Statistical and Machine Learning Methods for Multi-Study Prediction and Causal Inference

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Statistical and Machine Learning Methods for Multi-Study Prediction and Causal Inference"

presented by Cathy Wang

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature ....

Typed name: Prof. Giovanni Parmigiani

Signature Prasad Patil
Prasad Patil (Jun 9, 2022 15:18 EDT)

Typed name: Prof. Prasad Patil

Signature Pragya Sur
Pragya Sur (Jun 9, 2022 16:12 EDT)

Typed name: Prof. Pragya Sur

Signature ....

Typed name:

Date: May 9, 2022

Statistical and Machine Learning Methods for Multi-Study Prediction and Causal Inference

A dissertation presented
by

Cathy Wang

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
May 2022

Dissertation Advisor: Professor Giovanni Parmigiani Cathy Wang

Statistical and Machine Learning Methods for Multi-Study Prediction and Causal Inference

Abstract

In many areas of biomedical research, exponential advances in technology and facilitation of systematic data-sharing increased access to multiple studies. This dissertation proposes and compares methods to address three challenges in multi-study learning. First, personalized cancer risk assessment is key to early prevention, but studies typically report aggregated risk information. We address this challenge by proposing a method that integrates and deconvolves aggregated risk, allowing for heterogeneity in study populations, design, and risk measures, to provide personalized risk estimates that comprehensively reflect the best available data. Second, prediction models are widely used to evaluate disease risk and inform decisions about treatment, but models trained on a single study generally perform worse on out-of-study samples. To address this challenge, we compare two strategies for training prediction models on multiple studies to improve generalizability: merging and ensembling; in practice, our theory can help guide decisions on choosing the ideal strategy. Third, heterogeneous treatment effect estimation is central to personalizing treatment and improving clinical practice, but existing approaches on synthesizing evidence across multiple studies do not account for between-study heterogeneity. We address this challenge by proposing a flexible method that estimates heterogeneous treatment effects from multiple studies, including evidence from randomized controlled trials and real world data, while appropriately accounting for between-study differences in the propensity score and outcome models.

In Chapter 1, we propose a meta-analytic approach for deconvolving aggregated risks to provide age-, gene-, and sex-specific cancer risk. Carriers of pathogenic variants in mismatch repair (MMR) genes benefit from reliable information about their cancer risk to better inform targeted surveillance strategies for colorectal cancer (CRC), but published estimates vary. Variation in published estimates could arise from differences in study designs, selection criteria for molecular testing, and statistical adjustments for ascertainment. Previous meta-analyses of CRC risk are based on studies that report gene- and sex-specific risk. This may exclude studies that provide aggregated cancer risk across sex and genes and lead to bias. To address this challenge, our meta-analytic approach has the ability to deconvolve aggregated risks, allowing us to use all of the information available in the literature and provide more comprehensive penetrance estimates. This method can be applied in the future to other gene/cancer combinations without restriction on the mutation.

In Chapter 2, we compare methods for training gradient boosting models on multiple studies. When training and test studies come from different distributions, prediction models trained on a single study generally perform worse on out-of-study samples due to heterogeneity in study design, data collection methods, and sample characteristics. Training prediction models on multiple studies can address this challenge and improve cross-study replicability of predictions. We focus on two strategies for training cross-study replicable models: 1) merging all studies and training a single model, and 2) multi-study ensembling, which involves training a separate model on each study and combining the resulting predictions. We study boosting algorithms in a regression setting and compare cross-study replicability of merging vs. multi-study ensembling both empirically and theoretically. In particular, we characterize an analytical transition point beyond which ensembling exhibits lower prediction error than merging for boosting with linear learners. We verify the theoretical transition point empirically and illustrate how it may guide practitioners' choice regarding merging vs. ensembling in a breast cancer application.

In Chapter 3, we propose an approach for estimating heterogeneous treatment effects in multiple studies. Heterogeneous treatment effect estimation is central to many modern statistical applications, such as precision medicine. Despite increased access to multiple studies, existing methods on heterogeneous treatment effect estimation are largely rooted in theory based on a single study. These methods generally rely on the assumption that the heterogeneous treatment effect is the same across studies. However, this assumption may be untenable under potential heterogeneity in study design, data collection methods, and sample characteristics across multiple studies. To address this challenge, we propose the multi-study R-learner for estimating heterogeneous treatment effects under the presence of between-study heterogeneity. This method allows information to be borrowed across multiple studies and flexible modeling of the nuisance components with machine learning methods. We show analytically that optimizing the multi-study R-loss is equivalent to optimizing the oracle loss up to an error that diminishes at a relatively fast rate with the sample size. Under the series estimation framework, we derive a pointwise normality result for the multi-study R-learner estimator. Empirically, we show that as between-study heterogeneity increases, the multi-study R-learner results in lower estimation error than the R-learner via simulations and a breast cancer application.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my advisor, Giovanni Parmigiani, for his incredible mentorship, unwavering support, and superb guidance throughout my graduate studies.

My dissertation committee members, Prasad Patil and Pragya Sur, played an integral role in shaping my thesis, and I am grateful for their time and helpful suggestions.

I would also like to thank Danielle Braun for mentoring me and providing instrumental contributions to Chapter 1.

I am very grateful to the members of the BayesMendel Lab and the Multi-Study Learning Lab. In particular, Boyu Ren provided excellent suggestions for Chapter 3.

Last but not least, I would like to express my gratitude to my family and friends. My husband, Derek Shyr, provided endless support and love that helped me tremendously. Thank you for being my bedrock and cheering me on. I am extremely grateful for my parents' and parents'-in-law support, love, and guidance.

Chapter 1

Penetrance of Colorectal Cancer Among Mismatch Repair Gene Mutation Carriers: A Meta-Analysis

Cathy Wang[1, 2], Yan Wang [3, 4], Kevin S. Hughes [4], Giovanni Parmigiani [1, 2], Danielle Braun [1, 2]

[1] *Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA*

[2] *Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA*

[3] *Division of Surgical Oncology, Massachusetts General Hospital, Boston, MA, USA*

[4] *Department of Breast Surgery, Shanghai Cancer Hospital, Fudan University, Shanghai, People's Republic of China*

This work was originally published in the *Journal of National Cancer Institute: Cancer Spectrum* at https://doi.org/10.1093/jncics/pkaa027.

Citation: Wang, C., Wang, Y., Hughes, K.S., Parmigiani, G. and Braun, D., 2020. Penetrance of colorectal cancer among mismatch repair gene mutation carriers: a meta-analysis. JNCI cancer spectrum, 4(5), p.pkaa027.

Abstract

Background: Lynch syndrome, the most common colorectal cancer (CRC) syndrome, is caused by germline mutations in mismatch repair (MMR) genes. Precise estimates of age-specific risks are crucial for sound counseling of individuals managing a genetic predisposition to cancer, but published risk estimates vary. The objective of this work is to provide gene-, sex-, and age-specific risk estimates of CRC for MMR mutation carriers that comprehensively reflect the best available data.

Methods: We conducted a meta-analysis to combine risk information from multiple studies on Lynch-syndrome-associated CRC. We used a likelihood-based approach to integrate reported measures of CRC risk and deconvolved aggregated information to estimate gene- and sex-specific risk.

Results: Our comprehensive search identified 10 studies (8 on MLH1, 9 on MSH2, and 3 on MSH6). We estimated the cumulative risk of CRC by age and sex in heterozygous mutation carriers. At age 70, for males and females respectively, risks for MLH1 are 43.9% (95% CI: 39.6, 46.6) and 37.3% (95% CI: 32.2, 40.2); for MSH2 54% (95% CI: 49, 56.3) and 38.6% (95% CI: 34.1, 42); and for MSH6 12% (95% CI: 2.4, 24.6) and 12.3% (95% CI: 3.5, 23.2).

Conclusion: Our results provide up-to-date and comprehensive age-specific CRC risk estimates for counseling and risk prediction tools. These will have a direct clinical impact by improving prevention and management strategies for both individuals who are MMR mutation carriers and those considering testing.

1.1 Background

Lynch syndrome, also known as hereditary nonpolyposis colorectal cancer syndrome (HNPCC), accounts for approximately 3-5% of all colorectal cancers (CRC) and is an autosomal dominant condition caused by germline pathogenic variants in mismatch repair (MMR) genes (Rustgi (2007); Jass (2006)). Carriers of pathogenic variants in any of the MMR genes; MLH1, MSH2, MSH6, PMS2, or EPCAM have an increased risk of developing several types of cancers, including colorectal, endometrial, stomach, small bowel, and biliary tract cancers (Umar *et al.* (2004)). Lynch syndrome is generally identified following investigation of familial aggregation of multiple and/or early-onset cancers based on the Amsterdam II criteria, NCCN guidelines, Bethesda guidelines (Umar *et al.* (2004); Provenzale *et al.* (2016); Vasen *et al.* (1999)) or more quantitative risk assessment (Kastrinos *et al.* (2018)). More recently, it is also being found incidentally through panel genetic testing and by MSI or IHC testing of all colorectal cancers. In addition, Hampel *et al.* have recently called for sequencing of all colorectal cancers (Hampel *et al.* (2018)).

Carriers of pathogenic variants in MMR genes can benefit from reliable information about their cancer risk to better inform effective management and targeted surveillance strategies. Published estimates of penetrance (age-specific risk of cancer for carriers) vary. Studies typically provide different measures of CRC risk, including cumulative penetrance, relative risks (RR), or standardized incidence ratios (SIR) from family-based studies, and odds ratios (OR) from case-control studies.

The objective of this work is to combine results from published studies to provide more accurate age- and sex-specific penetrance estimates of MLH1, MSH2, and MSH6 on CRC for individuals with Lynch syndrome. Cumulative lifetime penetrance estimates of CRC range from 30% to 74% for MLH1 and MSH2 gene mutation carriers, and from 10% to 22% for MSH6 mutation carriers (Giardiello *et al.* (2014)). Variation in published estimates could arise from differences in study designs, selection criteria for molecular testing, and statistical adjustments for ascertainment (Kraft and Thomas (2000)). Without adjustment, estimated lifetime risk in studies of high-risk families can be higher than that estimated from population-based studies. In sensitivity analyses, studies have shown that different ascertainment schemes can lead to inconsistent risk estimates (Stoffel *et al.* (2009); Mukherjee *et al.* (2011)). To address these concerns, we explicitly consider properly adjusting for ascertainment as an inclusion criterion for our meta-analysis. Previous meta-analyses of CRC risk in individuals with Lynch syndrome are based on studies that report gene- and sex-specific cumulative penetrance estimates (Chen *et al.* (2006); Jenkins *et al.* (2014)). This excludes additional published risk measures from studies that provide aggregated information across sex and genes. In our analysis, we do not make these exclusions, as they may miss important information and may lead to bias.

1.2 Methods

1.2.1 Literature Search

We performed three separate PubMed searches for MLH1, MSH2, and MSH6, with the following queries: **MLH1/Colorectal**: ("MutL Protein Homolog 1"[Mesh] OR "MLH1"[TIAB] OR "Lynch syndrome"[TIAB]) AND ("Risk"[Mesh] OR "Risk"[TI] OR "Penetrance"[TIAB] OR "Hazard ratio"[TIAB]) AND ("Colorectal Neoplasms"[Mesh] OR "Colorectal Neoplasms, Hereditary Nonpolyposis"[Mesh] OR "colorectal cancer"[TIAB]), **MSH2/Colorectal**: ("MutS Homolog 2 Protein"[Mesh] OR "MSH2"[TIAB] OR "Lynch syndrome"[TIAB]) AND ("Risk"[Mesh] OR "Risk"[TI] OR "Penetrance"[TIAB] OR "Hazard ratio"[TIAB]) AND ("Colorectal Neoplasms"[Mesh] OR "Colorectal Neoplasms, Hereditary Nonpolyposis"[Mesh] OR "colorectal cancer"[TIAB]), **MSH6/Colorectal**: ("G-T mismatch-binding protein"[Supplementary Concept] OR "MSH6"[TIAB]) AND ("Risk"[Mesh] OR "Risk"[TI] OR "Penetrance"[TIAB] OR "Hazard ratio"[TIAB]) AND ("Colorectal Neoplasms"[Mesh] OR "Colorectal Neoplasms, Hereditary Nonpolyposis"[Mesh] OR "colorectal cancer"[TIAB]). We performed a similar search in EMBASE with the following query: ('MutL protein homolog 1'/exp OR 'DNA mismatch repair protein MSH2'/exp OR 'protein MutS'/exp OR MLH1:ab,ti OR MSH2:ab,ti OR MSH6:ab,ti OR Lynch:ab,ti)AND('rectum tumor'/exp OR 'colon tumor'/exp OR ((colon

OR rectal OR rectum OR colorectal) NEAR/3 (cancer* OR neoplasm* OR carcinoma* OR tumor* OR tumour*)):ab,ti)AND('risk'/exp OR risk*:ab,ti OR penetrance:ab,ti OR 'hazard ratio':ab,ti).

References from relevant articles and previous meta-analyses were reviewed to identify additional studies that were not captured by the PubMed or EMBASE searches. In selecting articles from those found by the query, we required the following inclusion criteria: studies must (1) report risk (and corresponding 95% confidence interval) of CRC for carriers of germline mutations in MLH1, MSH2, or MSH6, (2) adjust for ascertainment if cohort is not population based or design is not case-control, and (3) include non-overlapping participants with other studies (Fig. 1.1). We excluded studies that focus on patients with polymorphisms and/or CRC as a secondary cancer. We chose not to include the PMS2 gene, though it is also involved in mismatch repair and associated with Lynch syndrome. In a PubMed literature search similar to that performed for our main analysis (for MLH1, MSH2, and MSH6), three studies report the risk of CRC for PMS2 mutation carriers (Win *et al.* (2012); Sanne *et al.* (2014); Guindalini *et al.* (2015)) and only one of these provides disaggregated data for PMS2 (Sanne *et al.* (2014)). PMS2 carriers generally have a later age of onset than their MLH1/MSH2 counterparts, resulting in lower numbers of events for comparable observation years. Moreover, the low sensitivity of clinical criteria and less widespread diagnostic testing for identifying PMS2 carriers (Sjursen *et al.* (2010); Møller *et al.* (2017)) make it challenging to extend our meta-analysis to PMS2 at the present time.

Studies were first assessed based on title and abstract using a natural language processing (NLP) algorithm (Bao *et al.* (2019)). This algorithm uses a support vector machine (SVM), which learns a linear decision rule based on the bag-of-ngrams representation of each title and abstract. At least two reviewers independently examined the study abstracts, and those deemed relevant underwent full-text review. For studies that remained relevant after full-text review, we extracted the following information: first author's last name, year of publication, study population, ascertainment method, number of events, number of carriers, gene type, and relevant risk estimates with corresponding confidence intervals.

1.2.2 Statistical analysis

Common approaches for combining evidence across multiple studies include fixed effects models, which assume an underlying true effect size for all included studies, and random effects models, which allow for the true effect size to vary from study to study. Typically, these approaches cannot be used directly to combine heterogeneous measures of CRC risk that result from different study designs. Marabelli *et al.* Marabelli *et al.* (2016) developed a likelihood-based method allowing meta-analytic integration of different types of cancer risk estimates (e.g. penetrance, RR, SIR, and OR). This method, however, does not address the challenge of combining studies that report gene-aggregated (a combination of two or more MMR genes) or sex-aggregated cancer risks, which are common in the Lynch syndrome literature. The deconvolution of aggregated risk information is crucial for personalized prevention, as male and/or MLH1/MSH2 mutation carriers typically have higher risks of CRC than their female and/or MSH6 counterparts (Stoffel *et al.* (2009); Jenkins *et al.* (2006); Barrow *et al.* (2009); Choi *et al.* (2009); Bonadona *et al.* (2011); Vasen *et al.* (2001)). In this work, we utilize a more general likelihood-based approach that allows the integration of aggregated cancer risks to provide accurate age-, gene-, and sex-specific penetrance of CRC for MMR mutations carriers.

As a preliminary step, we used the Q^2 and I^2 values to explore between-study heterogeneity. A p-value of less than .05 was considered representative of statistically significant heterogeneity. All tests were two-sided and performed using the *meta* (Schwarzer (2007)) package in R (version 3.3) (R Core Team (2017)). To investigate potential publication bias, we created funnel plots and used a two-sided Egger's (Egger *et al.* (1997)) test to assess asymmetry. We then conducted our meta-analyses based on two complementary approaches. In approach (1) we used the DerSimonian and Laird random effects model (DerSimonian and Laird (1986)) (see details in appendix A) to perform separate meta-analyses of cumulative risk by decade of age. We assume the underlying penetrances are heterogeneous, with between-study variance captured by the Δ^2 parameter in DerSimonian and Laird (1986). The DerSimonian and Laird random effects model does not provide a way to handle aggregated estimates and does not lend itself to extrapolation of estimates to older ages, as required in genetic counseling and decision support tools.

To address these issues, in approach (2) we used a likelihood-based approach to obtain penetrance estimates by yearly age. This approach extends the method Marabelli *et al.* (2016), which allows the meta-analytic integration of different risk measures into age-specific penetrance curves, and is described in detail in appendix A. Briefly, we modeled the penetrance in mutation carriers as a probability distribution function characterized by two parameters. We specified the likelihood terms based on the study design and the risk estimates reported, and estimated the parameters by maximizing the likelihood. Penetrance was assumed to follow a log-logistic distribution. The log-logistic distribution was chosen because, among the commonly used parametric distributions, it was the most similar to penetrance curves reported in the literature (Jenkins *et al.* (2006); Kopciuk *et al.* (2009)) and to the trend indicated by the meta-analytic results of the DerSimonian and Laird random effects model in approach (1). Parameter estimates based on the log-logistic distribution are provided in appendix A. In addition, we conducted leave-one-study-out sensitivity analyses to better understand the sources of heterogeneity. We used the *meta* (Schwarzer (2007)) and the *stats4* (R Core Team (2017)) packages in R to perform the DerSimonian and Laird random effects model analysis and the maximum likelihood estimation for the likelihood-based approach, respectively.

We extended the Marabelli *et al.* method to incorporate studies that provide aggregated risk information. For studies that report sex-aggregated risk, we modeled the penetrance function as a weighted average of the male- and female-specific penetrance functions, which can be estimated separately as long as we have at least some studies that provide sex-specific risk. Weights correspond to the proportion of male or female carriers in the study. Similarly, for studies that report gene-aggregated risk, we modeled the penetrance as a weighted average based on the proportion of different carriers in the study. By allowing studies that report aggregated risk estimates to borrow information from those that report gene- or sex-specific risk estimates, this likelihood-based method combines both direct (gene/sex-specific) and indirect (aggregated) evidence from the literature to provide comprehensive risk estimates of CRC.

Studies typically report risk estimates for carriers who are less than 80 years old. Penetrance estimates from ages 81 to 110 were obtained by multiplying the risk of non-carriers at each age by the risk ratio comparing the risk of carriers to that of non-carriers at age 80 (RR), i.e. $RR = \frac{\text{Carrier penetrance estimate at age 80 from likelihood approach}}{\text{Non-carrier penetrance estimate at age 80 from SEER}}$. We obtained the risk of CRC for non-carriers from the Surveillance, Epidemiology, and End Results Program database (SEER) (SEE), which provides the combined risk of CRC for carriers and non-carriers. As mutations are sufficiently rare, we assume that the general population risk provided by SEER approximates the CRC risk for non-carriers (de la Chapelle (2005)).

1.3 Results

Overall, our searches resulted in 4759 abstracts as of March 8, 2019. Among the 4759 abstracts, 586 were deemed relevant by the natural language processing (NLP) algorithm. After human review, 576 were excluded due to the following criteria: unclear/inappropriate ascertainment adjustment (23), not relevant for MMR or CRC (129), overlap with included studies (16), reports penetrance modified by other risk factors (10), second cancer (50), missing full-text (7), non-pathogenicity (2), polymorphisms (74), not relevant for penetrance (265). For our final meta-analysis, we included 10 studies (Fig. 1.1). Table 1.1 shows a synopsis of the included studies, along with a description of the study design, ascertainment mechanism, and risk estimation methods. Studies vary in terms of population, ascertainment, and design. Among the studies, one reported aggregated risk for sex, three reported aggregated risk for the MMR genes, and one reported both sex- and gene-aggregated risk. Eight studies reported risk for MLH1 carriers, nine reported risk for MSH2 carriers, and three reported risk for MSH6 carriers.

To quantify the between-study variation, we performed tests of heterogeneity and calculated the corresponding I^2 values. With three genes, six age intervals (ages 30, 40, ..., 80), and two sexes, a total of 36 tests were performed. For MLH1, the p-values were <0.001 at ages 40-70 for both sexes. The corresponding I^2 values ranged from 83.3% (68.5, 91.1) to

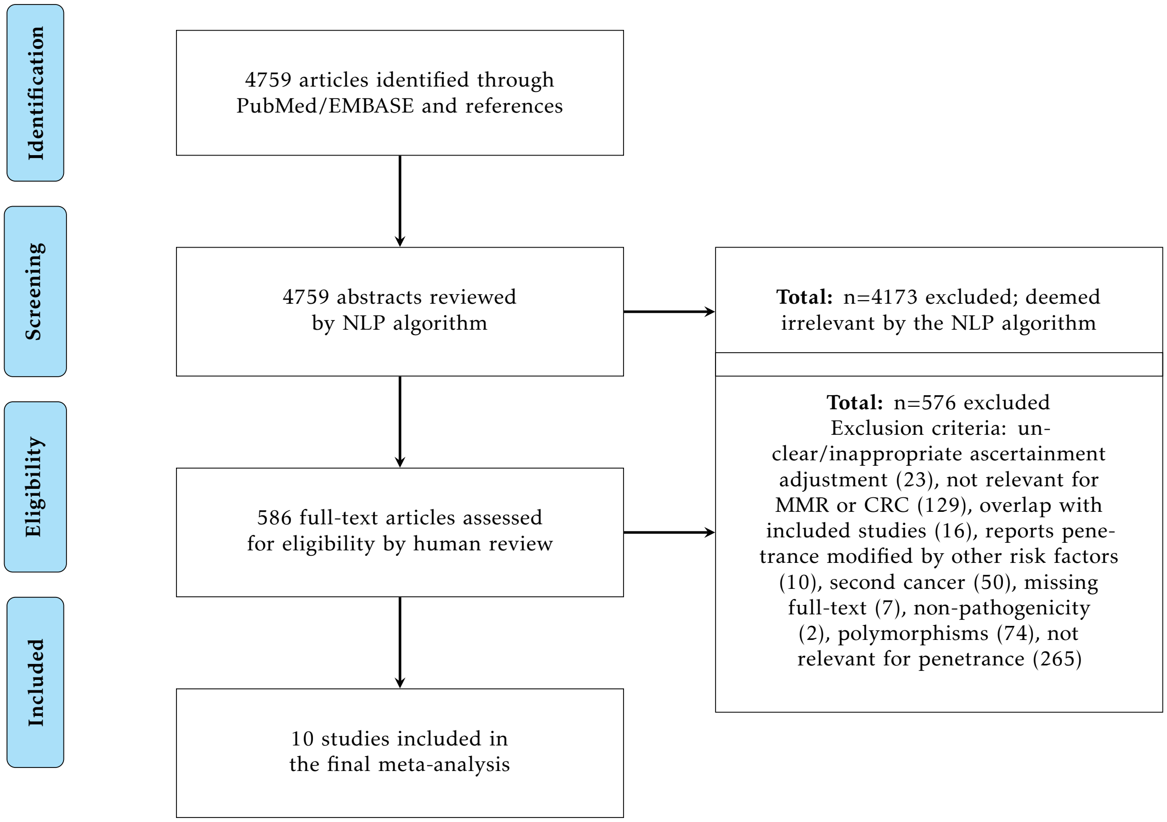

Identification
Screening
Eligibility
Included
4759 articles identified through PubMed/EMBASE and references
4759 abstracts reviewed by NLP algorithm
586 full-text articles assessed for eligibility by human review
10 studies included in the final meta-analysis
Total: n=4173 excluded; deemed irrelevant by the NLP algorithm
Total: n=576 excluded
Exclusion criteria: unclear/inappropriate ascertainment adjustment (23), not relevant for MMR or CRC (129), overlap with included studies (16), reports penetrance modified by other risk factors (10), second cancer (50), missing full-text (7), non-pathogenicity (2), polymorphisms (74), not relevant for penetrance (265)

Figure (1.1) *PRISMA flow diagram of the literature review for our meta-analysis*

Study	Population	Ascertainment	Estimation	No. of Events	No. of Carriers	Gene(s)	Condition for unbiasedness
Aaltonen *et al.*	Regional hospitals, Finland	FD relatives of CRC cases	Kaplan-Meier analysis where relatives were censored at ascertainment, emigration, or last contact with the proband	91	242	MLH1, MSH2	No additional familial aggregation other than MLH1/MSH2
Bonadona *et al.*	ERISCAM study France	Relatives of CRC cases identified from cancer genetics clinics and mutated for MMR genes	Genotype restricted likelihood conditioning on the phenotypes of all relatives and genotype of the proband	768	1633	MLH1, MSH2, MSH6	No additional familial aggregation other than MLH1/MSH2/MSH6
Borras *et al.*	Genetic counseling clinic Spain	Relatives of CRC cases with MMR mutation	Modified segregation analysis conditioning on the genotype and phenotype of the proband and phenotype of all relatives	28	180	MLH1	No additional familial aggregation other than MLH1
Dowty *et al.*	CCFR	FD and SD, or all relatives of cases with MMR mutation, for population- and clinic-based families, respectively	Modified segregation analysis conditioning on the genotype and phenotype of the proband, or on the genotype of the proband and phenotypes of all relatives, for population- and clinic-based families, respectively	1112	2253	MLH1, MSH2	No additional familial aggregation other than MLH1/MSH2
Dunlop *et al.*	SNCR, Scotland	Relatives of early-onset CRC cases identified from population-based registries and mutated for MMR genes	Kaplan-Meier analysis excluding probands	25	67	MLH1, MSH2	No effect from size-based sampling; or, risks to patient carrier cases and relatives are no higher than carrier non-patient cases
Kopciuk *et al.*	Medical Genetics Clinic Canada	Multiple-case families with MMR mutation	Modified segregation analysis conditioning on the phenotypes of all FDR	101	145	MSH2	No additional familial aggregation other than MSH2
Moller *et al.*	Prospective multi center database by Europe "Majorica group"	Mutation carriers with increased risk of CRC identified by each center	Cumulative incidence rate excluding individuals with prior cancer	711	1942	MLH1, MSH2, MSH6	None
Mukherjee *et al.*	MECC, CHS	All participants, or carrier families with history of LS, identified from population study and cancer clinics, respectively	Modified segregation analysis conditioning on the genotype and phenotype of the proband, or on the genotype and phenotype of the proband and the phenotype of affected FD relatives	74	88	MSH2	No additional familial aggregation other than MSH2
Quehenberger *et al.*	Dutch HNPCC family registry	Multiple-case families with MMR mutation	Modified segregation analysis conditioning on observed phenotypes and on the event that at least one case in the family was a carrier	104	397	MLH1, MSH2	No additional familial aggregation other than MLH1/MSH2
Stoffel *et al.*	DFCI, UMichigan	Multiple-case families with MMR mutation	Modified segregation analysis conditioning on the genotype and phenotype of the proband and phenotype of all relatives	99	307	MLH1, MSH2, MSH6	No additional familial aggregation other than MLH1/MSH2/MSH6

Table (1.1) *Summary of studies included in our meta-analysis*

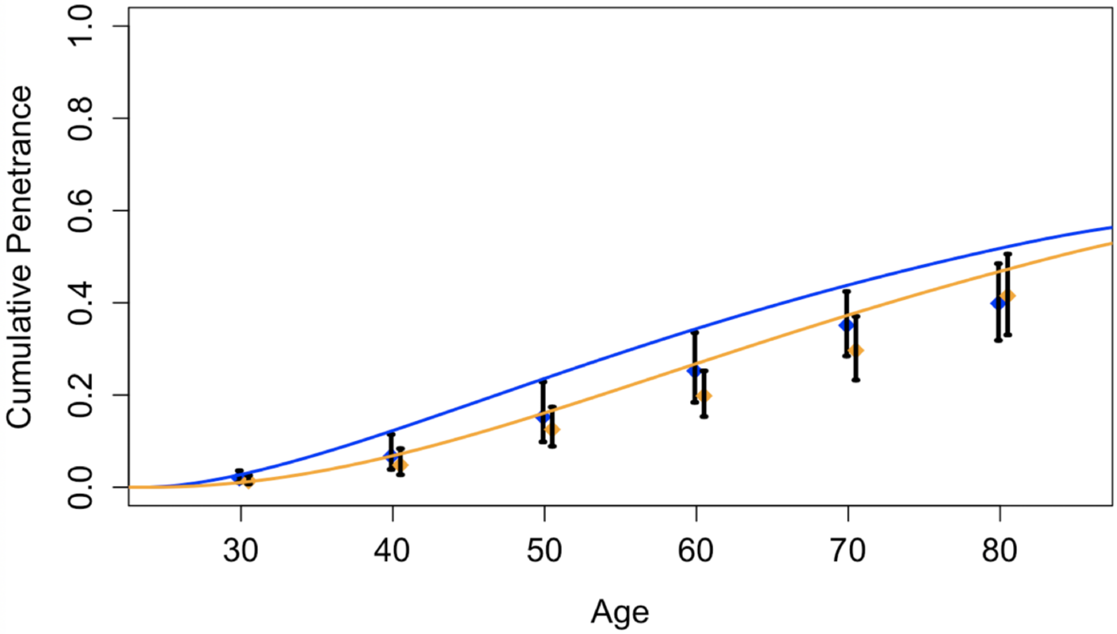
90.3% (83.2%, 94.4%) for males and from 78.4% (57.5%, 89%) to 86.6% (75.7%, 92.6%) for females. The p-values at age 80 for males and females, respectively, were 0.005 (I^2 : 83.0% (56.5, 93.3)) and 0.002 (I^2 : 84.8% (62.0, 93.9)). For MSH2 male carriers, the p-values were <0.0001 at all age intervals with corresponding I^2 values ranging from 89.2% (81.7, 93.6) to 94.1% (89.8, 96.6). For MSH2 female carriers, the p-value was 0.0415 (I^2: 50.2% (0.0, 76.7)) at age 50 and <0.0001 at ages 60 (I^2: 76.0% (54.0, 87.5)) and 70 (I^2: 77.4% (57.1, 88.1)). For MSH6, the only significant p-value at the 0.05 level was that of female mutation carriers at age 70 (p-value = 0.0423, I^2: 68.4% (0.0, 90.8)). Overall, there is evidence for heterogeneity in the risk estimates across the decades for MLH1 and MSH2 mutation carriers but less so for MSH6. Results from tests of asymmetry in the funnel plots suggest that there is little evidence of publication bias. Details on publication bias assessment can be found in appendix A.

Next, we examined sources of heterogeneity from various aspects of study characteristics. This between-study heterogeneity could arise due to differences in study design, mutation type, study population, and estimation strategy. Among the 10 included studies, Moller *et al.* (Møller *et al.* (2017)) was the only study that conducted a prospective cohort analysis, whereas the rest focused on retrospective cohorts. Regarding mutation type, Borras *et al.* (Borràs *et al.* (2010)), Kopciuk *et al.* (Kopciuk *et al.* (2009)), and Mukherjee *et al.* (Mukherjee *et al.* (2011)) are studies that exclusively focused on founder mutations. All other studies included carriers of mixed mutation types, so it was not feasible to separate the effects of mutations from these studies at the present time. As a result, the findings from our meta-analysis represent the average risk among a group of carriers with a representative mix of mutations. Regarding study populations, it is likely that different populations may segregate different mutations. Though there are studies containing more than one subpopulation (Møller *et al.* (2017); Baglietto *et al.* (2010); Dowty *et al.* (2013)), they provide limited evidence of population-specific variation in penetrance. As shown in Table 1, each study used an analysis method that addressed ascertainment mechanism in its design. Studies that were not population-based (Stoffel *et al.* (2009); Mukherjee *et al.* (2011); Bonadona *et al.* (2011); Kopciuk *et al.*

(2009); Borràs *et al.* (2010); Dowty *et al.* (2013); Aaltonen *et al.* (2007); Quehenberger *et al.* (2005)) typically used estimation strategies that condition on information of the phenotype or genotype of included individuals to adjust for ascertainment.

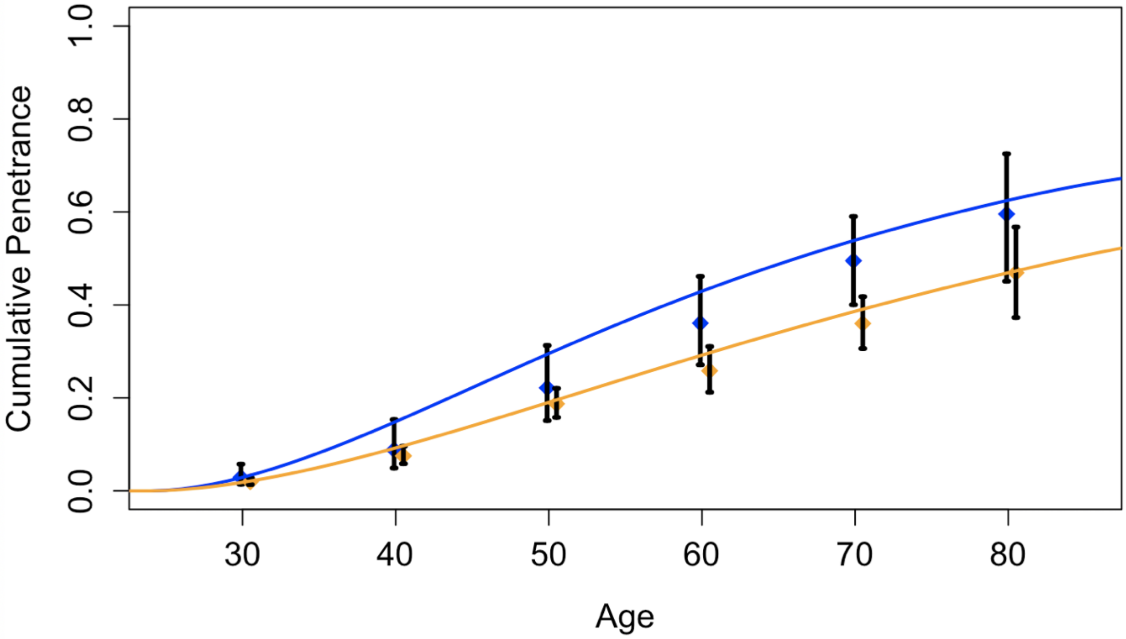
Fig. 1.2 shows the following: (1) the means and 95% CI of the meta-analytic penetrances at each 10-year age interval that were estimated using the DerSimonian and Laird method; (2) the smoothed curves obtained from the likelihood-based approach that represent our final estimates by yearly age. Estimated cumulative penetrance by age 70 from both approaches are displayed in Table 1.2 by sex and gene. Using the likelihood based approach, the penetrances by age 70 were estimated, for males and females respectively, to be 43.9% (95% CI: 39.6, 46.6) and 37.3% (95% CI: 32.2, 40.2) for MLH1 carriers 54% (95% CI: 49, 56.3) and 38.6% (95% CI: 34.1, 42) for MSH2 carriers, and 12% (95% CI: 2.44, 24.6) and 12.3% (95% CI: 3.5, 23.2) for MSH6 carriers. In general, male carriers of MLH1 and MSH2 have higher risk of developing colorectal cancer compared to their female counterparts. Estimates of MSH6 penetrance on CRC show increased variability (wider CIs) due to smaller sample sizes. Visual comparison of the CIs within each 10-year age interval indicates overlap across studies for all three genes. Because all studies reported cumulative penetrance, we were able to include the same studies (8 on MLH1, 9 on MSH2, and 3 on MSH6) for both the DerSimonian and Laird and the likelihood-based approaches. In addition to Fig. 1.2, Fig. A.1 in appendix A shows the study-specific estimates and 95% CI by decade of age.

