Practicable Characterization of Systematic Heterogeneity

Practicable Characterization of Systematic Heterogeneity

a dissertation presented
by
Sarah Chika Anoke
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
April 2017

Dissertation advisor: Professor Corwin M. Zigler Sarah Chika Anoke

Practicable Characterization of Systematic Heterogeneity

Abstract

In public health, personalized medicine is the ideal. For example, an effective strategy for improving the health of a population is to measure the health of constituent subpopulations and intervene where the treatment is most needed. Alternatively, a member of a subpopulation presents with an ailment and relevant covariates are used to determine the appropriate treatment. Such strategies reasonably assume that there is heterogeneity in the effect of the treatment on health across subpopulations. However identification of heterogeneity tends to be expensive, understandably so due to the demands we are making of our data. Costs appear when having to make strong *a priori* assumptions about the number and identifying characteristics of the subpopulations across which the treatment effect differs, in the increased sample size required for the data to fill in gaps left by the absence of assumptions, and/or in the manual evaluation of large numbers of covariates. This dissertation discusses different approaches to reducing this cost. Chapter 1 compares a Lot Quality Assurance Sampling (LQAS) survey conducted in southwestern Uganda to an unaffiliated but coincident Demographic Health Survey (DHS) and shows that if we redefine our goal in terms of the programmatic decisions we need to make, we can come to the same conclusions at a fraction of the cost. In Chapter 2 I consider just how expensive it is to identify heterogeneity in the absence of *a priori* assumptions, and draw some general conclusions about the capabilities and limitations of extant modern methods of causal inference. I conclude with Chapter 3, where I leverage the insights of Chapter 2 to build a visualization application that facilitates the exploratory, hypothesis-generating analysis of treatment effect heterogeneity (TEH), particularly for large datasets where a manual evaluation of covariates is not practicable.

Contents

List of figures

List of tables

Acknowledgments

In addition to what you will read on the following pages, along the way I have learned so much about mentorship as the mentee of incredible scholars. I am deeply grateful to my advisor Cory Zigler for his support, teaching, and guidance on life within and outside my dissertation. Sincere thanks also goes to Committee Member Giovanni Parmigiani, whose depth and breadth of knowledge I aspire to. I would also like to thank Committee Member Sherri Rose whose thoughtful questions trained me to think like a statistician, and to use my background as an asset. It is not possible to fully describe the immense support of Marcello Pagano, also a Committee Member, during my time as a graduate student, but I will forever be appreciative. Dr. Christine Choirat deserves a very special and sincere thanks for being the bridge between the content I learned in my courses and the skills & considerations required of a software developer.

My path to this point would not have been possible without the support of my undergraduate advisor Dr. Gregg Tucci. Another special note of thanks to Dr. Nicholas Horton, who has also been very supportive of my development as a biostatistician. Dr. Joe Blitzstein was also instrumental in my success.

Finally, my family (Mom, Dad, Jane, brothers, JayBear), friends (309, PKB), and colleagues (Anders Huitfeldt, Jessica Gronsbell) have supported me throughout my time as a student, and I am very thankful.

1

Comparing two survey methods of measuring health-related indicators: Lot Quality Assurance Sampling and Demographic Health Surveys

Sarah C. Anoke[1], Paul Mwai[1], Caroline Jeffery[2], Joseph J. Valadez[2], Marcello Pagano[1]

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

2 *Department of International Public Health, Liverpool School of Tropical Medicine, Liverpool, UK*

Abstract

Two common methods used to measure indicators for health program monitoring and evaluation are the demographic and health surveys (DHS) and lot quality assurance sampling (LQAS); each one has different strengths. We report on both methods when utilized in comparable situations. We compared 24 indicators in south-west Uganda, where data for prevalence estimations were collected independently for the two methods in 2011 (LQAS: $n = 8876$; DHS: $n = 1200$). Data were stratified (e.g., gender and age) resulting in 37 comparisons. We used a two-sample two-sided z-test of proportions to compare both methods. The average difference between LQAS and DHS for 37 estimates was 0.062 (SD = 0.093; median = 0.039). The average difference among the 21 failures to reject equality of proportions was 0.010 (SD = 0.041; median = 0.009); among the 16 rejections, it was 0.130 (SD = 0.010, median = 0.118). Seven of the 16 rejections exhibited absolute differences of <0.10, which are clinically (or managerially) not significant; 5 had differences >0.10 and <0.20 (mean = 0.137, SD = 0.031) and four differences were >0.20 (mean = 0.261, SD = 0.083). There is 75.7% agreement across the two surveys. Both methods yield regional results, but only LQAS provides information at less granular levels (e.g. the district level) where managerial action is taken. The cost advantage and localization make LQAS feasible to conduct more frequently, and provides the possi-

bility for real-time health outcomes monitoring. This work was supported by NIH Grants 5T32AI007358-24, 5T32AI007358-25.

1.1 Introduction

The importance of monitoring and evaluation (M&E) to assess interventional programs, inform allocation of resources and improve evidence-based policy has been commented on by several authors [13, 16, 48]. Two common sampling and survey methodologies used to track health program indicators for M&E are the demographic and health surveys (DHS) [7] and lot quality assurance sampling (LQAS) [74].

DHS and LQAS differ in structure because they serve different purposes: DHS for international comparisons and benchmarking, LQAS for intranational comparisons, benchmarking and health system management. A unique benefit of LQAS is the 'locality' of the methodology. LQAS gives local (e.g., subdistrict, county or subcounty) information, which, if need be, can subsequently be further aggregated into district and regional information. The disaggregation helps overcome the ecological fallacy problem, the assumption that all subregions perform at the regional mean. Additionally, LQAS gives more distributive information about how the subregional estimates vary across the region, which allows for identification of geographical disparities.

Further, LQAS surveys are shorter, cheaper to implement, and the data obtained are readily available. With regard to this last point, LQAS data are hand tabulated within a week of data collection to permit district managers to classify subdistrict units according to predetermined coverage targets; also, more formal reports with districts and regional prevalence measures can be produced within 6 weeks of data collection. Thus, the surveys can be done more frequently, perhaps within the three-to five-year interim between DHS implementations. This increased frequency of measurement allows LQAS data to be used for health system management whereas DHS data, because of the need for international consistency, take several months after collection to process and several additional months to compile into a final report. The increased frequency of LQAS surveys also positively impacts the building of local capacity, because local district teams incorporate LQAS data collection into their regular health system responsibilities, whereas a DHS may temporarily employ individuals every few years.

An LQAS survey also is flexible and can be adapted to obtain information most useful for program management; survey items relevant to the region of implementation are easily added or removed, and these modifications do not hinder either the data collection process or the data analysis. Comparatively, a DHS is a large and expensive undertaking, making it difficult to modify the data collection and analysis process.

This inertia, combined with the DHS' occasional reference as a 'gold standard', underscores the importance of identifying the best use of a specific survey tool, rather than assuming it serves all informational purposes.

Finally, another advantage of LQAS is that the data are almost real time in that the data collectors see the immediate and local impact of the data they collect, as opposed to a detached central 'black box' repository and its distant possible impact on health policy. This may favorably affect the quality of the data, and it certainly influences the cost of providing national, or aggregated summaries, as the inputs to such summaries are the data that were gathered to provide local information, an aim that presumably justifies the cost of obtaining the data. Thus, the marginal cost of aggregation is minimal compared to the cost of acquiring the data.

The goal of this study is to provide substantive evidence to support the above claims about LQAS' relative utility, by conducting a formal statistical comparison of indicators common between the two surveys. These indicators cover several aspects of Ugandan public health, such as HIV prevention, malaria treatment and prevention, family planning and reproductive health, sanitation, maternal, newborn and child health, and nutrition.

1.2 Methods

1.2.1 Selection of region and indicators for comparison

We selected Uganda for this comparison because data exist from both DHS and LQAS surveys collected around the same time: between July and August 2011 for the LQAS, and between June and December 2011 for the DHS.

DHS is a national survey; the sample collected represents all 112 districts in Uganda. Seventy-eight of these districts are engaged in USAID-funded projects that use LQAS for their monitoring. The best geographic overlap between the two surveys is in the DHS-defined south-west region, where LQAS surveys were conducted in each of this region's constituent districts. In this study, we compare indicators calculated for the south-west region.

The choice of indicators to compare started with a 'core set' of 59 national indicators created to track social service performance in Uganda. This list was created by a Technical Working Group of the USAID-funded STAR-E LQAS project comprising representatives from several Ugandan institutions, projects and programs. Twenty-five LQAS indicators had definitions comparable to those contained in the DHS Final Report; we report on 24 of these comparisons. We replicated all but one DHS result using the DHS data set supplied by Inner City Fund (ICF) International. The indicator for which we could not reproduce the reported DHS estimate and the 33 LQAS indicators we did not find within the DHS Final Report were omitted.

1.2.2 Sampling schemes and data collection

The DHS Program is implemented by ICF International under contract from the U.S. Agency for International Development (USAID) [7]. The Program administers several surveys internationally, including the eponymous demographic and health survey (DHS). Although there is a general structure, each survey is tailored to the needs of the specific country. Here, we discuss the structure of the 2011 Ugandan DHS (UDHS), which was implemented jointly with the Uganda Bureau of Statistics (UBOS).

As discussed in the 2011 UDHS Final Report, the sample for the 2011 UDHS was designed 'to provide population and health indicator estimates for the country as a whole and for urban and rural areas separately' as well as for 10 regions, whose boundaries are administratively defined by the DHS Program [73]. This two-stage stratified cluster sample was selected by sampling households in each of 405 clusters, where stratification was by urban/rural status and region. The sampling frame for the selection of the clusters was the 2002 Population Census provided by UBOS. A three-month household listing operation was conducted in the 405 selected clusters, starting in April 2011. Data collection took place over a six-month period, from the end of June 2011 to early December 2011. Women aged 15-49 years in all households and men aged 15-54 in one-third of households were eligible for interview.

In the first stage of sampling within the south-west region, 40 clusters were selected from a total of 8369 with 7983 being rural and 386 urban. The 40 selected clusters comprise five urban and 35 rural areas. In the second stage of sampling, the DHS sampled 1200 of 685,695 households; 150 were urban and 1050 rural. The expected number of completed interviews for the region was 1097 (96.3% completed) for women 15-49 years and 477 (92.6% completed) for men 15-54 years. We report the actual sample sizes with the results. The national DHS first stage of sampling comprised 405 clusters selected from 48,715 clusters (42,675 rural, 6040 urban), and included 119 urban and 286 rural areas. The second stage comprised 12,150 households (8580 rural, 3570 urban) of 5,076,534 households. The expected number of completed interviews was 9885 for women 15-49 and 3628 for men 15-54.

The three subsurveys of interest are the household survey, the women's survey (asked in all households), and the men's survey (asked in approximately every third household). All three subsurveys were conducted within the same household. The LQAS methodology is a health science derivative of Statistical Quality Control, a set of tools developed by Dodge and Romig, and Shewhart [50]. The data are sampled from a local administrative unit called a supervision area (SA; e.g., county, subcounty or parish within a district), which is classified as 'acceptable' or 'unacceptable' according to a coverage target. Although the goal is classification, it is also possible to aggregate SA-level data to construct prevalence estimates for the respective districts and regions; here, the classification decisions do not in any way impact the estimation of indicators [74, 50].

LQAS in Uganda during 2011 included more than 11,400 interviewees; the sample of 8676 households in the south-west was also selected using a stratified two-stage process. Districts in south-west region were divided into SAs based on how the district managed health services. Within each SA, a sample of 19 or 24 villages was selected with probability of selection into the sample proportional to the village population size (PPS). To maintain an approximate minimum district sample size of 96, districts with only 4 SAs required an SA sample size of 24 (4 SAs 9 $\times$ 24). In each selected village, the interviewer constructed a map of the village with the help of a chief or other local leader, and divided the map into equivalent segments based on visible landmarks and the number of households in each segment. One segment was selected randomly. The interviewer then enumerated the households in the selected segment and selected one randomly. If the selected segment had 30 or more households, it was further segmented and a subsegment selected randomly; all households in the final segment were enumerated and one chosen randomly. To accommodate the fact that there could be a nearby household with zero probability of selection (e.g., it was omitted from the map because it was hidden behind vegetation), the next house with the closest door was selected for the first interview. Thereafter, the household with the next closest door was selected for each subsequent subpopulation. Only one individual from each subpopulation was interviewed in the sampled village.

The five subsurveys of interest correspond to particular subpopulations: mothers of children 0-11 months, mothers of children 12-23 months, women 15-49 years, men 15-54 years and youth 15-24 years. All five subsurveys were conducted in different households, comparatively different from what was employed by DHS. To accomplish this, from a randomly selected house, an interviewee is selected who is either a woman aged 15-49 years, a man aged 15-54 years, the mother of an infant aged 0-11 months, the mother of an infant aged 12-23 months or a youth aged 15-24 years. Subsequent households were selected to find interviewees from the remaining populations, taking care not to select two interviewees from the same household.

1.2.3 Weighting

Within both the UDHS and LQAS data sets, individuals had different probabilities of being sampled. To construct valid, representative estimates from these data, we calculated sampling weights based on each sampling design.

DHS In the 2011 UDHS, sampling weights were calculated based on the two-stage stratified cluster design used to sample households (see Appendix A.4 of [73] for details). These weights are provided within the 2011 UDHS data set.

LQAS In the LQAS data, we calculated weights based on the two-stage stratified design used to sample households. Within each SA, a fixed number (either 19 or 24) was sampled irrespective of the SA population size. To adjust for differences in SA sample sizes, individual observations are weighted by the number of individuals a response represents. For example, if an observation is one of 19 sampled from an SA with a population of 2000, then each observation is weighted by 2000/19. In another SA, if an observation is one of 24 sampled from an SA with a population of 4500, then each observation is weighted by 4500/24. We use these weights to construct a representative district point estimate, and a representative regional point estimate (Figure 1.1).

1.2.4 Sampling errors

DHS The DHS Program provides a formula in Appendix B of the 2011 UDHS Final Report [73] for calculating sampling errors based on the two-stage stratified cluster design used to sample individuals. For indicators considered to be of ‘primary interest’ by the DHS Program, sampling errors are provided in the report. Where possible, we use these sampling errors. For indicators where sampling errors are not provided, we calculated them using the formulae provided.

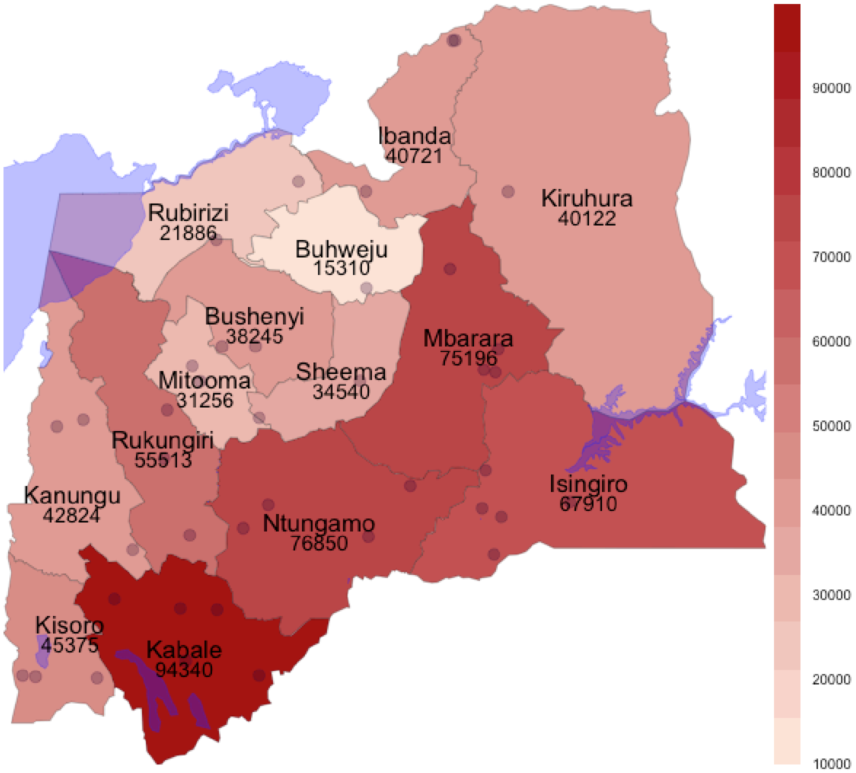


Figure 1.1 Population distribution by district, across the 14 districts of the south-west region. These population counts were calculated from LQAS sampling frames created during sampling of the data used in this writing, and were used to calculate the weighted regional prevalence estimate for each LQAS indicator. The 40 clusters that were sampled by the DHS Program for inclusion in their survey are denoted by translucent circles. The LQAS population counts are congruous with the distribution of DHS clusters, which were selected based on a distribution proportional to the population density.

LQAS The survey data software within Stata® 13 was used to calculate standard errors at both the district and regional levels [65]. For details on the formulae used, refer to the Stata Survey Data Reference Manual [64]. At the regional level, we used the Wilson score interval to construct confidence intervals [22].

1.2.5 Statistical comparison of indicators

A two-sample two-sided z-test of proportions was used to test whether the proportions as estimated from DHS data and LQAS data were statistically equivalent. Standard errors for test statistics were calculated by taking the square root of the sum of the squared standard errors from the two estimated proportions. In two cases (Table A.4 on page 60), it was necessary to calculate a weighted average and accompanying standard error of two LQAS subpopulation estimates for comparison to a single DHS measure. The weights used were the proportion of the aggregated sample that belonged to a particular subpopulation. For example, for an aggregated sample consisting of members from two subpopulations with 1353 and 752 members, respectively, the corresponding weights are 1353/(1353+752) and 752/(1353+752).

1.3 Results

1.3.1 Regional comparisons

The 24 selected indicators cover several aspects of Ugandan public health; including HIV knowledge, counseling, and behavior (8 indicators), malaria treatment and prevention (3), family planning & reproductive health (4), child health (3), nutrition (4) and sanitation (2). The results of the 37 comparisons are summarized as a forest plot (Figure 1.2).

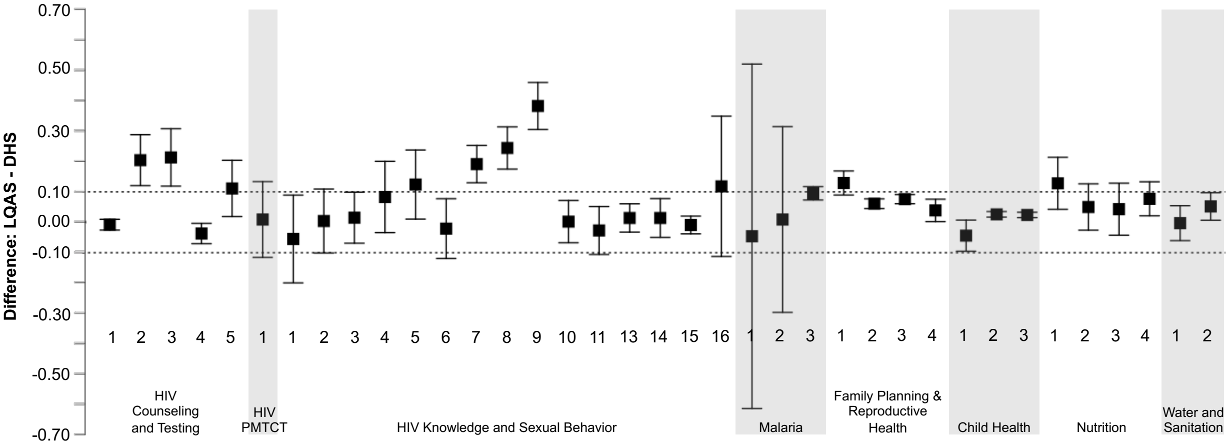


Figure 1.2 A forest plot of 37 comparisons of DHS and LQAS data collected in south-west Uganda during 2011.

Point estimates, confidence intervals and the results of statistical comparisons are shown in the Appendix (Tables A.1-A.8, starting on page 57). In Tables A.9 and A.10 (also in the Appendix, on pages 64 and 65, respectively), we summarize our comparisons. For 6 indicators (Table A.1 on page 57 and Table A.3 on page 59), we refine the comparison by making subpopulation comparisons (e.g., men, women, male youths, female youths) resulting in additional comparisons. In total, we assessed 38 comparisons; 1 comparison using a cohort of male youths (Table A.3 on page 59) was eliminated due to the UDHS having insufficient comparable data, thereby reducing the number of comparisons to 37. We did not reject equality of the proportions in 21 of 37 (56.8%). The average difference between LQAS and DHS estimates for the 37 comparisons was 0.062 (SD = 0.093; median = 0.039). The average difference among the 21 failures to reject equality of proportions was 0.010 (SD = 0.041; median = 0.009); among the 16 rejections, it was 0.130 (SD = 0.010, median = 0.118). As the large standard deviation, and lower median value compared to the mean indicate considerable variation among these rejections, we examined the variation further. Seven of the 16 rejections exhibited differences of <0.10, which are clinically (or managerially) not significant; five more had differences >0.10 and <0.20 (mean = 0.137, SD = 0.031) and 4 differences were >0.20 (mean = 0.261, SD = 0.083). We consider the more interesting of the 16 rejections in the Discussion below.

1.3.2 Distribution of prevalences across districts

The limit of inference when using UDHS data is at the regional level; however, district health system managers cannot use such results without making the strong assumption that the districts within the region perform similarly, with the regional estimate reflective of the overall mean. This assumption is unnecessary, and indeed, becomes a testable hypothesis, when making inferences using LQAS data, because we are able to provide information at both the regional and subregional (i.e., district) levels. This information includes identification of highly and poorly performing districts (and highly and poorly performing SAs within the district), and a measure of the geographic variability of the regional estimator.

To illustrate this point, in Figure 1.1, Figure 1.3, and Figure 1.4 are maps of south-west region displaying the 14 constituent districts with population sizes, and prevalence estimates calculated using LQAS data from that district for two indicators (contraceptive prevalence, and fully vaccinated children 12-23 months of age). Each smaller filled circle represents one of the 40 DHS clusters sampled from this region; note that the DHS prevalence is estimated such that the comparative map would contain a single color covering the whole region. In the lower portion of each of these maps is the overall regional prevalence from both surveys.

