Statistical Developments for Network Meta-Analysis and Methane Emissions Quantification

by

Augustine Wigle

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2025

© Augustine Wigle 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Thomas Lumley Chair, Biostatistics, Dept. of Statistics, University of Auckland
Supervisor:	Audrey Béliveau Associate Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
Internal Members:	Christian Boudreau Research Associate Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
	Paul Marriott Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
	Zelalem Negeri Assistant Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
Internal-External Member:	Ashok Chaurasia Associate Professor, School of Public Health Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Augustine Wigle is the sole author of Chapters 1 and 6 which were not written for publication.

Chapter 2

Augustine Wigle is the first author of the publication associated with the research in Chapter 2, which was carried out under the supervision of Dr. Audrey Béliveau. See Wigle and Béliveau (2022).

Chapter 3

Augustine Wigle is the first author of the manuscript and the R package associated with the research in Chapter 3. The research was carried out under the supervision of Dr. Adriani Nikolakopoulou and Dr. Audrey Béliveau. Dr. Georgia Salanti, Dr. Gerta Rücker, Dr. Guido Schwarzer and Dr. Dimitris Mavridis provided intellectual input on revising the manuscript. Dr. Guido Schwarzer assisted in the writing and documentation of the R package. The manuscript is currently being revised for *Statistics in Medicine*. See Wigle et al. (2025) for the manuscript and Wigle and Schwarzer (2024) for the R package.

Chapter 4

Augustine Wigle is the first author of the manuscript and the R package associated with the research in Chapter 4. This research was conducted under the supervision of Dr. Audrey Béliveau. The manuscript is currently being revised for *Journal of the Royal Statistical Society Series A*. See Wigle and Béliveau (2024) for the manuscript and github.com/augustinewigle/methaneInventory for the R package.

Chapter 5

Augustine Wigle is the first author of the publication associated with the research in Chapter 5. This research was carried out under the supervision of Dr. Kyle J. Daun, Dr. Christiane Lemieux, and Dr. Audrey Béliveau. The survey and data collection was conducted by Augustine Wigle, Daniel Blackmore, Dr. Paule Lapeyre, Dr. Kyle J. Daun, Kevin Fritz, Jack Johnson, Kirk Osadetz, and technology providers. Augustine Wigle developed the Bayesian methods, analysed the data, and drafted the manuscript. The co-authors assisted in revising the manuscript. See Wigle et al. (2024).

Abstract

This thesis provides statistical contributions to solve challenges in Network Meta-Analysis (NMA) and the quantification of methane emissions from the oil and gas industry.

NMA is an extension of pairwise meta-analysis which facilitates the simultaneous comparison of multiple treatments using data from randomized controlled trials. Some treatments may involve combinations of components, such as one or more drugs given in different combinations. Component NMA (CNMA) is an extension of NMA which allows the estimation of the relative effects of components. In Chapter 2, we compare the popular Bayesian and frequentist approaches to additive CNMA and show that there is an important difference in the assumptions underlying these commonly used models. We prove that the most popular Bayesian CNMA model is prone to misspecification, while the frequentist approach makes a less restrictive assumption. We develop novel Bayesian CNMA models which avoid the restrictive assumption and are robust, and demonstrate in a simulation study that the proposed Bayesian models have favourable statistical properties compared to the existing Bayesian model. The use of all CNMA approaches is demonstrated on a published network.

A commonly reported item in an NMA is a list of treatments ranked from most to least preferred, also known as a treatment hierarchy. In Chapter 3, we present the Precision Of Treatment Hierarchy (POTH), a metric which quantifies the level of certainty in a treatment hierarchy from Bayesian or frequentist NMA. POTH summarises the level of certainty into a single number between 0 and 1, making it simple to interpret regardless of the number of treatments in the network. We propose modifications of POTH which can be used to investigate the role of individual treatments or subsets of treatments in the level of certainty in the hierarchy. We calculate POTH for a database of published NMAs to investigate its distribution and relationships with network characteristics. We also provide an in-depth worked example to demonstrate the methods on a real dataset.

In the second part of the thesis, we focus on some problems in the quantification of methane emissions from the oil and gas industry. Measurement-based methane inventories, which involve surveying oil and gas facilities and compiling data to estimate methane emissions, are becoming the gold standard for quantifying emissions. However, there is a current lack of statistical guidance for the design and analysis of such surveys. In Chapter 4, we propose the novel application of multi-stage survey sampling techniques to analyse measurement-based methane survey data, providing estimators of total and stratum-level emissions and an interpretable variance decomposition. We also suggest a potentially more efficient approach involving the Hájek estimator, and outline a simple Monte Carlo approach which can be combined with the multi-stage approach to incorporate measurement

error. We investigate the performance of the multi-stage estimators in a simulation study and apply the methods to aerial survey data of oil and gas facilities in British Columbia, Canada, to estimate the methane emissions in the province.

In Chapter 5, we introduce a Bayesian model for measurements from a methane quantification technology given a true emission rate. The models are fit using data collected in controlled releases (CR) of methane for six different technology types. We use a weighted bootstrap algorithm to provide the distribution of the true emission rate given a new measurement, which synthesizes the new measurement data with the CR data and external information about the possible true emission rate. We present results for the measurement uncertainty of six quantification technologies. Finally, we demonstrate the use of the weighted bootstrap algorithm with different priors and data.

Acknowledgements

First, I want to thank my supervisor, Dr. Audrey Béliveau. You pushed me to be my best and I would not be where I am today without your guidance and support. Thank you for believing in me!

I also want to thank my committee, Dr. Zelalem Negeri, Dr. Paul Marriott, Dr. Christian Boudreau, Dr. Ashok Chaurasia, and Dr. Thomas Lumley, for the time and energy you have dedicated to providing feedback on my thesis. To Dr. Kyle Daun and his lab, thank you for the opportunity to work on so many interesting projects.

There are so many wonderful people in the department who make it a great place to be a student. Thank you to Mary Lou, Carla, Greg Preston, Divya, and Jenny for your help and everything you do for SAS. Dr. Christiane Lemieux, Dr. Alex Stringer, and Dr. Yingli Qin, thank you for your support and guidance. Thank you to the Waterloo Methods in Meta-Analysis group, which has been a welcome source of community in the last 6 months of my degree. Caitlin, thank you for all the work you put into this group, not to mention the great advice and conversations.

I am so thankful for the time I spent at the University of Freiburg and the people I met there. To the whole meta-analysis team at IMBI, thank you for revitalizing my enthusiasm for research. I couldn't have asked for better colleagues and I learned so much working with you all. Adriani, thank you especially for your support and mentorship.

I am so lucky for the amazing people I get to call my friends. Thank you Meixi for being my first friend in grad school, and Agassi, Marcus, and Chris for being a source of laughs and support since 2019. To Minzee, I'm grateful for our deep chats, and I will miss being your coworker. Tatiana and Grace, I wouldn't have made it through the first years of our PhD if it weren't for our shared laughs and tears at 52 Amos. I'm so glad we got to go through this together. Thank you to Elena, Rachelle, Nina, Emily, and Angus for your friendship and helping me to enjoy life outside of research. I love you so much.

Tanner - thank you for listening to me complain about my simulations, calling me every night when I was on another continent, never letting me be too hard on myself, and the millions of other things you do that make life better. Your patience, support, and love helped me get through the final years of my PhD.

Finally, thank you to my parents and my sister Thorn for their support. Coming home for a meal when I was stuck and frustrated always helped me to decompress and gain the motivation to keep going. Thank you for encouraging me and believing in me.

Dedication

To my mother.

Table of Contents

Examining Committee ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vii

Dedication viii

List of Figures xv

List of Tables xxi

1 Introduction 1

1.1 Network Meta-Analysis 1

1.1.1 Bayesian NMA Models 3

1.1.2 Frequentist NMA Model 5

1.1.3 Multicomponent Treatments 7

1.1.4 Ranking Treatments in NMA 8

1.2 Methane Emissions Quantification 10

1.3 Outline of Contributions 11

2 Bayesian Unanchored Additive Models for Component Network Meta-Analysis **14**

2.1 Introduction . . . 14

2.2 Comparison of Existing CNMA Models . . . 15

2.2.1 Bayesian CNMA Model of Welton et al. (2009) . . . 16

2.2.2 Frequentist CNMA Model of Rücker et al. (2019) . . . 17

2.3 The Anchor in Additive CNMA . . . 20

2.3.1 Role of the Anchor Treatment . . . 20

2.3.2 Anchored and Unanchored Models . . . 21

2.3.3 The Anchor and the Placebo Effect . . . 24

2.4 Novel Bayesian Unanchored CNMA Models . . . 27

2.4.1 Novel Bayesian Unanchored Model for Contrast-level Data . . . 27

2.4.2 Novel Bayesian Unanchored Model for Arm-level Data . . . 28

2.5 Simulation Study . . . 29

2.5.1 Network Structure and Parameter Values . . . 30

2.5.2 Performance Measures . . . 31

2.5.3 Results: Network 1 . . . 36

2.5.4 Results: Network 2 . . . 37

2.6 Application . . . 39

2.6.1 Dataset and Methods . . . 39

2.6.2 Results . . . 40

2.7 Discussion and Future Work . . . 42

3 Quantifying Certainty in Treatment Hierarchies from Network Meta-Analysis **45**

3.1 Introduction . . . 45

3.2 Methodology . . . 48

3.2.1 Maximum and Minimum Certainty in a Hierarchy . . . 48

3.2.2 The Variance of SUCRA Values and Expected Ranks 49
3.2.3 The Precision of Treatment Hierarchy (POTH) 50
3.2.4 POTH and the Probability Distribution of Ranks 52
3.2.5 POTH with Groups of Equivalent Treatments 53
3.2.6 Deriving POTH for Subsets of Treatments 55
3.3 What is the Precision of Treatment Hierarchies in Published Networks? Empirical Evidence from 267 Published Networks 58
3.4 Real Data Example: Safety of Immune Checkpoint Inhibitors as Cancer Treatments 62
3.5 Discussion and Future Work 65

4 Estimating Methane Emissions from the Upstream Oil and Gas Industry Using a Multi-Stage Framework 67
4.1 Introduction 67
4.2 Aerial Survey Data from British Columbia 69
4.3 Methods 73
4.3.1 Measurement-Based Surveys As Multi-Stage Sampling 74
4.3.2 Estimating Methane Emissions Using Multi-Stage Sampling 76
4.3.3 Increasing Efficiency Using the Hájek Estimator 78
4.3.4 First and Second Stage Designs for BC Data 80
4.3.5 Estimation Accounting for Measurement Error 83
4.3.6 Simulation Methodology 85
4.4 Results 87
4.4.1 Simulation Results 87
4.4.2 Data Analysis 90
4.5 Discussion and Future Work 96

5 A Bayesian Approach for Estimating Uncertainty in Measurements from Methane Emission Quantification Technologies **100**

5.1 Introduction . . . 100

5.2 Methods . . . 102

5.2.1 A Flexible Model for the Measured Emission Rate Given the True Rate . . . 102

5.2.2 Algorithm for Quantifying Measurement Uncertainty in New Measurements . . . 106

5.3 Data . . . 107

5.4 Results . . . 108

5.4.1 Uncertainty Model Results . . . 108

5.4.2 Example of Quantifying Uncertainty in New Measurements . . . 115

5.5 Discussion and Future Work . . . 118

6 Conclusion **120**

References **123**

APPENDICES **134**

A Chapter 2 Appendix **135**

A.1 Implications of the Anchor . . . 135

A.2 Equivalence of Novel Bayesian Models . . . 136

A.3 Simulation Data Generation Mechanism . . . 137

A.4 Parameter Calculation in Simulations . . . 137

B Chapter 3 Appendix **139**

B.1 Ranking Probabilities in the Minimum-Certainty Hierarchy Scenario . . . 139

B.2 Obtaining A Simulation-Based Reference Distribution for POTH . . . 141

B.3 Proof that $V(N)$ is Uniquely Minimized Under Maximum Certainty in the Hierarchy . . . 142

B.4 Networks Producing Smallest and Largest POTH in NMA Database 142
B.5 Extra Data Examples 144
B.5.1 Antifungal Treatments to Prevent Mortality for Solid Organ Transplant Recipients 145
B.5.2 Efficacy of Pharmacological Treatments for Persistent Depressive Disorder 147
B.6 What is the Distribution of Precision of Treatment Hierarchies in Networks with No Treatment Differences? A Simulation Study 150
B.6.1 Set-up 150
B.6.2 Results 151

C Chapter 4 Appendix **154**
C.1 Note on Sample and Population Sizes 154
C.2 Treatment of Wells 155
C.3 Summary of Notation Used in Chapter 4 155
C.4 Variance Formulas for Multi-Stage Approach 157
C.5 Proof of Equivalency of Sampling Designs 159
C.6 Derivation of Variance Estimators 161
C.6.1 Variance Estimator for the IPW Approach with SRS 161
C.6.2 Variance Estimator for the Hájek Approach with SRS 162
C.7 Derivation of Total Variance of $\tilde{\tau}$ 163
C.8 Additional Data Analysis Results 165
C.8.1 Results for Multi-Stage Hájek with Stage II Uncertainty and Measurement Error 165
C.8.2 Results for Multi-stage IPW with No Stage II Uncertainty and Measurement Error 166
C.8.3 Results for Multi-stage Hájek with No Stage II Uncertainty and Measurement Error 167
C.8.4 Results for Multi-stage IPW with Stage II Uncertainty and No Measurement Error 168

C.8.5 Results for Multi-stage Hájek with Stage II Uncertainty and No Measurement Error . . . 169
C.8.6 Results for Multi-stage IPW with No Stage II Uncertainty and No Measurement Error . . . 170
C.8.7 Results for Multi-stage Hájek with No Stage II Uncertainty and No Measurement Error . . . 171
C.8.8 Results for Multi-Stage IPW with Measurement and Stage II Uncertainty for Only Southern Facilities . . . 172

D Chapter 5 Appendix **173**

D.1 Prior Distributions . . . 173
D.2 Detailed Algorithm . . . 175
D.2.1 Step 1 . . . 175
D.2.2 Step 2 . . . 176
D.2.3 Step 3 . . . 178

List of Figures

1.1 An example of a network of RCTs studying treatments for Diabetes. The thickness of the edges is proportional to the number of RCTs which directly make each comparison. The network was created with the BUGSnet package (Béliveau et al., 2019). 2

1.2 An example of a SUCRA plot for treatments aimed at reducing mortality from Diabetes. This plot was created using the BUGSnet package (Béliveau et al., 2019). 9

2.1 Illustration of the average absolute effects of different treatments. Panel i): "no treatment" ($\varnothing$), placebo (P), component A, and component B. Panel ii): treatment $A + B$ if the anchor is placebo, and $A + B$ if the anchor is "no treatment". 25

2.2 Network plots of the two scenarios considered in the simulation study. The number of trials making each comparison is written on the corresponding edge. Network 1) is shown on the left and Network 2) on the right. 30

2.3 Bias of estimates of relative effect parameters and heterogeneity for Network 1 (left, pink) and Network 2 (right, green). The relative effect parameters shown correspond to treatment effects relative to the treatment used to generate the data for that scenario. 33

2.4 Coverage probability and average length of 95% credible or confidence intervals for the relative effect parameters monitored in the simulation for network 1 (top row, pink) and network 2 (bottom row, green). 34

2.5 Network 1 average rankings based on SUCRA/P-scores and percentage of datasets where the value of $\Delta = DIC(\text{Model 2.4.2}) - DIC(\text{Model 2.2.1})$ is less than -2, between -2 and 2, and greater than 2. Negative values of Δ indicate that the unanchored model has a better fit. Differences of DIC greater than 2 are considered meaningful (Spiegelhalter et al., 2002). 35

2.6 Network 2 average rankings based on SUCRA/P-scores and percentage of datasets where the value of $\Delta = DIC(\text{Model 2.4.2}) - DIC(\text{Model 2.2.1})$ is less than -2, between -2 and 2, and greater than 2. Negative values of Δ indicate that the unanchored model has a better fit. Differences of DIC greater than 2 are considered meaningful (Spiegelhalter et al., 2002). 38

2.7 Log-odds ratios of the five psychological components versus usual care and 95% confidence/credible intervals for the four additive CNMA models. Note that a point estimate and confidence interval is missing from the frequentist model for the relaxation component due to this not being readily available from the R package netmeta. 41

3.1 The hypothetical posterior distributions for the relative effect of treatments A, C, and D compared to treatment B that created Sets 1 (left, SUCRA(A) = 0.411, SUCRA(B) = 0.472, SUCRA(C) = 0.530, SUCRA(D) = 0.586) and Set 2 (right, SUCRA(A) = 0.005, SUCRA(B) = 0.334, SUCRA(C) = 0.667, and SUCRA(D) = 0.994). 46

3.2 The simulated posterior distributions (top row), POTH residuals $r_{\text{POTH}}(j)$ (middle row), and $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ (bottom row) for two different hypothetical scenarios. Smaller values indicate a better treatment. 59

3.3 Histogram of POTHs from random-effects NMAs performed on 267 datasets available from Petropoulou et al. (2017); Papakonstantinou (2019). 60

3.4 Left: Scatterplot of the POTH versus the number of treatments in the network in the 267 random effects NMAs. Right: Scatterplot of the POTH versus the proportion of treatment comparisons with a p-value less than 0.05 in the 267 random effects NMAs. That is, the x-axis is the proportion of treatment comparisons where there is evidence that the treatments being compared have different effects. 62

3.5 Plots of POTH versus the estimated heterogeneity parameter σ for the three most common effect measures in **nmadb**: log odds ratio, log relative risk, and mean difference. There is an outlying point that was omitted from the mean difference plot which had an estimated σ of 68.06 and POTH equal to 0.597. 63

3.6 Above: Estimated confidence intervals (CI) for the odds ratio of all-grade treatment-related adverse events for ICI versus conventional therapy in the cancer network. P-scores and residual POTH are shown to the left of CIs. ORs less than one favour the intervention and ORs greater than one favour conventional therapy. The treatments are organized from largest (highest ranking in hierarchy) to smallest (lowest ranking in hierarchy) P-scores. A positive residual indicates that the treatment contributes to the certainty in the hierarchy, while a negative residual indicates that the treatment reduces the certainty in the hierarchy. Below, left: Best k treatments POTH for $k = 2$ to N treatments in the ICI network. Below, right: POTH residuals $r_{\text{POTH}}(j)$ for all treatments in the network. 64

3.7 Distribution of POTH values observed in the same network as the ICI example but with no treatment differences. 65

4.1 Boxplots of non-zero measured emission rates for facility types (strata). There are two outliers above 200 kg/h in the dataset for the sweet gas plant facility type. The number of non-zero measurements for each stratum is shown in the box below each boxplot. Acronyms: GP = gas plant, MWB = multi-well battery, LNGP = liquid natural gas plant, CO = crude oil, CTF = custom treating facility, GGS = gas gathering system, SWB = single-well battery, MS = meter station, TF = tank farm, NGL = natural gas liquids. 71

4.2 Percent bias for different estimators of the stratum totals $T(h)$ and provincial total T. 88

4.3 Mean squared error for different estimators of the stratum totals $T(h)$ and provincial total T when considering stage II uncertainty, including the contribution from bias and from variance. 89

4.4 Coverage probability of 95% Wald CIs for all estimation approaches by stratum. The dotted line represents the nominal coverage probability. 90

4.5 Point estimates and 95% CIs for all methods compared on the BC data. In the legend, considering stage II indicates that $D = 365$ whereas ignoring stage II indicates that $D = d_p$. Similarly, considering measurement uncertainty means the measurement uncertainty MC algorithm was used while ignoring measurement uncertainty means that the bias-correction method without the algorithm was used. 92

4.6 The estimated variance contributions for the provincial total from different uncertainty sources, with and without considering measurement error through the MC algorithm. Results are from the multi-stage IPW approach where stage II uncertainty was considered. 94

4.7 Percent contribution to the estimated total variance by uncertainty source for every stratum in the BC data. Strata are organized from highest (top left) to lowest (bottom right) emitting. The results are from the multi-stage IPW approach which includes stage II and measurement uncertainty. Acronyms: GP = gas plant, MWB = multi-well battery, LNGP = liquid natural gas plant, CO = crude oil, CTF = custom treating facility, GGS = gas gathering system, SWB = single-well battery, MS = meter station, TF = tank farm, NGL = natural gas liquids. 95

5.1 Uncertainty quantification model results for QOGI Provider A. The model was fit to data from the first campaign. No external data were available. . 110

5.2 Uncertainty quantification model results for QOGI Provider B. The model was fit to data from the second campaign. No external data were available. 111

5.3 Uncertainty quantification model results from QOGI Provider C. The model was fit to data from the second campaign. CR data from the first campaign was used as external data. 112

5.4 Uncertainty quantification model results for truck-based TDLAS data. The model was fit to data from the second campaign. CR data from the first campaign was used as external data. 113

5.5 Side-by-side comparison of Bayesian model proposed in Table 5.3 and that presented in Conrad et al. (2023b) for aerial TDLAS. 114

5.6 Uncertainty quantification model results for aerial NIR HSI data. The model was fit to data from the second campaign. CR data from the first campaign were used as external data. 115

5.7 Prior and posterior distributions for the true value of the emission rate given the observed measurement(s), prior distribution, and CR data. Posterior 90% CIs are shown with dashed lines. 117

B.1 Reference distributions for POTH where there are no true treatment differences in the network with the lowest (left) and highest (right) POTH observed in the database analysis. 143

B.2 Forest plots showing estimates and 95% CIs for relative effects of network producing the lowest (top) and highest (bottom) POTH in the database of NMAs. 144

B.3 Above: Estimated confidence intervals (CI) for the relative risk of mortality for anti-fungal treatments versus control in the organ transplant network. P-scores and residual POTH are shown to the left of CIs. RRs less than one favour the intervention and RRs greater than one favour control. The treatments are organized from largest (highest ranking in hierarchy) to smallest (lowest ranking in hierarchy) P-scores. A positive residual indicates that the treatment contributes to the certainty in the hierarchy, while a negative residual indicates that the treatment reduces the certainty in the hierarchy. Below, left: Best k treatments POTH for $k = 2, ..., N$ in the organ transplant network. Below, right: POTH residuals $r_{\text{POTH}}(j)$ for all treatments in the network. 145

B.4 Reference distribution for antifungals example where no treatment differences exist. 146

B.5 Reference distribution for POTH in the depression example, where there are no true treatment differences. 148

B.6 Estimated CIs for the odds ratio of 50% improvement of symptom severity for pharmacological interventions versus placebo in the depression network. P-scores and residual POTH are shown to the left of CIs. ORs greater than one favour the intervention and ORs less than one favour placebo. The treatments are organized from largest to smallest P-scores. A positive residual indicates that the treatment contributes to the certainty in the hierarchy, while a negative residual indicates that the treatment reduces the certainty in the hierarchy. 149

B.7 Left: Best k treatments POTH for $k = 2$ to N treatments in the depression network. Right: POTH residuals $r_{\text{POTH}}(j)$ for all treatments in the network. 150

B.8 Mean of POTH over the simulated datasets for different combinations of heterogeneity (represented by τ), network type, and standard error in the data. 152
B.9 Histograms showing the distribution of POTH observed for different settings in the simulation study. 153

List of Tables

1.1 A hypothetical network involving a multicomponent treatment. Studies 1-3 involve only single-component treatments, while the treatment in arm 2 of study 4 is multi-component (education and support group). 8

2.1 Summary of the model elements used in the new unified notation for additive CNMA models in this thesis. 19

2.2 Summary table of the four CNMA models described in this chapter, labeled by the subsection in which they were introduced. Note that the definition of $\boldsymbol{\epsilon}_i$ for the model in Section 2.2.1 and its dimension is slightly altered if trial i includes the reference treatment. 29

2.3 Parameter values used in the simulation for Network 1 and Network 2. Values of the parameters under different anchors are also given. The value of the single component relative effects and the heterogeneity σ are the same for different anchors. The multicomponent relative effects and the baseline responses are different for different anchor treatments. 32

2.4 Point estimates and 95% CIs for the elements of $\boldsymbol{d}$ from the Bayesian unanchored arm-based model. 42

3.1 The probability mass function for rank(k). 48

3.2 Tables of ranking probabilities corresponding to the most certainty in a hierarchy (left) and illustrating the least certainty in a hierarchy (right). The values in the tables correspond to the probability of treatment k taking rank u, p_{ku}. The bottom rows show the SUCRA values and expected ranks for each treatment. 50

3.3 Table of ranking probabilities corresponding to the hypothetical scenario where half of the treatments perform equally well and half equally poorly ($G = 2$). The values in the table correspond to p_{ku}, the probability of treatment k taking rank u. 54

4.1 Description of facilities, Petrinex subtype codes, and stratum definitions used in the data analysis. 72

4.2 Estimates of the total methane emissions in BC for 2021, variance estimates by uncertainty source, and total variance estimates by stratum using the multi-stage approach with IPW considering all three stages and measurement uncertainty. Due to rounding, small non-zero values are shown as 0.00. Acronyms: GP = gas plant, MWB = multi-well battery, LNGP = liquid natural gas plant, CO = crude oil, CTF = custom treating facility, GGS = gas gathering system, SWB = single-well battery, MS = meter station, TF = tank farm, NGL = natural gas liquids. 93

5.1 Summary of the units for each parameter in the likelihood, important to keep in mind when specifying prior distributions. Exclusion criteria: * = may be removed if technology does not have false positives in CR data, † = may be removed if its removal leads to a simpler model which has adequate fit and predictive performance. 105

5.2 Summary of technologies, providers, and number of non-zero measurements from the CR campaigns conducted at the CMC facility in Alberta. N_1 and N_2 refer to the number of observations collected for a given technology during the first and second campaigns, respectively. 108

5.3 Summary of selected models for each methane quantification technology provider. 109

5.4 Summary of 90% credible intervals (CI) resulting from weighted bootstrap algorithm for different priors and numbers of new measurements made. . . 118

B.1 Two scenarios illustrating different joint ranking distributions where all SUCRAs are equal to 0.5. In Scenario 1, the certainty in the rank of treatment 3 is increased, with an equal and opposite decrease in the certainty of the rank of treatment 4. In Scenario 2, both treatments 3 and 4 have multimodal ranking distributions which have equal certainty to those treatments with flat ranking distributions. 140

B.2 Between-study heterogeneity (σ^2), I^2, and POTH for the two extreme networks in the database. 144

C.1 A summary of the notation and terms used in Chapter 4. 155

C.2 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 165

C.3 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 166

C.4 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 167

C.5 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 168

C.6 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 169

C.7 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 170

C.8 Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum 171

C.9 Estimates of the total methane emissions for 2021 from facilities in BC south of the 58 degree north latitude line, variance estimates by uncertainty source, and total variance estimates by stratum 172

Chapter 1

Introduction

This thesis examines current data-driven problems and provides statistical developments to address them. In the first part of the thesis, we propose new models and methodology for Network Meta-Analysis (NMA). In the second part of the thesis, we develop novel applications of statistical methods to improve the quantification of methane emissions from oil and gas (OG) producers. We provide the necessary background on NMA in Section 1.1 and introduce key concepts in methane emissions quantification in Section 1.2. In Section 1.3, we provide an outline of the structure and contributions of the following chapters of this thesis.

1.1 Network Meta-Analysis

NMA is an extension of pairwise meta-analysis which synthesizes information on the safety or efficacy of two or more treatments from a set of Randomized Controlled Trials (RCTs) (Lumley, 2002; Psaty et al., 2003). The set of RCTs forms a network of evidence. An example of a network is shown in Figure 1.1, which is composed of trials comparing treatments for Diabetes (Elliott and Meyer, 2007). Treatments which are directly compared in a trial are connected by an edge, for example, blocker and ARB. NMA enables the estimation of the relative effects of all treatments in a network, even if they were not directly compared in an RCT, such as ARB and ACE inhibitor in Figure 1.1.

Consistency is a core concept in NMA models that relates direct and indirect evidence. Consistency can be expressed via the consistency equations,

$$d_{k,l} = d_{m,l} - d_{m,k} \text{ for } k \neq l \neq m, \tag{1.1}$$

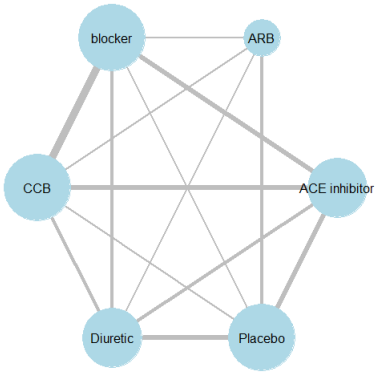


Figure 1.1: An example of a network of RCTs studying treatments for Diabetes. The thickness of the edges is proportional to the number of RCTs which directly make each comparison. The network was created with the BUGSnet package (Béliveau et al., 2019).

where $d_{k,l}$ is the true relative effect of treatment l versus treatment k. Consistency describes logical relationships between the population relative effects and reduces the dimension of the parameter space (Lu and Ades, 2009). In a standard NMA it is only necessary to estimate $\{d_{t,k}\colon\ k = 1, \ldots, N\}$, where N is the number of treatments in the network and t is some reference treatment, because consistency enables the computation of any other comparison from the relative effects in this set (Lu and Ades, 2009). The relative effects $\{d_{t,k} : k = 1, \ldots, N\}$ are referred to as the basic parameters and the other relative effects which are functions of the basic parameters are called functional parameters (Dias et al., 2018).

Both frequentist and Bayesian approaches to NMA are commonly used, although there are advantages to a Bayesian approach (Dias et al., 2018). Bayesian models are flexible and can be adapted to a variety of data types. Bayesian approaches that use arm-level trial data can even accommodate binomial data where some trials have zero counts without requiring adjustment (Dias et al., 2018). Frequentist approaches require non-standard methods in this situation, such as adding a small value to all cells which may introduce bias or the use of non-standard effect scales or link functions (Efthimiou, 2018; Efthimiou et al., 2019). Another advantage is that Bayesian approaches automatically account for un-

certainty in the estimation of heterogeneity in random-effects models, whereas frequentist approaches typically estimate the heterogeneity parameter and then treat it as fixed and known because this assumption is required for the relative effect estimates to be unbiased (Walter and Balakrishnan, 2022). The uncertainty in the heterogeneity from a Bayesian approach is propagated to the posterior distributions of the relative effects, allowing for a better reflection of the true uncertainty in the estimates. In the following subsections, we describe the Bayesian and frequentist approaches to NMA which are built upon throughout Chapters 2 and 3, although this is not an exhaustive list of models for NMA.

1.1.1 Bayesian NMA Models

Dias et al. (2018) provide two Bayesian models for NMA which differ in the type of data they use; the first uses data summarized at the arm level of trials, and the second uses data summarized at the contrast level. First, consider that we have a set of I trials, each of which has a_i treatment arms for which the outcome is binary, for example, cured or not, and assume that the number of outcomes in arm j of trial i, R_{ij}, is reported for $j = 1, ..., a_i$ and $i = 1, ..., I$. We can model R_{ij} using the binomial likelihood, that is, $R_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij})$ where n_{ij} is the number of patients in arm j in trial i and π_{ij} is the probability of a patient in trial i who was in arm j having the outcome.

Note that we distinguish between the arm number and the treatment employed in that arm. For example, a trial may have three treatment arms, and the arms are labelled 1, 2, and 3. The treatments employed in each arm may be treatment numbers 2, 4, and 5. Thus, let t_{ij} be the index of the treatment in arm j of trial i.