Among the ten studies, four focused on individuals who have not been screened or have not had prior surgery by censoring participants at the age of colonoscopy screening or prophylactic surgery (Bonadona *et al.* (2011); Kopciuk *et al.* (2009); Dowty *et al.* (2013); Quehenberger *et al.* (2005)). For the remainder of the studies, it was unclear whether screened individuals were included. While screening and surgery were not part of the recruitment criteria, it is reasonable to assume that a number of participants from these six studies (Stoffel *et al.* (2009); Mukherjee *et al.* (2011); Møller *et al.* (2017); Borràs *et al.* (2010); Aaltonen *et al.* (2007); Dunlop *et al.* (1997)) may have undergone screening/surgery according to current screening recommendations (Vasen *et al.* (2007)). We divided the studies into two groups:



# of events	34	259	579	916	1244	913
# of studies	7	8	8	8	8	4

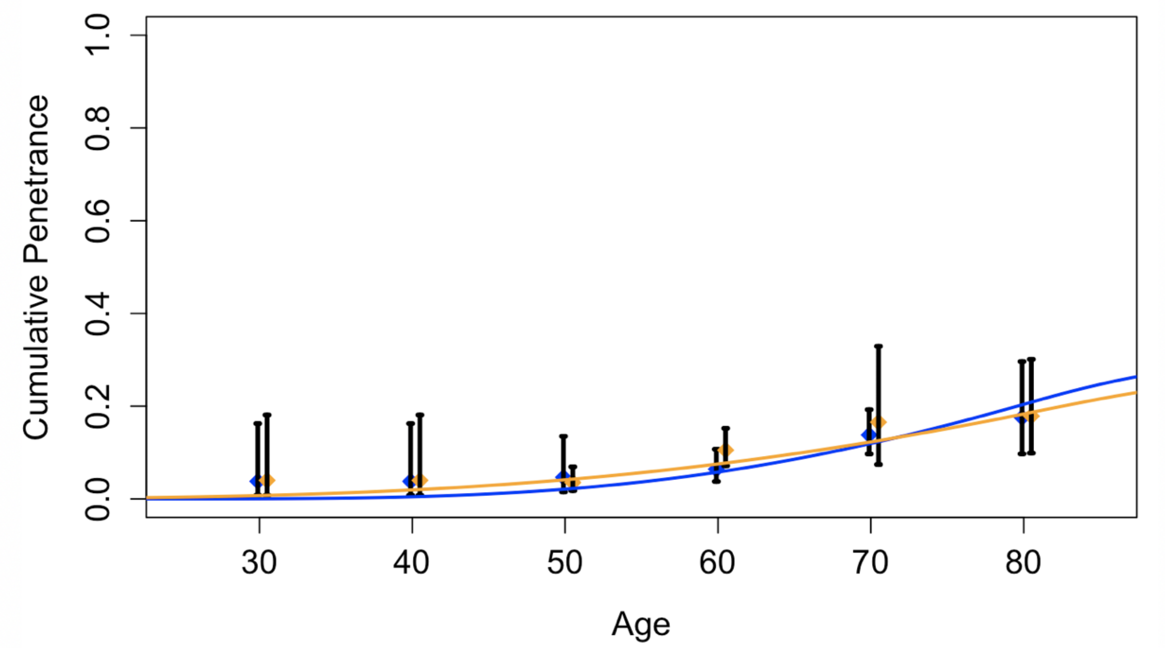
(**a**) *MLH1*



# of events	62	285	712	1106	1451	1334
# of studies	8	9	9	9	9	6

(**b**) *MSH2*

(Continued)



# of events	4	4	6	37	79	20
# of studies	2	2	2	3	3	1

(c) *MSH6*

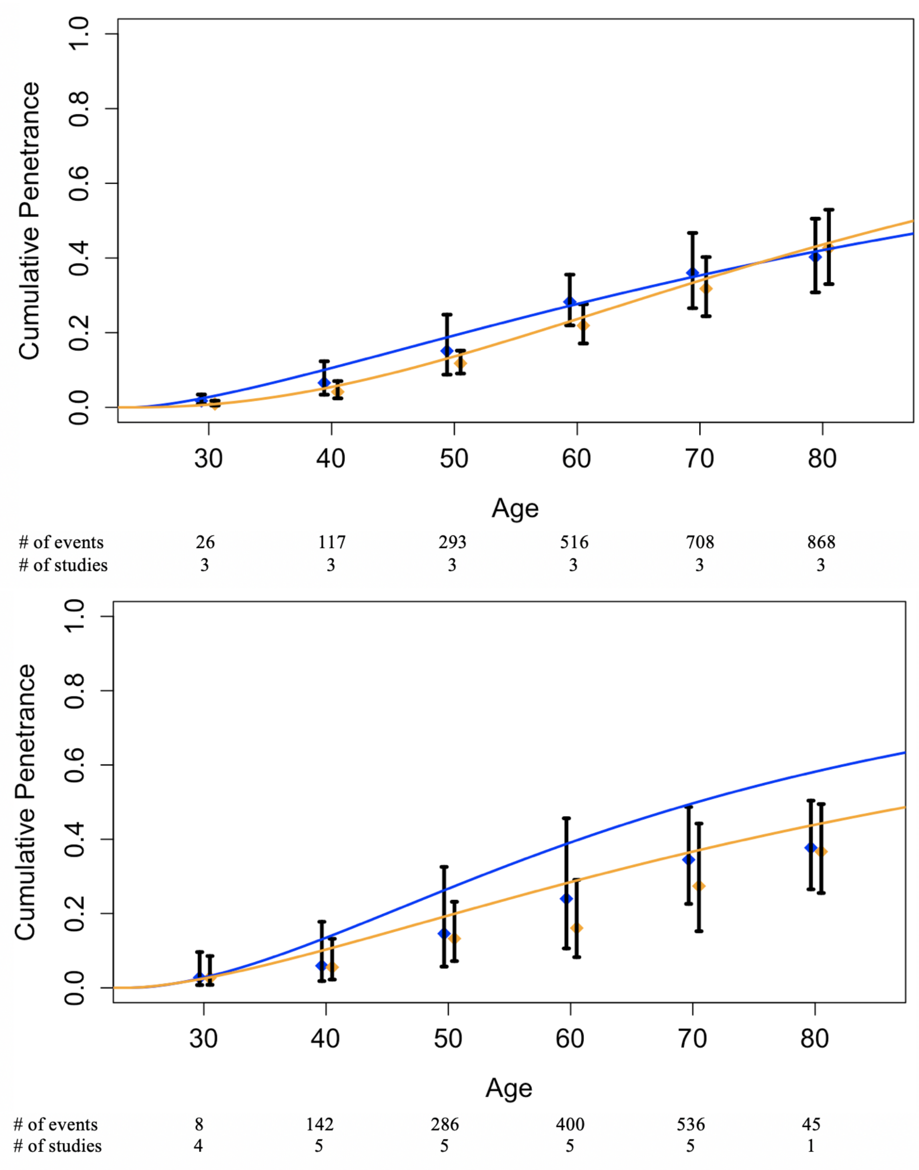
Figure (1.2) *Age-specific colorectal cancer risk for mismatch repair gene mutation carriers*

Panels (a), (b), and (c) correspond to MLH1, MSH2, and MSH6 mutation carriers, respectively. ***DerSimonian and Laird random effects model results:*** *The age range is divided into 10-year intervals. Within each we show the meta-analytic estimate from the DerSimonian and Laird random effects model (thick vertical black bars). The height of vertical bars represents 95% CIs.* ***Likelihood-based approach results:*** *Smooth blue and orange lines represent penetrance estimated from the likelihood-based approach by yearly age. Blue corresponds to male carriers, and orange corresponds to female carriers.*

Sex	Gene	Method	Study Population	Cum. Penetrance (%) with 95% CI
Males	MLH1	DerSimonian and Laird	All	35.1 (28.5, 42.4)
			Unscreened	36 (26.6, 46.7)
			Unspecified	34.5 (22.6, 48.7)
		Likelihood-based	All	43.9 (39.6, 46.6)
			Unscreened	35.3 (29.4, 40)
			Unspecified	50 (43.3, 54.2)
	MSH2	DerSimonian and Laird	All	49.6 (40, 59)
			Unscreened	51.8 (36.4, 66.9)
			Unspecified	47.3 (35.7, 59.1)
		Likelihood-based	All	54 (49, 56.3)
			Unscreened	53.2 (47.1, 57.4)
			Unspecified	57 (49.2, 62.3)
	MSH6	DerSimonian and Laird	All	13.8 (9.7, 19.3)
			Unscreened	14 (7.18, 25.6)
			Unspecified	13.7 (9.01, 20.3)
		Likelihood-based	All	12 (2.4, 24.6)
			Unscreened	19.2 (5.06, 32.8)
			Unspecified	13.2 (0.6, 76.2)
Females	MLH1	DerSimonian and Laird	All	29.7 (23.2, 37.1)
			Unscreened	31.8 (24.4, 40.2)
			Unspecified	27.4 (15.2, 44.2)
		Likelihood-based	All	37.3 (32.2, 40.2)
			Unscreened	34 (27.1, 39.4)
			Unspecified	36.7 (29.6, 42.4)
	MSH2	DerSimonian and Laird	All	36 (30.6, 41.8)
			Unscreened	34.6 (26.9, 43.2)
			Unspecified	37.5 (28.8, 47.2)
		Likelihood-based	All	38.6 (34.1, 42)
			Unscreened	37.3 (33, 40.6)
			Unspecified	41 (34.4, 46.3)
	MSH6	DerSimonian and Laird	All	16.6 (7.4, 32.9)
			Unscreened	10.7 (4.89, 21.9)
			Unspecified	22.3 (10.5, 41.2)
		Likelihood-based	All	12.3 (3.5, 23.2)
			Unscreened	5.3 (0.002, 16.5)
			Unspecified	29.6 (2.5, 79.5)

Table (1.2) *Estimated cumulative penetrance (%) by age 70 of colorectal cancer*

(1) studies that focus on unscreened populations (Bonadona *et al.* (2011); Kopciuk *et al.* (2009); Dowty *et al.* (2013); Quehenberger *et al.* (2005)) and (2) studies that do not provide details on screening and therefore are assumed to be a mix of screened and unscreened populations (Stoffel *et al.* (2009); Mukherjee *et al.* (2011); Møller *et al.* (2017); Borràs *et al.* (2010); Aaltonen *et al.* (2007); Dunlop *et al.* (1997)). Fig. 1.3 shows the cumulative penetrance of CRC for MLH1, MSH2, and MSH6 mutation carriers, after stratifying studies by screening status. Estimated cumulative penetrance by age 70 from both the DerSimonian and Laird and likelihood-based approach are displayed in Table 1.2 by sex, gene, and screening status. For the four studies that included unscreened participants, the penetrance by age 70 was estimated, for males and females respectively, to be 35.3% (95% CI: 29.4, 40) and 34% (95% CI: 27.1, 39.4) for MLH1 carriers, 53.2% (95% CI: 47.1, 57.4) and 37.3% (95% CI: 33, 40.6) for MSH2 carriers, and 19.2% (95% CI: 5.06, 32.8) and 5.3% (95% CI: 0.002, 16.5) for MSH6 carriers. For the six studies that potentially included both screened and unscreened participants (unspecified), the penetrance by age 70 was estimated, for males and females respectively, to be 50% (95% CI: 43.3, 54.2) and 36.7% (95% CI: 30, 42.4) for MLH1 carriers, 57% (95% CI: 49.2, 62.3) and 41% (95% CI: 34.4, 46.3) for MSH2 carriers, and 13.2% (95% CI: 0.6, 76.2) and 29.6% (95% CI: 2.5, 79.5) for MSH6 carriers (Fig. 1.3). Studies on unscreened populations report lower cumulative risk for MLH1 and female MSH6 mutation carriers compared to studies on both screened and unscreened populations. However, the converse is true for male MSH6 mutation carriers. Among the MSH6 studies that report CRC risk in both screened and unscreened populations, Stoffel *et al.* (Stoffel *et al.* (2009)) made conservative ascertainment adjustments, which could lead to lower risk estimates. While differences in CRC risk between the cohorts appear to be more pronounced for MSH6 mutation carriers, this could be attributed to the lack of studies in the unscreened group at age 80. Overall, there is considerable overlap in the 95% CIs across all three genes and both sexes, indicating insufficient evidence to substantiate differences in CRC risk between unspecified (likely a mix of screened and unscreened) and unscreened populations. In addition to Fig. 1.3, Fig. A.2 in appendix A shows the study-specific estimates and 95% CI by decade of age.

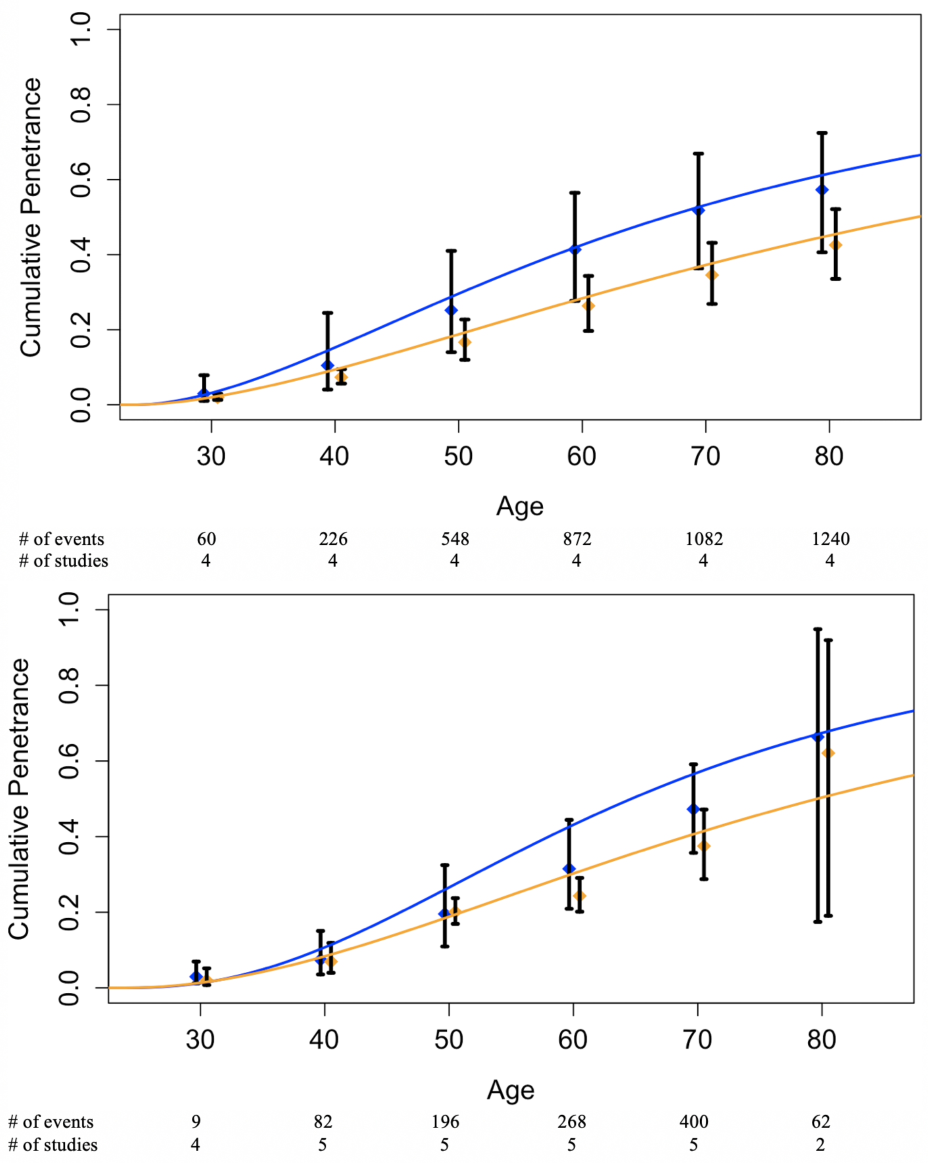


# of events	26	117	293	516	708	868
# of studies	3	3	3	3	3	3

# of events	8	142	286	400	536	45
# of studies	4	5	5	5	5	1

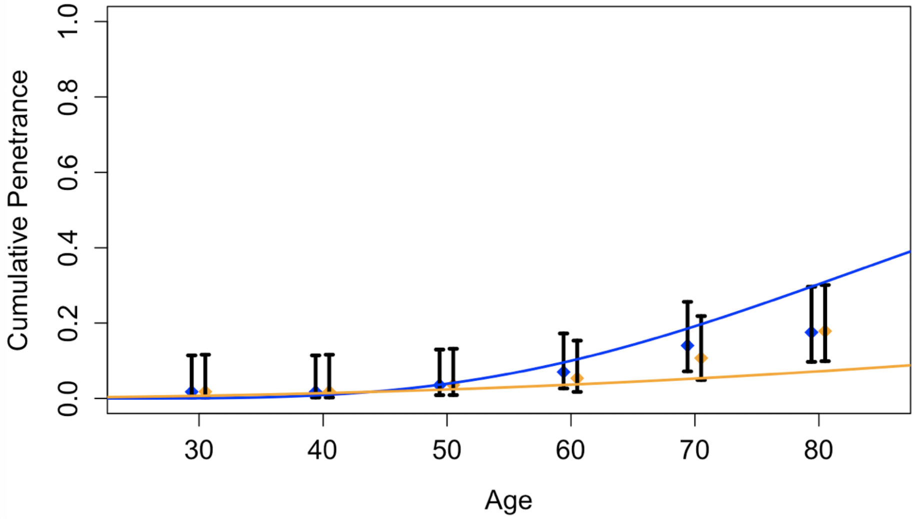
(a) *MLH1*

(Continued)

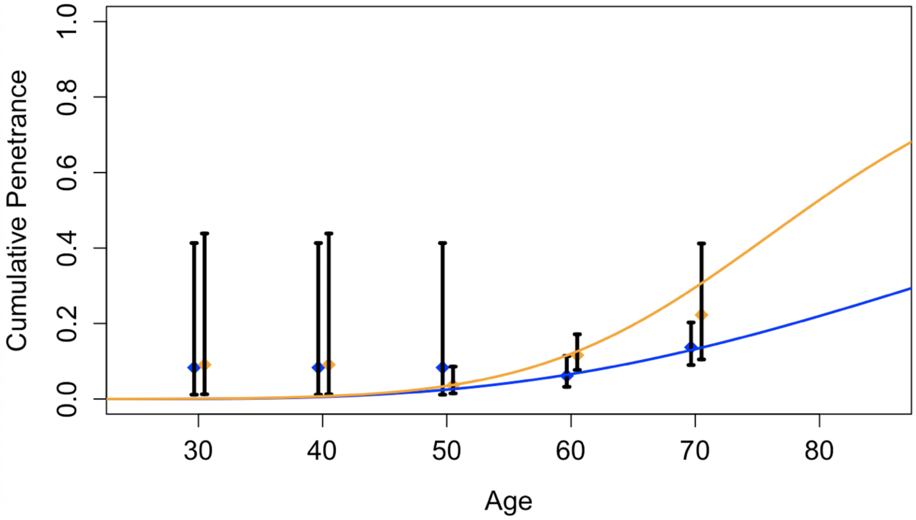


(**b**) *MSH2*

(Continued)



# of events	2	2	4	7	14	20
# of studies	1	1	1	1	1	1



# of events	2	2	7	30	65	0
# of studies	1	1	2	2	2	0

(c) *MSH6*

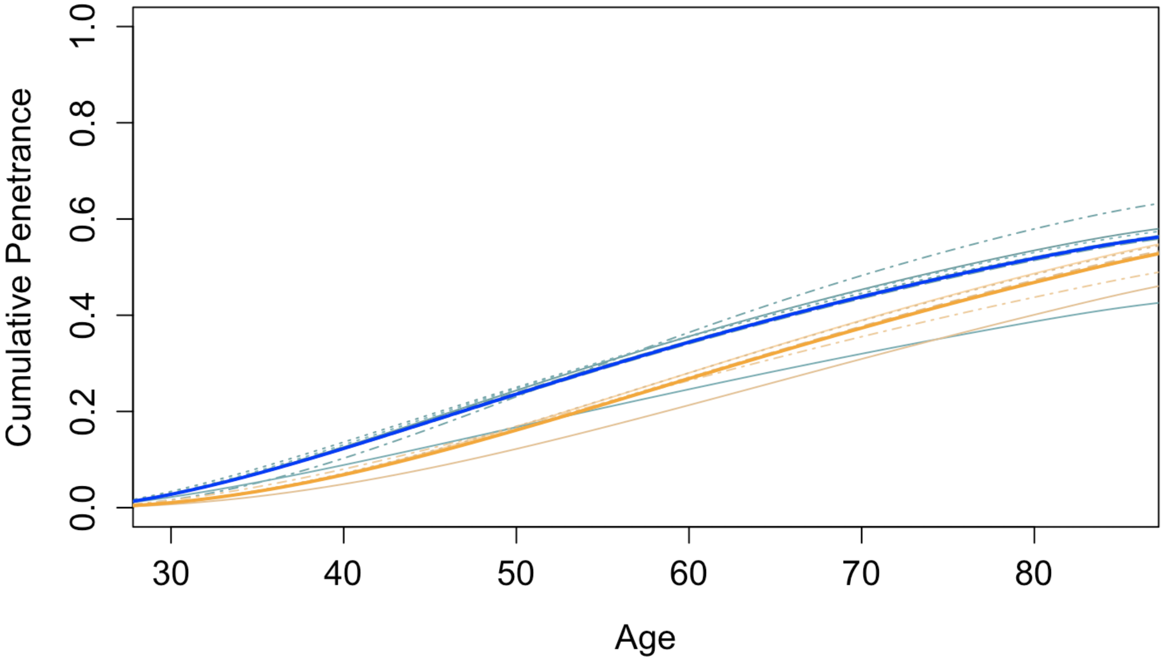
Figure (1.3) *Colorectal cancer risk stratified by screening status*

Colorectal cancer risk stratified by studies on unscreened/no prior surgery population (top panel) or unspecificed (i.e. likely a mix of screened and unscreened populations) (bottom panel) for (a) MLH1 carriers, (b) MSH2 carriers, and (c) MSH6 carriers. ***DerSimonian and Laird random effects model results:*** *The age range is divided into 10-year intervals. Within each we show the meta-analytic estimate from the DerSimonian and Laird random effects model (thick vertical black bars). The height of vertical bars represents 95% CIs.* ***Likelihood-based approach results:*** *Smooth blue and orange lines represent penetrance estimated from the likelihood-based approach by yearly age. Blue corresponds to male carriers, and orange corresponds to female carriers.*

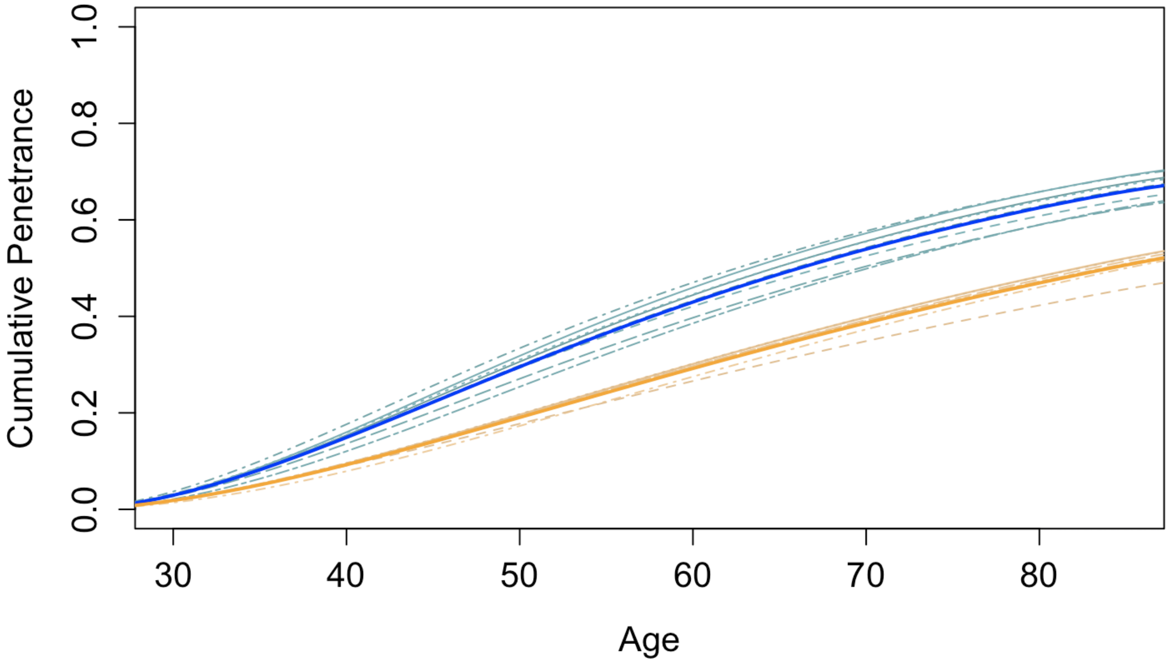
Next, we conducted sensitivity analysis by design/analysis strategy, study population, and mutation type. Mukherjee *et al.* (Mukherjee *et al.* (2011)) focused on founder mutations in MSH2 for individuals of Ashkenazi Jewish descent. Because previous evidence shows that there is an increased risk of CRC in Ashkenazi Jews (Locker and Lynch (2004)), we conducted our meta-analysis with and without this study. Removal of Mukherjee *et al.* had little effect on the combined penetrance estimates for MSH2 mutation carriers. Similarly, we conducted a systematic leave-one-study-out sensitivity analysis and concluded that the meta-analytic results of MLH1 and MSH2 mutation carriers are quite robust to leave-one-study-out sensitivity analysis (Fig. 1.4). Estimated penetrance for female MSH6 mutation carriers is sensitive to the removal of studies by Bonadona *et al.* (Bonadona *et al.* (2011)) and Moller *et al.* (Møller *et al.* (2017)). Penetrance for male MSH6 mutation carriers is sensitive to the removal of Moller *et al* (Møller *et al.* (2017)). Because these two studies were weighted more heavily in the analysis due to their sample sizes, it is not surprising that removing one would affect the risk estimates. This variation in penetrance estimates for MSH6 carriers can be attributed to the smaller sample size (both in number of included studies and in number of mutation carriers) compared to their MLH1/MSH2 counterparts. Moreover, because MSH6 mutation carriers tend to have a later age of onset, the risk information reported by studies is limited to ages 50 and above. Among the three studies that report sex-specific risk for MSH6 mutation carriers, two indicate that female risks are associated with more variability than male risks (Stoffel *et al.* (2009); Møller *et al.* (2017)), resulting in more variable maximum likelihood estimates for the female carriers. Overall, the meta-analytic risk estimates for MLH1/MSH2 carriers are robust to the removal of studies, whereas the estimates for MSH6 are more easily affected due to the smaller number of available studies.

1.4 Discussion

We performed a systematic review of the risk of CRC in mutation carriers of MLH1, MSH2, and MSH6, and combined evidence from 10 studies to provide age-, gene-, and sex-specific risk estimates. These comprehensively reflect the best available data. We conclude that the

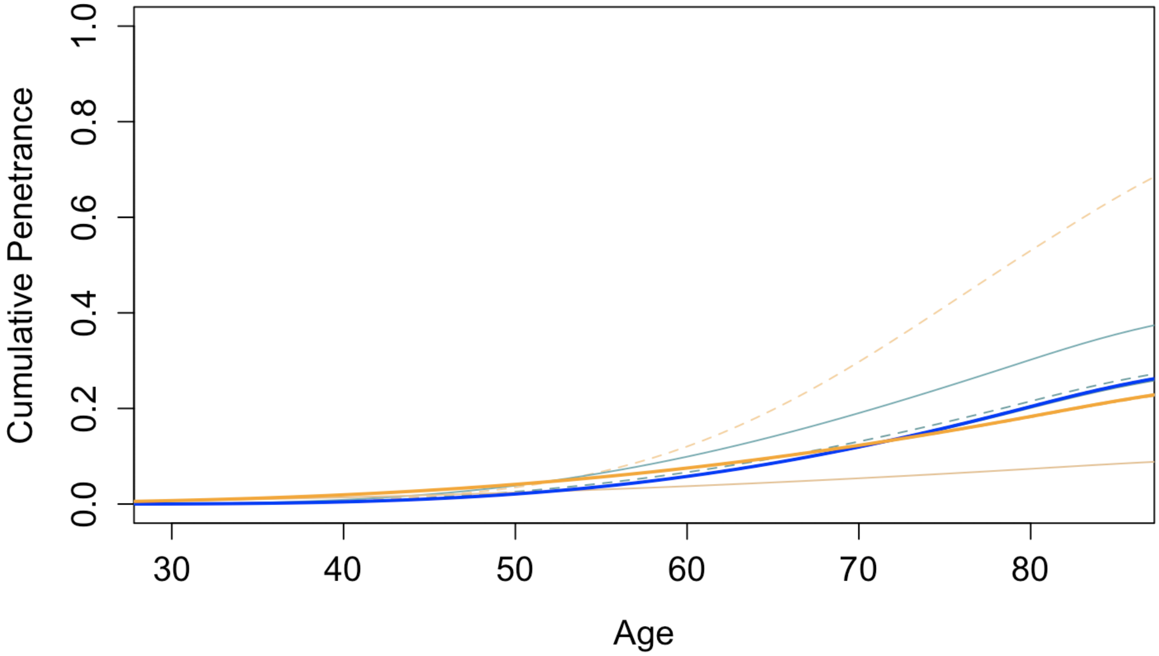

Cumulative Penetrance
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Age

(**a**) *MLH1*


Cumulative Penetrance
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Age

(**b**) *MSH2*

(Continued)



(c) *MSH6*

Figure (1.4) *Leave-one-study-out sensitivity analysis*

Panels (a), (b), and (c) correspond to MLH1, MSH2, and MSH6 mutation carriers, respectively. Bolded solid lines: Cumulative penetrance estimates of CRC based on our likelihood-based approach. Dashed lines: Cumulative penetrance estimates by yearly age of CRC from leave-one-study-out tests of sensitivity. Blue corresponds to male carriers, and orange corresponds to female carriers. Visually, small deviation of a dashed line from the solid line suggests our meta-analysis is robust to the removal of that study.

lifetime cumulative penetrance to age 70 of CRC for males and female carriers, respectively, are 43.9% (95% CI: 39.6, 46.6) and 37.3% (95% CI: 32.2, 40.2) for MLH1 carriers, 54% (95% CI: 49, 56.3) and 38.6% (95% CI: 34.1, 42) for MSH2 carriers, and 12% (95% CI: 2.4, 24.6) and 12.3% (95% CI: 3.5, 23.2) for MSH6 carriers. The smaller number of MSH6 mutation carriers in our analysis led to less certain estimates for that gene, especially at younger ages. Interestingly, more recent studies tend to have narrower CIs, suggesting increased precision in their penetrance estimates. While more conservative ascertainment adjustment mechanisms in recent studies are at play, it is difficult to establish whether those may impact the study estimates or the CIs. The narrower CIs may be attributed to carrier sample size, as recent studies including Bonadona *et al.* (Bonadona *et al.* (2011)), Dowty *et al.* (Dowty *et al.* (2013)) and Moller *et al.* (Møller *et al.* (2017)) have the three largest carrier sample sizes among the included studies.

The differences in the penetrance estimates between the DerSimonian and Laird random effects model and our likelihood-based approach could be attributed to the parametric assumption of the likelihood-based approach. Overall, because the majority of the likelihood-based estimates fall within the meta-analytic 95% CI of the random effects model, we conclude that our findings are likely to be robust to the choice of statistical approach.

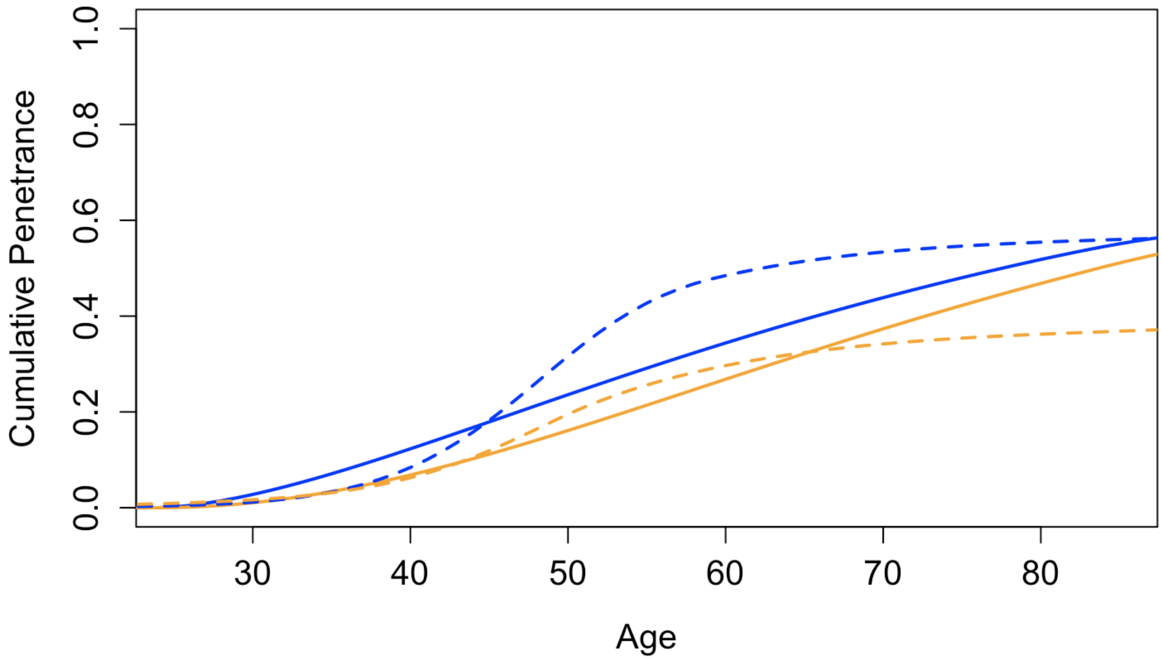
To the best of our knowledge, this meta-analysis is the first to provide age-, gene-, and sex-specific penetrance estimates of MLH1, MSH2, and MSH6 mutations for CRC. A previous meta-analysis by Jenkins *et al.* (Jenkins *et al.* (2014)) focused on combining evidence from four papers that report gene- and sex-specific penetrance for MLH1 and MSH2 mutation carriers to provide short-term (5 years) CRC risk. While there is some overlap in included studies, the risk estimates provided by our meta-analysis are age-specific, are based on several more studies, and include MSH6 mutation carriers.

A strength of the likelihood-based approach used here lies in its ability to deconvolve aggregated risks, allowing us to use all of the information available in the literature and provide more comprehensive penetrance estimates. Of note, our meta-analysis included only studies that made adjustments for ascertainment if the participants were recruited through high-risk families, so reported risk estimates were less likely to be biased upward. At the same time, many studies were excluded as a result. Our method can be applied in the future to address other Lynch syndrome genes/cancers, such as PMS2, EPCAM, endometrial cancer, and more generally to other gene/cancer combinations with no restriction on the mutation type as long as there are enough studies. A potential limitation of this approach is the use of a parametric distribution to model the penetrance; this assumption, while difficult to check, can be relaxed with richer data. For example, a leave-one-study-out sensitivity analysis can be used to assess the parametric modeling choice. Currently, our meta-analysis includes only papers that report cumulative penetrance. Extensions of our devonvolution method could potentially be designed to include studies that report other risk measures (e.g odds ratio, hazard ratio, etc.). Regarding systematic sources of study heterogeneity, our meta-analysis includes studies of mixed mutation types and populations. While ideally one would desire to assess mutation- or population-specific variation in penetrance, present information is insufficient, and it is not feasible to separate the these effects. Overall, the meta-analytic results for MLH1 and MSH2 mutation carriers are robust according to the sensitivity analysis and show little evidence of publication bias. On the other hand, the same can not be said for MSH6 mutation carriers due to the small number of studies.

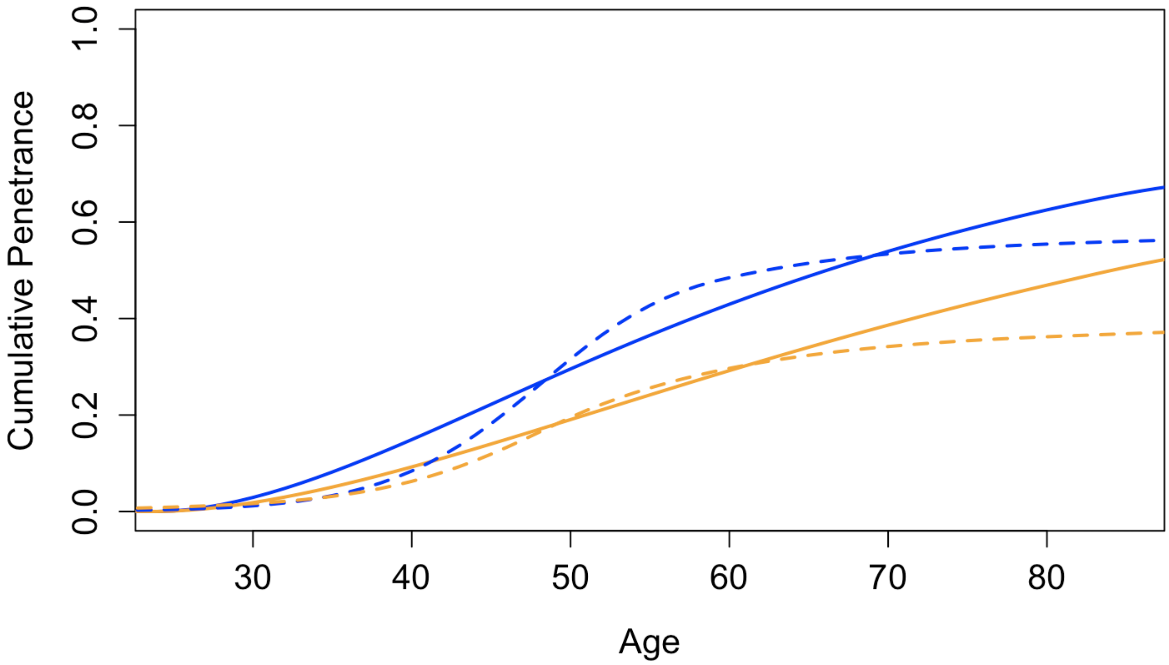
It is well known that colonoscopic surveillance serves as an effective prevention strategy for individuals managing their CRC risk (Järvinen *et al.* (1995)). Our results show that cancer penetrance estimated from populations that are a mix of unscreened and screened individuals is similar to that estimated from unscreened populations for MSH2 mutation carriers. However, the former is higher for MLH1 and female MSH6 mutation carriers. This may be due to the fact that individuals with a family history of CRC are more likely to undergo screening. Thus, the remaining individuals who are unscreened in these studies may have a lower risk of cancer. Moreover, mutation carriers from clinics or population-based registries were referred for enhanced surveillance with colonoscopy, so cancers detected by colonoscopies may increase the cumulative lifetime risk in populations that are a mix of unscreened and screened individuals. While results indicate otherwise for male MSH6 carriers, there is substantial overlap in CIs across all ages, suggesting a lack of evidence to support differences in penetrance between the two groups. It is challenging to compare study results stratified by screening, as the majority of the studies did not fully clarify whether surveillance was part of the patient selection criteria. More refined data would be needed to extend our analysis to incorporate colonoscopic surveillance as a modifier of CRC risk, along with other environmental factors previously shown to affect cancer risk, such as aspirin use (Burn *et al.* (2011)), smoking (Watson *et al.* (2004); Pande *et al.* (2010)), and body mass index (Win *et al.* (2011)).

MMRpro is a genetic counseling and Clinical Decision Support (CDS) tool that estimates the probability of carrying MMR mutations and of developing CRC for mutation carriers. It relies on meta-analytic penetrance estimates (Chen *et al.* (2006)). Chen *et al.* assume the penetrance for MLH1 and MSH2 carriers are the same and that of MSH6 male and female carriers are the same, whereas our meta-analysis contains more studies to substantiate the estimation of gene- and sex-specific risk (Fig. 1.5). In comparison, our results show higher lifetime penetrance estimates for MSH2 and female MLH1 carriers, lower estimates for female MSH6 carriers, and similar estimates for male MLH1 and MSH6 carriers, compared to Chen *et al.* (Fig. 1.5). Of the five studies included in the meta-analysis by Chen *et al.*, we included two in our current analysis (Quehenberger *et al.* (2005); Dunlop *et al.* (1997)). We excluded one due to overlap in study participants (Jenkins *et al.* (2006)), one due to lack of ascertainment adjustment (Hampel *et al.* (2005)) and another because it does not provide colorectal-specific risks (Buttin *et al.* (2004)).

In conclusion, our analysis provides a principled empirical assessment of the risk of Lynch-syndrome-associated CRC by combining evidence from relevant studies. For individuals with Lynch syndrome, the risk of cancer is dependent on sex and type of MMR mutation, with male MLH1 or MSH2 mutation carrier risk at age 70 approximately 4 times higher than that of his female MSH6 counterpart. Risk estimates from our meta-analysis will be incorporated in the 2019 version of the risk prediction tool MMRpro (Chen *et al.* (2006)), and the clinical decision support tool, ASK2ME (All Syndrome Known to Man Evaluator)

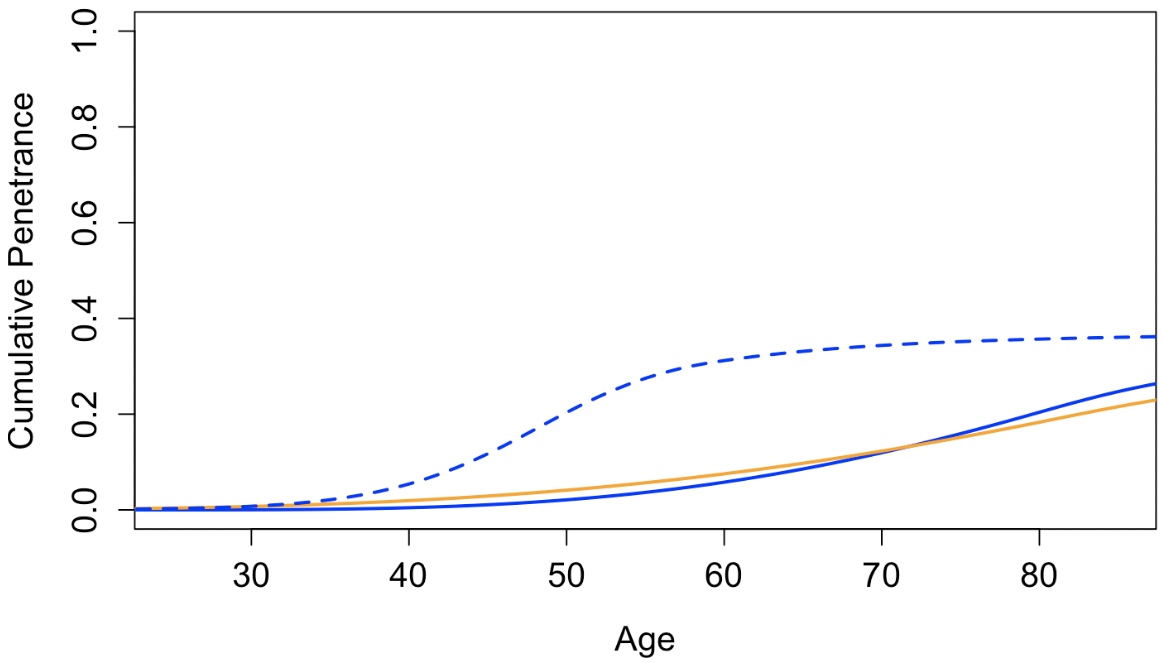

Cumulative Penetrance
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Age

(**a**) *MLH1*


Cumulative Penetrance
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Age

(**b**) *MSH2*

(Continued)



(c) *MSH6*

Figure (1.5) *Cumulative penetrance estimates of colorectal cancer from current meta-analysis and MMR-pro*

Panels (a), (b), and (c) correspond to MLH1, MSH2, and MSH6 mutation carriers, respectively. Estimates from current meta-analysis and MMRpro are denoted by solid and dotted lines, respectively. Blue corresponds to male carriers, and orange corresponds to female carriers.