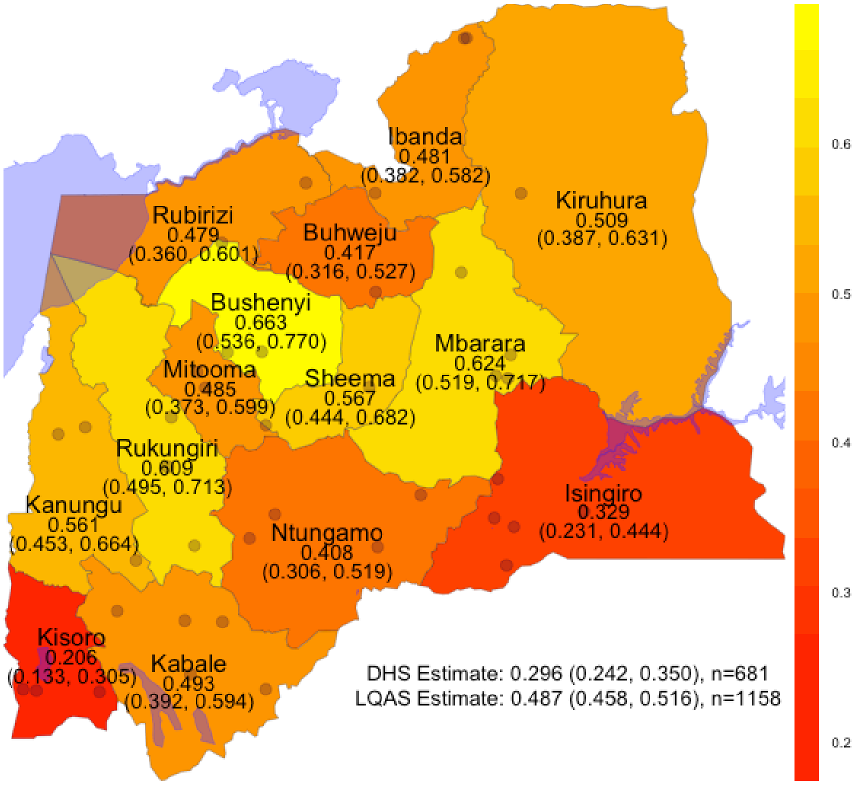


Figure 1.3 Distribution of the indicator '% of currently married women who are using any family planning method' across the 14 districts of the south-west region. Test for homogeneity of prevalences: p-value < 0.0005. The 40 clusters that were sampled by the DHS Program for inclusion in their survey are denoted by translucent circles. Refer to Table A.5 on page 61 for more detailed information on these prevalence estimates.

1.4 Discussion

1.4.1 Discrepancies between prevalence estimates

When comparing two indicators, we first need to ensure that the indicators are measuring the same phenomenon. This is often difficult to ensure when the two are defined in different surveys by different individuals. Our choice of indicators to compare was influenced by how closely we could achieve comparability of indicators. Secondly, if two indicators are supposedly estimating the same quantity and the results differ, it is not possible, without importing extra information into the argument which we do not have available, to determine which indicator yields an answer that is closer to the 'truth'. With these caveats, we failed to find disagreement in 21 comparisons and another 7 show clinically insignificant difference (75.7%). However,

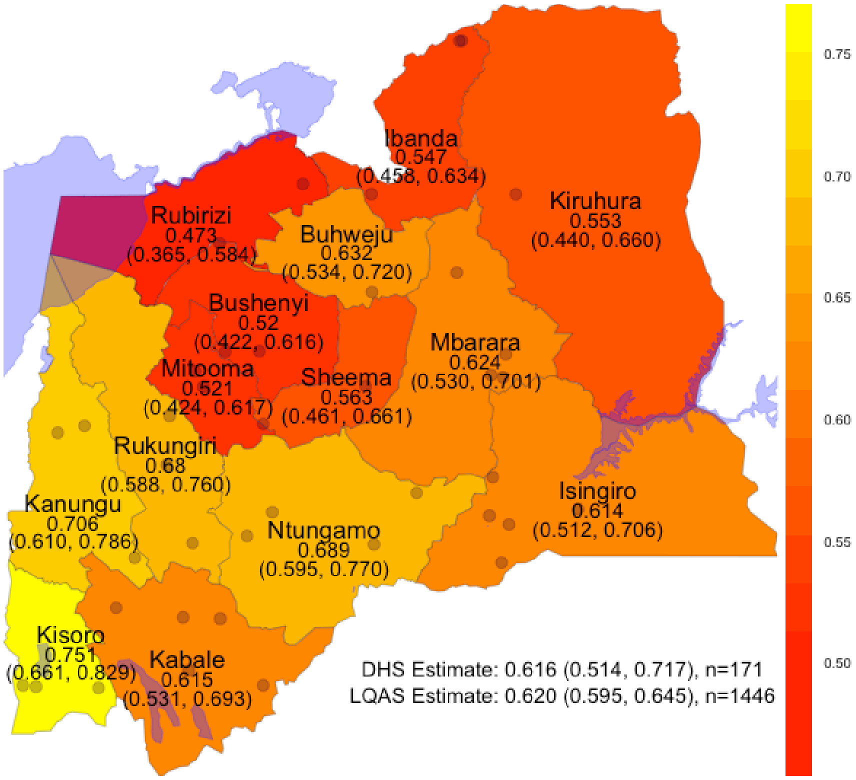


Figure 1.4 Distribution of the indicator '% of children 12-23 months who are fully vaccinated' under *Definition 2* (1 BCG + 3 DPT + 3 POLIO + MEASLES) across the 14 districts of the south-west region. Test for homogeneity of prevalences: p-value $= 0.002$. The 40 clusters that were sampled by the DHS Program for inclusion in their survey are denoted by translucent circles. Refer to Table A.6 on page 61 for more detailed information on these prevalence estimates.

there are discrepancies that reveal subtle differences between the UDHS and LQAS surveys. We discuss only a selection of extreme discrepancies to perhaps find explanation for these and other differences. For example, consider the 'HIV Counseling and Testing' indicators (Table A.1 on page 57), where, across all subpopulations, three of the five comparisons failed to disagree. While two indicators were found to be statistically different, their values are still reasonably close and clinically insignificant. For the five 'HIV Knowledge and Sexual Behavior' indicators (Table A.3 on page 59), four failed to disagree for almost all subpopulation comparisons. For the indicator reporting the percentage of individuals who have had sexual intercourse with a non-marital or noncohabiting sexual partner, the LQAS estimates were higher for all subpopulations. However, three of the four differences were clinically insignificant. In this example, the statistical difference masks the similarity of the prevalence estimates when considered from the point of view of the health system manager.

For the ‘Prevention of Mother-to-Child Transmission’ (PMTCT) indicator (Table A.2 on page 58), there was a significant difference. We believe this is attributable to the differing construction of the two indicators; the DHS asks several questions of respondents about receiving specific information related to PMTCT, while the LQAS survey asks a general question about whether the mother has received information about PMTCT.

Next, consider the indicator ‘% of mothers of children 0-11 months who received two of more doses of SP/Fansidar during their last pregnancy’ (Table A.4 on page 60). From the way the corresponding DHS women’s questionnaire item is structured (Item #425), respondents are asked to volunteer the name of their antimalarial; if the respondent does not know the name of the antimalarial, they are shown the packages of medications to support their response. In the LQAS survey interview, respondents are also asked to volunteer the name of their antimalarial, but the packets of medications are not shown.

Another discrepancy is ‘% of households using iodized salt’ (Table A.7 on page 62), where the DHS estimate is higher than the LQAS estimate in two circumstances. This difference could reasonably be attributed to the methods used by the interviewer to determine the presence of iodized salt. During the DHS interview, the interviewer asks the respondent for a teaspoonful of cooking salt and performs a chemical test for presence of iodine (Household Questionnaire Item #140). During the LQAS survey interview (Mothers of children 12-23 months Questionnaire Item #514), the interviewer requests the household’s salt packet and checks the packaging for indication of iodization. In short, there is no chemical testing and the package may underreport the presence of iodine. We must also take into account that the DHS uses a representative sample of all households whereas the LQAS uses a representative sample of households with mothers of children 12-23 months of age. The former comprises a population with more variation and could include a confounder associated with purchasing of iodized salt. Nevertheless, the populations are not equivalent. When we extract the households with children 12-23 months from the DHS for comparison with the LQAS, the results are closer (95.9% vs. 92.2%) but we compare an LQAS sample of $n = 1371$ with a DHS cluster sample of $n = 171$. The power in the LQAS sample to detect small differences may be the reason for this statistically significant but clinically insignificant result.

An additional discrepancy is ‘% of households with safe water supply’ (Table A.8 on page 63, but an explanation for the difference is not as readily available as for the previous three examples. In comparing the available option responses in the two surveys for ‘source of drinking water’, we see that they are largely the same with two exceptions: ‘public tap/standpipe’ and ‘protected spring’. Both of these safe water sources that are included as DHS response items but are not LQAS response items. Exclusion of these response items from the numerator of the DHS indicator only further exacerbates the difference between the two prevalence measures.

One discrepancy worth mentioning concerns an indicator we omit from the final analysis due to lack of definition compatibility, namely '% of mothers of children aged 0-11 months who took iron supplementary tablets for at least 90 days during last pregnancy'. The estimate from DHS data yielded 0.044 [95% CI (0.007, 0.082) with $n = 205$] while the LQAS data yielded 0.776 [95% CI (0.754, 0.797) with $n = 1446$]. We believe this discrepancy is caused by the way the questions are asked of the respondents. Within the DHS Women's Questionnaire, respondents are asked 'How many days did you take iron tablets during your last pregnancy?' and provide an integer. Within the LQAS Mothers of children aged 0-11 months questionnaire, respondents are asked 'Did you take iron tablets for at least 90 days during your last pregnancy?' and provide a yes or no. The estimation goals of the two questions are different; the DHS wanted to report an average number of days, and the LQAS wanted a binary classification.

Differences between prevalence estimates, such as those discussed above, do not mean that one estimate is correct and the other is not. Rather, these differences expose differences in questionnaire items and interviewer protocols that can lead to the improvement of both surveys. Prevalence estimates that are similar lend support to the other, leading us to believe that the calculated estimate may be close to reality.

1.4.2 Comparison of costs

It is interesting but difficult to compare the costs of LQAS with those of DHS as the purpose for their respective uses is different. One clear difference in this Ugandan case is that the DHS is designed to measure indicators at a regional level while the LQAS survey utilizes the measures at the district level. Hence, many more district-level samples are collected with the LQAS survey. The only financial data in the literature concerning DHS costs come from the 1991, 1994, 1996, and 1999 Tanzania surveys [55]. That study took all expected recurrent and non-capital costs, divided by the number of participating households times the national estimate of average household size for 2000-01. Oddly, as this results in a lower cost estimate, all members of the household were considered as participants, rather than just those interviewed. The cost was \$19.57 per participant (or \$25.25 in 2013 dollars). Using this information to estimate the cost per interview in the 2011 UDHS, which includes a household and a women's survey in the same household, and men in every third sampled house, the cost per interview was $57.94 (or $130.37 per household in 2013 dollars).

The cost data for LQAS come from a detailed cost study in Costa Rica [74] and a comparative assessment from 2002 of three USAID projects in Nepal, Nicaragua and Armenia [30]. LQAS promotes the engagement of District Health Managers as a cost-saving mechanism as their costs are already paid by the Ministry of Health. These in-kind costs are included in this analysis as an LQAS cost. Taking into account that LQAS uses parallel sampling of interviewees (all in different households), the cost per interview is $11.17, using the first index household as the reference (or $29.28 per household in 2013 dollars). In these examples, LQAS was at least 4.5 times less expensive than DHS for each household participating in the survey and 5.2 times less expensive for each interview. We note though the UDHS used a questionnaire more extensive than that of the LQAS survey, and included height and weight measurement, blood specimen collection for on-site anaemia and laboratory vitamin A testing. An extensive questionnaire and biological measurement does increase the costs of a DHS.

1.4.3 Surveys are complementary, not redundant

From the prevalence comparisons, we see that as a secondary by-product the LQAS survey provides very similar information to that of the DHS. Twenty-one of 37 comparisons for the 25 selected indicators failed tests of statistical difference, including important measures of HIV knowledge and sexual behavior, malarial prophylaxis, child vaccination and nutrition. Seven statistical differences were clinically insignificant resulting in a failure to find meaningful difference in 75.7% of the comparisons. Many of the prevalence estimates that did not agree across the two surveys have reasonable explanations. Other comparisons of LQAS with demographic surveillance systems have proved to have an excellent agreement of results, but in those occasions the indicators were identical [13]. Similarly, reliability studies of LQAS have recently compared data collected by managers who use LQAS results to improve their own programs with data collected by disinterested data collectors; the concordance of the two data sets was very high [8].

In fact, the information provided by the LQAS survey is a superset of the information provided by the DHS; it provides similar information to that of the DHS, and more. In general, for a fixed sample size, a stratified sampling strategy produces more precise estimates than a cluster sampling strategy. In the case of this particular regional study, where the LQAS survey sample was stratified and the DHS used a cluster sample, for all indicators the LQAS sample size was larger than that of the DHS. This suggests that the LQAS measures are more precise but we do again note that the surveys were designed and conducted for different purposes, so a comparison of sample size is not so straightforward. The large sample sizes also may have led to the statistical differences between the surveys that are not important from a health system management perspective.

It is indeed true that the purposes and intended use of data generated by the two surveys are different. For example, the DHS is designed to collect information on the population of living mothers with children under five years of age, so there is five years of history in every resulting measure. The LQAS survey is designed to collect information on the population of mothers with younger children such as under one year of age, or 12 to 23 months of age so this survey gives information on health system performance from the recent past. This short time frame lends flexibility to the survey, so questionnaire items can be modified and updated based on the most effective direction of healthcare delivery.

However, the stratification of the LQAS sample allows us to investigate the geographic variability of the regional point estimate, exemplified in Figures 1.3 and 1.4. Use of such information, in conjunction with demographic information like the population distribution of Figure 1.1, provides the structure needed for the evidence-based allocation of resources. The LQAS results provide a further and more granular depiction of variability when considering the classification of subdistrict-level supervision areas according to a coverage target. The subdistrict areas (counties, subcounties and parishes in the case of Uganda) are not presented in Figures 1.3 and 1.4, but are the main reason for using LQAS, to empower subdistrict managers to manage by quickly available classification results reflecting the current condition of the area for which they are responsible. This is in contrast to DHS data, which are able to give a single estimate for the region that cannot be disaggregated [13]. Although an analyst could consider, alternatively to LQAS, a design akin to a stratified DHS, the analyst would lose many of the advantages particular to LQAS, including the ease of data collection, the timeliness of results, and relatively low financial and human cost.

To our knowledge, this formal comparison of indicators as calculated using LQAS data and DHS data collected within similar time periods is the first of its kind. However, a comparison on the basis of an emulation was reported in [12]. Our findings are quite similar to other comparisons to the LQAS sampling procedure seen in the M&E literature. For example, Singh et al. [63] report consonance of immunization coverage estimates in a region of India as calculated from data using the LQAS sampling method and from data using the 30-cluster survey method of the World Health Organization's Expanded Program on immunization [33].

Bhuiya et al. [11] also report agreement of estimates from LQAS data and 'health and demographic system' data collected in Matlab, Bangladesh.

Several individuals involved in global health policy have commented on the need of data at different levels for policy-making and management [16]. As evidenced by our study and similar studies discussed above, the LQAS methodology provides these multilevel data, whereas the DHS, by nature of its design, cannot. The DHS has built a reputation of providing high-quality data for international comparisons; we have shown that LQAS gives the same accuracy, but is programmatically more relevant [13, 48]. Further, LQAS builds local capacity, because regular data collection will lead to its institutionalization. Chan et al. [16] describe this institutionalization as 'essential', because it strengthens a country's ability to collect, process, analyze and use health data. Also, by virtue of using local health workers to collect LQAS data, it is cheaper than the DHS.

1.5 Conclusion

The LQAS sampling method is a viable, timely, and informative complement to the DHS that can be used in interstitial years. It is more-management oriented because of the quick turnaround of data collection and analysis, allowing for targeted, data-driven decisions to be made quickly. This results in timely and local evidence of the value of the data collected and it might also convince local data gatherers of the value of the data gathering effort and result in higher quality data.

1.6 Acknowledgements

We thank Stephen Lwanga, Management Sciences for Health (MSH) Country Director for providing logistical support for this research; William Vargas, Charles Nkolo and Joseph Ouma for their support during the implementation of this research; the District Health Officers and health workers of the south-west region of Uganda for their commitment and dedication to the health needs of the population they serve. We would like to thank all other staff members of the STAR-E LQAS project and MSH staff members without whom this work would not have been possible. This work was supported by the American people through USAID under terms of the Cooperative agreement with Management Sciences for Health.

2

Approaches to treatment effect heterogeneity in the presence of confounding

Sarah C. Anoke[1], Sharon-Lise Normand[1,2], and Corwin M. Zigler[1]

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

2 *Department of Health Care Policy, Harvard Medical School, Boston, MA, USA*

Abstract

The literature on causal effect estimation tends to focus on the population mean estimand, which is less informative as medical treatments are becoming more personalized and there is increasing awareness that subpopulations of individuals may experience a group-specific effect that differs from the population average. In fact, it is possible that there is underlying systematic effect heterogeneity that is obscured by focus on the population mean estimand. In this context, understanding which covariates contribute to this treatment effect heterogeneity (TEH) and how these covariates determine the differential treatment effect is an important consideration. Towards such an understanding, this chapter briefly reviews three approaches used in making causal inferences and conducts a simulation study to compare these approaches according to their performance in an exploratory evaluation of TEH when the heterogeneous subgroups are not known *a priori*. Performance metrics include the detection of any heterogeneity, the identification and characterization of heterogenerous subgroups, and unconfounded estimation of the treatment effect within subgroups. The methods are then deployed in a comparative effectiveness evaluation of drug-eluting versus bare-metal stents among 54,099 Medicare beneficiaries in the continental United States admitted to a hospital with acute myocardial infarction in 2008.

This work was supported by grants NIH 5T32AI007358-27, NIH NCI P01-CA134294, and NIH NGIMS R01-GM111339.

2.1 Introduction

Literature on estimation of the causal effect of a treatment on an outcome tends to focus on the population mean estimand, which is appropriate for many research questions. However, technological advances and subsequent increases in the quantity and quality of biomedical data has led to an interest in personalized medicine, the mining of large observational data sources to construct treatments tailored to the covariate distribution of a population [38]. The resulting research question then involves determination of *treatment effect heterogeneity* (TEH), the existence of an underlying partition of the population into subgroups across which the treatment effect varies systematically. Although the causal effect research question has evolved from one of a population-level average effect to one of subgroup-specific effects, methods for population-level average effect estimates still dominate the literature on causal effect estimation. The goal of this discussion is to evaluate the extent to which several common methods for causal effect estimation simultaneously adjust for confounding and allow for an exploratory investigation of subgroup-specific treatment effects, in potentially high-dimensional settings without any prior knowledge of the number or specific characteristics of these subgroups.

Subgroup analysis methods originated in the clinical trial setting [5, 52], and have been generalized for use in observational studies [51] with limitations [1, 5, 42, 52, 58]. Many methods for subgroup detection in observational studies have grown out of the genetics and bioinformatics literature, but are not designed for comparative evaluation or causal inference [21, 24, 39, 54, 62, 79, 82, 83]. In light of these issues there have been a number of proposed applications of modern machine-learning methods such as regression trees [15, 32] to TEH, including the development of non-parametric causal forests comprised of "honest" trees [6, 77], the use of trees to identify members of a subgroup with an "enhanced" treatment effect [26], and a weighted ensemble of estimators [28].

This paper contributes the ongoing discussion by considering exploratory subgroup detection and TEH estimation in high-dimensional settings when manual evaluation of effect modifiers is not feasible, accomplished in conjunction with confounding adjustment. Achievement of these goals is defined as correct estimation of the number of underlying subgroups, interpretable characterization of the subgroups by observed covariates, and unconfounded estimation of the treatment effect within each subgroup. The discussion continues in §2.1.1 with a brief overview of causal inference and treatment effect estimation. In §2.1.2 we more formally define TEH and distinguish it from other but related causal concepts, and in §2.2 we introduce regression trees as a modeling procedure well-suited for our treatment of TEH identification as a classification problem.

We then discuss three general classes of modeling approach that have been used for causal effect estimation: 1) modeling the outcome conditional on covariates and treatment (e.g., linear regression), 2) modeling the treatment conditional on covariates (e.g., propensity score estimation), and 3) modeling the outcome and treatment jointly conditional on covariates. Evaluation of the ability of each approach to identify TEH is done by describing, implementing, and comparing representative modern methods from each model class: Bayesian Additive Regression Trees (BART) [35], propensity scores estimated with Generalized Boosted Models (GBM) [45], and the Facilitating Score (FS) [66], respectively. Note that these specific methods are not investigated based on any judgment of optimality or superiority over other methods, and such judgment is not the focus of this paper. Rather, evaluation of each of these representative methods is meant to assess the relative strengths and weaknesses of its respective class of modeling approach for the purposes of identifying and estimating TEH. BART and propensity scores with GBM were chosen as modern methods that have recently emerged as popular approaches to overall causal effect estimation. FS is a recently-proposed approach from the machine learning literature that is similarly rooted in tree-based approaches, designed specifically for the purposes of estimating TEH. In §2.3 each method is compared qualitatively, in §2.4 via simulation study, and in §2.5 in the context of an actual CER investigation. The discussion is concluded in §2.6.

2.1.1 Notation and Estimation of the Overall Average Treatment Effect

Let i index individuals within a sample of size n, randomly sampled from a much larger population of interest. T_i is a binary indicator of an individual's point exposure status and $\boldsymbol{X}_i$ a p-dimensional vector of measured pre-treatment covariates. Lowercase t_i, $\boldsymbol{x}_i$ are realizations of their uppercase counterparts. Y_{1i} and Y_{0i} represent the *potential outcomes* that would have been observed had individual i been assigned to treatment or control, respectively. On the difference scale, the causal effect of treatment on individual i is $Y_{1i} - Y_{0i}$. The Fundamental Problem [37] precludes observation of this individual treatment effect (ITE), so we instead consider the average treatment effect (ATE) as our estimand, defined in (2.1).

$$\begin{aligned} \mathrm{E}[Y_1 - Y_0] &= \mathrm{E}[Y_1] - \mathrm{E}[Y_0] & (2.1) \\ &= \mathrm{E}[Y \mid T = 1, \boldsymbol{X}] - \mathrm{E}[Y \mid T = 0, \boldsymbol{X}]. & (2.2) \end{aligned}$$

There are variables associated with both T and Y such that the quantity measured in (2.2) is not a treatment effect, but a spurious measure of association that is in part due to dissimilarities in their distribution across treatment arms. These problematic covariates are referred to as *confounders*, defined here as the subset of $\boldsymbol{X}$ required for strong ignorability [59] to hold. For the purposes of this discussion, we assume that the available covariates measured in $\boldsymbol{X}$ contain (at least) all confounders required to satisfy the assumption of strong ignorability and estimate causal treatment effects. The notation above also implies the Stable Unit Treatment Value Assumption [60].