As discussed previously, we need only estimate a set of basic parameters that are relative to a common reference treatment. In the Bayesian approach, one of the treatments in the network is chosen as the reference treatment, which further reduces the number of parameters that must be estimated by one, since $d_{k,k} = 0$ for any treatment k. The choice of the reference treatment is arbitrary in standard NMA models but is nevertheless important in the model set-up and interpretation of the basic parameters (Lu et al., 2011). Without loss of generality, we assume that the reference treatment is the treatment labelled as 1. Thus, in this model we will estimate the set of relative effects $\{d_{1,k} : k = 2, \ldots, N\}$, which form the basic parameters in this model.

Returning to the model, the probability of the outcome occurring in arm j of trial i, π_{ij}, is modelled as follows:

$$\text{logit}(\pi_{ij}) = \mu_i + \mathbf{I}\{j \neq 1\}\delta^i_{t_{i1},t_{ij}} \text{ for } j = 1, \ldots, a_i, \tag{1.2}$$

where $\mathbf{I}\{j \neq 1\} = 1$ if $j \neq 1$ and 0 otherwise, μ_i is the log odds of response from patients in trial i if they were treated with the treatment in arm 1 of trial i, and $\delta^i_{t_{i1},t_{ij}}$ is the trial-specific relative effect of the treatment in arm j of trial i compared to the treatment in arm 1 of trial i, that is, $\delta^i_{t_{i1},t_{ij}}$ is the log-odds ratio of t_{ij} versus t_{i1}. Note that $\delta^i_{t_{i1},t_{i1}} = 0$ by definition. The parameters μ_i, $i = 1, ..., I$ are called the trial-specific baselines, and are generally considered nuisance parameters.

At the next level of the hierarchical model, we have

$$\boldsymbol{\delta^i} = \begin{pmatrix} \delta^i_{t_{i1},t_{i2}} \\ \vdots \\ \delta^i_{t_{i1},t_{ia_i}} \end{pmatrix} \sim \text{MVN}_{(a_i-1)} \left(\begin{pmatrix} d_{t_{i1},t_{i2}} \\ \vdots \\ d_{t_{i1},t_{ia_i}} \end{pmatrix}, \sigma^2 \boldsymbol{\Sigma}_{(a_i-1)\times(a_i-1)} \right), \tag{1.3}$$

where $\boldsymbol{\Sigma}_{(a_i-1)\times(a_i-1)}$ is a square matrix of dimension $(a_i - 1) \times (a_i - 1)$ with ones on the diagonal and 0.5 on the off-diagonal entries. The parameter σ^2 represents the common between-trial variance of the relative effects, that is, the between-trial variance for $\delta^i_{t_{i1},t_{ij}}$ is equal to σ^2 for any t_{i1} and t_{ij}. More complex variance structures are also possible (Lu and Ades, 2009). A common-effect model is also possible by setting $\sigma^2 = 0$. Recall that due to consistency, we have $d_{t_{i1},t_{ij}} = d_{1,t_{ij}} - d_{1,t_{i1}}$.

It can be more convenient to work with the univariate distribution of $\delta^i_{t_{i1},t_{ij}}$ conditional on $\delta^i_{t_{i1},t_{i2}}$, ..., $\delta^i_{t_{i1},t_{ij-1}}$ (Dias et al., 2018). This distribution is

$$\delta^i_{t_{i1},t_{ij}} \mid \begin{pmatrix} \delta^i_{t_{i1},t_{i2}} \\ \vdots \\ \delta^i_{t_{i1},t_{ij-1}} \end{pmatrix} \sim \text{N}\left(d_{1,t_{ij}} - d_{1,t_{i1}} + \frac{1}{j-1} \sum_{\ell=1}^{j-1} (\delta^i_{t_{i1},t_{i\ell}} - (d_{1,t_{i\ell}} - d_{1,t_{i1}})), \frac{j}{2(j-1)} \sigma^2 \right).$$

Priors must be set on the parameters μ_i for $i = 1$, ..., I, $d_{1,k}$ for $k = 2$, ..., N, and σ. Typically vague priors are used, because we wish for the observed data to exert the most influence on the estimated relative effects (Dias et al., 2018). These are typically $\mu_i \sim \text{N}(0, 10^2)$, $d_{1,k} \sim \text{N}(0, 10^2)$, and $\sigma \sim \text{Unif}(0, \text{upper bound})$ where the upper bound is chosen based on the scale of the data and the expected range of observed treatment effects.

The model described above uses data summarised at the arm level, that is, the aggregated event totals for each arm of each trial. Some trials may only report differences between arms, such as the log-odds ratio of arm 2 versus 1, 3 versus 1, and so on. We refer to this type of data as contrast-level or contrast-based. We must use a different likelihood for this type of data. In the typical Bayesian approach, we work with $a_i - 1$ log-odds ratios from each trial, $i = 1, ..., I$. Let y_{ij}, $j = 2$, ..., a_i be the log-odds ratio of the treatment

t_{ij} compared to t_{i1} observed in trial i. Additionally, let SE_{ij} be the standard error of the log-odds ratio y_{ij}, and $SE_{t_{i1}}$ be the standard error of the log-odds in arm 1 of trial i. Then we use a normal likelihood as follows:

$$\begin{pmatrix} y_{i2} \\ \vdots \\ y_{ia_i} \end{pmatrix} \sim \text{MVN}_{a_i-1}\left(\boldsymbol{\delta^i}, \boldsymbol{S}_i\right),$$

where

$$\boldsymbol{S}_i = \begin{pmatrix} SE_{i2}^2 & \dots & SE_{t_{i1}}^2 \\ \vdots & \ddots & \vdots \\ SE_{t_{i1}}^2 & \dots & SE_{ia_i}^2 \end{pmatrix},$$

noting that the diagonal elements of $\boldsymbol{S}_i$ represent variances of contrasts and the off-diagonal elements represent the variance of the responses in arm 1 of trial i. $\boldsymbol{\delta^i}$ has the same random-effects distribution as given in (1.3), and similar priors can be used for the $d_{1,k}$'s and σ.

Finally, note that although we focus on binomial data and the log-odds ratios in this section, it is simple to adapt these models to other types of data by using a different likelihood and link function on the lefthand side of Equation (1.2), or different summary data for contrast-based methods. For example, continuous data can be modelled in an arm-based Bayesian model by using the means of each trial arm as the outcome and the identity link function in Equation (1.2).

1.1.2 Frequentist NMA Model

The frequentist approach to NMA uses weighted least squares regression and contrast-based data, but various approaches to accounting for multi-arm trials exist (Lu et al., 2011; Rücker and Schwarzer, 2014). In this section, we focus on the method proposed in Rücker and Schwarzer (2014).

In the previous section, we discussed a contrast-based model which used $a_i - 1$ contrasts from each trial, where the contrasts were all relative to a common comparator (arm 1 of the trial). For this model, we use all $m_i = \binom{a_i}{2}$ contrasts that can be made from the a_i arms. Note that the m_i contrasts are not independent, in the sense that some contrasts are functions of other contrasts. Let y_{ij} be the jth contrast in trial i, and $\boldsymbol{y} = (y_{11}, \dots, y_{1m_1}, \dots, y_{I1}, \dots, y_{Im_I})'$ be the vector of all observed contrasts in the network, which is of dimension $M \times 1$ where $M = \sum_{i=1}^{I} m_i$. Also, let t_{j1} be the treatment index

of the minuend[1] of contrast j, and let t_{j2} be the treatment index of the subtrahend[2] of contrast j.

Let $\boldsymbol{B}$ be the $M \times N$ edge-vertex incidence matrix. In each row, $\boldsymbol{B}_{jl} = 1$ if contrast j has treatment l in the minuend, -1 if treatment j is the subtrahend of contrast l, and 0 otherwise. Additionally, let $\boldsymbol{W}$ be a diagonal matrix with $\boldsymbol{W}_{jj}$ equal to the inverse of the squared standard error of contrast j (analogous to $1/SE_{ij}^2$ using the notation from the Bayesian contrast-based method). If the model includes multi-armed trials, the matrix $\boldsymbol{W}$ needs to be adjusted to account for the correlations between the contrasts within multi-armed trials. The Bayesian model accounts for this in the likelihood by using a covariance matrix which is not diagonal, $\boldsymbol{S}_i$. However, Rücker and Schwarzer (2014) take a different approach in which the weights in $\boldsymbol{W}$ are reduced for contrasts from multi-arm trials which we describe below (Rücker and Schwarzer, 2014).

The procedure and notation described in Rücker and Schwarzer (2014) for adjusting for multi-arm trials can be confusing, so we introduce our own slightly different notation and explain the procedure here. Let $\boldsymbol{B}_i$ be a $m_i \times a_i$ matrix which describes all the possible contrasts between arms within trial i similarly to how $\boldsymbol{B}$ describes all the contrasts in the network. $\boldsymbol{V}_i$ is a $a_i \times a_i$ symmetric matrix were $V_{i,kl}$ equals the variance of the contrast between treatment arm k and treatment arm l in trial i, or 0 if $k = l$. We also introduce $\boldsymbol{L}_i$ and $\boldsymbol{L}_i^+$, the trial-specific Laplacian matrix and its Moore-Penrose pseudoinverse, respectively. It can be shown that $\boldsymbol{L}_i^+$ can be written as

$$-\frac{1}{2a_i^2}\boldsymbol{B}_i'\boldsymbol{B}_i\boldsymbol{V}\boldsymbol{B}_i'\boldsymbol{B}_i \tag{1.4}$$

(Rücker and Schwarzer, 2014). To adjust the weights of a multi-arm trial, $\boldsymbol{L}_i^+$ is computed according to (1.4), and then $\boldsymbol{L}_i$ is determined by $\boldsymbol{L}_i = (\boldsymbol{L}_i^+)^+$. The new adjusted weights which account for within-trial correlation are the absolute value of the negative non-diagonal entries of $\boldsymbol{L}_i$. Then $\boldsymbol{W}_{adj}$ is the matrix with the adjusted weights as calculated using the above procedure for all multi-armed trials on the diagonal (and the inverse of the observed variances for two-armed trials).

For a common-effect model, we assume

$$\boldsymbol{y} \sim \text{MVN}_M\left(\begin{pmatrix} d_{t_{12},t_{11}} \\ \vdots \\ d_{t_{M2},t_{M1}} \end{pmatrix}, \boldsymbol{W}_{adj}^{-1}\right),$$

[1] A quantity from which another quantity is subtracted from. For example, in the equation $5 - 1 = 4$, 5 is the minuend.

[2] A quantity which is subtracted from another quantity. For example, in the equation $5 - 1 = 4$, 1 is the subtrahend.

where $\boldsymbol{W}_{adj} = \boldsymbol{W}$ for a network with no multi-armed trials.

Then a weighted regression estimate of $(d_{t_{12},t_{11}}, \ldots, d_{t_{M2},t_{M1}})'$, the vector of the population means of the contrasts in $\boldsymbol{y}$, is given by

$$\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{W}_{adj}\boldsymbol{B})^{+}\boldsymbol{B}'\boldsymbol{W}_{adj}\boldsymbol{y}. \tag{1.5}$$

The Moore-Penrose pseudoinverse is needed because for most network structures, $\boldsymbol{B}'\boldsymbol{W}_{adj}\boldsymbol{B}$ is not invertible. Since $\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{W}_{adj}\boldsymbol{B})^{+}\boldsymbol{B}'\boldsymbol{W}_{adj}$ plays the role of the "hat" matrix in regression, we let $\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{W}_{adj}\boldsymbol{B})^{+}\boldsymbol{B}'\boldsymbol{W}_{adj} = \boldsymbol{H}$. Note that the vector estimated in (1.5) could have repeated elements. It also does not necessarily provide estimates for a comprehensive set of parameters with a common reference treatment in the same way as the Bayesian method. For example, consider a small network where trial 1 compares treatments 1 and 2, trial 2 compares treatments 2 and 3, and trial 3 compares treatments 3 and 4. Then Equation (1.5) will provide estimates of $d_{1,2}$, $d_{2,3}$, and $d_{3,4}$ only.

Finally, a random-effects model is also possible using this frequentist approach. An estimate of the between-trial variance is calculated and then added to the observed variances of each comparison before computing $\boldsymbol{W}_{adj}$. The heterogeneity variance σ^2 can be estimated using a method-of-moments;

$$\hat{\sigma}^2 = \max\left(\frac{\boldsymbol{Q} - df}{tr\{(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{U}\boldsymbol{W}\}}, 0\right),$$

where $\boldsymbol{Q}$ is also known as Cochran's Q, and is computed according to $\boldsymbol{Q} = (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})'\boldsymbol{W}_{adj}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})$ (Rücker and Schwarzer, 2014), $\boldsymbol{U}$ is an $M \times M$ block diagonal matrix derived from the matrix $\boldsymbol{B}(\boldsymbol{B}^{1/2})'$ (see Rücker and Schwarzer (2014) for details), and df is the degrees of freedom, computed as

$$df = \sum_{a=1}^{\max a_i, i=1,\ldots I} (a-1)n_a - (N-1),$$

where a represents the number of arms and n_a is the number of trials in the network with a arms (Rücker and Schwarzer, 2014).

1.1.3 Multicomponent Treatments

Multicomponent treatments refer to treatment protocols which are composed of combinations of other treatments. We refer to the treatments which act as sub-units for multicomponent treatments as components. For example, consider the hypothetical network of

evidence described in Table 1.1. The components education and support group are employed as treatments on their own in studies 2 and 3 respectively, and in combination in a multicomponent treatment in study 4. To analyse the data with a standard NMA model, one must decide whether the treatment "Education & support group" should be lumped with one of the nodes "Education" or "Support group", or treated as a new node in the network. Lumping of treatments can lead to increased heterogeneity, while defining new nodes for all multicomponent treatments can lead to a sparse or even disconnected network (Higgins et al., 2019). A solution is to analyse networks involving multicomponent treatments with a Component NMA (CNMA) model, which facilitates the estimation of the relative effects of components and multicomponent treatments simultaneously (Petropoulou et al., 2021; Welton et al., 2009; Rücker et al., 2019; Efthimiou et al., 2022).

Table 1.1: A hypothetical network involving a multicomponent treatment. Studies 1-3 involve only single-component treatments, while the treatment in arm 2 of study 4 is multi-component (education and support group).

Study	Arm 1	Arm 2
1	Usual care	Counselling
2	Usual care	Education
3	Usual care	Support group
4	Usual care	Education & support group

1.1.4 Ranking Treatments in NMA

A popular outcome reported in NMA is a ranking of treatments from best to worst, also known as a treatment hierarchy (Papakonstantinou et al., 2022). Treatment hierarchies are produced by calculating ranking metrics which summarise the estimated relative effect of each treatment, ideally while taking into account their precision. The ranking metrics can then be ordered from largest to smallest to obtain the treatment hierarchy. We refer to the ordered list of treatments from best to worst as a hierarchy instead of a ranking to avoid confusion between the ranking metrics and the final output based on the ranking metrics (the hierarchy).

Common ranking metrics used to produce hierarchies include the Surface Under the Cumulative RAnking curve (SUCRA), the P-score (which is a frequentist version of SUCRA), the expected rank, and the probability of having the best outcome value (Rücker and Schwarzer, 2015; Salanti et al., 2022). The predictive P-score is a Bayesian ranking

metric that measures the expected performance of each treatment in a new study (Rosenberger et al., 2021). SUCRAs and P-scores have been shown to be less sensitive to different levels of precision in the effect estimates compared to the probability of having the best value and are commonly reported (Chiocchia et al., 2020; Davies and Galla, 2021; Salanti et al., 2022). The expected rank is a function of SUCRA and will result in the same hierarchy (Rücker and Schwarzer, 2015).

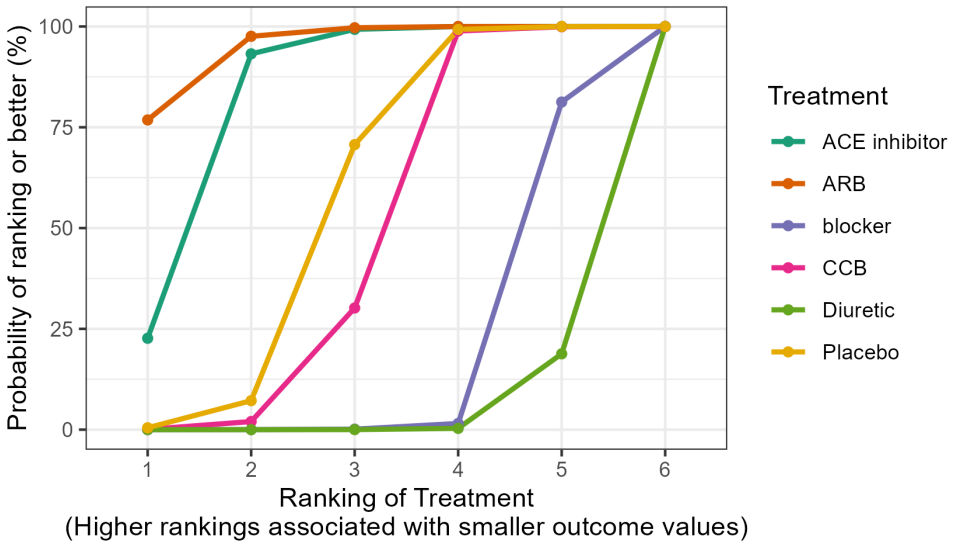


Figure 1.2: An example of a SUCRA plot for treatments aimed at reducing mortality from Diabetes. This plot was created using the BUGSnet package (Béliveau et al., 2019).

The formula for the SUCRA of treatment k is

$$\text{SUCRA}(k) = \frac{\sum_{r=1}^{N-1} \sum_{u=1}^{r} p_{ku}}{N-1}, \quad k = 1, \ldots, N, \tag{1.6}$$

where p_{ku} is the probability that treatment k takes rank u and N is the number of treatments in the network. SUCRA values can also be visualized by plotting the probability that each treatment ranks r or better versus the ranking r. The SUCRA value for treatment k is then the area under this curve, scaled by $1/(N-1)$. An example of such a plot using the Diabetes data is shown in Figure 1.2. The corresponding treatment hierarchy and SUCRA values, from most to least preferred, are ARB (SUCRA = 0.948), ACE inhibitor (SUCRA = 0.830), placebo (SUCRA = 0.555), CCB (SUCRA = 0.462), blocker (SUCRA = 0.166), and diuretic (SUCRA = 0.038).

SUCRA values are calculated in a Bayesian context where samples from the posterior can be used to estimate the rank probabilities p_{ku}. SUCRA values are more challenging to calculate in a frequentist NMA as the rank probabilities p_{ku} are less readily available in this context, but they can be obtained using resampling methods (White, 2011).

P-scores were proposed as a frequentist version of SUCRA values which do not require resampling (Rücker and Schwarzer, 2015). P-scores are calculated using the relative effect estimate of k versus l, $\hat{d}_{l,k}$, and their standard errors $SE(\hat{d}_{l,k})$, assuming a larger response is favourable, according to

$$\mathrm{P}(k) = \frac{1}{N-1} \sum_{\substack{l=1 \\ l \neq k}}^{N} \Phi \left(\frac{\hat{d}_{l,k}}{SE(\hat{d}_{l,k})} \right), \tag{1.7}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. If smaller responses are favourable, $\hat{d}_{l,k}$ is replaced with $-\hat{d}_{l,k} = \hat{d}_{k,l}$. P-scores and SUCRAs are equivalent in a Bayesian analysis when the posterior distribution of the relative effects are independent and normally distributed (Rücker and Schwarzer, 2015).

1.2 Methane Emissions Quantification

Methane is a strong greenhouse gas with a short lifetime in the atmosphere compared to carbon dioxide, meaning reductions can have an impact in a short timeframe (Ripple et al., 2024). Methane emissions from the OG sector account for approximately 40% of anthropogenic sources globally, making it a focus for emissions reductions (IEA, 2023). Initiatives aimed at reducing methane emissions include the setting of emissions reduction targets and the compilation of methane inventories where emissions are estimated and attributed to different sources to understand major contributors to emissions and track progress towards targets (Climate and Clean Air Coalition, 2021; Environment and Climate Change Canada, 2016; Intergovernmental Panel On Climate Change, 2023). Previous approaches to estimating inventories from the exploration and production sector of the OG industry, also known as upstream OG, typically relied on multiplying emissions factors (EFs) by the number of facilities in a given jurisdiction, but these approaches have been found to be inadequate (MacKay et al., 2021; Tyner and Johnson, 2021; Zavala-Araiza et al., 2018; Vollrath et al., 2024). Reasons include ignoring the underlying skewed distribution of emissions, limited data, and ignoring the uncertainty associated with the sampling and measurement methods in the calculation of EFs. An alternative to outdated EF approaches is a measurement-based inventory, which is derived by surveying OG facilities

with methane detection and quantification technologies (MDQTs) and using these measurements to inform the inventory (Tyner and Johnson, 2021). In a review of approaches to measuring methane emissions from OG facilities, Vollrath et al. (2024) highlighted the need for statistical frameworks that can integrate measurement data while considering associated uncertainties. Important sources of uncertainty include detection and measurement uncertainty from MDQTs, uncertainty from the sampling design of measurement-based surveys, and temporal variability of emissions.

Another initiative designed to reduce methane emissions from OG production and storage facilities is the implementation of regulatory requirements for Leak Detection And Repair (LDAR) programs (Atherton et al., 2017; Fox et al., 2021b). The idea is that OG producers must regularly survey their facilities with technology to detect leaks and repair identified leaks within a certain timeframe to reduce unnecessary emissions. Initial regulations tend to recommend the use of Optical Gas Imaging (OGI) technology for surveys, which is time-consuming as an individual operator must manually survey an entire facility with an OGI camera (Fox et al., 2021b; Atherton et al., 2017). Alternative approaches have become more popular which take advantage of emerging MDQTs which can operate at larger scales, such as by mounting on a truck or aircraft. These approaches can quickly survey an entire facility, but may require follow-up with OGI in order to localize detected leaks before they can be repaired. Several software programs have been developed to model the emissions reductions and cost-effectiveness of LDAR programs with various MDQTs, frequencies, and follow-up strategies, such as FEAST (Kemp and Ravikumar, 2021), LDAR-Sim (Fox et al., 2021a), and AroFEMP (Lavoie et al., 2021). However, not all approaches consider the measurement error of the technologies being deployed, despite this factor potentially impacting the effectiveness of programs (Fox et al., 2021b). Accurate models for measurement uncertainty of MDQTs are needed to ensure deployed LDAR programs reduce methane emissions as much as possible. These models are also needed to integrate measurement error into methane inventory estimation techniques.

1.3 Outline of Contributions

CNMA is an increasingly popular method for synthesizing evidence which involves multicomponent treatments (Petropoulou et al., 2021), but there is a lack of rigourous comparison between existing Bayesian and frequentist approaches. In Chapter 2 of this thesis, we identify weaknesses in the definition of additivity used in CNMA models and highlight an important difference in the assumptions underlying common Bayesian and frequentist CNMA models that has been overlooked in the literature until now. In particular, we show

that the most popular Bayesian CNMA model makes a stronger assumption than the frequentist approach to CNMA, causing it to be prone to misspecification. We introduce the notion of "anchored" (less flexible) and "unanchored" (more flexible) CNMA models. We propose novel Bayesian CNMA models which are unanchored, and show in a simulation study that the anchored Bayesian approach leads to biased estimates, poor coverage probability, and inaccurate treatment hierarchies when it is misspecified, while the unanchored frequentist approach and our novel unanchored Bayesian models are robust. The anchored and unanchored CNMA models are fit to a published network of treatments for coronary heart disease and we compare the results.

Once an NMA model has been fit, a common output of interest is a treatment hierarchy. Treatment hierarchies from NMA have important implications as they can aid in decision-making, however they are often reported without any measure of uncertainty, which can be misleading. In Chapter 3, we propose an approach to quantifying uncertainty in producing a treatment hierarchy from a frequentist or Bayesian NMA. By defining maximum and minimum certainty in a hierarchy in terms of the SUCRA or P-scores, we show that the variance of the SUCRA or P-scores captures useful information about the level of certainty in a treatment hierarchy. We use this insight to propose the Precision Of Treatment Hierarchy (POTH), which summarises the level of certainty in a treatment hierarchy into a single number between 0 and 1. We propose modifications to POTH to summarise the level of certainty in the hierarchy of a subset of treatments of interest, and approaches to identifying the relative contribution of each treatment to uncertainty in the hierarchy. To investigate the distribution of POTH and any associations with network characteristics, we calculate POTH for a database of over 200 published NMAs. We conclude with an in-depth worked example to demonstrate the methods on a real dataset.

In the second part of the thesis, we turn our attention to some problems in the quantification of methane emissions from the OG industry. Measurement-based methane inventories can provide an accurate estimate of methane emissions from a given jurisdiction, but existing approaches are computationally intensive and challenging to implement and interpret. In Chapter 4, we show how multi-stage survey sampling techniques can be used to produce a measurement-based methane inventory. We provide estimators of the total and stratum-level emissions, corresponding variance estimators, and an interpretable decomposition of the variance which can be useful in planning future surveys. We also provide an efficient approach based on the Hájek estimator which requires an adjustment to the second stage design. To address measurement error in the survey data, we describe a Monte Carlo approach which incorporates a known measurement model for the technology used to collect the data. In a simulation study, we show that the multi-stage estimators have good statistical properties. We apply the methods to aerial survey data of OG facilities

in British Columbia, Canada and compare the inventory to that of a previously published analysis.

An accurate model for the uncertainty associated with component-level measurements from MDQTs is needed as input for the methane inventory calculation approach proposed in Chapter 4, as well as modelling LDAR programs and other applications. In Chapter 5, we introduce a Bayesian method to develop probability distributions of measurements given a true emission rate using data collected in controlled releases (CR) of methane for different technologies. The approach includes flexible likelihoods which can capture complex relationships in the data. We also use a weighted bootstrap algorithm to provide the distribution of the true emission rate given a new measurement, which synthesizes the measurement with the CR data and external information about the possible true emission rate. We present results for six quantification technologies and show that flexible models that accommodate complex non-linear behaviour are needed to adequately model measurement error. We also demonstrate the use of the weighted bootstrap algorithm with different priors and show how uncertainty can be reduced by making repeated independent measurements.

To conclude, we summarise all contributions of this thesis in Chapter 6.

Chapter 2

Bayesian Unanchored Additive Models for Component Network Meta-Analysis

2.1 Introduction

In a recent review of methods for multicomponent interventions, Petropoulou et al. (2021) recommended the use of additive CNMA models (when the assumptions are satisfied) over other types of CNMA models. An additive CNMA model assumes that the relative effect of a multicomponent treatment is the sum of its components' relative effects, that is,

$$d_{A,X} = \sum_{x \in X} d_{A,x}, \tag{2.1}$$

where X is a multicomponent treatment, x indexes the components of X, and A is a common comparator treatment. Additive CNMA models were preferred in the review because they allow for comparison of relative efficacy of all possible combinations of components, and because of the availability of code and R packages which facilitate their implementation (Petropoulou et al., 2021). Existing additive CNMA models have been applied in practice by e.g. Mills et al. (2011, 2012); Madan et al. (2014); Pompoli et al. (2018); and Freeman et al. (2018).

There are two established additive CNMA methods; one proposed by Welton et al. (2009) and another proposed by Rücker et al. (2019). The model of Welton et al. (2009)

uses Bayesian inference and takes trial data summarised at the arm level. The approach of Rücker et al. (2019) is a frequentist model which uses data summarised at the contrast level. Similarly to standard NMA approaches, the two methods also differ in the way multi-arm trials are accounted for. Recently, several Bayesian CNMA models were proposed by Efthimiou et al. (2022). Since the work in this thesis chapter was completed before publication of Efthimiou et al. (2022), the terms "existing" or "established" methods in this chapter refer to the methods of Welton et al. (2009) and Rücker et al. (2019). Efthimiou et al. (2022) focuses on extending CNMA models to allow for the use of individual patient data and facilitating selection of component interactions. They describe their model III as Welton et al. (2009)'s model, however we note (from inspecting their model formulation and code) that it is not exactly Welton et al. (2009)'s model, but rather analogous to a model we propose in Section 2.4.1.

There has been limited work done assessing existing CNMA models. Thorlund and Mills (2012) investigated the performance of the standard Bayesian NMA model compared to the additive CNMA model of Welton et al. (2009) in a simulation study. Rücker et al. (2021) compared the frequentist CNMA model to a matching method on a dataset of multiple myeloma. However, no direct comparison of these CNMA methods has been made so far in the literature. The two models use different notation for different elements of the models, which makes comparison difficult.

In this chapter, we introduce a novel notation and compare existing CNMA models in Section 2.2. In Section 2.3, we introduce the definition of the anchor treatment and show that it plays an important role in CNMA models. This is a new insight and the idea of the anchor treatment was overlooked in the literature until now. We emphasize the role of the anchor treatment and make use of the novel notation of the previous section to show that existing CNMA models differ in their use of the anchor treatment. We also discuss the interpretation of placebo or "no treatment" as the anchor. In Section 2.4, we propose two novel Bayesian models for CNMA. We compare the existing and novel CNMA models in a simulation study in Section 2.5 and apply the models to a published dataset in Section 2.6. We conclude with a discussion of the results and possible avenues for future work in Section 2.7.

2.2 Comparison of Existing CNMA Models

In this section, we introduce the CNMA models of Welton et al. (2009) and Rücker et al. (2019) using new unified notation. The notation is summarized in Table 2.1.

2.2.1 Bayesian CNMA Model of Welton et al. (2009)

Welton et al. (2009) proposed a Bayesian method for analysing multicomponent treatments and the effects of their components using data summarised at the arm level. This model is the first CNMA model in the literature. The model is typically fit using Markov Chain Monte Carlo (MCMC) sampling via BUGS software. BUGS code corresponding to the model as originally presented is available in their paper.

In this model, similarly to the standard NMA model presented in Section 1.1.1, a reference treatment must be chosen. Without loss of generality, let the reference treatment be the treatment labelled as "1". If there are C components in the network, the model estimates $C-1$ relative effects where the comparator is the reference treatment. Elements of the model whose structure depends on the treatment chosen to be treatment 1 are written with a superscript (1).

Suppose we have I RCTs, indexed with i, each with a total of a_i arms. Assuming the outcome of interest is binary, such as death, then the trial data are aggregated as the counts of the outcomes in arm j of trial i, R_{ij}. Similarly to the standard model, we have

$$R_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij}),$$

given n_{ij}, the total number of subjects observed in arm j of trial i, and π_{ij}, the probability of an outcome in the respective arm and trial.