(Braun *et al.* (2018)) to improve risk prediction and management strategies for individuals who have mutations in MLH1, MSH2, and MSH6. Our results can support the development of effective prevention strategies and personalized counseling.

Chapter 2

Multi-study Boosting: Theoretical Considerations for Merging vs. Ensembling

Cathy Wang[1, 2], Pragya Sur[3], Giovanni Parmigiani[1, 2], Prasad Patil[4]

[1] *Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA*

[2] *Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA*

[3] *Department of Statistics, Harvard University, Cambridge, MA, USA*

[4] *Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA*

Abstract

Cross-study replicability is a powerful model evaluation criterion that emphasizes generalizability of predictions. When training cross-study replicable prediction models, it is critical to decide between merging and treating the studies separately. We study boosting algorithms in the presence of potential heterogeneity in predictor-outcome relationships across studies and compare two multi-study learning strategies: 1) merging all the studies and training a single model, and 2) multi-study ensembling, which involves training a separate model on each study and ensembling the resulting predictions. In the regression setting, we provide theoretical guidelines based on an analytical transition point to determine whether it is more beneficial to merge or to ensemble for boosting with linear learners. In addition, we characterize a bias-variance decomposition of estimation error for boosting with component-wise linear learners. We verify the theoretical transition point result in simulation and illustrate how it can guide the decision on merging vs. ensembling in an application to breast cancer data.

2.1 Background

In settings where comparable studies are available, it is critical to simultaneously consider and systematically integrate information across multiple studies when training prediction models. Multi-study prediction is motivated by applications in biomedical research, where exponential advances in technology and facilitation of systematic data-sharing increased access to multiple studies (Kannan *et al.* (2016); Manzoni *et al.* (2018)). When training and test studies come from different distributions, prediction models trained on a single study generally perform worse on out-of-study samples due to heterogeneity in study design, data collection methods, and sample characteristics. (Castaldi *et al.* (2011); Bernau *et al.* (2014); Trippa *et al.* (2015)). Training prediction models on multiple studies can address these challenges and improve the cross-study replicability of predictions.

Recent work in multi-study prediction investigated two approaches for training cross-study replicable models: 1) merging all studies and training a single model, and 2) multi-study ensembling, which involves training a separate model on each study and combining the resulting predictions. When studies are relatively homogeneous, Patil and Parmigiani (2018) showed that merging can lead to improved replicability over ensembling due to the increase in sample size; as between-study heterogeneity increases, multi-study ensembling demonstrated preferable performance. While the trade-off between these approaches has been explored in detail for random forest (Ramchandran *et al.* (2020)) and linear regression (Guan *et al.* (2019)), none have examined this for boosting, one of the most successful and popular supervised learning algorithms.

Boosting combines a powerful machine learning approach with classical statistical modeling. Its flexible choice of base learners and loss functions makes it highly customizable to many data-driven tasks including binary classification (Freund and Schapire (1997)), regression (Friedman (2001)) and survival analysis (Wang and Wang (2010)). To the best of our knowledge, this work is the first to study boosting algorithms in a setting with multiple and potentially heterogeneous training and test studies. Existing findings on boosting are largely rooted in theories based on a single training study, and extensions of the algorithm to a multi-study setting often assume a subset of the training study shares the same distribution as the test study. Bühlmann (2006) and Tutz and Binder (2007) studied boosting with linear base learners and characterized an exponential bias-variance trade-off under the assumption that the training and test studies have the same predictor distribution. Habrard *et al.* (2013) proposed a boosting algorithm for domain adaptation with a single training study. Dai Wenyuan *et al.* (2007) proposed a transfer learning framework for boosting that uses a small amount of labeled data from the test study in addition to the training data to make classifications on the test study. This approach was extended to handle data from multiple training studies (Yao and Doretto (2010); Bellot and van der Schaar (2019)) and modified for regression (Pardoe and Stone (2010)) and survival analysis (Bellot and van der Schaar (2019)).

In this paper, we study boosting algorithms in a regression setting and compare cross-study replicability of merging versus multi-study ensembling. We assume a flexible mixed effects model with potential heterogeneity in predictor-outcome relationships across studies and provide theoretical guidelines to determine whether merging is more beneficial than ensembling. In particular, we characterize an analytical transition point beyond which ensembling exhibits lower mean squared prediction error than merging for boosting with linear learners. Conditional on the selection path, we characterize a bias-variance decomposition for the estimation error of boosting with component-wise linear learners. We verify the theoretical transition point results via simulations, and illustrate how it may guide practitioners' choice regarding merging vs. ensembling in a breast cancer application.

2.2 Methods

2.2.1 Multi-study Setup

We consider K training studies and V test studies that measure the same outcome and the same p predictors. Each study has size n_k with a combined size of $N = \sum_{k=1}^{K} n_k$ for the training studies and $N^{\text{Test}} = \sum_{k=K+1}^{K+V} n_k$ for the test studies. Let $Y_k \in \mathbb{R}^{n_k}$ and $X_k \in \mathbb{R}^{n_k \times p}$

denote the outcome vector and predictor matrix for study k, respectively. The linear mixed effects model is of the form

$$Y_k = X_k\beta + Z_k\gamma_k + \epsilon_k, \quad k = 1,\ldots,K+V \tag{2.1}$$

where $\beta \in \mathbb{R}^p$ are the fixed effects and $\gamma_k \in \mathbb{R}^q$ the random effects with $E[\gamma_k] = 0$ and $Cov(\gamma_k) = \text{diag}(\sigma_1^2,\ldots,\sigma_q^2) =: G$. If $\sigma_j^2 > 0$, then the effect of the jth predictor varies across studies; if $\sigma_j^2 = 0$, then the predictor has the same effect in each study. The matrix $Z_k \in \mathbb{R}^{n_k \times q}$ is a subset of X_k that corresponds to the random effects, and ϵ_k are the residual errors where $E[\epsilon_k] = 0, Cov(\epsilon_k) = \sigma_\epsilon^2 I$, and $Cov(\gamma_k, \epsilon_k) = 0$. We consider an extension of (2.1) and assume the study data are generated under the mixed effects model of the form

$$Y_k = f(X_k) + Z_k\gamma_k + \epsilon_k, \quad k = 1,\ldots,K+V \tag{2.2}$$

where $f(\cdot)$ is a real-valued function. Compared to (2.1), the model in (2.2) provides more flexibility in fitting the mean function $E(Y_k)$.

For any study k, we assume Y_k is centered to have zero mean and X_k standardized to have zero mean and unit ℓ_2 norm, i.e., $\|X_{jk}\|_2 = 1$ for $j = 1,\ldots,p$, where $X_{jk} \in \mathbb{R}^N$ denotes the jth predictor in study k. Unless otherwise stated, we use $i \in \{1,\ldots,N\}$ to index the observations, $j \in \{1,\ldots,p\}$ the predictors, and $k \in \{1,\ldots,K+V\}$ the studies. For example, $X_{ijk} \in \mathbb{R}$ is the value of the jth predictor for observation i in study k. We formally introduce boosting on the merged study (Y,X) in the next section, but the formulation is the same for the kth study if one were to replace (Y,X) with (Y_k,X_k). In particular, we focus on boosting with linear learners due to its analytical tractability. We denote a linear learner as an operator $H : \mathbb{R}^N \to \mathbb{R}^N$ that maps the responses Y to fitted values $\hat{Y}$. Examples of linear learners include ridge regression and more general projectors to a class of basis functions such as regression or smoothing splines. We denote the basis-expanded predictor matrix by $\tilde{X} \in \mathbb{R}^{N\times P}$ and the subset of predictors with random effects by $\tilde{Z} \in \mathbb{R}^{N\times Q}$. We define the basis-expanded predictor matrix as

$$\tilde{X} = [h(X_i) \quad \cdots \quad h(X_N)]^T \in \mathbb{R}^{N\times P},$$

where

$$h(X_i) = \Big(h_{11}(X_{i1}), \ldots, h_{U_1 1}(X_{i1}), \ldots, h_{1p}(X_{ip}), \ldots, h_{U_p p}(X_{ip})\Big) \in \mathbb{R}^P, \quad i = 1, \ldots, N$$

is the vector of $P = \sum_p U_p$ one-dimensional basis functions evaluated at the predictors $X_i \in \mathbb{R}^p$. As an example, suppose we have $p = 2$ covariates, X_{i1}, X_{i2}, and we want to model X_{i1} linearly and X_{i2} with a cubic spline at knots $\xi_1 = 0$ and $\xi_2 = 1.5$. The basis-expanded predictor matrix $\tilde{X}$ contains the following vector of $P = 6$ basis functions:

$$h(X_i) = (h_{11}(X_{i1}), h_{12}(X_{i2}), h_{22}(X_{i2}), h_{32}(X_{i2}), h_{42}(X_{i2}), h_{51}(X_{i2})), \quad i = 1, \ldots, N$$

where

$$\begin{aligned} h_{11}(X_{i1}) &= X_{i1} & \qquad h_{32}(X_{i2}) &= X_{i2}^3 \\ h_{12}(X_{i2}) &= X_{i2} & \qquad h_{42}(X_{i2}) &= (X_{i2} - 0)_+^3 \\ h_{22}(X_{i2}) &= X_{i2}^2 & \qquad h_{52}(X_{i2}) &= (X_{i2} - 1.5)_+^3 \end{aligned}$$

and $(X_{i2} - \xi)_+^3 = max\Big\{(X_{i2} - \xi)^3, 0\Big\}$. For $\lambda \geq 0$, our goal is to minimize the objective

$$||Y - \tilde{X}\beta||_2^2 + \lambda \beta^T \beta$$

with respect to parameters $\beta \in \mathbb{R}^P$. We denote the vector of coefficient estimates and fitted values by $\hat{\beta} := BY$ and $\hat{Y} := HY$, respectively, where

$$B := (\tilde{X}^T \tilde{X} + \lambda I)^{-1} \tilde{X}^T \in \mathbb{R}^{P \times N}$$

and

$$H := \tilde{X}(\tilde{X}^T \tilde{X} + \lambda I)^{-1} \tilde{X}^T = \tilde{X}B \in \mathbb{R}^{N \times N}.$$

2.2.2 Boosting with linear learners

Given the basis-expanded predictor matrix $\tilde{X} \in \mathbb{R}^{N \times P}$, the goal of boosting is to obtain an estimate $\hat{F}(\tilde{X})$ of the function $F(\tilde{X})$ that minimizes expected loss $E\Big[\ell(Y, F(\tilde{X}))\Big]$ for a given loss function $\ell(\cdot, \cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+^N$, where outcome $Y \in \mathbb{R}^N$ can be continuous (regression problem) or discrete (classification problem). Examples of $\ell(Y, F)$ include exponential loss

$exp(YF)$ for AdaBoost (Freund (1995)) and ℓ_2 (squared error) loss $(Y-F)^2/2$ for ℓ_2 boosting (Bühlmann and Yu (2003)). In finite samples, estimation of $F(\cdot)$ is done by minimizing the empirical risk via functional gradient descent where the base learner $g(\tilde{X};\hat{\theta})$ is repeatedly fit to the negative gradient vector

$$r = \frac{-\partial\ell(Y,F)}{\partial F}\bigg|_{F=\hat{F}_{(m)}(\tilde{X})}$$

evaluated at $\hat{F}_{(m)}(\tilde{X}) = \hat{F}_{(m-1)}(\tilde{X}) + \eta g(\tilde{X};\hat{\theta}_m)$ across $m = 1,\ldots,M$ iterations. Here, $\eta \in (0,1]$ denotes the learning rate, and $\hat{\theta}_m$ denotes the estimated finite or infinite-dimensional parameter that characterizes g (i.e., if g is a regression tree, then θ denotes the tree depth, minimum number of observations in a leaf, etc.). Under ℓ_2 loss, the negative gradient at iteration m is equivalent to the residuals $Y - \hat{F}_m(\tilde{X})$. Therefore, ℓ_2 boosting produces a stage-wise approach that iteratively fits to the current residuals (Bühlmann and Yu (2003); Friedman (2001)).

Let $\hat{\beta}_{(m)} \in \mathbb{R}^P$ and $\hat{Y}_{(m)} \in \mathbb{R}^N$ denote the coefficient estimates and fitted values at the mth boosting iteration, respectively. We describe ℓ_2 boosting with linear learners in **Algorithm 1**.

Algorithm 1 ℓ_2 boosting with linear learners.

```
Initialization:
    $\hat{\beta}_{(0)} = 0, \quad \hat{Y}_{(0)} = 0$
Iteration: For $m = 1,2,\ldots,M$, fit a linear learner to the residuals $r_{(m)} = Y - \hat{Y}_{(m-1)}$ and obtain the estimated coefficients
    $\hat{\beta}^{\text{current}}_{(m)} = Br_{(m)}$
    and fitted values
    $\hat{Y}^{\text{current}}_{(m)} = Hr_{(m)}.$
    The new coefficient estimates are given by:
    $\hat{\beta}_{(m)} = \hat{\beta}_{(m-1)} + \eta\hat{\beta}^{\text{current}}_{(m)}$
    The new fitted values are given by:
    $\hat{Y}_{(m)} = \hat{Y}_{(m-1)} + \eta\hat{Y}^{\text{current}}_{(m)}$
    where $\eta \in (0,1]$ is the learning rate.
```

By Proposition 1 in Bühlmann and Yu (2003), the ℓ_2 boosting coefficient estimates at iteration M can be written as:

$$\hat{\beta}_{(M)}^{\text{Merge}} = \sum_{m=1}^{M} \eta B (I - \eta H)^{m-1} Y. \tag{2.3}$$

Equation (2.3) represents $\hat{\beta}_{(M)}^{\text{Merge}}$ as the sum across coefficient estimates obtained from repeatedly fitting a linear learner H to residuals $r_{(m)} = (I - \eta H)^{m-1} Y$ at iteration $m = 1, \dots, M$. The ensemble estimator, based on pre-specified weights w_k such that $\sum_{k=1}^{K} w_k = 1$, is

$$\hat{\beta}_{(M)}^{Ens} = \sum_{k=1}^{K} w_k \hat{\beta}_{k(M)} = \sum_{k=1}^{K} w_k \left[\sum_{m=1}^{M} \eta B_k (I - \eta H_k)^{m-1} Y_k \right] \tag{2.4}$$

where B_k and H_k $(k = 1, \dots, K)$ are study-specific analogs of B and H, respectively.

2.2.3 Boosting with component-wise linear learners

Boosting with component-wise linear learners (Bühlmann *et al.* (2007); Bühlmann and Yu (2003)), also known as LS-Boost (Friedman (2001)) or least squares boosting (Freund *et al.* (2017)), determines the predictor $\tilde{X}_{\hat{j}_{(m)}} \in \mathbb{R}^N$ that results in the maximal decrease in the univariate least squares fit to the current residuals $r_{(m)}$. The algorithm then updates the $\hat{j}_{(m)}$th coefficient and leaves the rest unchanged. Let $\hat{\beta}_{(m)j} \in \mathbb{R}$ denote the jth coefficient estimate at the mth iteration and $\hat{\beta}_{\hat{j}_{(m)}} \in \mathbb{R}$ the estimated coefficient of the selected covariate in iteration m **Algorithm** 2 describes boosting with component-wise linear learners.

Proposition 1. *Let $e_{\hat{j}_{(m)}} \in \mathbb{R}^P$ denote a unit vector with a 1 in the $\hat{j}_{(m)}$-th position,*

$$B_{(m)} = e_{\hat{j}_{(m)}} \left(\tilde{X}_{\hat{j}_{(m)}}^T \tilde{X}_{\hat{j}_{(m)}} \right)^{-1} \tilde{X}_{\hat{j}_{(m)}}^T,$$

and

$$H_{(m)} = \tilde{X}_{\hat{j}_{(m)}} \left(\tilde{X}_{\hat{j}_{(m)}}^T \tilde{X}_{\hat{j}_{(m)}} \right)^{-1} \tilde{X}_{\hat{j}_{(m)}}^T.$$

The coefficient estimates for ℓ_2 boosting with component-wise linear learners at iteration M can be written as:

$$\hat{\beta}_{(M)}^{\textit{Merge, CW}} = \sum_{m=1}^{M} \eta B_{(m)} \left(\prod_{\ell=0}^{m-1} \left(I - \eta H_{(m-\ell-1)} \right) \right) Y. \tag{2.5}$$

Algorithm 2 ℓ_2 boosting with component-wise linear learners.

```
Initialization:
    $\hat{\beta}_{(0)} = 0, \quad \hat{Y}_{(0)} = 0$
Iteration: For $m = 1, 2, \ldots, M$, compute the residuals
    $r_{(m)} = Y - \hat{Y}_{(m-1)}.$
    Determine the covariate $\tilde{X}_{\hat{j}_{(m)}}$ that results in the best univariate least squares fit to $r_{(m)}$ :
    $\hat{j}_{(m)} = \underset{1 \le j \le P}{\arg\min} \sum_{i=1}^{N} \left( r_{(m)i} - \tilde{X}_{ij} \hat{\beta}_{(m)j} \right)^2 .$
    Calculate the corresponding coefficient estimate:
    $\hat{\beta}_{\hat{j}_{(m)}} = \left( \tilde{X}^T_{\hat{j}_{(m)}} \tilde{X}_{\hat{j}_{(m)}} \right)^{-1} \tilde{X}^T_{\hat{j}_{(m)}} r_{(m)}.$
    Update the fitted values and the coefficient estimate for the $\hat{j}_{(m)}$th covariate
    $\hat{Y}_{(m)} = \hat{Y}_{(m-1)} + \eta \tilde{X}_{\hat{j}_{(m)}} \hat{\beta}_{\hat{j}_{(m)}}$
    $\hat{\beta}_{(m)\hat{j}_{(m)}} = \hat{\beta}_{(m-1)\hat{j}_{(m)}} + \eta \hat{\beta}_{\hat{j}_{(m)}}$
    where $\eta \in (0, 1]$ is a learning rate.
```

A proof is provided in the appendix. Proposition 1 represents $\hat{\beta}^{\text{Merge,CW}}_{(M)}$ as the sum across coefficient estimates obtained from repeatedly fitting a univariate linear learner $H_{(m)}$ to the current residuals $r_{(m)} = (\prod_{\ell=0}^{m-1}(I - \eta H_{(m-\ell-1)}))Y$ at iteration m. As $M \to \infty$, $\hat{\beta}^{\text{Merge,CW}}_{(M)}$ converges to a least squares solution which is unique if the predictor matrix has full rank (Bühlmann *et al.* (2007)). The ensemble estimator, based on pre-specified weights w_k, is

$$\hat{\beta}^{\text{Ens, CW}}_{(M)} = \sum_{k=1}^{K} w_k \hat{\beta}^{\text{CW}}_{(M)k} = \sum_{k=1}^{K} w_k \left[\sum_{m=1}^{M} \eta B_{(m)k} \left(\prod_{\ell=0}^{m-1} \Big(I - \eta H_{(m-\ell-1)k} \Big) \right) Y_k \right] \tag{2.6}$$

where $B_{(m)k}$ and $H_{(m)k}$ are study-specific analogs of $B_{(m)}$ and $H_{(m)}$, respectively.

2.2.4 Performance comparison

We compare merging and ensembling based on mean squared prediction error (MSPE) of V unseen test studies $\tilde{X}_0 \in \mathbb{R}^{N^{\text{Test}} \times P}$ with unknown outcome vector $Y_0 \in \mathbb{R}^{N^{\text{Test}}}$,

$$E[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M)}\|_2^2]$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. To properly characterize the performance of boosting with component-wise linear learners (**Algorithm 2**), we account for the algorithm's adaptive nature by conditioning on its selection path. To make progress analytically, we assume Y is normally distributed with mean $\mu := f(\tilde{X})$ and covariance $\Sigma := \text{blkdiag}(\{Z_k G Z_k^T + \sigma_\epsilon^2 I\}_{k=1}^K)$. At iteration m, selecting the covariate $\tilde{X}_{\hat{j}_{(m)}}$ that results in the best univariate least squares fit to $r_{(m)}$ can be expressed as

$$\|(I - H_{(\hat{j}_{(m)})}) r_{(m)}\|_2^2 \le \|(I - H_{(j)}) r_{(m)}\|_2^2,$$

which is equivalent to

$$(sgn_{(m)} \tilde{X}^T_{\hat{j}(m)} / \|\tilde{X}_{\hat{j}_{(m)}}\|_2 \pm \tilde{X}_j^T / \|\tilde{X}_j\|_2) r^{(m)} \ge 0 \tag{2.7}$$

$\forall j \neq \hat{j}_{(m)}$, $sgn_{(m)} = \text{sign}(\tilde{X}^T_{\hat{j}_{(m)}} r_{(m)})$, where

$$r_{(m)} = \prod_{\ell=0}^{m-1} (I - \eta H_{(m-\ell-1)}) Y := \Upsilon_{(m)} Y.$$

With fixed $\tilde{X}$, the inequalities in (2.7) can be compactly represented as the polyhedral representation $\Gamma Y \geq 0$ for a particular matrix $\Gamma \in \mathbb{R}^{2M(P-1)\times N}$, where the $(\tilde{m}+2(j-\omega(j))-1)$th and $(\tilde{m}+2(j-\omega(j)))$th rows are given by

$$(sgn_{(m)}\tilde{X}_{\hat{j}_{(m)}}^T/\|\tilde{X}_{\hat{j}_{(m)}}\|_2 \pm \tilde{X}_j^T/\|\tilde{X}_j\|_2)\Upsilon^{(m)}$$

$\forall j \neq \hat{j}_{(m)}$ with $\tilde{m} = 2(P-1)(m-1)$ and $\omega(j) = \mathbb{1}\{j > \hat{j}_{(m)}\}$ (Rügamer and Greven (2020)). The jth regression coefficient in **Algorithm** 2 can be written as

$$\hat{\beta}_{(M)j}^{\text{Merge, CW}} = v_j^T Y,$$

where $v_j = (\sum_{m=1}^{M} \eta B_{(m)}(\prod_{\ell=0}^{m-1}(I-\eta H_{(m-\ell-1)}))^T e_j$ and $e_j \in \mathbb{R}^P$ is a unit vector. The distribution of $\hat{\beta}_{(M)j}^{\text{Merge, CW}}$ conditional on the selection path is given by the polyhedral lemma in Lee *et al.* (2016).

Lemma 1 (Polyhedral lemma from Lee *et al.* (2016)). *Given the selection path*

$$\mathcal{P} := \{Y : \Gamma Y \geq 0, z_j = z\},$$

where $z_j := (I - c_j v_j^T)Y$ *and* $c_j := \Sigma v_j (v_j^T \Sigma v_j)^{-1}$,

$$\hat{\beta}_{(M)j}^{Merge,\ CW} | \mathcal{P} \sim TruncatedNormal\Big(v_j^T\mu, v_j \Sigma v_j^T, a_j, b_j\Big),$$

where

$$a_j = \max_{\ell:(\Gamma c_j)_\ell > 0} \frac{0-(\Gamma z_j)_\ell}{(\Gamma c_j)_\ell}$$
$$b_j = \min_{\ell:(\Gamma c_j)_\ell < 0} \frac{0-(\Gamma z_j)_\ell}{(\Gamma c_j)_\ell}.$$

A proof is provided in the appendix. The conditioning is important because it properly accounts for the adaptive nature of **Algorithm** 2. Conceptually, it measures the magnitude of $\hat{\beta}_{(M)j}^{\text{Merge, CW}}$ among random vectors Y that would result in the selection path $\Gamma Y \geq 0$ for a fixed value of z_j. When $\Sigma = \sigma^2 I$, $z_j = (I - v_j(v_j^T v_j)^{-1} v_j^T)Y$ is the projection onto the orthocomplement of v_j. Accordingly, the polyhedron $\Gamma Y \geq 0$ holds if and only if $v_j^T Y$ does not deviate too far from z_j, hence trapping it between bounds a_j and b_j (Tibshirani *et al.* (2016)).

Moreover, because a_j and b_j are functions of z_j alone, they are independent of $v^T Y$ under normality. The result in Lemma 1 allows us to analytically characterize the mean squared error of the estimators $\hat{\beta}^{\text{Merge, CW}}_{(M)j}$ and $\hat{\beta}^{\text{Ens, CW}}_{(M)j}$ conditional on their respective selection paths.

2.2.5 Implicit regularization and early stopping

In **Algorithm 1** and **Algorithm 2**, the learning rate η and stopping iteration M together control the amount of shrinkage and training error. A smaller learning rate η leads to slower overfitting but requires a larger M to reduce the training error to zero. With a small η, it is possible to explore a larger class of models, which often leads to models with better predictive performance (Friedman (2001)). While boosting algorithms are known to exhibit slow overfitting behavior with small values of η, it is necessary to implement early stopping strategies to avoid overfitting (Schapire *et al.* (1998)). The boosting fit for **Algorithm 1** in iteration m (assuming $\eta = 1$) is

$$\mathcal{B}_{(m)} Y := (I - (I - H)^{m+1}) Y,$$

where $\mathcal{B}_{(m)} : \mathbb{R}^N \to \mathbb{R}^N$ is the boosting operator. For a base learner that satisfies $\|I - H\| \leq 1$ for a suitable norm, we have $\mathcal{B}_{(m)} Y \to Y$ as $m \to \infty$. That is, if left to run forever, the boosting algorithm converges to the fully saturated model Y (Bühlmann *et al.* (2007)). A similar argument can be made for **Algorithm 2** where

$$\mathcal{B}^{\text{CW}}_{(m)} = I - (I - H_{(\hat{j}_m)})(I - H_{(\hat{j}_{m-1})}) \cdots (I - H_{(\hat{j}_1)})$$

is the component-wise boosting operator. We define the degrees of freedom at iteration m as $tr(\mathcal{B}_{(m)})$ and use the corrected AIC criterion (AIC_c) (Bühlmann (2006)) to choose the stopping iteration M. Compared to cross-validation (CV), AIC_c-tuning is computationally efficient as it does not require running the boosting algorithm multiple times. For **Algorithm 1**, the AIC_c at iteration m is given by

$$AIC_c(m) = \log(\underline{\hat{\sigma}}^2) + \frac{1 + tr(\mathcal{B}_{(m)})/N}{1 - (tr(\mathcal{B}_{(m)}) + 2)/N}, \tag{2.8}$$

where $\underline{\hat{\sigma}}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - (\mathcal{B}_{(m)}Y)_i)^2$. The stopping iteration is

$$M = \underset{1 \leq m \leq m_{upp}}{\arg\min}\, AIC_c(m),$$

where m_{upp} is a large upper bound for the candidate number of boosting iterations (Bühlmann (2006)). For **Algorithm 2**, the AIC_c is computed by replacing $\mathcal{B}_{(m)}$ with $\mathcal{B}_{(m)}^{\text{CW}}$. We allow the stopping iterations to differ between the merged and ensemble learners. In our results, we denote them by M and $M_{\text{Ens}} = \{M_k\}_{k=1}^K$, respectively.

2.3 Results

We summarize the degree of heterogeneity in predictor-outcome relationships across studies by the sum of the variances of the random effects divided by the number of fixed effects: $\overline{\sigma^2} := tr(G)/P$, where $G \in \mathbb{R}^{Q \times Q}$. For boosting with linear learners, let $\tilde{R} = \sum_{m=1}^{M} \eta B(I - \eta H)^{m-1}$ and $\tilde{R}_k = \sum_{k=1}^{K} w_k [\sum_{m=1}^{M_k} \eta B_k (I - \eta H_k)^{m-1}]$. Let $b_{\text{Merge}} = Bias(\hat{\beta}_{(M)}^{\text{Merge}}) = \tilde{R}f(\tilde{X}) - f(\tilde{X}_0)$ denote the bias of the boosting coefficients for the merged estimator and $b_{\text{Ens}} = Bias(\hat{\beta}_{(M_{\text{Ens}})}^{\text{Ens}}) = \sum_{k=1}^{K} w_k \tilde{R}_k f(\tilde{X}_k) - f(\tilde{X}_0)$ the bias for the ensemble estimator. Let $Z' = blkdiag(\{Z_k\}_{k=1}^K)$ and $G' = blkdiag(\{G_k\}_{k=1}^K)$ where $G_k = G$ for $k = 1, \ldots, K$.

2.3.1 Boosting with linear learners

Theorem 1. *Suppose*

$$tr(Z'^T \tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R} Z') - \sum_{k=1}^{K} w_k^2 tr(Z_k^T \tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k Z_k) > 0 \tag{2.9}$$

Define

$$\tau = \frac{Q}{P} \times \frac{\sigma_\epsilon^2 (\sum_{k=1}^K w_k^2 tr(\tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k) - tr(\tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R})) + (b^{Ens})^T b^{Ens} - (b^{Merge})^T b^{Merge}}{tr(Z'^T \tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R} Z') - \sum_{k=1}^K w_k^2 tr(Z_k^T \tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k Z_k)} \tag{2.10}$$

Then $E[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M_{Ens})}^{Ens}\|_2^2] \leq [\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M)}^{Merge}\|_2^2]$ *if and only if* $\overline{\sigma}^2 \geq \tau$.

A proof is provided in the appendix. Under the equal variances assumption, Theorem 1 characterizes a transition point τ beyond which ensembling outperforms merging for

Algorithm 1. τ is characterized by differences in the predictive performance of merging vs. ensembling driven by within-study variability and bias in the numerator and between-study variability in the denominator. The condition in (2.9), which ensures τ is well defined, holds when the between-study variability of $\hat{\beta}^{\text{Merge}}_{(M)}$ is greater than that of $\hat{\beta}^{\text{Ens}}_{(M_{Ens})}$. This is generally true because merging does not account for between-study heterogeneity. τ depends on the population mean function f through the bias term. Therefore, an estimate of f is required to estimate the transition point unless the bias is equal to zero. One example of an unbiased estimator is ordinary least squares, which can be obtained by setting $H = \tilde{X}(\tilde{X}^T\tilde{X})\tilde{X}^T$ and $M = \eta = 1$. In general, for any linear learner $H : \mathbb{R}^N \to \mathbb{R}^N$, the transition point in Guan *et al.* (2019) (cf., Theorem 1) is a special case of (2.10) when $M = \eta = 1$.

Corollary 1. *Suppose* $tr(Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z') \neq 0$. *As* $\sigma^2 \to \infty$,

$$\frac{E[\|Y_0 - \tilde{X}\hat{\beta}^{Ens}_{(M_{Ens})}\|_2^2]}{E[\|Y_0 - \tilde{X}_0\hat{\beta}^{Merge}_{(M)}\|_2^2]} \longrightarrow \frac{\sum_{k=1}^K w_k^2 tr(Z_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k)}{tr(Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z')}.$$

This result follows immediately from Theorem 1. According to Corollary 1, the asymptote of the MSPE ratio comparing ensembling to merging equals the ratio of between-study variability. Because the merged estimator does not account for between-study variability, the asymptote is less than one.

Let $\sigma^2_{(1)}, \ldots, \sigma^2_{(D)}$ denote the distinct values of variances of the random effects, and let J_d denote the number of random effects with variance $\sigma^2_{(d)}$.

Theorem 2. *Suppose*

$$\max_d \sum_{i:\sigma_i^2=\sigma_{(d)}^2} \left[\sum_{k=1}^K \left(Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z'\right)_{i+Q\times(k-1),i+Q\times(k-1)} - w_k^2\left(Z_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k\right)_{i,i}\right] > 0$$

and define

$$\tau_1 = \frac{\sigma_\epsilon^2(\sum_{k=1}^K w_k^2 tr(\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k) - tr(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R})) + (b^{Ens})^T b^{Ens} - (b^{Merge})^T b^{Merge}}{P\max_d \frac{1}{J_d}\sum_{i:\sigma_i^2=\sigma_{(d)}^2}[\sum_{k=1}^K (Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z')_{i+Q\times(k-1),i+Q\times(k-1)} - w_k^2(Z_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k)_{i,i}]}. \tag{2.11}$$

Then $E[\|Y_0 - \tilde{X}_0\hat{\beta}^{Ens}_{(M_{Ens})}\|_2^2] \geq E[\|Y_0 - \tilde{X}_0\hat{\beta}^{Merge}_{(M)}\|_2^2]$ *when* $\overline{\sigma}^2 \leq \tau_1$.

Suppose

$$\min_{d} \sum_{i:\sigma_i^2=\sigma_{(d)}^2} \left[\sum_{k=1}^{K} \left(Z'^T \tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R} Z' \right)_{i+Q\times(k-1), i+Q\times(k-1)} - w_k^2 \left(Z_k^T \tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k Z_k \right)_{i,i} \right] > 0$$

and define

$$\tau_2 = \frac{\sigma_\epsilon^2 (\sum_{k=1}^K w_k^2 tr(\tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k) - tr(\tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R})) + (b^{Ens})^T b^{Ens} - (b^{Merge})^T b^{Merge}}{P \min_d \frac{1}{J_d} \sum_{i:\sigma_i^2=\sigma_{(d)}^2} [\sum_{k=1}^K (Z'^T \tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R} Z')_{i+Q\times(k-1), i+Q\times(k-1)} - w_k^2 (Z_k^T \tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k Z_k)_{i,i}]}. \tag{2.12}$$

Then $E[\|Y_0 - \tilde{X}_0 \hat{\beta}^{Ens}_{(M_{Ens})}\|_2^2] \leq E[\|Y_0 - \tilde{X}_0 \hat{\beta}^{Merge}_{(M)}\|_2^2]$ *when* $\overline{\sigma}^2 \geq \tau_2$.

A proof is provided in the appendix. Theorem 2 generalizes Theorem 1 to account for unequal variances along the diagonal of G. It characterizes a transition interval $[\tau_1, \tau_2]$ where merging outperforms ensembling when $\overline{\sigma}^2 \leq \tau_1$ and vice versa when $\overline{\sigma}^2 \geq \tau_2$. The transition interval provided by Guan *et al.* (2019) (cf. Theorem 2) is a special case of (2.11, 2.12) when $M = \eta = 1$.

2.3.2 Boosting with component-wise linear learners

To properly characterize the performance of the boosting estimator in **Algorithm** 2, we condition on its selection path. To this end, we provide the conditional MSE of the merged and ensemble estimators in Proposition 2. Assuming $Y \sim MVN(\mu, \Sigma)$, it follows that Y_k is normal with mean $\mu_k := f(\tilde{X}_k)$ and covariance $\Sigma_k := Z_k G Z_k^T + \sigma_\epsilon^2 I$ for $k = 1, \ldots, K$. Let

$$\mathcal{P} = \{Y : \Gamma Y \geq 0, z_j = z\}$$

and

$$\mathcal{P}^{\text{Ens}} = \{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$$

denote the conditioning events for the merged and ensemble estimators, respectively, where

$$\mathcal{P}_k := \{Y_k : \Gamma_k Y_k \geq 0, z_{jk} = z_k\}$$

summarizes the boosting path from fitting **Algorithm 2** to the data in study k. Let $\bar{\mu}_j = v_j^T \mu$ and $\vartheta_j^2 = v_j \Sigma v_j^T$ denote the mean and variance of $\hat{\beta}^{\text{CW, Merge}}_{(M)j} = v_j^T Y$, respectively. And let

$\alpha_j = \frac{a_j - \bar{\mu}_j}{\vartheta_j}$ and $\xi_j = \frac{b_j - \bar{\mu}_j}{\vartheta_j}$ denote the standardized lower and upper truncation limits. We denote the study-specific versions of $\bar{\mu}_j, \theta_j, \alpha_j$ and ξ_j by $\bar{\mu}_{jk}, \theta_{jk}, \alpha_{jk}$, and ξ_{jk}, respectively.

Proposition 2. *Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative distribution functions of a standard normal variable, respectively. The conditional mean squared error (MSE) of the merged estimator is*

$$E\left[\left(\hat{\beta}_{(M)j}^{Merge,\ CW} - \beta_j\right)^2 \middle| \mathcal{P}\right] = \left(\bar{\mu}_j - \vartheta_j\left(\frac{\phi(\xi_j) - \phi(\alpha_j)}{\Phi(\xi_j) - \Phi(\alpha_j)}\right) - \beta_j\right)^2 + \vartheta_j^2\left(1 - \frac{\xi_j\phi(\xi_j) - \alpha_j\phi(\alpha_j)}{\Phi(\xi_j) - \Phi(\alpha_j)} - \left(\frac{\phi(\xi_j) - \phi(\alpha_j)}{\Phi(\xi_j) - \Phi(\alpha_j)}\right)^2\right).$$

The conditional MSE of the ensemble estimator is

$$E\left[\left(\hat{\beta}_{(M_{Ens})j}^{Ens,\ CW} - \beta_j\right)^2 \middle| \mathcal{P}^{Ens}\right] = \left(\sum_{k=1}^{K} w_k\left(\bar{\mu}_{jk} - \vartheta_{jk}\left(\frac{\phi(\xi_{jk}) - \phi(\alpha_{jk})}{\Phi(\xi_{jk}) - \Phi(\alpha_{jk})}\right)\right) - \beta_j\right)^2 + \sum_{k=1}^{K} w_k^2\vartheta_{jk}^2\left(1 - \frac{\xi_{jk}\phi(\xi_{jk}) - \alpha_{jk}\phi(\alpha_{jk})}{\Phi(\xi_{jk}) - \Phi(\alpha_{jk})} - \left(\frac{\phi(\xi_{jk}) - \phi(\alpha_{jk})}{\Phi(\xi_{jk}) - \Phi(\alpha_{jk})}\right)^2\right).$$

A proof is provided in the appendix. Proposition 2 characterizes the conditional MSE of boosting estimators via the bias-variance decomposition. By the polyhedral lemma (Lee *et al.* (2016)), the selection path $\Gamma Y \geq 0$ is equivalent to truncating $\hat{\beta}_{(M)j}^{\text{Merge, CW}} = v_j^T Y$ to an interval $[a_j, b_j]$ around z_j. When there is no between-study heterogeneity, $z_j = (I - v_j(v_j^T v_j)^{-1} v^T)Y$ is the residual from projecting Y onto v_j. Loosely speaking, the selection path is equivalent to $v_j^T Y$ not deviating too far from z_j. As shown in Section 2.2.4, we can rewrite the selection path as a system of $2M(P-1)$ inequalities with the variable $v_j^T Y$:

$$\{\Gamma Y \geq 0\} = \{\Gamma c_j(v_j^T Y) \leq -\Gamma z_j\}. \tag{2.13}$$

For fixed P, as the number of boosting iterations M increases, the number of linear inequalities (or constraints) in (2.13) also increases; as a result, the size of the polyhedron $\Gamma Y \geq 0$ decreases. A smaller polyhedron generally leads to a narrower truncation interval $[a_j, b_j]$ around $v_j^T Y$. Intuitively, a tighter truncation interval leads to reduced variance. When between-study heterogeneity is low, at a fixed learning rate η, the merged model generally requires a later stopping iteration than the study-specific model due to the increase in sample size. Therefore, $\hat{\beta}_{(M)j}^{\text{Merge, CW}}$ tends to have a tighter truncation region, and as a result, smaller variance than $\hat{\beta}_{(M_{\text{Ens}})j}^{\text{Ens, CW}}$. As between-study heterogeneity increases, the merged model often has an earlier stopping iteration to avoid overfitting, so $Var(\hat{\beta}_{(M)j}^{\text{Merge, CW}}) > Var(\hat{\beta}_{(M_{\text{Ens}})j}^{\text{Ens, CW}})$. In practice, the variance component in Proposition 2 can be computed given estimates of σ^2 and f.