2.1.2 Treatment Effect Heterogeneity

Variables $\boldsymbol{E} = \{E_1, E_2, \ldots\} \subseteq \boldsymbol{X}$ that define subpopulations across which the treatment effect differs are called *effect modifiers* [34, p. 42], with *effect modification* being synonymous with TEH. As noted by [34, p. 42] and [57, p. 199-201], whether a variable is an effect modifier depends on the scale on which the effect is being measured, be it additive as used here, multiplicative, or the odds ratio. To reflect this dependence, some authors use the terminology *effect-measure modification*. It is possible for a variable to be both an effect modifier and a confounder, which further emphasizes the importance of simultaneous confounder adjustment and TEH identification.

Mapping these statements back to our notation, effect modifiers $\boldsymbol{E}$ comprise a subset of $\boldsymbol{X}$, a collection of variables which deserve some clarification. First note that identifying TEH requires that causal effect identifiability assumptions (e.g., strong ignorability in §2.1.1) defined at the population level must hold within each subgroup. An implication is that achievement of ignorability is particular to a subpopulation, meaning that each subpopulation has its own set of confounders, and its own set of relationships among the confounders, treatment, and outcome. For example, a particular variable can be a confounder in more than one subpopulation but have different relationships with the outcome in each. Thus let variables $\boldsymbol{X}$ be the union of effect modifiers $\boldsymbol{E}$ and confounder sets from each subpopulation.

Our conceptualization of effect modification is different from that of causal [76, p. 268] or biologic [57, p. 202] interaction, the combined effect of treatment T and a second exposure on the outcome Y. In this context, it is of interest whether the effect of T depends on the value of this second treatment (or vice versa in the symmetric argument) [75, Definition 2]. Contrastingly, effect modifiers are characteristics of observational units used to define subpopulations [75, Definition 1]. Our conceptualization of effect modification is also different from that of mediation, a causal concept that aims to understand “how an effect occurs” [76] by considering the pathways between T and Y and variables on those pathways, termed *mediators*. Effect modification is a causal concept that aims to understand “for whom an effect occurs” [76].

2.2 Regression Trees for Characterizing TEH

Assuming that strong ignorability is satisfied by measured covariates $\boldsymbol{X}$ and a modeling approach has been selected (as will be discussed in §2.3), there are different estimation procedures that an analyst could consider in fitting a model to data. One such procedure is the fitting of a step function with a *classification* or *regression tree* (CART). CART does not require the prespecification of any relationships between Y, T, and/or $\boldsymbol{X}$, and the relationships that it does estimate are quite flexible. Further, CART is potentially nonparametric and is able to yield valid estimates when data are missing at random. CART has relevance to the goal of identifying TEH because the classifications can be thought of as the detection and characterization of subgroups by effect modifiers. As discussed in the previous section, each subgroup has its own set of confounders, and its own set of relationships among Y, T, and $\boldsymbol{X}$. These subgroup-specific relationships can be thought of as (statistical) interactions between treatment, confounders, and effect modifiers. Achievement of subgroup-specific ignorability can be thought of as the detection of many (potentially) high-order interactions, making CART a natural choice for TEH estimation. The Supplementary Material outlines some terminology used in the regression tree literature.

2.3 Treatment Effect Modeling Approaches

There are several approaches an analyst could pursue in modeling the treatment effect, and within each approach, several methods to choose from. In this section we describe three classes of approach and from each, a representative method that has been previously used for the explicit purpose of causal effect estimation. For each representative method, we outline its statistical framework and discuss how it may be used to identify subgroups.

The tree-based methods highlighted here average across many simple or unstable models, improving variance but losing direct interpretability of the estimated subgroups. Further, the analyst must *a priori* specify a maximum number of possible subgroups to investigate, relying on the ability to collapse across subgroups if the data indicate fewer. After deciding on a method and how to use that method to assign group membership, interpretation of which characteristics define each subgroup relies on the analyst's ability to inspect covariates distributions within each group to infer what characterizes them. Assuming ignorability is satisfied by some subgroup-specific confounder covariate set $\boldsymbol{X}_{\text{subgroup}}$, unconfounded subgroup-specific treatment effect estimation is possible. TEH is present if the estimated treatment effects vary across the subgroups.

A necessary condition for such estimation is that the analyst can extrapolate the relationships estimated from the treatment arm to the control arm, and vice versa. This extrapolation relies on *positivity*, the assumption that within every level of $\boldsymbol{X}_{\text{subgroup}}$, the probability of treatment is bounded away from 0 and 1. This is an assumption that our causal interpretations rely on, and should be empirically justified. Our ability to provide such justification depends on the estimation method, so the discussion of each estimation method includes an examination of positivity.

2.3.1 Modeling Class 1: Outcome conditional on covariates and treatment

By far the most common modeling approach is to model the conditional mean of $Y \mid \boldsymbol{X}, T$. The overwhelmingly most popular class of models within this approach is parametric linear regression, which models the conditional mean as a linear combination of covariates. To see how this modeling approach allows for causal inference, consider the linear regression model

$$\mathrm{E}[Y \mid \boldsymbol{X}, T] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \gamma_0 T + \gamma_1 T X_1. \tag{2.3}$$

SUTVA and ignorability allow the difference in conditional means to represent the conditional average treatment effect (ATE):

$$\gamma_0 + \gamma_1 X_1 = \underbrace{\mathrm{E}[Y \mid \boldsymbol{X}, T = 1] - \mathrm{E}[Y \mid \boldsymbol{X}, T = 0]}_{\text{difference in conditional means}} = \mathrm{E}[Y_1 \mid \boldsymbol{X}] - \mathrm{E}[Y_0 \mid \boldsymbol{X}] = \underbrace{\mathrm{E}[Y_1 - Y_0 \mid \boldsymbol{X}]}_{\text{conditional ATE}}. \tag{2.4}$$

First considering the case where $\gamma_1 \equiv 0$, we see that the conditional ATE is γ_0, constant for all values of $\boldsymbol{X}$. Because of the collapsibility of the mean, γ_0 is also the marginal average treatment effect. If $\gamma_1 \neq 0$, then the linear predictor contains a (statistical) interaction term, which embeds the *a priori* belief that the average treatment effect is not additive and its magnitude depends on the value of X_1. Use of (statistical) interaction terms is one way of specifying possible TEH, but requires knowledge of the covariates that define the underlying subgroups, infeasible when considering a large number of covariates.

Another set of methodologies within this modeling class estimate the *disease risk score* [2, 3], a special case of Mietennen's *confounder score* [47] and used when the outcome is binary. A disease risk score is the conditional probability of experiencing the outcome while unexposed to treatment Pr$(Y = 1 \mid \boldsymbol{X}, T = 0)$ and is a tool for ranking subjects on how "case-like" they are [47, p. 611]. Related to the disease risk score is the *prognostic score*, a recasting of the disease risk score in the language of potential outcomes [31]. In both cases, observations are stratified by score and the treatment effect estimated within each stratum.

Representative Method: Bayesian Additive Regression Trees (BART) A popular alternative to parametric regression for estimating (2.3) is Bayesian Additive Regression Trees (BART) [18, 19], which we will explore in some detail for its potential to provide exploratory analysis of TEH. Following the notation of Chipman et al. [18], let $\mathcal{T}_j$, $j = 1, \ldots, m$ represent a tree with B_j terminal nodes; that is, a partition of the population into B_j subgroups. Let $M_j = \{\mu_{b_j} \mid b_j = 1, \ldots, B_j\}$ be the set of mean outcomes across the subpopulations defined the terminal nodes of tree $\mathcal{T}_j$. Also let $g(\boldsymbol{x}_i, t_i \mid \mathcal{T}_j, M_j)$ represent the mapping of observed covariate and treatment pair $(\boldsymbol{x}_i, t_i)$ to a terminal node within tree $\mathcal{T}_j$ with mean $\mu_{b_j} \in M_j$. An individual's outcome is then modeled as the sum of m trees

$$Y_i = \varepsilon_i + \sum_{j=1}^{m} g(\boldsymbol{x}_i, t_i \mid \mathcal{T}_j, M_j), \quad \varepsilon_i \sim N(0, \sigma^2) \tag{2.5}$$

where m is fixed. This is an example of *boosting*, the construction of a large tree by summing together simple trees. In the Bayesian model, the m simple trees and outcome variance σ^2 are the unknown parameters; the data are $\boldsymbol{y}$ the vector of observed outcomes, and $\boldsymbol{X}$ the matrix of observed covariate data. A Gibbs sampler is used to sample from the posterior distribution of the boosted tree. Each posterior draw is used to generate a predicted outcome $\widehat{Y}_i$ for all observations i. Notationally, let the kth posterior draw be denoted as $\widehat{\mathcal{T}}^{(k)} = \Big((\hat{\mathcal{T}}_1^{(k)}, \hat{M}_1^{(k)}), \ldots, (\hat{\mathcal{T}}_m^{(k)}, \hat{M}_m^{(k)}), \hat{\sigma}^{(k)}\Big)$, and the vector of n predicted outcomes generated from this tree denoted as $\widehat{\boldsymbol{Y}}^{(k)}$. The set $\widehat{\mathcal{T}} = \Big\{\widehat{\mathcal{T}}^{(1)}, \widehat{\mathcal{T}}^{(2)}, \ldots, \widehat{\mathcal{T}}^{(K)}\Big\}$ denotes the K posterior draws.

As Hill [35] contributes in her reframing of BART from a predictive methodology into a methodology for causal effect estimation, $\widehat{\boldsymbol{Y}}^{(k)}$ is the vector of predicted potential outcomes, corresponding to the treatment actually received. $\widehat{\boldsymbol{Y}}^{(k)}_{\text{counterfactual}}$ is another vector of predicted potential outcomes, but corresponding to the treatment not received. The ITE predicted from $\widehat{\mathcal{T}}^{(k)}$ is then the appropriate difference in the predicted potential outcomes. This prediction of ITEs is repeated $\forall\ \widehat{\mathcal{T}}^{(k)} \in \widehat{\mathcal{T}}$ (envision a $K \times n$ matrix, with the i^{th} column representing K samples from the posterior distribution of the ITE for the i^{th} observation). Estimates of the ITE for each individual (as well as σ^2) are then obtained by summarizing across the K posterior draws, for example, by averaging to take the posterior mean estimate. It is this averaging across K boosted trees (rather than inference based on one boosted tree) that differentiates BART from traditional boosted methods.

Although subgroup estimation is not explicitly part of the model specification or estimation output, the analyst is still able to investigate the empirical distribution of ITEs for clues. For example Foster et al. [26] refer to $\widehat{\boldsymbol{Y}}^{(k)}_{\text{counterfactual}}$ as the “virtual twin” of $\widehat{\boldsymbol{Y}}^{(k)}$, and suggest regressing the predicted ITEs on $\boldsymbol{X}$, towards finding a single subgroup with a treatment effect that is “enhanced“ relative to the ATE. Hill [35] proffers visualization of the modes of the predicted ITEs (by histogram for example) for hints about the underlying number of subgroups. Alternatively, the analyst can *a priori* set the number of subgroups to ten (say), and group observations based on deciles of the empirical distribution and estimate a TE within each subgroup.

When using BART or other outcome regression models, valid causal inference relies on positivity, but empirical positivity violations (e.g., a level of $\boldsymbol{X}_{\text{subgroup}}$ with only treated observations) will not be automatically evident. In fact, outcome regression will yield causal effect estimates whether or not positivity is violated. This issue can be partially overcome by assuming any empirical violations are random rather than deterministic [80], and checking that the unconditional probability of treatment within the finite subgroup sample is bounded away from 0 and 1.

2.3.2 Modeling Class 2: Treatment conditional on covariates

Modeling $T \mid \boldsymbol{X}$ is referred to as propensity score estimation, where the *propensity score* $e(\boldsymbol{X}) = \Pr(T = 1 \mid \boldsymbol{X})$ is the conditional probability of treatment [56]. This modeling approach is typically employed as part of a covariate dimension reduction strategy, where the analyst attempts to satisfy ignorability by conditioning on $e(\boldsymbol{X})$ rather than the covariates individually [56, Theorem 3]. The use of $e(\boldsymbol{X})$ allows us to adjust for confounding while averting the need to model how each covariate relates to the outcome of interest or at least alleviating the consequences of misspecifying such a model [36].

Propensity score methods are used in designing observational comparative studies, that is, the structuring "to obtain, as closely as possible, the same answer that would have been obtained in a randomized experiment comparing the same analogous treatment and control conditions in the same population" [61]. This can be accomplished, for example, by matching or subclassifying treated and untreated observations with similar values of the propensity score. A key benefit of this approach is that it permits the analyst to empirically judge the plausibility of the hypothetical study design before analysis of any outcome. After grouping observations on the propensity score, the analyst can empirically assess covariate balance, the similarity of covariate distributions among treated and control units with similar values of the propensity score.

In addition to empirical verification of the hypothetical study design and covariate balance, use of propensity scores to adjust for confounding empirically alerts the analyst to violations of the positivity assumption: when the propensity score model discriminates treatment groups too well, it yields subgroups that are homogeneous with respect to treatment, thus empirical justification of causal interpretation is absent, the treatment effect is undefined, and overall inference must be restricted to observations with defined treatment effect estimates.

Estimation of the propensity score itself has traditionally been done via logistic regression, but the literature shows movement towards more flexible alternatives. For example, Woo et al. [81] evaluate the use of generalized additive models in propensity score estimation, where the linear predictor is replaced with a flexible additive function. Ghosh [27] generalize propensity score estimation as an example of confounder dimension reduction and discuss the theoretical validity of "covariate sufficiency" in causal inference.

Representative Method: Generalized Boosted Models (GBM) One of the most popular flexible alternatives to traditional logistic regression is the use of generalized boosted models (GBM) [45]. We use this method to estimate the conditional log odds of treatment logit[Pr$(T_i = 1 \mid \boldsymbol{X}_i)$], by summing together many low-depth regression trees. Again following the notation of Chipman et al. [18], let $g(\boldsymbol{x}_i \mid \mathcal{T}_j, M_j)$ represent a mapping of an observed covariate value to a terminal node within tree $\mathcal{T}_j$ with mean $\mu_{b_j} \in M_j = \{\mu_{b_j} \mid b_j = 1, \ldots, B_j\}$. In this model, μ_{b_j} is the mean log odds of treatment for observations in terminal node b_j.

The estimation algorithm is initialized at tree $\mathcal{T}_0$ with $B_0 = 1$ node and $M_0 = \{\text{logit}(\bar{t})\}$ where $\bar{t}$ is the unconditional proportion of treated individuals in the sample. During the jth of m iterations, a low-depth tree fit to residuals $[r_i = t_i - \text{expit}\,[g(\boldsymbol{x}_i \mid \mathcal{T}_{j-1}, M_{j-1})]]$, where expit $[g(\boldsymbol{x}_i \mid \mathcal{T}_{j-1}, M_{j-1})]$ is the predicted probability of treatment based on the tree from the previous iteration. Let the number of nodes on this residual tree be denoted by B^*_j and $b^*_{j\ell}$ represent the set of observations in terminal node $\ell \in \{1, \ldots, B^*_j\}$. For each terminal node $b^*_{j\ell}$ an update is calculated and added to $\mathcal{T}_{j-1}$ to generate $\mathcal{T}_j$. The end result of this algorithm is a sequence of trees with increasingly better fit to the data. To prevent overfitting, the algorithm is stopped at the iteration that minimizes some average measure of covariate imbalance across the two treatment arms (e.g., the average standardized absolute mean difference, or the Kolmogorov-Smirnov statistic).

The resulting estimated propensity score can be used to group individuals, but such grouping requires an *a priori* specification of the number of subgroups. The analyst could set the number of subgroups to ten (say) and group observations based on deciles of the empirical propensity score distribution, then estimate a TE within each subgroup.

Recall, however, that our goal is to group observations that are similar, where the desired similarity is in the ITE. While estimating differential effects across groups defined by the estimated propensity score is commonplace, note that these groups are not defined based on observations' ITE. Rather, observations are grouped based on the likelihood of receiving treatment, so assessment of TEH is typically restricted to whether the treatment effect varies with values of the estimated propensity score. This can provide an overall assessment of the presence of TEH, but as the propensity score is a scalar summary of a multivariate covariate vector, deriving clinical or scientific interpretability from knowing that the treatment effect varies with the propensity score is challenging. Thus TEH across values of the propensity score does not provide the specificity of heterogeneity we are interested in. If an effect modifier is associated with Y and not T – in other words, if an effect modifier is not also a confounder – then propensity score methods will not be able to detect it. By ignoring outcome data and focusing solely on the relationship between T and $\boldsymbol{X}$, the propensity score model has difficulty learning about who experiences the treatment effect differently.

2.3.3 Modeling Class 3: Outcome and treatment jointly, conditional on covariates

A relatively recent approach is considering the conditional joint distribution of the outcome and treatment by modeling $(Y, T) \mid \boldsymbol{X}$. Nelson and Noorbaloochi [49] define a multidimensional *sufficient summary* $\boldsymbol{S}(\boldsymbol{X})$, a balancing score such that $(Y, T) \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{S}(\boldsymbol{X})$. Wang et al. [78] take a Bayesian variable selection perspective, defining a Bayesian adjustment for confounding (BAC) methodology for estimating average treatment effects with linear regression models by averaging over the posterior probability of covariate inclusion in a joint model for (Y, T).

Representative Method: Facilitating Score (FS) The representative approach from this model class that we investigate in detail is that of Su et al. [66], who define the multidimensional *facilitating score* $\boldsymbol{a}_0(\boldsymbol{X})$ as a statistic that satisfies the following conditional independence:

$$\boldsymbol{X} \perp\!\!\!\perp (Y_0, Y_1, T) \mid \boldsymbol{a}_0(\boldsymbol{X}) \overset{\text{relaxation}}{\Longrightarrow} \underbrace{\boldsymbol{X} \perp\!\!\!\perp T \mid \boldsymbol{a}_0(\boldsymbol{X})}_{\text{addresses confounding}} \text{ and } \underbrace{\boldsymbol{X} \perp\!\!\!\perp (Y_0, Y_1) \mid \boldsymbol{a}_0(\boldsymbol{X})}_{\text{addresses effect modification}}. \tag{2.6}$$

Estimation of $\boldsymbol{a}_0(\boldsymbol{X})$ involves joint modeling of (Y_0, Y_1, T), precluded by the Fundamental Problem [37]. Thus Su et al. [66] instead propose a multidimensional *weak facilitating score* $\boldsymbol{a}(\boldsymbol{X})$, that satisfies the following as derived from the above relaxation:

$$\underbrace{\boldsymbol{X} \perp\!\!\!\perp T \mid \boldsymbol{a}(\boldsymbol{X})}_{\text{addresses confounding}} \text{ and } \underbrace{\mathrm{E}[Y_1 - Y_0 \mid \boldsymbol{X}] = \mathrm{E}[Y_1 - Y_0 \mid \boldsymbol{a}(\boldsymbol{X})]}_{\text{addresses effect modification}}. \tag{2.7}$$

The weak FS $\boldsymbol{a}(\boldsymbol{X})$ is therefore a balancing score, and conditioning on $\boldsymbol{a}(\boldsymbol{X})$ defines a subpopulation within which the average treatment effect is constant. This is the first method discussed thus far that has explicitly addressed the issue of TEH.

To estimate the weak FS, Su et al. [66] use the conditional independence

$$(Y, T) \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{h}(\boldsymbol{X}) \tag{2.8}$$

for a statistic $\boldsymbol{h}(\boldsymbol{X})$. The validity of this independence is a consequence of a factorization theorem applied to the joint distribution of observed data $f_{Y,T \mid \boldsymbol{X}}(y, t \mid \boldsymbol{x})$ [66, Theorem 7]. By this theorem, the statistic $\boldsymbol{h}(\boldsymbol{X})$

that fulfills the preceding conditional independence also fulfills definition (2.7) of a weak facilitating score [66, Theorem 3]. This then allows for indirect estimation of the weak FS by jointly modeling (Y, T). Regression trees [15] are used for this modeling, where the fact that the joint conditional density $f_{Y,T \,|\, \boldsymbol{X}}(y, t \,|\, \boldsymbol{x})$ is constant within a terminal node implies $(Y, T) \perp\!\!\!\perp \boldsymbol{X}$ within that node. This within-node independence implies (2.8); that is, that the facilitating score is constant within a given node. Because a single tree model is known to be unstable (i.e., a small change in the data can result in a large change in the final tree structure) [32, p. 312], Su et al. [66] propose an *aggregated grouping* strategy (similar to bagging) to average across K possible tree structures. Again adopting the notation of Chipman et al. [18], to generate one possible tree structure, a bootstrap sample is generated to grow and prune tree $\mathcal{T}_k$ to B_k terminal nodes. This tree is then applied to the original data. An $n \times n$ pairwise distance matrix $\boldsymbol{D}_k$ is generated from the resulting tree classifications, where matrix element

$$d_{ii'}^{(k)} = \begin{cases} 1 & \text{if observations } \{i, i'\} \text{ fall into the same terminal node of } \mathcal{T}_k \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

The distance matrices are then averaged to obtain $\boldsymbol{D} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{D}_k$, and a clustering algorithm (e.g., multidimensional scaling, partitioning around mediods) applied to $\boldsymbol{D}$ to obtain the final data stratification. The end product is the assignment of each observation to a subgroup, and a TE can be estimated within each.

Similar to BART, issues of sparsity may preclude empirical justification of positivity, but this can be partially overcome through assumptions made of the larger subpopulation that the sample represents and checking that the unconditional probability of treatment within each subgroup is bounded away from 0 and 1. We note that although the estimation algorithm of Su et al. [66] ensures empirical subgroup positivity through particular stopping rules within the node-splitting procedure, these rules also increase the potential to conceal true subgroups. For example, if there is a tree node that contains observations from two subgroups but there are too few treatment observations, the procedure will not split that node and inference will be made on the whole node.

Approach	Representative Method	Summary + Assumption
$Y \mid \boldsymbol{X}, T$	Bayesian Additive Regression Trees (BART) [18, 19] and application to causal effect estimation [35]	Tree-based modeling of individual potential outcomes.
$T \mid \boldsymbol{X}$	Propensity score estimation with Generalized Boosted Regression (GBM) [45]	Tree-based modeling of logit$[e(\boldsymbol{x})]$.
$Y, T \mid \boldsymbol{X}$	Tree-based Facilitating Score (FS) estimation [66]	Tree-based modeling of $f_{Y,T \mid \boldsymbol{X}}(Y, T \mid \boldsymbol{X})$. $\boldsymbol{X} \perp\!\!\!\perp T \mid \boldsymbol{h}(\boldsymbol{X})$ and $\mathrm{E}[Y_1 - Y_0 \mid \boldsymbol{X}] = \mathrm{E}[Y_1 - Y_0 \mid \boldsymbol{h}(\boldsymbol{X})]$

Table 2.1 Summary of statistical methods being compared.