Let $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ia_i})'$ be the vector of binomial probabilities in trial i. Without loss of generality, assume that if trial i employed the reference treatment, it is in the arm labelled 1. Then $\text{logit}(\boldsymbol{\pi}_i)$, the vector of the log-odds of outcomes in the arms of trial i, can be modelled using

$$\text{logit}(\boldsymbol{\pi}_i) = \alpha_i \mathbf{1}_{1\times a_i} + \boldsymbol{V}_i^{(1)} \boldsymbol{d}^{(1)} + \boldsymbol{\epsilon}_i, \tag{2.2}$$

where α_i is the baseline response in trial i to the reference treatment (this differs from the baselines in the standard NMA model, which are relative to treatment t_{i1}); the vector $\mathbf{1}_{1\times a_i}$ is an a_i-length vector of ones; $\boldsymbol{d}^{(1)} = (d_{1,2}, \ldots, d_{1,C})'$ is a vector of relative effects of length $C-1$ for each component except for the reference treatment (since $d_{1,1} = 0$). $\boldsymbol{V}_i^{(1)}$ is a design matrix with dimension $a_i \times (C-1)$ where $V_{i,jk}^{(1)} = 1$ if arm j involves component $k+1$ and 0 otherwise. That is, if arm j employs only the reference treatment, the row sum of $\boldsymbol{V}_i^{(1)}$ is zero, otherwise the row sum is equal to the number of components employed in the treatment protocol of arm j. Multiplying the j^{th} row of $\boldsymbol{V}_i^{(1)}$ by $\boldsymbol{d}^{(1)}$ results in Equation (2.1), where A is the reference treatment and X is the treatment protocol in arm j of trial i.

Further, $\boldsymbol{\epsilon}_i$ are random errors representing between-trial variation. If trial i includes treatment 1, then $\boldsymbol{\epsilon}_i = (0, \epsilon_{i2}, \ldots, \epsilon_{ia_i})'$, and

$$\boldsymbol{\epsilon}_{i,2:a_i} \sim \text{MVN}_{a_i-1}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i-1 \times a_i-1}),$$

where $\boldsymbol{\Sigma}_{p \times p}$ has the same definition as in Section 1.1.1. Otherwise, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ia_i})'$ and $\boldsymbol{\epsilon}_i \sim \text{MVN}_{a_i}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i \times a_i})$. This model is a random effects model. To obtain a common-effect model, σ is set to zero, which is equivalent to omitting $\boldsymbol{\epsilon}_i$ from Equation (2.2).

Independent diffuse normal priors are used for the baselines α_i and the relative effects $\boldsymbol{d}^{(1)}$, such as $\text{N}(\mathbf{0}, 1000\boldsymbol{I})$. A flat uniform prior is used for σ, where the upper limit is chosen based on the scale of the data, similarly to the standard NMA model.

Estimation is performed by using MCMC sampling. A posterior sample is obtained, usually by using BUGS (Welton et al., 2009). A point estimate for $d_{1,k}$, $k = 2, \ldots, C$ and σ can then be obtained by taking the mean or the median of the posterior samples. The posterior distribution of multicomponent treatment effects can be obtained by applying the additivity Equation (2.1) to posterior samples. The posterior distribution for the relative effect of any two treatments k and l can then be determined by applying the consistency equation to the posterior samples of $d_{1,l}$ and $d_{1,k}$. Credible intervals can also be obtained using the appropriate quantiles of the posterior samples, for example, 2.5% and 97.5% for a 95% credible interval.

2.2.2 Frequentist CNMA Model of Rücker et al. (2019)

Rücker et al. (2019) proposed a frequentist approach to CNMA which takes contrast-based summary data. The R package netmeta implements this method and is freely available (Balduzzi et al., 2023). We introduce their common and random effects models using the unified notation which is summarized in Table 2.1, which differs from the notation used in introducing the standard frequentist NMA model in Section 1.1.2.

Similarly to the standard frequentist approach, this model uses all $m_i = \binom{a_i}{2}$ contrasts from each a_i-arm trial as data. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})'$ be the vector of m_i comparisons from trial i with associated standard errors SE_{ij}. The model is given by

$$\boldsymbol{y}_i \sim \text{MVN}_{m_i}(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i),$$

with

$$\boldsymbol{\mu}_i = \boldsymbol{U}_i \boldsymbol{V}_i \boldsymbol{d},$$

where $\boldsymbol{d}$ is a vector of length C which contains relative effects for each component 1 to C. $\boldsymbol{U}_i$ is a $m_i \times a_i$ design matrix which creates a vector of m_i contrasts from the a_i arms of trial i. In particular, $U_{i,jk} = 1$ if the k^{th} arm is the minuend of the j^{th} contrast, -1 if the k^{th} arm is the subtrahend of the j^{th} contrast, and 0 otherwise. $\boldsymbol{V}_i$ is a design matrix with dimension $a_i \times C$ where $V_{i,jk} = 1$ if arm j involves component k and 0 otherwise.

In this model, we use the notation $\boldsymbol{d}$ and $\boldsymbol{V}_i$ without a superscript to emphasize the fact that the structure of these model elements do not depend on a chosen reference treatment, and thus they involve all C components rather than $C - 1$ as in the Bayesian model from the previous section.

$\boldsymbol{\Omega}_i$ is a diagonal covariance matrix that reflects the standard errors in the data and, for random effects models, the degree of between-study heterogeneity estimated using a method-of-moments. Between-study heterogeneity can be omitted for a common-effect model. $\boldsymbol{\Omega}_i$ is analogous to $\boldsymbol{W}^{-1}$ from the standard frequentist model, but on the trial scale. The adjustments needed for multi-arm trials and random-effects follow the same procedure as for the standard model (Rücker et al., 2019).

To obtain a regression estimate of $\boldsymbol{d}$, the model can be rewritten by letting $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_I')'$, $\boldsymbol{\Omega}$ be a block-diagonal matrix with $(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I)$ along the diagonal, and

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{U}_1 \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{U}_I \boldsymbol{V}_I \end{pmatrix}. \tag{2.3}$$

We can then write

$$\boldsymbol{y} \sim \text{MVN}_M(\boldsymbol{X}\boldsymbol{d}, \boldsymbol{\Omega}), \tag{2.4}$$

where $\boldsymbol{y}$ and $\boldsymbol{X}$ are the same as those defined in Rücker et al. (2019). Weighted least squares is used to estimate $\boldsymbol{d}$, that is

$$\widehat{\boldsymbol{d}} = (\boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{X})^+ \boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{y}, \tag{2.5}$$

where $(\boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{X})^+$ is the pseudoinverse of $\boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{X}$ (Rücker et al., 2019; Rücker and Schwarzer, 2014). For some network structures, the matrix $\boldsymbol{X}^T \boldsymbol{\Omega}^{-1} \boldsymbol{X}$ is invertible, but in general the pseudoinverse is needed.

To estimate any treatment comparison, the appropriate contrast of relative effects in $\hat{\boldsymbol{d}}$ can be taken, by applying additivity (Equation (2.1)) and the consistency equation (Equation (1.1)). For example, if the relative effect of treatment $1 + 2$ compared to 3 was desired, it would be calculated as

$$\hat{d}_{3,1+2} = \hat{d}_1 + \hat{d}_2 - \hat{d}_3, \tag{2.6}$$

where $\hat{d}_k$ represents the k^{th} entry in the vector $\hat{\boldsymbol{d}}$.

Table 2.1: Summary of the model elements used in the new unified notation for additive CNMA models in this thesis.

Model Element	Role	Description
$\boldsymbol{U}_i$, $\boldsymbol{U}_i^*$	Design Matrix	Forms contrasts between arms - only for contrast-based data
$\boldsymbol{V}_i$, $\boldsymbol{V}_i^{(1)}$	Design Matrix	Forms component structure of treatment arms
α_i	Parameter	Represents trial-specific response to the anchor treatment - only for arm-level data
σ^2	Parameter	Represents between-trial heterogeneity in random effects models
$\boldsymbol{d}^{(1)}$	Parameter	Vector of population relative effects for components $2, \ldots, C$ where treatment 1 (reference) is the comparator - only for anchored model
$\boldsymbol{d}$	Parameter	Vector of relative effects for components $1, \ldots, C$ where the unknown anchor is the comparator - only for unanchored models
$\mathbf{1}_{1\times a_i}$	Structural Element	Vector of ones of length a_i - only for arm-level data
$\boldsymbol{\Sigma}_{p\times p}$	Structural Element	Square matrix of ones and 1/2's of dimension $p \times p$ which describes correlation structure - only for Bayesian random effect models
$\boldsymbol{y}_i$, $\boldsymbol{y}_i^*$	Data	Vector of contrast data - only for contrast-level data
R_{ij}, n_{ij}	Data	Binomial counts and sample sizes for j^{th} arm in trial i - only for arm-level data
$\boldsymbol{\Omega}_i$	Covariance Matrix	Covariance matrix incorporating uncertainty in data and from between-trial heterogeneity - only for frequentist model
$\boldsymbol{S}_i^*$	Covariance Matrix	Covariance matrix quantifying uncertainty in data - only for Bayesian contrast-based model
$*$	Notation	Indicates elements of the model which are of reduced dimension due to the inclusion of $a_i - 1$ contrasts instead of $\binom{a_i}{2}$ dependent contrasts
(1)	Notation	Indicates elements of the model which are of reduced dimension because the model is anchored, and therefore the element's structure depends on which treatment was chosen as reference

2.3 The Anchor in Additive CNMA

2.3.1 Role of the Anchor Treatment

In additive CNMA models, the assumption which defines multicomponent treatment relative effects is stated mathematically in Equation (2.1). We call the treatment A in Equation (2.1) the anchor treatment. Simply put, Equation (2.1) indicates that additivity is implemented on the relative effect scale, and specifically on the scale where effects are relative to the anchor treatment A. We introduce the new term "anchor treatment" because the role of this treatment is distinct from the reference treatment in an analysis.

The anchor treatment defines the relative effects on which additivity can be applied. Different forms of additivity are possible based on which treatment is the anchor. For example, if treatment A is the anchor treatment, then the population multicomponent relative effects are defined according to $d_{A,X} = \sum_{x \in X} d_{A,x}$ for all multicomponent treatments X. However, this does not imply that $d_{B,X}$ is equal to $\sum_{x \in X} d_{B,x}$. In fact, if treatment B was the anchor and thus multicomponent relative effects were determined according to $\sum_{x \in X} d_{B,x}$, this would represent a different set of population relative effects and thus a different data-generating mechanism. We can state this as a theorem.

Theorem 1. *If additivity holds for anchor treatments A and B simultaneously, then treatments A and B have indistinguishable effects.*

Proof. Suppose A and B are single component treatments and X is a multicomponent treatment in an arbitrary network. Let $|X| > 1$ denote the number of components making up the multicomponent treatment X. Assume that the additivity equation holds for both A and B as the anchor simultaneously, that is,

$$d_{A,X} = \sum_{x \in X} d_{A,x} \tag{2.7}$$

and

$$d_{B,X} = \sum_{x \in X} d_{B,x}. \tag{2.8}$$

Since we have consistency, we can write

$$\begin{aligned} d_{A,X} &= \sum_{x \in X} (d_{B,x} - d_{B,A}) \\ &= \sum_{x \in X} d_{B,x} - \sum_{x \in X} d_{B,A} \\ &= d_{B,X} - |X| d_{B,A}, \end{aligned}$$

where the last line is obtained by using Equation (2.8). Applying consistency to the left-hand side, we have

$$d_{B,X} - d_{B,A} = d_{B,X} - |X|d_{B,A},$$

which implies that $d_{B,A} = 0$, and thus treatments A and B must have indistinguishable effects. □

Theorem 1 shows that the anchor treatment defines the relative effects on which additivity can be applied, and that anchor treatments that have different effects correspond to different assumptions about multicomponent relative effects. This means that the parameter definitions, and thus their estimates, will differ depending on the anchor used to fit the model. Two analysts presented with the same network data could thus come to different conclusions about the relative effectiveness of treatments simply by using models with different anchors. This is important to emphasize as it has been overlooked so far in the literature. This is further exemplified in Appendix A.1.

A further point worth emphasizing is the distinction between the anchor treatment and the reference treatment used in a Bayesian NMA. It is difficult to differentiate between the two concepts because in the existing Bayesian CNMA model, we are restricted to using the same treatment as the anchor and the reference treatment (see Section 2.3.2). Despite this, they play distinct roles which is why we refer to them with different terms. It is true that using any treatment as the reference treatment is equivalent as long as consistency is assumed, that is,

$$d_{A,C+D} = d_{X,C+D} - d_{X,A} = d_{Y,C+D} - d_{Y,A}.$$

In contrast, from Theorem 1 we know that using different anchor treatments leads to different models for additivity in general, so different anchors are not equivalent. Thus we differentiate between their two roles: The reference treatment is simply the treatment we use as the comparator for the population relative effects in the model, which is arbitrary due to the consistency equations. The anchor treatment is the treatment which defines the relative effect scale on which additivity is applied and thus defines our assumptions about multicomponent relative effects. It happens that in some models, the same treatment may play both roles, but the roles are nevertheless distinct.

2.3.2 Anchored and Unanchored Models

We argue that the anchor treatment is chosen in two different ways in the models presented in Section 2.2. It is either specified by the modeller a priori or automatically selected by the model. The first method corresponds to assuming that a particular treatment in the

network is the true anchor, and we define this to be an anchored approach. The second method corresponds to assuming that additivity holds for some unknown true anchor which is determined during model fitting, and we define this as an unanchored approach. Due to a lack of discussion on this topic in the literature, it is unlikely that the true anchor treatment would be known. Further, an anchor specified in an analysis may not fit the data well, leading to a misspecified model which may introduce bias to the estimated relative treatment effects and rankings. Therefore, unanchored models are of particular interest.

To illustrate whether existing CNMA approaches are anchored or unanchored, consider a network with components placebo, treatment A, and treatment B. Let us examine the design matrices of our novel notation and the resulting parameters for a two-arm trial with placebo in the first arm and treatment A in the second arm. If we use the model from Section 2.2.1, we must choose which treatment to use as the reference treatment (treatment 1). If placebo is chosen as the reference treatment, the design matrix $\boldsymbol{V}_i^{(1)}$ is as follows:

$$\boldsymbol{V}_i^{(1)} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

and the resulting vector $\text{logit}(\boldsymbol{\pi}_i)$ is

$$\begin{pmatrix} \alpha_i \\ \alpha_i \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} d_{1,A} \\ d_{1,B} \end{pmatrix} = \begin{pmatrix} \alpha_i + 0 \\ \alpha_i + d_{1,A} \end{pmatrix}.$$

This implies that α_i is the response to the placebo (the reference treatment), and the elements of $\boldsymbol{d}^{(1)}$ are relative to placebo. In contrast, if treatment A was chosen as reference, we would obtain

$$\begin{pmatrix} \alpha_i + d_{1,Placebo} \\ \alpha_i + 0 \end{pmatrix},$$

showing that in general α_i is the response to the reference treatment and $\boldsymbol{d}^{(1)}$ is a vector of effects relative to the reference treatment. Since multicomponent treatment effects are computed as sums of the elements of $\boldsymbol{d}^{(1)}$, the treatment chosen as the reference is forced to be the anchor in the analysis. The structure of the design matrix in this model assumes that a particular treatment is the anchor, thus the model of Welton et al. (2009) is an anchored model.

On the other hand, consider using the model from Section 2.2.2. The design matrices are as follows:

$$\boldsymbol{U}_i = \begin{pmatrix} -1 & 1 \end{pmatrix}; \boldsymbol{V}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The resulting parameter contrast is

$$\boldsymbol{U}_i\boldsymbol{V}_i\boldsymbol{d} = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} d_{k,Placebo} \\ d_{k,A} \\ d_{k,B} \end{pmatrix} = d_{k,A} - d_{k,Placebo}.$$

In this model, multicomponent treatment effects are sums of the elements of $\boldsymbol{d}$, which are relative to the unspecified anchor treatment k. The unspecified anchor is selected based on what gives the best fit to the data. Therefore, the anchor treatment is estimated during model fitting and not specified by the modeller, making this an unanchored model.

In an unanchored model, the estimates of $\boldsymbol{d}$ can also be used to identify how similar the estimated anchor is to the treatments present in the network. We can examine the point estimates for the elements of $\boldsymbol{d}$ that are identifiable. If the estimate for treatment k is close to zero, that indicates that the anchor treatment is similar in effect to treatment k. This can provide insight into the true anchor in unanchored models.

A key difference between the anchored and unanchored models is in the dimension of the population relative effect parameter vector. In the anchored model, the vector $\boldsymbol{d}^{(1)}$ is of length $C-1$. In the unanchored model, the vector $\boldsymbol{d}$ is of length C. This is because in the anchored model, $\boldsymbol{d}^{(1)}$ is a vector of effects relative to treatment 1, and $d_{1,1} = 0$ by definition, so only the remaining $C-1$ relative effects need to be estimated. In contrast, in the unanchored model, the relative effects in $\boldsymbol{d}$ are not restricted to be relative to a particular treatment, so it is necessary to estimate a relative effect for all C components. Note that the elements of $\boldsymbol{d}$ are not uniquely identifiable in general, although some elements may be depending on the network structure. This is what necessitates the use of the pseudoinverse in Equation (2.5).

We emphasize that anchored models impose the anchor treatment, which may or may not result in an additivity assumption which fits the data well. From this perspective, anchored models have the potential to be biased due to misspecifying the anchor, and thus, anchored models should be used with caution. In addition to using the anchor which gives the best fit to the data, we remark that unanchored models are more flexible in other ways. First, the estimated anchor may not be a treatment in the network. Further, the estimated anchor may be zero or "no treatment", corresponding to additivity on the absolute effect scale, rather than a relative effect scale. To capitalise on the advantages of the unanchored approach, we introduce two Bayesian CNMA models that are unanchored in Section 2.4.

2.3.3 The Anchor and the Placebo Effect

We argued that unanchored models will use the anchor which fits the data best while in anchored models, the anchor is specified a priori. Intuitive choices for the anchor in an anchored model may be either "no treatment" or "placebo". In this section, we investigate the practical implications of assuming these anchors. We argue that although they may be conceptually meaningful, the interpretations rely on strong assumptions which we may not be able to check in practice.

To investigate the conceptual meaning of placebo or "no treatment" as anchors, we should consider the conceptual definition of the relative effect parameter $d_{k,l}$, which is the difference in outcome of treatment l compared to treatment k on average, or in other words,

$$d_{k,l} = \mathbf{E}(Y_l - Y_k) = \mathbf{E}(Y_l) - \mathbf{E}(Y_k), \tag{2.9}$$

where $\mathbf{E}(Y_l)$ is the average outcome observed for patients who receive treatment l, and treatments l and k could be single or multicomponent. Let $\varnothing$ and P represent "no treatment" and placebo, respectively. Nikolakopoulou et al. (2024) discussed concepts related to the placebo effect in CNMA, and we make use of their notation here. Let f represent the response to "no treatment", that is, $\mathbf{E}(Y_\varnothing) = f$. This can be viewed as the natural progression of the disease and other mechanisms not related to any treatment (Nikolakopoulou et al., 2024). Additionally, let π represent the effect of placebo compared to "no treatment". We assume that these effects are additive so that $\mathbf{E}(Y_\varnothing) = f$ and $\mathbf{E}(Y_P) = \pi + f$ (Nikolakopoulou et al., 2024). We also assume that these effects are constant and do not interact with any components, that is, $\mathbf{E}(Y_k) = f + \pi + d_{P,k}$ for any component k (Nikolakopoulou et al., 2024; Boussageon et al., 2022).

In Figure 2.1, Panel i), we visualize the average absolute effects of "no treatment" ($\varnothing$), placebo (P), treatment A (A), and treatment B (B) under the assumption described above. If the anchor treatment is placebo, we have

$$d_{P,A+B} = d_{P,A} + d_{P,B}, \tag{2.10}$$

and by rearranging Equation (2.9), we find that this implies

$$\mathbf{E}(Y_{A+B}) = \pi + f + d_{P,A} + d_{P,B} = \mathbf{E}(Y_P) + d_{P,A} + d_{P,B}, \tag{2.11}$$

illustrated in Panel ii) of Figure 2.1. This means that if the placebo is the anchor, the absolute effect of $A + B$ is the addition of the effect of A over placebo plus the effect of B over placebo plus the effect of placebo. That is, we get one placebo effect per *treatment*,

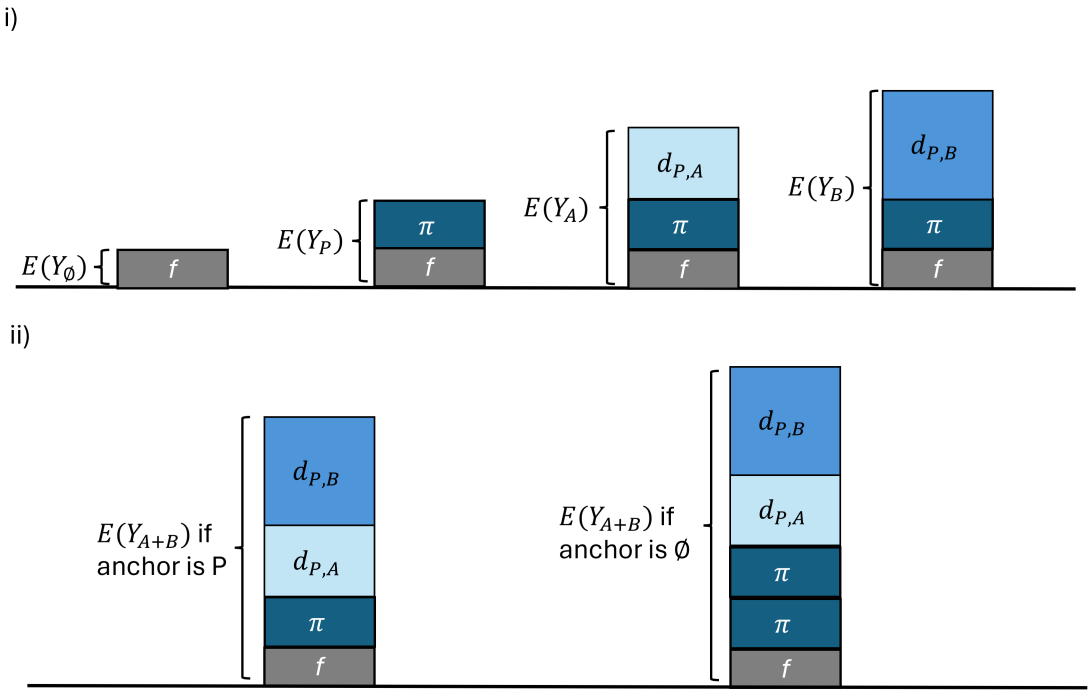


Figure 2.1: Illustration of the average absolute effects of different treatments. Panel i): "no treatment" ($\varnothing$), placebo (P), component A, and component B. Panel ii): treatment $A + B$ if the anchor is placebo, and $A + B$ if the anchor is "no treatment".

not per component. This makes sense if we do not expect each component to contribute its own placebo effect. For example, if all treatments (single and multicomponent) are given in the form of a topical cream, we may expect that the placebo effect is the same regardless of the number of active ingredients included in the cream.

Alternatively, consider "no treatment" as the anchor, that is,

$$d_{\varnothing,A+B} = d_{\varnothing,A} + d_{\varnothing,B}. \tag{2.12}$$

Plugging this in and rearranging Equation (2.9) suggests

$$\mathbf{E}(Y_{A+B}) = 2\pi + f + d_{P,A} + d_{P,B}. \tag{2.13}$$

This is illustrated in Panel ii) of Figure 2.1, and implies that the placebo effects for components are additive. That is, under the anchor $\varnothing$, we get one placebo effect per *component*. This is sensible when it is likely that each component contributes a new placebo effect. For

example, if each component is delivered as a separate pill, so that the treatment $A + B$ involves taking two pills instead of one, we may hypothesize that each pill could contribute its own placebo effect.

Both "no treatment" and "placebo" as the anchor have distinct conceptual meanings that may be useful in different settings. Placebo is a sensible anchor when we expect the placebo effect to be present for multicomponent treatments, and it does not change as more components are added. Alternatively, "no treatment" is a useful anchor when each component contributes its own placebo effect. It may be tempting to use an anchored model where we assume the anchor which best models our assumptions about the placebo effect in the problem at hand. Alternatively, we may want to fit an unanchored model and then use the estimates of $\boldsymbol{d}$ to provide insight into the behaviour of the placebo effect in a certain dataset, for example, if $\hat{d}_{k,P}$ is close to zero, we may want to conclude that the placebo effect is not additive for components in our dataset. However, it is not guaranteed that placebo, or particularly "no treatment", will be present as a node in the network, which may prevent us from using the desired anchor or from estimating the desired relative effects. Moreover, the interpretations derived above hinge on strong assumptions, discussed in further detail in the following paragraph.

While in theory these anchors have interpretations related to the placebo effect, they rely on strong assumptions which may not be met. First, the interpretations of Figure 2.1 are all subject to the assumption that the placebo effect and the specific treatment effects are additive and do not interact. It has been found that this assumption is often violated and it is rarely possible to be checked (Boussageon et al., 2022). Second, we assumed that the placebo effect is the same for all components, but placebo effects may differ between components and treatments, such as if some components are taken as a pill while some are given by an injection. If placebo effects differ between components, not only could this cause bias in the meta-analysis, but it also invalidates the interpretations in this section (Nikolakopoulou et al., 2024). Third, there may be scenarios where the true behaviour of the placebo effect is somewhere in between "one placebo effect per treatment" and "one placebo effect per component". For example, there may be additive effects for the first 2-3 components, but this effect may diminish as the number of components gets larger. In this case, there would be a different "true anchor" for different treatments, and either the placebo or "no treatment" would not work as a one-size-fits-all anchor. Since it is likely that one or more of these assumptions is violated, we caution against over-interpretation of the anchor treatment and its implications for the placebo effect.

To investigate the suitability of different anchors, an unanchored model can be fit and the relative effects of "no treatment" and placebo versus the anchor can be investigated to see where the anchor lies (if they are present and identifiable in the network). Although

the unanchored model will use the anchor which best fits the data without relying on some of the previous assumptions needed to justify the use of a specific treatment as the anchor, it still requires that there is one "true anchor" for all treatments, and so the results can still be misleading. Another approach would be to fit anchored models with various anchors to see how sensitive the conclusions are to the specified anchor. In all cases, the fit of the model should be assessed, such as by comparing the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002).

2.4 Novel Bayesian Unanchored CNMA Models

Due to the more flexible nature of unanchored models, practitioners may wish to avoid the use of an anchored CNMA model. Although the frequentist approach of Rücker et al. (2019) is already an option for practitioners to implement an unanchored additive model, a Bayesian approach may be preferred. Additionally, the model of Rücker et al. (2019) only accommodates contrast-level data. Thus, in sections 2.4.1 and 2.4.2 we introduce novel unanchored Bayesian additive CNMA models to handle contrast-level and arm-level summary data, respectively. Table 2.2 contains a summary of the CNMA models included in this thesis.

2.4.1 Novel Bayesian Unanchored Model for Contrast-level Data

We introduce a random effects model which follows a similar formulation to Rücker et al. (2019) in that it does not require the specification of an anchor for additivity and uses contrast-level data, however it uses a Bayesian framework. Additionally, in this model, we will use the same subset of the contrasts as the standard contrast-level Bayesian NMA model from Section 1.1.1, that is, for trial i with a_i arms, we use $a_i - 1$ contrasts which are all relative to the treatment in arm 1, t_{i1}. Due to the different subset of data being used, some elements in our model have reduced dimension which is indicated with superscript $*$ compared to similar elements in the frequentist version.

Let $\boldsymbol{y}_i^* = (y_{i2}^*, \ldots, y_{ia_i}^*)'$ be the $(a_i - 1) \times 1$ vector of independent contrasts relative to the treatment in arm 1 of study i for $i = 1, \ldots, I$. Let SE_{ij} be the standard error of the contrast associated with the jth treatment arm compared to arm 1 in the ith study, and let $SE_{t_{i1}}$ be the standard error of the raw response in arm 1 in study i. Then we have

$$\boldsymbol{y}_i^* \mid \boldsymbol{\mu}_i^*, \sigma^2 \sim \mathrm{N}(\boldsymbol{\mu}_i^*, \boldsymbol{S}_i^*), i = 1, \ldots, I, \tag{2.14}$$

where $\boldsymbol{S}_i^*$ has the same definition as $\boldsymbol{S}_i$ in Section 1.1.1, but we add the superscript $*$ to emphasize the fact that it is of dimension $(a_i - 1) \times (a_i - 1)$. Further,

$$\boldsymbol{\mu}_i^* = \boldsymbol{U}_i^* \boldsymbol{V}_i \boldsymbol{d} + \boldsymbol{\epsilon}_i^*, \tag{2.15}$$

where $\boldsymbol{U}_i^*$ is a $(a_i - 1) \times a_i$ design matrix, where $U_{i,jk}^* = 1$ if the k^{th} arm is the minuend of the j^{th} contrast, -1 if the k^{th} arm is the subtrahend in the j^{th} contrast, and 0 otherwise. The rows of $\boldsymbol{U}_i^*$ sum to zero and the first column of $\boldsymbol{U}_i^*$ contains all -1's. $\boldsymbol{V}_i$ is the same design matrix as introduced in Section 2.2.2. $\boldsymbol{\epsilon}_i^*$ represents between-study heterogeneity, and is distributed according to $\boldsymbol{\epsilon}_i^* \sim \text{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i - 1 \times a_i - 1})$. Here, we also note that the inclusion of multicomponent treatments does not impact the compound symmetry of treatments. That is, the correlation between the heterogeneity terms of any two treatments in the same trial, regardless of if they are single or multicomponent, is 1/2.

To finish the Bayesian treatment of the model, we place vague priors on $\boldsymbol{d}$ and σ. Similarly to the model in Section 2.2.2, in this model $\boldsymbol{d}$ represents a set of relative effects with an unknown reference point, corresponding to an unanchored model. We use priors similar to those used on the relative effects in the model described in Section 2.2.1, $\boldsymbol{d} \sim N(\boldsymbol{0}, 1000\boldsymbol{I})$, and a uniform prior on the heterogeneity parameter σ with a suitably large upper limit based on the scale of the data.

2.4.2 Novel Bayesian Unanchored Model for Arm-level Data

In order to accommodate arm-level data, we adapt the model introduced in Section 2.4.1. We assume that the arm-level response consists of binomial counts of an outcome of interest for each arm of each trial, similarly to the previous arm-based Bayesian models in this thesis. Recall that the binomial distribution family and logit link functions we use can be modified to accommodate other types of outcomes, such as event counts or continuous outcomes, following Dias et al. (2018). Let the observed binomial counts in the j^{th} arm of trial i be R_{ij}, and as before we have $R_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij})$. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ia_i})'$ be the vector of binomial probabilities in trial i. Then we model the binomial probabilities according to

$$\text{logit}(\boldsymbol{\pi}_i) = \alpha_i \boldsymbol{1}_{1 \times a_i} + \boldsymbol{V}_i \boldsymbol{d} + \boldsymbol{\epsilon}_i \tag{2.16}$$

where

$$\boldsymbol{\epsilon}_i \sim \text{MVN}_{a_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i \times a_i}),$$

and α_i, $\boldsymbol{1}_{1 \times a_i}$, $\boldsymbol{V}_i$, $\boldsymbol{d}$, σ and $\boldsymbol{\Sigma}_{a_i \times a_i}$ are as defined previously. We use the same priors as in Section 2.2.1. We can show that this model is equivalent to the contrast-based model in Section 2.4.1, with the exception that the contrast-based method requires a normality assumption when calculating the standard errors in the matrix $\boldsymbol{S}_i^*$ (see Appendix A.2).

Table 2.2: Summary table of the four CNMA models described in this chapter, labeled by the subsection in which they were introduced. Note that the definition of $\boldsymbol{\epsilon}_i$ for the model in Section 2.2.1 and its dimension is slightly altered if trial i includes the reference treatment.