2.4 Simulations

We conducted simulations to evaluate the performance of boosting with four base learners: ridge, component-wise least squares (CW-LS), component-wise cubic smoothing splines (CW-CS) and regression trees. We sampled predictors from the `curatedOvarianData` R package (Ganzfried *et al.* (2013)) to reflect realistic and potentially heterogeneous predictor distributions. The true data-generating model contains $p = 10$ predictors of which 5 have random effects. The outcome for individual i in study k is

$$Y_{ik} = f(X_{ik}) + Z_{ik}\gamma_k + \epsilon_{ik}, \tag{2.14}$$

where $\gamma_k \sim MVN(0, G)$ with $G = diag(\sigma_1^2, \ldots, \sigma_5^2)$, $Z_{ik} = (X_{3ik}, X_{4ik}, X_{5ik}, X_{6ik}, X_{7ik})$, and $\epsilon_{ik} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 1$ for $i = 1, \ldots, n_k, k = 1, \ldots, K$. The mean function f has the form

$$\begin{aligned} f(X_{ik}) = &-0.28h_{11}(X_{1ik}) - 0.12h_{21}(X_{1ik}) - 0.78h_{31}(X_{1ik}) + 0.035h_{41}(X_{1ik}) - 0.23X_{2ik} \\ &+ 1.56X_{3ik} - 0.0056X_{4ik} + 0.13X_{5ik} + 0.0013X_{6ik} - 0.00071X_{7ik} - 0.0023X_{8ik} \\ &- 0.69X_{9ik} + 0.016X_{10ik} \end{aligned} \tag{2.15}$$

where $h_{11}, \ldots, h_{41}$ are cubic basis splines with a knot at 0, and the coefficients were generated from $N(0, 0.5)$. The coefficients for $X_{2ik}, X_{3ik}, X_{5ij}$ and X_{9ik} were generated from $N(0, 1)$, and those for $X_{4ik}, X_{6ik}, X_{7ik}, X_{8ik}$, and X_{10ik} were generated from $N(0, 0.01)$.

We generated $K = 4$ training and $V = 4$ test studies of size 100. For each simulation replicate $s = 1, \ldots, 500$, we generated outcomes for varying levels of $\overline{\sigma}^2$, trained merged and multi-study ensemble boosting models and evaluated them on the test studies.The outcome was centered to have zero mean and predictors standardized to have zero mean and unit ℓ_2 norm. The regularization parameter λ for ridge boosting and stopping iteration M for tree boosting were chosen using 3-fold cross validation. The stopping iteration for linear base learners (ridge, CW-LS, and CW-CS) were chosen based on the AIC_c-tuning procedure described in Section 2.2.5. All hyperparameters were tuned on a held-out data set of size 400 with σ^2 set to zero. For tree boosting, we set the maximum tree-depth to two. A learning rate of $\eta = 0.5$ was used for all boosting models. For the ensemble estimator, equal weight was assigned to each study. We considered two cases for the structure of $G:$ 1) equal variance and 2) unequal variance. In the first case, Figure 3.1 shows the relative predictive performance comparing multi-study ensembling to merging for varying levels of $\overline{\sigma}^2$. When $\overline{\sigma}^2$ was small, the merged learner outperformed the ensemble learner. As $\overline{\sigma}^2$ increased, there exists a transition point beyond which ensembling outperformed merging. The empirical transition point based on simulation results confirmed the theoretical transition point (2.10) for boosting with linear learners. As $\overline{\sigma}^2$ tended to infinity, the log relative performance ratio tended to -0.81 by Corollary 1. Figure 3.2 shows the relative predictive performance under the unequal variance case. For boosting with linear learners, there exists a transition interval $[\tau_1, \tau_2]$ where merging outperformed ensembling when $\overline{\sigma}^2 \leq \tau_1$ and vice versa when $\overline{\sigma}^2 \geq \tau_2$. Compared to boosting with linear or tree learners, boosting with component-wise learners had an earlier transition point.

For boosting with component-wise linear learners, we compared the performance of merging and multi-study ensembling based on results in Proposition 2. In each simulation replicate, we generated outcomes based on (2.15) and estimated $\beta_{(M)}^{\text{CW, Merge}}$ and $\beta_{(M)}^{\text{CW, Ens}}$ with M set to 30. We assumed equal variance along the diagonal of G. At each boosting iteration $m = 1, ..., M$, we evaluated the MSE for both estimators with respect to $\beta_6 = 1.72$ conditional on the boosting path up to iteration m. We chose to evaluate the coefficient associated with X_6 because the true data-generating coefficient β_6 had the largest magnitude, and as a result, the component-wise boosting algorithm was more likely to select X_6. Figure 3.3 shows the MSE associated with the merged and ensemble estimators at $\overline{\sigma}^2 = 0.01$ and 0.05. We chose these values because the empirical transition point for boosting with component-wise linear

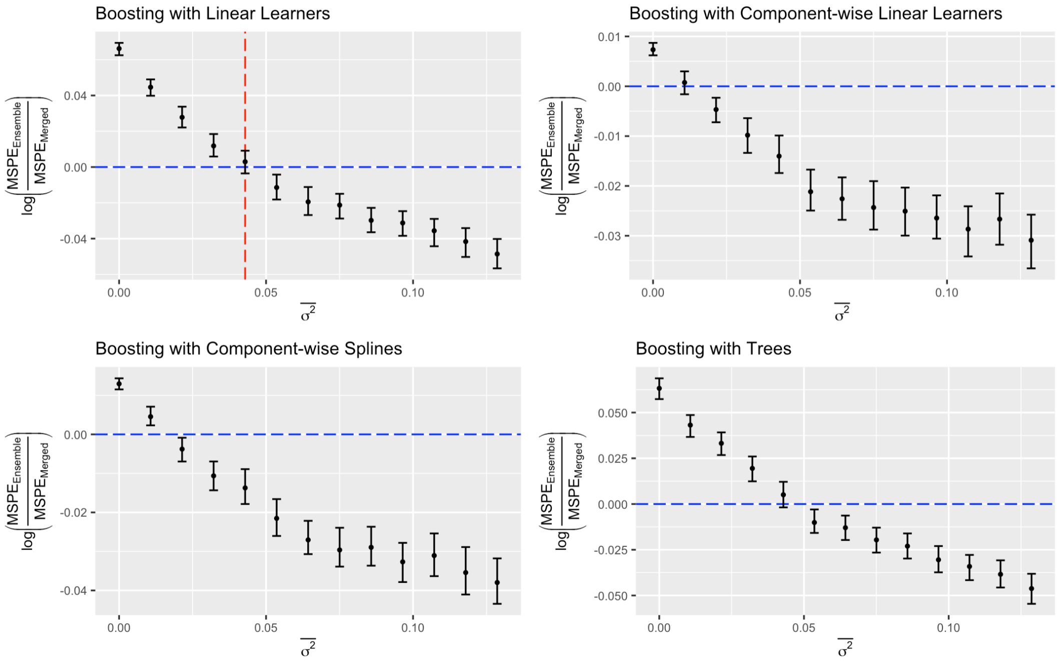


Figure (2.1) *Prediction performance of multi-study ensembling vs. merging for boosting under the equal variance assumption*

Log relative mean squared prediction error (MSPE) of multi-study ensembling vs. merging for boosting with different base learners under the equal variance assumption. The red vertical dashed line indicates the transition point τ. The solid circles represent the average performance ratios comparing multi-study ensembling to merging, and vertical bars the 95% bootstrapped intervals.

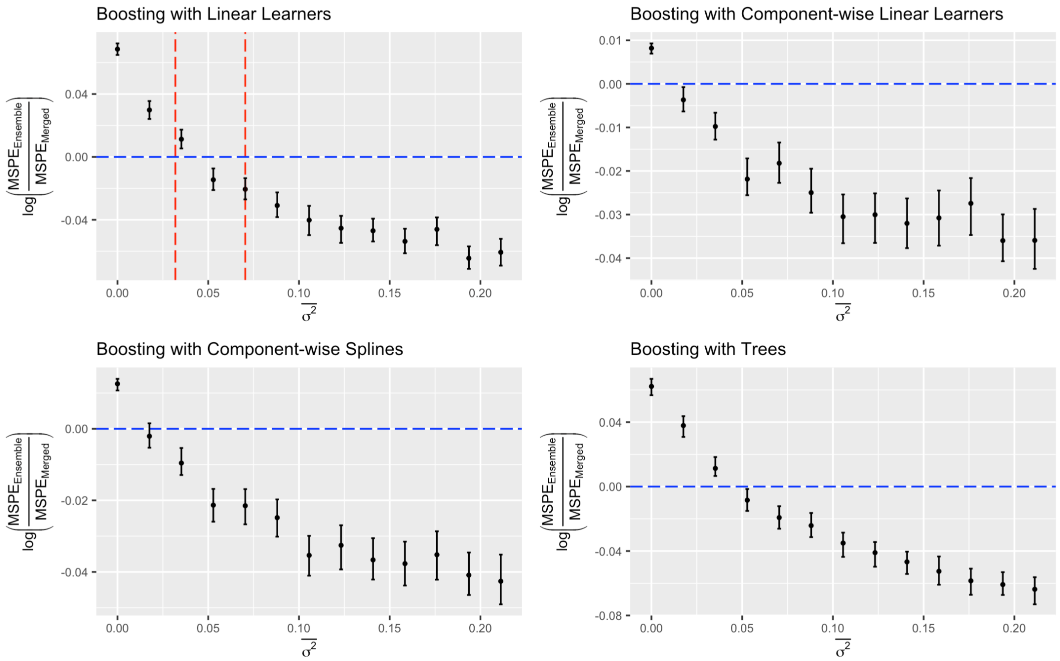


Figure (2.2) *Prediction performance of multi-study ensembling vs. merging for boosting under the unequal variance assumption.*

Log relative mean squared prediction error (MSPE) of multi-study ensembling vs. merging for boosting with different base learners under the unequal variance assumption. The red vertical dashed lines indicates the transition interval $[\tau_1, \tau_2]$. The solid circles represent the average performance ratios comparing multi-study ensembling to merging, and vertical bars the 95% bootstrapped intervals.

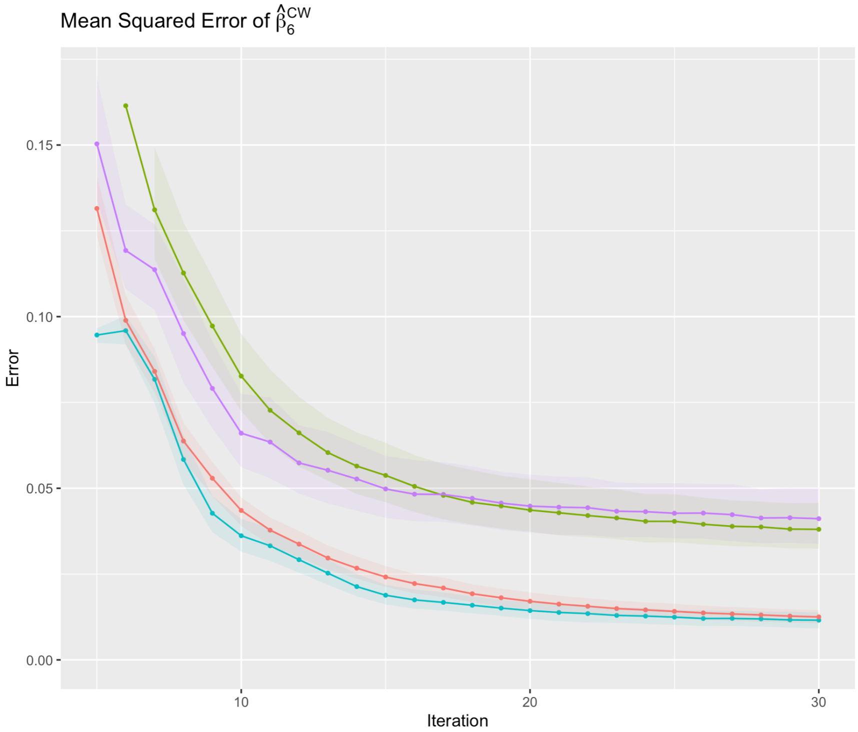


Figure (2.3) *Mean squared error of ensembling and merging estimators across different levels of* $\overline{\sigma}^2$

Blue and red lines correspond to the merged and ensemble estimators at $\overline{\sigma}^2 = 0.01$*, respectively. Purple and green lines correspond to the merged and ensemble estimators at* $\overline{\sigma}^2 = 0.05$*, respectively.*

learners in Figure 3.1 lay between 0.01 and 0.05. When $\overline{\sigma}^2 = 0.01$, merging outperformed ensembling. As the number of boosting iterations increased, both performed similarly. At $\overline{\sigma}^2 = 0.05$, merging outperformed ensembling up until $M = 20$, beyond which ensembling began to show preferable performance.

2.5 Breast Cancer Application

Using data from the `curatedBreastData` R package (Planey (2020)), we illustrated how the transition point theory could guide decisions on merging vs. ensembling. This R package contains 34 high-quality gene expression microarray studies from over 16 clinical trials on individuals with breast cancer. The studies were normalized and post-processed using the `processExpressionSetList()` function. In practice, a key determinant of breast cancer prognosis and staging is tumor size (Fleming (1997)). Clinicians use the TNM (tumor, node, metastasis) system to describe how extensive the breast cancer is. Under this system, "T" plus a letter or number (0 to 4) is used to describe the size (in centimeters (cm)) and location of the tumor. While the best way to measure the tumor is after it's been removed from the breast, information on tumor size can help clinicians develop effective treatment strategies. Common treatment options for breast cancer include surgery (e.g., mastectomy or lumpectomy), drug therapy (e.g., chemotherapy or immunotherapy) or a combination of both (Gradishar *et al.* (2021)).

In our data illustration, the goal was to predict tumor size (cm) before treatment and surgery. We trained boosting models on $K = 5$ training studies with a combined size of $N = 643$: ID 1379 ($n = 60$), ID 2034 ($n = 281$), ID 9893 ($n = 155$), ID 19615 ($n = 115$) and ID 21974 ($n = 32$) and evaluated them on $V = 4$ test studies with a combined size of $N^{\text{Test}} = 366$: ID 21997 ($n = 94$), ID 22226 ($n = 144$), ID 22358 ($n = 122$), and ID 33658 ($n = 10$). We selected the top $p = 40$ gene markers that were most highly correlated with tumor size in the training studies as predictors and randomly selected $q = 8$ to have random effects with unequal variance. To calculate the transition interval from Theorem 2, we trained boosting models with ridge learners using two strategies: merging and ensembling. We also estimated the variances of the random effects ($\sigma_1^2, \ldots, \sigma_8^2$) and residual error ($\sigma_\epsilon^2$) by fitting a linear mixed effects model using restricted maximum likelihood. The estimate of $\overline{\sigma}^2$ and σ_ϵ^2 were 4.32×10^{-2} and 1.053, respectively, and the transition interval was $[0.020, 0.026]$. Aside from ridge, we trained boosting models with three other base learners: CW-LS, CW-CS and regression trees. Results comparing the predictive performance of ensembling vs. merging are shown in Figure 3.4. By Theorem 2, merging would be preferred over ensembling for boosting with ridge learners because the estimate of $\overline{\sigma}^2$ was smaller than the lower bound of the transition interval. This result was corroborated by the boxplot of performance ratios in Figure 3.4.

Among the boosting algorithms that perform variable selection, ensembling outperformed merging when boosting with regression trees, and both performed similarly when boosting with component-wise learners. Table 2.1 summarizes the top three genes selected by each

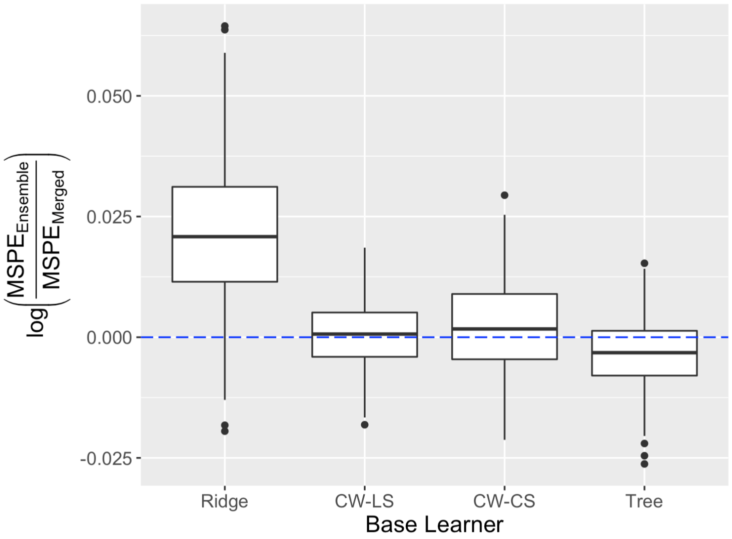


Figure (2.4) *Prediction performance of multi-study ensembling vs. merging for breast cancer application*

Log relative mean squared prediction error (MSPE) of multi-study ensembling vs. merging for boosting with different base learners under the equal variance assumption. Ridge = ridge regression; CW-LS = component-wise least squares; CW-CS = component-wise cubic smoothing splines; tree = regression tree.

algorithm. Genes were ordered by decreasing variable importance, which was defined as the reduction in training error attributable to selecting a particular gene. In the merged study, both boosting with CW-CS and trees selected the same three genes: *S100P*, *MMP11*, and *E2F8*, whereas boosting with CW-LS selected *S100P*, *ASPN*, and *STY1*. This may be attributed to the fact that, compared to CW-LS, CW-CS and trees are more flexible and can capture non-linear trends in the data. Overall, there was some overlap in the genes that were selected by the three base learners across studies. In study ID 1379, all three base learners selected *S100P*, and all but the tree learner selected *AEBP1*. In studies ID 9893, 19615 and 21974, all three learners selected *PPP1R3C*, *CD9*, and *CD69*, respectively. Tree boosting selected a single gene in studies ID 1379, 9893, and 21974 because the optimal number of boosting iterations determined by 3-fold CV was one. In general, CV-tuning leads to earlier stopping iterations than AIC_c-tuning as CV approximates the test error on a smaller sample.

Learner	ID 1379	ID 2034	ID 9893	ID 19615	ID 21974	Merged
CW-LS	S100P (0.135)	MMP11 (0.0455)	PPP1R3C (0.0421)	CENPN (0.111)	CD69 (0.193)	S100P (0.0215)
	AEBP1 (0.129)	CENPA (0.0241)	IGF1 (0.0208)	CD9 (0.0767)	MMP11 (0.108)	ASPN (0.0184)
	CENPA (0.0652)	CAMP (0.0204)	SYT1 (0.0183)	ASPN (0.0733)	ESR1 (0.0358)	SYT1 (0.0133)
CW-CS	AEBP1 (0.133)	TNFSF4 (0.0477)	PPP1R3C (0.0463)	CENPN (0.103)	MMP11 (0.183)	S100P (0.021)
	C10orf116 (0.115)	S100A9 (0.0405)	GRP (0.0342)	CD9 (0.0865)	CD69 (0.182)	MMP11 (0.0195)
	S100P (0.100)	CLU (0.0321)	POSTN (0.0256)	COL1A1 (0.0848)	S100P(0.0889)	E2F8 (0.0185)
Tree	S100P (0.111)	S100A9 (0.0699)	PPP1R3C (0.0438)	COL1A1 (0.131)	CD69 (0.147)	MMP11 (0.0286)
	N/A	MMP11 (0.0588)	N/A	CD9 (0.108)	N/A	S100P (0.0266)
	N/A	N/A	N/A	ADRA2A (0.0732)	N/A	E2F8 (0.0249)

Table (2.1) *Selected genes ordered by decreasing variable importance across different training studies*

Each entry in the table consists of the gene name followed by the amount of reduction in training error that is attributed to selecting the gene in parentheses. An entry is N/A if there were fewer than three selected genes. CW-LS = component-wise least squares and CW-CS = component-wise cubic smoothing splines.

2.6 Discussion

In this paper, we studied boosting algorithms in a regression setting and compared merging and multi-study ensembling for improving cross-study replicability of predictions. We assumed a flexible mixed effects model with potential heterogeneity in predictor-outcome relationships across studies and provided theoretical guidelines for determining whether it was more beneficial to merge or to ensemble. In particular, we extended the transition point theory from Guan *et al.* (2019) to boosting with linear learners. For boosting with component-wise linear learners, we characterized a bias-variance decomposition of estimation error conditional on the selection path.

Boosting under ℓ_2 loss is computationally simple and analytically attractive. In general, performance of the algorithm is inextricably linked with the choice of learning rate η and stopping iteration M. Common tuning procedures include AIC_c tuning, cross-validation, and restricting the total step size (Zhang and Yu (2005)). When both η and M are set to one, the transition point results on boosting coincide with those on ordinary least squares and ridge regression from Guan *et al.* (2019). A smaller η corresponds to increased shrinkage of the effect estimates and decreased complexity of the boosting fit. For fixed M, decreasing η results in a smaller transition point τ, suggesting that multi-study ensembling would be preferred over merging at a lower threshold of heterogeneity. This can be attributed to the fact that for a fixed M, merging would require a larger η due to the increase in sample size. Because of the interplay between η and M, for a fixed η, decreasing M also leads to a smaller τ. Bühlmann (2006) noted that a smaller η resulted in a weaker learner with reduced variance, and this was empirically shown to be more successful than a strong learner.

We focused on ℓ_2 boosting with linear learners for the opportunity to pursue closed form solutions. With an appropriate choice of basis function, these learners can in theory approximate any sufficiently smooth function to any level of precision (Stone (1948)). In our simulations, the empirical transition points of boosting with ridge learners and boosting with regression trees were similar, suggesting that in certain scenarios it may be reasonable to consider the transition point theory in Theorems 1 and 2 as a proxy when comparing merging and ensembling for boosted trees. It is important to note, however, that such an approximation may not be warranted in settings where the choice of hyperparameters differ from that of our simulations. Although this paper focuses on boosting algorithms, we acknowledge important connections with other machine learning methods. A close relative of boosting with component-wise linear learners is the incremental forward stagewise algorithm (FS), which selects the covariate most correlated (in absolute value) with the residuals $r_{(m)}$ (Efron *et al.* (2004)). Because the covariates are standardized, both algorithms lead to the same variable selection for a given $r_{(m)}$.

A potential limitation of Theorems 1 and 2 is that the tuning parameters (e.g., η and M) are treated as fixed. These quantities are typically chosen by tuning procedures that introduce additional variability. Although we assumed the same η for merging and ensembling in simulations, the transition point τ can be estimated with different values of η, which may be more realistic in practice. For the ensembling approach, we assigned equal weight to each study, which is equivalent to averaging the predictions. The equal-weighting strategy is a special case of stacking (Breiman (1996); Ren *et al.* (2020)) and is preferred in settings where studies have similar sample sizes.

Many areas of biomedical research face a replication crisis in which scientific studies are difficult or impossible to replicate (Ioannidis (2005); National Academies of Sciences *et al.* (2019)). An equally important but less commonly examined issue is the replicability of prediction models. To improve cross-study replicability of predictions, our work provides a theoretical rationale for choosing multi-study ensembling over merging when between-study heterogeneity exceeds a well-defined threshold. As many areas of science are becoming data-rich, it is critical to simultaneously consider and systematically integrate multiple studies to improve cross-study replicability of predictions.

Chapter 3

Multi-Study *R*-Learner for Heterogeneous Treatment Effect Estimation

Joint work with Dr. Boyu Ren, Dr. Prasad Patil, and Dr. Giovanni Parmigiani

Abstract

Flexible estimation of heterogeneous treatment effects is central to precision medicine. While efforts in systematic data sharing and data curation initiatives have increased access to multiple datasets, existing methods for estimating heterogeneous treatment effects are largely rooted in theory based on a single study. We propose a general class of two-step algorithms for treatment effect estimation in multiple studies. The approach is easy to use and allows flexible modeling of the nuisance functions with machine learning techniques. Under the series estimation framework, we show that the resulting estimator is asymptotically normal. We illustrate via simulations and a breast cancer data application that the multi-study R-learner can result in lower estimation error than the R-learner under the presence of between-study heterogeneity.

3.1 Background

Heterogeneous treatment effect estimation is central to many modern statistical applications ranging from precision medicine (Collins and Varmus (2015)) to optimal policy-making (Hitsch and Misra (2018)). In settings where comparable studies are available, it is critical to simultaneously consider and systematically integrate information across multiple studies when estimating treatment effects. Multi-study heterogeneous treatment effect estimation is motivated by applications in biomedical research, where exponential advances in technology and facilitation of systematic data sharing have increased access to multiple studies (Kannan *et al.* (2016); Manzoni *et al.* (2018)). In this work, we introduce a general approach for estimating heterogeneous treatment effects by leveraging information from multiple studies.

Despite increased access to multiple datasets, existing methods on heterogeneous treatment effect estimation are largely rooted in theory based on a single study. These approaches range from inverse probability weighting estimators (Abrevaya *et al.* (2015)) to flexible methods based on random forest (Wager and Athey (2018)), boosting (Powers *et al.* (2018)), and combinations of generic machine learning techniques (Künzel *et al.* (2019)). Recently, Nie and Wager (2021) proposed the R-learner, a general class of two-step algorithms that allows flexible modeling of nuisance functions using machine learning models. The R-learner was motivated by Robinson's transformation, which was originally used to estimate parametric components in partially linear models (Robinson (1988)). Chernozhukov *et al.* (2018) studied these models and proposed an approach that leverages machine learning and sample splitting for estimating treatment effects. Wu and Yang (2021) proposed the integrative R-learner that leverages data from two studies: 1) a randomized clinical trial for identification and 2) an observational study for boosting efficiency when estimating heterogeneous treatment effects. To incorporate information from both sources, the integrative R-learner relies on the assumption that the heterogeneous treatment effect is the same in both studies. This assumption is common in the data integration literature to allow transporting causal inference across studies (Buchanan *et al.* (2018); Dahabreh and Hernán (2019); Dahabreh *et al.* (2019)). When multiple comparable studies are available, heterogeneous treatment effects may differ across studies due to heterogeneity in study design, data collection methods, and sample characteristics. While this issue challenges existing approaches for heterogeneous treatment effect estimation, it also creates opportunities for more general paradigms to account for between-study heterogeneity. In this work, we propose the multi-study R-learner for estimating heterogeneous treatment effects under the presence of between-study heterogeneity. Nie and Wager (2021)'s R-learner is a special case of our approach when there is no between-study heterogeneity. Similar to the R-learner, the multi-study version uses a two-step algorithm that allows flexible modeling of the nuisance components with machine learning methods. In addition to these nuisance components, the multi-study R-learner incorporates the probability of study ascertainment given baseline covariates, allowing strength to be borrowed across studies. We show analytically that optimizing the multi-study R-loss is equivalent to optimizing the oracle loss up to an error that diminishes at a relatively fast rate with the sample size. Under the series estimation framework, we derive a pointwise normality result for the multi-study R-learner estimator. Empirically, we show via simulations and a breast cancer data application that as between-study heterogeneity increases, the multi-study R-learner results in lower estimation error than the R-learner.

3.2 Methods

Suppose we have data from a collection of studies $\mathcal{S}$ indexed by $k = 1, \dots, K$. Study k consists of n_k independent and identically distributed samples $(Y_i, X_i, A_i, S_i = k)$ $(i = 1, \dots, n_k)$, where $Y_i \in \mathbb{R}$ denotes the observed outcome, $X_i \in \mathcal{X} \subset \mathbb{R}^p$ the baseline covariates, and $A_i \in \{0, 1\}$ the treatment assignment. The total number of observations from K studies is $n = \sum_{k=1}^{K} n_k$. We adopt the potential outcomes framework (Rubin (1974)) and let $\{Y_i(1), Y_i(0)\}$ denote the counterfactual outcomes that would have been observed given the treatment assignments $A_i = 1$ and $A_i = 0$, respectively. The heterogeneous treatment effect in study k is characterized by $\tau_k(x) = E[Y_i(1) - Y_i(0) | X_i = x, S_i = k]$. To estimate $\tau_k(x)$, we make the following assumptions:

Assumption 1. $Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i)$ for $i = 1, \ldots, n$.

Assumption 2. $E[Y_i(a)|A_i = a, X_i = x, S_i = k] = E[Y_i(a)|X_i = x, S_i = k]$ for all $k = 1, \ldots, K$ and treatment $a \in \{0, 1\}$.

Assumption 1, commonly referred to as the consistency assumption, states that the observed outcome is equal to the potential outcome under the treatment actually received. *Assumption* 2 posits no unmeasured confounding in the mean response function conditional on the baseline covariates in study $k = 1, \ldots, K$. We rewrite the conditional mean response $m(x) := E[Y_i|X_i = x]$ as a weighted average,

$$m(x) = \sum_{k=1}^{K} m_k(x) p(k|x),$$

where $m_k(x) = E[Y_i|X_i = x, S_i = k]$ denotes the conditional mean response in study k, and $p(k|x) = P(S_i = k|X_i = x)$ the ascertainment probability for study k given baseline covariates x. For study k, we denote the treatment propensity by $e_k(x) = P(A_i = 1|X_i = x, S_i = k)$ and the counterfactual mean response function by $\mu_{k(a)}(x) = E[Y_i(a)|X_i = x, S_i = k]$ for treatment $a \in \{0, 1\}$. Under *Assumptions* 1 and 2, we re-express the conditional mean response in study k as $m_k(x) = \mu_{k(0)}(x) + e_k(x)\tau_k(x)$ and write

$$Y_i - m(X_i) = \sum_{k=1}^{K} \{A_i - e_k(X_i)\} \tau_k(X_i) p(k|X_i) + \epsilon_i, \tag{3.1}$$

where $\epsilon_i := Y_i(A_i) - \sum_{k=1}^{K} \left\{ \mu_{k(0)}(X_i) + A_i \tau_k(X_i) \right\} p(k|X_i)$.

Claim 1. *Under Assumption 2,* $E[\epsilon_i|A_i, X_i] = 0$.

A proof is provided in the appendix. When there is no between-study heterogeneity, i.e., $m(\cdot) = m_k(\cdot)$, $e(\cdot) = e_k(\cdot)$ and $\tau(\cdot) = \tau_k(\cdot) \;\; \forall k$, (3.1) is equal to

$$Y_i - m(X_i) = \{A_i - e(X_i)\} \tau(X_i) + \epsilon_i, \tag{3.2}$$

where $\epsilon_i := Y_i(A_i) - \{\mu_{(0)}(X_i) + A_i \tau(X_i)\}$ and $E[\epsilon_i|A_i, X_i] = 0$. The decomposition in (3.1) extends (3.2) to a multi-study setting ($K > 1$) where $m_k(\cdot)$, $e_k(\cdot)$, $\mu_{k(0)}(\cdot)$, and $\tau_k(\cdot)$ potentially differ across studies $k = 1,\dots,K$. It motivates the mean squared error loss function minimization problem for estimating $\{\tau_1(\cdot),\dots,\tau_K(\cdot)\}$

$$\{\tau_1(\cdot),\dots,\tau_K(\cdot)\} = \underset{\tilde{\tau}_1,\dots,\tilde{\tau}_K}{\arg\min}\left\{E\left(\left[\{Y_i - m(X_i)\} - \sum_{k=1}^{K}\{A_i - e_k(X_i)\}p(k|X_i)\tilde{\tau}_k(X_i)\right]^2\right)\right\}, \tag{3.3}$$

which leads to the empirical minimization,

$$\{\widehat{\tau}_1(\cdot),\dots,\widehat{\tau}_K(\cdot)\} = \underset{\tilde{\tau}_1,\dots,\tilde{\tau}_K}{\arg\min}\Big\{L_n(\{\tau_k(\cdot)\}_{k=1}^{K}) + \Lambda_\tau\Big\}, \tag{3.4}$$

where

$$L_n\Big(\{\tau_k(\cdot)\}_{k=1}^{K}\Big) = \frac{1}{n}\sum_{i=1}^{n}\left[\{Y_i - m(X_i)\} - \sum_{k=1}^{K}\{A_i - e_k(X_i)\}p(k|X_i)\tau_k(X_i)\right]^2, \tag{3.5}$$

is the oracle multi-study R-loss and Λ_τ is a regularizer on the complexity of the $\tau_1(\cdot),\dots,\tau_K(\cdot)$ functions to avoid overfitting. In practice, the optimization in (3.4) may be infeasible because the nuisance functions $m(\cdot), e_k(\cdot)$, and $p(k|\cdot)$ are generally unknown. An exception is that the propensity score function $e_k(\cdot)$ is known in randomized clinical trials (RCTs). Outside of this setting, we estimate the nuisance functions from data and perform the following optimization

$$\{\widehat{\tau}_1(\cdot),\dots,\widehat{\tau}_K(\cdot)\} = \underset{\tilde{\tau}_1,\dots,\tilde{\tau}_K}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\{Y_i - \widehat{m}(X_i)\} - \sum_{k=1}^{K}\{A_i - \widehat{e}_k(X_i)\}\hat{p}(k|X_i)\tilde{\tau}_k(X_i)\right]^2 + \Lambda_\tau\right\}. \tag{3.6}$$

Equation (3.6) is an approximation for optimizing the oracle loss function in (3.5). We use cross-fitting to estimate the nuisance functions and propose a general class of two-step algorithms for treatment effect estimation in multiple studies.

1. We randomly divide the data into Q evenly-sized folds, where Q is typically set to 5 or 10. Let $q(i)$ denote the index set of the fold where the subject i belongs. We use the samples that do not belong to $q(i)$ to fit $\widehat{e}_k$ ($k = 1,\dots,K$), $\widehat{m}$, and $\widehat{p}$. We denote the predictions made without using the data fold that the ith subject belongs to as $\widehat{e}_k^{-q(i)}(X_i)$, $\widehat{m}^{-q(i)}(X_i)$ and $\widehat{p}^{-q(i)}(k|X_i)$.

2. Estimate treatment effects by minimizing $\widehat{L}_n(\{\tau_k(\cdot)\}_{k=1}^K) + \Lambda_\tau$, where

$$\widehat{L}_n\left(\{\tau_k(\cdot)\}_{k=1}^K\right) = \frac{1}{n}\sum_{i=1}^n \left[\{Y_i - \widehat{m}^{-q(i)}(X_i)\} - \sum_{k=1}^K \{A_i - \widehat{e}_k^{-q(i)}(X_i)\}\hat{p}^{-q(i)}(k|X_i)\tilde{\tau}_k(X_i)\right]^2 \tag{3.7}$$

   is the multi-study R-loss.

3.2.1 Theoretical analysis

The goal of our theoretical analysis is two-fold: first, we show that the difference between the oracle multi-study R-loss $L_n\left(\{\tau_k(\cdot)\}_{k=1}^K\right)$ and the plug-in version $\widehat{L}_n\left(\{\tau_k(\cdot)\}_{k=1}^K\right)$ diminishes with a relatively fast rate with n; second, we show the multi-study R-learner is asymptotically normal and unbiased. To achieve this goal, we approximate $\tau_k(\cdot)$ $(k = 1,\ldots,K)$ based on d_k-basis functions, where d_k is allowed to grow with the sample size n_k to balance the trade-off between bias and variance. This is a nonparametric regression method known as series estimation (Wasserman (2006); Belloni *et al.* (2015)). While we rely on series estimation for the opportunity to derive theoretical results, we emphasize that the multi-study R-learner is a general estimation framework for heterogeneous treatment effects. From equation (3.1), we approximate the function

$$x \mapsto g(x) := \sum_{k=1}^K (A_i - e_k(x))\tau_k(x)p(k|x)$$

by linear forms $x \mapsto u(x)^\intercal\beta$, i.e.,

$$g(x) = u(x)^\intercal\beta + r(x), \tag{3.8}$$

where $\beta = (\beta_1^\intercal, \ldots, \beta_K^\intercal)^\intercal$ is a d–dimensional vector of regression coefficients and $\beta_k \in \mathbb{R}^{d_k}$ is the vector of regression coefficients for study k. We let $r(x) := r_g(x) := g(x) - u(x)^\intercal\beta$ denote the approximation error. We assume $g \in \mathcal{G}$ where $\mathcal{G}$ is some class of functions. We let

$$u(x) := W(x)Z(x)v(x),$$

where

$$\begin{aligned}
W(x) &:= \text{blkdiag}(\{W_k(x)\}_{k=1}^K) \in \mathbb{R}^{d\times d}, \quad W_k(x) := \text{diag}((A_i - e_k(x))\mathbb{1}_{d_k}) \in \mathbb{R}^{d_k\times d_k}, \\
Z(x) &:= \text{blkdiag}(\{Z_k(x)\}_{k=1}^K) \in \mathbb{R}^{d\times d}, \quad Z_k(x) := \text{diag}(p(k|x)\mathbb{1}_{d_k}) \in \mathbb{R}^{d_k\times d_k},
\end{aligned}$$

and

$$v(x) := (v_1(x)^\intercal, \dots, v_K(x)^\intercal)^\intercal,$$

where $v_k(x) := (v_{k,1}(x), \dots, v_{k,d_k}(x))^\intercal$ is a vector of approximation functions that can change with n_k. That is, d_k can increase with n_k. We denote the regressors as

$$u_i := u(X_i) = W_i Z_i v_i$$

where

$$\begin{aligned}
W_i &:= W(X_i) \\
Z_i &:= Z(X_i) \\
v_i &:= (v_1(X_i)^\intercal, \dots, v_K(X_i)^\intercal)^\intercal.
\end{aligned}$$

We denote the plug-in multi-study R-loss by

$$\widehat{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^n \Big[\{Y_i - \widehat{m}^{-q(i)}(X_i)\} - \widehat{u}_i^\intercal \beta\Big]^2 \tag{3.9}$$

where

$$\begin{aligned}
\widehat{u_i} &:= \widehat{W_i} Z_i v(X_i), \\
\widehat{W_i} := \text{blkdiag}(\{\widehat{W}_{k,i}\}_{k=1}^K), &\quad \widehat{W}_{k,i} := \text{diag}((A_i - \widehat{e}_k^{-q(i)}(X_i))\mathbb{1}_{d_k}).
\end{aligned}$$

If we know the nuisance functions a priori, the oracle multi-study R-loss is

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^n \Big[\{Y_i - m(X_i)\} - u_i^\intercal \beta\Big]^2. \tag{3.10}$$

To make progress analytically, we assume the following:

Assumption 3. $\|\tau_k(x)\|_\infty$ and $E\Big[(Y_i - m(X_i))^2|X_i = x\Big]$, and $E\Big[(A_i - e_k(X_i))^2|X_i = x\Big]$ are bounded for any $x \in \mathcal{X}$ and $k = 1, \dots, K$.

Assumption 4. $E\left[(m(X_i) - \widehat{m}(X_i))^2\right] = O(a_n^2), E\left[(e_k(X_i) - \widehat{e_k}(X_i))^2\right] = O(a_n^2)$ $(k = 1, \dots, K)$, and $E\left[(e_k(X_i) - \widehat{e_k}(X_i))(e_{k'}(X_i) - \widehat{e_{k'}}(X_i))\right] = O(a_n^2)$ where $k \neq k'$, a_n is some sequence such that $a_n = O(n^{-r})$ with $r > 1/4$.

Assumption 5. Uniformly over all n, eigenvalues of $Q := E[u_i u_i^\intercal]$ are bounded above and away from zero.