2.4 Comparison of Methodologies: Simulation Studies

We evaluate the ability of the three approaches discussed in §2.3 to identify TEH by considering the representative method from each approach; these methods are summarized in Table 2.1.

2.4.1 Simulated Data

Data Structure and Analysis Letting $\ell_1 \in \{A, B, C, D\}$ denote a particular simulation scenario, Table 2.2 defines possible underlying correlation structures for $\{Y, T, X_1, \ldots, X_6, E_1, E_2, E_3\}$. Let $Y \sim N(\mu_{\ell_1}, 1)$ denote the continuous outcome, and $T \sim \text{Bern}(p_{\ell_1})$ the binary treatment. There is one covariate associated with the treatment only, $X_5 \sim \text{Bern}(0.5)$. There is one covariate associated with the outcome only, $X_6 \sim N(0,1)$. There are four confounders of the effect of treatment on outcome, $(X_1, X_2, X_3, X_4) \overset{\text{i.i.d.}}{\sim} N(0,1)$. In addition, there are three binary effect modifiers, $(E_1, E_2, E_3) \overset{\text{i.i.d.}}{\sim} \text{Bern}(0.5)$. These three variables define eight subgroups (Group 1, …, Group 8), with six unique treatment effects among them. As determined by μ_{ℓ_1} and the eight unique values of (E_1, E_2, E_3), the subgroup-specific ATEs are 1, 2, 5, 5, 6, 6, 9, and 10. Regardless of the underlying data generation mechanism, models are fit using all available covariates (similar to what might be done in practice).

Within one of the 100 simulation iterations, a dataset of size $n = 1500$ is generated according to §2.4.1 and TEH evaluated using each of the methods described in §3. In the case of BART and GBM, such evaluation is done by partitioning the estimated probability distribution (be it of the outcome or treatment) into deciles, then estimating a TE within each subgroup defined by the deciles. In the case of the facilitating score (FS), the dissimilarity matrix $\boldsymbol{D}$ is partitioned into 10 subgroups using the *partitioning around mediods* method and a TE estimated within each. In practice the analyst will know neither the number nor relative sizes of the true underlying subgroups; thus success of a method is judged in part by its ability to group similar observations together. Typical measures of concordance are complicated by our estimation of more subgroups (ten) than truly exist in the population (eight), as well as the unequal sizes of the true subgroups, leading to members of true subgroups necessarily split across estimated subgroups. This process is explained in detail below, as is our proposed measure of concordance.

ℓ_1	**Data Generation Scenario**	**Definition**
A	confounding and no effect modification	$p_{\ell_1} = \text{expit}(\ 0.1X_1 - 0.1X_2 + 1.1X_3 - 1.1X_4$ $+0.4X_5)$ $\mu_{\ell_1} = -3.85 + 5T + 0.5X_1 - 2X_2 - 0.5X_3$ $+2X_4 + X_6$
B	effect modification and no confounding	$p_{\ell_1} = \text{expit}(\ 0.4X_5)$ $\mu_{\ell_1} = -3.85 + 5T + X_6 - E_1 - 2E_3 + TE_1$ $+4TE_2 - 4TE_3$
C	effect modification and confounding	$p_{\ell_1} = \text{expit}(\ 0.1X_1 - 0.1X_2 + 1.1X_3 - 1.1X_4$ $+0.4X_5)$ $\mu_{\ell_1} = -3.85 + 5T + 0.5X_1 - 2X_2 - 0.5X_3$ $+2X_4 + X_6 - E_1 - 2E_3 + TE_1$ $+4TE_2 - 4TE_3$
D	effect modification and confounding by effect modifiers	$p_{\ell_1} = \text{expit}(\ 0.1X_1 - 0.1X_2 + 1.1X_3 - 1.1X_4$ $+0.4X_5 - 0.1E_1 + 1.1E_2 - 4E_3)$ $\mu_{\ell_1} = -3.85 + 5T + 0.5X_1 - 2X_2 - 0.5X_3$ $+2X_4 + X_6 - E_1 - 2E_3 + TE_1$ $+4TE_2 - 4TE_3$

Table 2.2 Summary of data generation scenarios for Simulation Study 2.4.1.

Results To summarize the results of a single simulation iteration, subgroups are numbered in ascending order by the estimated treatment effect. If an individual is in a subgroup for which there is an undefined treatment effect (e.g., its subgroup is entirely comprised of individuals on treatment), it is assigned to an "undefined" subgroup. For example, if only seven of ten subgroups have a defined treatment effect, then the 8th, 9th, and 10th subgroups are empty and the individuals in the three subgroups are all reassigned to the same "undefined" subgroup. In the ordering of subgroups, the "undefined" group is placed last.

The estimated partition is cross-tabulated with the eight true groupings, which are also ordered by the magnitude of their average treatment effect (envision a 8×11 table where each row corresponds to true subgroup

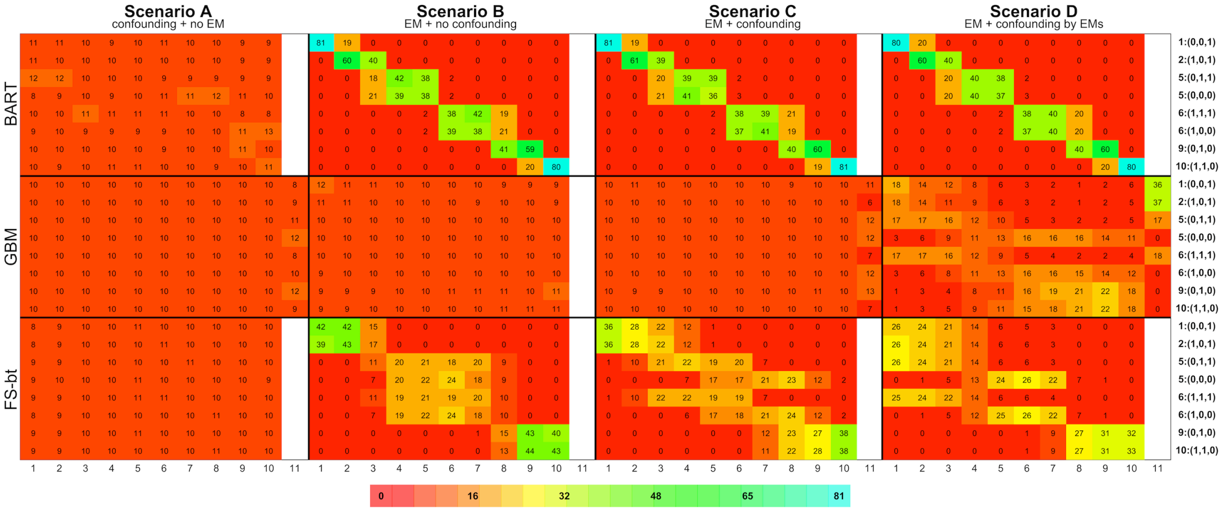


Figure 2.1 Visualization of results from Simulation Study 2.4.1. Details regarding how the figure was constructed, and how to interpret the figure, are given in §2.4.1.

membership and each column corresponds to estimated subgroup membership). Within this cross-tabulation table, row percentages for each of the $q = 1, 2, \ldots, 8$ rows are calculated representing the proportion of units in true treatment group q that are assigned each of the estimated subgroups (columns). This table of row percentages is constructed for each simulation iteration. The resulting collection of tables is averaged over the 100 simulation iterations to yield a single table of cell-specific averages. A single average indicates how often a method places an individual from true subgroup q in to each of the 11 estimated subgroups. Averages for the different data generation scenarios and the different estimation methodologies are visualized in Figure 2.1.

To ease explanation, consider the fourth block in the second row of Figure 2.1 – the table summarizing data generated under Scenario D and using GBM to address TEH. The last row of this table summarizes results for observations in Group 8, the true subgroup with the largest ATE (as indicated by the numbers in the right-hand margin). The first number in this row is "1", the average percentage of units in Group 8 that were assigned to the estimated subgroup with the smallest treatment effect. This average is taken across the simulations for which this estimated subgroup had membership. For this first cell, the average is taken across all 100 simulation iterations, because by design, there is always membership in the smallest group. Consider, however, the 10th cell in this row containing the value "18". On average across the 8 simulations for which there was membership in the 10th subgroup, 18% of units in Group 8 were assigned to the subgroup with the tenth (i.e., largest) treatment effect. Of special note is the "36" in the 11th column of the first row; on average across the 92 simulation iterations for which there was at least one estimated subgroup with an undefined treatment effect, 36% of observations in Group 1 were placed in an estimated subgroup having an undefined treatment effect. Said simply, using GBM to estimate TEH, over one-third of observations in true Group 1 can be expected to have an undefined treatment effect. We note that GBM was the only estimation procedure that yielded subgroups with an undefined treatment effect; the cell-specific averages presented for BART and FS were across all 100 simulation iterations. The denominators for the cell-specific averages presented for GBM are not explicitly provided within the figure, but are contained in Table B.1 of the Appendix.

Measures of concordance between Figure 2.1 and what we expect to see are given in Table 2.3. Defining "truth" as the color arrangement we expect to see for a particular estimation method and simulation scenario, we calculate the Euclidean distance of each observed cell color (i.e., a block in Figure 2.1) from its expected cell color, in red-green-blue (RGB) color space. We then average these cell-specific distances over the 80 cells to get a summary measure of how far our observed data are from what we expect. (Note that the 11th *undefined* column was omitted from these calculations, under the assumption that membership in this column is reflected by absence in the remaining 10 columns included in the calculation.) Scaling these distances by the maximum distance (the distance between random assignment of and perfect assignment), we get a measure of distance from what we expect to see on the $[0, 1]$ scale. Letting (R_c, G_c, B_c) represent the expected color of cell c and (r_c, g_c, b_c) the observed color, we define this distance in Equation (2.10) below.

$$\frac{1}{80 \text{ cells}} \sum_{c \in \{80 \text{ cells}\}} \sqrt{(R_c - r_c)^2 + (B_c - b_c)^2 + (G_c - g_c)^2} \tag{2.10}$$

Subtracting this quantity from 1, we get a measure of concordance that is similar to the traditional definition of "sensitivity", in that we are conditioning on the truth and measuring agreement with this truth. "Sensitivity“ for each estimation procedure and simulation scenario is presented in Table 2.3.

Figure 2.2 presents a second summary of the simulation results. The structure of this grid is pattered after Figure 2.1, where each row is an estimation method, and each column is a data generation scenario. Letting $j = 1, \ldots, 100$ denote the simulation iteration, the ten treatment effect estimates generated during the jth iteration are plotted, with the x-axis corresponding to magnitude; this is repeated for all 100 simulation iterations. For each estimated treatment group (e.g., `TE(1)`), a boxplot is used to help visualize the distribution of estimates in that group. Recall from §2.4.1, the estimated treatment groups are always sorted from smallest

	Scenario A confounding and no effect modification	**Scenario B** effect modification and no confounding	**Scenario C** effect modification and confounding	**Scenario D** effect modification and confounding by EMs
Bayesian Additive Regression Trees (BART) + partitioning $\widehat{ITE}$ distribution into deciles	94.5	97.9	97.6	98.4
Generalized Boosted Models (GBM) + partitioning $\widehat{e(\boldsymbol{X})}$ distribution into deciles	98.5	1.0	0.7	10.9
Facilitating Score (FS) + partitioning $\boldsymbol{D}$ into 10 subgroups using PAM	96.4	42.1	30.1	23.5

Table 2.3 Summary of "sensitivity" calculations for Simulation Study 2.4, with quantities calculated according to Equation (2.10) and given as percentages.

TE to largest; so `TE(1)` will always be the estimated subgroup with the smallest treatment effect. For clarity, the x-axes of the forest plots have been omitted, but there are vertical dashed lines denoting the true average treatment effects $(1, 2, 5, 6, 9, 10)$.

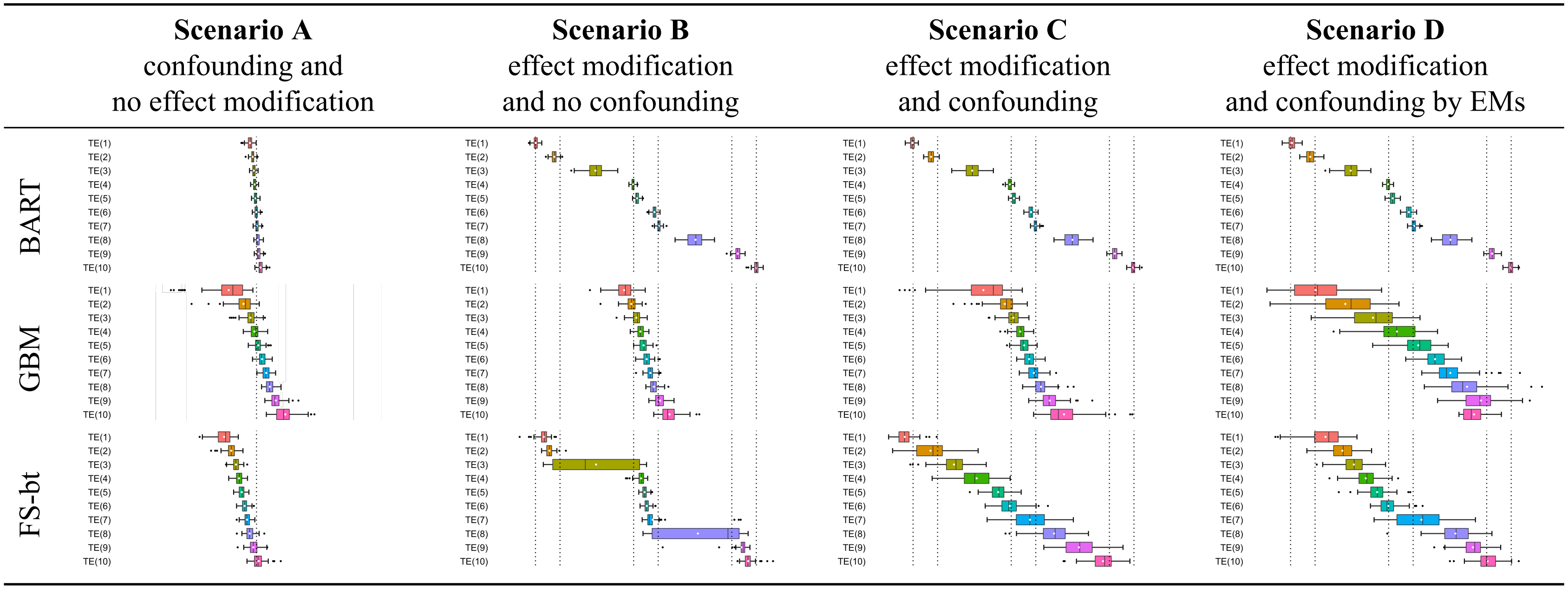


Figure 2.2 Box plots of point estimates of the ATE across 100 replicates of each simulation scenario, with layout analogous to that of Figure 2.1 in the main text. For clarity, the x-axes of the plots have been omitted, but there are vertical dashed lines denoting the true average treatment effects $(1, 2, 5, 6, 9, 10)$.

Discussion Figure 2.1 is a qualitative metric that allows for broad comparisons of the "performance" of each TEH identification strategy under several different data generation scenarios, where "performance" refers to the ability of the method to group truly similar observations together. For BART and GBM, deciles are used to define the estimated subgroups; for our sample size of 1500, each of the 10 estimated subgroups is expected to contain 150 individuals. While there is no analogous sample size imposed on the subgroups estimated by FS, the groups tend to contain between 110 and 180 individuals. If we consider that individuals in true Group 1 have value $(E_1, E_2, E_3) = (0, 0, 1)$ and Pr$\{(E_1, E_2, E_3) = (0, 0, 1)\} = 1/2^3 = 0.125$, then we expect $0.125 \times 1500 \approx 188$ individuals to be in Group 1. In a procedure that does a good job of grouping truly similar observations together, we'd expect to see Group 1 concentrated within the first estimated decile and the remaining $188 - 150 = 38$ observations in the contiguous decile.

Looking at the results from BART we see exactly what we expect given the particular data generation scenario. Under Scenario A, there is no effect modification, and observations from each of the true Groups are evenly distributed across the estimated subgroups. Under Scenarios B, C, and D, where effect modification is present, $150/188 = 80\%$ of observations in true Group 1 are in the first estimated decile, and $38/188 = 20\%$ are in the contiguous decile. These percentages are reflected in the underlying cell color; the full range of colors is given by the scale at the bottom of the figure. In true Group 2 there are 188 individuals, with $150 - 38 = 112$ in the second decile and $188 - 112 = 76$ in the third. Again, we see exactly what we expect, with $112/188 = 60\%$ of true Group 2 in the second decile and $76/188 = 40\%$ in the third, and these percentages reflected in the underlying cell color. As demonstrated by these percentages, because the true subgroup sizes are not multiples of the estimated subgroup sizes, the estimated subgroups are heterogenous with respect to the true subgroups.

This heterogeneity is also made obvious in Figure 2.2. We expect the first estimated subgroup to be completely comprised of individuals from true Group 1 and the second estimated subgroup to have $38/150 = 25\%$ from true Group 1 and $112/150 = 75\%$ from true Group 2. Thus we would expect the treatment effect of the first estimated subgroup (i.e., `TE(1)`) to be 1, and the estimated treatment effect of the second estimated subgroup to be $25\% \times 1 + 75\% \times 2 = 1.75$, and this is exactly what we see under Scenarios B, C, and D for BART. Of note is the boxplot associated with the third estimated subgroup, which we expect to be comprised of $76/150 = 51\%$ from true Group 2 and $74/150 = 49\%$ from true Group 3, with an expected treatment effect of $51\% \times 2 + 49\% \times 5 = 3.5$. This particular boxplot echoes the feature of Figure 2.1

where the estimated subgroups are heterogenous with respect to the true subgroups. This also emphasizes the importance of estimating a large number of subgroups relative to the true number, because the analyst will want the estimated subgroups to be homogeneous.

Returning to the BART results in Figure 2.1, not only is the distribution of true Groups 1 and 2 as expected, the distribution of all eight true Groups is as we would expect. In short, BART does an excellent job of grouping truly similar observations. This is demonstrated by the clustering of large percentages in a given row, and the diagonal pattern of cell coloring. As explained in detail above, the spread over three versus two columns is purely a function of the true group size; larger groups will spread over more columns. Table 2.3 quantifies this concordance, reporting a high "sensitivities" across all simulation scenarios.

Next considering the GBM analysis, the results confirm our earlier hypothesis, that an effect modifier must be associated with treatment for the propensity score to have any chance of detecting the resulting TEH. Under Scenario B where there is effect modification but the EMs are not associated with treatment (i.e., the EMs are not confounders), GBM does no better than random assignment of observations. This is demonstrated by the equal ($\approx 10\%$) allocation of each true subgroup (rows) across the ten estimated subgroups (columns), by the relatively uniform red coloring across the summary table, and by the relatively small "sensitivity" reported in Table 2.3. However in Scenario D where all three EMs are associated with treatment, GBM is able to detect some of the underlying data structure. Looking generally at the distribution of cell percentages/coloring in this table, we see two blocks of orange & yellow coloring, in the upper left and the lower right. What is being manifested in this separation is GBM's detection of the one EM (E_3) that has a strong association with treatment relative to the other two EMs (logistic regression coefficients of $(E_1, E_2, E_3) = (-0.1, 1.1, -4)$) and the other covariates in the dataset (see Table 2.2). The four rows (the first, second, third, and fifth) with more orange coloring on the left represent the four true subgroups with $E_3 = 1$ and the remaining four rows represent the subgroups with $E_3 = 0$.

Perhaps the most interesting aspect of the results from GBM is the 11th "undefined" column, where the method is alerting us to positivity violations. There are certain combinations of EMs that lead to extreme average propensity score values within that subgroup. For example, observations with $(E_1, E_2, E_3) = (0, 0, 1)$ and $(E_1, E_2, E_3) = (1, 0, 1)$ have average propensity scores of 0.06 and 0.04, respectively. Such a low average propensity score means that in finite settings like this one, where GBM is able to (somewhat) correctly group these observations together, often the estimated subgroups will not have any treated observations and the TE will be undefined. Figure 2.1 quantifies this, with brighter colors in the 11th column indicating more extreme positivity issues.

Now looking towards the results generated by FS, under Scenario B where there is effect modification and no confounding, the blocks of color suggest that the method is able to group observations by the magnitude of their treatment effect, rather than their covariate values: true subgroups with ATEs of 1 and 2 are grouped together (rows 1 and 2), as are true subgroups with ATEs of 5 and 6 (rows 3 through 6), and true subgroups with ATEs 9 and 10 (rows 7 and 8). A preliminary investigation of PAM partitioning the data into 20 subsamples revealed four large groups, implying that specifying more subgroups would have detected more structure, but that not all structure would have been detected due to the relatively weak association between E_1 and Y.

Comparing the FS results of Scenario C to Scenario B, we see the beginnings of a bifurcation of the middle block of observations. Because we have now introduced covariates that have treatment and outcome associations of similar strength to E_2, the decision trees constructed by FS choose E_2 less often as a variable that defines a decision rule, instead choosing the other covariates. Because E_2 has become relatively less important in defining subgroups, the movement of observations is towards the groups defined by strong EM E_3. Thus, the third and fifth rows, which have value $E_3 = 1$, are moving towards the grouping of the first and second rows which also have value $E_3 = 1$; similarly for the fourth and sixth rows with value $E_3 = 0$.

Our final comments on FS are about Scenario D, which demonstrates a potential pitfall of empirical positivity enforcement through algorithmic stopping rules. As expected, FS is able to group observations by strong EM E_3. Different from Scenario C where E_3 is not associated with treatment, here E_3 has a strong negative association with treatment. Thus observations with $E_3 = 1$, those in rows 1, 2, 3, and 5, have very low values of the propensity score. Decision tree nodes with these observations cannot be split any further because the stopping rule requiring a minimum number of treated observations will have been triggered. Comparatively, observations with $E_3 = 0$ have moderate propensity score values, so are able to be further divided by E_2.

While the above simulation study was intentionally simplistic in its data generation, the Supplementary Material present an analogous simulation where data are generated to more directly mimic the covariate distribution observed in the CER investigation of §2.5. The general conclusions of this supplementary simulation study are the same as those presented here.

