G., and Wang, J.-L. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2022. R package version 0.5.9.

[58] Zweig, M. H., and Campbell, G. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry 39*, 4 (1993), 561–577.