Assumption 6. (a) For each n and d, there are finite constants c_d and l_d such that for each $f \in \mathcal{F}$,

$$\left\|r_f\right\|_{F,2} := \sqrt{\int_{x \in \mathcal{X}} r_f^2(x) dF(x)} \leq c_d$$

and

$$\left\|r_f\right\|_{F,\infty} := \sup_{x \in \mathcal{X}} \left|r_f(x)\right| \leq l_d c_d.$$

(b) Uniformly over n,

$$\sup_{x \in \mathcal{X}} E[\epsilon_2^2 I\{|\epsilon_i| > M\}|X_i = x] \to 0$$

as $M \to \infty$

(c)

$$\underline{\sigma}^2 \gtrsim 1 \text{ where } \underline{\sigma}^2 = \inf_{x \in \mathcal{X}} E[\epsilon_i^2 | X_i = x].$$

(d) Let $\xi_D = \sup_x \|u(x)\|$,

$$\xi_d^2 \log(d)/n\left(1 + \sqrt{d} l_d c_d\right) \to 0$$

and

$$l_d c_d \to 0.$$

(e)

$$\sqrt{n/d} \cdot l_d c_d \to 0$$

Assumption 3 and *Assumption 4* ensure the difference between (3.9) and (3.10) diminishes with a relatively fast rate with n. In particular, *Assumption 4* requires the convergence rate of estimators for the nuisance functions to be faster than $n^{-1/4}$. This is plausible because the Neyman orthogonality of the loss function renders the impact of the estimated nuisance functions negligible (Chernozhukov *et al.* (2018)). *Assumption 5-6* are required for the pointwise normality result for the multi-study R-learner estimator (c.f. Theorem 4.2 in Belloni *et al.* (2015)). *Assumption 5* is a regularity condition that ensures the regressors u_i are not too co-linear. In *Assumption 6a*, $\mathcal{F}$ is some class of functions and r_f is r_g with g replaced by f. The finite constants c_d and l_d together characterize the approximation properties of the underlying class of functions. *Assumption 6b* is a mild uniform integrability condition, and it holds if for some $m > 2, \sup_{x \in \mathcal{X}} E[|\epsilon_i|^m | X_i = x] \lesssim 1$. *Assumption 6c* is used to properly normalize the estimator. *Assumption 6d* ensures that the impact of unknown design and approximation error on the sampling error of the estimator is negligible, and *Assumption 6e* ensures the approximation error is negligible relative to the estimation error.

3.3 Results

Lemma 1. *Under Assumptions 3-4,* $\widehat{L}_n(\beta) = L_n(\beta) + O_p(a_n^2)$.

Theorem 1. *Under Assumptions 1-7, for any* $x \in \mathcal{X} \subseteq \mathbb{R}^p$,

$$\sqrt{n}\frac{\widehat{\tau}(x) - \tau(x)}{\|s(x)\|} \xrightarrow{d} N(0,1) + o_P(1)$$

where $s(x) = \Omega^{1/2} Z(x) v(x)$ *and* $\Omega = Q^{-1} E[\epsilon_i^2 u_i u_i^\intercal] Q^{-1}$.

Proofs of the above results are provided in the appendix. Lemma 1 states that the difference between the oracle and plug-in loss functions diminishes with rate a_n^2, where $a_n = O(n^{-r})$ and $r > 1/4$. This implies that $\widehat{\beta} = \arg\min_b \frac{1}{n}\sum_{i=1}^n \{Y_i - m(X_i) - u_i^\intercal b\}^2 + O_p(a_n^2)$. Using this result, Theorem 1 provides the pointwise convergence in distribution result for the multi-study R-learner estimator $\widehat{\tau}(x)$ at any $x \in \mathcal{X}$.

3.3.1 Simulations

We perform simulations to evaluate the performance of the multi-study R-learner. We simulate $K = 4$ studies of sample size $n = 400$ and sample $p = 40$ covariate from the `curatedOvarianData` R package (Ganzfried *et al.* (2013)) to reflect realistic and potentially heterogeneous covariate distributions. We generate

$$\begin{aligned} A_i | X_i, S_i = k &\sim Ber(e_k(X_i)), \quad \epsilon_i | X_i \sim N(0,1), \\ Y_i &= \sum_{k=1}^{K} [\mu_{k(0)}(X_i) + A_i \tau_k(X_i)] p(k|X_i) + \epsilon_i. \end{aligned}$$

We allow the ascertainment probability to depend on x:

$$p(k|X_i) \sim \text{Multinom}\left(n = 400, K = 4, p_k = \frac{exp(X_i \beta_k)}{1 + \sum_{k=1}^{K-1} exp(X_i \beta_k)} \text{ for } k = 2,3,4 \right)$$

where $\sum_{k=1}^{K} p_k = 1$ and $\beta_k \sim \text{MVN}(0, I)$. For $k = 1, \ldots, K$, we generate

$$\mu_{k(0)}(X_i) = X_i \beta_k^{\mu(0)}, \quad \tau_k(X_i) = X_i \beta_k^{\tau} + Z_i \gamma_k^{\tau},$$

where $\beta_k^{\mu(0)} \sim \text{MVN}(0, I)$ and $\beta_k^{\tau} \sim \text{MVN}(2, I)$. $Z_i \in \mathbb{R}^q$ is a subset of X_i that corresponds to the random effects $\gamma_k^{\tau} \in \mathbb{R}^q$. The random effects have $E[\gamma_k^{\tau}] = 0$ and $Cov(\gamma_k^{\tau}) = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$. If $\sigma_j^2 > 0$, then the effect of the jth covariate varies across studies; if $\sigma_j^2 = 0$, then the covariate has the same effect in each study. We assume equal variance, that is, $\sigma_j^2 = \sigma_{\tau}^2$ for $j = 1, \ldots, q$. Each study has eight confounders, and each confounder has a random effect; the regression coefficients for the other 32 covariates are set to 0. We simulate between-study heterogeneity two ways: 1) heterogeneity in the magnitude of the confounder coefficients by varying σ_{τ}^2, and 2) heterogeneity in the degree of overlap in the confounders' support by varying the proportion of overlap p_o. If $p_o = 1$, then all studies share the same eight confounders; if $p_o = 0.5$, then all studies have four common confounders, and each study has four study-specific ones; if $p_o = 0$, then the studies do not share any confounders. In Scenario I, the studies are randomized trials, so $e_k(x) = 0.5$ for all x. In Scenario II, the studies are observational, and we generate $e_k(X_i) = \text{expit}(X_i \beta_k^e)$, where $\beta_k^e \sim MVN(1, I)$. The values of

$\beta_k, \beta_k^{\mu(0)}, \beta_k^e$, and β_k^τ used in the simulations are provided in the appendix. We estimate the nuisance functions $m(\cdot), e_k(\cdot)$, and $p(k|x)$ with elastic net (and logistic/multinomial elastic net).

Figure 3.1 shows the $\log_2$ mean squared error (MSE) ratio comparing multi-study R-learner to the R-learner. The top panels correspond to Scenario I and the bottom Scenario II. Under the oracle setting, we optimize the loss function in (3.9) with the true nuisance functions (left panel); in practice, the oracle nuisance functions are typically unknown (especially $m(\cdot)$), so we estimate them with elastic net and optimize the loss function in (3.6) (right panel). Overall, as p_o decreases, the multi-study R-learner outperforms the R-learner. Under Scenario II's oracle setting, the multi-study R-learner has lower MSE across all levels of p_o, with improvements of at least 84% when there is complete overlap in the confounders. Within each level of p_o, the multi-study R-learner shows favorable performance as σ_τ increases, suggesting that the multi-study R-learner is preferred when between-study heterogeneity is high.

3.4 Data Application to Breast Cancer

We illustrate the multi-study R-learner by applying it to breast cancer data from the `curatedBreastData` R package (Planey *et al.* (2015)). Female breast cancer is a molecularly heterogeneous disease consisting of four main subtypes: 1) HR+/HER2-, 2) HR-/HER2-, 3) HR+/HER2+, and 4) HR-/HER2+ (Hwang *et al.* (2019)). HR+ means that tumor cells have hormone receptors (HR) for estrogen or progesterone, which promote the growth of HR+ turmors; on the other hand, HR- means that tumor cells do not have those receptors. HER2+ means that tumor cells produce high levels of the protein HER2 (human epidermal growth factor receptor 2), which has been shown to be associated with aggressive tumor behavior (Slamon *et al.* (1987)); likewise, HER2- means that tumor cells do not produce high levels of HER2. On a molecular level, different breast cancers behave and proliferate in various ways. Therefore, it is important to characterize and understand treatment effect heterogeneity for patients with different breast cancer subtypes.

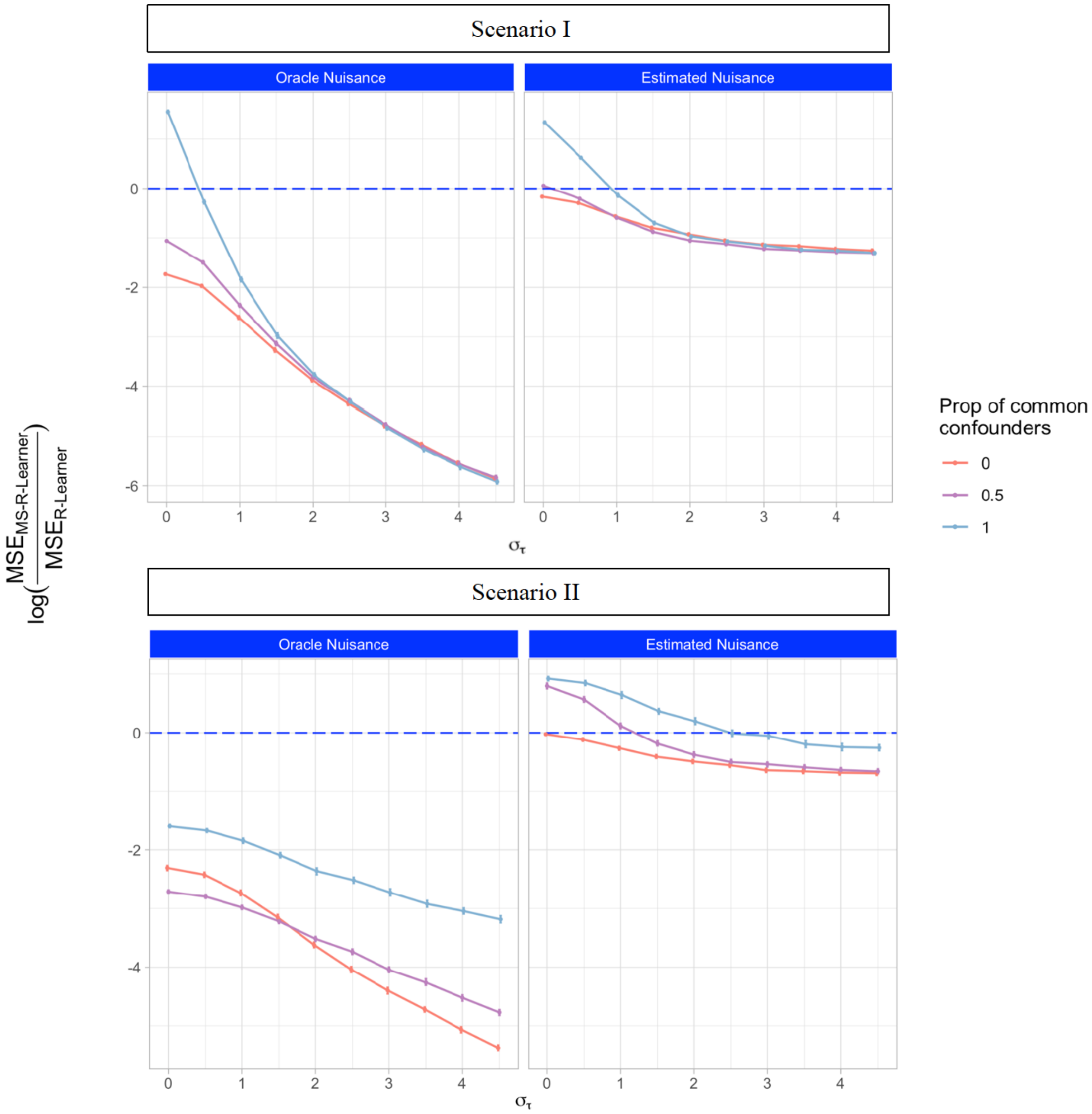


Figure (3.1) *Performance of multi-study R-learner vs. R-learner*

$\log_2$ *mean squared error ratio comparing multi-study R-learner (MS-R-Learner) to R-learner. Scenario I focuses on randomized trials, and Scenario II focuses on observational studies. The solid circles represent the average performance ratios, and vertical bars the 95% bootstrapped intervals. The differently colored lines correspond to varying proportion of overlap in confounders:* $p_o = 0$ *(red),* $p_o = 0.5$ *(purple), and* $p_o = 1$ *(blue).*

Chemotherapy is a common treatment option for breast cancer. Generally, practitioners may recommend chemotherapy in two situations: before or after surgery. First, neoadjuvant chemotherapy can be used to reduce the size or extent of breast cancer before surgery. Its purpose is to downstage the extent of disease in the breast and/or regional lymph nodes and provide information regarding treatment response to direct adjuvant therapies (Sikov *et al.* (2020)). Second, adjuvant chemotherapy can be used to try to kill any cancer cells that might have been left behind or have spread but can't be seen on imaging tests. While both chemotherapies carry the same risks and side effects, patients who don't respond to neoadjuvant chemotherapy run the additional risk of having delayed their main treatment. Therefore, the purpose of our data application is to characterize the treatment effect heterogeneity of anthracyline/taxane (A/T), a neoadjuvant chemotherapy regimen for early breast cancer.

The outcome of interest is pathological complete response (pCR), which is defined as disappearance of all invasive cancer in the breast after completion of neoadjuvant chemotherapy (1 = responded, 0 = otherwise). We identified $K = 2$ studies where patients were administered A/T neoadjuvant chemotherapy ($1 = A$, $0 = T$). The first study (GSE21997) is a randomized trial of $n_1 = 94$ women aged between 18 and 79 with stage II-III breast cancer (Martin *et al.* (2011)). Patients were assigned to receive four cycles of either A or T before surgery. The second study (GSE25065) is an observational study of $n_2 = 168$ women who were HER2- with stage I-III breast cancer (Hatzis *et al.* (2011)). We focus on $p = 4$ covariates: age, histology grade (1-3), HR+ (1 = yes, 0 = 0), HER2+ (1 = yes, 0 = no). Table 3.1 summarizes the treatment and covariate information. We fit a logistic regression model to the data and use the estimated probabilities as the ascertainment probability. In particular, $P(S_i = 2|age_i, HR+_i, HER2+_i) = \text{expit}(2.308 - 0.0309 age_i + 0.223(HR+_i) - 3.456(HER2+_i))$. A challenge with illustrating heterogeneous treatment effect estimators on real data is that we do not have access to both counterfactuals. Therefore, we generate study-specific treatment effects to make the task of estimating heterogeneous treatment effects non-trivial. We generate $\tau_1(X_i) = 1 - age_i/100 - 0.5(HR+) - 0.5(HER2+)$ and $\tau_2(X_i) = 1 - age_i/100 - 0.5(HR+)$ for studies 1 and 2, respectively. If $\tau(X_i) > 0$, then we set $\{Y_i(1), Y_i(0)\}$ to (0, 1); otherwise, we set it to $(1, 0)$. Finally, we set $Y_i = Y_i(A_i)$. Figure 3.2 summarizes the heterogeneous treatment effect generating mechanism. Women who are HR-/HER2- have triple-negative breast cancer (TNBC), which is an aggressive subtype of breast cancer that tends to have a worse prognosis. While T is often used to treat TNBC, patients usually develop resistance to it (Gómez-Miragaya *et al.* (2017); Maloney *et al.* (2020)). Therefore, we generated the outcomes so that women with TBNC would not respond to T. In a similar vein, anthracycline-induced cardiotoxicity may interfere with treatment response (Cardinale *et al.* (2020)), so women who were either HR+/HER2- or HR-/HER2+ would not respond to A if they were over the age of 50 (Figure 3.2).

Table (3.1) *Descriptive statistics of breast cancer data*

	n	$S = 1$	$S = 2$
		$N = 94$	$N = 168$
Histologic Grade	262		
1		3% (3)	5% (9)
2		55% (52)	34% (57)
3		41% (39)	61% (102)
HER2+	262		
0		72% (68)	99% (166)
1		28% (26)	1% (2)
Treatment	262		
T		43% (40)	50% (84)
A		57% (54)	50% (84)
HR+	262		
0		35% (33)	31% (52)
1		65% (61)	69% (116)
Age	262		
≤ 50 Years		50% (47)	56% (94)
> 50 Years		50% (47)	44% (74)

We randomly divide the data into a training ($n_{\text{train}} = 162$) and a test set ($n_{\text{test}} = 100$). To implement the R-learner and multi-study R-learner, we estimate $\widehat{m}(\cdot)$ and $\widehat{e}(\cdot)$ from the

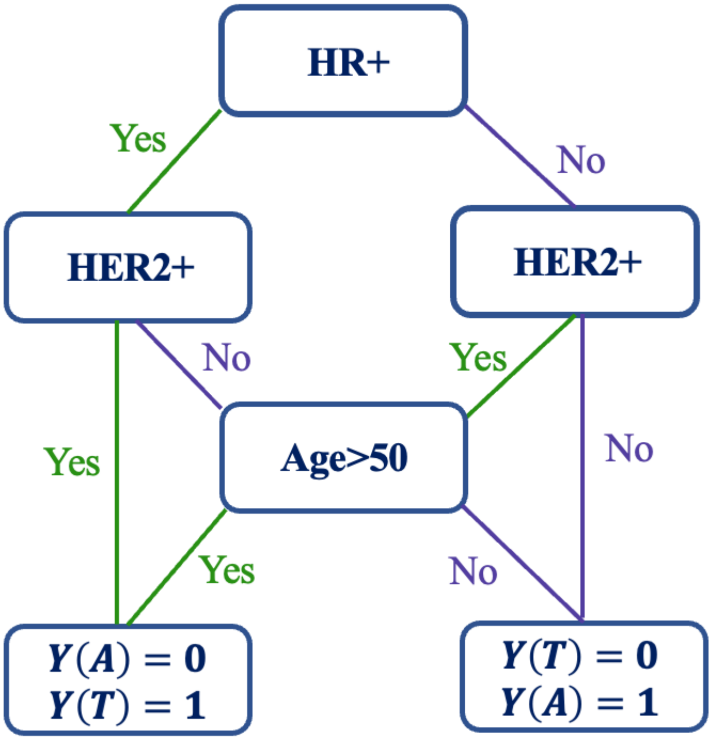


Figure (3.2) *Outcome-generating mechanism for the breast cancer data illustration.*

training data using lasso with tuning parameters selected by cross-validation. Because study 1 was a randomized trial, we use 0.5 for its propensity score, and we estimate $\widehat{e_2}$ using training data from study 2. Next, we optimized the R-loss and multi-study R-loss functions to estimate the treatment effect and calculated the log mean square error on the test set. Figure 3.3 compares the true treatment effects $\tau(X_i)$ with estimated treatment effects $\widehat{\tau}(X_i)$ for the R-learner and multi-study R-learner. Both approaches perform well overall but tend to underestimate $\tau(X_i)$ in the $[-1,-0.5]$ range. This can be attributed to the small sample size $(n = 7)$ of women who are HR+/HER+. To simulate between-study heterogeneity, we add normally-distributed random effects with mean 0 and covariance $\sigma_\tau I$ to the coefficients for HR+ and HER2+. Figure 3.4 shows the $\log_2$ mean squared error ratio comparing lasso-based multi-study R-learner to R-learner on the test set. When between-study heterogeneity (σ_τ) increases, the multi-study R-learner demonstrates preferable performance.

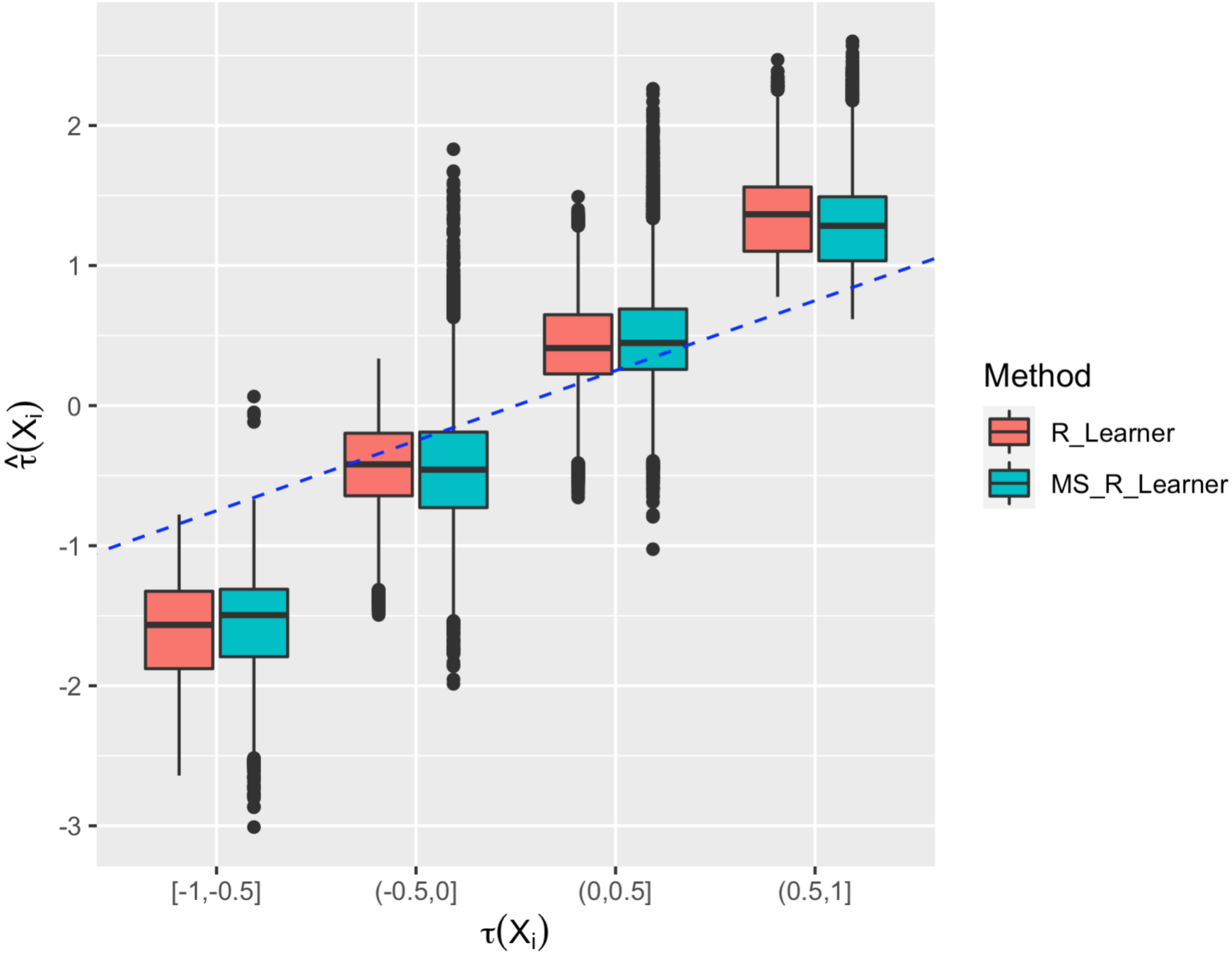


Figure (3.3) *Heterogeneous treatment effect estimates $\hat{\tau}(X_i)$ on the test set*

The treatment effect estimates are obtained from lasso-based R-learner (red boxplots) and multi-study R-learner (green boxplots) compared to the true $\tau(X_i)$.

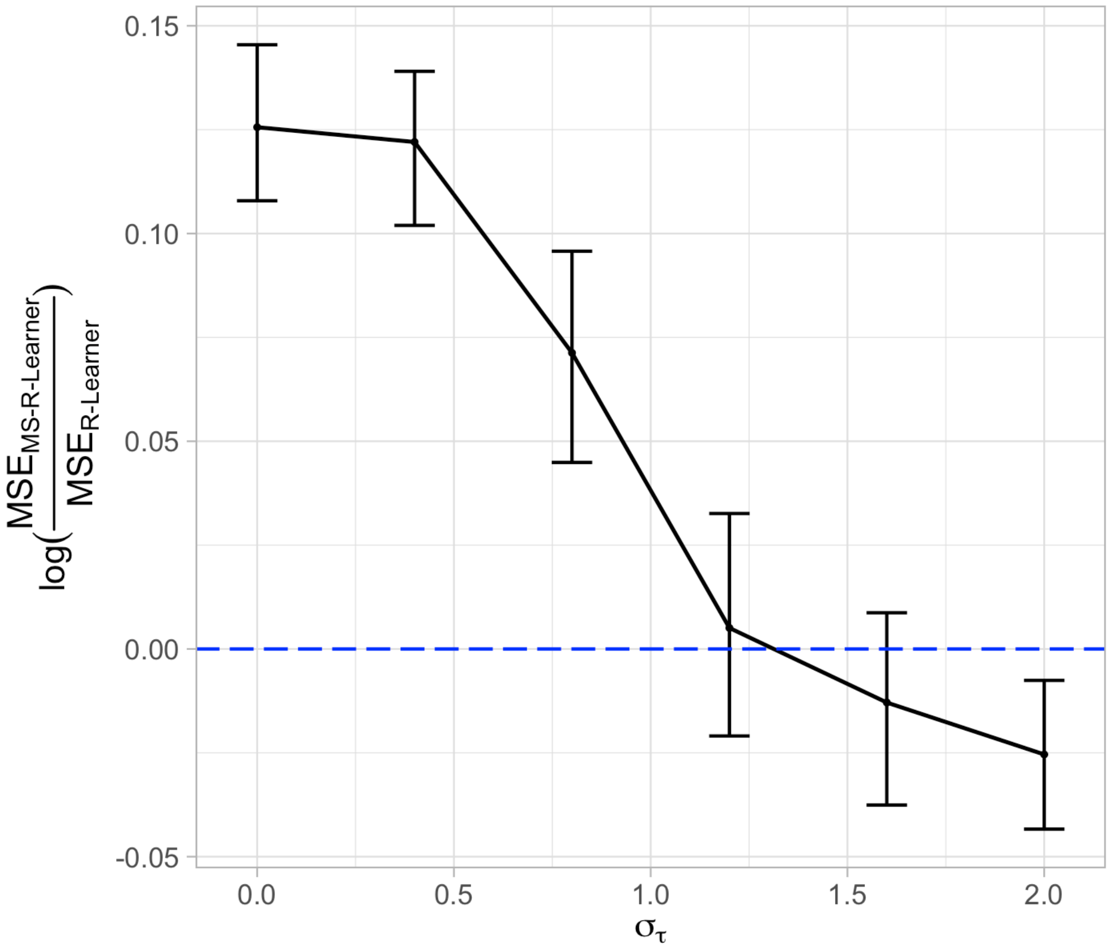


Figure (3.4) *Performance of multi-study R-learner vs. R-learner on breast cancer data*

$\log_2$ *mean squared error ratio comparing lasso-based multi-study R-learner (MS-R-Learner) to R-learner in the breast cancer data application. The solid circles represent the average performance ratios, and vertical bars the 95% bootstrapped intervals.*

3.5 Discussion

We propose the multi-study R-learner for estimating heterogeneous treatment effects under the presence of between-study heterogeneity. It is a general approach that is easy to implement in practice and allows flexible modeling of the nuisance functions using machine learning techniques. From the perspective of multi-study learning, our approach can be seen as a unifying framework for estimating heterogeneous treatment effects that admits two special cases. First, when there is no between-study heterogeneity, the multi-study R-learner is equivalent to the R-learner. Second, if the study designs are so different such that $p(k|\cdot) = 1$ for all individuals in study k $\forall k$, then optimizing the multi-study R-loss is equivalent to optimizing the R-loss on each study separately. An example of this is when a study's exclusion criteria matches the inclusion criteria of another study completely (e.g., one study recruits participants of ages ≤ 18, whereas another recruits those > 18).

It would be interesting to extend the current framework to accommodate multi-valued treatments. In this case, the multi-study R-learner estimator for $t > 2$ treatment levels can be constructed based on the multivariate version of Robin's transformation as suggested by Nie and Wager (2021). Moreover, it would be interesting to extend the multi-study R-learner to handle other types of outcomes, e.g. binary, multinomial, and survival.

In many areas of biomedical research, exponential advances in technology and facilitation of systematic data-sharing have led to increased access to multiple studies. To account for potential between-study heterogeneity in personalized treatment effects, our approach extends the R-learner to a multi-study setting and allows flexible modeling of the nuisance components. As many areas of science are becoming data-rich, it is critical to simultaneously consider and systematically integrate multiple studies when estimating heterogeneous treatment effects.

References

(). SEER Incidence Data, 1973-2015. https://seer.cancer.gov/data/, accessed: 2017-03-20.

Aaltonen, L., Johns, L., Järvinen, H., Mecklin, J.-P. and Houlston, R. (2007). Explaining the familial colorectal cancer risk associated with mismatch repair (mmr)-deficient and mmr-stable tumors. *Clinical Cancer Research*, **13** (1), 356–361.

Abrevaya, J., Hsu, Y.-C. and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, **33** (4), 485–505.

Baglietto, L., Lindor, N. M., Dowty, J. G., White, D. M., Wagner, A., Gomez Garcia, E. B., Vriends, A. H., Group, D. L. S. S., Cartwright, N. R., Barnetson, R. A. *et al.* (2010). Risks of lynch syndrome cancers for msh6 mutation carriers. *Journal of the National Cancer Institute*, **102** (3), 193–201.

Bao, Y., Deng, Z., Wang, Y., Kim, H., Armengol, V. D., Acevedo, F., Ouardaoui, N., Wang, C., Parmigiani, G., Barzilay, R., Braun, D. and Hughes, K. S. (2019). Using machine learning and natural language processing to review and classify the medical literature on cancer susceptibility genes.

Barrow, E., Robinson, L., Alduaij, W., Shenton, A., Clancy, T., Lalloo, F., Hill, J. and Evans, D. (2009). Cumulative lifetime incidence of extracolonic cancers in lynch syndrome: a report of 121 families with proven mutations. *Clinical genetics*, **75** (2), 141–149.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, **186** (2), 345–366.

Bellot, A. and van der Schaar, M. (2019). Boosting transfer learning with survival data from heterogeneous domains. In *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 57–65.

Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L. and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, **30** (12), i105–i112.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, vol. 6. Athena Scientific Belmont, MA.

Bonadona, V., Bonaïti, B., Olschwang, S., Grandjouan, S., Huiart, L., Longy, M., Guimbaud, R., Buecher, B., Bignon, Y.-J., Caron, O. *et al.* (2011). Cancer risks associated with germline mutations in mlh1, msh2, and msh6 genes in lynch syndrome. *Jama*, **305** (22), 2304–2310.

Borràs, E., Pineda, M., Blanco, I., Jewett, E. M., Wang, F., Teulé, À., Caldés, T., Urioste, M., Martínez-Bouzas, C., Brunet, J. *et al.* (2010). Mlh1 founder mutations with moderate penetrance in spanish lynch syndrome families. *Cancer research*, pp. 0008–5472.

Braun, D., Yang, J., Griffin, M., Parmigiani, G. and Hughes, K. S. (2018). A clinical decision support tool to predict cancer risk for commonly tested cancer-related germline mutations. *Journal of genetic counseling*, pp. 1–13.

Breiman, L. (1996). Stacked regressions. *Machine learning*, **24** (1), 49–64.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J. and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181** (4), 1193–1209.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, **34** (2), 559–583.

Bühlmann, P., Hothorn, T. *et al.* (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical science*, **22** (4), 477–505.

— and Yu, B. (2003). Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, **98** (462), 324–339.

Burn, J., Gerdes, A.-M., Macrae, F., Mecklin, J.-P., Moeslein, G., Olschwang, S., Eccles, D., Evans, D. G., Maher, E. R., Bertario, L. *et al.* (2011). Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the capp2 randomised controlled trial. *The Lancet*, **378** (9809), 2081–2087.

Buttin, B. M., Powell, M. A., Mutch, D. G., Babb, S. A., Huettner, P. C., Edmonston, T. B., Herzog, T. J., Rader, J. S., Gibb, R. K., Whelan, A. J. *et al.* (2004). Penetrance and expressivity of msh6 germline mutations in seven kindreds not ascertained by family history. *The American Journal of Human Genetics*, **74** (6), 1262–1269.

Cardinale, D., Iacopo, F. and Cipolla, C. M. (2020). Cardiotoxicity of anthracyclines. *Frontiers in cardiovascular medicine*, **7**, 26.

Castaldi, P. J., Dahabreh, I. J. and Ioannidis, J. P. (2011). An empirical assessment of validation practices for molecular classifiers. *Briefings in bioinformatics*, **12** (3), 189–202.

Chen, S., Wang, W., Lee, S., Nafa, K., Lee, J., Romans, K., Watson, P., Gruber, S. B., Euhus, D., Kinzler, K. W. *et al.* (2006). Prediction of germline mutations and cancer risk in the lynch syndrome. *Jama*, **296** (12), 1479–1487.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Choi, Y.-H., Cotterchio, M., McKeown-Eyssen, G., Neerav, M., Bapat, B., Boyd, K., Gallinger, S., McLaughlin, J., Aronson, M. and Briollais, L. (2009). Penetrance of colorectal cancer among mlh1/msh2 carriers participating in the colorectal cancer familial registry in ontario. *Hereditary cancer in clinical practice*, **7** (1), 14.

Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England journal of medicine*, **372** (9), 793–795.

Dahabreh, I. J. and Hernán, M. A. (2019). Extending inferences from a randomized trial to a target population. *European journal of epidemiology*, **34** (8), 719–722.

—, Robertson, S. E., Petito, L. C., Hernán, M. A. and Steingrimsson, J. A. (2019). Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *arXiv preprint arXiv:1908.09230*.

Dai Wenyuan, Y. Q., Guirong, X. *et al.* (2007). Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning, Corvallis, USA*, pp. 193–200.

de la Chapelle, A. (2005). The incidence of lynch syndrome. *Familial cancer*, **4** (3), 233–237.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, **7** (3), 177–188.

Dowty, J. G., Win, A. K., Buchanan, D. D., Lindor, N. M., Macrae, F. A., Clendenning, M., Antill, Y. C., Thibodeau, S. N., Casey, G., Gallinger, S. *et al.* (2013). Cancer risks for mlh1 and msh2 mutation carriers. *Human mutation*, **34** (3), 490–497.

Dunlop, M. G., Farrington, S. M., Carothers, A. D., Wyllie, A. H., Sharp, L., Burn, J., Liu, B., Kinzler, K. W. and Vogelstein, B. (1997). Cancer risk associated with germline dna mismatch repair gene mutations. *Human molecular genetics*, **6** (1), 105–110.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, **32** (2), 407–499.

Egger, M., Smith, G. D., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, **315** (7109), 629–634.

Fleming, I. D. (1997). Ajcc cancer staging manual. *American Joint Committee on Cancer*.

Freund, R. M., Grigas, P. and Mazumder, R. (2017). A new perspective on boosting in linear regression via subgradient optimization and relatives. *The Annals of Statistics*, **45** (6), 2328–2364.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and computation*, **121** (2), 256–285.

— and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, **55** (1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G. *et al.* (2013). curatedovariandata: clinically annotated data for the ovarian cancer transcriptome. *Database*, **2013**.

Giardiello, F. M., Allen, J. I., Axilbund, J. E., Boland, C. R., Burke, C. A., Burt, R. W., Church, J. M., Dominitz, J. A., Johnson, D. A., Kaltenbach, T. *et al.* (2014). Guidelines on genetic evaluation and management of lynch syndrome: a consensus statement by the us multi-society task force on colorectal cancer. *Gastrointestinal Endoscopy*, **80** (2), 197–220.

Gómez-Miragaya, J., Palafox, M., Paré, L., Yoldi, G., Ferrer, I., Vila, S., Galván, P., Pellegrini, P., Pérez-Montoyo, H., Igea, A. *et al.* (2017). Resistance to taxanes in triple-negative breast cancer associates with the dynamics of a cd49f+ tumor-initiating population. *Stem cell reports*, **8** (5), 1392–1407.

Gradishar, W. J., Moran, M. S., Abraham, J., Aft, R., Agnese, D., Allison, K. H., Blair, S. L., Burstein, H. J., Dang, C., Elias, A. D. *et al.* (2021). Nccn guidelines® insights: Breast cancer, version 4.2021: Featured updates to the nccn guidelines. *Journal of the National Comprehensive Cancer Network*, **19** (5), 484–493.

Guan, Z., Parmigiani, G. and Patil, P. (2019). Merging versus ensembling in multi-study machine learning: Theoretical insight from random effects. *arXiv preprint arXiv:1905.07382*.

Guindalini, R. S. C., Win, A. K., Gulden, C., Lindor, N. M., Newcomb, P. A., Haile, R. W., Raymond, V., Stoffel, E., Hall, M., Llor, X. *et al.* (2015). Mutation spectrum and risk of colorectal cancer in african american families with lynch syndrome. *Gastroenterology*, **149** (6), 1446–1453.

Habrard, A., Peyrache, J.-P. and Sebban, M. (2013). Boosting for unsupervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 433–448.

Hampel, H., Pearlman, R., Beightol, M., Zhao, W., Jones, D., Frankel, W. L., Goodfellow, P. J., Yilmaz, A., Miller, K., Bacher, J. *et al.* (2018). Assessment of tumor sequencing as a replacement for lynch syndrome screening and current molecular tests for patients with colorectal cancer. *JAMA oncology*, **4** (6), 806–813.

—, Stephens, J. A., Pukkala, E., Sankila, R., Aaltonen, L. A., Mecklin, J.-P. and de la Chapelle, A. (2005). Cancer risk in hereditary nonpolyposis colorectal cancer syndrome: later age of onset. *Gastroenterology*, **129** (2), 415–421.

Hatzis, C., Pusztai, L., Valero, V., Booser, D. J., Esserman, L., Lluch, A., Vidaurre, T., Holmes, F., Souchon, E., Wang, H. *et al.* (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *Jama*, **305** (18), 1873–1881.

Hitsch, G. J. and Misra, S. (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*.

Hwang, K.-T., Kim, J., Jung, J., Chang, J. H., Chai, Y. J., Oh, S. W., Oh, S., Kim, Y. A., Park, S. B. and Hwang, K. R. (2019). Impact of breast cancer subtypes on prognosis of women with operable invasive breast cancer: a population-based study using seer database. *Clinical Cancer Research*, **25** (6), 1970–1979.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, **2** (8), e124.

Järvinen, H. J., Mecklin, J.-P. and Sistonen, P. (1995). Screening reduces colorectal cancer rate in families with hereditary nonpolyposis colorectal cancer. *Gastroenterology*, **108** (5), 1405–1411.

Jass, J. R. (2006). Hereditary non-polyposis colorectal cancer: the rise and fall of a confusing term. *World journal of gastroenterology: WJG*, **12** (31), 4943.

Jenkins, M. A., Baglietto, L., Dowty, J. G., Van Vliet, C. M., Smith, L., Mead, L. J., Macrae, F. A., John, D. J. B. S., Jass, J. R., Giles, G. G. *et al.* (2006). Cancer risks for mismatch repair gene mutation carriers: a population-based early onset case-family study. *Clinical gastroenterology and hepatology*, **4** (4), 489–498.

—, Dowty, J. G., Ait Ouakrim, D., Mathews, J. D., Hopper, J. L., Drouet, Y., Lasset, C., Bonadona, V. and Win, A. K. (2014). Short-term risk of colorectal cancer in individuals with lynch syndrome: a meta-analysis. *Journal of Clinical Oncology*, **33** (4), 326–331.

Kannan, L., Ramos, M., Re, A., El-Hachem, N., Safikhani, Z., Gendoo, D. M., Davis, S., Gomez-Cabrero, D., Castelo, R., Hansen, K. D. *et al.* (2016). Public data and open source tools for multi-assay genomic investigation of disease. *Briefings in bioinformatics*, **17** (4), 603–615.

Kastrinos, F., Idos, G. and Parmigiani, G. (2018). Prediction Models for Lynch Syndrome. In L. Valle, S. B. Gruber and G. Capella (eds.), *Hereditary Colorectal Cancer: Genetic Basis and Clinical Implications*, Cham: Springer International Publishing, pp. 281–303.

Kopciuk, K. A., Choi, Y.-H., Parkhomenko, E., Parfrey, P., McLaughlin, J., Green, J. and Briollais, L. (2009). Penetrance of hnpcc-related cancers in a retrolective cohort of 12 large newfoundland families carrying a msh2 founder mutation: an evaluation using modified segregation models. *Hereditary cancer in clinical practice*, **7** (1), 16.