	Existing models	
	2.2.1 (arm, anchored)	2.2.2 (contrast, unanchored)
Data	$R_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij})$	$\boldsymbol{y}_i \sim \text{MVN}_{m_i}(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i)$
Treatment effects	$\text{logit}(\boldsymbol{\pi}_i) = \alpha_i \mathbf{1}_{1\times a_i} + \boldsymbol{V}_i^{(1)} \boldsymbol{d}^{(1)} + \boldsymbol{\epsilon}_i$	$\boldsymbol{\mu}_i = \boldsymbol{U}_i \boldsymbol{V}_i \boldsymbol{d}$
Heterogeneity	$\boldsymbol{\epsilon}_i \sim \text{MVN}_{a_i}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i \times a_i})$	$\hat{\sigma}^2$, embedded in $\boldsymbol{\Omega}_i$
	New models	
	2.4.2 (arm, unanchored)	2.4.1 (contrast, unanchored)
Data	$R_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$	$\boldsymbol{y}_i^* \mid \boldsymbol{\mu}_i^*, \sigma^2 \sim \text{MVN}_{a_i-1}(\boldsymbol{\mu}_i^*, \boldsymbol{S}_i^*)$
Treatment effects	$\text{logit}(\boldsymbol{\pi}_i) = \alpha_i \mathbf{1}_{1\times a_i} + \boldsymbol{V}_i \boldsymbol{d} + \boldsymbol{\epsilon}_i$	$\boldsymbol{\mu}_i^* = \boldsymbol{U}_i^* \boldsymbol{V}_i \boldsymbol{d} + \boldsymbol{\epsilon}_i^*$
Heterogeneity	$\boldsymbol{\epsilon}_i \sim \text{MVN}_{a_i}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i \times a_i})$	$\boldsymbol{\epsilon}_i^* \sim \text{MVN}_{a_i-1}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_{a_i-1 \times a_i-1})$

2.5 Simulation Study

The simulations in this section are designed to illustrate the biases which can arise from misspecifying the anchor in an anchored model and demonstrate the performance and flexibility of the new unanchored models in Section 2.4. To this end, we consider eight different simulation scenarios (described in Section 2.5.1), and simulate 1000 arm-level datasets from each scenario using the model presented in Section 2.2.1 with the anchor treatment specific to the scenario. We choose this model to generate data because it allows for the generation of arm-level data and the specification of the true anchor in additivity.

Each dataset is analysed using the four CNMA models covered previously in this thesis chapter: Bayesian anchored CNMA model with arm-level data (referred to as Bayes Arm Anchored), Frequentist unanchored CNMA model with contrast-level data (Freq Contrast Unanchored), Bayesian unanchored CNMA model with contrast-level data (Bayes Contrast Unanchored), and Bayesian unanchored CNMA model with arm-level data (Bayes Arm Unanchored). For models Freq Contrast Unanchored and Bayes Contrast Unanchored, the arm-level data were converted to contrast-level data using the procedure described in Appendix A.3.

The priors specified for the Bayesian models were $\boldsymbol{d}$ or $\boldsymbol{d}^{(1)} \sim \text{N}(\mathbf{0}, 1000\boldsymbol{I})$ for the unanchored and anchored models respectively, $\sigma \sim \text{Unif}(0, 2)$, and $\alpha \sim \text{N}(0, 1000)$ for the arm-based models.

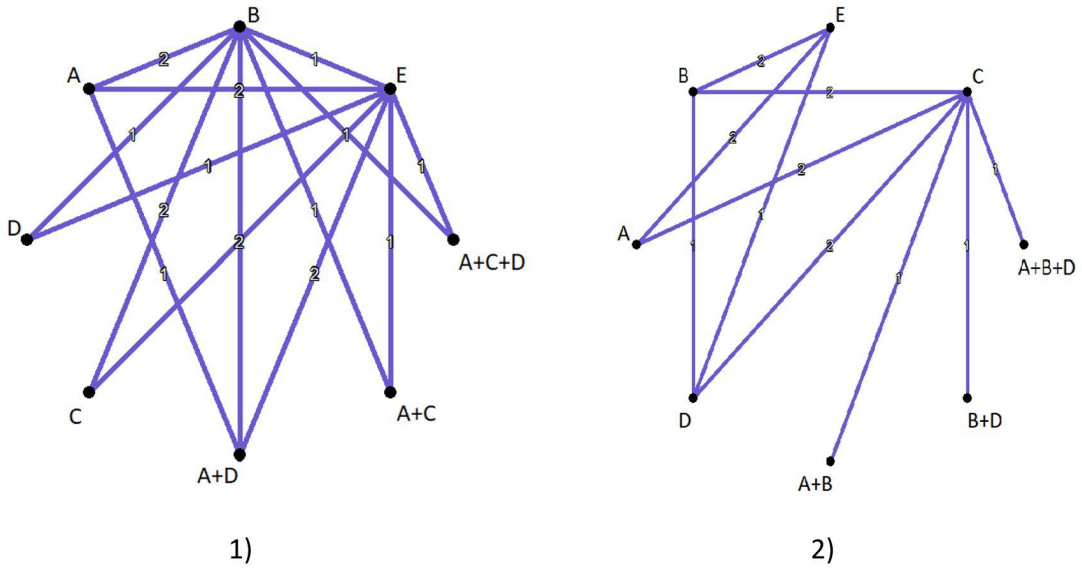


Figure 2.2: Network plots of the two scenarios considered in the simulation study. The number of trials making each comparison is written on the corresponding edge. Network 1) is shown on the left and Network 2) on the right.

Bayes Arm Anchored requires specifying an anchor treatment. In the simulations, the specified anchor is sometimes correctly specified and sometimes misspecified. This is indicated by including the data-generating, true anchor in the plot labels and the anchor specified in the anchored model in brackets in the model name.

2.5.1 Network Structure and Parameter Values

Two different network structures were used and are presented in Figure 2.2. Both network structures include five single-component interventions A, B, C, D, and E, and three multicomponent interventions. The treatments are labelled alphabetically according to decreasing effectiveness in each network structure, assuming a large response is good. All trials have two arms only. In each network, additivity was considered relative to two different anchor treatments. In each of the two networks and for each of the two additivity assumptions, two different amounts of data were considered: 50 and 500 patients per arm.

In the first network (referred to as network 1), the relative effects are set so that the

worst treatment (treatment E) is much less effective than the other four treatments. The value of the heterogeneity (σ) is set relatively low. The parameter values used in simulating data for this network are presented in the first three columns of Table 2.3. We consider two different ways to simulate the multicomponent data: one where the true anchor is the treatment E; the other specifying the true anchor to be a different treatment, B, which has a similar effect size to the remaining treatments A, C and D. The additivity assumption implied by using E as the anchor treatment is quite different from the additivity assumption implied by using treatment B as the anchor. This can be seen in the multicomponent treatment effects in Table 2.3 and is due to the fact that treatments E and B have quite different effect sizes. From this table, we can see that misspecifying the anchor as E when its true value is B could give poor results - our assumption (that the true anchor is E) expects the multicomponent treatment effects shown in the third column of Table 2.3, but the data-generating mechanism reflects the multicomponent treatment effects in the second column. Therefore, network 1 represents a scenario where we expect that misspecifying the anchor treatment will have a strong impact on model fit. This could correspond to a scenario where a set of trials compared several very effective drugs to a placebo treatment (which is not very effective), and the true anchor is one of the effective drugs but the data is analysed assuming the placebo is the anchor.

The second network (network 2) was designed to represent a more moderate scenario in terms of the consequences of misspecifying the anchor for additivity. The parameter values used in simulating data for this network are presented in the last three columns of Table 2.3. The heterogeneity in the network (σ) was increased compared to network 1. We consider two different ways to simulate the multicomponent data: one specifying the true anchor to be the moderately effective treatment C; the other specifying the true anchor to be the worst treatment, E. The absolute relative effect of those two treatments is smaller (0.5) compared to the absolute difference of treatments E and B in network 1 (0.9) and there are fewer direct and indirect trials providing data on the relative effect. Network 2 represents a less severe scenario in terms of misspecifying the anchor treatment because the two additivity assumptions in network 2 give more similar multicomponent treatment effects than the two additivity assumptions used in network 1. For details on how the values in Table 2.3 were determined and how they were used to simulate the data, see Appendix A.3 and A.4.

2.5.2 Performance Measures

For the Bayesian models, the posterior mean was used as the point estimate. To quantify the bias in each model, we calculate the difference between the mean of the estimated value

Table 2.3: Parameter values used in the simulation for Network 1 and Network 2. Values of the parameters under different anchors are also given. The value of the single component relative effects and the heterogeneity σ are the same for different anchors. The multicomponent relative effects and the baseline responses are different for different anchor treatments.

Network 1			Network 2		
Parameter	Anchor B	Anchor E	Parameter	Anchor C	Anchor E
$d_{E,A}$	1.2	1.2	$d_{C,A}$	0.70	0.70
$d_{E,B}$	0.9	0.9	$d_{C,B}$	0.35	0.35
$d_{E,C}$	0.8	0.8	$d_{C,C}$	0.00	0.00
$d_{E,D}$	0.7	0.7	$d_{C,D}$	-0.20	-0.20
$d_{E,E}$	0.0	0.0	$d_{C,E}$	-0.50	-0.50
$d_{E,A+C}$	1.1	2.0	$d_{C,A+B}$	1.05	1.55
$d_{E,A+D}$	1.0	1.9	$d_{C,B+D}$	0.15	0.65
$d_{E,A+C+D}$	0.9	2.7	$d_{C,A+B+D}$	0.90	1.85
α	0.05	-0.85	α	-0.60	-1.10
σ	0.1	0.1	σ	0.40	0.40

and the true value used in data generation for each parameter. Coverage probability was calculated for 95% confidence and credible intervals (CIs) for the frequentist and Bayesian models respectively. Additionally, length of the CIs was calculated.

To assess the ability of the models to give accurate treatment hierarchies, we computed the average SUCRA value for each treatment in the Bayesian models or the average P-score for each treatment in the frequentist model. SUCRA values were calculated using the dmetar R package and P-scores were calculated using the netmeta R package.

Methods to assess model fit in terms of additivity and choice of anchor are of interest. We investigate the use of DIC to assess and compare model fit between anchored and unanchored models. In particular, we compare the anchored arm-level model 2.2.1 to the unanchored arm-level model 2.4.2 using the DIC. Let $\Delta = DIC(\text{Model 2.4.2}) - DIC(\text{Model 2.2.1})$. Then negative values of Δ indicate better fit for the unanchored model compared to the anchored model. Differences in DIC of two or more are generally considered meaningful (Spiegelhalter et al., 2002).

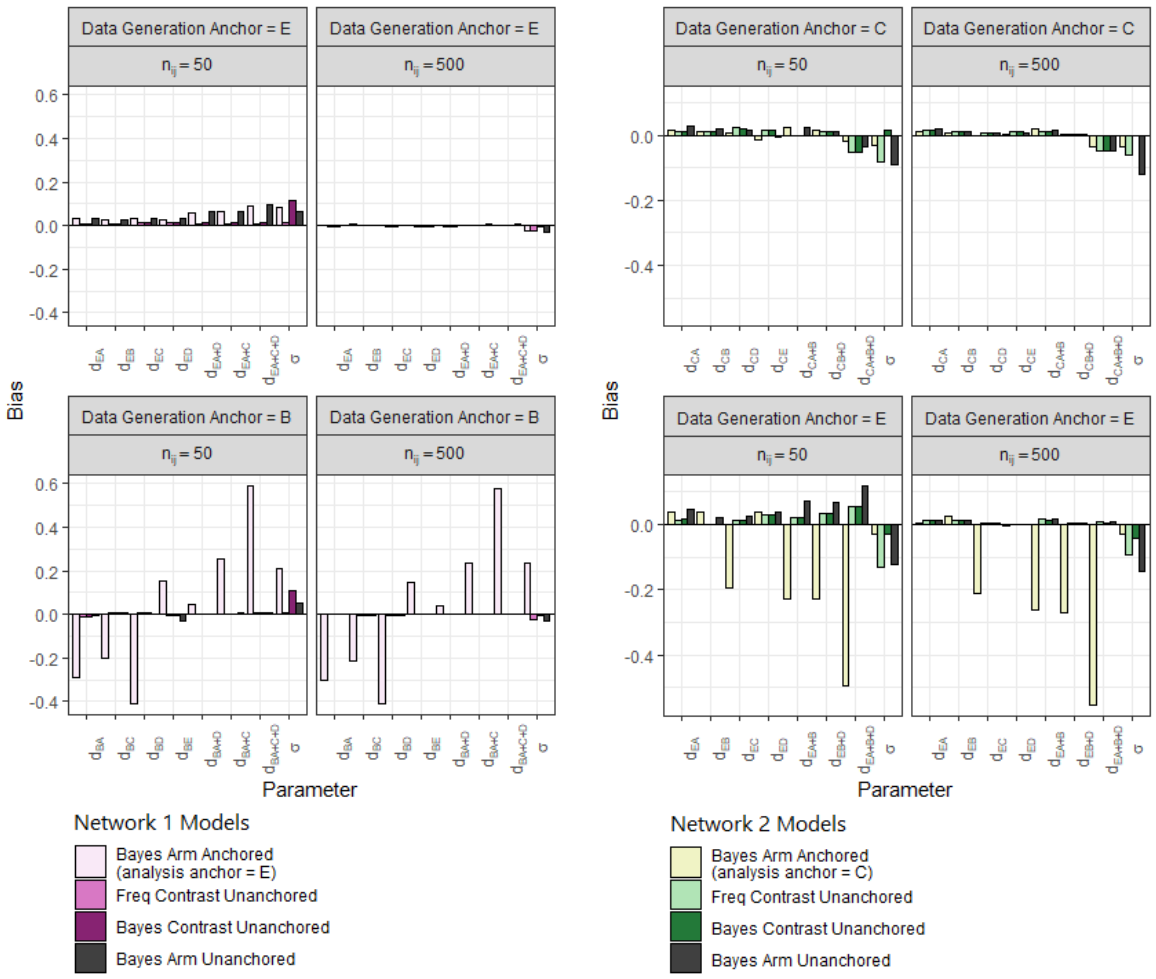


Figure 2.3: Bias of estimates of relative effect parameters and heterogeneity for Network 1 (left, pink) and Network 2 (right, green). The relative effect parameters shown correspond to treatment effects relative to the treatment used to generate the data for that scenario.

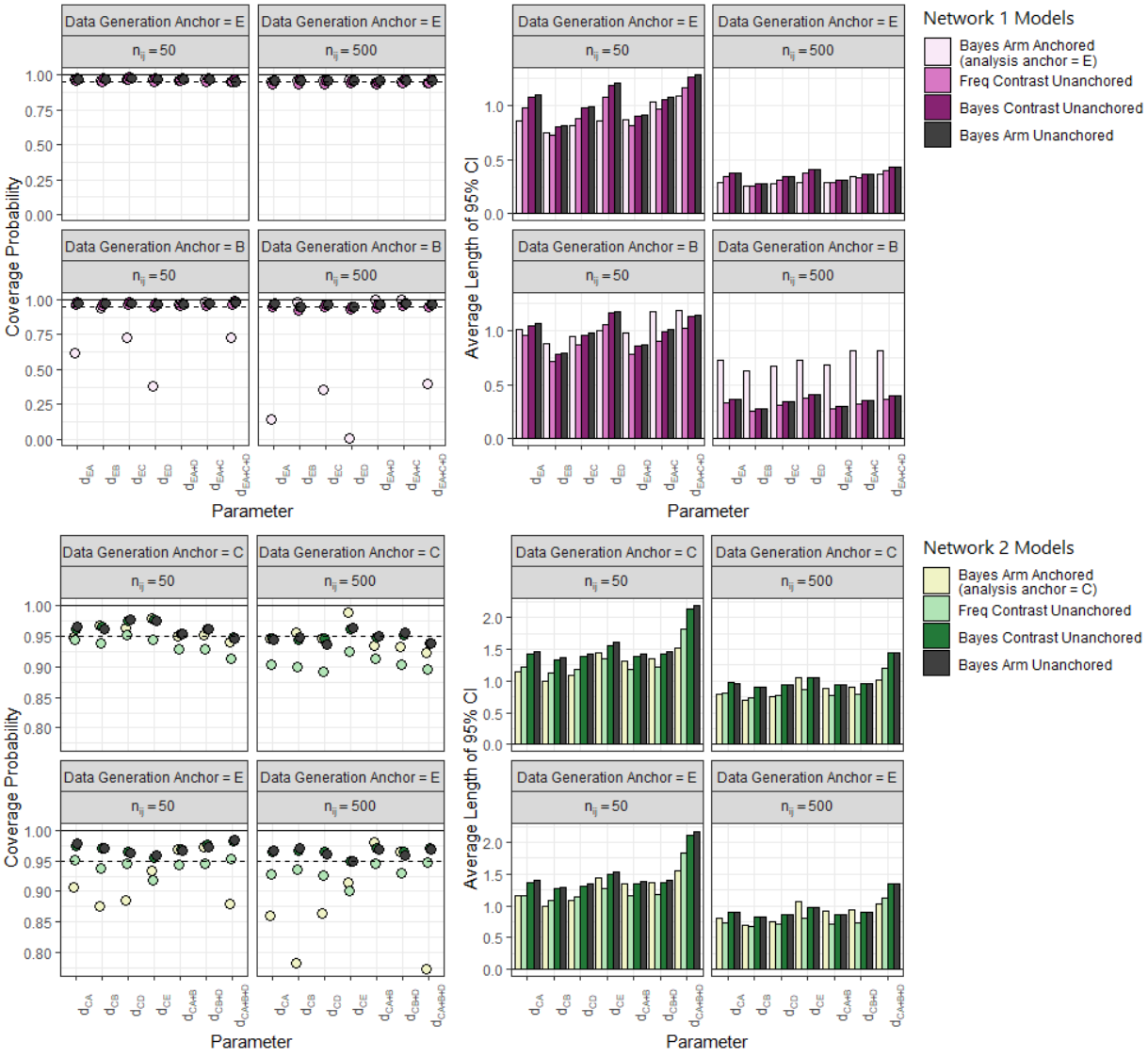


Figure 2.4: Coverage probability and average length of 95% credible or confidence intervals for the relative effect parameters monitored in the simulation for network 1 (top row, pink) and network 2 (bottom row, green).

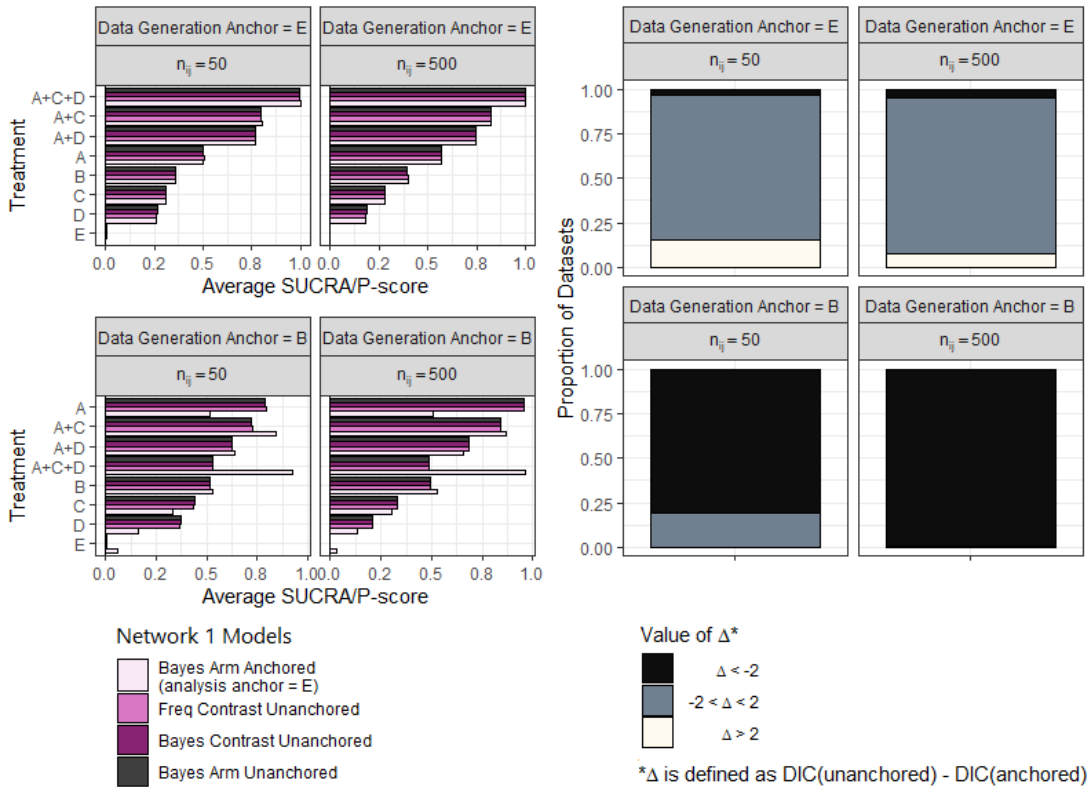


Figure 2.5: Network 1 average rankings based on SUCRA/P-scores and percentage of datasets where the value of $\Delta = DIC(\text{Model 2.4.2}) - DIC(\text{Model 2.2.1})$ is less than -2, between -2 and 2, and greater than 2. Negative values of Δ indicate that the unanchored model has a better fit. Differences of DIC greater than 2 are considered meaningful (Spiegelhalter et al., 2002).

2.5.3 Results: Network 1

The biases for relative effect parameters and heterogeneity are shown for network 1 on the left of Figure 2.3. When the anchor used in data generation is E, the bias for the relative effects is small for all models and sample sizes. When the treatment B is used as the anchor for data generation, the bias for the three unanchored models remains small, however the Bayes Arm Anchored (analysis anchor = E) model has much larger magnitude biases, regardless of the sample size. The bias for σ decreases as the sample size increases for the Bayesian models, but it increases for the frequentist model. Overall, the bias is always relatively low for the unanchored models, but when the anchor is misspecified, the anchored model shows severe bias. Here we also note that due to the assumption of consistency, if the set of relative effect estimates with one treatment as the comparator are unbiased, then the sets of relative effect estimates with any other treatment as the comparator will also be unbiased. Similarly, if the set of relative effect estimates with one treatment as the comparator are biased, then the sets of relative effect estimates with any other treatment as the comparator will also likely be biased.

The coverage probability and average length of 95% CIs for network 1 are shown on the top row of Figure 2.4. For the unanchored models, the coverage probability is close to the nominal level of 0.95 for all relative effect parameters, sample sizes, and data generation anchors. However, when the analysis anchor does not match the data generation anchor, the anchored model performs poorly. In that scenario, $d_{E,A}$, $d_{E,C}$, $d_{E,D}$ and $d_{E,A+C+D}$ all have coverage probabilities well below the nominal level of 0.95 which decrease as the sample size is increased, falling to zero for $d_{E,D}$. Interestingly, the single component treatments whose relative effects have poor coverage are the components involved in the multicomponent treatments.

Turning to the average lengths of the CIs, when the data generation anchor is E, models Bayes Arm Anchored (analysis anchor = E) and Freq Contrast Unanchored have the shortest CIs and the lengths are shorter for the larger sample size. When the data generation anchor is B, Bayes Arm Anchored (analysis anchor = E) model tends to have longer CIs, particularly for the larger sample size compared to the other models.

The average SUCRA or P-score value for each treatment in the different scenarios is shown on the left of Figure 2.5. The order of the treatments on the Y-axis indicates the true treatment hierarchy, where the treatment with the largest effect is at the top and the smallest effect is at the bottom. The true order changes depending on which treatment was used as the anchor in generating the data. When treatment E is the true anchor, all four models produce accurate treatment hierarchies on average based on the SUCRA and P-scores. However, when treatment B is the true anchor, the Bayes Arm Anchored

(analysis anchor = E) model results in average SUCRA scores which reflect a different hierarchy of the treatments than the truth. In particular, the treatment hierarchy implied by the average SUCRA scores from this model is the same as the treatment hierarchy which is true when E is the true anchor.

A breakdown of the values of Δ for each scenario comparing the DIC of the unanchored and anchored Bayesian arm-level models is given in the right panel of Figure 2.5. When the true anchor is E, in the majority of the models, neither the anchored or unanchored models have a meaningfully better fit than the other model, for both sample sizes. When the true anchor is B, the unanchored model tends to be preferred by the DIC the majority of the time. The Bayes Arm Anchored (analysis anchor = E) model is recommended only 0.03% of the time when $n_{ij} = 50$ and is never recommended when $n_{ij} = 500$.

2.5.4 Results: Network 2

The bias of estimates in network 2 is shown on the right side of Figure 2.3. The biases for the relative effect parameters show similar patterns to network 1. In particular, the anchored model, Bayes Arm Anchored (analysis anchor = C), shows large biases when the analysis anchor is misspecified. For both data generation anchors, the magnitudes of the bias for the heterogeneity are larger for the Freq Contrast Unanchored and Bayes Arm Unanchored models compared to the other two models.

The coverage probability and average length of 95% CIs for relative effects are shown in the bottom row of Figure 2.4. When the data generation anchor is C, the coverage probabilities are close to the nominal level for all models and sample sizes, although the Freq Contrast Unanchored model tends to fall slightly below 0.95 and the Bayesian models slightly above 0.95. When the data generation anchor is E, the Bayes Arm Anchored (analysis anchor = C) model has coverage probabilities which fall below the nominal level. The lengths of CIs tend to be shortest for models Bayes Arm Anchored (analysis anchor = C) and Freq Contrast Unanchored when the data generation anchor is C. As the sample size increases from 50 to 500, the average lengths of the CIs decrease for all models. A similar pattern is seen when the data generation anchor is E.

The left panel of Figure 2.6 shows the average SUCRA or P-score values for each treatment in the different scenarios. Similar results are seen for both sample sizes for both anchor scenarios. When the true anchor is C, all four models give average SUCRA/P-scores that correspond to the true treatment hierarchy. The Bayes Arm Anchored (analysis anchor = C) tends to give higher average SUCRA scores to the best treatments ($A + B$, $A+B+D$) and lower scores to the worse treatments (C, D, E) than the other three models.

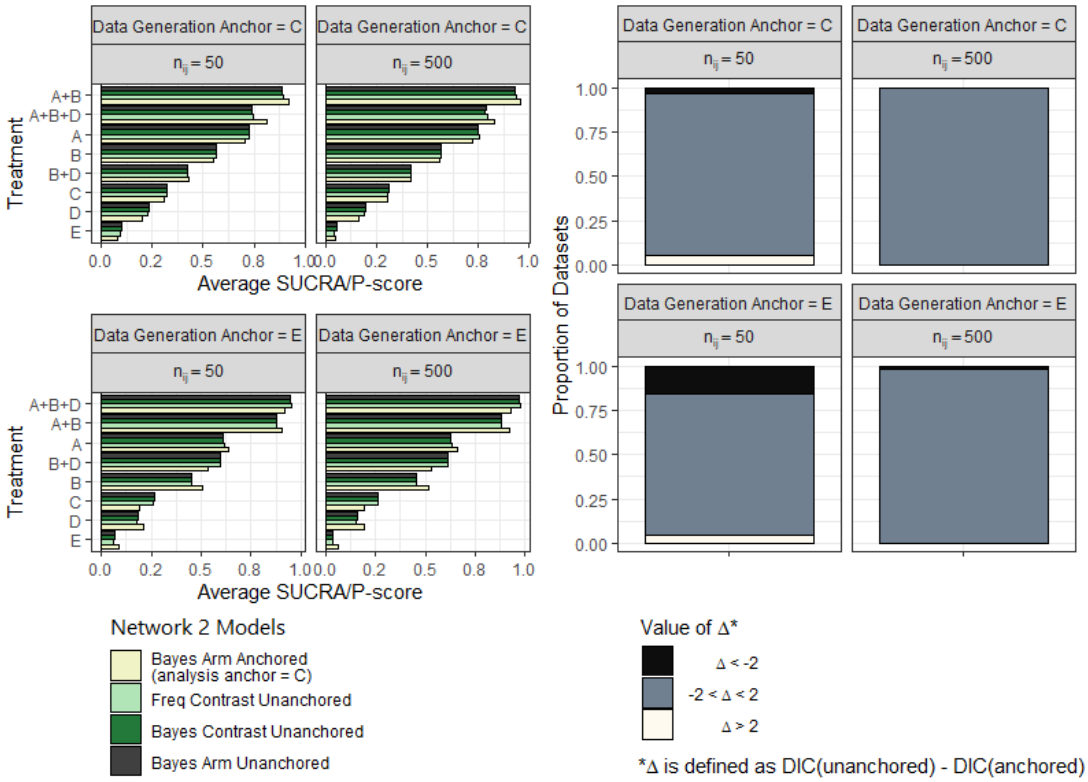


Figure 2.6: Network 2 average rankings based on SUCRA/P-scores and percentage of datasets where the value of $\Delta = DIC(\text{Model 2.4.2}) - DIC(\text{Model 2.2.1})$ is less than -2, between -2 and 2, and greater than 2. Negative values of Δ indicate that the unanchored model has a better fit. Differences of DIC greater than 2 are considered meaningful (Spiegelhalter et al., 2002).

When the true anchor is E, all models give average SUCRA/P-scores corresponding to the correct hierarchy, with the exception of Bayes Arm Anchored (analysis anchor = C), which incorrectly ranks D slightly higher than C on average. The right panel of Figure 2.6 shows a breakdown of the values of Δ for the different anchor and sample size scenarios for network 2. In all scenarios, the most common outcome is that the DIC does not recommend one model over the other.

2.6 Application

2.6.1 Dataset and Methods

To illustrate the CNMA models discussed in this chapter, we use the dataset analysed in Welton et al. (2009) which is freely available from http://www.bristol.ac.uk/population-health-sciences/centres/cresyda/mpes/code/. The data regards clinical trials of psychological interventions for adults with coronary heart disease (CHD). They include the data for a variety of outcomes; we will focus on trials which report all-cause mortality, a binary outcome. Welton et al. (2009) classified the psychological interventions employed in the clinical trials as being educational (Edu), behavioural (Beh), cognitive (Cog), relaxation (Rel), supportive (Sup), or some multicomponent combination of these, for example, educational and relaxation (Edu + Rel). In all trials, the psychological interventions were compared to a control arm which employed usual care. Several of the trials recorded zero events, which requires adjustment in the frequentist model.

For illustration purposes, we show the design matrices for the multi-arm trial numbered 23 for models 2.4.1 and 2.4.2. This trial has three treatment arms. In the first arm, a psychological intervention composed of educational, cognitive, and relaxation components was investigated. In the second arm, a psychological intervention composed of educational and relaxation components was used. The third arm used usual care only. For the contrast-based unanchored model introduced in Section 2.4.1, we require two design matrices: $\boldsymbol{U}_i^*$ and $\boldsymbol{V}_i$. $\boldsymbol{U}_i^*$ forms the contrasts between the arms, where the columns represent arms and the rows represent contrasts. Recall that in this model, $a_i - 1$ contrasts are used. Assuming that the two contrasts used are arm two versus arm one and arm three versus arm one, $\boldsymbol{U}_i^*$ for this trial is given by

$$\boldsymbol{U}_i^* = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The design matrix $\boldsymbol{V}_i$ shows the component structure of the arms, and is given by

$$\boldsymbol{V}_i = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

In this matrix, the columns correspond to the components Usual, Edu, Beh, Cog, Rel, and Sup. Model 2.4.2 uses the same $\boldsymbol{V}_i$ as model 2.4.1.