2.5 Comparison of Methodologies: CER for Cardiovascular stents in Medicare beneficiaries

Drug-eluting stents (DES) have been widely adopted as a non-inferior alternative to bare-metal stents (BMS) for treatment following myocardial infarction (MI) [44], with clinical-trials evidence indicating important effect modification by diabetes [9] and age [17, 41, 53]. In this data analysis, we consider the comparison of drug-eluting stents (DES) to bare-metal stents (BMS) as treatment of myocardial infarction (MI), by looking at the association of each with the two-year revascularization rate. Our goal in this exploratory analysis is to evaluate whether TEH is present, as determined by the three estimation methods under consideration and using our knowledge of their operating characteristics.

2.5.1 Data Structure

De-identified inpatient data on 38 covariates were generated by 54,099 Medicare beneficiaries hospitalized in the continental United States in 2008 with their first MI. An unadjusted comparison of the two-year revascularization rate in DES and BMS patients yields a risk difference of –0.055, indicative of a worse outcome with BMS and thus consistent with the literature, but thought to be confounded by patient characteristics that help determine treatment choice. As summarized in Table 2.4, patients receiving BMS generally have a higher baseline risk profile.

To evaluate TEH, each of the three estimation methods were applied to the data. In the case of BART and GBM, such evaluation was done by partitioning the estimated probability distribution (be it of the outcome or treatment) into 500 quantiles, then estimating a TE within each subgroup defined by these quantiles. In the case of FS, the dissimilarity matrix $\boldsymbol{D}$ is partitioned into 500 subgroups using PAM and a TE estimated within each.

2.5.2 Results

The results of this data analysis are summarized in Figure 2.3. For each estimation method, the subgroup-specific treatment effects (the risk difference) and associated uncertainty intervals are plotted in ascending order. The red vertical line denotes the marginal estimated risk difference. None of the estimation methods

	DES	**BMS**
	(30,562)	**(23,537)**
Race: white	90.2	90.1
Male	57.2	57.8
Age (years)	74.9	76.2
Region		
west	16.0	13.7
midwest	27.7	30.0
south	40.6	38.8
northeast	15.7	17.5
COPD	15.7	16.9
Asthma	2.6	2.5
Prior Coronary artery bypass graft (CABG) performed	0.4	1.1
Prior congestive heart failure	6.8	7.1
Prior myocardial infarction	3.0	2.8
Unstable angina	3.2	2.3
Chronic atherosclerosis	90.6	88.4
Respiratory failure	2.3	2.6
Hypertension	69.2	65.8
Prior stroke	1.0	1.3
Cerebrovascular disease (non stroke)	2.7	3.1
Renal failure	5.4	6.2
Pneumonia	6.7	8.5
Malnutrition	1.2	2.2
Dementia	3.3	5.3
Functional disability	1.3	1.7
Peripheral vascular disease	4.4	4.7
Trauma in the past year	3.5	4.3
Major psychiatric disorder	1.2	1.6
Liver disease	0.2	0.5
Severe hematological disorder	0.4	0.7
Anemia	14.6	18.2
Depression	4.8	4.9
Parkinsons/Huntington	0.9	1.0
Seizure disorder	1.1	1.4
Chronic fibrosis	1.4	1.6
Vertebral fractures	0.6	0.7
Cancer	3.6	6.3
Eligible for Medicaid	12.0	13.2
Diabetes	29.9	26.3
Revascularization within two years	22.0	27.5

Table 2.4 Baseline characteristics (% experiencing unless otherwise indicated) and one-year hospital readmission rate for DES (“treated”) and BMS (“untreated”) patients (columns 1 and 2). See §2.5.1 for details on the population that generated these data.

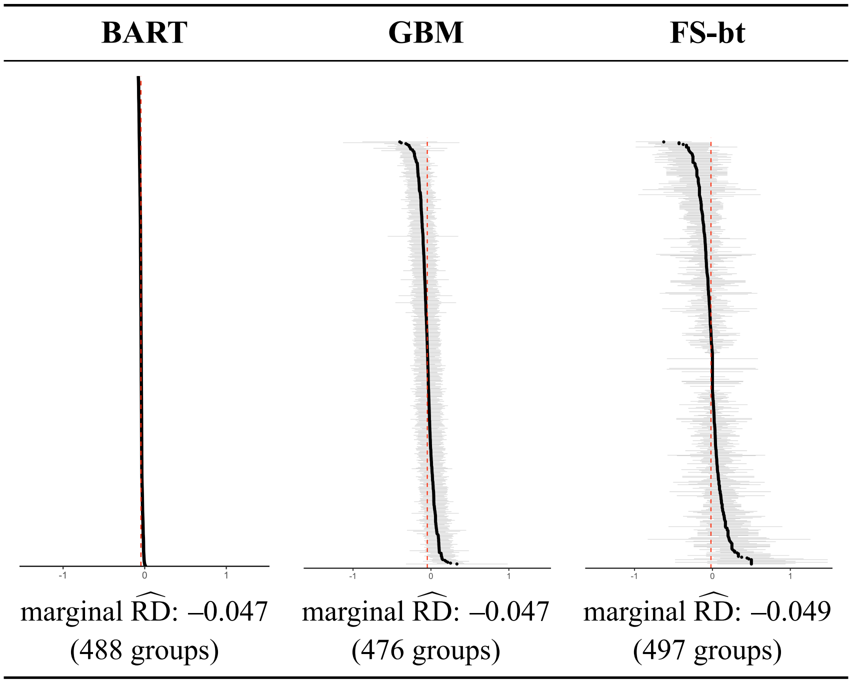


Figure 2.3 Visualization of results from Data Analysis 2.5. Details regarding how the figure was constructed, and how to interpret the figure, are given in §2.5.2.

were able to estimate 500 subgroups, for differing reasons. In the case of BART, there were several hundreds of patients with the same estimated ITE, and those subgroups could not be disaggregated. For GBM, 24 of the 500 subgroups had an undefined treatment effect (i.e., at least one treatment arm with less than 2 observations), 3 for FS.

Focusing on the results from BART (which proved most promising in the simulation studies), we investigate whether any individual covariates exhibit a clear association with subgroup membership. Towards this goal, we plot ITEs and subgroup average ITEs across the distribution of each covariates, as illustrated for four covariates in Figure 2.4. To ease explanation, consider the top-most plot marked *age*. The y-axis represents the subgroup-specific average age in years, and the x-axis represents the subgroup-specific ATE (again measured on the risk-difference scale). A red dot represents a subgroup, generated as described earlier: the $54,099$ posterior means are partitioned into 500 subgroups, and the average age and average TE (taken as the average of the posterior mean ITEs within that subgroup) are plotted. Thus, we expect 500 red dots in this single plot. To generate the values that the gray dots represent, we applied the subgrouping process used on the posterior means (i.e., partitioned into 500 subgroups and calculated subgroup-specific averages) to each of the 1000 posterior draws of 54,099 ITEs, and plotted a random subset of 100.

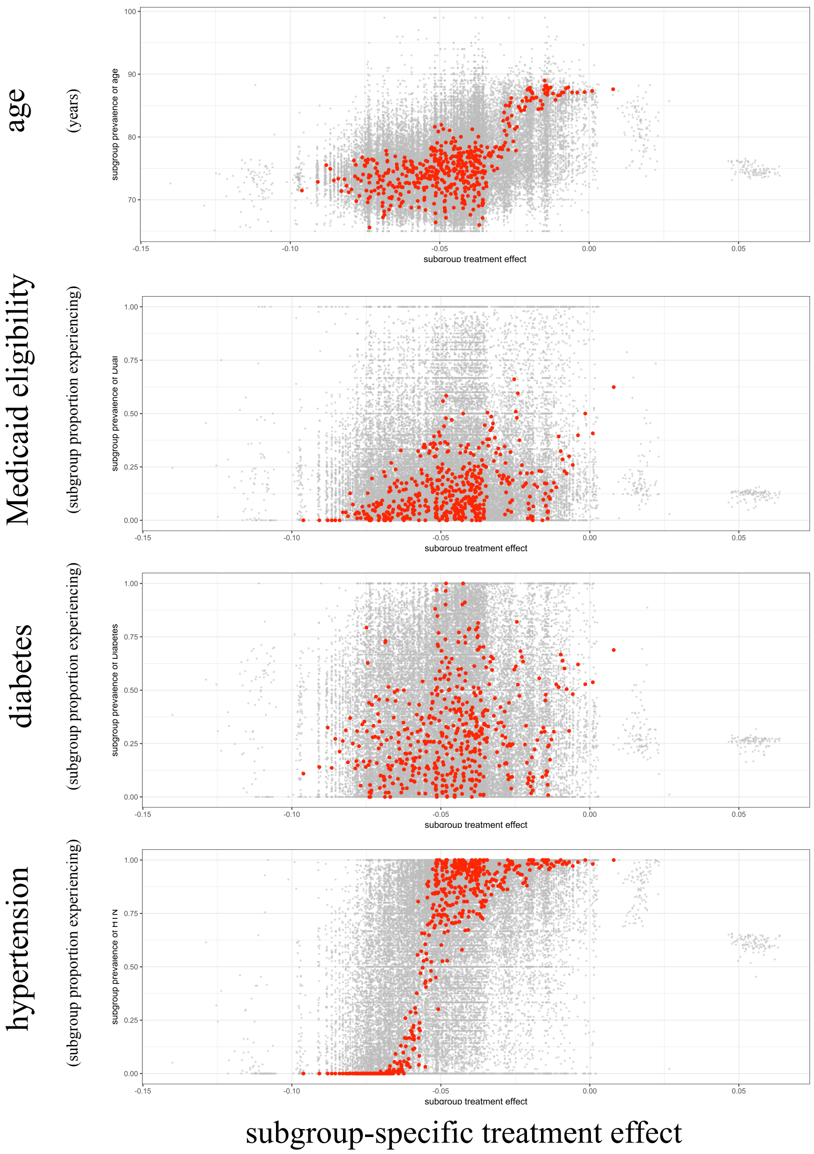


Figure 2.4 Visualization of results from Data Analysis 2.5. Covariates displayed are *age*, *Medicaid eligibility*, prior *diabetes* diagnosis, and prior *hypertension* diagnosis. The y-axis represents the subgroup-specific average, and the x-axis represents the subgroup-specific ATE (on the risk-difference scale). A red dot represents a subgroup, generated as described in §2.5.2: the $54\,099$ posterior means are partitioned into 500 subgroups, and the average covariate measure (e.g., average age) and average TE (taken as the average of the posterior mean ITEs within that subgroup) are plotted. Thus, we expect 500 red dots in a single plot. To generate the values that the gray dots represent, we applied the subgrouping process used on the posterior means (i.e., partitioned into 500 subgroups and calculated subgroup-specific averages) to each of the 1000 posterior draws of 54 099 ITEs, and plotted a random subset of 100.

2.5.3 Discussion

A qualitative analysis of the forest plots in in Figure 2.3 suggests that none of the three estimation procedures detect any treatment effect heterogeneity; the variability in subgroup-specific estimates is as one might expect from sampling variability. However the successful performance of BART in the simulation studies leads us to further investigation, as displayed in Figure 2.4.

Considering Figure 2.4, there is evidence of quantitative effect modification by *age*; it appears that DES lead to better outcomes in younger patients, a benefit that decreases to nearly zero in older patients. These conclusions are supported by the literature, where it is known that DES generally leads to better outcomes than BMS but the increased comordibities, bleeding risk, and frailty of elderly patients may negate the beneficial effects [17, 41, 53]. There is also compelling evidence of quantitative effect modification by *hypertension*, where absence of hypertension is associated with better outcomes within DES patients, as compared to BMS patients. The figure does not imply effect modification by *Medicaid eligibility* or *diabetes*.

The lack of evidence of effect modification by *diabetes* may come as a surprise because of the physiological [4, 40] and randomized clinical trial [9] evidence that supports the general understanding among clinicians that the effect of stent type on adverse cardiovascular outcomes is different within diabetic patients. However what we are seeing is a well-known issue with measurement of diabetes prevalence: it is subject to high rates of misclassification [25]. High blood pressure, on the other hand, is positively associated with diabetes [43] and is much easier to determine. So in fact, it is possible that we are seeing the effect modification of diabetes through a proxy.

2.6 Conclusion

The goal of a TEH estimation method is to provide a partitioning of the covariate space into interpretable subgroups, identifiable by covariate values. Such a partition would be extracted from the data, rather than specified *a priori* by the researcher. Historically, detection of TEH has involved identification of effect modifiers by subject matter experts, then an evaluation of the estimated treatment effect within each subgroup. This sort of *a priori* specification precluded exploratory analyses of TEH (for good reason, out of a desire to constrain the type I and II error rates), treated confounding and TEH as separate issues, and was not scalable to high-dimensional data. Uncertainty in confounder and/or effect modifier selection was not addressed by these methods. Thus out of necessity, we have seen an evolution of estimation procedures to match the increased complexity of our research questions and our data.

We contribute to the ongoing discussion by briefly reviewing and evaluating three general classes of modeling approach, through a performance comparison of representative modern methods from each class. We considered the ability of each method to detect subgroups in an exploratory, hypothesis-generating manner. Our simulation studies revealed that GBM, as a representation of using propensity scores to estimate causal effects, is not able to detect effect modifiers that are not associated with treatment; that is, effect modifiers that are not also confounders. However, GBM is able to alert the analyst to positivity violations whereas the representative methods from the other modeling classes extrapolate, possibly inappropriately. For example, FS, as a representation of the joint modeling of outcome and treatment conditional on covariates, potentially fails to disaggregate when the treatment prevalence is extreme, leading the analyst to draw conclusions on the aggregate. BART, as a representation of modeling the outcome conditional on covariates, does not require observations in both treatment arms to calculate the subgroup ATE so positivity violations may go unnoticed. The ability of each method to estimate an unbiased subgroup-specific treatment effect is related to its ability to group similar observations together, and to the number of subgroups that the sample is initially partitioned into by the analyst. When the initial partition is too coarse, the resulting subgroups are still heterogeneous with respect to the true subgroups, leading to biased treatment effect estimates. The conclusions drawn from our simulation studies were used to evaluate the results of a comparative effectiveness analysis, looking at the effect of stent type on an adverse cardiovascular outcome. Diabetes status and age are known in the cardiovascular literature as an effect modifier, and presented itself as such in our analysis, although through the correlated *hypertension* covariate.

Our analyses do have some limitations, that present opportunities for future work. Our heuristic study was designed to gain some intuition about the more mathematically-rigorous evaluative measures, so we do not address measurement of uncertainty in the subgroup-specific treatment effect estimates, nor any of the classical statistical performance metrics (e.g., consistency). There are also potential problems with using the same data to estimate subgroups and treatment effects, and future work would explore ways to avoid this. Future work would also explore ways to estimate the number of underlying subgroups from the data, rather than *a priori* specification by the analyst. Lastly, we explicitly explored methods designed to “automatically” detect which of the measured variables are confounders and/or effect modifiers. This was motivated by the desire to address settings where the sheer number of measured covariates or limited contextual knowledge precluded prior specification of such variables. However, such analyses entail important limitations. All methods we consider rely on the assumption that the entirety of $\boldsymbol{X}$ is measured pretreatment and thus unaffected by T. If this assumption was violated, such automated procedures could be susceptible to forms of bias such as posttretment selection bias or M-bias. Furthermore more, if the available variables in $\boldsymbol{X}$ do not contain important confounders or proxies, then the methods explored here would still suffer from unobserved confounding bias.

3

Visualization software for exploring treatment effect heterogeneity

Sarah C. Anoke[1], Christine Choirat[1], and Corwin M. Zigler[1]

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

Abstract

Identification of treatment effect heterogeneity (TEH) involves comparisons of covariate distributions, both within and across subgroups. Feasibility quickly decreases as the number of covariates and/or subgroups increases. This chapter describes use of the web application framework `shiny` and statistical programming language `R` to construct visualization software for exploratory, evidenced-based hypothesis-generating analyses of treatment effect heterogeneity. The software consists of three main features: (1) a forest plot displaying all subgroup-specific treatment effects, (2) subgroup profile plots displaying subgroup-specific covariate means in a way that highlights the distinguishing features of subgroups, and (3) covariate profile plots that facilitate the identification of effect modifiers by displaying covariate distributions as a function of the subgroup-specific treatment effect. The ability of these tools to contribute to the identification of TEH is demonstrated in a comparative effectiveness evaluation of drug-eluting versus bare-metal stents among 54,099 Medicare beneficiaries in the continental United States admitted to a hospital with acute myocardial infarction in 2008.

To download `hetviz`, see instructions in the User Manual at

`https://github.com/sanoke/hetviz/wiki`.

3.1 Introduction

A pioneer of visualization, statistician John W. Tukey has said that "the greatest value of a picture is when it forces us to notice what we never expected to see" [72]. Statistics, as quantitative storytelling, and data science, the combination of such storytelling with computer science, is enhanced by the use of data visualizations because such illustrations allow the analyst to tell a more complete story [46].

Visualization has important relevance to the identification of treatment effect heterogeneity (TEH), the existence of an underlying partition of a population into subgroups across which the effect of a treatment varies systematically. Complete ascertainment of TEH is defined as correct estimation of the number of underlying subgroups, interpretable characterization of the subgroups by observed covariates, and unconfounded estimation of the treatment effect within each subgroup. As discussed in Chapter 2, there are several common causal inference methods with the population causal effect as the target estimand, but they are able to be repurposed for exploratory identification of TEH. The analyst must *a priori* specify a maximum number of possible subgroups to investigate, relying on the ability to collapse across subgroups if the data indicate fewer. After deciding on an estimation method and how to use the method to assign subgroup membership, interpretation of which characteristics define each subgroup relies on the analyst's ability to inspect covariates distributions within each group to infer what characterizes them. Assuming ignorability is satisfied by some subgroup-specific confounder covariate set $\mathbf{X}_{\text{subgroup}}$, unconfounded subgroup-specific treatment effect estimation is possible. TEH is present if the estimated treatment effects vary across the subgroups, and the covariates that define this final partition are referred to as *effect modifiers*.

In high dimensional settings visualizations are important because they help the analyst overcome the infeasibility of manual evaluation, allows the analyst to double-check their intuition regarding what they expect to see, and to again quote Dr. Tukey, can reveal insights the analyst did not expect to see. The manual comparison of several subgroup-specific high-dimensional covariate distributions is incredibly difficult, at best. Thus we use visualization to assist in our ability to reason about the distinguishing features of these distributions, and lead to the generation of hypotheses about which, out of many covariates, are important effect modifiers. Manual evaluation is still important, because, at least in the public health setting, there are some considerations that cannot be delegated to a computer. But we need methods that will help us process this larger picture, because humans are susceptible to information blindness, a phenomenon where the brain stops absorbing information when there is too much to take in [23].

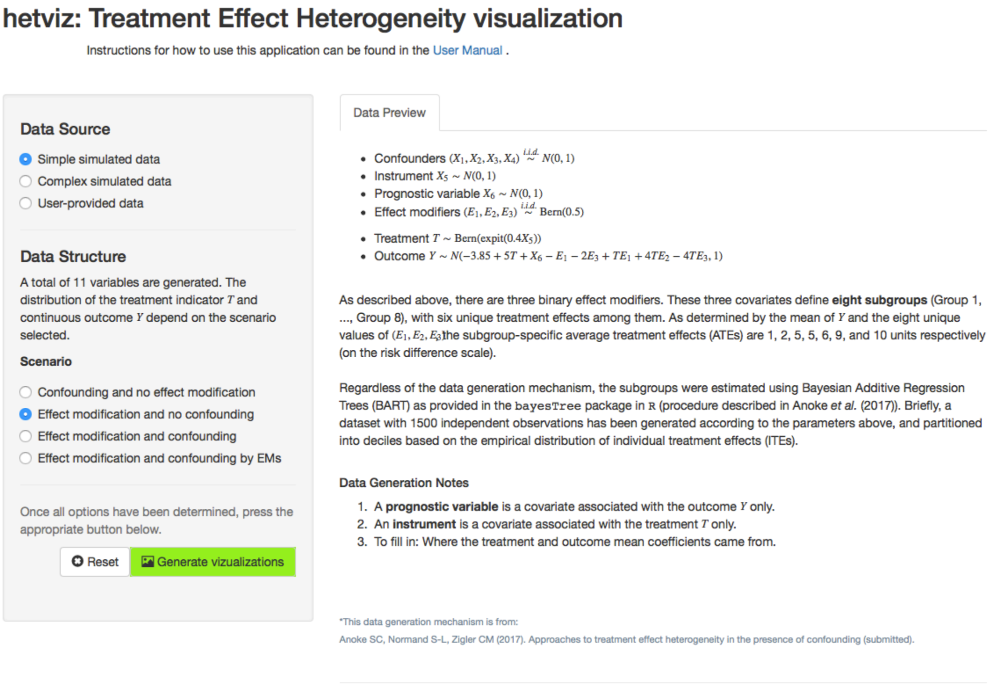
Thus "data visualization" allows for the coherent organization of large amounts of information. Although this term has enjoyed recent popularity as the associated technology has become more powerful and less expensive, the use of high-resolution summary graphics dates back many decades. Referred to by Tufte [70] as possibly "the best statistical graphic ever drawn", the illustration of Napoleon's 1812 march through Russia by French engineer Charles Joseph Minard is renowned for its communication of complex, multivariate information in a clear and coherent manner. French cartographer Jacques Bertin [10] and statisticians William Cleveland [20] and John Tukey [72] are among the first to add academic rigor to the field of information design and are pioneers of the minimalist, high resolution statistical graphics that are common today. In fact, the ubiquity of computers and available data have broadened the audience interested in the construction and consumption of such graphics, and Edward Tufte [67, 68, 69, 70, 71] and Garrett Grolemond and Hadley Wickam [29] have taken the theoretical foundations of graphics design and made them accessible to laypersons.

It is on the shoulders of these giants that we address the problem at hand: the creation of software for making determinations of TEH by automating the comparison of several high-dimensional covariate distributions and facilitating identification of distinguishing features, if any, of each. Noting that such features are easier to distinguish visually, but not wanting the analyst to be debilitated by information blindness [23] or cognitive overload [29], our proposed software `hetviz` is thoughtful in how this information is presented. Most existing data visualization software tends to be spatially-concerned; geography and mapping lend themselves to visualization. However to our knowledge, this is the first visualization software that allows for exploratory, hypothesis-generating analyses of subgroup-specific distributions in causal inference. It allows the user to engage with their data in a way that numerical summaries preclude. This idea of engagement with the data to inform future analyses is by no means new, advocated by Tukey [72], Tufte [70], Cleveland [20], and Breiman [14], among others. To quote Cleveland [20, p. 219], the goal of our software is to "…convey…the empirical distribution of the data and not to make formal statistical inferences about a population distribution from which the data might have come".