Kraft, P. and Thomas, D. C. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. **66** (3), 1119–1131.

Künzel, S. R., Sekhon, J. S., Bickel, P. J. and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, **116** (10), 4156–4165.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E. *et al.* (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, **44** (3), 907–927.

Locker, G. Y. and Lynch, H. T. (2004). Genetic factors and colorectal cancer in ashkenazi jews. *Familial cancer*, **3** (2), 215–221.

Maloney, S. M., Hoover, C. A., Morejon-Lasso, L. V. and Prosperi, J. R. (2020). Mechanisms of taxane resistance. *Cancers*, **12** (11), 3323.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A. and Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, **19** (2), 286–302.

Marabelli, M., Cheng, S.-C. and Parmigiani, G. (2016). Penetrance of atm gene mutations in breast cancer: A meta-analysis of different measures of risk. *Genetic epidemiology*, **40** (5), 425–431.

Martin, M., Romero, A., Cheang, M., López García-Asenjo, J. A., García-Saenz, J. A., Oliva, B., Román, J. M., He, X., Casado, A., De La Torre, J. *et al.* (2011). Genomic predictors of response to doxorubicin versus docetaxel in primary breast cancer. *Breast cancer research and treatment*, **128** (1), 127–136.

Møller, P., Seppälä, T., Bernstein, I., Holinski-Feder, E., Sala, P., Evans, D. G., Lindblom, A., Macrae, F., Blanco, I., Sijmons, R. *et al.* (2017). Cancer incidence and survival in lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective lynch syndrome database. *Gut*, **66** (3), 464–472.

Mukherjee, B., Rennert, G., Ahn, J., Dishon, S., Lejbkowicz, F., Rennert, H. S., Shiovitz, S., Moreno, V. and Gruber, S. B. (2011). High risk of colorectal and endometrial cancer in ashkenazi families with the msh2 a636p founder mutation. *Gastroenterology*, **140** (7), 1919–1926.

National Academies of Sciences, E., Medicine *et al.* (2019). *Reproducibility and replicability in science*. National Academies Press.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, **108** (2), 299–319.

Pande, M., Lynch, P. M., Hopper, J. L., Jenkins, M. A., Gallinger, S., Haile, R. W., LeMarchand, L., Lindor, N. M., Campbell, P. T., Newcomb, P. A. *et al.* (2010). Smoking and colorectal cancer in lynch syndrome: results from the colon cancer family registry and the university of texas md anderson cancer center. *Clinical cancer research*, **16** (4), 1331–1339.

Pardoe, D. and Stone, P. (2010). Boosting for regression transfer. In *ICML*.

Patil, P. and Parmigiani, G. (2018). Training replicable predictors in multiple studies. *Proceedings of the National Academy of Sciences*, **115** (11), 2578–2583.

Planey, K. (2020). *curatedBreastData: Curated breast cancer gene expression data with survival and treatment information*. R package version 2.18.0.

—, Planey, M. K., biocViews ExperimentData, E. and TissueMicroarrayData, G. (2015). Package ‘curatedbreastdata’.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T. and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, **37** (11), 1767–1787.

Provenzale, D., Gupta, S., Ahnen, D. J., Bray, T., Cannon, J. A., Cooper, G., David, D. S., Early, D. S., Erwin, D., Ford, J. M. *et al.* (2016). Genetic/familial high-risk assessment: colorectal version 1.2016, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, **14** (8), 1010–1030.

Quehenberger, F., Vasen, H. and Van Houwelingen, H. (2005). Risk of colorectal and endometrial cancer for carriers of mutations of the hmlh1 and hmsh2 gene: correction for ascertainment. *Journal of medical genetics*, **42** (6), 491–496.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramchandran, M., Patil, P. and Parmigiani, G. (2020). Tree-weighting for multi-study ensemble learners. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, NIH Public Access, vol. 25, p. 451.

Ren, B., Patil, P., Dominici, F., Parmigiani, G. and Trippa, L. (2020). Cross-study learning for generalist and specialist predictions. *arXiv preprint arXiv:2007.12807*.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, **66** (5), 688.

Rügamer, D. and Greven, S. (2020). Inference for l2-boosting. *Statistics and Computing*, **30** (2), 279–289.

Rustgi, A. K. (2007). The genetics of hereditary colon cancer. *Genes & development*, **21** (20), 2525–2538.

Sanne, W., Brohet, R. M., Tops, C. M., van der Klift, H. M., Velthuizen, M. E., Bernstein, I., Munar, G. C., Garcia, E. G., Hoogerbrugge, N., Letteboer, T. G. *et al.* (2014). Lynch syndrome caused by germline pms2 mutations: delineating the cancer risk. *J Clin Oncol*, **33**, 319–325.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S. *et al.* (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, **26** (5), 1651–1686.

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, **7** (3), 40–45.

Sikov, W. M., Boughey, F. J. C., Al-Hilli, Z. and Chen, W. (2020). General principles of neoadjuvant management of breast cancer.

Sjursen, W., Haukanes, B. I., Grindedal, E. M., Aarset, H., Stormorken, A., Engebretsen, L. F., Jonsrud, C., Bjørnevoll, I., Andresen, P. A., Ariansen, S. *et al.* (2010). Current clinical criteria for lynch syndrome are not sensitive enough to identify msh6 mutation carriers. *Journal of medical genetics*, pp. jmg–2010.

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A. and McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *science*, **235** (4785), 177–182.

Stoffel, E., Mukherjee, B., Raymond, V. M., Tayob, N., Kastrinos, F., Sparr, J., Wang, F., Bandipalliam, P., Syngal, S. and Gruber, S. B. (2009). Calculation of risk of colorectal and endometrial cancer among patients with lynch syndrome. *Gastroenterology*, **137** (5), 1621–1627.

Stone, M. H. (1948). The generalized weierstrass approximation theorem. *Mathematics Magazine*, **21** (5), 237–254.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, **111** (514), 600–620.

Trippa, L., Waldron, L., Huttenhower, C., Parmigiani, G. *et al.* (2015). Bayesian nonparametric cross-study validation of prediction methods. *Annals of Applied Statistics*, **9** (1), 402–428.

Tutz, G. and Binder, H. (2007). Boosting ridge regression. *Computational Statistics & Data Analysis*, **51** (12), 6044–6059.

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., Chapelle, A. d. l., Rüschoff, J., Fishel, R., Lindor, N. M., Burgart, L. J., Hamelin, R. *et al.* (2004). Revised bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, **96** (4), 261–268.

Vasen, H., Stormorken, A., Menko, F., Nagengast, F., Kleibeuker, J., Griffioen, G., Taal, B., Moller, P. and Wijnen, J. (2001). Msh2 mutation carriers are at higher risk of cancer than mlh1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families. *Journal of Clinical Oncology*, **19** (20), 4074–4080.

Vasen, H. F., Möslein, G., Alonso, A., Bernstein, I., Bertario, L., Blanco, I., Burn, J., Capella, G., Engel, C., Frayling, I. *et al.* (2007). European guidelines for the clinical management of lynch syndrome (hnpcc). *Journal of medical genetics*.

—, Watson, P., Mecklin, J.-P., Lynch, H. T. *et al.* (1999). New clinical criteria for hereditary nonpolyposis colorectal cancer (hnpcc, lynch syndrome) proposed by the international collaborative group on hnpcc. *Gastroenterology*, **116** (6), 1453–1456.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, **113** (523), 1228–1242.

Wang, Z. and Wang, C. (2010). Buckley-james boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*, **9** (1).

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watson, P., Ashwathnarayan, R., Lynch, H. T. and Roy, H. K. (2004). Tobacco use and increased colorectal cancer risk in patients with hereditary nonpolyposis colorectal cancer (lynch syndrome). *Archives of internal medicine*, **164** (22), 2429–2431.

Win, A., Dowty, J., English, D., Campbell, P., Young, J., Winship, I., Macrae, F., Lipton, L., Parry, S., Young, G. *et al.* (2011). Body mass index in early adulthood and colorectal cancer risk for carriers and non-carriers of germline mutations in dna mismatch repair genes. *British journal of cancer*, **105** (1), 162.

Win, A. K., Young, J. P., Lindor, N. M., Tucker, K. M., Ahnen, D. J., Young, G. P., Buchanan, D. D., Clendenning, M., Giles, G. G., Winship, I. *et al.* (2012). Colorectal and other cancer risks for carriers and noncarriers from families with a dna mismatch repair gene mutation: a prospective cohort study. *Journal of Clinical Oncology*, **30** (9), 958.

Wu, L. and Yang, S. (2021). Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *First Conference on Causal Learning and Reasoning*.

Yao, Y. and Doretto, G. (2010). Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*, IEEE, pp. 1855–1862.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, **33** (4), 1538–1579.

Appendix A

Appendix to Chapter 1

A.1 DerSimonian and Laird's Random Effects Model

Motivated by the heterogeneity in study population, design, and ascertainment mechanism, we combined decade-specific risk estimates from the published studies using the DerSimonian and Laird random-effects model, which considers both within- and between-study variation DerSimonian and Laird (1986). We weighted the study-specific risk estimates by the inverse of their variance to calculate a summary estimate and its 95% confidence interval. For studies that reported sex- or gene-aggregated penetrance estimates without specifying the number of sex- or gene-specific carriers, we weighted these estimates by the ratio of male:female penetrance in the general population (based on the SEER registry) or the ratio of MLH1:MSH2:MSH6 penetrance estimates from Bonadona *et al.* Bonadona *et al.* (2011), respectively. We chose Bonadona *et al.* because among the studies that reported penetrance estimates for all 3 MMR genes of interest, its estimates were derived from the largest carrier sample spanning over 7 decades.

A.2 Meta-Analytic Approach

The general model allowing the meta-analytic integration of different types of cancer risk estimates has been described previously Marabelli *et al.* (2016). Let C denote mutation carrier status where $C = 1$ represents carriers and $C = 0$ non-carriers. Suppose there are $k = 1,\ldots,K$ studies included in the meta-analysis.

A.2.1 Studies Reporting Penetrance

In our meta-analysis, all 10 papers reported penetrance. For studies $k = 1,\ldots,10$, we assume the time to cancer T_k follows a probability distribution function F, characterized by n unknown parameters of interest (where n is determined based on the distributional assumptions for F), $\boldsymbol{\theta_c} = (\theta_{c1},...,\theta_{cn})$, such that;

$$P(\text{cancer by age } a|c) = P(T_k \leq a|c) = F_c(a|\theta_{c1},\ldots,\theta_{cn}) \tag{A.1}$$

In study k, we let

$$\mu_{kc}(a) \equiv F_c(a|\theta_{c1},\ldots,\theta_{cn})$$

denote the probability of cancer by age a given carrier status and

$$\boldsymbol{\mu_{kc}(a)} \equiv F_c(\boldsymbol{a}|\theta_{c1},...,\theta_{cn})$$

denote the probability of cancer for a vector of ages $\boldsymbol{a} = (a_1,...,a_m)$. For study-reported cumulative penetrance values $\boldsymbol{X^{\text{Pen}}_{kc}(a)} = (X^{\text{Pen}}_{kc}(a_1), X^{\text{Pen}}_{kc}(a_2),..., X^{\text{Pen}}_{kc}(a_m))$, we assume $\boldsymbol{X^{\text{Pen}}_{kc}(a)}$ follows a multivariate normal (MVN) distribution centered at $\boldsymbol{\mu_{kc}}$ with corresponding covariance matrix $\hat{\Sigma}^{\text{Pen}}_{\boldsymbol{k}}$, i.e. $\boldsymbol{X^{\text{Pen}}_{kc}(a)} \sim \text{MVN}(\boldsymbol{\mu_{kc}}, \hat{\Sigma}^{\text{Pen}}_{\boldsymbol{k}})$. In practice, studies generally do not report the covariance matrix $\hat{\Sigma}_k$ $(k = 1,\ldots,K)$. To estimate $\hat{\Sigma}_k$, we simulate penetrance values from a normal distribution with means and standard deviations provided by the original paper. These values are simulated under the constraint that the penetrance values were non-decreasing as age increases. The covariance matrix is estimated between the penetrance values for each pair of ages.

Assuming that the studies are independent, the overall likelihood can be written as the product of the study-specific contributions:

$$L_{\text{Overall}}(\boldsymbol{\theta}_c) = \prod_{k=1}^{10} L_k(\boldsymbol{\theta}_c | \boldsymbol{X}_{kc}^{\text{Pen}}) \tag{A.2}$$

where the study-specific likelihood $L_k(\boldsymbol{\theta}_c | \boldsymbol{X}_{kc}^{\text{Pen}})$ can be written in terms of the sampling distribution of $\boldsymbol{X}_{\boldsymbol{kc}}^{\text{Pen}}$. The parameter estimates $\hat{\boldsymbol{\theta}}_c$ are obtained by maximizing the overall likelihood.

This likelihood is then extended to include gene- and/or sex- aggregated risk information. We assume that the cancer penetrance follows a distribution function F_{cgs}, where $g \in \{1,2,3\}$ denotes the three Lynch syndrome genes (MLH1, MSH2, and MSH6, respectively), and $s \in \{1,2\}$ denotes male and female, respectively. Depending on the study, $\boldsymbol{\mu_{kc}}$ can be written as one of the following:

- Study k reports gene- and sex- specific penetrance estimates

$$\boldsymbol{\mu_{kc}}(\boldsymbol{a}) = F_{cgs}(\boldsymbol{a} | \theta_{cgs1}, \ldots, \theta_{cgsn})$$

- Study k reports gene-specific but sex-aggregated penetrance estimates

$$\boldsymbol{\mu_{kc}}(\boldsymbol{a}) = F_{cg}(\boldsymbol{a} | \theta_{cg1}, \ldots, \theta_{cgn})$$

  where $F_{cg} = p_{g1}F_{cg1} + (1 - p_{g1})F_{cg2}$ and p_{g1} is the proportion of male carriers.

- Study k reports sex-specific but gene-aggregated penetrance estimates

$$\boldsymbol{\mu_{kc}}(\boldsymbol{a}) = F_{cs}(\boldsymbol{a} | \theta_{cs1}, \ldots, \theta_{csn})$$

  where $F_{cs} = p_{1s}F_{c1s} + p_{2s}F_{c2s} + p_{3s}F_{c3s}$ and p_{1s}, p_{2s}, p_{3s} are the proportions of MLH1, MSH2, and MSH6 mutation carriers, respectively.

- Study k reports both gene- and sex-aggregated penetrance estimates

$$\boldsymbol{\mu_{kc}}(\boldsymbol{a}) = F_c(\boldsymbol{a} | \theta_{c1}, \ldots, \theta_{cn})$$

  where $F_c = p_{11}F_{c11} + p_{12}F_{c12} + p_{21}F_{c21} + p_{22}F_{c22} + p_{31}F_{c31} + p_{32}F_{c32}$ and p_{11} is the proportion of male MLH1 mutation carriers, p_{12} is the proportion of female MLH1 mutation carriers, and so forth.

We assume that the cancer penetrance can be modeled as a log-logistic distribution function, i.e., $\boldsymbol{\mu_{kc}(a)} = F_{cgs}(\boldsymbol{a};\lambda_{gs},\kappa_{gs}) = \frac{1}{1+\frac{a}{\lambda_{gs}}^{(-\kappa_{gs})}}$. The unknown vector of parameters, $\theta_{\mathbf{c}} = \{\lambda_{\mathbf{gs}},\kappa_{\mathbf{gs}}\}$ has a total of 2 (shape and scale parameters) for each of the 3 (genes) × 2 (sexes), resulting in 12 parameters. For study k, the study-specific likelihood is

$$L_c(\lambda_{gs},\kappa_{gs}|\boldsymbol{X}^{\text{Pen}}_{kc}) = \frac{1}{(2\pi)^{\frac{m}{2}} det(\hat{\boldsymbol{\Sigma}}^{\text{Pen}}_{\boldsymbol{kc}})^{\frac{1}{2}}} exp\Big(-\frac{1}{2}(\boldsymbol{x}^{\text{Pen}}_{\boldsymbol{kc}} - \boldsymbol{\mu}_{\boldsymbol{k_c}})^T \hat{\boldsymbol{\Sigma}}^{\text{Pen}-1}_{\boldsymbol{kc}}(\boldsymbol{x}^{\text{Pen}}_{\boldsymbol{kc}} - \boldsymbol{\mu}_{\boldsymbol{kc}})\Big).$$

Maximum likelihood estimates of the parameters of the log-logistic distribution function $(\lambda_{\mathbf{gs}},\kappa_{\mathbf{gs}})$ were estimated to be: $\hat{\lambda}_{11} = 3.98, \hat{\kappa}_{11} = 0.61, \hat{\lambda}_{12} = 4.09, \hat{\kappa}_{12} = 0.50, \hat{\lambda}_{21} = 3.74, \hat{\kappa}_{21} = 0.56, \hat{\lambda}_{22} = 4.10, \hat{\kappa}_{22} = 0.58, \hat{\lambda}_{31} = 4.44, \hat{\kappa}_{31} = 0.31, \hat{\lambda}_{32} = 4.81$, and $\hat{\kappa}_{32} = 0.28$. To obtain measures of uncertainty, we calculated 95% credible intervals based on posterior probability distributions.

Fig. S1 shows the following: (1) the means and 95% CI of the meta-analytic penetrances at each 10-year age interval that were estimated using the DerSimonian and Laird method; (2) the smoothed curves obtained from the likelihood-based approach that represent our final estimates by age. Fig. S2 shows the cumulative penetrance of CRC for MLH1, MSH2, and MSH6 mutation carriers after stratifying studies by screening status.

A.2.2 Studies Reporting Relative Risk

The relative risk (RR) X^{RR}_k is assumed to be normally distributed with mean μ^{RR}_k and variance $\hat{\sigma}^{\text{RR}^2}_k$. The variance is provided by the study. We re-express the mean as a function of the penetrance Marabelli *et al.* (2016):

$$\mu^{\text{RR}}_k \equiv \frac{P(\text{cancer}|C=1)}{P(\text{cancer}|C=0)} \approx \frac{\int P(\text{cancer by age } a|C=1)g_1(a)da}{\int P(\text{cancer by age } a|C=0)g_0(a)da} = \frac{\int \mu_{k1}(a)g_{k1}(a)da}{\int \mu_{k0}(a)g_{k0}(a)da}, \qquad \text{(A.3)}$$

where $g_{kc}(a)$ denotes the density of age a given carrier status c in study k. In practice, when g_c is not available, we can use the density of age of onset q_{kc} instead:

$$P(\text{cancer}|c) \approx \int P(\text{cancer at age } a|c)q_c(a)da = \int f_{kc}(a|\boldsymbol{\theta}_c)q_{kc}(a)da.$$

A.2.3 Studies Reporting Odds Ratio

The relative risk (OR) X_k^{OR} is assumed to be normally distributed with mean μ_k^{OR} and variance $\sigma_k^{\mathrm{OR}^2}$. The variance is provided by the study. We re-express the mean as a function of the penetrance Marabelli *et al.* (2016):

$$\mu_k^{\mathrm{OR}} \equiv \frac{\frac{P(\mathrm{C}=1|\mathrm{cancer})}{1-P(\mathrm{C}=1|\mathrm{cancer})}}{\frac{P(\mathrm{C}=1|\mathrm{no\ cancer})}{1-P(\mathrm{C}=1|\mathrm{no\ cancer})}} = \frac{\frac{P(\mathrm{C}=1)P(\mathrm{cancer}|\mathrm{C}=1)}{P(\mathrm{C}=0)P(\mathrm{cancer}|\mathrm{C}=0)}}{\frac{P(\mathrm{C}=1)P(\mathrm{no\ cancer}|\mathrm{C}=1)}{P(\mathrm{C}=0)P(\mathrm{no\ cancer}|\mathrm{C}=0)}} \approx \frac{\frac{\int f_{k1}(a)r_{k1}(a)da}{\int f_{k0}(a)r_{k0}(a)da}}{\frac{\int(1-\mu_{k1}(a))s_{k1}(a)da}{\int(1-\mu_{k0}(a))s_{k0}(a)da}}, \tag{A.4}$$

where given carrier status c, r_{kc} and s_{kc} denote the density of age of onset among cases and the density of age of inclusion among controls, respectively.

A.2.4 Studies Reporting Standardized Incidence Ratio

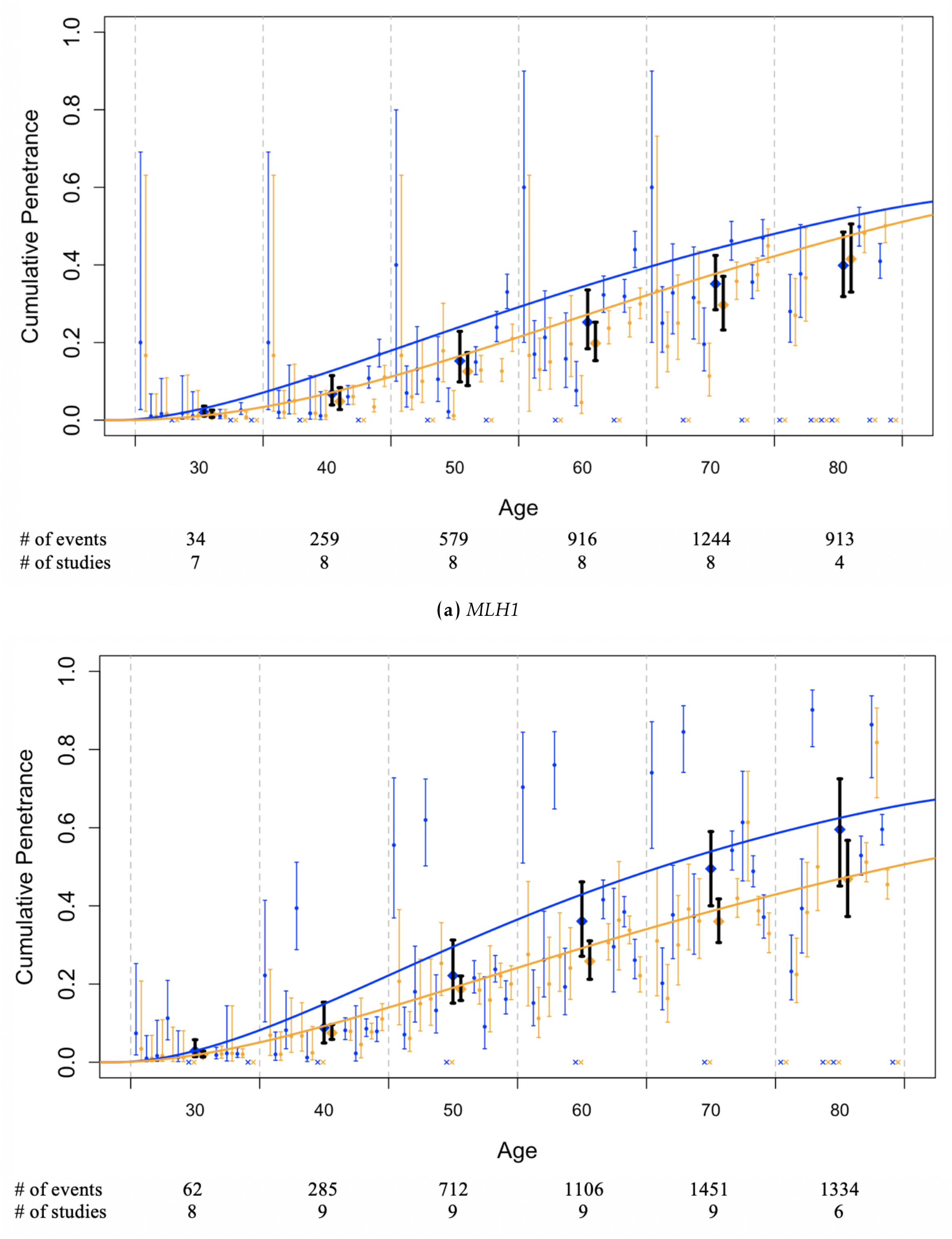
The standardized incidence ratio (SIR) X_k^{SIR} is assumed to be normally distributed with mean μ_k^{SIR} and variance $\sigma_k^{\mathrm{SIR}^2}$. The variance is provided by the study. We re-express the mean as a function of the penetrance Marabelli *et al.* (2016):

$$\mu_k^{\mathrm{SIR}} \approx \frac{\int f_{k1}(a)q_{k1}(a)da}{P(C=1)\int f_{k1}(a)q_{k1}(a)da + P(C=0)\int f_{k0}(a)q_{k0}(a)da}. \tag{A.5}$$

A.3 Publication bias

We created funnel plots to assess the potential issue of publication bias in our meta-analysis. Fig. A.3 shows the corresponding funnel plots for studies on MLH1, MSH2, and MSH6 mutation carriers by 10-year age intervals from age 30 to 80. For female MLH1 mutation carriers, there is some evidence of asymmetry in the funnel plots at ages 60 (p-value = 0.021) and 70 (p-value = 0.043) (Fig. a bottom row). For female MSH2 mutation carriers, there is marginal evidence of asymmetry at age 40 (p-value = 0.041) (Fig.b bottom row). The number of studies is too small to test for evidence of asymmetry for MSH6 mutation carriers. All p-values associated with male mutation carriers for MLH1 and MSH2 were not significant at the 0.05 level. Due to the small number of studies across all three genes (8 on MLH1, 9 on

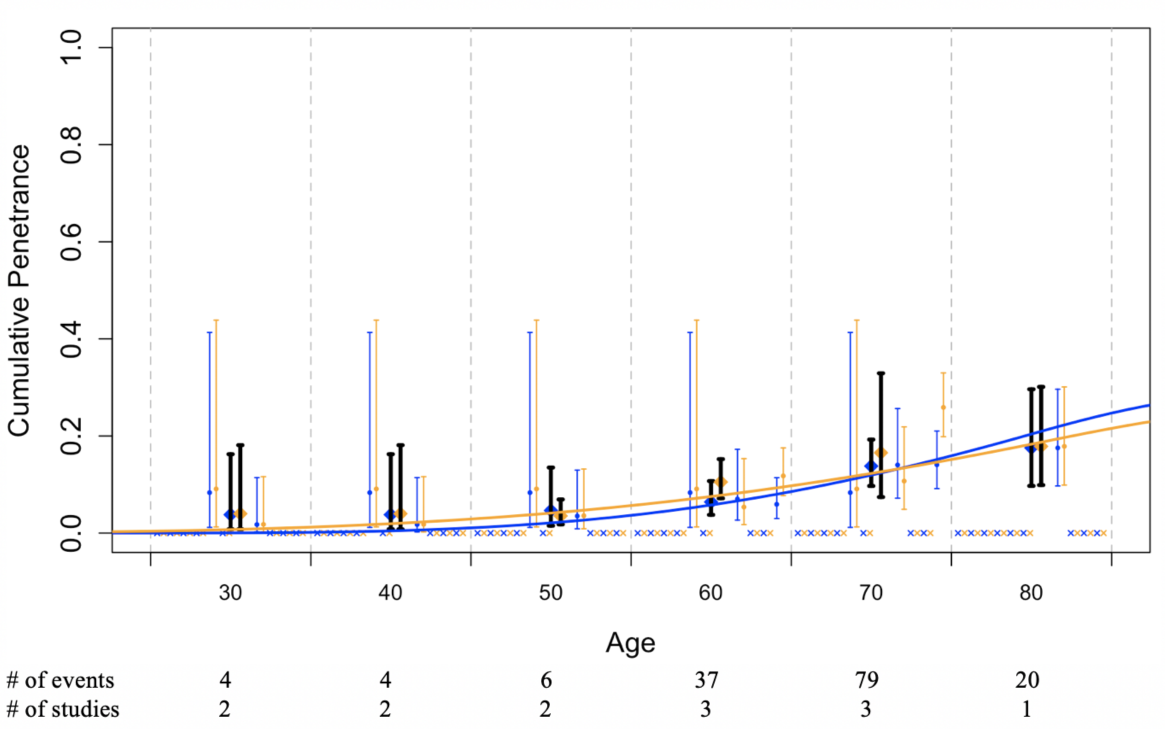
MSH2, 3 on MSH6), the Egger test may be underpowered to distinguish chance from real asymmetry.


Cumulative Penetrance
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Age
# of events
34
259
579
916
1244
913
# of studies
7
8
8
8
8
4
Cumulative Penetrance
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Age
# of events
62
285
712
1106
1451
1334
# of studies
8
9
9
9
9
6

(**a**) *MLH1*

(**b**) *MSH2*

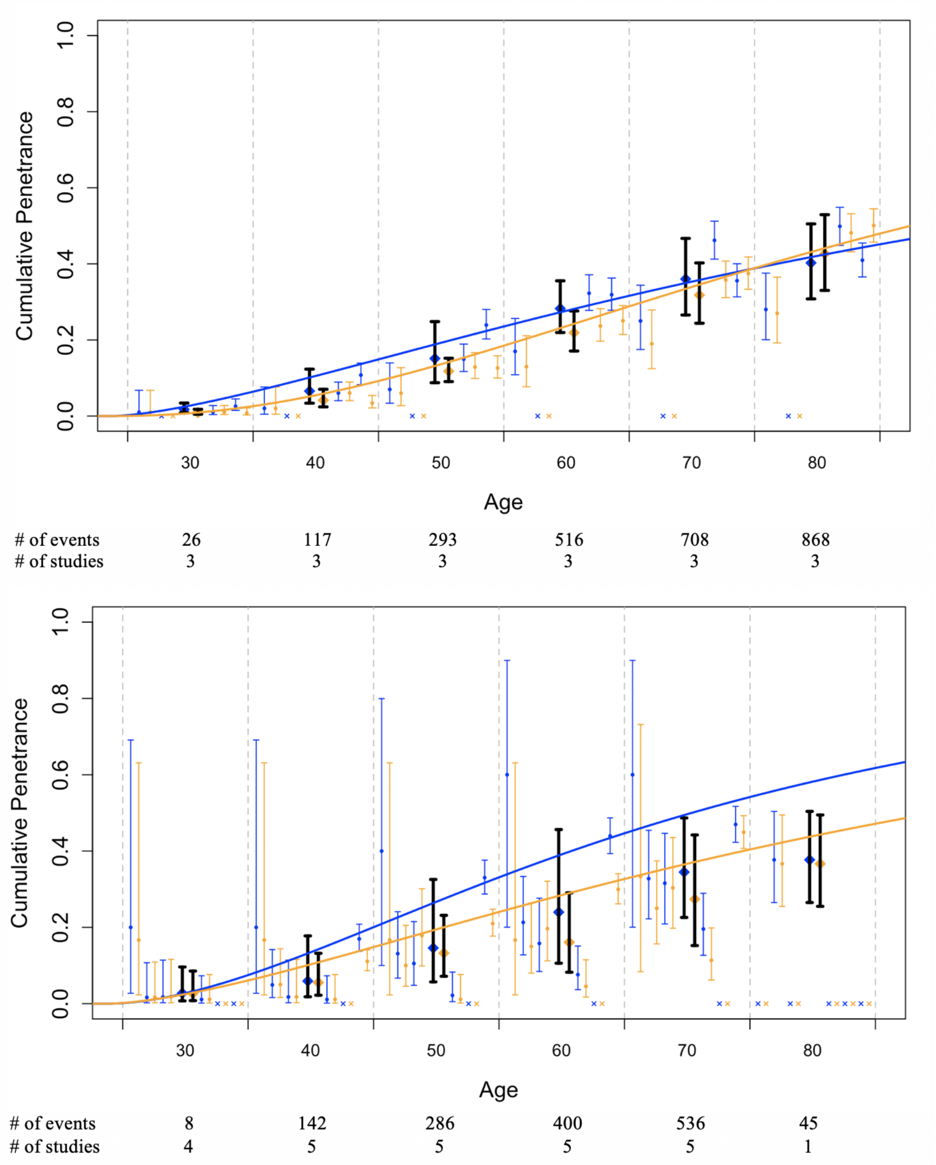
(Continued)



(c) *MSH6*

Figure (A.1) *Age-specific colorectal cancer risk for mismatch repair gene mutation carriers by study*

Panels (a), (b), and (c) correspond to MLH1, MSH2, and MSH6 mutation carriers, respectively. ***DerSimonian and Laird random effects model results:*** *The age range is divided into 10-year intervals. Within each we show cumulative risk estimates from individual studies (thin vertical blue/orange bars) and the meta-analytic estimate from the DL random effects model (thick vertical black bars). The height of vertical bars represents 95% CIs. Within each 10-year age interval, the published studies are arranged by publication year (left to right):* Dunlop et al. *Dunlop et al. (1997),* Quehenberger et al. *Quehenberger et al. (2005),* Aaltonen et al. *Aaltonen et al. (2007),* Kopciuk et al. *Kopciuk et al. (2009),* Stoffel et al. *Stoffel et al. (2009),* Borras et al. *Borràs et al. (2010),* Bonadona et al. *Bonadona et al. (2011),* Mukherjee et al. *Mukherjee et al. (2011),* Dowty et al. *Dowty et al. (2013), and* Moller et al. *Møller et al. (2017). An "x" indicates that the age interval was not available.* ***Likelihood-based approach results:*** *Smooth blue and orange lines represent penetrance estimated from the likelihood-based approach by yearly age. Blue corresponds to male carriers, and orange corresponds to female carriers.*

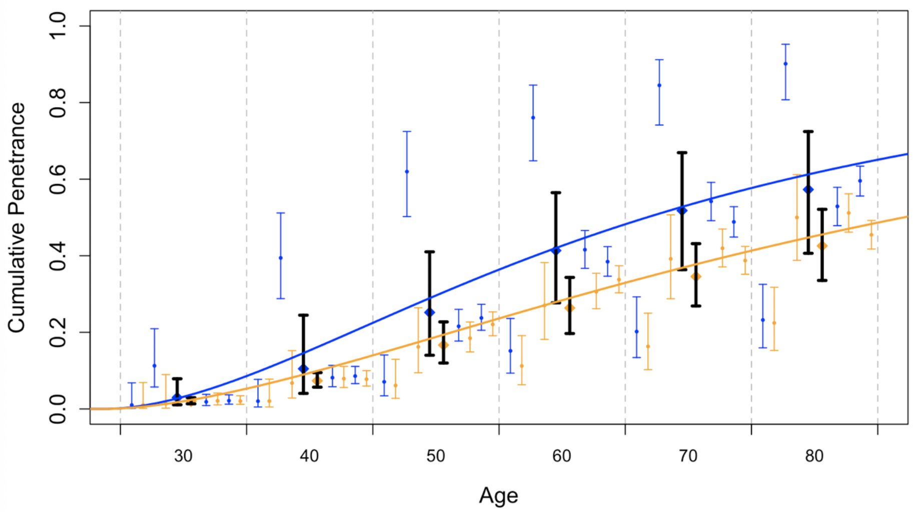


# of events	26	117	293	516	708	868
# of studies	3	3	3	3	3	3

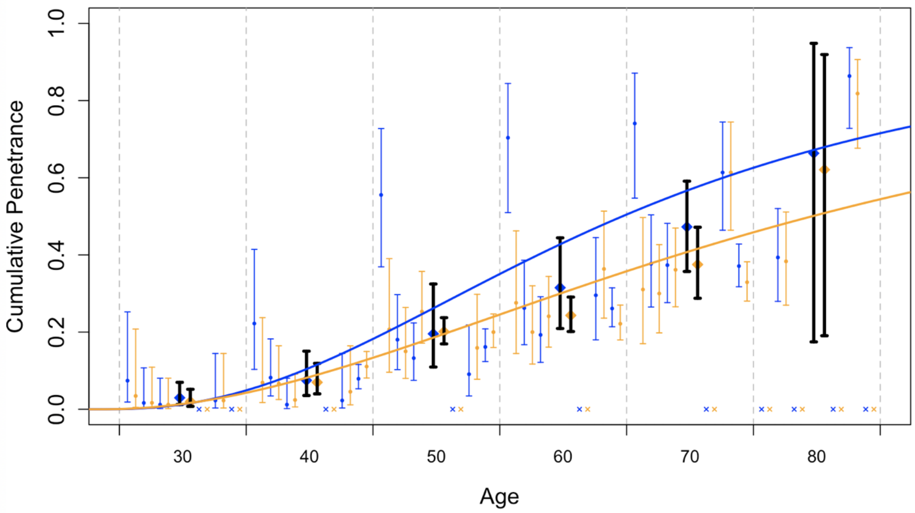
# of events	8	142	286	400	536	45
# of studies	4	5	5	5	5	1

(**a**) *MLH1*

(Continued)



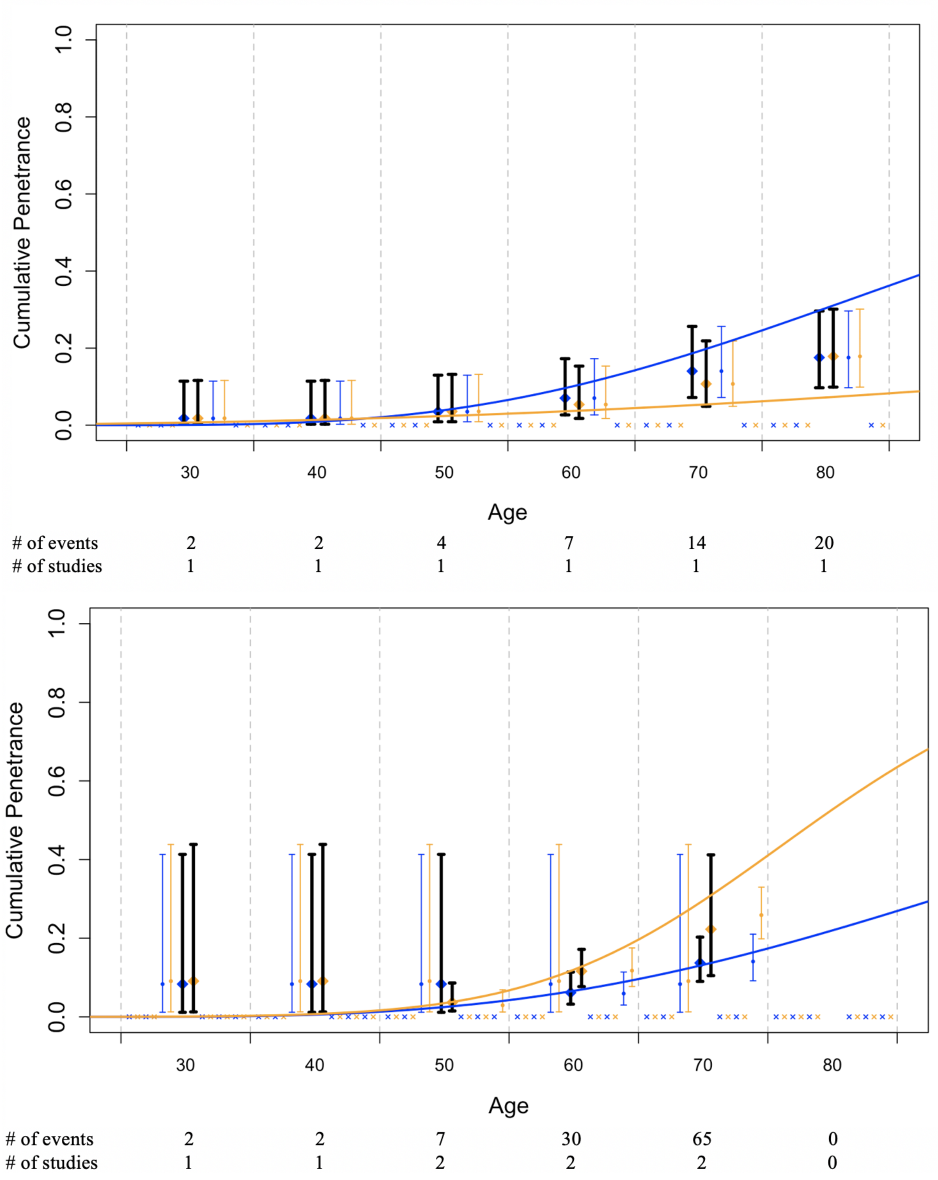
# of events	60	226	548	872	1082	1240
# of studies	4	4	4	4	4	4