Random-effects additive CNMA models were fit to the CHD dataset, including the Bayesian anchored model, Bayesian unanchored models (arm and contrast-level data), and the frequentist unanchored model. The Bayesian models were fit to the data in R using JAGS for MCMC sampling via the runjags R package (Denwood, 2016). The models ran two parallel chains and at least 10,000 burn-in iterations followed by at least 40,000 samples from the posterior. Convergence was assessed by inspecting trace plots. The R package netmeta was used to fit the random-effects additive CNMA model introduced in Section 2.2.2 (Balduzzi et al., 2023). An adjustment for zero event counts was performed by setting the arguments `incr = 0.5` and `allstudies = T` to the function `pairwise` before performing the CNMA analysis using netmeta.

For the purposes of comparing the anchored and unanchored model fit, DIC was calculated for the Bayes Arm Anchored and Bayes Arm Unanchored model using JAGS. Although it is typically recommended that the DIC be calculated manually for hierarchical models, no leverage can be manually calculated for study arms which experience zero events, and thus the DIC cannot be calculated manually for this data example (Dias et al., 2018). The DIC calculated automatically by JAGS is thus reported.

2.6.2 Results

The results for the relative effects of the psychological components compared to usual care are shown in Figure 2.7 in a forest plot. We note that due to the way the data was coded in the original model used by Welton et al. (2009), the multi-arm correction was not correctly applied to trials with three arms. By recoding the data so that usual care is implemented in the first arm of each study, this issue is corrected. Our results presented here therefore differ slightly from those in the original paper. The most important finding from this plot is that while all three unanchored models find no statistically significant differences between psychological interventions and usual care as the CIs include zero, the anchored model finds that behavioural interventions significantly reduce all-cause mortality when compared with usual care. Additionally, the point estimates from the anchored model

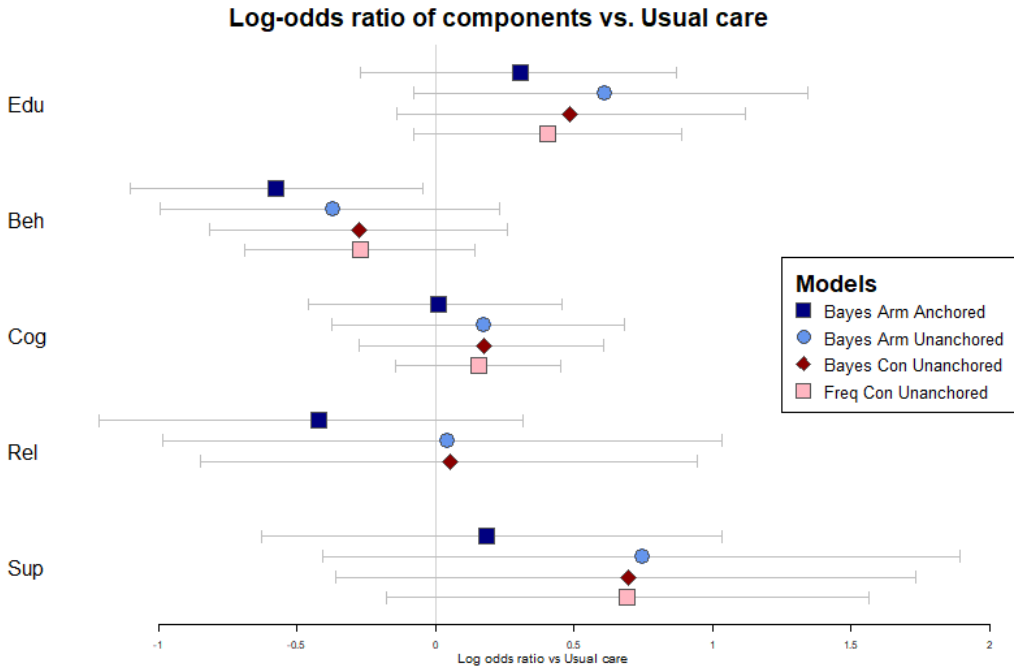


Figure 2.7: Log-odds ratios of the five psychological components versus usual care and 95% confidence/credible intervals for the four additive CNMA models. Note that a point estimate and confidence interval is missing from the frequentist model for the relaxation component due to this not being readily available from the R package netmeta.

tend to be shifted downwards compared to those from the unanchored models, while those from the unanchored models are closer to each other.

Although the conclusions from all three unanchored models are similar, some differences remain. Firstly, the frequentist model gives narrower confidence intervals than the Bayesian models' credible intervals. This could be due to the fact that the Bayesian models account for the uncertainty in the estimation of heterogeneity while the frequentist model treats the estimated heterogeneity parameter as fixed and known. Secondly, the Bayesian arm-level model 2.4.2 gives longer credible intervals than either of the other unanchored models and gives slightly different point estimates than the other unanchored models, particularly for educational and behavioural components. One possible explanation for this is that the contrast-based models make a normal approximation in calculating the standard errors of the contrasts, whereas the arm-level methods do not. This could result in loss

of information in the data, particularly for studies for whom the normal approximation is poor, such as those with small sample sizes.

We compare the DIC for the two Bayesian models which use arm-level data. The anchored and unanchored models had a DIC of 363.4 and 363.5 respectively. From this, there is no evidence that one model has a better fit over the other. Overall, all three unanchored models give similar results and we note that the anchored model finds a significant effect of behavioural components versus usual care whereas the unanchored models do not find any significant differences relative to usual care.

Table 2.4: Point estimates and 95% CIs for the elements of $\boldsymbol{d}$ from the Bayesian unanchored arm-based model.

Component	Posterior Mean of $d_{Anchor,k}$ (95% CI)
Usual care	-0.535 (-1.335, 0.243)
Educational	0.053 (-0.587, 0.743)
Behavioural	-0.910 (-1.635, -0.219)
Cognitive	-0.350 (-1.062, 0.331)
Supportive	-0.532 (-1.333, 0.205)
Relaxation	0.172 (-0.682, 0.988)

In Table 2.4, we illustrate how the estimates of $\boldsymbol{d}$ from an unanchored model can tell us about the estimated anchor. The table shows the point estimate and 95% CI for the relative effect of all components versus the anchor from the Bayes Arm Unanchored model. The component with the point estimate the closest to zero is Educational with a value of 0.053, indicating that the anchor which best fits the data has a very similar effect to the educational component on average. However, CIs indicate we do not have statistical evidence to differentiate the anchor from Usual care, Educational, Cognitive, Supportive or Relaxation.

2.7 Discussion and Future Work

In this chapter, we introduced the notion of the anchor treatment and anchored and unanchored CNMA models. We introduced unified notation and clarified the nature of the additivity assumption in two existing CNMA models. We also presented arguments for why unanchored models are more robust than anchored models. We provided two novel Bayesian unanchored additive CNMA models to capitalize on the advantages of a Bayesian

approach. A simulation study was completed to demonstrate the performance of the unanchored and anchored additive CNMA models under different scenarios in terms of relative effects, true anchor, and sample sizes. A case study was presented where all additive CNMA models were demonstrated on a real dataset.

Anchored models such as Welton et al. (2009)'s can lead to serious problems if the anchor is misspecified. In the simulation study, we found that if the analysis anchor does not match the true anchor used to generate data, important biases are introduced and coverage probability can be poor. We also showed that the biases introduced can lead to inaccurate treatment hierarchies based on SUCRA values. Anchored models should be used with caution, while unanchored models such as that of Rücker et al. (2019) and the Bayesian models introduced in this chapter are preferable. However, in the simulation study when the analysis anchor was correctly specified, the anchored model can present narrower CIs than the unanchored models. Thus, practitioners may still wish to use an anchored model if they are confident in their choice of the anchor treatment.

Given that anchored models can provide benefits when the analysis anchor agrees with the data generating mechanism, but can lead to serious problems if it is misspecified, a method to test the suitability of a specific treatment as the anchor in the anchored model may still be desirable. Welton et al. (2009) use the DIC to choose between standard NMA and CNMA models. In our simulation study, we explored the use of DIC for detecting an unsuitable anchor by comparing the anchored model to an unanchored model. However, we found that in network 2, despite notable biases and poor coverage probability of the misspecified anchor model, the DIC did not reliably recommend the unanchored model over the anchored model. An alternative approach would be to look only at the posterior mean deviance, which is not penalized for complexity, rather than the DIC, to assess whether the anchored model is appropriate. No frequentist anchored CNMA model has been developed at this time, so no tests are available to assess the anchor choice if frequentist models are used. Development of methods to assess the suitability of the chosen anchor in anchored additive CNMA models is a possible avenue of future work.

We restricted our investigation to additive CNMA models in this thesis. The additivity assumptions made in the two existing CNMA models discussed in this chapter can be relaxed by including interaction terms (Welton et al., 2009; Rücker et al., 2019). It is not known how the interpretation of the relative effect parameters change when using an anchored interaction model vs. an unanchored interaction model. Expanding our understanding of anchored and unanchored models in settings beyond additivity is an interesting area of future work.

Finally, there is potential for further development of the unanchored Bayesian CNMA

models proposed in this chapter. We could incorporate interaction terms to relax additivity, along with further developing the interpretation of unanchored interaction models. In addition, the Bayesian models could be applied to disconnected networks as described in Rücker et al. (2019) and demonstrated in Rücker et al. (2021). A potential benefit of the Bayesian CNMA framework is that the additivity assumption could be relaxed by introducing structure in the priors for the relative effects of multicomponent treatments. This approach would not require interaction parameters to be estimated, which is attractive in situations where data are sparse, such as in disconnected networks.

Chapter 3

Quantifying Certainty in Treatment Hierarchies from Network Meta-Analysis

3.1 Introduction

A hierarchy of treatments is a useful output of an NMA where treatments are ordered from most to least preferred. A treatment hierarchy is often the desired output of a systematic review (Papakonstantinou et al., 2022). To avoid over-interpretation, treatment hierarchies must be considered alongside the underlying uncertainty stemming from the value and precision of the estimated relative effects. For example, consider two hypothetical sets of SUCRA values. Set 1 gives treatments A, B, C, and D SUCRA values of 0.411, 0.472, 0.530, and 0.586, while Set 2 gives SUCRA values of 0.005, 0.334, 0.667, and 0.994. Both sets of SUCRA values would produce the same treatment hierarchy with A the least preferred and D the most preferred. However, the interpretations of the different sets of SUCRAs are very different. The differences in SUCRAs between treatments are much larger in Set 2 than in Set 1. The posterior distribution of the relative treatment effects versus B which produced the SUCRA values for Sets 1 and 2 are shown in Figure 3.1. The posterior distributions which produced Set 1 have similar centres and have a lot of overlap with each other and with zero (representing equivalence with treatment B). From a statistical perspective, the posteriors on the left show little evidence of a difference in effects between any of the treatments. On the other hand, the posteriors which produced Set 2 have much less overlap with each other and with zero. In this scenario, we have more confidence

that D produces a greater effect than C, and so on. We can therefore interpret Set 1 as producing a hierarchy with low certainty, whereas Set 2 produces a hierarchy with high certainty.

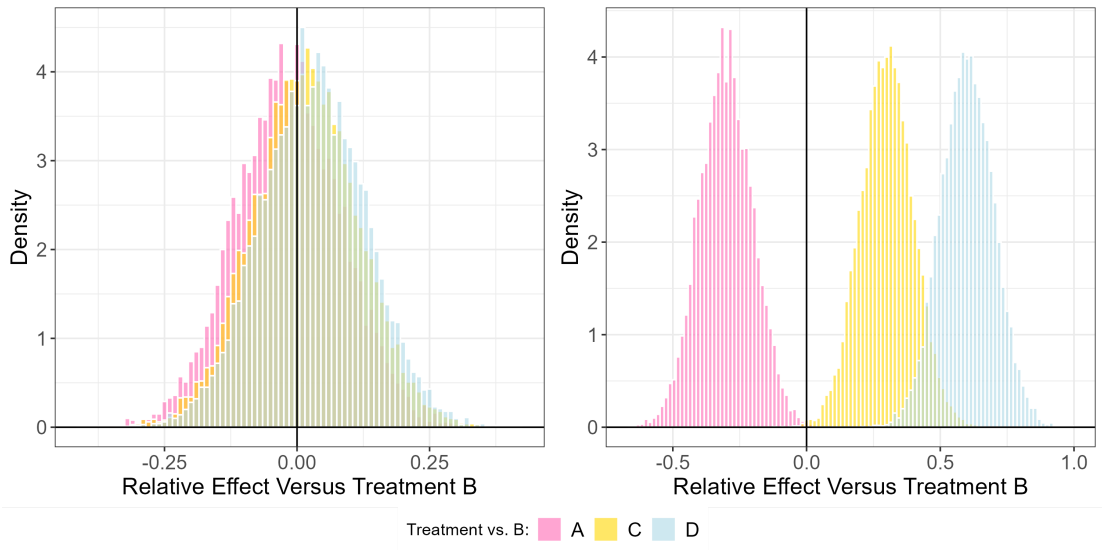


Figure 3.1: The hypothetical posterior distributions for the relative effect of treatments A, C, and D compared to treatment B that created Sets 1 (left, SUCRA(A) = 0.411, SUCRA(B) = 0.472, SUCRA(C) = 0.530, SUCRA(D) = 0.586) and Set 2 (right, SUCRA(A) = 0.005, SUCRA(B) = 0.334, SUCRA(C) = 0.667, and SUCRA(D) = 0.994).

It is necessary to differentiate between the uncertainty in creating a treatment hierarchy and the uncertainty in the ranking metrics used to produce the hierarchy, such as the SUCRAs or the P-scores. The uncertainty in SUCRAs or P-scores was discussed by Veroniki et al. (2018). The only uncertainty in SUCRA(k) comes from the ranking probabilities p_{ku}, which are usually assumed to be known. The P-score can be interpreted as an average of p-values, and uncertainty in p-values is rarely considered (Veroniki et al., 2018). On the other hand, the uncertainty in a given treatment hierarchy is determined by the uncertainty and overlap of estimated relative effects. Moreover, we emphasize that in this work we focus on the uncertainty in the treatment hierarchy without considering the uncertainty in estimating the ranking metrics.

Approaches to quantify the uncertainty in producing a treatment hierarchy can broadly be classified as operating at the treatment or hierarchy level. Treatment-level approaches quantify the uncertainty in the rank of one treatment at a time. In a network with N

treatments, a treatment-level approach produces N measures of uncertainty, one specific to each treatment. Confidence intervals (CIs) for SUCRA values or predictive P-scores are examples of treatment-level approaches (Rosenberger et al., 2021). Current approaches to calculating CIs for SUCRA values are restricted to taking discrete values and provide little insight, ranging from zero to one, whereas CIs for predictive P-scores can take continuous values (Veroniki et al., 2018; Wu et al., 2021; Rosenberger et al., 2021). SUCRA and predictive P-score CIs have only been implemented in a Bayesian setting. More recently, it was proposed to use the normalized entropy to quantify uncertainty in each treatment's place in a hierarchy (Wu et al., 2021). In a given network, there will be N SUCRA CIs, predictive P-score CIs, or normalized entropy metrics, each of which may express a different level of certainty. Therefore, treatment-level approaches are not adequate to summarise the overall uncertainty in creating a treatment hierarchy.

Hierarchy-level approaches to quantifying uncertainty have also been suggested. Salanti et al. (2014) provided procedures for evaluating the confidence in a hierarchy in the domains defined by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. These include integrating information from the contribution matrix and visualizing plots of the ranking probabilities for each treatment (rankograms) (Salanti et al., 2014). The methods are not tied to a particular ranking metric. Papakonstantinou et al. (2022) suggest approaches for quantifying uncertainty in hierarchies developed from clinically relevant criteria as an alternative to SUCRAs or P-scores. All suggested methods can be subjective or depend on the number of treatments in the network. Finally, since the normalized entropy was proposed as a treatment-level approach, the average normalized entropy among the treatments in the network was used to quantify uncertainty in the treatment hierarchy in an analysis of published NMAs (Wu and Tu, 2022). However, the properties of the average normalized entropy have not yet been explored in depth. No single hierarchy-level approach to quantifying uncertainty has yet been broadly adopted by NMA practitioners.

In this chapter, we introduce the Precision of Treatment Hierarchy (POTH), a measure of certainty in producing a treatment hierarchy. POTH is derived from SUCRAs or P-scores and operates at the hierarchy-level, summarising the uncertainty in a hierarchy into a single value between zero and one. In Section 3.2, we define scenarios which have maximum and minimum certainty. We introduce the variance of SUCRAs and use this to derive POTH. We also show how POTH is related to the average variance of the treatments' rank probability distributions. We show how POTH can be calculated for subsets of treatments of interest, and discuss two useful applications. In Section 3.3, we calculate the POTH in a database of NMAs and investigate associations between the POTH and various network characteristics. In Section 3.4, we provide a worked example to demonstrate the usefulness

and interpretation of POTH. An R package **poth** and code to reproduce the database analysis and examples are available (Wigle and Schwarzer, 2024). We conclude with a discussion of our findings and future work in Section 3.5.

3.2 Methodology

3.2.1 Maximum and Minimum Certainty in a Hierarchy

First, let us clarify the interpretation of ranks and ranking probabilities p_{ku}, which are used to calculate SUCRA(k) in Equation (1.6). In a Bayesian framework, the p_{ku} are approximated by sampling from the posterior distributions of the relative effects. In a frequentist framework, they are approximated by resampling from the sampling distribution of the relative effects' estimates. A treatment's rank in a given draw from the posterior or sampling distribution is determined by the value of all treatments' relative effects and whether the outcome is beneficial or harmful. In each draw, treatment k will assume one of the $u = 1, 2, \ldots, N$ possible ranks. The rank of treatment k, rank(k), is a random variable. The ranking probability p_{ku} is the probability that treatment k takes the u^{th} rank in a random draw from the posterior. Together, p_{ku} for $u = 1, \ldots, N$ forms the discrete probability distribution for rank(k), as shown in Table 3.1. We treat these probabilities as perfectly known.

Table 3.1: The probability mass function for rank(k).

Rank	1	2	...	N
Probability	p_{k1}	p_{k2}	...	p_{kN}

Consider a situation where we are completely certain about the hierarchy of the treatments. In general, this situation is represented by the configuration of ranking probabilities p_{ku} shown on the left in Table 3.2. We are completely certain about the hierarchy in this case because the probability of observing the ranking of (1, 2, ..., N) in a random draw from the posterior or sampling distribution is one. Note that this situation is unlikely to occur in practice because the order of the relative effects would have to be identical in every sample drawn from the posterior or sampling distribution. In this hypothetical scenario, the SUCRA values are evenly spaced between zero and one by increments of $1/(N-1)$. This is related to the fact that SUCRA(k) can be expressed in terms of the expected rank of treatment k, $\mathbf{E}\left[\text{rank}(k)\right]$ (Rücker and Schwarzer, 2015), which is calculated using the

probability distribution in Table 3.1 as

$$\mathbf{E}\left[\text{rank}(k)\right] = \sum_{u=1}^{N} u p_{ku},$$

and the SUCRA value of treatment k can be expressed as

$$\text{SUCRA}(k) = \frac{N - \mathbf{E}\left[\text{rank}(k)\right]}{N-1}. \tag{3.1}$$

The mean of the SUCRA values in any given network is 0.5 and the mean of the expected ranks in any given network is $(N+1)/2$ (Rücker and Schwarzer, 2015).

On the other extreme, consider the situation shown on the right in Table 3.2. In this scenario, each treatment takes rank u with probability $1/N$ and all SUCRA values are equal to 0.5. This represents a highly uncertain scenario because we have no evidence to suggest ranking any treatment over any others. Other configurations of ranking probabilities are possible which will result in all SUCRA values equal to 0.5. In general, when all SUCRA values are 0.5, any increase in the certainty of the place that one treatment takes in the hierarchy is offset by a decrease in the certainty of another treatment's place in the hierarchy. An in-depth explanation is available in Appendix B.1. We therefore define the minimum certainty in a hierarchy as any set of ranking probabilities which results in all SUCRA values equal to 0.5.

In the most certain scenario, the SUCRA values are spread out along the interval $[0, 1]$, whereas in the least certain scenario, the SUCRA values are all equal to their mean. This suggests that deviation of SUCRA values (or expected ranks) from their mean is an indicator of the certainty in a hierarchy.

3.2.2 The Variance of SUCRA Values and Expected Ranks

To investigate the deviation of SUCRAs from their mean, let us consider the variance of the SUCRA values, or equivalently, the scaled variance of the expected ranks. Let

$$S^2(N) = \frac{1}{N} \sum_{k=1}^{N} \left[\text{SUCRA}(k) - 0.5\right]^2 \tag{3.2}$$

be the variance of the SUCRA values, which can also be expressed in terms of the variance of the expected ranks using Equation (3.1) to give

$$S^2(N) = \frac{1}{N(N-1)^2} \sum_{k=1}^{N} \left\{ \mathbf{E}\left[\text{rank}(k)\right] - \frac{N+1}{2} \right\}^2. \tag{3.3}$$

Table 3.2: Tables of ranking probabilities corresponding to the most certainty in a hierarchy (left) and illustrating the least certainty in a hierarchy (right). The values in the tables correspond to the probability of treatment k taking rank u, p_{ku}. The bottom rows show the SUCRA values and expected ranks for each treatment.

Rank (u)	Treatment (k)			
	1	2	...	N
1	1	0	...	0
2	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	0	0	...	1
SUCRA	1	$(N-2)/(N-1)$	...	0
$\mathbf{E}\left[\text{rank}(k)\right]$	1	2	...	N

Rank (u)	Treatment (k)			
	1	2	...	N
1	$1/N$	$1/N$	...	$1/N$
2	$1/N$	$1/N$	...	$1/N$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	$1/N$	$1/N$	...	$1/N$
SUCRA	0.5	0.5	...	0.5
$\mathbf{E}\left[\text{rank}(k)\right]$	$(N+1)/2$	$(N+1)/2$	...	$(N+1)/2$

Here, we interpret $S^2(N)$ as a measure of variability of SUCRAs between treatments. It is clear that when there is minimum certainty in the hierarchy, $S^2(N) = 0$ because all SUCRA values (or expected ranks) are equal to their mean. In other words, minimum certainty in the hierarchy is identified when the variance of the SUCRA values is zero. It can also be shown that maximum certainty in the treatment hierarchy coincides with maximum variance of the SUCRA values by considering the distribution of each treatment's rank (see Sec. 3.2.4).

Let $S^2_{\max}(N)$ be the upper bound of $S^2(N)$, that is, $S^2_{\max}(N)$ is the value of $S^2(N)$ evaluated when the certainty in the hierarchy is maximized and the number of treatments is equal to N. Using the fact that in the maximum certainty scenario the SUCRA values form a discrete uniform distribution on $[0, 1/(N-1), 2/(N-1), \ldots, 1]$, the maximum variance as a function of N is

$$S^2_{\max}(N) = \frac{(N+1)}{12(N-1)}.$$

As the number of treatments approaches infinity, this converges to the variance of the continuous uniform distribution from zero to one, which equals 1/12.

3.2.3 The Precision of Treatment Hierarchy (POTH)

The upper bound of $S^2(N)$ is $S^2_{\max}(N)$, which depends on the number of treatments in the network. This makes $S^2(N)$ alone difficult to interpret. Therefore, let the Precision of

Treatment Hierarchy (POTH) be defined as

$$\begin{aligned}\text{POTH}(N) = \text{POTH} &= \frac{S^2(N)}{S^2_{\max}(N)} \\ &= \frac{12(N-1)}{(N+1)} S^2(N). \end{aligned} \tag{3.4}$$

Since $S^2(N)$ is a variance, POTH is non-negative. Thus, we have that

$$0 \leq \text{POTH} = \frac{S^2(N)}{S^2_{\max}(N)} \leq 1,$$

with equality to zero when there is minimum certainty in the hierarchy and equality to one when there is maximum certainty in the hierarchy. The POTH is therefore a measure of the degree of certainty in a hierarchy, with higher POTH corresponding to higher certainty. Furthermore, the bounds of POTH do not depend on the number of treatments in the network, facilitating its interpretation.

Note that the key properties of SUCRA values which facilitate the derivation of POTH, namely, a mean of 0.5, are also possessed by frequentist P-scores. Thus, POTH can be calculated using the variance of P-scores by simply replacing SUCRA(k) with P(k) in Equation (3.2). Additionally, knowledge of the ranking probabilities p_{ku} is not needed to compute POTH as long as one of the SUCRA values, P-scores, or relative differences and standard errors are available.

Interpreting POTH

POTH represents the extent of certainty in a treatment hierarchy. While a generic categorisation of POTH values may not always be appropriate, we tentatively assign POTH values less than 0.5 as low certainty and above 0.8 as high certainty based on the relationship between POTH and the proportion of statistically significant treatment differences in the database analysis presented in Section 3.3 and early simulation results reported in Appendix B.6 which found that POTH values below 0.5 are very common in networks with no treatment differences.

To further aid in the interpretation of an observed POTH value, we suggest comparing it to the distribution of POTH that could be generated in a similar network but where all observed treatment differences are due to chance, or where treatment differences observed between active treatments are due to chance. In particular, the reference distribution

is obtained by repeatedly sampling from a network with the same structure, standard errors, and heterogeneity as our network of interest, but where the true relative effects of all (active) treatments are zero. The observed POTH value can then be compared to a histogram of the reference distribution to gain an understanding of how our observed POTH value compares to typical values seen in the reference scenario.

This distribution can be used to calculate the probability of obtaining a POTH value that is at least as large as we observed given that there are no true treatment differences (or no treatment differences between active treatments). If X% of the reference distribution has larger POTH than we observed, then there is an X% chance that a treatment hierarchy obtained under the reference scenario will be more certain than the treatment hierarchy we observed. In summary, POTH describes the extent of certainty in the hierarchy, and the probability describes how likely we are to have a higher level of certainty in the reference scenario. In Appendix B.2, we describe a simulation-based approach to obtaining a relevant reference distribution and corresponding probability. This approach is implemented in the function 'refdist' of R package **poth**.

3.2.4 POTH and the Probability Distribution of Ranks

We have shown that POTH = 0 uniquely corresponds to the scenario with minimum certainty in the hierarchy. In this section we show that POTH can be calculated from the average variance of the probability distributions of rank(k) for $k = 1, \ldots, N$ and prove that POTH = 1 corresponds to the maximum certainty in hierarchy scenario.

Intuitively, the variance of the distributions of individual ranks are inversely linked to the certainty of a hierarchy. When there is high certainty in the hierarchy, most of the probability will be concentrated at one particular rank for each treatment, leading to distributions of individual ranks with low variances. On the other hand, if the hierarchy is very uncertain, the distributions of individual ranks will be flat with larger variances. Assuming the probabilities p_{ku} are exactly known, the variance of rank(k) is given by

$$\text{Variance}\left[\text{rank}(k)\right] = \sum_{u=1}^{N} \left\{u - \mathbf{E}\left[\text{rank}(k)\right]\right\}^2 p_{ku}.$$

Since we are interested in the uncertainty in the whole hierarchy, we can look at the average of the variances for all observed treatments,

$$V(N) = \frac{1}{N} \sum_{k=1}^{N} \sum_{u=1}^{N} \left\{u - \mathbf{E}\left[\text{rank}(k)\right]\right\}^2 p_{ku}.$$

Under maximum certainty in the hierarchy, $p_{ku} = 1$ for $u = \mathbf{E}[\text{rank}(k)]$ and $p_{ku} = 0$ for all $u \neq \mathbf{E}[\text{rank}(k)]$, for $k = 1, \ldots, N$. $V(N)$ is zero under these conditions, and in Appendix B.3, we show that this is a unique minimum. Therefore, $V(N)$ is uniquely minimized when certainty in the hierarchy is maximized.

Further, we show that POTH can be expressed in terms of $V(N)$. We can find the relationship between $V(N)$ and $S^2(N)$ using a weighted sum of squares decomposition as follows:

$$\underbrace{\sum_{u=1}^{N}\left(u - \frac{N+1}{2}\right)^2}_{\frac{N(N+1)(N-1)}{12}} = \underbrace{\sum_{k=1}^{N}\sum_{u=1}^{N}\left\{u - \mathbf{E}[\text{rank}(k)]\right\}^2 p_{ku}}_{NV(N)} + \underbrace{\sum_{k=1}^{N}\sum_{u=1}^{N}\left\{\mathbf{E}[\text{rank}(k)] - \frac{N+1}{2}\right\}^2 p_{ku}}_{N(N-1)^2 S^2(N)}, \tag{3.5}$$

noting that the cross term $2\sum_{k=1}^{N}\sum_{u=1}^{N}(u - \mathbf{E}[\text{rank}(k)])\left(\mathbf{E}[\text{rank}(k)] - \frac{N+1}{2}\right)p_{ku}$ is equal to zero. Then by dividing Equation (3.5) by $N(N+1)(N-1)/12$, we conclude that

$$\text{POTH} = \frac{12(N-1)}{(N+1)}S^2(N) = 1 - \frac{12}{(N+1)(N-1)}V(N), \tag{3.6}$$

or in other words, POTH can be derived from $V(N)$ and is equivalent to POTH from the variance of SUCRAs or expected ranks. Since we have shown that $V(N)$ is uniquely minimized when there is maximum certainty in the hierarchy, we also have that POTH and $S^2(N)$ are uniquely maximized when there is maximum certainty in the hierarchy.

3.2.5 POTH with Groups of Equivalent Treatments

To exemplify POTH, consider a hypothetical network with N treatments, where there are G equally sized groups of equivalent treatments, and G is a factor of N. The largest possible number of groups is $N/2$. For example, for $G = 2$ and even N, we have a group of $N/2$ equivalently "good" treatments and a group of $N/2$ equivalently "bad" treatments. Table 3.3 shows the ranking probabilities for this case. In this section, we calculate POTH for arbitrary G.

The difference between the expected ranks in group p and q is equal to

$$|p - q|\frac{N}{G},$$

which means the difference in SUCRAs between treatments in groups p and q is equal to

$$|p - q|\left(\frac{N}{G(N-1)}\right). \tag{3.7}$$

Table 3.3: Table of ranking probabilities corresponding to the hypothetical scenario where half of the treatments perform equally well and half equally poorly ($G = 2$). The values in the table correspond to p_{ku}, the probability of treatment k taking rank u.