This introduction of the software will proceed as follows. In §3.2 we use simulated data to demonstrate the general structure of data provided to the software for analysis and the three main functionalities of the software: the creation of a forest plot summarizing the estimated treatment effect in each subgroup (§3.2.2), the visualization of subgroup profiles (§3.2.3), and the visualization of covariate distributions across subgroups (§3.2.4). §3.3 details a real data application to demonstrate how this software can aid in the identification of TEH in practice. We conclude in §3.4 with a summary and general guidance on how to include the software in practice.

3.2 Functionality

When the application is first opened, the user will see an interface as displayed in Figure 3.1. The gray left-hand sidebar allows the user to provide their data, and the right-hand panel displays information about the provided data.



hetviz: Treatment Effect Heterogeneity visualization

Instructions for how to use this application can be found in the User Manual .

Data Source

● Simple simulated data
○ Complex simulated data
○ User-provided data

Data Structure

A total of 11 variables are generated. The distribution of the treatment indicator T and continuous outcome Y depend on the scenario selected.

Scenario

○ Confounding and no effect modification
● Effect modification and no confounding
○ Effect modification and confounding
○ Effect modification and confounding by EMs

Once all options have been determined, press the appropriate button below.

- Confounders $(X_1, X_2, X_3, X_4) \overset{i.i.d.}{\sim} N(0,1)$
- Instrument $X_5 \sim N(0,1)$
- Prognostic variable $X_6 \sim N(0,1)$
- Effect modifiers $(E_1, E_2, E_3) \overset{i.i.d.}{\sim} \text{Bern}(0.5)$
- Treatment $T \sim \text{Bern}(\text{expit}(0.4X_5))$
- Outcome $Y \sim N(-3.85 + 5T + X_6 - E_1 - 2E_3 + TE_1 + 4TE_2 - 4TE_3, 1)$

As described above, there are three binary effect modifiers. These three covariates define **eight subgroups** (Group 1, ..., Group 8), with six unique treatment effects among them. As determined by the mean of Y and the eight unique values of $(E_1, E_2, E_3)$the subgroup-specific average treatment effects (ATEs) are 1, 2, 5, 5, 6, 9, and 10 units respectively (on the risk difference scale).

Regardless of the data generation mechanism, the subgroups were estimated using Bayesian Additive Regression Trees (BART) as provided in the `bayesTree` package in `R` (procedure described in Anoke *et al.* (2017)). Briefly, a dataset with 1500 independent observations has been generated according to the parameters above, and partitioned into deciles based on the empirical distribution of individual treatment effects (ITEs).

Data Generation Notes

1. A **prognostic variable** is a covariate associated with the outcome Y only.
2. An **instrument** is a covariate associated with the treatment T only.
3. To fill in: Where the treatment and outcome mean coefficients came from.

*This data generation mechanism is from:

Anoke SC, Normand S-L, Zigler CM (2017). Approaches to treatment effect heterogeneity in the presence of confounding (submitted).

Figure 3.1 View of the initial `hetviz` user interface when the application is first opened.

3.2.1 Dataset provision

The major functionalities of `hetviz` will be demonstrated using simulated data described in §2.4.1 and Table 2.2 on pages 30 and 31, respectively. These are the same data as provided in the “simple simulated data” option of the interface (Figure 3.1, top left). To briefly summarize, these data are comprised of 1,500 observations and 8 approximately equal-sized subgroups with 6 unique treatment effect values, $(1, 2, 5, 6, 9, 10)$. Additionally the software provides “complex simulated data”, data simulated to mimic observed health data. These simulated datasets are provided to the user as examples that can be used to interactively explore the functionality of `hetviz`.

The user is able to provide their own data for visualization within the software, either as (1) a local comma-delimited (i.e., CSV) file, (2) a remote comma-delimited file with its location specified with a URL, or (2) a table within a PostgreSQL database. At minimum, the dataset should contain an outcome variable (which can be binary or continuous), a treatment variable (binary, coded as 0 for ‘control’ and ‘1’ as treated), and a variable containing each observation’s subgroup assignment (integer-valued). Due to scalability issues of third-party software, `hetviz` can accommodate up to 48 additional covariates (although there are plans to extend this). Upon providing the location of the dataset, the user will be given an opportunity to provide the name of the variables that correspond to the (binary) treatment, the outcome, and the subgroup specification. Treatment effects are calculated within the software on the difference scale (i.e., risk differences).

Once the user has completed specification of their dataset to the software, s/he can press the green “Generate visualizations” button (Figure 3.1, bottom left). The first visualization that will appear is the forest plot.

3.2.2 Feature 1: Forest plot

The forest plot visualization allows the analyst to compare the subgroup-specific treatment effects and identify any general patterns that would suggest effect heterogeneity. Examples of the graphs generated are provided in Figures 3.2(a) and 3.2(b). On the y-axis of each graph is the subgroup-specific treatment effect, and on the x-axis is an integer denoting subgroup membership. Internally, the software reassigns subgroup membership labels so that subgroup 1 has the smallest estimated TE, subgroup 2 has the second smallest estimated TE, etc.

Figure 3.2(a) displays the default forest plot. There are 10 red circles denoting the estimated treatment effect for each of the 10 subgroups, as well as a thick error bar ranging from the 25th to the 75th quantile of estimated ITEs for that subgroup (much like the central box of a box plot) and a thin error bar with a maximum length of 1.5 times the interquartile range (much like the whiskers of a box plot). If the user does not provide estimated ITEs for each observation, then the error bars correspond to those of a 95% confidence interval. This presentation was chosen to be more accommodating to large numbers of subgroups than the traditional boxplot.

If the dataset is small (i.e., $\leq$ 5,000 observations and $\leq$ 30 subgroups) then the user is able to select an alternative forest plot presentation displayed in Figure 3.2(b), that includes a traditional boxplot presentation in addition to the specific estimated ITE values.

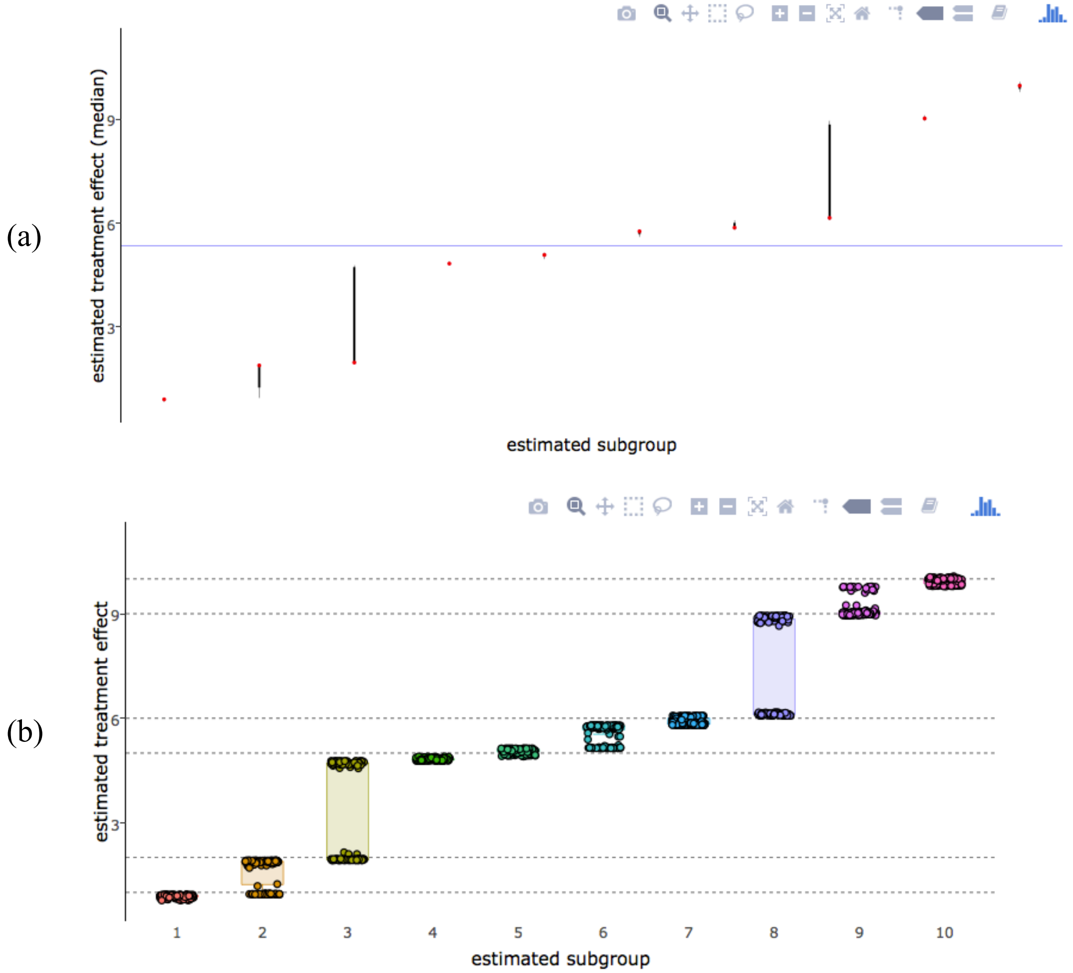


Figure 3.2 Forest plot visualization allows the user to compare the subgroup-specific treatment effects and identify any general patterns that suggest effect heterogeneity. Subfigure (a) is the default forest plot presentation, that includes vertical lines denoting the locations of the true subgroup-specific treatment effects. Subfigure (b) is an additional, alternative presentation for small datasets.

3.2.3 Feature 2: Subgroup profiles

While the associated plots will be explained in detail in the data analysis example of §3.3, a brief summary is given here. An example of the plots generated by this functionality is given in Figure 3.4.

The main subgroup profile plot is displayed in Figure 3.4(a). On the x-axis is each covariate in the dataset. On the y-axis is distance from each covariate's marginal mean, measured in marginal standard errors[*].

Thus in Figure 3.4(a), each line corresponds to a subgroup – each line is a subgroup profile. Plotting these profiles together allows the user to visually compare the subgroups for distinguishing features. The user is also able to hover their mouse pointer over any point in the graph to see what subgroup the datum corresponds to. When an extreme value has been identified, as well as the associated subgroup, the user is able to use the visualization in 3.4(b) to isolate that subgroup and look at entire subgroup-specific distributions (rather than just the mean) of each covariate in that subgroup.

3.2.4 Feature 3: Covariate profiles

Again, the associated plots will be explained in detail in the data analysis example of §3.3, but a brief summary is given here. An example of the plots generated by this functionality is given in Figure 3.5.

The main covariate profile plot is displayed in Figure 3.5(a). On the x-axis is an integer denoting subgroup membership. Recall that internally `hetviz` numbers observations such that subgroup 1 corresponds to the subgroup with the smallest estimated treatment effect, so movement along the x-axis means that the subgroup treatment effect is increasing. On the y axis is distance from each covariate's marginal mean, measured in marginal standard errors[†]. Note that this y-axis is the same as the y-axis in the subgroup profile plot.

Thus in Figure 3.5(a), each line corresponds to a covariate – each line is a covariate profile. Plotting these profiles together allows the user to visually compare how covariate distributions change as a function of the subgroup-specific treatment effect. If the covariate is an effect modifier, we would expect to see a nonlinear profile, or a profile with a slope different from zero.

As with the subgroup profile plot, the user is able to hover their mouse pointer over any point in the graph to see what covariate the datum corresponds to. When an extreme value has been identified, as well as the

[*] This will be changed to marginal standard deviations, so that for a given covariate, all subgroups have the same unit of distance (rather than the unit of distance being dependent on subgroup size).

[†] This will be changed to marginal standard deviations, so that for a given covariate, all subgroups have the same unit of distance (rather than the unit of distance being dependent on subgroup size).

associated covariate, the user is able to use the visualization in 3.5(b) to isolate that covariate and look more closely at how the distribution of that covariate changes as a function of the subgroup-specific treatment effect. On the x-axis is the subgroup-specific treatment effect, and on the y-axis is the subgroup-specific covariate average. Within the plot itself, there is a datum (i.e., a black point) for each subgroup. Each of these subgroup points has red error bars, denoting the subgroup specific standard error (the subgroup-specific standard deviation of that covariate divided by the number of observations in that subgroup).

A major benefit of such figures is being able to identify nonlinear TEH without having to know *a priori* the relationship between the covariate and the subgroup-specific treatment effect.

3.3 Data Analysis Example

In this section we demonstrate the use of `hetviz` in a treatment effect heterogeneity analysis. In particular, we will investigate the Medicare data described in §2.5.1 and Table 2.4 on pages 39 and 40, respectively.

When the data are first provided to the software, the first visualization that the user will see is Figure 3.3. The red dots denote the subgroup-specific treatment effects, and the black lines are the error bars. Within the software itself, we can zoom in to see that the error bars are so narrow that they can only be seen at a high magnification. Figure 3.3 suggests that there may be heterogeneity, although the relatively narrow range of the subgroup-specific risk differences ($\approx$ 0 to -10%) imply that what we see is sampling variability, rather than systematic.

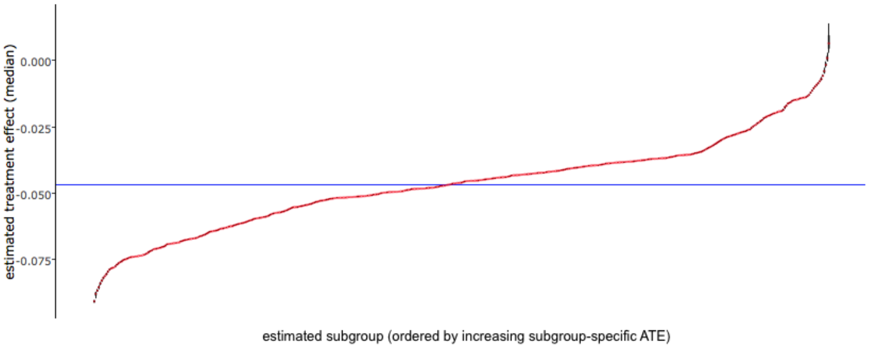
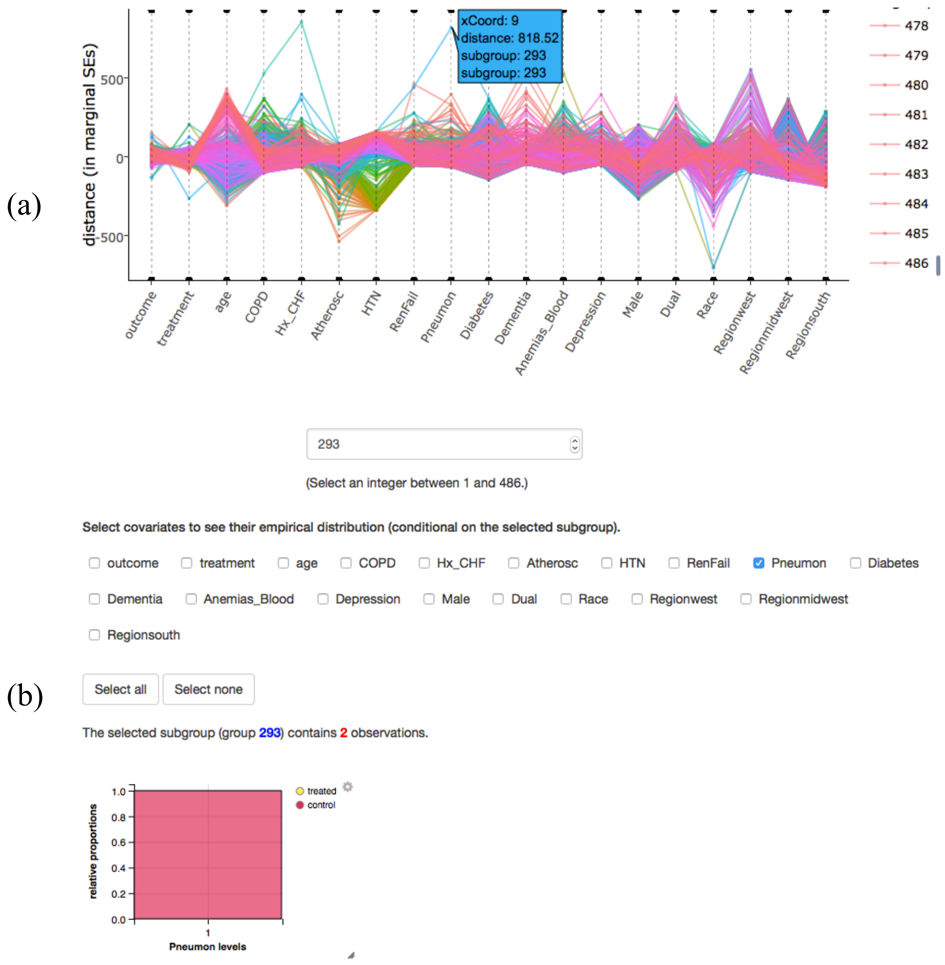


Figure 3.3 Forest plot visualization of Medicare data analysis described in §2.5.

The next visualization is the subgroup profiles, displayed in Figure 3.4(a). We see one subgroup that stands out in blue. When we hover our mouse over this subgroup, we see that this extreme value occurs with the `Pneumon` covariate, an indicator of prior pneumonia diagnosis, and this profile belongs to subgroup 293. We then use the visualization in Figure 3.4(b) to see that this extreme value occurs because this subgroup only has two observations, who both have pneumonia and are both in the “control” group.



293

(Select an integer between 1 and 486.)

Select covariates to see their empirical distribution (conditional on the selected subgroup).

☐ outcome ☐ treatment ☐ age ☐ COPD ☐ Hx_CHF ☐ Atherosc ☐ HTN ☐ RenFail ☑ Pneumon ☐ Diabetes

☐ Dementia ☐ Anemias_Blood ☐ Depression ☐ Male ☐ Dual ☐ Race ☐ Regionwest ☐ Regionmidwest

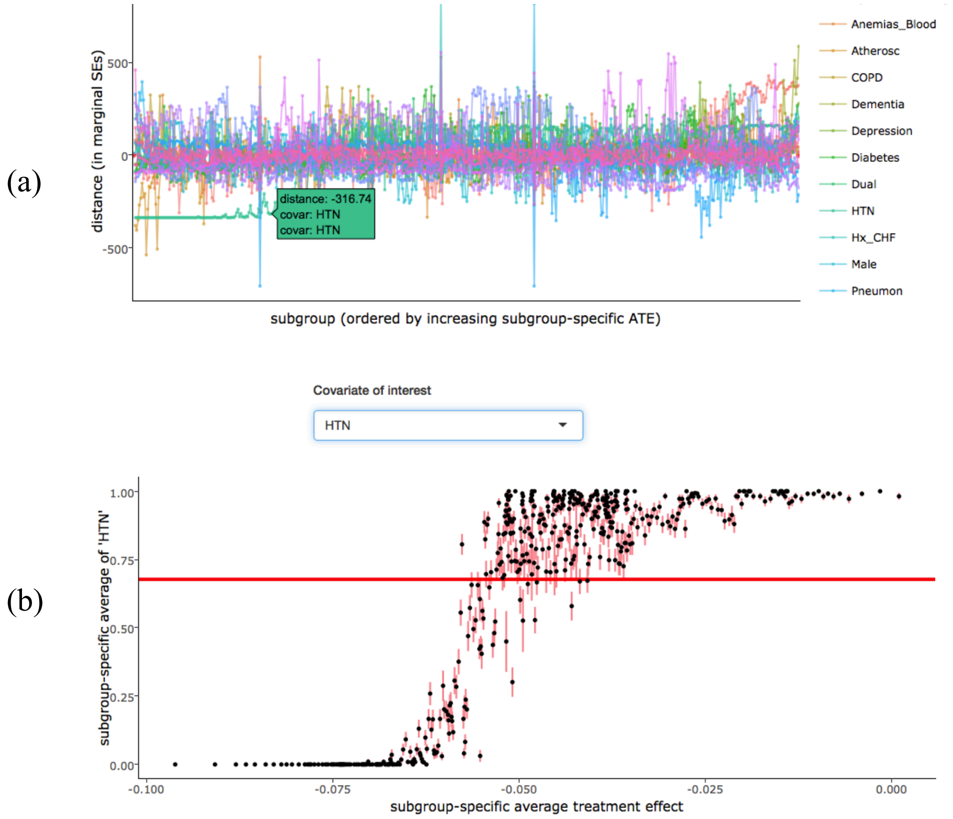
☐ Regionsouth

Select all | Select none

The selected subgroup (group 293) contains 2 observations.

Figure 3.4 Subgroup profile visualization allows the user to visually compare the subgroups for distinguishing features. Subfigure (a) gives an overall comparison of all subgroups, allowing the user to select subgroups that stand out as well as distinguishing covariates. When the user has selected a subgroup of interest, s/he is able to focus on that subgroup using the visualization in Subfigure (b), and view the full subgroup-specific distribution of each covariate.

We then decide to look at the profile of the `Pneumon` covariate, in Figure 3.5(a). The sharp blue spikes correspond to this covariate, but occur when the subgroup size is very small, so there is not much information in the spikes. Another covariate profile catches our eye, in the green – this covariate seems to be consistently and significantly distant from its marginal mean when the treatment effect is small, then sharply increases its distance as the treatment effect increases. When we place our mouse over a datum in that covariate profile, we see it belongs to `HTN`, an indicator of prior hypertension diagnosis. We use the visualization in Figure 3.5(b) to investigate this particular covariate more closely, and we find substantial evidence that `HTN` is an effect modifier.



Covariate of interest

HTN

Figure 3.5 Covariate profile visualization allows the user to visually compare how a covariate's distribution (through its mean) changes as a function of the subgroup-specific treatment effect. If the covariate is an effect modifier, we would expect to see a nonlinear profile, or a profile with a slope different from zero. The visualization of Subfigure (a) gives an overall comparison of all covariates, and Subfigure (b) allows the user to investigate a specific covariate.

3.4 Conclusion

Visualization helps analysts reason about their data and turn information into conclusions. Recent increases in the quality and quantity of available data present novel opportunities to strengthen the conclusions we are able to draw from them. However in exploratory analyses where hypotheses are not defined at the outset, we must take special care to familiarize ourselves with the content of our dataset. This familiarity leads to more informative conclusions when conducting formal analyses and making inferences.