# of events	9	82	196	268	400	62
# of studies	4	5	5	5	5	2

(**b**) *MSH2*

(Continued)



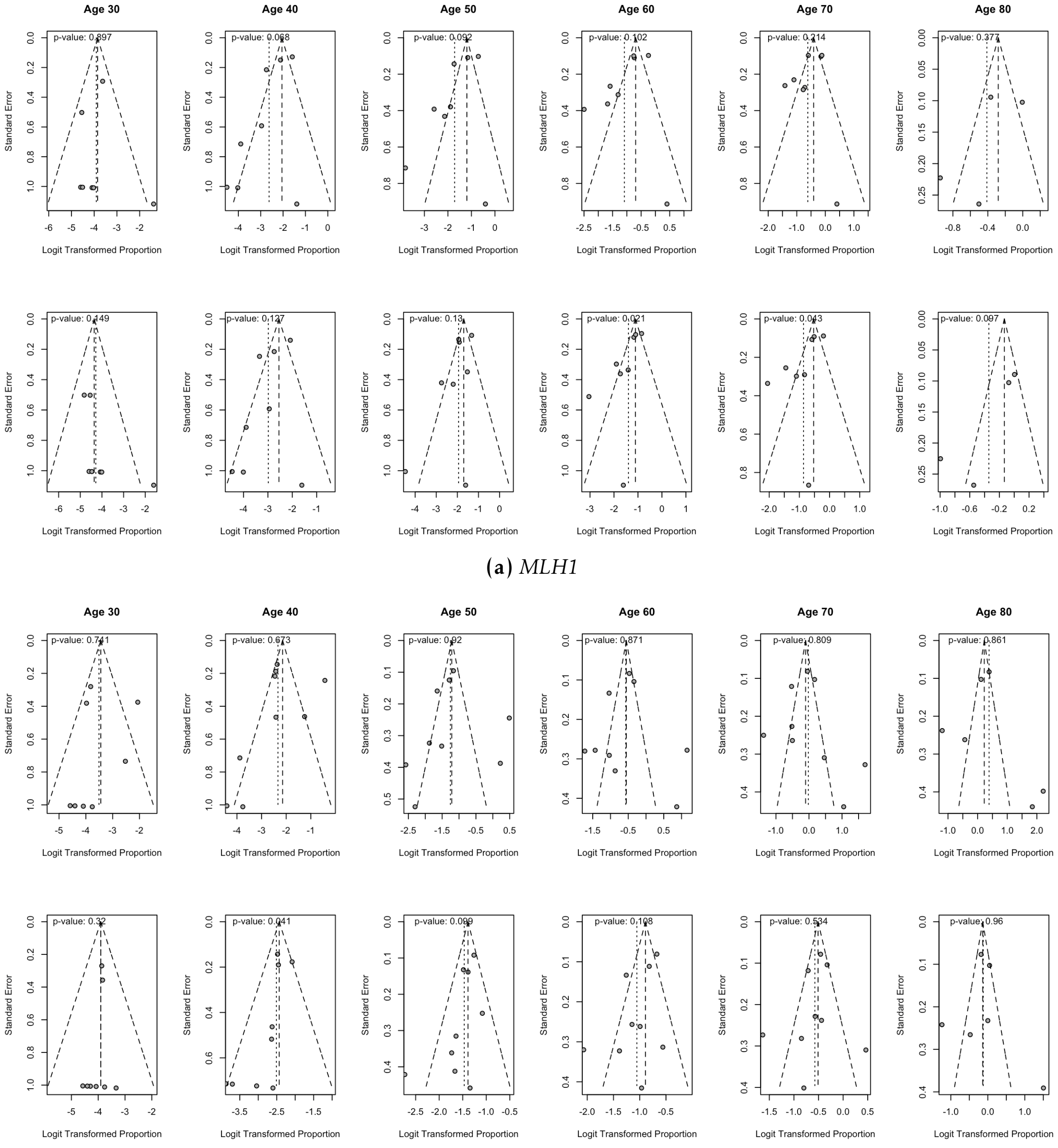
# of events	2	2	4	7	14	20
# of studies	1	1	1	1	1	1

# of events	2	2	7	30	65	0
# of studies	1	1	2	2	2	0

(c) *MSH6*

Figure (A.2) *Colorectal cancer risk stratified by screening status and study*

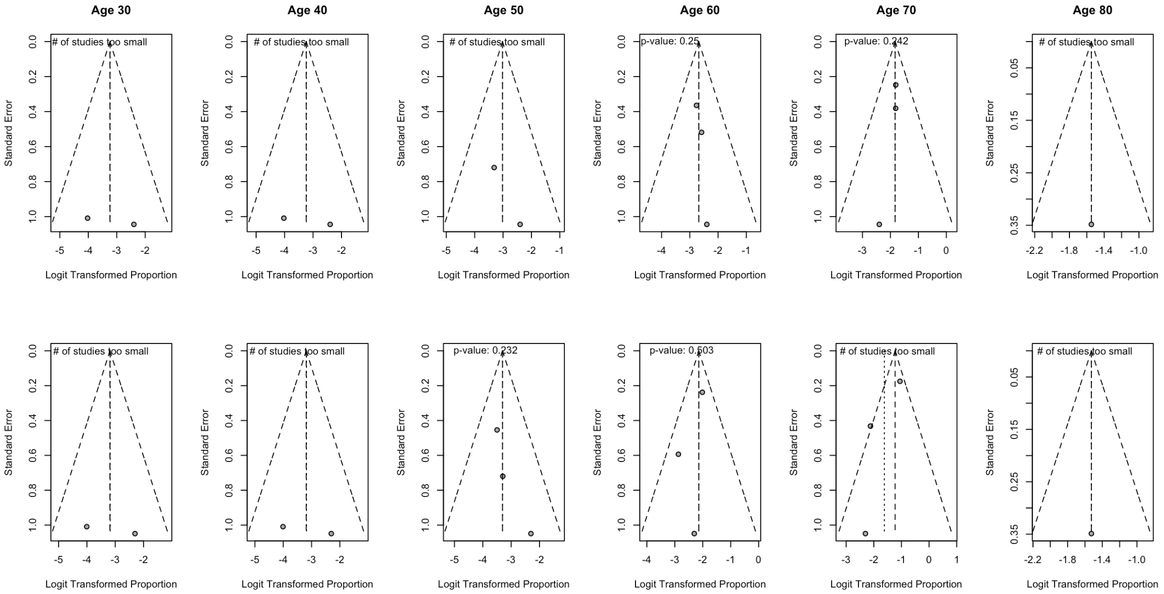
Colorectal cancer risk stratified by studies on unscreened/no prior surgery population (top panel) or unspecificed (i.e. likely a mix of screened and unscreened populations) (bottom panel) for (a) MLH1 carriers, (b) MSH2 carriers, and (c) MSH6 carriers. ***DerSimonian and Laird random effects model results:*** *The age range is divided into 10-year intervals. Within each we show cumulative risk estimates from individual studies (thin vertical blue/orange bars) and the meta-analytic estimate from the DerSimonian and Laird random effects model (thick vertical black bars). The height of vertical bars represents 95% CIs. Within each 10-year age interval, the published studies are arranged by publication year: Quehenberger et al.Quehenberger et al. (2005), Kopciuk et al., Kopciuk et al. (2009), Bonadona et al. Bonadona et al. (2011), and Dowty et al. Dowty et al. (2013) are arranged from left to right in the top panels. Dunlop et al. Dunlop et al. (1997), Aaltonen et al. Aaltonen et al. (2007), Stoffel et al. Stoffel et al. (2009), Borras et al. Borràs et al. (2010), Mukherjeeet al. Mukherjee et al. (2011), and Molleret al. Møller et al. (2017) are arranged from left to right in the bottom panels. An "x" represents not available.* ***Likelihood-based approach results:*** *Smooth blue and orange lines represent penetrance estimated from the likelihood-based approach by yearly age. Blue corresponds to male carriers, and orange corresponds to female carriers.*


Age 30
Age 40
Age 50
Age 60
Age 70
Age 80
p-value: 0.897
p-value: 0.068
p-value: 0.092
p-value: 0.102
p-value: 0.214
p-value: 0.377
p-value: 0.149
p-value: 0.127
p-value: 0.13
p-value: 0.021
p-value: 0.043
p-value: 0.097
Age 30
Age 40
Age 50
Age 60
Age 70
Age 80
p-value: 0.741
p-value: 0.673
p-value: 0.92
p-value: 0.871
p-value: 0.809
p-value: 0.861
p-value: 0.32
p-value: 0.041
p-value: 0.069
p-value: 0.108
p-value: 0.534
p-value: 0.96
Standard Error
Logit Transformed Proportion

(a) *MLH1*

(b) *MSH2*

(Continued)



(c) *MSH6*

Figure (A.3) *Funnel plots by study*

Panels (a), (b), and (c) correspond to MLH1, MSH2, and MSH6 mutation carriers, respectively. From left to right, the plots correspond to ages 30, 40, ..., 80. Within each panel, the top row corresponds to male, and bottom female.

Appendix B

Appendix to Chapter 2

B.1 Proof of Proposition 1

Proof of Proposition 1. We show $r_{(m)} = \prod_{\ell=0}^{m-1} \left(I - \eta H_{(m-\ell-1)}\right) Y$ by induction. Without loss of generality, we assume $\eta = 1$. At iteration 1, the residual vector is

$$\begin{aligned} r_{(1)} &= Y - \hat{Y}_{(0)} \\ &= \left(I - H_{(0)}\right) Y \end{aligned}$$

At iteration $m-1$, we assume the induction hypothesis:

$$r_{(m-1)} = \prod_{\ell=0}^{m-2} \left(I - H_{(m-\ell-1)}\right) Y \tag{B.1}$$

At iteration m, the residual vector is

$$\begin{aligned}
r_{(m)} &= Y - \hat{Y}_{(m-1)} \\
&= Y - \left(\hat{Y}_{(m-2)} + H_{(m-1)} r_{(m-1)}\right) \\
&= r_{(m-1)} - H_{(m-1)} r_{(m-1)} \\
&= \left(I - H_{(m-1)}\right) r_{(m-1)} \\
&\overset{(B.1)}{=} \left(I - H_{(m-1)}\right)\left(I - H_{(m-2)}\right) \cdots \left(I - H_{(1)}\right)\left(I - H_{(0)}\right) Y \\
&= \prod_{\ell=0}^{m-1} \left(I - H_{(m-\ell-1)}\right) Y.
\end{aligned}$$

It follows that $\left(\tilde{X}_{\hat{j}_{(m)}}^T \tilde{X}_{j_m}\right)^{-1} \tilde{X}_{\hat{j}_{(m)}}^T r_{(m)} \in \mathbb{R}$ is the coefficient estimate of $\tilde{X}_{\hat{j}_{(m)}}$. Multiplying the coefficient estimate by $e_{\hat{j}_{(m)}} \in \mathbb{R}^P$ results in an U-dimensional vector with $\left(\tilde{X}_{\hat{j}_{(m)}}^T \tilde{X}_{j_m}\right)^{-1} \tilde{X}_{\hat{j}_{(m)}}^T r_{(m)}$ in the $\hat{j}_{(m)}$-th position and 0 everywhere else. The final coefficient estimates are given by the sum across iteration-specific vectors $e_{\hat{j}_{(m)}} \left(\tilde{X}_{\hat{j}_{(m)}}^T \tilde{X}_{j_m}\right)^{-1} \tilde{X}_{\hat{j}_{(m)}}^T r_{(m)}$ for $m = 1, \dots, M$. □

B.2 Proof of Lemma 1

Proof of Lemma 1. We decompose Y into

$$Y = c_j (v_j^T Y) + z_j$$

and rewrite the polyhdron as

$$
\begin{aligned}
\{\Gamma Y \geq 0\} &= \left\{\Gamma\left(c_j v_j^T Y + z_j\right) \geq 0\right\} \\
&= \left\{\Gamma c_j (v_j^T y) \geq 0 - \Gamma z_j\right\} \\
&= \left\{\left(\Gamma c_j\right)_\ell \left(v_j^T Y\right) \geq 0 - (\Gamma z_j)_\ell \quad \text{for all } \ell = 1, \ldots, 2M(P-1)\right\} \\
&= \left\{\begin{pmatrix} v_j^T Y \geq \frac{0-(\Gamma z_j)_\ell}{(\Gamma c_j)_\ell}, & \text{for } \ell : (\Gamma c_j)_\ell > 0 \\ v_j^T Y \leq \frac{0-(\Gamma z_j)_\ell}{(\Gamma c)_i}, & \text{for } \ell : (\Gamma c_j)_\ell < 0 \\ 0 \geq 0 - (\Gamma z_j)_\ell & \text{for } \ell : (\Gamma c_j)_i = 0 \end{pmatrix}\right\} \\
&= \left\{\begin{pmatrix} v_j^T Y \geq \max\limits_{\ell:(\Gamma c)_\ell > 0} \frac{0-(\Gamma z_j)_\ell}{(\Gamma c)_i} \\ v_j^T Y \leq \min\limits_{\ell:(\Gamma c_j)_\ell < 0} \frac{0-(\Gamma z_j)_\ell}{(\Gamma c_j)_\ell} \\ 0 \geq \max\limits_{\ell:(\Gamma c)_\ell = 0} 0 - (\Gamma z_j)_\ell \end{pmatrix}\right\}
\end{aligned}
$$

where in the last step, we have divided the components into three categories depending on whether $(\Gamma c_j)_\ell \gtreqless 0$, since this affects the direction of the inequality (or whether we can divide at all). Since $v_j^T Y$ is the same quantity for all ℓ, it must be at least the maximum of the lower bounds, which is a_j, and no more than the minimum of the upper bounds, which is b_j. Since a_j, b_j, and c_j are independent of $v_j^T Y$, then $v_j^T Y$ is conditionally a normal random variable, truncated to be between a_j and b_j. By conditioning on the value of z_j,

$$v_j^T Y | \{\Gamma Y \geq 0, z_j = z\}$$

is a truncated normal.

□

B.3 Proof of Theorems 1 and 2

Proof of Theorems 1 and 2.

$$
\begin{aligned}
Bias\left(\tilde{X}_0\hat{\beta}^{\text{Merge}}_{(M)}\right) &= E\left(\tilde{X}_0 \sum_{m=1}^{M} \eta B\left(I-\eta H\right)^{m-1} Y\right) - f\left(\tilde{X}_0\right) \\
&= \tilde{X}_0 \tilde{R} f\left(\tilde{X}\right) - f\left(\tilde{X}_0\right) \\
Bias\left(\tilde{X}_0\hat{\beta}^{\text{Ens}}_{(M)}\right) &= E\left(\tilde{X}_0 \sum_{k=1}^{K} w_k \left[\sum_{m=1}^{M} \eta B_k\left(I-\eta H_k\right)^{m-1} Y_k\right]\right) - f\left(\tilde{X}_0\right) \\
&= \sum_{k=1}^{K} w_k \tilde{X}_0 \tilde{R}_k f\left(X_k\right) - f\left(\tilde{X}_0\right) \\
Cov\left(\tilde{X}_0\hat{\beta}^{\text{Merge}}_{(M)}\right) &= Cov\left(\tilde{X}_0 \sum_{m=1}^{M} \eta B\left(I-\eta H\right)^{m-1} Y\right) \\
&= \tilde{X}_0 \tilde{R} Cov(Y) \tilde{R}^T \tilde{X}_0^T \\
&= \tilde{X}_0 \tilde{R} \text{blkdiag}\left(\left\{Cov\left(Y_k\right)\right\}_{k=1}^{K}\right) \tilde{R}^T \tilde{X}_0^T \\
&= \tilde{X}_0 \tilde{R} \text{blkdiag}\left(\left\{Z_k G Z_k^T + \sigma_\epsilon^2 I\right\}_{k=1}^{K}\right) \tilde{R}^T \tilde{X}_0^T \\
Cov\left(\tilde{X}_0\hat{\beta}^{\text{Ens}}_{(M)}\right) &= Cov\left(\tilde{X}_0 \sum_{k=1}^{K} w_k \left[\sum_{m=1}^{M} \eta B_k\left(I-\eta H_k\right)^{m-1} Y_k\right]\right) \\
&= Cov\left(\tilde{X}_0 \sum_{k=1}^{K} w_k \tilde{R}_k Y_k\right) \\
&= \sum_{k=1}^{K} w_k^2 \tilde{X}_0 \tilde{R}_k \left(Z_k G Z_k^T + \sigma_\epsilon^2 I\right) \tilde{R}_k^T \tilde{X}_0^T \\
&= \sum_{k=1}^{K} w_k^2 \tilde{X}_0 \tilde{R}_k Z_k G Z_k^T \tilde{R}_k^T \tilde{X}_0^T + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 \tilde{X}_0 \tilde{R}_k \tilde{R}_k^T \tilde{X}_0^T
\end{aligned}
$$

Let $b^{\text{Merge}} = Bias\left(\tilde{X}_0\hat{\beta}^{\text{Merge}}_{(M)}\right)$. The MSPE of $\hat{\beta}^{\text{Merge}}_{(M)}$ is

$$\begin{aligned}
E\left[\|Y_0-\tilde{X}_0\hat{\beta}^{\text{Merge}}_{(M)}\|_2^2\right] &= \text{tr}\left(Cov\left(\tilde{X}_0\hat{\beta}^{\text{Merge}}_{(M)}\right)\right)+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \text{tr}\left(\tilde{X}_0\tilde{R}\text{blkdiag}\left(\{Cov(Y_k)\}_{k=1}^K\right)\tilde{R}^T\tilde{X}_0^T\right)+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \text{tr}\left(\text{blkdiag}\left(\left\{Z_kGZ_k^T+\sigma_\epsilon^2 I\right\}_{k=1}^K\right)\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \text{tr}\left(\text{blkdiag}\left(\{Z_kGZ_k^T\}_{k=1}^K\right)\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)+\sigma_\epsilon^2\text{tr}\left(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \text{tr}\left(Z'G'Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)+\sigma_\epsilon^2\text{tr}\left(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \text{tr}\left(G'Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z'\right)+\sigma_\epsilon^2\text{tr}\left(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \sum_{d=1}^D\sigma_{(d)}^2\left\{\sum_{i:\sigma_i^2=\sigma_{(d)}^2}\left[\sum_{k=1}^K\left(Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0RZ'\right)_{i+Q\times(k-1),i+Q\times(k-1)}\right]\right\}+\sigma_\epsilon^2\text{tr}\left(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)\\
&\quad+\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]
\end{aligned}$$

Let $b^{\text{Ens}} = Bias\left(\tilde{X}_0\hat{\beta}^{\text{Ens}}_{(M_{\text{Ens}})}\right)$. The MSPE of $\hat{\beta}^{Ens}_{(M_{\text{Ens}})}$ is

$$\begin{aligned}
E\left[\|Y_0-\tilde{X}_0\hat{\beta}^{\text{Ens}}_{(M_{\text{Ens}})}\|_2^2\right] &= \text{tr}\left(Cov\left(\tilde{X}_0\hat{\beta}^{\text{Ens}}_{(M_{\text{Ens}})}\right)\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \text{tr}\left(\tilde{X}_0Cov\left(\sum_{k=1}^K w_k\tilde{R}_kY_k\right)\tilde{X}_0^T\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \sum_{k=1}^K w_k^2\text{tr}\left(Z_kGZ_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k\right)+\sigma_\epsilon^2\sum_{k=1}^K w_k^2\text{tr}\left(\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \sum_{k=1}^K w_k^2\text{tr}\left(GZ_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k\right)+\sigma_\epsilon^2\sum_{k=1}^K w_k^2\text{tr}\left(\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]\\
&= \sum_{d=1}^D\sigma_{(d)}^2\left\{\sum_{i:\sigma_i^2=\sigma_{(d)}^2}\left[\sum_{k=1}^K w_k^2\left(Z_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k\right)_{i,i}\right]\right\}\\
&\quad+\sigma_\epsilon^2\sum_{k=1}^K w_k^2\text{tr}\left(\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}+E\left[\|Y_0-f(\tilde{X}_0)\|_2^2\right]
\end{aligned}$$

If $\sigma_1^2=\sigma_2^2=\ldots=\sigma_J^2$ (Theorem 1), then

$$\begin{aligned}
&\overline{\sigma}^2 \geq \frac{Q}{P}\times\frac{\sigma_\epsilon^2\left(\sum_{k=1}^K w_k^2\text{tr}\left(\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k\right)-\text{tr}\left(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}-\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}}{\text{tr}\left(Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z'\right)-\sum_{k=1}^K w_k^2\text{tr}\left(Z_k^T\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k\right)}\\
&\Rightarrow \sigma^2\left(\text{tr}\left(Z'^T\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}Z'\right)-\sum_{k=1}^K w_k^2\text{tr}\left(Z_k\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_kZ_k\right)\right)\\
&\geq \sigma_\epsilon^2\left(\sum_{k=1}^K w_k^2\text{tr}\left(\tilde{R}_k^T\tilde{X}_0^T\tilde{X}_0\tilde{R}_k\right)-\text{tr}\left(\tilde{R}^T\tilde{X}_0^T\tilde{X}_0\tilde{R}\right)\right)+\left(b^{\text{Ens}}\right)^T b^{\text{Ens}}-\left(b^{\text{Merge}}\right)^T b^{\text{Merge}}\\
&\Leftrightarrow E\left[\|Y_0-\tilde{X}_0\hat{\beta}^{\text{Merge}}_{(M)}\|_2^2\right]\geq E\left[\|Y_0-\tilde{X}_0\hat{\beta}^{\text{Ens}}_{(M_{\text{Ens}})}\|_2^2\right].
\end{aligned}$$

If $\sigma_j^2 \neq \sigma_{j'}^2$ for at least one $j \neq j'$ (Theorem 2), then let

$$a_d = \sum_{i:\sigma_i^2=\sigma_{(d)}^2} \left[\sum_{k=1}^{K} \left(Z'^T \tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R} Z' \right)_{i+Q\times(k-1),i+Q\times(k-1)} - w_k^2 \left(Z_k^T \tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k Z_k \right)_{i,i} \right]$$

and

$$c = \sigma_\epsilon^2 \left(\sum_{k=1}^{K} w_k^2 \mathrm{tr}\left(\tilde{R}_k^T \tilde{X}_0^T \tilde{X}_0 \tilde{R}_k \right) - \mathrm{tr}\left(\tilde{R}^T \tilde{X}_0^T \tilde{X}_0 \tilde{R} \right) \right) + \left(b^{\mathrm{Ens}} \right)^T b^{\mathrm{Ens}} - \left(b^{\mathrm{Merge}} \right)^T b^{\mathrm{Merge}}.$$

Since

$$E\left[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M)}^{\mathrm{Merge}}\|_2^2 \right] \geq E\left[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M_{\mathrm{Ens}})}^{\mathrm{Ens}}\|_2^2 \right] \Longleftrightarrow \sum_{d=1}^{D} \sigma_{(d)}^2 a_d \geq c$$

and

$$\left(\min_d \frac{a_d}{J_d} \right) \sum_{d=1}^{D} \sigma_{(d)}^2 J_d \leq \sum_{d=1}^{D} \sigma_{(d)}^2 \leq \left(\max_d \frac{a_d}{J_d} \right) \sum_{d=1}^{D} \sigma_{(d)}^2 J_d,$$

assuming $a_d > 0$ for all d, then

$$\begin{aligned}
\overline{\sigma}^2 = \frac{\sum_{d=1}^{D} \sigma_{(d)}^2 J_d}{P} &\leq \frac{c}{P \max_d \frac{a_d}{J_d}} = \tau_1 \\
&\Rightarrow \sum_{d=1}^{D} \sigma_{(d)}^2 a_d \leq \max_d \frac{a_d}{J_d} \sum_{d=1}^{D} \sigma_{(d)}^2 J_d \leq c \\
&\Longleftrightarrow E\left[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M)}^{\mathrm{Merge}}\|_2^2 \right] \leq E\left[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M_{\mathrm{Ens}})}^{\mathrm{Ens}}\|_2^2 \right].
\end{aligned}$$

and

$$\begin{aligned}
\overline{\sigma}^2 = \frac{\sum_{d=1}^{D} \sigma_{(d)}^2 J_d}{P} &\geq \frac{c}{P \max_d \frac{a_d}{J_d}} = \tau_2 \\
&\Rightarrow \sum_{d=1}^{D} \sigma_{(d)}^2 a_d \geq \min_d \frac{a_d}{J_d} \sum_{d=1}^{D} \sigma_{(d)}^2 J_d \geq c \\
&\Longleftrightarrow E\left[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M)}^{\mathrm{Merge}}\|_2^2 \right] \geq E\left[\|Y_0 - \tilde{X}_0 \hat{\beta}_{(M_{\mathrm{Ens}})}^{\mathrm{Ens}}\|_2^2 \right].
\end{aligned}$$

□

B.4 Proof of Proposition 2

Proof of Proposition 2.

$$
\begin{aligned}
Var\left(\hat{\beta}^{\text{Merge, CW}}_{(M)j}\middle|\mathcal{P}\right) &= \vartheta_j^2\left(1-\frac{\xi_j\phi(\xi_j)-\alpha_j\phi(\alpha_j)}{\Phi(\xi_j)-\Phi(\alpha_j)}-\left(\frac{\phi(\xi_j)-\phi(\alpha_j)}{\Phi(\xi_j)-\Phi(\alpha_j)}\right)^2\right)\\
Bias^2\left(\hat{\beta}^{\text{Merge, CW}}_{(M)j}\middle|\mathcal{P}\right) &= \left(\bar{\mu}_j-\vartheta_j\left(\frac{\phi(\xi_j)-\phi(\alpha_j)}{\Phi(\xi_j)-\Phi(\alpha_j)}\right)-\beta_j\right)^2\\
Var\left(\hat{\beta}^{\text{Ens, CW}}_{(M)j}\middle|\mathcal{P}^{\text{Ens}}\right) &= Var\left(\sum_{k=1}^{K}w_k\hat{\beta}^{\text{CW}}_{(M_k)jk}\middle|\mathcal{P}^{\text{Ens}}\right)\\
&= \sum_{k=1}^{K}w_k^2 Var\left(\hat{\beta}^{\text{CW}}_{(M_k)jk}\middle|\mathcal{P}_k\right) \qquad \text{(because } Y_k\text{'s are independent)}\\
&= \sum_{k=1}^{K}w_k^2\vartheta_{jk}^2\left(1-\frac{\xi_{jk}\phi(\xi_{jk})-\alpha_{jk}\phi(\alpha_{jk})}{\Phi(\xi_{jk})-\Phi(\alpha_{jk})}-\left(\frac{\phi(\xi_{jk})-\phi(\alpha_{jk})}{\Phi(\xi_{jk})-\Phi(\alpha_{jk})}\right)^2\right)\\
Bias^2\left(\hat{\beta}^{\text{Ens, CW}}_{(M)j}\middle|\mathcal{P}^{\text{Ens}}\right) &= \left(\sum_{k=1}^{K}w_k E\left(\hat{\beta}^{\text{CW}}_{(M_k)jk}\middle|\mathcal{P}^{\text{Ens}}\right)-\beta_j\right)^2\\
&= \left(\sum_{k=1}^{K}w_k E\left(\hat{\beta}^{\text{CW}}_{(M_k)jk}\middle|\mathcal{P}_k\right)-\beta_j\right)^2 \qquad \text{(because } Y_k\text{'s are independent)}\\
&= \left(\sum_{k=1}^{K}w_k\left(\bar{\mu}_{jk}-\vartheta_{jk}\left(\frac{\phi(\xi_{jk})-\phi(\alpha_{jk})}{\Phi(\xi_{jk})-\Phi(\alpha_{jk})}\right)\right)-\beta_j\right)^2
\end{aligned}
$$

□

B.5 Truncation region for component-wise boosting coefficients

Claim 2 (Truncation region for component-wise boosting coefficients)**.** *Let $Y \in \mathbb{R}^N$ denote the outcome vector where $Y \sim N(\mu, \Sigma)$. The boosting coefficients can be written as*

$$
\begin{aligned}
\hat{\beta}^{\textit{CW, Merge}}_{(M)} &= V^T Y\\
&:= \sum_{m=1}^{M}\eta B_{(m)}\left(\prod_{\ell=0}^{m-1}(I-\eta H_{(m-\ell-1)})\right)Y,
\end{aligned}
$$

where $V \in \mathbb{R}^{N \times P}$ depends on Y through variable selection. We decompose Y into

$$Y = C(V^T Y) + Z^*,$$

where

$$C = \Sigma V \left(V^T \Sigma V\right)^{-1},$$

is a $N-$dimensional vector and

$$Z^* = \left(I - \Sigma V \left(V^T \Sigma V\right)^{-1} V^T\right) Y$$

is a $\ell_P := 2M(P-1)$ dimensional vector. We claim the polyhedral set $\{\Gamma Y \geq 0\}$ can be re-written as a truncation region where the coefficients $\hat{\beta}_{(M)}^{CW,\,Merge}$ have non-rectangular truncation limits.

Proof. We define the projection $\Pi_k(S)$ of a set $S \subset \mathbb{R}^n$ by letting

$$\Pi_k(S) = \{(x_1, \ldots, x_k) | \exists x_{k+1}, \ldots, x_n \text{ s.t. } (x_1, \ldots, x_n) \in S\}.$$

Given a polyhedron $\mathcal{P}$ in terms of linear inequality constraints of the form

$$Ax \geq b,$$

we state the Fourier Motzkin elimination algorithm from Bertsimas and Tsitsiklis (1997).

We note the following:

1. The projection $\Pi_k(\mathcal{P})$ can be generated by repeated application of the elimination algorithm (Theorem 2.10 in Bertsimas and Tsitsiklis (1997))

2. The elimination approach always produces a polyhedron (definition of the elimination algorithm in Bertsimas and Tsitsiklis (1997)).

Therefore, it follows that a projection $\Pi_k(\mathcal{P})$ of a polyhedron is also a polyhedron.

The polyhedral set $\mathcal{P} := \{Y : \Gamma Y \geq 0\}$ is a system of $\ell_P := 2M(P-1)$ linear inequalities, with P variables $V^T Y_1, \ldots, V^T Y_P$. Let $(A)_{ij}$ denote the i,j-th entry in matrix A. We let $I_P = \{1, 2, \ldots, \ell_P\}$ denote the row index set for the system of inequalities with P variables and partition it into subsets I_P^+, I_P^-, and I_P^0, where $I_P^+ = \{i : (\Gamma C)_{ip} > 0\}, I_P^- = \{i : (\Gamma C)_{ip} < 0\}$, and

Algorithm 3 Elimination algorithm for a system of linear inequalities

```
1: Rewrite each constraint $\sum_{j=1}^{N} a_{ij}x_j \geq b_i$ in the form
       $a_{iNx_N} \geq -\sum_{j=1}^{N-1} a_{ij}x_j + b_i, \quad i = 1,\dots,m$
   if $a_{iN} \neq 0$, divide both sides by $a_{iN}$. By letting $\bar{x} = (x_1,\dots,x_{n-1})$, we obtain an equivalent
   representation of $\mathcal{P}$ involving the following constraints
       $x_N \geq d_i + f_i'\bar{x}, \qquad$ if $a_{iN} > 0$
       $d_j + f_j'\bar{x} \geq x_N, \qquad$ if $a_{jN} < 0$
       $0 \geq d_k + f_k'\bar{x}, \qquad$ if $a_{kN} = 0$
   Each $d_i, d_j, d_k$ is a scalar, and each $f_i, f_j, f_k$ is a vector in $\mathbb{R}^{N-1}$.
2: Let $\mathcal{Q}$ be the polyhedron in $\mathbb{R}^{N-1}$ defined by the constraints
       $d_j + f_j'\bar{x} \geq d_i + f_i'\bar{x} \qquad$ if $a_{iN} > 0$ and $a_{jN} < 0$
       $0 \geq d_k + f_k'\bar{x}, \qquad$ if $a_{kN} = 0$
```

$I_P^0 = \{i : (\Gamma C)_{ip} = 0\}$. Then we have

$$
\begin{aligned}
\{\Gamma Y \geq 0\} &= \left\{\Gamma\left(CV^TY + Z^*\right) \geq 0\right\} \\
&= \left\{\underbrace{\Gamma C}_{\ell_P \times P}\ \underbrace{V^TY}_{P\times 1} \geq \underbrace{0 - \Gamma Z^*}_{\ell_P \times 1}\right\} \\
&= \left\{\sum_{j=1}^{P} (\Gamma C)_{ij}(V^TY)_j \geq 0 - (\Gamma Z^*)_i \quad i = 1,\dots,\ell_P\right\} \\
&= \left\{(\Gamma C)_{ip}(V^TY)_p \geq -\sum_{j=1}^{P-1} (\Gamma C)_{ij}(V^TY)_j - (\Gamma Z^*)_i \quad i = 1,\dots,\ell_P\right\} \\
&= \left\{\begin{array}{ll}
(V^TY)_P \geq \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{qj}(V^TY)_j - (\Gamma Z^*)_q}{(\Gamma C)_{qp}}, & \text{for } q \in I_P^+ \\
(V^TY)_P \leq \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{rj}(V^TY)_j - (\Gamma Z^*)_r}{(\Gamma C)_{rp}}, & \text{for } r \in I_P^- \\
0 \geq -\sum_{j=1}^{P-1}(\Gamma C)_{sj}(V^TY)_j - (\Gamma Z^*)_s & \text{for } s \in I_P^0
\end{array}\right\} \\
&= \left\{\begin{array}{ll}
\max_{q \in I^+} \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{qj}(V^TY)_j - (\Gamma Z^*)_q}{(\Gamma C)_{qp}} \leq \left(V^TY\right)_p \leq \min_{r \in I^-} \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{rj}(V^TY)_j - (\Gamma Z^*)_r}{(\Gamma C)_{rp}} & \\
\qquad 0 \geq -\sum_{j=1}^{P-1}(\Gamma C)_{sj}(V^TY)_j - (\Gamma Z^*)_s & \text{for } s \in I_P^0
\end{array}\right\}
\end{aligned}
$$

We reduce this to a system of inequalities with $P-1$ variables after eliminating $(V^TY)_P$:

$$\left\{\begin{matrix} \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{qj}(V^TY)_j-(\Gamma Z^*)_q}{(\Gamma C)_{qp}} \leq \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{rj}(V^TY)_j-(\Gamma Z^*)_r}{(\Gamma C)_{rp}} \text{ for } q \in I_P^+, r \in I_P^- \\ 0 \geq -\sum_{j=1}^{P-1}(\Gamma C)_{sj}(V^TY)_j - (\Gamma Z^*)_s \text{ for } s \in I_P^0 \end{matrix}\right\} \tag{B.2}$$

The set in (B.2) is a system of $\ell_{P-1} := |I_P^+| \times |I_P^-| + |I_P^0|$ inequalities. It is a polyhedral set in $\mathbb{R}^{P-1}$, which can be seen by rewriting (B.2) as follows:

$$
\begin{aligned}
&\left\{\begin{matrix} \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{qj}(V^TY)_j-(\Gamma Z^*)_q}{(\Gamma C)_{qp}} \leq \frac{-\sum_{j=1}^{P-1}(\Gamma C)_{rj}(V^TY)_j-(\Gamma Z^*)_r}{(\Gamma C)_{rp}} \text{ for } q \in I^+, r \in I^- \\ 0 \geq -\sum_{j=1}^{P-1}(\Gamma C)_{sj}(V^TY)_j-(\Gamma Z^*)_s \text{ for } s \in I^0 \end{matrix}\right\} \\
=&\left\{\begin{matrix} -\sum_{j=1}^{P-1}(\Gamma C)_{rp}(\Gamma C)_{qj}(V^TY)_j-(\Gamma C)_{rp}(\Gamma Z^*)_q \geq -\sum_{j=1}^{P-1}(\Gamma C)_{qp}(\Gamma C)_{rj}(V^TY)_j-(\Gamma C)_{qp}(\Gamma Z^*)_r \text{ for } q \in I^+, r \in I^- \\ \sum_{j=1}^{P-1}(\Gamma C)_{sj}(V^TY)_j \geq -(\Gamma Z^*)_s \text{ for } s \in I^0 \end{matrix}\right\} \\
=&\left\{\begin{matrix} \sum_{j=1}^{P-1}\big((\Gamma C)_{qp}(\Gamma C)_{rj}-(\Gamma C)_{rp}(\Gamma C)_{qj}\big)(V^TY)_j \geq (\Gamma C)_{rp}(\Gamma Z^*)_q-(\Gamma C)_{qp}(\Gamma Z^*)_r \text{ for } q \in I^+, r \in I^- \\ \sum_{j=1}^{P-1}(\Gamma C)_{sj}(V^TY)_j \geq -(\Gamma Z^*)_s \text{ for } s \in I^0 \end{matrix}\right\}.
\end{aligned}
$$

Let A_{p-k} denote a $\ell_{p-k} \times (p-k)$ matrix, $(V^TY)_{1:p-k}$ a vector that contains the first $p-k$ coordinates of (V^TY), and $b_{p-k}(Z^*)$ a ℓ_{p-k}-dimensional vector, where $k \in \{0, \ldots, P-1\}$, and ℓ_{p-k} is the number of linear constraints in $\Pi_{p-k}(\mathcal{P})$, which is the projection of $\mathcal{P}$. Note that $A_P = \Gamma C$ and $b_P(Z^*) = 0 - \Gamma Z^*$.

We repeat the elimination process $P-1$ times to obtain $\Pi_1(\mathcal{P})$:

$$
\begin{aligned}
\{\Gamma Y \geq 0\} &= \{A_p(V^TY) \geq b_p(Z^*)\} \\
\Pi_{P-1}(\mathcal{P}) &= \{A_{P-1}(V^TY)_{1:P-1} \geq b_{P-1}(Z^*)\} \\
&\vdots \\
\Pi_1(\mathcal{P}) &= \{A_1(V^TY)_1 \geq b_1(Z^*)\}.
\end{aligned}
$$

Induction base case for $\Pi_2(\mathcal{P})$: Without loss of generality, we assume the variable in $\Pi_1(\mathcal{P})$ is $(V^TY)_1$. We can obtain its lower and upper truncation limits, $\mathcal{V}_1^{\text{lo}}(Z^*)$ and $\mathcal{V}_1^{\text{up}}(Z^*)$, and $\mathcal{V}_1^0(Z^*)$ using the same argument as the one in Lee *et al.* (2016), where

$$
\begin{aligned}
\mathcal{V}_1^{\text{lo}}(Z^*) &= \max_{i:(A_1)_i>0} \frac{(b_1(Z^*))_i}{(A_1)_i} \\
\mathcal{V}_1^{\text{up}}(Z^*) &= \min_{i:(A_1)_i<0} \frac{(b_1(Z^*))_i}{(A_1)_i} \\
\mathcal{V}_1^0(Z^*) &= \max_{i:(A_1)_i=0} (b_1(Z^*))_i.
\end{aligned}
$$

We conclude that $\Pi_1(\mathcal{P}) = \{(\mathcal{V}_1^{\text{lo}}(Z^*) \leq (V^TY)_1 \leq \mathcal{V}_1^{\text{up}}(Z^*), \mathcal{V}_1^0(Z^*) \leq 0\}$.

By the definition of $\Pi_2(\mathcal{P})$, we have

$$\begin{aligned}\Pi_2(\mathcal{P}) &= \left\{A_2(V^T Y)_{1:2} \geq b_2(Z^*)\right\} \\ &= \left\{\begin{matrix} A_2(V^T Y)_{1:2} \geq b_2(Z^*) \\ \mathcal{V}_1^{\text{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\text{up}}(Z^*) \\ \mathcal{V}_1^0(Z^*) \leq 0 \end{matrix}\right\}\end{aligned}$$

because reducing the system from $\Pi_2(\mathcal{P})$ to $\Pi_1(\mathcal{P})$ does not change the range of $(V^T Y)_1$ that satisfy the linear constraints in $\Pi_2(\mathcal{P})$.