Rank (u)	Treatment (k)					
	1	...	$N/2$	$N/2+1$	...	N
1	$2/N$	...	$2/N$	0	...	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	0	...	$\vdots$
$N/2$	$2/N$	...	$2/N$	0	...	0
$N/2+1$	0	...	0	$2/N$	...	$2/N$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$2/N$	...	$\vdots$
N	0	...	0	$2/N$	...	$2/N$
SUCRA	$(3N-2)/(4(N-1))$			$(N-2)/(4(N-1))$		

To calculate POTH for arbitrary G, we use the fact that the variance can be written as a sum of pairwise differences, that is, we can write

$$\begin{aligned} S^2(N) &= \frac{1}{N}\frac{1}{2N}\sum_{k=1}^{N}\sum_{l=1}^{N}\left[\text{SUCRA}(k) - \text{SUCRA}(l)\right]^2 \\ &= \frac{1}{N}\frac{1}{2N}\left(\frac{N}{G}\right)^2\sum_{p=1}^{G}\sum_{q=1}^{G}(p-q)^2\left(\frac{N}{G(N-1)}\right)^2 \\ &= \frac{G^2-1}{G^2}\frac{N^4}{12(N-1)^2}. \end{aligned}$$

The POTH for a network with G equally-sized groups of treatments is therefore

$$\text{POTH} = \frac{12(N-1)}{(N+1)}S^2(N) = \frac{G^2-1}{G^2}\frac{N^2}{N^2-1}. \tag{3.8}$$

For example, for $G = 2$ groups of size $\frac{N}{2}$, the POTH is equal to

$$\text{POTH} = \frac{3N^2}{4(N^2-1)}. \tag{3.9}$$

If we fix the number of groups and increase the number of treatments, the value of POTH goes to $(G^2-1)/G^2$. For the case of $G = 2$ groups, as N approaches infinity,

the POTH approaches 0.75 from above. This is because the certainty about the rankings within groups decreases as the number of treatments in the groups increase, but we always have certainty about the relative ranking of the two groups.

If we fix the number of treatments N and increase the number of groups G to its maximum value of $N/2$, the POTH becomes $(N^2-4)/N^2$. If we increase N and keep the maximum value of groups, the POTH goes to 1. As N increases, the number of distinguishable hierarchy positions (groups) increases, leading to higher certainty in the hierarchy.

3.2.6 Deriving POTH for Subsets of Treatments

Suppose we want to know the certainty in the hierarchy of an arbitrary subset of treatments in a network. For example, clinicians might be interested in the hierarchy of the top five treatments (five treatments with the highest SUCRA values) or the hierarchy of all treatments which are non-surgical. Notice that SUCRA values or P-scores can be re-calculated using subsets of the treatments observed in a network. The interpretation of a SUCRA value or P-score calculated using a subset of the observed treatments is the degree of certainty that a treatment beats other competing treatments, where "competing treatments" means the subset of treatments considered rather than all treatments in the network.

To re-calculate SUCRA values, the ranking probabilities p_{ku} must be re-calculated. Let $\boldsymbol{T}$ represent a subset of all treatments observed in a network. Let $p_{ku}^{\boldsymbol{T}}$ be the probability of treatment k taking rank u out of the treatments in $\boldsymbol{T}$. In a setting where the rank probabilities are calculated from samples of relative effects for treatments $k = 1, ..., N$, the probabilities $p_{ku}^{\boldsymbol{T}}$ can be calculated by ignoring the column of effect estimates for all treatments not in $\boldsymbol{T}$. Then the SUCRA value for treatment k when the set of competing treatments is defined as $\boldsymbol{T}$ is given by

$$\text{SUCRA}_{\boldsymbol{T}}(k) = \frac{1}{N_{\boldsymbol{T}} - 1} \sum_{r=1}^{N_{\boldsymbol{T}}-1} \sum_{u=1}^{r} p_{ku}^{\boldsymbol{T}}, \tag{3.10}$$

where $N_{\boldsymbol{T}}$ is the number of treatments in $\boldsymbol{T}$.

P-scores can be recalculated by simply omitting treatments not of interest from the sum in Equation (1.7), to give

$$\text{P}_{\boldsymbol{T}}(k) = \frac{1}{N_{\boldsymbol{T}} - 1} \sum_{\substack{l \in \boldsymbol{T} \\ l \neq k}} \Phi\left(\frac{\hat{d}_{l,k}}{SE(\hat{d}_{l,k})}\right). \tag{3.11}$$

Both $\text{SUCRA}_{\boldsymbol{T}}(k)$ and $\text{P}_{\boldsymbol{T}}(k)$ maintain a mean of 0.5.

We define the subset POTH of the set $\boldsymbol{T}$ as

$$\text{POTH}_{\boldsymbol{T}} = \frac{12(N_{\boldsymbol{T}} - 1)}{(N_{\boldsymbol{T}} + 1)} S^2_{\boldsymbol{T}}(N_{\boldsymbol{T}}), \tag{3.12}$$

where

$$S^2_{\boldsymbol{T}}(N_{\boldsymbol{T}}) = \frac{1}{N_{\boldsymbol{T}}} \sum_{k \in \boldsymbol{T}} (\text{SUCRA}_{\boldsymbol{T}}(k) - 0.5)^2.$$

$S^2_{\boldsymbol{T}}(N_{\boldsymbol{T}})$ could also be calculated with $\text{P}_{\boldsymbol{T}}(k)$. The interpretation of $\text{POTH}_{\boldsymbol{T}}$ is the certainty in the treatment hierarchy for the treatments included in $\boldsymbol{T}$. Subset POTH is always specific to the subset $\boldsymbol{T}$ and multiple $\text{POTH}_{\boldsymbol{T}}$ can be calculated for a given network as long as there are more than two treatments in the network. $\text{POTH}_{\boldsymbol{T}}$ can be higher, lower, or equal to POTH in a given network. In the next two subsections, we describe two useful applications of $\text{POTH}_{\boldsymbol{T}}$.

We note here that subset POTH is not the same as fitting a new NMA which only includes the treatments of interest. For example, suppose we have a network composed of A:B, A:C, A:D, and B:D trials. The subset POTH for the set $\boldsymbol{T} = \{A, B, C\}$ still leverages the information from the A:D and B:D studies, whereas re-fitting the NMA to only the A:B and A:C trials would lead to a loss of information. The difference between the subset POTH for $\boldsymbol{T}$ and the POTH from the NMA of only the A:B and A:C studies will depend on the information offered by the A:D and B:D trials. For example, if the A:D and B:D trials allow the treatment relative effects to be estimated more precisely, then $\text{POTH}_{\boldsymbol{T}}$ will be higher than POTH for the network that excludes those trials.

Quantifying Treatment Contributions to POTH

A given value of POTH can be achieved in multiple ways, except for the maximum certainty scenario. For example, a POTH of 0.7 could be observed when the $N-1$ treatments have a similar effect which is considerably different than that of the N-th treatment, or, on the other extreme, the N treatments have effects which are moderately different from each other (see example in Figure 3.2). To discern which of the two situations (or any other in-between scenario) is true, it is important to gauge the graphical presentation of the effects and observe the overlap in treatment effects. We also introduce here a numerical tool to complement the graphical approach: the leave-one-out POTH residual.

Let $\boldsymbol{T}^{-j}$ be the set of all treatments in the network except treatment j, for $j = 1, \ldots, N$. Then subset POTH can be calculated to give a set of N leave-one-out POTH values, each

one representing the precision in the hierarchy when treatment j is left out. We introduce the POTH residual for treatment j,

$$r_{\text{POTH}}(j) = \text{POTH} - \text{POTH}_{\boldsymbol{T}^{-j}}. \tag{3.13}$$

If

$$r_{\text{POTH}}(j) < 0$$

then treatment j reduces the certainty in the hierarchy as a whole. Similarly, if

$$r_{\text{POTH}}(j) > 0$$

then treatment j increases the certainty in the hierarchy as a whole. A POTH residual equal to zero means that the certainty in the hierarchy does not change if treatment j is ignored. Under maximum certainty in the hierarchy, the POTH residual for every treatment is zero. Additionally, $r_{\text{POTH}}(j)$ is bounded between -1 and 1.

Best k Treatments POTH

The certainty in the best positions in the hierarchy may be of greater interest than that of all treatments. In particular, we may be interested in the POTH of the best k treatments, where $k < N$. We suggest defining $\boldsymbol{T}^{\text{best } k}$ as the set of k treatments with the k highest SUCRA or P-score values in the network. Then we can calculate $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ for $k = 2, ..., N$, which represents the POTH for the best k treatments, and note that $\text{POTH}_{\boldsymbol{T}^{\text{best } N}} = \text{POTH}$. To investigate how the certainty in the hierarchy changes as k increases to N, we can plot $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ versus k. Significant changes in $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ after the addition of a treatment can help identify clusters of similarly performing treatments, as shown in the following hypothetical example.

Interpretation of POTH Residuals and Best k Treatments POTH: A Hypothetical Example

The POTH residuals and $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ in combination with a graphical presentation of the size and precision of estimated effects, such as a plot of the posterior distributions or a forest plot, can aid in interpreting the impact of individual treatments on the certainty in a hierarchy. Consider two networks with four treatments, A, B, C, and D. In one network, the four treatments are all moderately different from each other, while in the other, three treatments B, C, and D perform similarly while A stands out as having a much different

effect than the others (for example, A may be a new drug which represents a significant advance over treatment options B, C, and D). We simulate draws from hypothetical posterior distributions under these two scenarios, and calculate POTH, POTH residuals, and best k treatments POTH. We assume the outcome is harmful, so smaller responses are desirable. In Figure 3.2, we show the posterior distributions for a situation with moderate separation of all effects (left) and a situation where treatment A is significantly separated from the others (right). Both scenarios have similar POTH (0.709 versus 0.701).

Below each posterior plot in Figure 3.2 we show a graphical representation of the POTH residuals for each treatment. In the situation with similar separation between all treatments, the top and bottom treatments in the hierarchy increase POTH an equivalent amount while the middle two treatments decrease POTH an equivalent amount. The top and bottom treatments increase POTH because they are similar in effects to only one treatment each, while the middle treatments have two neighbouring treatments to which they have similar effects. In the scenario where A is quite different from B, C, and D, the bottom three treatments in the hierarchy reduce the POTH, while the best treatment A, which has a much different effect size than the others, increases the POTH an important amount.

In the bottom row of Figure 3.2 we present the best k treatments POTH for different values of k in the two networks. In the network where all treatments are moderately different, we see a steady increase in $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ as more treatments are considered. In the scenario where A is much different from the other treatments, we see the highest $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ when only the top 2 treatments, A and B, are considered. This is because the posterior distribution of the relative effect of A vs. B has almost no overlap with zero, giving high confidence in the treatment hierarchy for these two treatments. As more treatments are added, the $\text{POTH}_{\boldsymbol{T}^{\text{best } k}}$ falls since the treatments we are adding have a lot of overlap with each other.

3.3 What is the Precision of Treatment Hierarchies in Published Networks? Empirical Evidence from 267 Published Networks

Petropoulou et al. (2017) provided a comprehensive database of NMAs published between 1999 and 2015. We calculate POTH in this database to gain an understanding of POTH in published networks and investigate relationships between POTH and network characteristics. From the database, a total of 267 NMAs were identified with available data. The R

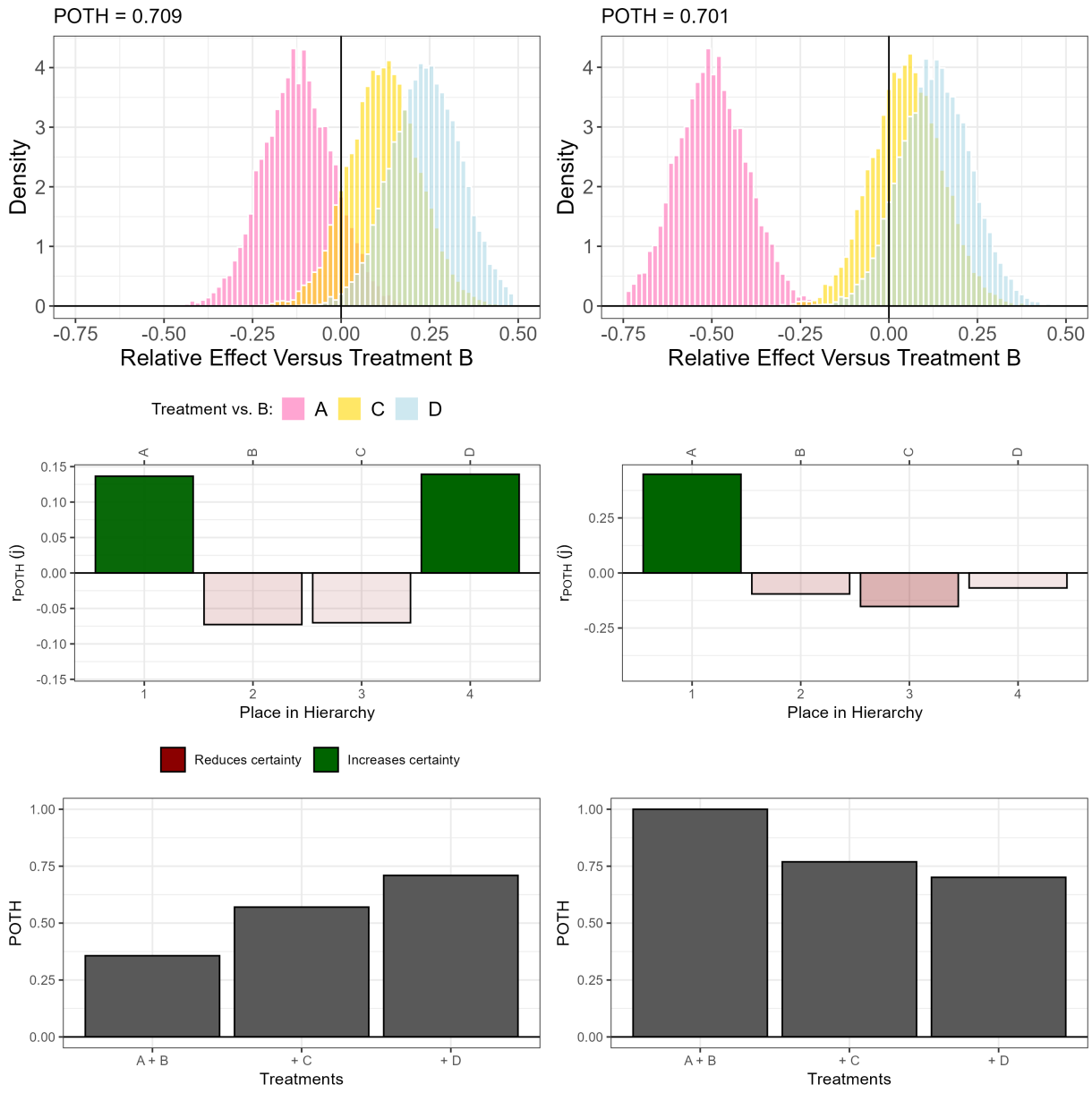


Figure 3.2: The simulated posterior distributions (top row), POTH residuals $r_{\text{POTH}}(j)$ (middle row), and $\text{POTH}_{\boldsymbol{T}^{\text{best }k}}$ (bottom row) for two different hypothetical scenarios. Smaller values indicate a better treatment.

package **nmadb** was used to access the data (Papakonstantinou, 2019). Each dataset was re-analysed using frequentist random-effects NMA using the R package **netmeta** (Balduzzi et al., 2023). The effect measure used in the initial publication of each dataset was used in the re-analysis. The REstricted Maximum Likelihood (REML) estimator was used to estimate the between-study heterogeneity parameter σ. P-scores for each treatment were calculated, then POTH was calculated for each network using the R package **poth** (Wigle and Schwarzer, 2024).

The distribution of POTH in the networks is shown in Figure 3.3. Note that due to the sample size and the lack of included networks since 2015, this distribution should not be used as a reference distribution for POTH. Still, it provides some insight into the different levels of certainty in treatment hierarchies observed in some published networks. The minimum POTH is 0.096 and the maximum is over 0.999. Results for the networks producing the minimum and maximum POTH are included in Appendix B.4. The median POTH is 0.671. The inter-quartile range is 0.528 to 0.833.

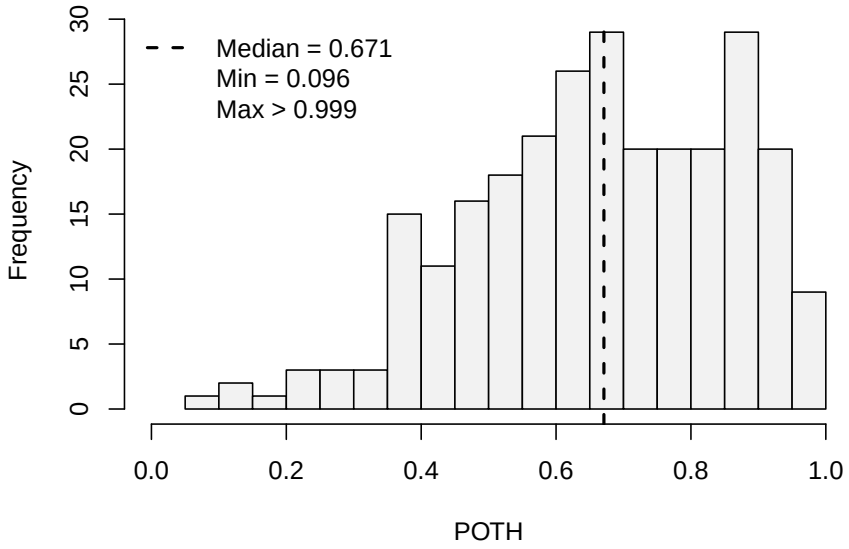


Figure 3.3: Histogram of POTHs from random-effects NMAs performed on 267 datasets available from Petropoulou et al. (2017); Papakonstantinou (2019).

A plot of the POTH versus the number of treatments in the network is shown in the left

panel of Figure 3.4. There are many more NMAs comparing ten or fewer treatments than NMAs comparing more than ten treatments. The Spearman's rank correlation between the number of treatments in the network and the POTH is -0.193, indicating a weak relationship. This suggests that we can compare POTH across different networks with varying numbers of treatments. However, there is limited data on the POTH for networks comparing more than 20 treatments, and a better understanding of the behaviour of POTH in networks comparing large numbers of treatments is needed before conclusions about its use in very large networks can be made.

Another interesting relationship is that between the POTH and the number of treatment comparisons which are statistically significant in an NMA. The right panel of Figure 3.4 shows a plot of the POTH versus the proportion of treatment comparisons that are significantly different at the 5% level. There is a clear relationship between the POTH and the proportion of significantly different treatment comparisons, with POTH values close to one tending to have a proportion of significant comparisons close to one as well. The converse is true to some extent, however there is more variation. For example, among NMAs where there were zero significantly different treatment comparisons, the range of POTH values is from 0.096 (very low certainty) to 0.571 (moderate certainty). Still, this association can aid in the interpretation of POTH values. For a network with POTH less than or equal to 0.5, the proportion of significant treatment comparisons is expected to be quite low - the largest proportion of significant treatment comparisons seen in networks with POTH less than 0.5 was 0.267. Networks with moderate POTH values between 0.5 and 0.8 had a proportion of significant treatment comparisons ranging between 0 and 0.533. We found that networks with POTH greater than 0.8 had at least one third of the treatment comparisons being significant. For a network with POTH close to one, we expect almost all treatment comparisons to be significant.

Additionally, we investigate the relationship between POTH and heterogeneity in the network. Since the interpretation of the heterogeneity parameter σ depends on the effect measure, we investigate the NMAs based on the effect measure used in the analysis. The three most common effect measures in the database were the log odds ratio (119 NMAs), log relative risk (68 NMAs), and mean difference (45 NMAs). There were 14 or fewer NMAs using each of the remaining effect measures. The POTH versus the estimate of σ is shown for the top three effect measures in Figure 3.5. The Pearson's correlation coefficient between POTH and σ for each effect measure are -0.143, -0.074, and -0.110 for log-odds ratio, log relative risk, and mean difference respectively. This indicates a weak negative correlation between POTH and the heterogeneity of a network.

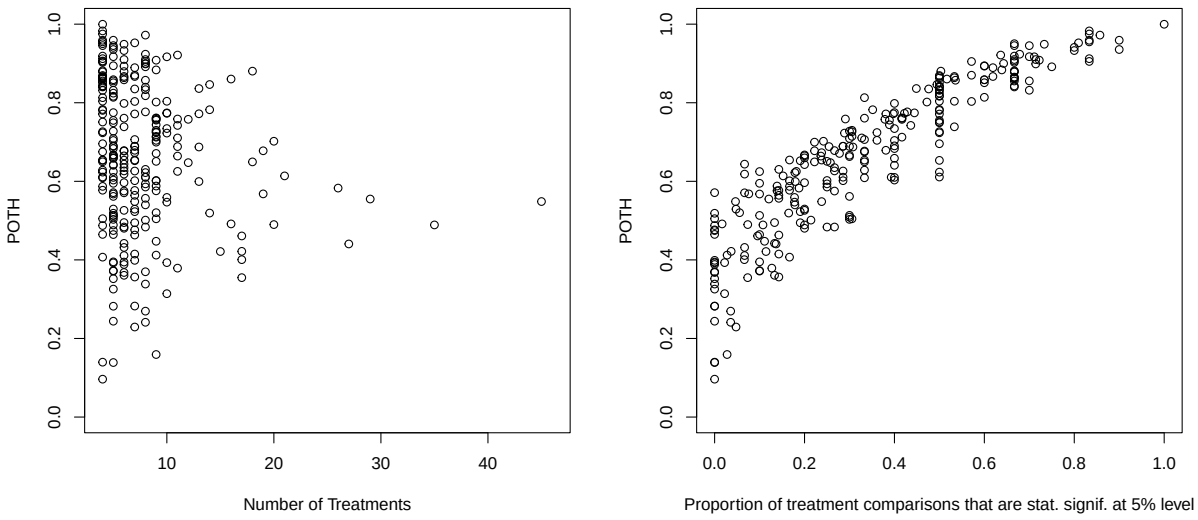


Figure 3.4: Left: Scatterplot of the POTH versus the number of treatments in the network in the 267 random effects NMAs. Right: Scatterplot of the POTH versus the proportion of treatment comparisons with a p-value less than 0.05 in the 267 random effects NMAs. That is, the x-axis is the proportion of treatment comparisons where there is evidence that the treatments being compared have different effects.

3.4 Real Data Example: Safety of Immune Checkpoint Inhibitors as Cancer Treatments

Here we present a published network and demonstrate the use and interpretation of POTH, POTH residuals and best k treatments POTH. A NMA of Immune Checkpoint Inhibitors (ICIs) was conducted by Xu et al. (2018) to assess the safety of ICI drugs as cancer treatment. Data were obtained from Rosenberger et al. (2021), who re-analysed the data. There are seven treatments and 23 studies. The outcome of interest is the number of treatment-related adverse events, so smaller values indicate a safer treatment. A random-effects NMA model was fit to the data using the REML method to estimate heterogeneity. The POTH was calculated using the R package **poth** (Wigle and Schwarzer, 2024).

The estimated treatment effects are shown in the forest plot in Figure 3.6. Atezolizumab is ranked as the safest in the hierarchy and ICI and conventional is ranked as the least safe. The POTH is equal to 0.838. A reference distribution for this scenario where there

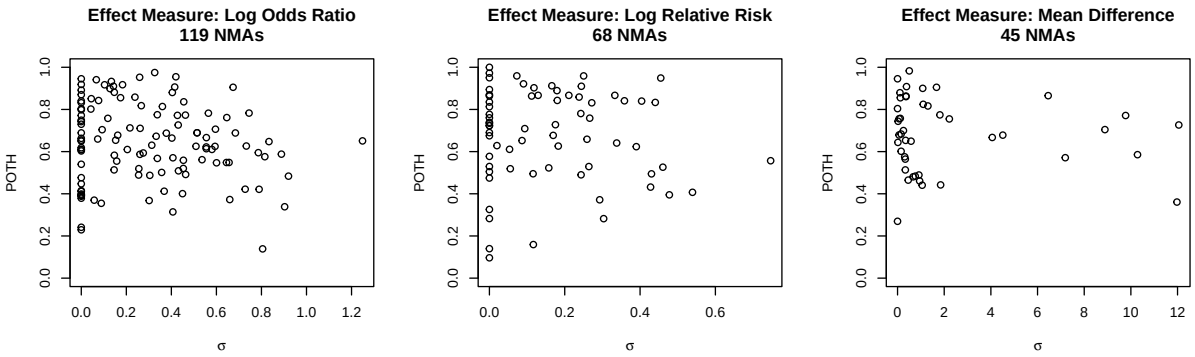


Figure 3.5: Plots of POTH versus the estimated heterogeneity parameter σ for the three most common effect measures in **nmadb**: log odds ratio, log relative risk, and mean difference. There is an outlying point that was omitted from the mean difference plot which had an estimated σ of 68.06 and POTH equal to 0.597.

are no true treatment differences is shown in Figure 3.7. We found that there was only a 0.6% chance of obtaining a larger POTH than 0.838 in the reference scenario, indicating that it is extremely rare to obtain a treatment hierarchy with this extent of certainty when there are no treatment differences. Thus, the POTH value of 0.838 indicates a high extent of certainty in the treatment hierarchy, and we would be very unlikely to have this high of an extent of certainty if there were no true treatment differences.

The POTH residuals are shown in the column beside the P-scores and in the bottom right panel. The treatments contributing the most to certainty in the hierarchy are atezolizumab, ICI and conventional, and conventional therapy. Ipilimumab and two ICIs reduce the certainty in the hierarchy. Ipilimumab has a similar point estimate to nivolumab, and two ICIs has the widest CI in the network.

The best k treatments POTH is shown for different values of k in the bottom panel of Figure 3.6. The POTH is the lowest when the top three treatments are considered. There is a noticeable increase in the POTH when going from the top three to the top four treatments. The certainty then increases steadily as more treatments are added. This illustrates that there is a cluster of treatments with similar safety profiles occupying the top three spots in the hierarchy made up of atezolizumab, pembrolizumab, and conventional therapy. Indeed, we can see the similarity in the CIs of atezolizumab and pembrolizumab versus conventional therapy in the forest plot in Figure 3.6.

Finally, consider creating a hierarchy for only the treatments which are single ICIs. That is, let $\boldsymbol{T}$ ={atezolizumab, pembrolizumab, ipilimumab, nivolumab}. The resulting $\text{POTH}_{\boldsymbol{T}}$

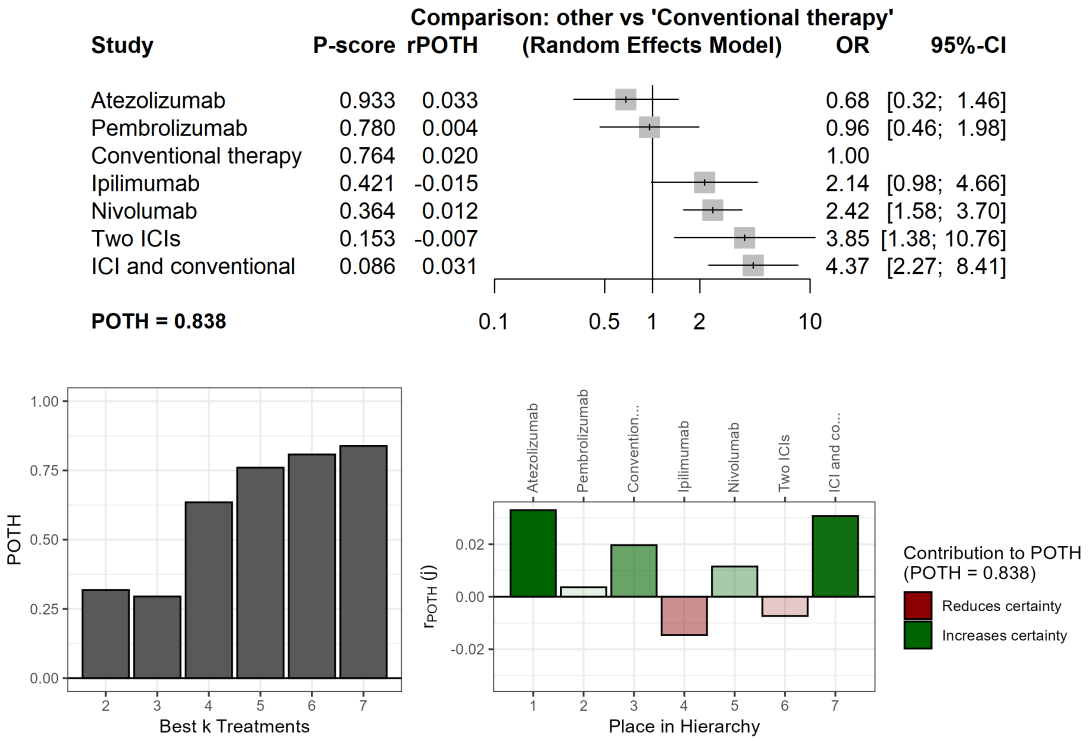


Figure 3.6: Above: Estimated confidence intervals (CI) for the odds ratio of all-grade treatment-related adverse events for ICI versus conventional therapy in the cancer network. P-scores and residual POTH are shown to the left of CIs. ORs less than one favour the intervention and ORs greater than one favour conventional therapy. The treatments are organized from largest (highest ranking in hierarchy) to smallest (lowest ranking in hierarchy) P-scores. A positive residual indicates that the treatment contributes to the certainty in the hierarchy, while a negative residual indicates that the treatment reduces the certainty in the hierarchy.
Below, left: Best k treatments POTH for $k = 2$ to N treatments in the ICI network.
Below, right: POTH residuals $r_{\text{POTH}}(j)$ for all treatments in the network.

is 0.777, indicating moderate certainty. This is lower than the POTH for all treatments in the network (0.838), indicating that there is more certainty in the hierarchy considering all treatments in the network than in the hierarchy of only the single ICIs. In Figure 3.6, we can see that the single ICIs roughly form two clusters (atezolizumab/pembrolizumab and ipilimumab/nivolumab). This example is reminiscent of the case when there are two groups of equivalent treatments described in Section 3.2.5, where we found that the POTH

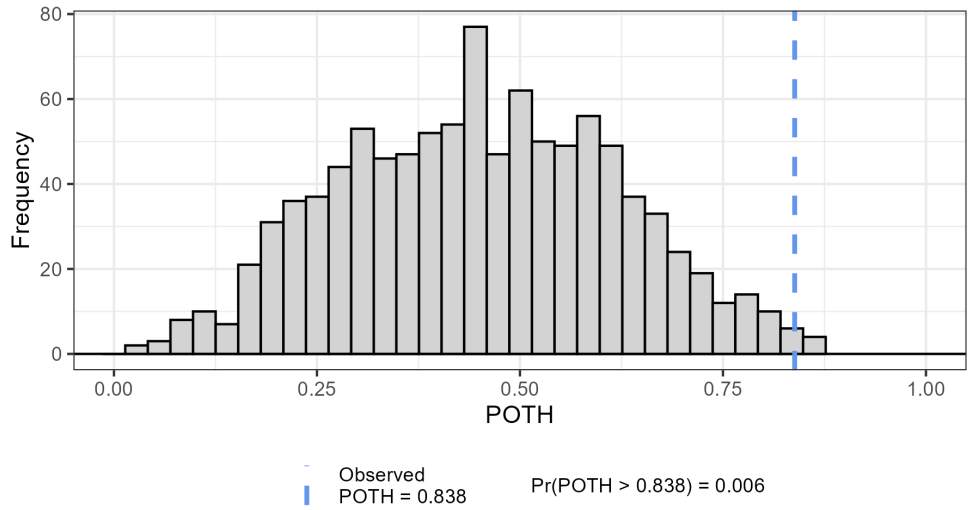


Figure 3.7: Distribution of POTH values observed in the same network as the ICI example but with no treatment differences.

is bounded below by 0.75.