The software `hetviz` facilitates this familiarity, by allowing for facile comparison of many subgroup-specific covariate distributions, towards the identification of effect modifiers in high-dimensional settings. This is accomplished through three major features: (1) a forest plot displaying all subgroup-specific treatment effects, (2) subgroup profile plots displaying subgroup-specific covariate means in a way that highlights the distinguishing features of subgroups, and (3) covariate profile plots that facilitate the identification of effect modifiers by displaying covariate distributions as a function of the subgroup-specific treatment effect. Furthermore, the publicly-available implementation of `hetviz` provides examples of the best practices of software design, including usability testing, measures of coverage, and reproducability.

This manuscript marks the first release of `hetviz`, and future releases will include improvements on current limitations. For example, `hetviz` can only visualize data stored in flat files. Large complex datasets are stored in relational databases, and in the future `hetviz` will support queries from such data sources. Additionally, it would be helpful for users to be able to generate and save customizable reports, containing figures generated during their exploratory analyses. However in its current form, `hetviz` is able to contribute to statistical analyses of large datasets by automating the generation of a holistic, broad investigation of the contents of large datasets when investigating treatment effect heterogeneity.

A

Detailed statistical results to accompany Chapter 1

Values marked with an asterisk (*) as reported in the 2011 Ugandan DHS Final Report. If a quantity is unmarked, it was calculated by the authors for this study. 'Women' are 15-49 years of age, 'men' are 15-54, and 'youth' are 15-24. Prevalences are reported with their associated 95% confidence intervals (CIs). Prevalences are compared using a two-sample two-sided *z*-test of proportions, and the associated *p*-value reported.

Indicator	**LQAS estimate (95% CI)**	**DHS estimate (95% CI)**	**Comparison (*p*-value)**
% of individuals who were counseled and received an HIV test in last 12 months and know their results.	Women ($n = 1445$): 0.440 (0.414, 0.465)	Women ($n = 1097$): 0.388* (0.351, 0.425)*	0.024
	Female youth ($n = 781$): 0.350 (0.317, 0.384)	Female youth ($n = 451$): 0.353 (0.306, 0.400)	0.921
	Men ($n = 1446$): 0.294 (0.271, 0.318)	Men ($n = 291$): 0.217 (0.166, 0.268)	0.006
	Male youth ($n = 633$): 0.204 (0.174, 0.237)	Male youth ($n = 116$): 0.161 (0.081, 0.240)	0.335
% of mothers of children 0-11 months who were counseled and received an HIV test during the last pregnancy and know the results.	Mothers ($n = 1446$): 0.870 (0.852, 0.886)	Mothers ($n = 205$): 0.820 (0.746, 0.893)	0.197

Table A.1 Comparison of *HIV Counseling and Testing* indicators.

Indicator	**LQAS estimate (95% CI)**	**DHS estimate (95% CI)**	**Comparison (*p*-value)**
% of mothers of children 0-11 months who were counseled for 'prevention of mother-to-child transmission' services during last pregnancy.	Mothers ($n = 1446$): 0.913 (0.898, 0.927)	Mothers ($n = 205$): 0.785 (0.701, 0.868)	0.003

Table A.2 Comparison of *Prevention of Mother-to-Child Transmission (PMTCT) of HIV* indicators.

Indicator	**LQAS estimate (95% CI)**	**DHS estimate (95% CI)**	**Comparison (*p*-value)**
% of individuals who had sex with more than one sexual partner in the last 12 months.	Women ($n = 1445$): 0.029 (0.021, 0.039)	Women ($n = 1097$): 0.005* (0.001, 0.009)*	0.496
	Female youth ($n = 781$): 0.028 (0.019, 0.042)	Female youth ($n = 451$): 0.002 (0, 0.006)	< 0.001
	Men ($n = 1446$): 0.111 (0.096, 0.128)	Men ($n = 291$): 0.155 (0.106, 0.204)	0.095
	Male youth ($n = 633$): 0.072 (0.056, 0.092)	Male youth ($n = 116$): 0.033 (0, 0.067)	0.058
% of individuals who have had sexual intercourse with a non-marital or non-cohabitating sexual partner.	Women ($n = 1445$): 0.086 (0.072, 0.101)	Women ($n = 1097$): 0.010 (0.004, 0.016)	< 0.001
	Female youth ($n = 781$): 0.066 (0.051, 0.086)	Female youth ($n = 451$): 0.005 (0, 0.011)	< 0.001
	Men ($n = 1446$): 0.198 (0.178, 0.219)	Men ($n = 291$): 0.069 (0.035, 0.102)	< 0.001
	Male youth ($n = 633$): 0.111 (0.096, 0.146)	Male youth ($n = 116$): 0.016 (0, 0.040)	< 0.001
% of individuals who have had sexual intercourse with a non-marital or non-cohabitating sexual partner in the last 12 months and used a condom at last higher-risk sex.	Women ($n = 128$): 0.317 (0.243, 0.402)	Women ($n = 11$): 0.308 (0.012, 0.605)	0.957
	Female youth ($n = 54$): 0.505 (0.376, 0.634)	Female youth ($n = 2$): 0.551 (0, 1)	0.894
	Men ($n = 186$): 0.426 (0.357, 0.498)	Men ($n = 20$): 0.308 (0.088, 0.527)	0.372
	Male youth ($n = 82$): 0.462 (0.358, 0.569)	Male youth ($n = 2$): 0	n/a
% of youth 15-24 years who have had sexual intercourse before the age of 15.	Female youth ($n = 781$): 0.045 (0.032, 0.061)	Female youth ($n = 451$): 0.054 (0.028, 0.080)	0.532
	Male youth ($n = 633$): 0.076 (0.058, 0.099)	Male youth ($n = 116$): 0.062 (0.001, 0.115)	0.641
% of men who are circumcised.	Men ($n = 1446$): 0.102 (0.087, 0.119)	Men ($n = 291$): 0.088 (0.044, 0.132)	0.561
	Male youth ($n = 633$): 0.072 (0.055, 0.095)	Male youth ($n = 116$): 0.099 (0.022, 0.176)	0.501

Table A.3 Comparison of *HIV Knowledge and Sexual Behavior* indicators.

Indicator	**LQAS estimate (95% CI)**	**DHS estimate (95% CI)**	**Comparison (*p*-value)**
% of children 0-59 months who had fever in the two weeks preceding the survey and received treatment with ACT within 24 h of onset of fever.	0-11 months ($n = 1353$): 0.044 (0.031, 0.064) 12-23 months ($n = 752$): 0.090 (0.071, 0.113) Weighted average for comparison: 0.070 (0.056, 0.084)	0-23 months ($n = 49$): 0.068 (0, 0.138)	0.961
% of mothers of children 0-11 months who received two of more doses of SP/Fansidar during their last pregnancy.	Mothers ($n = 1446$): 0.649 (0.635, 0.684)	Mothers ($n = 205$): 0.267 (0.191, 0.343)	$<$ 0.001
% of children 0-59 months who slept under an insecticide-treated net the night preceding the survey.	0-11 months ($n = 1446$): 0.658 (0.633, 0.682) 12-23 months ($n = 1446$): 0.657 (0.632, 0.681) Weighted average for comparison: 0.657 (0.639, 0.675)	0-23 months ($n = 412$): 0.413 (0.346, 0.481)	$<$ 0.001

Table A.4 Comparison of *Malaria* indicators. Weighted averages were calculated based on the proportion of the aggregated sample that belonged to a particular group.

Indicator	LQAS estimate (95% CI)	DHS estimate (95% CI)	Comparison (p-value)
% of currently married women who are using any family planning method.	Mothers ($n = 1158$): 0.487 (0.458, 0.516)	Mothers ($n = 681$): 0.296* (0.242, 0.350)*	< 0.001
% of mothers of children 0-11 months who attended ANC at least four times during their last pregnancy.	Mothers ($n = 1446$): 0.466 (0.441, 0.492)	Mothers ($n = 205$): 0.487 (0.392, 0.581)	0.677
% of mothers of children 0-11 months who delivered their last baby in a health facility.	Mothers ($n = 1446$): 0.668 (0.644, 0.692)	Mothers ($n = 205$): 0.544 (0.433, 0.655)	0.034
% of mothers of children 0-11 months who were assisted by a skilled health worker during their last delivery.	Mothers ($n = 1446$): 0.645 (0.618, 0.672)	Mothers ($n = 205$): 0.562 (0.448, 0.677)	0.161

Table A.5 Comparison of *Family Planning & Reproductive Health* indicators.

Indicator	LQAS estimate (95% CI)	DHS estimate (95% CI)	Comparison (p-value)
% of children 12-23 months who are fully vaccinated, according to Definition 1 (1 BCG + 3 DPT + 4 POLIO + MEASLES).	12-23 months ($n = 1446$): 0.286 (0.264, 0.310)	12-23 months ($n = 171$): 0.271 (0.190, 0.353)	0.729
% of children 12-23 months who are fully vaccinated, according to Definition 2 (1 BCG + 3 DPT + 3 POLIO + MEASLES).	12-23 months ($n = 1446$): 0.620 (0.595, 0.645)	12-23 months ($n = 171$): 0.616* (0.514, 0.717)*	0.940
% of children 0-11 months with diarrhea in the last two weeks receiving oral rehydration therapy (ORT).	0-11 months ($n = 393$): 0.176 (0.141, 0.216)	12-23 months ($n = 46$): 0.231 (0.091, 0.371)	0.454

Table A.6 Comparison of *Child Health* indicators.

Indicator	**LQAS estimate (95% CI)**	**DHS estimate (95% CI)**	**Comparison (*p*-value)**
% of children under six months of age who are exclusively breastfed.	0-5 months ($n = 783$): 0.540 (0.503, 0.576)	0-5 months ($n = 110$): 0.531 (0.412, 0.650)	0.887
% of children 12-23 months receiving vitamin A supplementation in the last six months.	12-23 months ($n = 1446$): 0.656 (0.631, 0.680)	12-23 months ($n = 171$): 0.545 (0.456, 0.635)	0.020
% of households using iodized salt.	Households with mothers of children 12-23 months ($n = 1372$): 0.922 (0.907, 0.935)	Households$^{(1)}$ ($n = 1049$): 0.984 (0.975, 0.993) $^{(1)}$out of houses that had salt that was tested (denominator includes only houses that had salt that was tested) - DHS uses this.	$^{(1)}$ < 0.001
		Households$^{(2)}$ ($n = 1128$): 0.915 (0.894, 0.937) $^{(2)}$out of all non-missing values (denominator includes houses with no salt, and with untested salt).	$^{(2)}$ 0.609
		Children$^{(3)}$ ($n = 171$): 0.959 (0.929, 0.989) $^{(3)}$out of all children 12-23 months.	$^{(3)}$ 0.027
% of mothers of children 0-11 months who received vitamin A supplementation within 2 months after delivery.	Mothers ($n = 1446$): 0.507 (0.482, 0.533)	Mothers ($n = 205$): 0.294 (0.203, 0.385)	< 0.001

Table A.7 Comparison of *Nutrition* indicators.

Indicator	LQAS estimate (95% CI)	DHS estimate (95% CI)	Comparison (*p*-value)
% of households with safe water supply.	Households ($n = 1445$): 0.634 (0.609, 0.658)	Households$^{(1)}$ ($n = 1128$): 0.311 (0.252, 0.370) $^{(1)}$LQAS safe water is piped, protected well, borehole, rainwater, tanker/truck, bottled water.	$^{(1)} < 0.001$
		Households$^{(2)}$ ($n = 1128$): 0.430 (0.350, 0.510) $^{(2)}$DHS also includes public tap/standpipe and protected spring, which is not in the LQAS questionnaire (may be classified differently within LQAS).	$^{(2)}$ 0.005
% of households with latrine or toilet.	Households ($n = 1445$): 0.970 (0.959, 0.977)	Households ($n = 1128$): 0.978 (0.964, 0.992)	0.381

Table A.8 Comparison of *Water and Sanitation* indicators.

	Indicator	Subpopulation	Result at 5%	Absolute difference
Table A.1: HIV Counseling and Testing	% of individuals who were counseled and received an HIV test in last 12 months and know their results.	Women	Different	0.052
		Female youth	Same	0.003
		Men	Different	0.077
		Male youth	Same	0.043
	% of mothers of children 0-11 months who were counseled and received an HIV test during the last pregnancy and know the results.		Same	0.050
Table A.2: Prevention of Mother-to-Child Transmission (PMTCT) of HIV	% of mothers of children 0-11 months who were counseled for 'prevention of mother-to-child transmission' services during last pregnancy.		Different	0.128
Table A.3: HIV Knowledge and Sexual Behavior	% of individuals who had sex with more than one sexual partner in the last 12 months.	Women	Same	0.024
		Female youth	Same	0.026
		Men	Same	0.044
		Male youth	Same	0.039
	% of individuals who have had sexual intercourse with a nonmarital or noncohabitating sexual partner.	Women	Different	0.076
		Female youth	Different	0.061
		Men	Different	0.129
		Male youth	Different	0.095
	% of individuals who have had sexual intercourse with a nonmarital or noncohabitating sexual partner in the last 12 months and used a condom at last higher-risk sex.	Women	Same	0.009
		Female youth	Same	0.046
		Men	Same	0.118
		Male youth	n/a	–
	% of youth 15-24 years who have had sexual intercourse before the age of 15.	Female youth	Same	0.014
		Male youth	Same	0.014
	% of men who are circumcised.	Men	Same	0.027
		Male youth	Same	0.002

Table A.9 Summary of comparisons, using indicators from Tables A.1, A.2, and A.3. Two indicators are concluded to be the 'same' if the hypothesis test of proportion equality failed to reject at the 5% level. Otherwise, the indicators are concluded to be 'different'. Refer to the indicated table for more detailed information on that indicator, including point estimates, confidence intervals, and *p*-values.

	Indicator	Result at 5%	Absolute difference
Table A.4: Malaria	% of children 0-23 months who had fever in the two weeks preceding the survey and received treatment with ACT within 24 h of onset of fever.	Same	0.002
	% of mothers of children 0-11 months who received two of more doses of SP/Fansidar during their last pregnancy.	Different	0.382
	% of children 0-23 months who slept under an insecticide-treated net the night preceding the survey.	Different	0.244
Table A.5: Family Planning & Reproductive Health	% of currently married women who are using any family planning method.	Different	0.191
	% of mothers of children 0-11 months who attended ANC at least four times during their last pregnancy.	Same	0.021
	% of mothers of children 0-11 months who delivered their last baby in a health facility.	Different	0.124
	% of mothers of children 0-11 months who were assisted by a skilled health worker during their last delivery.	Same	0.083
Table A.6: Child Health	% of children 12-23 months who are fully vaccinated, according to Definition 1 (1 BCG + 3 DPT + 4 POLIO + MEASLES).	Same	0.015
	% of children 12-23 months who are fully vaccinated, according to Definition 2 (1 BCG + 3 DPT + 3 POLIO + MEASLES).	Same	0.004
	% of children 0-11 months with diarrhea in the last two weeks receiving oral rehydration therapy (ORT).	Same	0.055
Table A.7: Nutrition	% of children under six months of age who are exclusively breastfed.	Same	0.009
	% of children 12-23 months receiving vitamin A supplementation in the last six months.	Different	0.111
	% of households using iodized salt.	Different	0.037
	% of mothers of children 0-11 months who received vitamin A supplementation within 2 months after delivery.	Different	0.213
Table A.8: Water and Sanitation	% of households with safe water supply.	Different	0.204
	% of households with latrine or toilet.	Same	0.008

Table A.10 Summary of comparisons, using indicators from Tables A.4, A.5, A.6, A.7, and A.8. Two indicators are concluded to be the 'same' if the hypothesis test of proportion equality failed to reject at the 5% level. Otherwise, the indicators are concluded to be 'different'. Refer to the indicated table for more detailed information on that indicator, including point estimates, confidence intervals, and *p*-values.

B

Supplementary material to accompany Chapter 2

B.1 Definition of terms from the regression tree literature

Below is a brief summary of terms used in Chapter 2, as well as the regression tree literature.

A *classifier* or *classification rule* is “a systematic way of predicting what class a case is in” [15]. In the most general sense, identifying TEH amounts to classifying observations according to how the treatment affects their outcome. A *tree* is one type of classifier, a sequence of binary covariate-based decision rules with its *depth* equal to the maximum number of decisions that have to be made to classify an observation. The tree represents a partitioning of the covariate space into terminal nodes or “leaves”, where within each leaf, observations are of the same class (i.e., a classification tree) or the predicted outcome is constant (i.e., a regression tree). Trees are summed into a new larger tree by adding an observation’s predictions from the summand trees. *Boosting* is the summing of many low-depth trees (i.e., “weak learners”) into a larger tree, a method known to improve predictive performance [32]. *Bootstrap aggregation (bagging)* is the averaging of many full-sized trees, as grown from bootstrap samples [32, Chapter 8.7]. To prevent overfitting (i.e., an overly-fine partitioning of the covariate space that is particular to the sample used to build the tree), an oversized tree is “grown” (constructed) then “pruned” (modified by removing “branches”, subtrees that do not contain the root node). For specific details on how to grow and prune a tree, the reader is directed to Breiman et al. [15].

B.2 Table from Simulation Study 2.4.1

	Scenario A confounding and no effect modification	**Scenario B** effect modification and no confounding	**Scenario C** effect modification and confounding	**Scenario D** effect modification and confounding by EMs
TE(1)	100	100	100	100
TE(2)	100	100	100	100
TE(3)	100	100	100	100
TE(4)	100	100	100	100
TE(5)	100	100	100	100
TE(6)	100	100	100	100
TE(7)	100	100	100	100
TE(8)	100	100	100	99
TE(9)	100	100	100	69
TE(10)	98	100	99	8
undefined TE	2	0	1	92

Table B.1 Summary of denominators used to calculate the cell-specific averages visualized in Figure 2.1 (on page 32) as part of Simulation Study 2.4.1; scenarios are defined in Table 2.2 (on page 31). "TE(1)" denotes the estimated subgroup with the smallest ATE. Note that GBM was the only estimation procedure that yielded subgroups with an undefined treatment effect; the cell-specific averages presented for BART and FS were across all 100 simulation iterations.

B.3 Figure from Simulation Study 2.4.1: True Treatment Effects

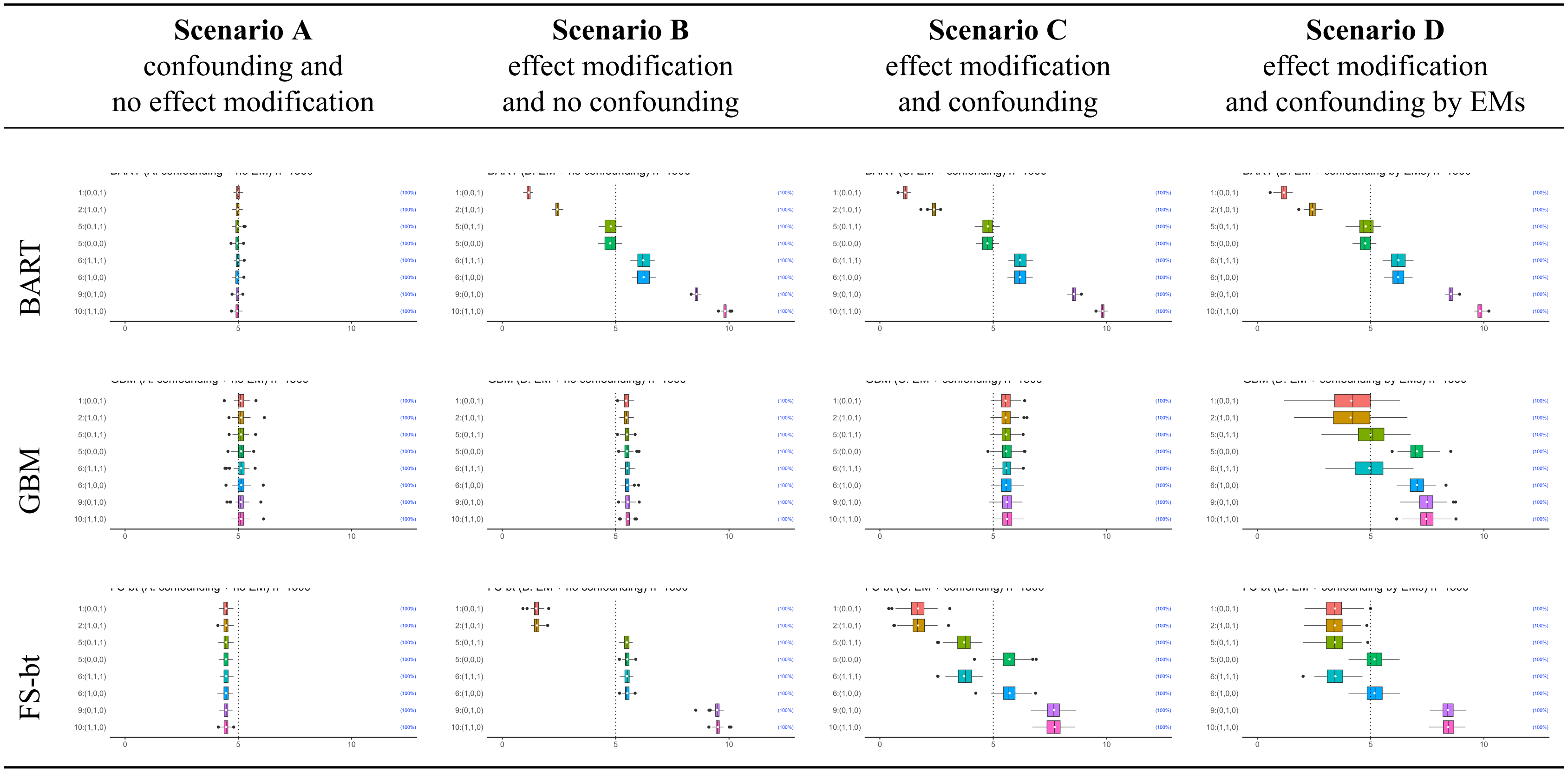


Figure B.1 Forest plots by true treatment effect, from Simulation Study 2.4.1. The structure of this grid is pattered after Figure 2.1, where each row is an estimation method, and each column is a data generation scenario. To ease explanation, consider the forest plot associated with the BART analysis of data generated under Scenario A. Letting $j = 1, \ldots, 100$ denote the simulation iteration, each observation during the jth iteration is assigned to a subgroup, within which an average treatment effect is estimated. Every observation in true Group 1 (say) has an associated ATE – the ATE estimated from the subgroup that the observation was assigned to. We can then take everyone in true Group 1, and take an average of these ATEs; in fact, we can do this for all eight true Groups, then plot the resulting eight averages. To generate the forest plots in this figure, these eight special averages were plotted, but for all 100 simulation iterations. For each true subgroup, a boxplot is used to help visualize the distribution of estimates in that group. For clarity, the x-axes of the forest plots have been omitted, but there are vertical dashed lines denoting the true average treatment effects $(1, 2, 5, 6, 9, 10)$.