We can obtain the lower and upper truncation limits for $(V^T Y)_2$ as a function of $(V^T Y)_1$.

$$\begin{aligned}
\Pi_2(\mathcal{P}) &= \left\{ A_2 (V^T Y)_{1:2} \geq b_2(Z^*) \right\} \\
&= \left\{ \begin{array}{c} A_2 (V^T Y)_{1:2} \geq b_2(Z^*) \\ \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \end{array} \right\} \\
&= \left\{ \begin{array}{c} \sum_{j=1}^{2} (A_2)_{ij} (V^T Y)_j \geq (b_2(Z^*))_i \quad \text{for } i = 1, \ldots, \ell_2 \\ \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \end{array} \right\} \\
&= \left\{ \begin{array}{c} (A_2)_{i2} (V^T Y)_2 \geq -(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i \quad \text{for } i = 1, \ldots, \ell_2 \\ \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \end{array} \right\} \\
&= \left\{ \begin{array}{c} (V^T Y)_2 \geq \frac{-(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i}{(A_2)_{i2}} \quad \text{for } i : (A_2)_{i2} > 0 \\ (V^T Y)_2 \leq \frac{-(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i}{(A_2)_{i2}} \quad \text{for } i : (A_2)_{i2} < 0 \\ 0 \geq -(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i \quad \text{for } i : (A_2)_{i2} = 0 \\ \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \end{array} \right\} \\
&= \left\{ \begin{array}{c} \max\limits_{i:(A_2)_{i2}>0} \frac{-(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i}{(A_2)_{i2}} \leq (V^T Y)_2 \leq \min\limits_{i:(A_1)_{i2}<0} \frac{-(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i}{(A_2)_{i2}} \\ 0 \geq \max\limits_{i:(A_2)_{i2}=0} -(A_2)_{i1} (V^T Y)_1 + (b_2(Z^*))_i \\ \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \end{array} \right\} \\
&= \left\{ \begin{array}{c} \mathcal{V}_2^{\mathrm{lo}}(Z^*, (V^T Y)_1) \leq (V^T Y)_2 \leq \mathcal{V}_2^{\mathrm{up}}(Z^*, (V^T Y)_1) \\ \mathcal{V}_2^{0}(Z^*, (V^T Y)_1) \leq 0 \\ \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^T Y)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \end{array} \right\}
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{V}_2^{\text{lo}}(Z^*,(V^TY)_1) &= \max_{i:(A_2)_{i2}>0} \frac{-(A_2)_{i1}(V^TY)_1(Z^*)_1 + (b_2(Z^*))_i}{(A_2)_{i2}} \\
\mathcal{V}_2^{\text{up}}(Z^*,(V^TY)_1) &= \min_{i:(A_2)_{i2}<0} \frac{-(A_2)_{i1}(V^TY)_1(Z^*)_1 + (b_2(Z^*))_i}{(A_2)_{i2}} \\
\mathcal{V}_2^{0}(Z^*,(V^TY)_1) &= \max_{i:(A_2)_{i2}=0} -(A_2)_{i1}(V^TY)_1 + (b_2(Z^*))_i.
\end{aligned}$$

<u>Inductive step for $\Pi_{P-1}(\mathcal{P})$</u>: Under the induction hypothesis, we assume

$$\Pi_{P-2}(\mathcal{P}) = \left\{ \begin{array}{c}
\mathcal{V}_1^{\text{lo}}(Z^*) \le (V^TY)_1 \le \mathcal{V}_1^{\text{up}}(Z^*) \\
\mathcal{V}_1^{0}(Z^*) \le 0 \\
\mathcal{V}_2^{\text{lo}}((V^TY)_1, Z^*) \le (V^TY)_2 \le \mathcal{V}_2^{\text{up}}((V^TY)_1, Z^*) \\
\mathcal{V}_2^{0}((V^TY)_1, Z^*) \le 0 \\
\vdots \\
\mathcal{V}_{P-2}^{\text{lo}}((V^TY)_{1:P-3}, Z^*) \le (V^TY)_{P-2} \le \mathcal{V}_{P-2}^{\text{up}}((V^TY)_{1:P-3}, Z^*) \\
\mathcal{V}_{P-2}^{0}((V^TY)_{1:P-3}, Z^*) \le 0
\end{array} \right\}$$

Then we have

$$
\begin{aligned}
\Pi_{P-1}(\mathcal{P}) &= \Big\{ A_{P-1}(V^TY)_{1:P-1} \geq b_{P-1}(Z^*) \Big\} \\
&= \left\{ \begin{array}{c}
A_{P-1}(V^TY)_{1:P-1} \geq b_{P-1}(Z^*) \\
\mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^TY)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\
\mathcal{V}_1^{0}(Z^*) \leq 0 \\
\mathcal{V}_2^{\mathrm{lo}}(Z^*,(V^TY)_1) \leq (V^TY)_2 \leq \mathcal{V}_2^{\mathrm{up}}(Z^*,(V^TY)_1) \\
\mathcal{V}_2^{0}(Z^*,(V^TY)_1) \leq 0 \\
\vdots \\
\mathcal{V}_{P-2}^{\mathrm{lo}}((V^TY)_{1:P-3},Z^*) \leq (V^TY)_{P-2} \leq \mathcal{V}_{P-2}^{\mathrm{up}}((V^TY)_{1:P-3},Z^*) \\
\mathcal{V}_{P-2}^{0}((V^TY)_{1:P-3},Z^*) \leq 0
\end{array} \right\} \\
&= \left\{ \begin{array}{c}
(A_{P-1})_{i(P-1)}(V^TY)_{P-1} \geq -\sum_{j=1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i \quad \text{for } i = 1, \ldots, \ell_{P-1} \\
\mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^TY)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\
\mathcal{V}_1^{0}(Z^*) \leq 0 \\
\mathcal{V}_2^{\mathrm{lo}}(Z^*,(V^TY)_1) \leq (V^TY)_2 \leq \mathcal{V}_2^{\mathrm{up}}(Z^*,(V^TY)_1) \\
\mathcal{V}_2^{0}(Z^*,(V^TY)_1) \leq 0 \\
\vdots \\
\mathcal{V}_{P-2}^{\mathrm{lo}}((V^TY)_{1:P-3},Z^*) \leq (V^TY)_{P-2} \leq \mathcal{V}_{P-2}^{\mathrm{up}}((V^TY)_{1:P-3},Z^*) \\
\mathcal{V}_{P-2}^{0}((V^TY)_{1:P-3},Z^*) \leq 0
\end{array} \right\} \\
&= \left\{ \begin{array}{c}
(V^TY)_{P-1} \geq \frac{-\sum_{j=1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i}{(A_{P-1})_{i(P-1)}} \quad \text{for } i : (A_{P-1})_{i(P-1)} > 0 \\
(V^TY)_{P-1} \leq \frac{-\sum_{j=1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i}{(A_{P-1})_{i(P-1)}} \quad \text{for } i : (A_{P-1})_{i(P-1)} < 0 \\
0 \geq -\sum_{j=1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i \quad \text{for } i : (A_{P-1})_{i(P-1)} = 0 \\
\mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^TY)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\
\mathcal{V}_1^{0}(Z^*) \leq 0 \\
\mathcal{V}_2^{\mathrm{lo}}(Z^*,(V^TY)_1) \leq (V^TY)_2 \leq \mathcal{V}_2^{\mathrm{up}}(Z^*,(V^TY)_1) \\
\mathcal{V}_2^{0}(Z^*,(V^TY)_1) \leq 0 \\
\vdots \\
\mathcal{V}_{P-2}^{\mathrm{lo}}((V^TY)_{1:P-3},Z^*) \leq (V^TY)_{P-2} \leq \mathcal{V}_{P-2}^{\mathrm{up}}((V^TY)_{1:P-3},Z^*) \\
\mathcal{V}_{P-2}^{0}((V^TY)_{1:P-3},Z^*) \leq 0
\end{array} \right\}. \\
&= \left\{ \begin{array}{c}
\mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^TY)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\
\mathcal{V}_1^{0}(Z^*) \leq 0 \\
\mathcal{V}_2^{\mathrm{lo}}((V^TY)_1,Z^*) \leq (V^TY)_2 \leq \mathcal{V}_2^{\mathrm{up}}((V^TY)_1,Z^*) \\
\mathcal{V}_2^{0}((V^TY)_1,Z^*) \leq 0 \\
\vdots \\
\mathcal{V}_{P-1}^{\mathrm{lo}}((V^TY)_{1:P-2},Z^*) \leq (V^TY)_{P-1} \leq \mathcal{V}_{P-1}^{\mathrm{up}}((V^TY)_{1:P-2},Z^*) \\
\mathcal{V}_{P-1}^{0}((V^TY)_{1:P-2},Z^*) \leq 0
\end{array} \right\}
\end{aligned}
$$

where

$$\mathcal{V}_{P-1}^{\mathrm{lo}}\left((V^TY)_{1:P-2}, Z^*\right) = \max_{i:(A_{P-1})_{i(P-1)}>0} \frac{-\sum_{j-1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i}{(A_{P-1})_{i(P-1)}}$$

$$\mathcal{V}_{P-1}^{\mathrm{up}}\left((V^TY)_{1:P-2}, Z^*\right) = \min_{i:(A_{P-1})_{i(P-1)}<0} \frac{-\sum_{j-1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i}{(A_{P-1})_{i(P-1)}}$$

$$\mathcal{V}_{P-1}^{0}\left((V^TY)_{1:P-2}, Z^*\right) = \max_{i:(A_{P-1})_{i(P-1)}=0} -\sum_{j=1}^{P-2}(A_{P-1})_{ij}(V^TY)_j + (b_{P-1}(Z^*))_i.$$

Therefore, we conclude that

$$\Pi_P(\mathcal{P}) = \{\Gamma Y \geq 0\}$$
$$= \left\{\begin{array}{c} \mathcal{V}_1^{\mathrm{lo}}(Z^*) \leq (V^TY)_1 \leq \mathcal{V}_1^{\mathrm{up}}(Z^*) \\ \mathcal{V}_1^{0}(Z^*) \leq 0 \\ \mathcal{V}_2^{\mathrm{lo}}((V^TY)_1, Z^*) \leq (V^TY)_2 \leq \mathcal{V}_2^{\mathrm{up}}((V^TY)_1, Z^*) \\ \mathcal{V}_2^{0}((V^TY)_1, Z^*) \leq 0 \\ \vdots \\ \mathcal{V}_{P-1}^{\mathrm{lo}}((V^TY)_{1:P-2}, Z^*) \leq (V^TY)_{P-1} \leq \mathcal{V}_{P-1}^{\mathrm{up}}((V^TY)_{1:P-2}, Z^*) \\ \mathcal{V}_{P-1}^{0}((V^TY)_{1:P-2}, Z^*) \leq 0 \\ \mathcal{V}_p^{\mathrm{lo}}((V^TY)_{1:P-1}, Z^*) \leq (V^TY)_p \leq \mathcal{V}_p^{\mathrm{up}}((V^TY)_{1:P-1}, Z^*) \\ \mathcal{V}_p^{0}((V^TY)_{1:P-1}, Z^*) \leq 0 \end{array}\right\}$$

where

$$\mathcal{V}_{P}^{\mathrm{lo}}\left((V^TY)_{1:P-1}, Z^*\right) = \max_{i:(A_P)_{ip}>0} \frac{-\sum_{j-1}^{P-1}(A_P)_{ij}(V^TY)_j + (b_P(Z^*))_i}{(A_P)_{ip}}$$

$$\mathcal{V}_{P}^{\mathrm{up}}\left((V^TY)_{1:P-1}, Z^*\right) = \min_{i:(A_P)_{ip}<0} \frac{-\sum_{j-1}^{P-1}(A_P)_{ij}(V^TY)_j + (b_P(Z^*))_i}{(A_P)_{ip}}$$

$$\mathcal{V}_{P}^{0}\left((V^TY)_{1:P-1}, Z^*\right) = \max_{i:(A_P)_{ip}=0} -\sum_{j=1}^{P-1}(A_P)_{ij}(V^TY)_j + (b_P(Z^*))_i.$$

□

Appendix C

Appendix to Chapter 3

C.1 Simulation Parameters

Ascertainment probabilitiy

$$
\begin{aligned}
X_i\beta_2 = &-0.841 + 1.384X_{i1} - 0.158X_{i13} - 1.072X_{i14} + 1.499X_{i27} \\
&- 1.010X_{i36} - 2X_{i37} - 0.143X_{i39} + 1.55X_{i40} \\
X_i\beta_3 = &-0.113 + 0.233X_{i2} - 1.272X_{i11} - 0.205X_{i13} + 0.347X_{i15} \\
&+ 0.0323X_{i16} + 0.121X_{i20} + 0.926X_{i33} - 1.062X_{i34} \\
X_i\beta_4 = &-1.351 - 0.202X_{i5} + 1.059X_{i10} + 1.243X_{i13} + 0.732X_{i20} \\
&+ 0.456X_{i21} + 0.648X_{i24} - 0.392X_{i35} + 0.752X_{i37}.
\end{aligned}
$$

C.2 Proof of Claim 1

Proof of Claim 1.

$$\begin{aligned}
E[\epsilon_i|X_i,A_i] =& E[Y_i(A_i)|A_i,X_i] - E\left[\sum_{k=1}^{K}\left\{\mu_{k(0)}(X_i) + A_i\tau_k(X_i)\right\}p(k|X_i)\middle|A_i,X_i\right]\\
=& E[Y_i(A_i)|A_i,X_i] - \sum_{k=1}^{K} p(k|X_i)E[\underbrace{E[Y_i(0)|X_i,S_i=k]}_{\mu_{k(0)}(X_i)}|A_i,X_i]\\
&- A_i\sum_{k=1}^{K} p(k|X_i)E[\underbrace{E[Y_i(1)-Y_i(0)|X_i,S_i=k]}_{\tau_k(X_i)}|A_i,X_i]\\
=& E[Y_i(A_i)|A_i,X_i] - (1-A_i)\sum_{k=1}^{K} p(k|X_i)E[E[Y_i(0)|X_i,S_i=k]|A_i,X_i]\\
&- A_i\sum_{k=1}^{K} p(k|X_i)E[E[Y_i(1)|X_i,S_i=k]|A_i,X_i]\\
=& E[Y_i(A_i)|A_i,X_i] - (1-A_i)\sum_{k=1}^{K} p(k|X_i)E[E[Y_i(0)|A_i=0,X_i,S_i=k]|A_i,X_i]\\
&- A_i\sum_{k=1}^{K} p(k|X_i)E[E[Y_i(1)|A_i=1,X_i,S_i=k]|A_i,X_i]\\
=& 0
\end{aligned}$$

The second to last equality holds by *Assumption* 2. The last equality holds because for $A_i = a \in \{0,1\}$, we have

$$\begin{aligned}
&E[Y_i(a)|A_i=a,X_i] - \sum_{k=1}^{K} p(k|X_i)E[E[Y_i(a)|A_i=a,X_i,S_i=k]|A_i=a,X_i]\\
=& E[Y_i(a)|A_i=a,X_i] - \sum_{k=1}^{K} p(k|X_i)E[Y_i(a)|A_i=a,X_i]\\
=& E[Y_i(a)|A_i=a,X_i] - E[Y_i(a)|A_i=a,X_i]\\
=& 0.
\end{aligned}$$

□

C.3 Proof of Lemma 1

Proof of Lemma 1. Recall that

$$\widehat{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big[\{Y_i - \widehat{m}^{-q(i)}(X_i)\} - \widehat{u}_i^\top \beta\Big]^2$$

and

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big[\{Y_i - m(X_i)\} - u_i^\top \beta\Big]^2.$$

We can re-write $\widehat{L}_n(\beta)$ as

$$\widehat{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\{Y_i - \widehat{m}^{-q(i)}(X_i)\} - \sum_{k=1}^{K}\{A_i - \widehat{e}_k^{-q(i)}(X_i)\}p(k|X_i)v_k(X_i)^\top \beta_k\right]^2,$$

where $\beta_k \in \mathbb{R}^{d_k}$ is the vector of coefficients for study k. Similarly, we can re-write $L_n(\beta)$ as

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\{Y_i - m(X_i)\} - \sum_{k=1}^{K}\{A_i - e_k(X_i)\}p(k|X_i)v_k(X_i)^\top \beta_k\right]^2.$$

We define the following notation:

$$\begin{aligned}
A_{m,i} &= m(X_i) - \widehat{m}^{-q(i)}(X_i) \\
A_{e_k,i} &= e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i) \qquad k = 1,\ldots,K \\
B_{m,i} &= Y_i - m(X_i) \\
B_{e_k,i} &= A_i - e_k(X_i) \qquad k = 1,\ldots,K \\
g_k(X_i, \beta_k) &= p(k|X_i)v_k(X_i)^\top \beta_k \\
A_{e,i} &= \sum_{k=1}^{K} A_{e_k,i}\, g_k(X_i;\beta_k) \\
B_{e,i} &= \sum_{k=1}^{K} B_{e_k,i}\, g_k(X_i;\beta_k),
\end{aligned}$$

By algebra, we have

$$\begin{aligned}
\widehat{L}_n(\beta) =& \frac{1}{n}\sum_{i=1}^{n}\left[B_{m,i} + A_{m,i} - \sum_{k=1}^{K}(B_{e_k,i} + A_{e_k,i})g_k(X_i;\beta_k)\right]^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}[B_{m,i} + A_{m,i} - B_{e,i} - A_{e,i}]^2 \\
=& L_n(\beta) + \frac{1}{n}\sum_{i=1}^{n}[A_{m,i} - A_{e,i}]^2 + \frac{2}{n}\sum_{i=1}^{n}(B_{m,i} - B_{e,i})(A_{m,i} - A_{e,i}) \\
=& L_n(\beta) + \frac{1}{n}\sum_{i=1}^{n}A_{m,i}^2 + \frac{1}{n}\sum_{i=1}^{n}A_{e,i}^2 - \frac{2}{n}\sum_{i=1}^{n}A_{m,i}A_{e,i} \\
& + \frac{2}{n}\sum_{i=1}^{n}B_{m,i}A_{m,i} - \frac{2}{n}\sum_{i=1}^{n}B_{m,i}A_{e,i} - \frac{2}{n}\sum_{i=1}^{n}B_{e,i}A_{m,i} + \frac{2}{n}\sum_{i=1}^{n}B_{e,i}A_{e,i}
\end{aligned}$$

First term: $\frac{1}{n}\sum_{i=1}^{n}A_{m,i}^2 = \frac{1}{n}\sum_{i=1}^{n}(m(X_i) - \widehat{m}^{-q(i)}(X_i))^2$

By Markov's inequality and *Assumption 4*, $\frac{1}{n}\sum_{i=1}^{n}A_{m,i}^2$ is $O_p(a_n^2)$.

Second term: $\frac{1}{n}\sum_{i=1}^{n}A_{e,i}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}A_{e_k,i}g_k(X_i;\beta_k)\right]^2$

We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}A_{e,i}^2 =& \sum_{k=1}^{K}\left[\frac{1}{n}\sum_{i=1}^{n}(e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))^2 g_k(X_i;\beta_k)^2\right] \\
& + \sum_{k \neq k'}\left[\frac{2}{n}\sum_{i=1}^{n}(e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)(e_{k'}(X_i) - \widehat{e}_{k'}^{-q(i)}(X_i))g_{k'}(X_i;\beta_{k'})\right]
\end{aligned}$$

By Markov's inequality, *Assumption 3*, and *Assumption 4*, the first sum is $O_p(a_n^2)$. Moreover, the cross term $\frac{2}{n}\sum_{i=1}^{n}(e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)(e_{k'}(X_i) - \widehat{e}_{k'}^{-q(i)}(X_i))g_{k'}(X_i;\beta_{k'})$ for $k = 1, \ldots, K$ is also $O_p(a_n^2)$. Therefore, we can conclude that $\frac{1}{n}\sum_{i=1}^{n}A_{e,i}^2$ is $O_p(a_n^2)$.

Third term: $\frac{1}{n}\sum_{i=1}^{n}A_{m,i}A_{e,i}$

We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} A_{m,i}A_{e,i} &= \frac{1}{n}\sum_{i=1}^{n}\left\{(m(X_i)-\widehat{m}^{-q(i)}(X_i))\left[\sum_{k=1}^{K}(e_k(X_i)-\widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)\right]\right\} \\
&\leq \frac{C_1}{n}\sum_{i=1}^{n}\left\{(m(X_i)-\widehat{m}^{-q(i)}(X_i))\left[\sum_{k=1}^{K}(e_k(X_i)-\widehat{e}_k^{-q(i)}(X_i))\right]\right\} \\
&\leq C_1\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left\{m(X_i)-\widehat{m}^{-q(i)}(X_i)\right\}^2\right)\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}(e_k(X_i)-\widehat{e}_k^{-q(i)}(X_i))\right]^2\right\}} \\
&= O_p(a_n^2)
\end{aligned}$$

for some positive constant C_1. The second line holds by *Assumption 3*, and the third line holds by Cauchy-Schwarz inequality. The last line holds by Markov's inequality, *Assumption 3*, and *Assumption 4*.

Fourth term: $\frac{1}{n}\sum_{i=1}^{n} B_{m,i}A_{m,i} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m(X_i))(m(X_i)-\widehat{m}^{-q(i)}(X_i))$

We define

$$B_{mm}^{q} = \frac{1}{|\{i : q(i) = q\}|}\sum_{i:q(i)=q} B_{m,i}A_{m,i}$$

to be the sample average of $B_{m,i}A_{m,i}$ in the qth cross-fitting fold. By the triangle inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - m(X_i))(m(X_i)-\widehat{m}^{-q(i)}(X_i))\right| \leq \sum_{q=1}^{Q}|B_{mm}^{q}|.$$

Therefore, it suffices to show that $B_{mm}^{q} = O_p(a_n^2)$. Let $\mathcal{I}^{-q} = \{X_i, A_i, Y_i, S_i : q(i) \neq q\}$ denote the set of observations that do not belong to the same data fold as observation i. B_{mm}^{q}'s expectation is

$$\begin{aligned}
E\left(B_{mm}^{q}\right) &= E(B_{m,i}A_{m,i}) \\
&= E\left(E\left[B_{m,i}A_{m,i}|\mathcal{I}^{-q}, X_i\right]\right) \\
&= E(A_{m,i}E\left[B_{m,i}|\mathcal{I}^{-q}, X_i\right]) \\
&= 0,
\end{aligned}$$

where the last line follows by the definition of $B_{m,i}$.

Next, its variance is

$$\begin{aligned}
Var\left(B^q_{mm}\right) &= E\left\{\left(B^q_{mm}\right)^2\right\} \\
&= \frac{E\left\{\sum_{i:q(i)=q} B^2_{m,i}A^2_{m,i} + \sum_{i\neq j:q(i)=q,q(j)=q} B_{m,i}B_{m,j}A_{m,i}A_{m,j}\right\}}{|\{i : q(i) = q\}|^2} \\
&= \frac{E\left(B^2_{m,i}A^2_{m,i}\right)}{|\{i : q(i) = q\}|} + \frac{\sum_{i\neq j:q(i)=q,q(j)=q} E\left(B_{m,i}B_{m,j}A_{m,i}A_{m,j}\right)}{|\{i : q(i) = q\}|^2}
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
E\left(B^2_{m,i}A^2_{m,i}\right) &= E\left(E\left[B^2_{m,i}A^2_{m,i}|\mathcal{I}^{-q}, X_i\right]\right) \\
&= E\left(A^2_{m,i}E\left[B^2_{m,i}|\mathcal{I}^{-q}, X_i\right]\right) \\
&\leq C_2E(A^2_{m,i}) \\
&= O_p(a^2_n)
\end{aligned}$$

for some positive constant C_2. The second to last line holds from *Assumption 3*. And the last line holds from *Assumption 4*.

For the second term, we have

$$\begin{aligned}
E\left(B_{m,i}B_{m,j}A_{m,i}A_{m,j}\right) &= E\left[E\left(B_{m,i}B_{m,j}A_{m,i}A_{m,j}|\mathcal{I}^{-q}, X_i\right)\right] \\
&= E\left[A_{m,i}A_{m,j}E\left(B_{m,i}B_{m,j}|\mathcal{I}^{-q}, X_i\right)\right] \\
&= E\left[A_{m,i}A_{m,j}E\left(B_{m,j}\right)E\left(B_{m,i}|\mathcal{I}^{-q}, X_i\right)\right] \\
&= 0
\end{aligned}$$

The second to last line follows because $B_{m,i}$ is independent of $B_{m,j}$ for $i \neq j$. The last line follows by the definition of $B_{m,i}$. Therefore, we have that

$$Var(B^q_{mm}) = \frac{Q}{n}O(a_n^2) = O(a_n^2/n),$$

where the first equality holds if the Q folds have equal number of observations (i.e., n/Q for each fold). Then by Chebychev's inequality, $\frac{1}{n}\sum_{i=1}^n B_{m,i}A_{m,i} = O_p(a_n^2/n)$.

Fifth term: $\frac{1}{n}\sum_{i=1}^n B_{m,i}A_{e,i} = \frac{1}{n}\sum_{i=1}^n \left[(Y_i - m(X_i))\left\{\sum_{k=1}^K (e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)\right\}\right]$

We define

$$B^q_{me} = \frac{1}{|\{i : q(i) = q\}|}\sum_{i:q(i)=q} B_{m,i}A_{e,i}$$

to be the sample average of $B_{m,i}A_{e,i}$ in the qth cross-fitting fold. By the triangle inequality,

$$\left|\frac{1}{n}\sum_{i=1}^n \left[(Y_i - m(X_i))\left\{\sum_{k=1}^K (e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)\right\}\right]\right| \leq \sum_{q=1}^Q |B^q_{me}|.$$

Therefore, it suffices to show that $B^q_{me} = O_p(a_n^2)$. Its expectation is

$$\begin{aligned}
E\left(B^q_{me}\right) &= E(B_{m,i}A_{e,i}) \\
&= E\left(E\left[B_{m,i}A_{e,i}|\mathcal{I}^{-q}, X_i\right]\right) \\
&= E(A_{e,i}E\left[B_{m,i}|\mathcal{I}^{-q}, X_i\right]) \\
&= 0,
\end{aligned}$$

where the last line follows by the definition of $B_{m,i}$.

Next, its variance is

$$\begin{aligned}
Var\left(B_{me}^{q}\right) &= E\left\{\left(B_{me}^{q}\right)^{2}\right\} \\
&= \frac{E\left\{\sum_{i:q(i)=q} B_{m,i}^{2} A_{e,i}^{2} + \sum_{i\neq j:q(i)=q,q(j)=q} B_{m,i} B_{m,j} A_{e,i} A_{e,j}\right\}}{|\{i : q(i) = q\}|^{2}} \\
&= \frac{E\left(B_{m,i}^{2} A_{e,i}^{2}\right)}{|\{i : q(i) = q\}|} + \frac{\sum_{i\neq j:q(i)=q,q(j)=q} E\left(B_{m,i} B_{m,j} A_{e,i} A_{e,j}\right)}{|\{i : q(i) = q\}|^{2}}
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
E\left(B_{m,i}^{2} A_{e,i}^{2}\right) &= E\left(E\left[B_{m,i}^{2} A_{e,i}^{2} | \mathcal{I}^{-q}, X_i\right]\right) \\
&= E\left(A_{e,i}^{2} E\left[B_{m,i}^{2} | \mathcal{I}^{-q}, X_i\right]\right) \\
&\leq C_2 E(A_{e,i}^{2}) \\
&= O_p(a_n^2)
\end{aligned}$$

for some positive constant C_2. The second to last line holds from *Assumption 3*. And the last line holds from *Assumption 4* and Markov's inequality.

For the second term, we have

$$\begin{aligned}
E\left(B_{m,i} B_{m,j} A_{e,i} A_{e,j}\right) &= E\left[E\left(B_{m,i} B_{m,j} A_{e,i} A_{e,j} | \mathcal{I}^{-q}, X_i\right)\right] \\
&= E\left[A_{e,i} A_{e,j} E\left(B_{m,i} B_{m,j} | \mathcal{I}^{-q}, X_i\right)\right] \\
&= E\left[A_{e,i} A_{e,j} E\left(B_{m,j}\right) E\left(B_{m,i} | \mathcal{I}^{-q}, X_i\right)\right] \\
&= 0
\end{aligned}$$

The second to last line follows because $B_{m,i}$ is independent of $B_{m,j}$ for $i \neq j$. The last line follows by the definition of $B_{m,i}$. Therefore, we have that

$$Var(B_{me}^{q}) = \frac{Q}{n} O(a_n^2) = O(a_n^2/n),$$

where the first equality holds if the Q folds have equal number of observations (i.e., n/Q for each fold). Then by Chebychev's inequality, $\frac{1}{n}\sum_{i=1}^{n} B_{m,i}A_{e,i} = O_p(a_n^2/n)$.

Sixth term: $\frac{1}{n}\sum_{i=1}^{n} B_{e,i}A_{m,i} = \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\sum_{k=1}^{K}(A_i - e_k(X_i))g_k(X_i;\beta_k)\right\}(m(X_i) - \widehat{m}^{-q(i)}(X_i))\right]$

We define

$$B_{em}^{q} = \frac{1}{|\{i : q(i) = q\}|}\sum_{i:q(i)=q} B_{e,i}A_{m,i}$$

to be the sample average of $B_{e,i}A_{m,i}$ in the qth cross-fitting fold. By the triangle inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\sum_{k=1}^{K}(A_i - e_k(X_i))g_k(X_i;\beta_k)\right\}(m(X_i) - \widehat{m}^{-q(i)}(X_i))\right]\right| \leq \sum_{q=1}^{Q}|B_{em}^{q}|.$$

Therefore, it suffices to show that $B_{em}^{q} = O_p(a_n^2)$. Its expectation is

$$\begin{aligned}
E(B_{em}^{q}) &= E(B_{e,i}A_{m,i}) \\
&= E(E\left[B_{e,i}A_{m,i}|\mathcal{I}^{-q}, X_i, S_i = k\right]) \\
&= E(A_{m,i}E\left[B_{e,i}|\mathcal{I}^{-q}, X_i, S_i = k\right]) \\
&= 0
\end{aligned}$$

where the last line follows because

$$\begin{aligned}
E[B_{e,i}|\mathcal{I}^{-q}, X_i, S_i = k] &= \sum_{k=1}^{K} g_k(X_i;\beta_k)E\left[A_i - e_k(X_i)|X_i, S_i = k\right] \\
&= \sum_{k=1}^{K} g_k(X_i;\beta_k)\left\{E\left[A_i|X_i, S_i = k\right] - e_k(X_i)\right\} \\
&= \sum_{k=1}^{K} g_k(X_i;\beta_k)\left\{e_k(X_i) - e_k(X_i)\right\} \\
&= 0
\end{aligned}$$

Next, its variance is

$$
\begin{aligned}
Var\left(B_{em}^{q}\right) &= E\left\{\left(B_{em}^{q}\right)^{2}\right\} \\
&= \frac{E\left\{\sum_{i:q(i)=q} B_{e,i}^{2}A_{m,i}^{2} + \sum_{i\neq j:q(i)=q,q(j)=q} B_{e,i}B_{e,j}A_{m,i}A_{m,j}\right\}}{|\{i:q(i)=q\}|^{2}} \\
&= \frac{E\left(B_{e,i}^{2}A_{m,i}^{2}\right)}{|\{i:q(i)=q\}|} + \frac{\sum_{i\neq j:q(i)=q,q(j)=q} E\left(B_{e,i}B_{e,j}A_{m,i}A_{m,j}\right)}{|\{i:q(i)=q\}|^{2}}
\end{aligned}
$$

For the first term, we have

$$
\begin{aligned}
E\left(B_{e,i}^{2}A_{m,i}^{2}\right) &= E\left(E\left[B_{e,i}^{2}A_{m,i}^{2}|\mathcal{I}^{-q},X_i\right]\right) \\
&= E\left(A_{m,i}^{2}E\left[B_{e,i}^{2}|\mathcal{I}^{-q},X_i\right]\right) \\
&\leq C_3 E(A_{m,i}^{2}) \\
&= O_p(a_n^{2})
\end{aligned}
$$

for some positive constant C_3. The second to last line holds from *Assumption 3*, and the last line holds from *Assumption 4*.

For the second term, we have

$$
\begin{aligned}
E(B_{e,i}B_{e,j}A_{m,i}A_{m,j}) &= E[E(B_{e,i}B_{e,j}A_{m,i}A_{m,j}|\mathcal{I}^{-q},X_i,S_i=k)] \\
&= E[A_{m,i}A_{m,j}E(B_{e,i}B_{e,j}|\mathcal{I}^{-q},X_i,S_i=k)] \\
&= E[A_{m,i}A_{m,j}E(B_{e,i})E(B_{e,j}|\mathcal{I}^{-q},X_i,S_i=k)] \\
&= 0
\end{aligned}
$$

The second to last line follows because $B_{e,i}$ is independent of $B_{e,j}$ for $i \neq j$. The last line follows by the definition of $B_{e,i}$. Therefore, we have that

$$
Var(B_{em}^{q}) = \frac{Q}{n}O(a_n^{2}) = O(a_n^{2}/n)
$$

where the first inequality holds if the Q folds have equal number of observations (i.e, n/Q for each fold). Then by Chebychev's inequality, $\frac{1}{n}\sum_{i=1}^{n} B_{e,i}A_{m,i} = O_p(a_n^2/n)$.

Seventh term:

$$\frac{1}{n}\sum_{i=1}^{n} B_{e,i}A_{e,i} = \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\sum_{k=1}^{K}(A_i - e_k(X_i))g_k(X_i;\beta_k)\right\}\left\{\sum_{k=1}^{K}(e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)\right\}\right]$$

We define

$$B_{ee}^q = \frac{1}{|\{i : q(i) = q\}|}\sum_{i:q(i)=q} B_{e,i}A_{e,i}$$

to be the sample average of $B_{e,i}A_{e,i}$ in the qth cross-fitting fold. By the triangle inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\sum_{k=1}^{K}(A_i - e_k(X_i))g_k(X_i;\beta_k)\right\}\left\{\sum_{k=1}^{K}(e_k(X_i) - \widehat{e}_k^{-q(i)}(X_i))g_k(X_i;\beta_k)\right\}\right]\right| \leq \sum_{q=1}^{Q}|B_{ee}^q|.$$

Therefore, it suffices to to show that $B_{ee}^q = O_p(a_n^2)$. Its expectation is

$$\begin{aligned}
E(B_{ee}^q) &= E(B_{e,i}A_{e,i}) \\
&= E(E[B_{e,i}A_{e,i}|\mathcal{I}^{-q}, X_i, S_i = k]) \\
&= E(A_{e,i}E[B_{e,i}|\mathcal{I}^{-q}, X_i, S_i = k]) \\
&= 0
\end{aligned}$$

Next, its variance is

$$\begin{aligned}
Var(B_{ee}^q) &= E\left\{(B_{ee}^q)^2\right\} \\
&= \frac{E\left\{\sum_{i:q(i)=q} B_{e,i}^2A_{e,i}^2 + \sum_{i\neq j:q(i)=q,q(j)=q} E(B_{e,i}B_{e,j}A_{e,i}A_{e,j})\right\}}{|\{i : q(i) = q\}|^2} \\
&= \frac{E\left(B_{e,i}^2A_{e,i}^2\right)}{|\{i : q(i) = q\}|} + \frac{\sum_{i\neq j:q(i)=q,q(j)=q} E(B_{e,i}B_{e,j}A_{e,i}A_{e,j})}{|\{i : q(i) = q\}|^2}
\end{aligned}$$

For the first term, we have

$$
\begin{aligned}
E(B_{e,i}^2 A_{e,i}^2) &= E(E[B_{e,i}^2 A_{e,i}^2 | \mathcal{I}^{-q}, X_i]) \\
&= E(A_{e,i}^2 E[B_{e,i}^2 | \mathcal{I}^{-q}, X_i]) \\
&\leq C_3 E(A_{e,i}^2) \\
&= O_p(a_n^2)
\end{aligned}
$$

The second to last line holds from *Assumption 3*, and the last line holds from *Assumption 3* and *Assumption 4*.

For the second term, we have

$$
\begin{aligned}
E(B_{e,i} B_{e,j} A_{e,i} A_{e,j}) &= E[E(B_{e,i} B_{e,j} A_{e,i} A_{e,j} | \mathcal{I}^{-q}, X_i, S_i = k)] \\
&= E[A_{e,i} A_{e,j} E(B_{e,i} B_{e,j} | \mathcal{I}^{-q}, X_i, S_i = k)] \\
&= E[A_{e,i} A_{e,j} E(B_{e,i}) E(B_{e,j} | \mathcal{I}^{-q}, X_i, S_i = k)] \\
&= 0
\end{aligned}
$$

The second to last line follows because $B_{e,i}$ is independent of $B_{e,j}$ for $i \neq j$. The last line follows by the definition of $B_{e,i}$. Therefore, we have that

$$
Var(B_{ee}^q) = \frac{Q}{n} O(a_n^2) = O(a_n^2/n)
$$

where the first inequality holds if the Q folds have equal number of observations (i.e, n/Q for each fold). Then by Chebychev's inequality, $\frac{1}{n}\sum_{i=1}^n B_{e,i} A_{e,i} = O_p(a_n^2/n)$.

Altogether, $\widehat{L}_n(\beta) - L_n(\beta)$ is dominated by the $O_p(a_n^2)$ term $\frac{1}{n}\sum_{i=1}^n A_{m,i}^2 + \frac{1}{n}\sum_{i=1}^n A_{e,i}^2 - \frac{2}{n}\sum_{i=1}^n A_{m,i} A_{e,i}$, so $\widehat{L}_n(\beta) - L_n(\beta) = O_p(a_n^2)$ □

C.4 Proof of Theorem 1

Proof of Theorem 1. By Lemma 1, we have

$$\widehat{\beta} = \arg\min_{b} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i) - u_i^{\intercal} b\}^2 + O_p(a_n^2).$$

Under Assumptions 5-6, Lemma 4.1 (Pointwise Linearization) in Belloni *et al.* (2015) states that for any $\alpha \in V^{d-1}$ where V^{d-1} is the space of vectors α such that $\|\alpha\| = 1$,

$$\sqrt{n}\alpha^{\intercal}(\widehat{\beta} - \beta) = \alpha^T \mathbb{G}_n[u_i \epsilon_i] + o_P(1) + O_p(\sqrt{n}a_n^2),$$

where $\mathbb{G}_n[f(w_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(w_i) - E[f(w_i)])$. Under *Assumption 4*, $O_p(\sqrt{n}a_n^2)$ is negligible compared to $o_p(1)$. Therefore, we have

$$\sqrt{n}\alpha^{\intercal}(\widehat{\beta} - \beta) = \alpha^T \mathbb{G}_n[u_i \epsilon_i] + o_P(1).$$

Recall $Q = E(u_i u_i^{\intercal})$. By Theorem 4.2 (Pointwise Normality) in Belloni *et al.* (2015), we have that for any $\alpha \in V^{d-1}$,

$$\sqrt{n}\frac{\alpha^{\intercal}(\widehat{\beta} - \beta)}{\left\|\alpha^{\intercal}\Omega^{1/2}\right\|} \xrightarrow{d} N(0,1) + o_P(1),$$

where $\Omega := Q^{-1}E[\epsilon_i^2 u_i u_i^{\intercal}]Q^{-1}$. Moreover, for any $x \in \mathcal{X}$, if we take $\alpha = Z(x)v(x)$ and $s(x) = \Omega^{1/2}Z(x)v(x)$, then

$$\sqrt{n}\frac{(Z(x)v(x))^{\intercal}(\widehat{\beta} - \beta)}{\|s(x)\|} \xrightarrow{d} N(0,1) + o_P(1).$$

Under *Assumption 6(e)*, we have

$$\sqrt{n}\frac{\widehat{\tau}(x) - \tau(x)}{\|s(x)\|} \xrightarrow{d} N(0,1) + o_P(1).$$

□