3.5 Discussion and Future Work

In this chapter, we introduce POTH, a metric which quantifies the certainty in producing a treatment hierarchy. We define maximum and minimum certainty scenarios and show that POTH is uniquely maximized in the case of maximum certainty and minimized in the case of minimum certainty. We propose $\text{POTH}_{\boldsymbol{T}}$ as a method to measure certainty in the hierarchy of a subset of treatments. The POTH answers the question "how certain is the presented hierarchy of all treatments?" while $\text{POTH}_{\boldsymbol{T}}$ answers the question "how certain is the hierarchy of a subset of treatments of interest?". We also introduce two useful applications of $\text{POTH}_{\boldsymbol{T}}$: POTH residuals using a leave-one-out approach to measure the contribution of each treatment to POTH, and best k treatments POTH to identify the certainty of the hierarchy for the top k treatments as k increases. The methods are easily applied in Bayesian and frequentist frameworks.

Although POTH was proposed specifically to summarise the certainty in treatment hierarchies derived using SUCRAs or P-scores, it can also be viewed simply as a way to summarise all the ranking probabilities in a given network. It is therefore an indicator of the

certainty in any treatment hierarchy derived using a ranking metric related to the ranking probabilities. A strength of POTH is that it can be applied to any extensions of SUCRA or P-values, as long as their range is zero to one and their mean is 0.5. For example, a version of POTH could be calculated using Bayesian predictive P-scores to quantify the certainty in the hierarchy of treatments when applied in a new study (Rosenberger et al., 2021). Another useful application would be to calculate POTH from SUCRA values that consider minimally important differences (Curteis et al., 2025). This approach would quantify the certainty in a hierarchy which only considers clinically meaningful treatment differences.

POTH is a quantitative approach, which sets it apart from some previously suggested approaches to quantifying uncertainty in treatment hierarchies (Papakonstantinou et al., 2022; Salanti et al., 2014). The average normalized entropy is also a quantitative hierarchy-level approach. An advantage of POTH over the average normalized entropy is that POTH can be calculated directly from SUCRAs or P-scores rather than relying on the individual ranking probabilities p_{ku}.

We performed an exploratory simulation study to investigate the distribution of POTH that can be expected in networks where no true treatment differences exist. Our preliminary results show that in networks with no true treatment differences, data with smaller standard errors leads to higher average POTH values, while increasing between-study heterogeneity can lead to lower average POTH values. We also found that the underlying network structure can have an important impact on the distribution of POTH observed in repeated sampling. Full results are available in Appendix B.6. A more in-depth simulation study which further investigates the impact of several network characteristics such as different magnitudes of treatment differences, precision in the data, number of treatments, network geometries, and heterogeneity is of interest. It would be difficult to capture a sufficient range of these factors to provide a reference distribution for a specific NMA whose properties were not known a priori. Instead, we provide an R function which provides a simulation-based reference distribution for the user's desired network characteristics. This ensures that a point of formal comparison is available to aid in the interpretation of an observed POTH value for any desired scenario.

It was previously highlighted that the quality of evidence in NMA should be evaluated separately for the treatment effect estimates and the hierarchy of treatments (Salanti et al., 2014). Formal guidance for assessing the certainty in NMA estimates has been provided, notably GRADE (Guyatt et al., 2008, 2011) and Confidence in Network Meta-Analysis (CINeMA) (Nikolakopoulou et al., 2020). However, no similar guidance for the hierarchies resulting from an NMA have been created. POTH and POTH for subsets of treatments could play a role in the future development of guidelines for evaluating the certainty in treatment hierarchies from NMA.

Chapter 4

Estimating Methane Emissions from the Upstream Oil and Gas Industry Using a Multi-Stage Framework

4.1 Introduction

In the present and following chapter of the thesis, we turn our attention to some challenges in accurately quantifying methane emissions from OG production and storage facilities. We introduce new notation in Chapters 4 and 5, and any commonalities between the notation used in Chapters 2 and 3 are coincidental.

Methane inventories provide critical insights into progress towards reductions targets, as well as highlighting potential areas for improvement, such as facility types or regions with higher emission rates (Conrad et al., 2023c; Johnson et al., 2023). Multiple recent studies have shown that official methane inventories in Canada significantly underestimate emissions and highlight the value of measurement-based inventories (Tyner and Johnson, 2021; MacKay et al., 2021; Johnson et al., 2023; Conrad et al., 2023a,c). Measurement-based inventories have the potential to provide accurate and data-driven estimates of emissions, but there is a lack of guidance on how measurement-based inventories should be compiled from survey data (Vollrath et al., 2024).

In the estimation of methane inventories for the upstream OG sector using measurement data, four sources of uncertainty must be considered. First, it is not practical due to cost and time constraints to measure methane at every facility in the population of interest,

such as an entire province or an entire country. This leads to sampling error - that is, the error in the estimate due to not measuring every facility in the population (results will vary if we sample different subsets of facilities). Similarly, it is also impractical to survey facilities every day within a time period of interest, leading to sampling error due to the days. Third, the measurement technology can fail to detect methane at emitting sources even when the emission rate is above its minimum detectable limit, leading to detection uncertainty. Finally, measurement technologies are subject to measurement uncertainty, that is, when methane is detected, the estimated emission rate is not exactly equal to the true emission rate, and is likely biased. These four sources of uncertainty must be considered when creating measurement-based estimates of methane emissions.

Measurement-based data has been collected in multiple surveys in Canada using Bridger Photonics, Inc's Gas-mapping LiDAR (GML) plane-mounted technology (Johnson et al., 2023; Conrad et al., 2023a,c). The GML technology has been studied through extensive controlled releases (Johnson et al., 2021; Tyner and Johnson, 2021) and both the probability of detection and measurement error of the technology have been characterised (Conrad et al., 2023b). These properties correspond to detection uncertainty and measurement uncertainty as described above. The probability of detection function and measurement error distribution can readily be used to quantify these sources of error in methane emissions inventories.

A simulation approach to creating measurement-based inventories from the aerial survey data collected using GML was recently developed and applied in Canada (Johnson et al., 2023; Conrad et al., 2023a,c). To the best of our knowledge, it is the only available approach to create measurement-based inventories from aerial survey data in the current literature. The approach separately estimates "measurable" emissions, which are emissions that are detectable by the GML aerial survey technology, and "unmeasurable" emissions, which are emissions below the minimum detectable limit of the measurement technology. Only the measurable portion of emissions can be estimated using the GML data, while the unmeasurable portion requires additional methods which rely on auxiliary data from surveys of facilities on the ground using OGI cameras.

The approach to estimating the measurable portion of emissions developed in Johnson et al. (2023) uses a nested Monte Carlo (MC) algorithm to account for measurement error, sampling error, and detection error. First, a MC procedure which accounts for measurement and detection uncertainty (the initial MC) is performed. Next, for every iteration of the initial MC, a mirror match bootstrap procedure (Sitter, 1992) is followed in order to scale the sample estimates to the population level and account for sampling error due to the facilities. The algorithm requires computationally intensive methods including numerical integration and root-finding methods to solve multiple equations at each iteration of

the initial MC. Additionally, both the initial MC and mirror-match bootstrap are complex and open-source code for their implementation is not available.

The primary objective of this work is to propose an alternative framework for estimating the measured portion of measurement-based methane inventories and their uncertainties which is easy to interpret and implement. We accomplish this goal by viewing the problem through a statistical survey lens, treating non-detection as survey non-response, and using established multi-stage weighting methods. We provide analytical formulas which account for sampling and detection uncertainty and have desirable statistical properties, such as unbiasedness. We also present decompositions of the variance formulas which are useful in identifying important contributors to the uncertainty of the inventory. Additionally, we propose a straightforward MC approach which incorporates measurement uncertainty. Further, an R package which implements the described analysis was written and is freely available. This alternative method and accompanying R package will make measurement-based inventories more accessible, in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable) endorsed by The Intergovernmental Panel on Climate Change (the United Nations body for assessing the science related to climate change) (Pirani et al., 2022). To the best of the authors' knowledge, this is the first attempt at addressing the development of measurement-based inventories from a statistical foundation.

The chapter is organized as follows: In Section 4.2, we provide the necessary background on the motivating dataset and the technology used to collect the data. In Section 4.3, we cast the problem under a design-based statistical framework and explain how weighting methods for multi-stage sampling can be applied in this situation, providing two options for estimators and the necessary modifications to the methodology required for each option. We provide a simple MC procedure that can be combined with the multi-stage framework to incorporate measurement uncertainty into the estimate and its variance. We present a simulation study evaluating the multi-stage estimators with respect to the design and apply the methods to the motivating dataset in Section 4.4, providing insights about key emitters and important sources of uncertainty. We conclude with a discussion of the results and potential future areas of development in Section 4.5.

4.2 Aerial Survey Data from British Columbia

Aerial survey data were collected using GML in the province of British Columbia (BC), Canada in 2021. The survey was the basis for developing a measurement-based methane inventory from the province's OG industry (Johnson et al., 2023). The British Columbia

Oil and Gas Research and Innovation Society (BC OGRIS) and the Methane Emissions Research Collaborative (MERC) provided anonymized aerial survey data to facilitate the development of new measurement-based methods for estimating methane emissions. Some anonymized data from the BC survey have since been published (Johnson et al., 2023), and subsequent surveys following similar designs to the BC survey have been performed in other provinces in Canada (Conrad et al., 2023a,c). A subset of the BC data provided by BC OGRIS and the MERC is available in the R package methaneInventory.

In the BC survey data, the airborne GML technology was used to detect and quantify methane from point sources at selected upstream OG facilities. Measurements made using airborne GML were attributed to specific components, where a component is a piece of equipment at an OG facility, such as a compressor building or a tank (Johnson et al., 2023). The plane visited some sites on multiple survey days, where a site is a collection of one or more facilities which are close together and were surveyed simultaneously. In addition, some facilities are large and required multiple passes by the plane to cover them completely on each survey occasion, with some components being visible to the plane in multiple passes. This resulted in multiple days of measurement data for some components, with potentially multiple data points per component per day. In passes where methane was detected, we have available the measured emission rate from GML, the plane's altitude, and the wind speed. The number of passes per day per component ranges between one and five, with a median of two passes per day per component. A total of 543 unique components are identified in the dataset, with a total of 1171 passes with successful detections.

Different kinds of facilities are expected to have different levels of emissions so it is useful to classify facilities by Petrinex subtype, where Petrinex is Canada's Petroleum Information Network (MacKay et al., 2021). Boxplots of the non-zero measurements for each facility type (stratum) are shown in Figure 4.1, with the number of non-zero methane measurements shown below each stratum's boxplot. Details about the definition of facility types and strata are given in Section 4.3.4 and Table 4.1. There is wide variation in the distribution of measurements between different strata, in terms of the average, variance, and skewness. There is also variation in the number of non-zero measurements, ranging from zero (enhanced recovery schemes and water and waste facilities, which are not shown in Figure 4.1) to 305 (gas multi-well batteries for effluent).

The GML's detection abilities are imperfect and sometimes the technology fails to detect emissions. As mentioned in the introduction, work has been done to describe both the probability of detection and the distribution of measurement error for the GML technology. A function which calculates the probability of detection (POD) of a given source by the GML was derived by Conrad et al. (2023b) using controlled release data. The POD is a function ϕ of the true emission rate (Y, in kg/h), altitude of the plane over the source (a,

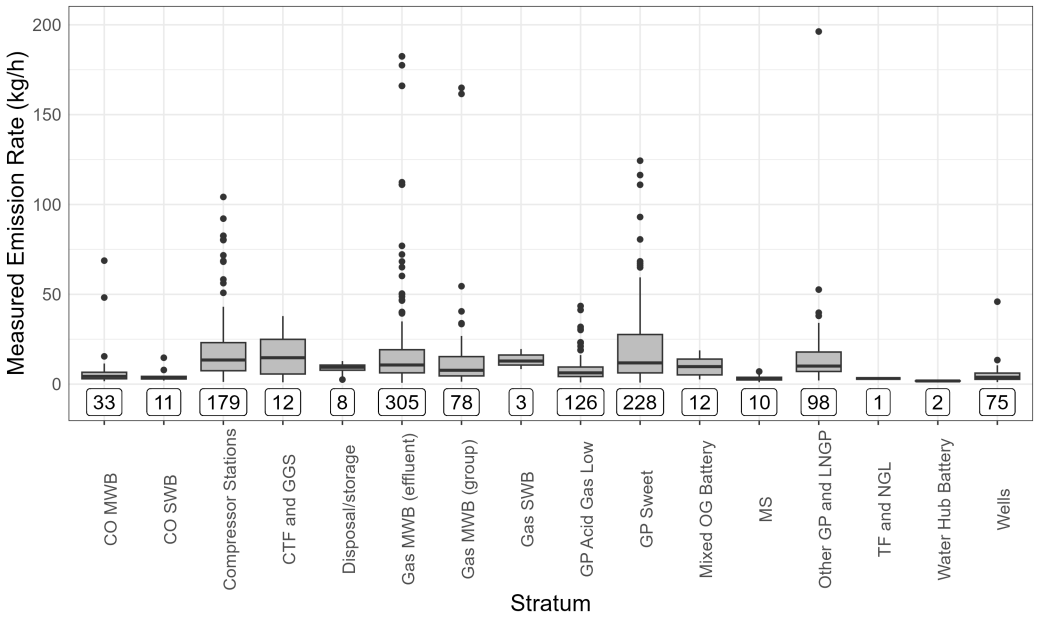


Figure 4.1: Boxplots of non-zero measured emission rates for facility types (strata). There are two outliers above 200 kg/h in the dataset for the sweet gas plant facility type. The number of non-zero measurements for each stratum is shown in the box below each boxplot. Acronyms: GP = gas plant, MWB = multi-well battery, LNGP = liquid natural gas plant, CO = crude oil, CTF = custom treating facility, GGS = gas gathering system, SWB = single-well battery, MS = meter station, TF = tank farm, NGL = natural gas liquids.

in m), and wind speed (u, in m/s). The POD of a point source is calculated as

$$\phi(Y, a, u) = \exp\left[-\left\{\frac{0.224 \times Y^{1.07}}{\left(\frac{a}{1000}\right)^{2.44}(u + 2.14)^{1.69}}\right\}^{-2.53}\right] \tag{4.1}$$

(Conrad et al., 2023b). We ignore any error or uncertainty in this function, in line with Johnson et al. (2023).

In addition, Conrad et al. (2023b) derived a probability density function for the true emission rate given a measurement from GML using controlled release data. Letting M be the measurement of the true emission rate Y, the distribution is

$$Y \mid M, M > 0 \sim \text{Log-logistic}(\text{scale} = d \times \alpha \times M, \text{shape} = \beta), \tag{4.2}$$

where $d = 0.918$, $\alpha = 0.891$, and $\beta = 3.82$. The mean of Y given M, by properties of the log-logistic distribution, is

$$d \times M \times \frac{\alpha\pi/\beta}{\sin(\pi/\beta)} = d \times M = 0.918 \times M. \tag{4.3}$$

This means that on average, the true emission rate is equal to 0.918 times the measured value. This insight provides a straightforward bias correction for measurements.

Table 4.1: Description of facilities, Petrinex subtype codes, and stratum definitions used in the data analysis.

Facility Description	Petrinex Code	Stratum Name	Sample size (n_h)	Population size (N_h)
Crude oil single-well battery	311	CO SWB	48	58
Crude oil multi-well group battery Crude oil multi-well proration battery	321 322	CO MWB	34	40
Gas single-well battery	351	Gas SWB	21	28
Gas multi-well group battery	361	Gas MWB (group)	54	79
Gas multi-well effluent measurement battery	362	Gas MWB (effluent)	113	140
Mixed oil and gas battery	393	Mixed OG Battery	16	16
Water hub battery	395	Water Hub Battery	22	34
Gas plant sweet	401	GP Sweet	21	25
Gas plant, acid gas flaring (< 1 t/d sulphur)	402	GP Acid Gas Low	19	23
Gas plant, acid gas flaring (> 1 t/d sulphur) Gas plant, acid gas injection Gas plant, sulphur recovery Gas plant, fractionation Liquefied natural gas plant Gas transporter	403 404 405 407 451 204	Other GP and LNGP	10	21
Enhanced recovery scheme	501	Recovery Scheme	23	27
Disposal Acid gas disposal Underground gas storage	503 504 505	Disposal and storage	48	77
Compressor station	601	Compressor station	45	254
Custom treating facility Gas gathering system	611 621	CTF and GGS	97	144

Facility Description	Petrinex Code	Stratum Name	Sample size (n_h)	Population size (N_h)
Field receipt meter station	631	MS	51	91
Interconnect receipt meter station	632			
NEB regulated field receipt meter station	637			
NED regulated interconnect receipt meter station	638			
Tank farm loading and unloading terminal	671	TF and NGL	13	24
Third party tank farm loading and unloading terminal	673			
Natural gas liquids hub terminal	676			
Surface waste facility	701	Water and Waste	13	20
Water source	901			
Water source battery	902			
Well - Gas	-	Wells	1004	9978
Well - oil	-			
Well - water	-			
Well - undefined	-			

4.3 Methods

In this section, we show how a multi-stage approach can be leveraged to estimate methane emissions from aerial survey data, accounting for the four sources of uncertainty in the survey data without substantial computational costs. In Section 4.3.1, we explain how a multi-stage framework can be applied to the survey data. Methods for estimating methane emissions that do not consider measurement uncertainty are developed in Sections 4.3.2, 4.3.3, and 4.3.4. In Section 4.3.5, we provide an algorithm to incorporate measurement uncertainty into the previous methods, allowing the consideration of all four sources of uncertainty in the estimation simultaneously. Methodology for a simulation study assessing the proposed multi-stage framework is described in Section 4.3.6. A summary of the notation used in this section is available in Appendix C.3.

4.3.1 Measurement-Based Surveys As Multi-Stage Sampling

Multi-stage sampling is a type of probability design where sampling is carried out in the following way:

1. A sample S^{I} of primary sampling units (PSUs) are selected from a population U^{I} according to a probability design.
2. For every PSU p in S^{I}, a sample of secondary sampling units (SSUs) S_p^{II} is drawn from the population U_p^{II} according to a probability design.
3. For every PSU p in S^{I} and SSU t in S_p^{II}, a sample of tertiary sampling units (TSUs) S_{pt}^{III} is drawn from the population U_{pt}^{III} according to a probability design,

and so on. Each stage is defined using a probability design which describes how the units at that stage were sampled, given the results of the sampling in the previous stage(s). In the context of measurement-based aerial surveys of OG facilities, a three-stage approach can be employed to describe the data collection process. We introduce stages to represent the selection of components to sample, the collection of data on different days, and the process of detection of methane in each pass.

In order to apply the methods presented here, the first and second order inclusion probabilities for all three stages of sampling must be known. The first order inclusion probability refers to the probability of a sampling unit being selected in a given stage. The second order inclusion probability refers to the probability of any two units both being selected at a given stage. However, data may not have been collected according to a statistical probability design. Practitioners must then work to find a probability design which matches the process by which data were collected as closely as possible. This may require assumptions and approximations which should be kept in mind when interpreting results.

In the methane data, the first stage of sampling represents the selection of components to survey. Therefore, the first stage population U^{I} represents the components at all OG facilities in the province of BC, and S^{I} represents the components which were surveyed. To support application of these methods to other measurement-based surveys, we introduce a generic probability design to represent the first stage in Section 4.3.2, while also providing results for the specific design assumed for the BC data in Section 4.3.4. The generic first stage probability design is described by π_p^{I}, the probability that component p is in S^{I}, and π_{pl}^{I}, the probability that components p and l are both in S^{I}.

The second stage of sampling is the selection of the days on which to survey each component selected into S^{I}. For each component $p \in S^{\mathrm{I}}$, we select a sample of days S_p^{II} on which to survey p. The second stage population U_p^{II} represents the entire time period for which we want to estimate emissions. Typically, methane emissions are estimated on a yearly basis, in which U_p^{II} is the year for all p. Similarly to the first stage, we use a generic second stage design, and later provide specific results for the BC data. The generic second stage design is described by $\pi_{t|p}^{\mathrm{II}}$, the probability that we survey component p on day t given that p was selected into S^{I}, and $\pi_{tu|p}^{\mathrm{II}}$, the probability that we survey component p on days t and u given that p is in S^{I}.

Finally, the third stage represents detection of methane in the passes on each day a component was surveyed. While in the first and second stages we use generic probability designs to allow for different modelling of the selection process for components and days, for the third stage we specify a probability design which is broadly applicable to modelling non-detection in measurement-based surveys. We adapt approaches for treating non-response in traditional surveys to treat passes with missed detections. We differentiate between missed detections and true non-detections in the same way as Johnson et al. (2023), by assuming that if methane is detected at component p in at least one pass on day t, any passes without detections on day t at component p are missed detections. The set S_{pt}^{III} is the set of passes in which methane was detected from component p on day t, while U_{pt}^{III} is the set of all passes made over component p on day t, regardless of detection status. The sampling at this stage is therefore governed by the probability of detection from component p on day t in pass q, $\pi_{q|pt}^{\mathrm{III}}$. We leverage Equation (4.1) to calculate the POD for GML, so that in the BC data, $\pi_{q|pt}^{\mathrm{III}} = \phi(Y_{ptq}, u_{ptq}, a_{ptq}) = \phi_{ptq}$, where Y_{ptq} is the true emission rate, u_{ptq} is the wind speed in m/s, and a_{ptq} is the height of the plane during pass q on day t over component p. Further, we assume that detection is independent between passes given the variables used to calculate ϕ_{ptq}, so that the second order inclusion probability $\pi_{qr|pt}^{\mathrm{III}}$ is given by $\phi_{ptq}\phi_{ptr}$. This design corresponds to Poisson sampling as discussed in Section 3.5 of Särndal et al. (1992) and 4.6 of Wu and Thompson (2020).

The three stages of sampling represent three important sources of uncertainty in estimating methane emissions from measurement-based survey data: uncertainty due to the facilities and components sampled (stage I), uncertainty due to the days on which sampling takes place (stage II), and detection uncertainty (stage III). The three-stage sampling framework therefore accounts for these three sources of error in the point estimate and its variance estimate. We refer to estimators which are unbiased with respect to stages I, II, and III as ptq-unbiased. The fourth source of uncertainty in the survey data, measurement uncertainty, must be incorporated outside of the multi-stage framework; see Section 4.3.5.

4.3.2 Estimating Methane Emissions Using Multi-Stage Sampling

We aim to estimate the total methane emission rate from all OG facilities in BC. We formulate the estimand as

$$T = \sum_{p \in U^{\mathrm{I}}} \overline{Y}_{p\cdot\cdot} = \sum_{p \in U^{\mathrm{I}}} \frac{\sum_{t \in U_p^{\mathrm{II}}} \overline{Y}_{pt\cdot}}{D} = \sum_{p \in U^{\mathrm{I}}} \frac{\sum_{t \in U_p^{\mathrm{II}}} \sum_{q \in U_{pt}^{\mathrm{III}}} Y_{ptq}/Q_{pt}}{D}, \tag{4.4}$$

where Q_{pt} is the total number of passes over component p on day t, and D is the total number of days in the period of interest. Further, $\overline{Y}_{pt\cdot}$ and $\overline{Y}_{p\cdot\cdot}$ are the emission rate from component p on day t averaged over the passes, and the emission rate from component p averaged over the days and the passes. T is an adaptation of the commonly used multi-stage total estimator with no averaging, $\sum_{p \in U^{\mathrm{I}}} \sum_{t \in U_p^{\mathrm{II}}} \sum_{q \in U_{pt}^{\mathrm{III}}} Y_{ptq}$ (Särndal et al., 1992, p. 137). In this section, we adapt existing results to accommodate T.

An (approximately) ptq-unbiased estimator for T is

$$\hat{T} = \sum_{p \in S^{\mathrm{I}}} \frac{1}{\pi_p^{\mathrm{I}}} \hat{\overline{Y}}_{p\cdot\cdot} = \sum_{p \in S^{\mathrm{I}}} \frac{1}{\pi_p^{\mathrm{I}}} \frac{\sum_{t \in S_p^{\mathrm{II}}} \hat{\overline{Y}}_{pt\cdot}/\pi_{t|p}^{\mathrm{II}}}{D} \tag{4.5}$$

(Särndal et al., 1992, pp. 146–148), where $\hat{\overline{Y}}_{p\cdot\cdot}$ is an estimator for the true emission rate from component p, averaged over all days and passes, and $\hat{\overline{Y}}_{pt\cdot}$ is an (approximately) unbiased estimator for the mean emission rate from component p on day t. Possible forms for $\hat{\overline{Y}}_{pt\cdot}$ are described later in this subsection and in Section 4.3.3.

A variance estimator for $\hat{T}$ which accounts for the uncertainty due to the three stages (component sampling, day sampling, and detection uncertainty) is

$$\hat{\mathrm{V}}^{\text{3 stage}} = \sum_{p \in S^{\mathrm{I}}} \sum_{l \in S^{\mathrm{I}}} \frac{\left(\pi_{pl}^{\mathrm{I}} - \pi_p^{\mathrm{I}} \pi_l^{\mathrm{I}}\right)}{\pi_{pl}^{\mathrm{I}}} \frac{\hat{\overline{Y}}_{p\cdot\cdot}}{\pi_p^{\mathrm{I}}} \frac{\hat{\overline{Y}}_{l\cdot\cdot}}{\pi_l^{\mathrm{I}}} + \sum_{p \in S^{\mathrm{I}}} \frac{\hat{\mathrm{V}}_{p\cdot\cdot}}{\pi_p^{\mathrm{I}}}, \tag{4.6}$$

(Särndal et al., 1992, pp.148–149) where

$$\hat{\mathrm{V}}_{p\cdot\cdot} = \frac{1}{D^2} \left[\sum_{t \in S_p^{\mathrm{II}}} \sum_{u \in S_p^{\mathrm{II}}} \frac{\left(\pi_{tu|p}^{\mathrm{II}} - \pi_{t|p}^{\mathrm{II}} \pi_{u|p}^{\mathrm{II}}\right)}{\pi_{tu|p}^{\mathrm{II}}} \frac{\hat{\overline{Y}}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}}} \frac{\hat{\overline{Y}}_{pu\cdot}}{\pi_{u|p}^{\mathrm{II}}} + \sum_{t \in S_p^{\mathrm{II}}} \frac{\hat{\mathrm{V}}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}}} \right] \tag{4.7}$$

is an unbiased estimator of the variance of $\hat{\overline{Y}}_{p\cdot\cdot}$, and $\hat{\mathrm{V}}_{pt\cdot}$ is an unbiased estimator for the variance of $\hat{\overline{Y}}_{pt\cdot}$. The true variance is given in Appendix C.4. A 95% Wald CI for the true total emissions is given by

$$\hat{T} \pm 1.96 \times \sqrt{\hat{\mathrm{V}}^{3\text{ stage}}}. \tag{4.8}$$

The variance can be broken down into contributions from each stage, corresponding to three of the four sources of uncertainty in the data. The estimated contribution due to detection uncertainty (stage III), sampling of days (stage II), and sampling of components (stage I) are given by

$$\hat{\mathrm{V}}^{\mathrm{III}} = \sum_{p \in S^{\mathrm{I}}} \frac{1}{(\pi_p^{\mathrm{I}})^2} \frac{1}{D} \sum_{t \in S_p^{\mathrm{II}}} \frac{\hat{\mathrm{V}}_{pt\cdot}}{(\pi_{t|p}^{\mathrm{II}})^2}, \tag{4.9}$$

$$\hat{\mathrm{V}}^{\mathrm{II}} = \max\left\{0, \sum_{p \in S^{\mathrm{I}}} \frac{\hat{\mathrm{V}}_{p\cdot\cdot}}{(\pi_p^{\mathrm{I}})^2} - \hat{\mathrm{V}}^{\mathrm{III}}\right\}, \tag{4.10}$$

and

$$\hat{\mathrm{V}}^{\mathrm{I}} = \max\left\{0, \hat{\mathrm{V}}^{3\text{ stage}} - \hat{\mathrm{V}}^{\mathrm{II}} - \hat{\mathrm{V}}^{\mathrm{III}}\right\}, \tag{4.11}$$

respectively (Särndal et al., 1992, pp.148–149), where the max operator avoids negative variance estimates which can arise especially with small sample sizes at stage II and III.

Now, a method for estimating $\overline{Y}_{pt\cdot}$ in the presence of missed detections must be chosen. We can use inverse probability weighting (IPW) to account for detection uncertainty and obtain an unbiased estimate of $\overline{Y}_{pt\cdot}$. As discussed in Section 4.3.1, detection in each pass is modelled as Poisson sampling, with inclusion probability ϕ_{ptq} for component p on day t in pass q. Under this design, the IPW estimate of $\overline{Y}_{pt\cdot}$ is

$$\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}} = \frac{\sum_{q \in S_{pt}^{\mathrm{III}}} Y_{ptq}/\phi_{ptq}}{Q_{pt}} \tag{4.12}$$

with corresponding variance estimate

$$\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{IPW}} = \frac{1}{Q_{pt}^2} \sum_{q \in S_{pt}^{\mathrm{III}}} \frac{(1-\phi_{ptq})}{\phi_{ptq}^2} Y_{ptq}^2 \tag{4.13}$$

(Wu and Thompson, 2020). If the ϕ_{ptq} are perfectly known, then the estimator in Equation (4.12) and its variance estimator are ptq-unbiased.

4.3.3 Increasing Efficiency Using the Hájek Estimator

A drawback to the use of the IPW estimator to correct for detection uncertainty is that it can result in paradoxical estimates and have a large variance in some situations (Horvitz and Thompson, 1952; Wu and Thompson, 2020). An alternative to the use of the IPW estimator in Equation (4.5) is the Hájek estimator, which has the form

$$\frac{\sum_{q \in S_{pt}^{\text{III}}} Y_{ptq}/\phi_{ptq}}{\sum_{q \in S_{pt}^{\text{III}}} 1/\phi_{ptq}}$$

(Hájek, 1971). The Hájek estimator is only approximately ptq-unbiased, but has a much smaller variance than the IPW estimator in a variety of cases and is often preferred (Särndal et al., 1992, pp. 181–184; Wu and Thompson, 2020, pp. 64-65). In particular, the Hájek estimator has been shown to have smaller variance in cases where 1) the Y_{ptq} are constant or have low variability, and 2) the sample size is variable. In the context of the motivating dataset, for most components, the emission rates are relatively homogeneous between passes in the same day, approximately satisfying case 1). Due to missed detections, we also have a variable stage III sample size, satisfying case 2). This leads us to believe that the Hájek estimator has the potential to reduce the variance contribution from detection uncertainty.

However, caution is needed when attempting to use the Hájek estimator in the context of estimating methane emissions. In typical sampling designs, samples of size zero are not allowed, however, due to the low number of passes per day and variation in detection probability, the size of the stage III sample S_{pt}^{III} may be zero for some components and days. When the size of S_{pt}^{III} is zero, the Hájek estimator is undefined (0/0). Therefore, in order to make use of the Hájek estimator in estimating methane inventories, one must adjust the stage II sampling design to ensure it is well-defined.

To ensure S_{pt}^{III} is non-empty for use with the Hájek estimator, we define $S_{p}^{\text{II}*}$ as the set of days where component p is surveyed AND methane is detected at least once, which ensures that all stage III samples are non-empty. This results in modifications to the stage II and III probability designs, which are discussed below. Stage I remains unchanged. This approach is equivalent to the previously described approach if the IPW estimator is used, which is shown in Appendix C.5.