B.4 Simulation Study: Simulated Treatment and Outcome

In this study, we consider the comparison of drug-eluting stents (DES) to bare-metal stents (BMS) as treatment of myocardial infarction (MI), by looking at the association of each with the two-year revascularization rate. We use real covariate data to simulate these two treatment options, as well as the two-year revascularization rate, to begin exploring the ability of these methods to identify TEH in real data.

B.4.1 Data Structure and Analysis

De-identified inpatient data on 38 covariates were generated by 169,539 Medicare beneficiaries hospitalized in the continental United States during 2009, 2010, or 2011 with their first MI. While the covariate summary given in Table 2.4 (on page 40) is of hospitalizations in 2008, the 2009-2011 covariate distribution is very similar. As treatment, these patients underwent percutaneous coronary intervention (PCI) for the placement of exactly one type of stent, either a DES or BMS. Let $\boldsymbol{x}$ denote the covariate vector of a patient and $T \sim \text{Bern}(p_{\ell_2})$ their binary treatment indicator, where $p_{\ell_2} = \text{expit}(\boldsymbol{x}^\top \widehat{\boldsymbol{\alpha}})$ is the probability of receiving a DES. The value of coefficient $\widehat{\boldsymbol{\alpha}}$ was set as the maximum likelihood estimate from the regression of observed treatment on the 38 covariates. Binary outcome $Y \sim \text{Bern}(\mu_{\ell_2})$ indicates that the patient had been readmitted for revascularization (via CABG or a second PCI) within two years of discharge from their original MI hospitalization, or died before they could experience the revascularization event. It is modeled by setting $\mu_{\ell_2} = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} - 0.308T + 0.6T(elig) - T(diabetes)$, where $(\widehat{\boldsymbol{\beta}}, -0.308)$ is the maximum likelihood estimate from the regression of the observed outcome on $(\boldsymbol{x}, t)$, and $(0.6, -1)$ the fixed interaction coefficients for effect modifiers $elig$, an indicator of Medicaid eligibility, and $diabetes$, an indicator of prior diabetes diagnosis. The main effects of these covariates are contained in $\widehat{\boldsymbol{\beta}}$. These two covariates define four subgroups, with ATEs $(-0.22, -0.08, 0.02, 0.16)$ measured on the risk difference scale.

A dataset was simulated by first sampling the observed covariate data of 10,000 Medicare beneficiaries from the 169,539, then generating T and Y from the distributions described above. To preclude effect estimation issues caused by empirical positivity violations, any covariate with a prevalence of less than 5% in either treatment arm was dropped. To preclude obfuscation of our argument within this artificial simulation scenario, any dataset that did not include either effect modifier (i.e., after having been dropped for low prevalence) was discarded. This process was repeated to generate 100 simulated datasets, and each

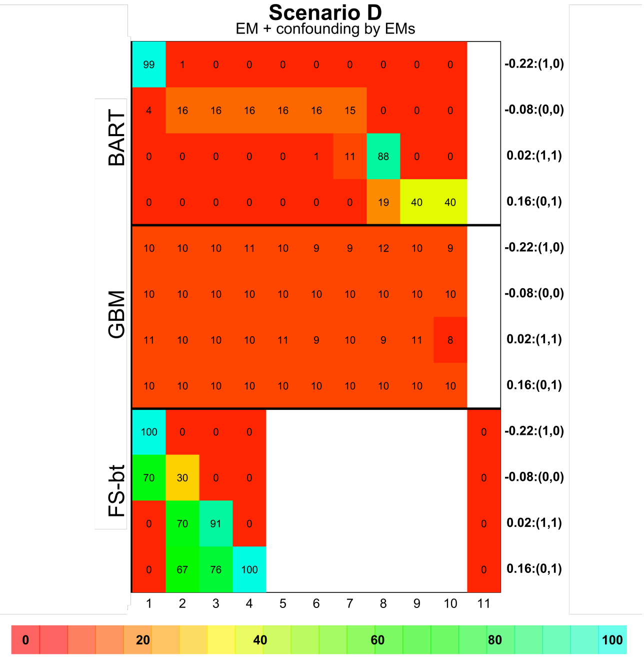


Figure B.2 Visualization of results from Simulation Study B.4. Details regarding how the figure was constructed, and how to interpret the figure, are given in §B.4.1.

analyzed as described in §4.1.1.

Results and Discussion Analysis results are summarized as described in §4.1.2 and visualized in Figure B.2 of these Supplementary Materials (page 71, for one data generation scenario). Looking at the results generated by FS, we draw conclusions similar to those drawn in Simulation Study 4.1; namely, FS is able to detect effect modifiers that are strongly associated with either the outcome or treatment, where strength is relative to the associations of the other covariates to the treatment and outcome. In this example, prior diabetes diagnosis has a strong association with the outcome, and we see that FS is able to group observations by the value of this covariate: the top two rows have a prior diabetes diagnosis, and the bottom two rows do not.

An interesting point is the spread of color in the last two rows of the FS result block, which is caused by the particulars of PAM's estimation procedure. While we were able to prespecify that 10 subgroups be estimated, and the "center" of each of these subgroups is an observation from the dataset (by design), if no other observations are close to a selected center, then that center will remain in an estimated subgroup of size one. As applied to Figure B.2 of these Supplementary Materials (on page 71), PAM is typically able

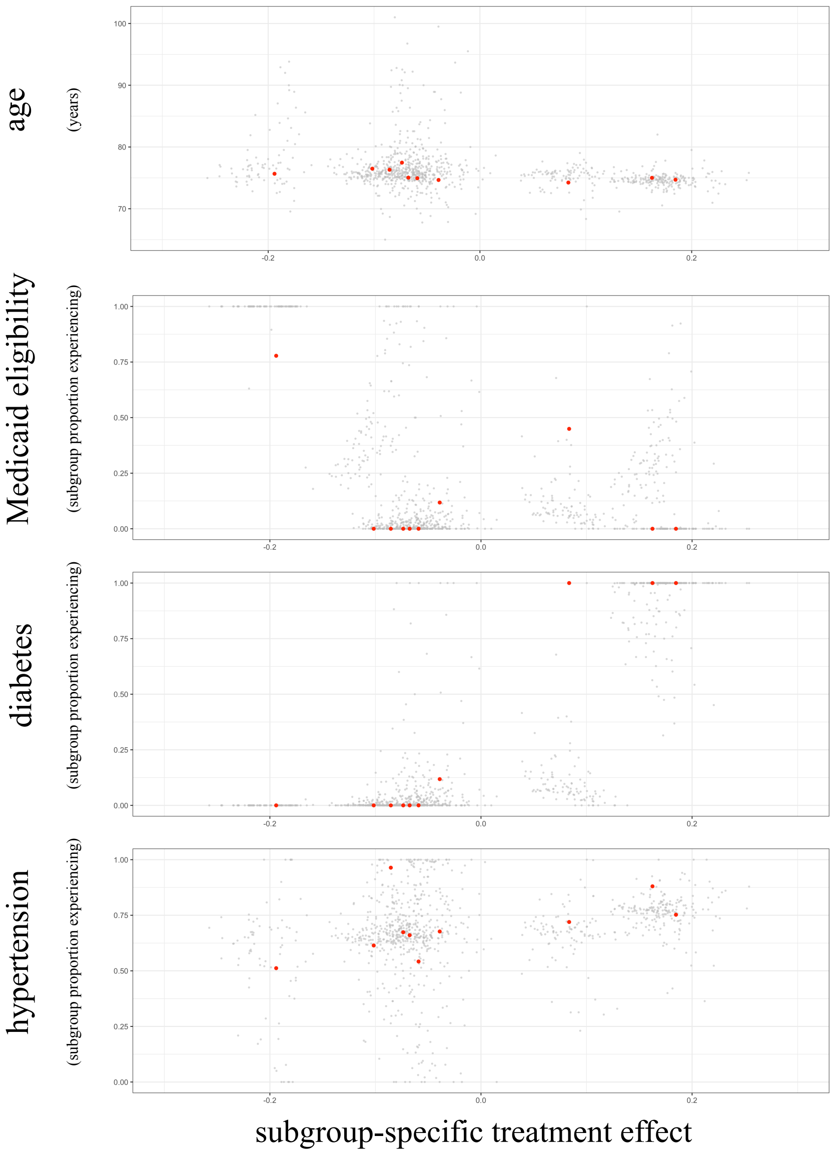


Figure B.3 Visualization of results from a single dataset from Simulation Study B.4. Covariates displayed are *age*, *Medicaid eligibility*, prior *diabetes* diagnosis, and prior *hypertension* diagnosis. Details regarding how the figure was constructed, and how to interpret the figure, are given in §B.4.1.

to detect 2-4 substantive subgroups. The remaining subgroups have just one observation, so have an undefined treatment effect; thus, empty columns 5 though 10 imply that PAM always had at least six estimated subgroups with one patient in each. The spread of color is actually an average across simulation iterations that detected differing numbers of subgroups.

Considering GBM, the results do not show evidence of being able to detect either EM, as demonstrated by the relative homogeneity of color and row percentages.

Finally considering the analysis using BART, this analysis procedure generated results that are as we would expect from correct identification of TEH. For example, the first row of this block represents observations in the subgroup with the smallest ATE, that are eligible for Medicaid and do not have a previous diabetes diagnosis, with a membership of $10,000 \times 0.11 \times 0.73 = 803$. Because a decile is 1000 observations, we expect 100% of observations to be column 1, and this is what we see in Figure B.2 (page 71; slightly different due to sampling variability). The remaining four rows also contain percentages as we expect.

A natural follow-up is to ask whether we can group individuals using their entire ITE posterior distribution, rather than a point summary, and Figure B.3 of these Supplementary Materials (on page 72) attempts to answer this. Here we visualize one of the 100 datasets summarized in Figure B.2 (page 71), to move our argument towards what we would expect in a real data analysis. To ease explanation, consider the top-most plot marked *age*. The y-axis represents the subgroup-specific average age in years, and the x-axis represents the subgroup-specific ATE (again measured on the risk-difference scale). A red dot represents a subgroup, generated as described earlier: the $10,000$ posterior means are partitioned into deciles, and the average age and average TE (taken as the average of the posterior mean ITEs within that decile) are plotted. Thus, we expect ten red dots in this single plot. These red dots are clustered into four groups because, by design, there are four true subgroups. The number of red dots in these clusters is, as with the columns Figure B.2 (page 71), proportional to the size of the true subgroup (e.g., the largest true subgroup has an ATE of -0.08).

The posterior means of Figure B.3 are somewhat redundant with Figure B.2; it is the gray dots that provide the additional information on the full distribution of each observation's ITE. To generate the values that these gray dots represent, we applied the subgrouping process used on the posterior means (i.e., partitioned into ten subgroups and calculated subgroup-specific averages) to each of the 1000 posterior draws, and plotted a random subset of 100. We note that the same cutpoints were used to partition each posterior draw; namely, the deciles that were used to partition the posterior means. Thus from this figure we are able to visualize the posterior means in red, in addition to some measure of uncertainty in gray.

By design, *age* is not an EM, and does not present itself as such; its distribution remains approximately constant as the subgroup-specific treatment effect increases. However *diabetes* is an EM, where its presence is associated with a larger treatment effect, and we see this manifest as an upward trend within its plot. We see analogous patterning in the plot for *Medicaid eligibility*. Interestingly, *hypertension* was not *a priori* specified as an EM, but due to its positive correlation with diabetes (a known clinical phenomenon), displays itself as one: the subgroup-specific ATE increases as the prevalence of *hypertension* increases.

References

[1] Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P., and for the CONSORT group, T. L. (2001). The revised consort statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*, 134(8):663–694.

[2] Arbogast, P. G. and Ray, W. A. (2008). Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research*, 18:67–80.

[3] Arbogast, P. G. and Ray, W. A. (2011). Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology*, 174(5):613–620.

[4] Armstrong, E. J., Waltenberger, J., and Rogers, J. H. (2014). Percutaneous coronary intervention in patients with diabetes: Current concepts and future directions. *Journal of Diabetes Science and Technology*, 8(3):581–589.

[5] Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355:1064–1069.

[6] Athey, S. and Imbens, G. W. (2015). Recursive partitioning for heterogeneous causal effects. *arXiv*, 1504(01132v3).

[7] Barrere, B., Fishel, J., McInturff, S., Pullum, T., Reinis, K., Rutstein, S., and Themme, A. (2014). The Demographic and Health Surveys (DHS) program.

[8] Beckworth, C. A., Davis, R. H., Faragher, B., and Valadez, J. J. (2015). Can health workers reliably assess their own work? a test-retest study of bias among data collectors conducting a lot quality assurance sampling survey in uganda. *Health Policy Plan*, 30(2):181–186.

[9] Berry, C., Tardif, J.-C., and Bourassa, M. G. (2007). Coronary heart disease in patients with diabetes Part ii: Recent advances in coronary revascularization. *Journal of the American College of Cardiology*, 49(6):643–656.

[10] Bertin, J. (1983). *Semiology of Graphics: Diagrams Networks Maps (English translation)*. University of Wisconsin Press, Madison, Wisconsin.

[11] Bhuiya, A., Hanifi, S., Roy, N., and Streatfield, P. K. (2007). Performance of the lot quality assurance sampling method compared to surveillance for identifying inadequately-performing areas in Matlab, Bangladesh. *Journal of Health, Population and Nutrition*, 25(1):37–46.

[12] Biedron, C., Pagano, M., Hedt, B. L., Kilian, A., Ratcliffe, A., Mabunda, S., and Valadez, J. J. (2010). An assessment of lot quality assurance sampling to evaluate malaria outcome indicators: Extending malaria indicator surveys. *International Journal of Epidemiology*, 39:72–79.

[13] Boerma, J. T. and Stansfield, S. K. (2007). Health statistics now: Are we making the right investments? *The Lancet*, 369(9563):779–786.

[14] Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215.

[15] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

[16] Chan, M., Kazatchkine, M., Lob-Levyt, J., Obaid, T., Schweizer, J., Sidibe, M., Veneman, A., and Yamada, T. (2010). Meeting the demand for results and accountability: A call for action on health data from eight global health agencies. *PLoS Med*, 7(1):e1000223.

[17] Chan, P.-H., Liu, S.-S., Tse, H.-F., Chow, W.-H., Jim, M.-H., HO, H.-H., and Siu, C. W. (2013). Long-term clinical outcomes of drug-eluting stents versus bare-metal stents in Chinese geriatric patients. *Journal of Geriatric Cardiology*, 10:330–335.

[18] Chipman, H. A., George, E. I., and McCulloch, R. E. (2007). *Advances in Neural Information Processing Systems*, chapter Bayesian ensemble learning. 19. MIT Press, Cambridge, MA.

[19] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298.

[20] Cleveland, W. S. (1985). *The Elements of Graphing*. Bell Telephone Laboratories, Murray Hill, New Jersey.

[21] Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404.

[22] Dean, N. E. and Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3(4):484–503.

[23] Duhigg, C. (2016). *Smarter Faster Better: The Secrets of Being Productive in Life and Business*. Random House, New York, NY.

[24] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

[25] Farmer, A. and Fox, R. (2011). Diagnosis, classification, and treatment of diabetes: Age of onset and body mass index are no longer a basis for classifying the cause. *BMJ*, 343(7824):597–598.

[26] Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–2880.

[27] Ghosh, D. (2011). Propensity score modelling in observational studies using dimension reduction methods. *Statistics & Probability Letters*, 81(7):813–820.

[28] Grimmer, J., Messing, S., and Westwood, S. J. (2014). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Unpublished manuscript*.

[29] Grolemund, G. and WIchkam, H. (2015). Visualizing complex data with embedded plots. *Journal of Computational and Graphical Statistics*, 24(1):26–43.

[30] Grundmann, C. (2002). The costs of using lqas for project management, monitoring and evaluation. *NGO Networks for Health Project*.

[31] Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.

[32] Hastie, T., Tibshirani, R., , and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2nd edition.

[33] Henderson, R. H. and Sundaresan, T. (1982). Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, 60(2):253–260.

[34] Hernán, M. A. and Robins, J. M. ((in progress)). *Causal Inference*. Chapman & Hall/CRC.

[35] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[36] Ho, D. E., Imai, K., King, G., and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236.

[37] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

[38] Jain, K. K. (2006). *Textbook of Personalized Medicine*. Springer, New York, NY.

[39] Kooperberg, C., LeBlanc, M., Dai, J. Y., and Rajapakse, I. (2009). Structures and assumptions: Strategies to harness gene×gene and gene×environment interactions in GWAS. *Statistical science: A review journal of the Institute of Mathematical Statistics*, 24(4):472–478.

[40] Kornowski, R., Mintz, G. S., Kent, K. M., Pichard, A. D., Satler, L. F., Bucher, T. A., Hong, M. K., Popma, J. J., and Leon, M. B. (1997). Increased restenosis in diabetes mellitus after coronary interventions is due to exaggerated intimal hyperplasia: A serial intravascular ultrasound study. *Circulation*, 95:1366–1369.

[41] Kurz, D. J., Bernheim, A. M., Tüller, D., Zbinden, R., Jeger, R., Kaiser, C., Galatius, S., Hansen, K. W., Alber, H., Pfisterer, M., and Eberli, F. R. (2015). Improved outcomes of elderly patients treated with drug-eluting versus bare metal stents in large coronary arteries: Results from the BAsel Stent Kosten-Effektivitäts Trial PROspective Validation Examination randomized trial. *American Heart Journal*, 170(4):787–795.

[42] Lagakos, S. W. (2006). The challenge of subgroup analyses – reporting without distorting. *The New England Journal of Medicine*, 354:1667–1669.

[43] Lago, R. M., Singh, P. P., and Nesto, R. W. (2007). Diabetes and hypertension. *Nature Clinical Practice Endocrinology & Metabolism*, 3(10):667.

[44] Malenka, D. J., Kaplan, A. V., Lucas, F. L., Sharp, S. M., and Skinner, J. S. (2008). Outcomes following coronary stenting in the era of bare-metal vs the era of drug-eluting stents. *Journal of the American Medical Association*, 299(24):2868–2876.

[45] McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.

[46] Microsoft Machine Learning & Data Science Summit (2016). Keynote Session: Dr. Edward Tufte - The Future of Data Analysis (video). `https://channel9.msdn.com/Events/Machine-Learning-and-Data-Sciences-Conference/Data-Science-Summit-2016/MSDSS11`.

[47] Miettinen, O. S. (1976). Stratification by a multivariate confounder score. *American Journal of Epidemiology*, 104(6):609–620.

[48] Murray, C. J. L. and Frenk, J. (2008). Health metrics and evaluation: Strengthening the science. *The Lancet*, 371(9619):1191–1199.

[49] Nelson, D. and Noorbaloochi, S. (2013). Information preserving sufficient summaries for dimension reduction. *Journal of Multivariate Analysis*, 115:347–358.

[50] Pagano, M. and Valadez, J. J. (2010). Commentary: Understanding practical lot quality assurance sampling. *Journal of Epidemiology*, 39:69–71.

[51] Pocock, S. J., Collier, T. J., Dandreo, K. J., de Stavola, B. L., Goldman, M. B., Kalish, L. A., Kasten, L. E., and McCormack, V. A. (2004). Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ*, page doi:10.1136/bmj.38250.571088.55.

[52] Pocock, S. J., Hughes, M. D., and Lee, R. J. (1987). Statistical problems in the reporting of clinical trials. *The New England Journal of Medicine*, 317:426–432.

[53] Puymirata, E., Mangiacapraa, F., Peacea, A., Ntarladimasa, Y., Contea, M., Bartuneka, J., Vanderheydena, M., Wijnsa, W., de Bruynea, B., and Barbatoa, E. (2013). Safety and effectiveness of drug-eluting stents versus bare-metal stents in elderly patients with small coronary vessel disease. *Archives of Cardiovascular Disease*, 106:554–561.

[54] Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147.

[55] Rommelmann, V., Setel, P. W., Hemed, Y., Angeles, G., Mponezya, H., Whiting, D., and Boerma, T. (2005). Cost and results of information systems for health and poverty indicators in the United Republic of Tanzania. *Bulletin of the World Health Organization*, 83(8):569–577.

[56] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[57] Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford University Press, 2nd edition.

[58] Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, 365:176–186.

[59] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

[60] Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

[61] Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

[62] Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511.

[63] Singh, J., Jain, D., Sharma, R., and Verghese, T. (1996). Evaluation of immunization coverage by lot quality assurance sampling compared with 30-cluster sampling in a primary health centre in india. *Bulletin of the World Health Organization*, 74(3):269–274.

[64] StataCorp (2013a). *Stata 13 Base Reference Manual*. StataCorp LP, College Station, TX.

[65] StataCorp (2013b). *Stata: Release 13. Statistical Software.* StataCorp LP, College Station, TX.

[66] Su, X., Kang, J., Fan, J., Levine, R. A., and Yan, X. (2012). Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*, 13:2955–2994.

[67] Tufte, E. R. (1974). *Data Analysis for Politics and Policy*. Prentice-Hall, Inc.

[68] Tufte, E. R. (1990). *Envisioning Information*. Graphics Press LLC, Cheshire, Connecticut.

[69] Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press LLC, Cheshire, Connecticut.

[70] Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press LLC, Cheshire, Connecticut, 2nd edition.

[71] Tufte, E. R. (2006). *Beautiful Evidence*. Graphics Press LLC, Cheshire, Connecticut.

[72] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company Inc, Reading, Massachusetts.

[73] Uganda Bureau of Statistics, I. I. I. (2012). *Uganda Demographic and Health Survey 2011*. Uganda Bureau of Statistics and ICF International Inc., Kampala, Uganda and Calverton, Maryland.

[74] Valadez, J. J. (1991). *Assessing Child Survival Programs in Developing Countries: Testing Lot Quality Assurance Sampling*. Harvard University Press, Cambridge.

[75] VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871.

[76] VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York, NY.

[77] Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv:1510.04342*.

[78] Wang, C., Parmigiani, G., and Dominici, F. (2012a). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68:661–686.

[79] Wang, H., Lo, S.-H., Zheng, T., and Hu, I. (2012b). Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics*, 28(21):2834–2842.

[80] Westreich, D. and Cole, S. R. (2010). Invited commentary: Positivity in practice. *American Journal of Epidemiology*, 171(6):674–677.

[81] Woo, M.-J., Reiter, J. P., and Karr, A. F. (2007). Estimation of propensity scores using generalized additive models. Technical Report 167, National Institute of Statistical Sciences.

[82] Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9):1167–1173.

[83] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.