Detection is statistically independent between passes of the plane given the true emission rates, so the probability of having any detections for component p on day t given that

$p \in S^{\mathrm{I}}$ and $t \in S_p^{\mathrm{II}}$ is given by

$$\phi_{pt\cdot} = 1 - \prod_{q \in U_{pt}^{\mathrm{III}}} (1 - \phi_{ptq}).$$

The first order inclusion probabilities for the adjusted stage II sampling design are given by

$$\pi_{t|p}^{\mathrm{II}\ *} = \phi_{pt\cdot}\pi_{t|p}^{\mathrm{II}}, \tag{4.14}$$

the probability that day t is in $S_p^{\mathrm{II}*}$ given $p \in S^{\mathrm{I}}$. The probability that t and u are both in $S_p^{\mathrm{II}*}$ is

$$\pi_{tu|p}^{\mathrm{II}\ *} = \begin{cases} \phi_{pt\cdot}\phi_{pu\cdot}\pi_{tu|p}^{\mathrm{II}} & \text{if } t \neq u \\ \pi_{t|p}^{\mathrm{II}\ *} & \text{if } t = u. \end{cases} \tag{4.15}$$

The stage III inclusion probabilities also change as a consequence of the stage II adjustment. In the original sampling design, the inclusion probability for pass q is equal to ϕ_{ptq}, the probability that methane is detected in pass q given that $p \in S^{\mathrm{I}}$ and $t \in S_p^{\mathrm{II}}$. In the modified sampling design, the inclusion probability for pass q is the probability that $q \in S_{pt}^{\mathrm{III}}$ given that $p \in S^{\mathrm{I}}$ and $t \in S_p^{\mathrm{II}*}$, that is, the probability of detecting methane in pass q given that p is surveyed on day t AND that methane is detected in at least one pass on that day. This additional condition in the modified design means the responses are no longer independent between passes, and so the stage III sampling scheme is no longer Poisson. Instead, the inclusion probabilities are given by

$$\pi_{q|pt}^{\mathrm{III}\ *} = \phi_{ptq}/\phi_{pt\cdot}, \tag{4.16}$$

and the passes are no longer independent, with

$$\pi_{qr|pt}^{\mathrm{III}\ *} = \phi_{ptq}\phi_{ptr}/\phi_{pt\cdot} \tag{4.17}$$

when $q \neq r$.

Let

$$\hat{\overline{Y}}_{pt\cdot}^{\mathrm{Hájek}} = \frac{\sum_{q \in S_{pt}^{\mathrm{III}}} Y_{ptq}\phi_{pt\cdot}/\phi_{ptq}}{\sum_{q \in S_{pt}^{\mathrm{III}}} \phi_{pt\cdot}/\phi_{ptq}} = \frac{\sum_{q \in S_{pt}^{\mathrm{III}}} Y_{ptq}/\phi_{ptq}}{\sum_{q \in S_{pt}^{\mathrm{III}}} 1/\phi_{ptq}}, \tag{4.18}$$

which is always defined for $t \in S_p^{\mathrm{II}*}$. In the methane data, an approximately unbiased variance estimate is

$$\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{Hájek}} = \frac{\phi_{pt\cdot}}{Q_{pt}^2}\left[\sum_{q \in S_{pt}^{\mathrm{III}}}(1 - \phi_{ptq})\left(\frac{Y_{ptq} - \hat{\overline{Y}}_{pt\cdot}^{\mathrm{Hájek}}}{\phi_{ptq}}\right)^2 + (\phi_{pt\cdot} - 1)\left(\sum_{q \in S_{pt}^{\mathrm{III}}} \frac{Y_{ptq} - \hat{\overline{Y}}_{pt\cdot}^{\mathrm{Hájek}}}{\phi_{ptq}}\right)^2\right]. \tag{4.19}$$

Note that $\phi_{pt\cdot}$ can only be calculated for components on days in which methane was detected in all Q_{pt} passes, since we require Y_{ptq} to calculate ϕ_{ptq}. In practice, many components have some passes where detections were missed. For components on days with missed detections, we use mean imputation (MI) to estimate $\phi_{pt\cdot}$. That is, in passes with missed detections, we impute the mean of the detection probabilities for the successfully detected passes on that day. Let $\mu_{pt\cdot} = \sum_{q \in S_{pt}^{\text{III}}} \phi_{ptq} / |S_{pt}^{\text{III}}|$ be the mean of the detection probabilities among the detected passes, where $|\cdot|$ denotes the number of elements in $\cdot$. Then, let

$$\hat{\phi}_{pt\cdot} = 1 - (1 - \mu_{pt\cdot})^{|U_{pt}^{\text{III}} \setminus S_{pt}^{\text{III}}|} \prod_{k \in S_{pt}^{\text{III}}} (1 - \phi_{ptq}). \tag{4.20}$$

In practice, we use this estimate in place of the unknown true value $\phi_{pt\cdot}$ in all estimators presented here.

To estimate the total methane emissions using the Hájek estimator for the component's daily mean emission rates, replace S_p^{II} with $S_p^{\text{II}*}$, $\pi_{t|p}^{\text{II}}$, $\pi_{u|p}^{\text{II}}$, and $\pi_{tu|p}^{\text{II}}$ with $\pi_{t|p}^{\text{II}\,*}$, $\pi_{u|p}^{\text{II}\,*}$, and $\pi_{tu|p}^{\text{II}\,*}$, $\hat{\overline{Y}}_{pt\cdot}^{\text{IPW}}$ with $\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}$, and $\hat{\text{V}}_{pt\cdot}^{\text{IPW}}$ with $\hat{\text{V}}_{pt\cdot}^{\text{Hájek}}$ in the methods described in Section 4.3.2.

4.3.4 First and Second Stage Designs for BC Data

In this section, we explain the particular stage I and II designs we use to analyse the BC survey data.

Stage I: Stratified Cluster Sampling

Sampling of components was done by first identifying all the active OG facilities in the province (Johnson et al., 2023). Then, in an effort to maximise sample size within a constrained budget and to accommodate on-site follow-up surveys needed to estimate the unmeasured inventory, all facilities south of the 58 degree north latitude line were included in the survey because they are more densely populated and easier to access than more remote facilities. All components at facilities included in the survey were sampled at least once, that is, surveyed at least once with one or more passes per survey.

While the true sampling mechanism was not stratified, in the analysis, we take a stratified approach to stage I. That is, we treat the sample of facilities as if it were an independent SRS of facilities in each stratum and the sample size in each stratum was fixed. Analysing the data as if the facilities were selected using stratified sampling produces a post-stratified

estimate which has improved precision compared to a SRS (Smith, 1991). We stratify based on the type of facility, since the emissions distribution differs by facility type, as seen in the data in Figure 4.1 and in previous studies, e.g. MacKay et al. (2021).

More specifically, strata were defined based on Petrinex facility subtypes. Some subtypes were merged to form larger strata based on expert knowledge and to improve sample sizes as described in Johnson et al. (2023). This scheme still leaves some strata with very small sample sizes, as low as one, posing a problem for variance estimation. Guidance has been presented in the statistical literature on the desirable minimum stratum sample size in post-stratification, ranging from 10 to 20 (Särndal et al., 1992, p. 267; Wu and Thompson, 2020; Cochran, 1977; Kish, 1965). In particular, Kish (1965) recommended a minimum sample size of 10 when the true stratum weights are known, such as in our application. Westfall et al. (2011) investigated the minimum sample size needed for accurate estimation in the context of forest inventories via simulation studies and found that stratum sizes of 10 or more gave stable mean estimates. Thus, we merge more strata than the initial analysis to ensure a minimum sample size of 10 facilities per stratum. Subtypes which had a sample size less than 10 were merged with strata with larger sample sizes that had similar characteristics. A remaining subtype with a small sample size which is not similar to any other subtypes (701 - surface waste battery) was merged with other strata which had small sample sizes even after merging with similar strata (901 - water source and 902 - water source battery) to form a stratum of sufficient size. This results in a total of $H = 18$ strata. Stratum definitions are given in Table 4.1. Sample and population sizes were derived from data provided by the MERC, and differ from those used in the other published analysis of these data (Johnson et al., 2023). Details on the sample and population sizes are given in Appendix C.1. Note that due to differences in how data for wells was reported, they require slightly different methods, see Appendix C.2.

Since a facility being surveyed means that all components at that facility are surveyed, the selection of components can be viewed as cluster sampling, where all elements of the cluster are surveyed. Let $h(p)$ represent the stratum that component p belongs to, and let Ω_h be the set of components belonging to stratum h. Let $n_{h(p)}$ be the number of facilities in stratum $h(p)$ in the stage I sample and $N_{h(p)}$ be the number of facilities in stratum $h(p)$ in the population. Further, let $f(p)$ represent the facility containing component p. Then the inclusion probabilities for stage I are given by

$$\pi_p^{\mathrm{I}} = \frac{n_{h(p)}}{N_{h(p)}} \tag{4.21}$$

and

$$\pi^{\mathrm{I}}_{pl} = \begin{cases} \frac{n_{h(p)}}{N_{h(p)}} & \text{if } f(p) = f(l) \\ \frac{n_{h(p)}(n_{h(p)}-1)}{N_{h(p)}(N_{h(p)}-1)} & \text{if } f(p) \neq f(l), h(p) = h(l) \\ \frac{n_{h(p)}}{N_{h(p)}} \frac{n_{h(l)}}{N_{h(l)}} & \text{if } h(p) \neq h(l). \end{cases} \tag{4.22}$$

Note that the assumed design implies that components in separate strata are independent. Because of this, we can easily calculate stratum-specific total estimates using

$$\hat{T}(h) = \sum_{p \in S^{\mathrm{I}} \cap p \in \Omega_h} \frac{1}{\pi^{\mathrm{I}}_p} \hat{\overline{Y}}_{p\cdot\cdot},$$

which has corresponding variance estimate

$$\hat{\mathrm{V}}^{3 \text{ stage}}(h) = \sum_{p \in S^{\mathrm{I}} \cap p \in \Omega_h} \sum_{l \in S^{\mathrm{I}} \cap l \in \Omega_h} \frac{\left(\pi^{\mathrm{I}}_{pl} - \pi^{\mathrm{I}}_p \pi^{\mathrm{I}}_l\right)}{\pi^{\mathrm{I}}_{pl}} \frac{\hat{\overline{Y}}_{p\cdot\cdot}}{\pi^{\mathrm{I}}_p} \frac{\hat{\overline{Y}}_{l\cdot\cdot}}{\pi^{\mathrm{I}}_l} + \sum_{p \in S^{\mathrm{I}} \cap p \in \Omega_h} \frac{\hat{\mathrm{V}}_{p\cdot\cdot}}{\pi^{\mathrm{I}}_p}.$$

We also have

$$\hat{\mathrm{V}}^{3 \text{ stage}} = \sum_{h=1}^{H} \hat{\mathrm{V}}^{3 \text{ stage}}(h), \tag{4.23}$$

which means $\hat{\mathrm{V}}^{3 \text{ stage}}(h)$ is an estimate of stratum h's contribution to the variance in the province-wide methane emissions estimate.

Stage II: Simple Random Sample

The stage II design describes which days are chosen to survey the components selected in stage I. We will let d_p be the number of days on which component p was surveyed, whereas D is the number of days in the population of interest. For example, if the data are being used to compile a yearly inventory, the population is $D = 365$ days in the year. We assume here that the days where sampling occurred represent a simple random sample from the population of D days. This is a simplification of the true underlying process, and the implications of this are discussed in Section 4.5.

For a simple random sample of d_p out of D days, the inclusion probabilities are given by

$$\pi^{\mathrm{II}}_{t|p} = \frac{d_p}{D} \tag{4.24}$$

and

$$\pi^{\text{II}}_{tu|p} = \begin{cases} \pi^{\text{II}}_{t|p} & \text{if } t = u; \\ \frac{d_p(d_p-1)}{D(D-1)} & \text{otherwise.} \end{cases} \tag{4.25}$$

The general formula for the variance estimate of $\hat{\overline{Y}}_{p\cdot\cdot}$ can be computationally intensive, so we derived a computationally simpler form specific to the SRS design at stage II when the IPW estimator is used:

$$\hat{\text{V}}^{\text{IPW}}_{p\cdot\cdot} = \frac{1}{D^2}\left[\sum_{t\in S^{\text{II}}_p} \frac{(D-d_p)D}{d_p(d_p-1)}(\hat{\overline{Y}}^{\text{IPW}}_{pt\cdot})^2 + \frac{D(d_p-D)}{d_p^2(d_p-1)}\left(\sum_{t\in S^{\text{II}}_p} \hat{\overline{Y}}^{\text{IPW}}_{pt\cdot}\right)^2 + \sum_{t\in S^{\text{II}}_p} \frac{D}{d_p}\hat{\text{V}}^{\text{IPW}}_{pt\cdot}\right].$$

An analogous variance formula for the Hájek estimator is

$$\hat{\text{V}}^{\text{Hájek}}_{p\cdot\cdot} = \frac{1}{D^2}\left[\sum_{t\in S^{\text{II}*}_p} \frac{D\left\{D-1-\hat{\phi}_{pt\cdot}(d_p-1)\right\}}{d_p(d_p-1)}\left(\frac{\hat{\overline{Y}}^{\text{Hájek}}_{pt\cdot}}{\hat{\phi}_{pt\cdot}}\right)^2 + \frac{D(d_p-D)}{d_p^2(d_p-1)}\left(\sum_{t\in S^{\text{II}*}_p} \frac{\hat{\overline{Y}}^{\text{Hájek}}_{pt\cdot}}{\hat{\phi}_{pt\cdot}}\right)^2 + \sum_{t\in S^{\text{II}*}_p} \frac{D}{d_p\hat{\phi}_{pt\cdot}}\hat{\text{V}}^{\text{Hájek}}_{pt\cdot}\right].$$

Derivations are available in Appendix C.6.

An estimation issue arises for components where $d_p = 1$ when using IPW, and components for which there is only one day with detections when using Hájek. The variance estimation formulas result in undefined expressions under these conditions. If no methane was detected at component p on any survey days, we conservatively assume that this component has an emission rate of zero for the population of D days and therefore contributes zero to the total and zero to the variance. If component p has only one day of measurement data (either due to $d_p = 1$ for IPW or by only having detections on one day for Hájek), we pool the $\hat{\text{V}}_{p\cdot\cdot}$ from all other components in the same stratum which were surveyed more than once and impute the average. In the BC data, there are 17 components that were surveyed only once ($d_p = 1$), and 208 components that have only one day of non-zero measurement data.

4.3.5 Estimation Accounting for Measurement Error

In the previous methods, in line with survey sampling literature, we treat the values of Y_{ptq} as fixed and known. In reality, the measurements vary around the true emission rate. To accommodate this, we introduce a superpopulation model by assuming the true emission rate Y_{ptq} follows a distribution given the measured emission rate M_{ptq}, and that the measured rate is non-zero. In particular, we can assume that the true emission rate

follows the distribution in Equation (4.2) when methane is detected. We denote this measurement model as m. Let $\mathbf{E}_{ptq}$ and $\mathbf{Var}_{ptq}$ denote the expectation and variance with respect to the ptq-design, that is, stages I, II, and III of the sampling design, treating the measured emission rates as fixed and known. Now, let $\mathbf{E}_m$ and $\mathbf{Var}_m$ denote the expectation and variance with respect to the measurement model m while the sampling design is fixed. Finally, let $\mathbf{E}$ and $\mathbf{Var}$ represent the total expectation and variance, that is, the expectation and variance with respect to both the ptq-design and measurement model m. Previously, our target of estimation was $T = \sum_{p \in U^{\mathrm{I}}} \overline{Y}_{p\cdot\cdot}$, which is the expectation of $\hat{T}$ with respect to the ptq-design. Now, we introduce the total expectation of $\hat{T}$,

$$\tau = \mathbf{E}_m(T) = \mathbf{E}_m[\mathbf{E}_{ptq}(\hat{T})] = \mathbf{E}(\hat{T}), \tag{4.26}$$

to be our new target of estimation. Given a sample of size B drawn from the distribution of Y_{ptq} given M_{ptq}, we can calculate B estimates of T, $\hat{T}^b$, $b = 1, \ldots, B$. Then we estimate τ with

$$\tilde{\tau} = \frac{\sum_{b=1}^{B} \hat{T}^b}{B}, \tag{4.27}$$

where the hat overscore denotes an estimator with respect to the ptq-design and the tilde overscore denotes an estimator with respect to both the ptq-design and model m.

The total variance of $\tilde{\tau}$, $\mathbf{Var}(\tilde{\tau})$, can be estimated according to

$$\tilde{\mathrm{V}}^{\mathrm{Total}} = \tilde{\mathrm{V}}^m + \tilde{\mathrm{V}}^{\mathrm{I}} + \tilde{\mathrm{V}}^{\mathrm{II}} + \tilde{\mathrm{V}}^{\mathrm{III}}, \tag{4.28}$$

where

$$\tilde{\mathrm{V}}^m = \frac{\sum_{b=1}^{B} (\hat{T}^b - \tilde{\tau})^2}{B - 1}$$

is the estimated contribution to the total variance of $\tilde{\tau}$ from measurement uncertainty, and

$$\tilde{\mathrm{V}}^j = \frac{\sum_{b=1}^{B} \hat{\mathrm{V}}^{j}{}_b}{B}$$

for $j = \mathrm{I}, \mathrm{II}, \mathrm{III}$ represents the estimated contribution due to stages I, II, and III of the ptq-design respectively. A derivation of Equation (4.28) is available in Appendix C.6. A 95% Wald CI for $\tilde{\tau}$ is given by

$$\tilde{\tau} \pm 1.96 \sqrt{\tilde{\mathrm{V}}^{\mathrm{Total}}}. \tag{4.29}$$

An algorithm to estimate the methane emissions total and its variance with respect to the ptq-design and measurement model m is given in Algorithm 1. This algorithm addresses the four key sources of uncertainty in methane survey data from upstream OG facilities: sampling uncertainty from components and days, detection uncertainty, and measurement uncertainty.

Algorithm 1 Monte Carlo procedure to account for measurement error.

for $p \in S^{\mathrm{I}}$ **do**
  for $t \in S_p^{\mathrm{II}}$ **do**
    for $q \in S_{pt}^{\mathrm{III}}$ **do**
      for b in $1, \ldots, B$ **do**
        Draw a true emission rate Y_{ptq}^b given M_{ptq} using the distribution in Equation (4.2).
      end for
      Calculate ϕ_{ptq} using Equation (4.1) where Y is the mean of the B Y_{ptq}^b's, $u = u_{ptq}$, and $a = a_{ptq}$.
    end for
    if Hájek estimator used **then**
      Calculate $\hat{\phi}_{pt\cdot}$ using Equation (4.20).
    end if
  end for
end for
for b in $1, \ldots, B$ **do**
  Calculate $\hat{T}^b$, $\hat{\mathrm{V}}^{\mathrm{I}b}$, $\hat{\mathrm{V}}^{\mathrm{II}b}$, and $\hat{\mathrm{V}}^{\mathrm{III}b}$ using Equations (4.5), (4.11), (4.10), and (4.9).
end for
Calculate $\tilde{\tau}$, $\tilde{\mathrm{V}}^{\mathrm{Total}}$, and CI for $\tilde{\tau}$ as described in Section 4.3.5.

4.3.6 Simulation Methodology

The dataset collected in aerial surveys in BC has several features which may challenge the normality assumption which is relied on in the construction of Wald CIs. The sample sizes at stages II and III are very small, and it is known that the distribution of emission rates from emitting OG facilities is right-skewed. A strength of the Wald CI is that it is considerably simpler to understand and requires almost no computational effort. In support of our aim to provide accessible methods for estimating methane inventories, we perform a simulation study to investigate the properties of the proposed estimators and Wald CIs under similar conditions to the BC data, in the absence of measurement error.

We created an artificial population, U^{I}, of facilities. We selected four strata from the BC dataset which represented different emission rate distributions and sampling fractions. These strata were crude oil single-well batteries (83% of facilities represented in S^{I}, fifth lowest emitting stratum in data analysis), meter stations (56% of facilities represented in S^{I}, third lowest emitting stratum), sweet gas plants (84% of facilities represented in S^{I},

second highest emitting stratum), and compressor stations (18% of facilities represented in S^{I}, highest emitting stratum). We use the same population sizes N_h as we have in the real population as shown in Table 4.1. Then, a number of components was randomly assigned to each facility in the artificial population, drawn from a discrete uniform distribution from one to 50. Each component was assigned as either emitting or not, drawn from a Bernoulli distribution with a probability of emitting equal to 0.055. These two distributions were chosen so that the simulated population had a similar true total to the estimated total from the four strata in the data analysis.

Next, for 365 days, emitting components were assigned a daily mean emission rate drawn from a stratum-specific distribution. To obtain stratum-specific emission rate distributions, the non-zero emission rates in the motivating dataset were bias-corrected as suggested in Equation (4.3). Then, a log-normal distribution was fit to the bias-corrected emission rates in the motivating dataset for each stratum to give unique log-normal distributions. Distributions were fit using the fitdistrplus package in R using the moment-matching method (Delignette-Muller and Dutang, 2015). The daily emission rate for component p was sampled from the log-normal distribution corresponding to the stratum of component p. Note that this is not exactly the same quantity as $\overline{Y}_{pt\cdot}$, where the mean is taken over the discrete number of passes.

Once daily average emission rates were assigned, for each emitting component and day, a number of passes is randomly assigned by drawing from the distribution of passes in the observed data. An instantaneous emission rate Y_{ptq}, altitude a_{ptq}, and wind speed u_{ptq} are assigned at each pass. The height and wind speed are drawn from normal distributions with mean and standard deviation equal to those observed in the data. The wind speed distribution is truncated at zero. Emission rates are drawn from a normal distribution with mean equal to the previously assigned daily mean and standard deviation proportional to the daily mean. The proportion is determined by taking the median of the ratio between mean and standard deviation observed in the data, by strata.

Sampling was carried out following the three-stage sampling design described in Section 4.3.4. First, a simple random sample of n_h facilities was selected out of N_h in the artificial population. All components belonging to selected facilities are part of S^{I}, the stage I sample. Next, a SRS of two out of 365 days was selected for each component in S^{I} to make up S_p^{II}. Finally, for all $p \in S^{\mathrm{I}}$ and $t \in S_p^{\mathrm{II}}$, detections were simulated in each pass $p = 1, \ldots, Q_{pt}$ according to a Bernoulli distribution where the success probability is calculated using Equation (4.1) with Y_{ptq}, u_{ptq}, and a_{ptq}. The passes with successful detections form S_{pt}^{III}.

For each sample, the stratum total was estimated using two types of estimators and

two assumed stage II designs, giving four different estimation approaches in all. Either IPW or Hájek estimators were used, assuming a stratified SRS at stage I, and for stage II, either a SRS of two out of 365 or a SRS of two out of two (ignoring stage II uncertainty) as described in Section 4.3.4. All estimation methods follow those described in Sections 4.3.2 and 4.3.3. The population total was estimated by summing the strata totals. 95% CIs for individual strata and for the population were calculating using the Wald method described in Equation (4.8). The sampling and estimation process was carried out 5000 times, giving a sample of 5000 stratum-level and overall population total estimates, denoted $\hat{T}(h)^c$ and $\hat{T}^c$ for $c = 1, \ldots, 5000$. The bias, variance, and mean-squared-error (MSE) of $\hat{T}$ for the four different methods were estimated according to

$$\mathbf{Bias}(\hat{T}) = \left(\frac{\sum_{c=1}^{5000} \hat{T}^c}{5000} \right) - T,$$

where T is the true value calculated for the artificial population,

$$\mathbf{Var}(\hat{T}) = \frac{\sum_{c=1}^{5000} (\hat{T}^c - \frac{\sum_{c=1}^{5000} \hat{T}^c}{5000})^2}{5000 - 1}$$

and

$$\mathbf{MSE}(\hat{T}) = \mathbf{Var}(\hat{T}) + \left\{ \mathbf{Bias}(\hat{T}) \right\}^2. \tag{4.30}$$

Percent bias is calculated by dividing the bias by the true value T and multiplying by 100%. The coverage probability of 95% CIs were calculated according to

$$\text{Coverage probability}(\hat{T}) = \frac{\sum_{c=1}^{5000} \mathcal{I}_{T \in (l^c, u^c)}}{5000} \tag{4.31}$$

where (l^c, u^c) represents the lower and upper bounds of the 95% Wald CI calculated for the c^{th} estimate, $\hat{T}^c$. The bias, MSE and coverage probability for the stratum-level estimator $\hat{T}(h)$ were calculated similarly.

4.4 Results

4.4.1 Simulation Results

The percent bias of the estimators for the four strata and the artificial population are shown in Figure 4.2. We differentiate only between the estimator type (IPW or Hájek)

and not between assumed stage II design in showing these results because the different designs only affect the variance, not the point estimate. For all strata and the population, the Hájek estimator results in a larger magnitude of bias than using IPW, and tends to be biased upwards. The largest magnitude of percent bias is seen when estimating the total for meter stations, at 4.2% bias. The percent bias tends to be larger for strata with lower true totals. This is likely due to the lower detection probabilities, leading to smaller sample sizes at stage III. For the population total, both approaches have low bias below 0.5%.

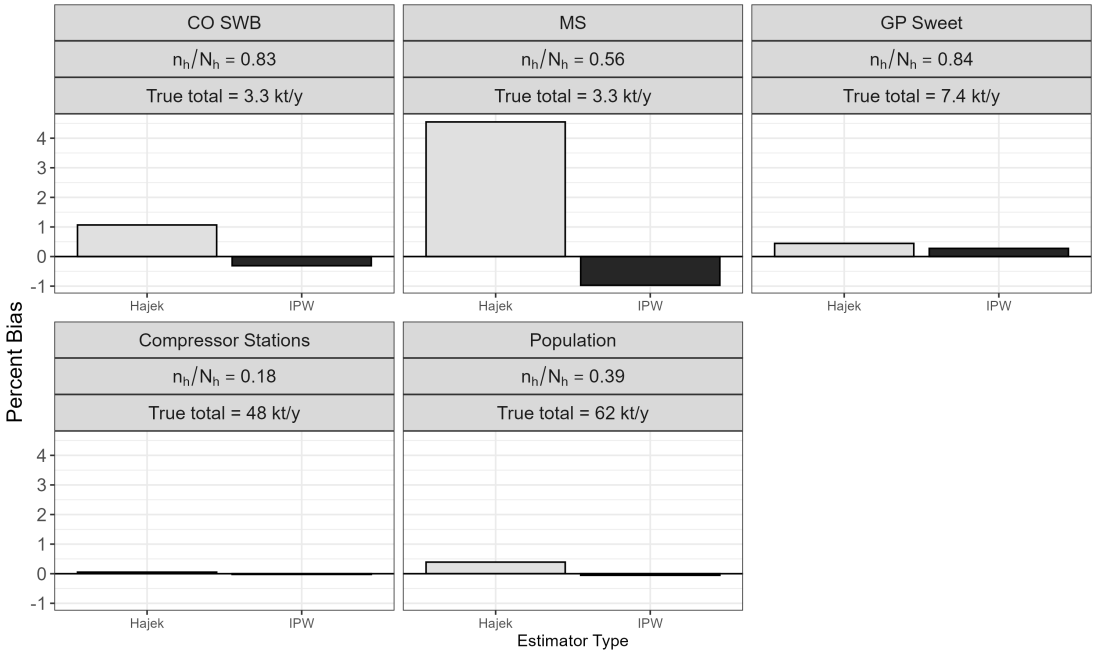


Figure 4.2: Percent bias for different estimators of the stratum totals $T(h)$ and provincial total T.

The MSE results of the simulation study are shown in Figure 4.3 for the approaches which considered uncertainty at stage II. The MSE are very similar between the Hájek and IPW approaches for all strata, except for the meter stations stratum, where IPW has noticeably lower MSE. In all cases, the biggest contributor to the MSE is the variance. Despite the Hájek estimator being more efficient than IPW in many situations, in this simulation study we see little to no efficiency gains in using the Hájek estimator. Conclusions are similar for the approaches which did not consider stage II uncertainty.

The coverage probabilities of the estimation approaches are shown in Figure 4.4. The

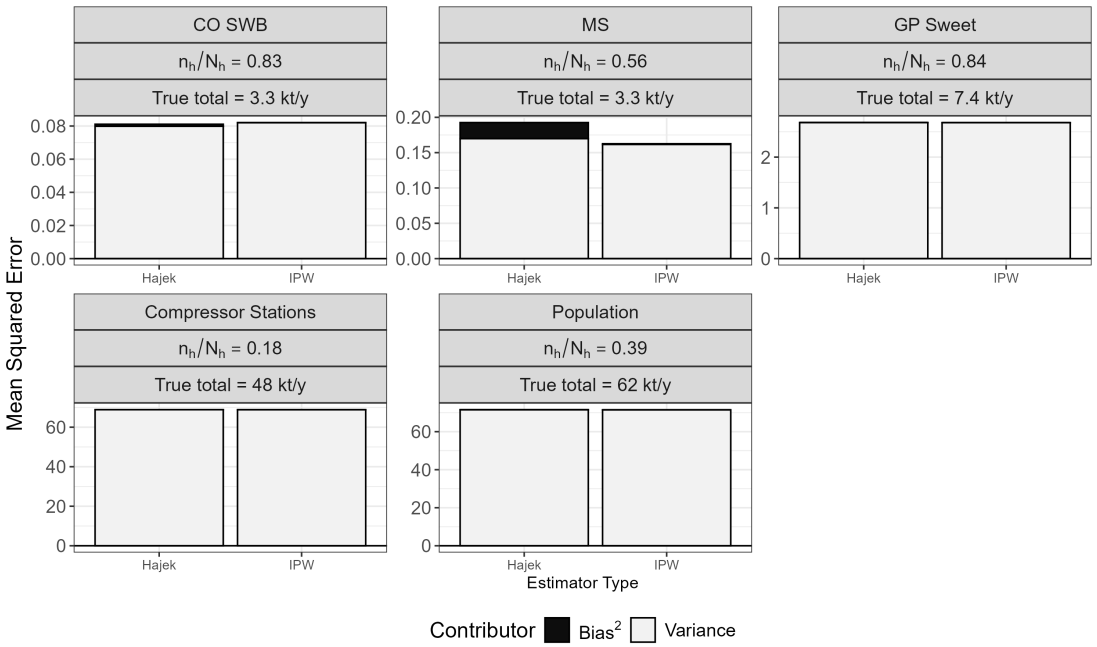


Figure 4.3: Mean squared error for different estimators of the stratum totals $T(h)$ and provincial total T when considering stage II uncertainty, including the contribution from bias and from variance.

approaches which used the true stage II sampling design ($D = 365$) give coverage the closest to the nominal level. The differences between IPW and Hájek are marginal. When considering the true stage II design, the coverage probability is at the nominal level for crude-oil single well batteries, only slightly below 95% for meter stations, compressor stations, and the population total, and below 90% for sweet gas plants. Ignoring the stage II uncertainty by setting $D = d_p$ in the analysis leads to reduced coverage probability for all strata. This is especially pronounced for the sweet gas plants stratum, which has the lowest coverage probabilities in general. This is likely due to the underlying emission distribution of this stratum. The log-normal distribution fit to the sweet gas plants data had the largest variance out of four stratum included in the simulation, which leads to skewness in the log-normal distribution, which could explain why the performance is worse for this stratum.

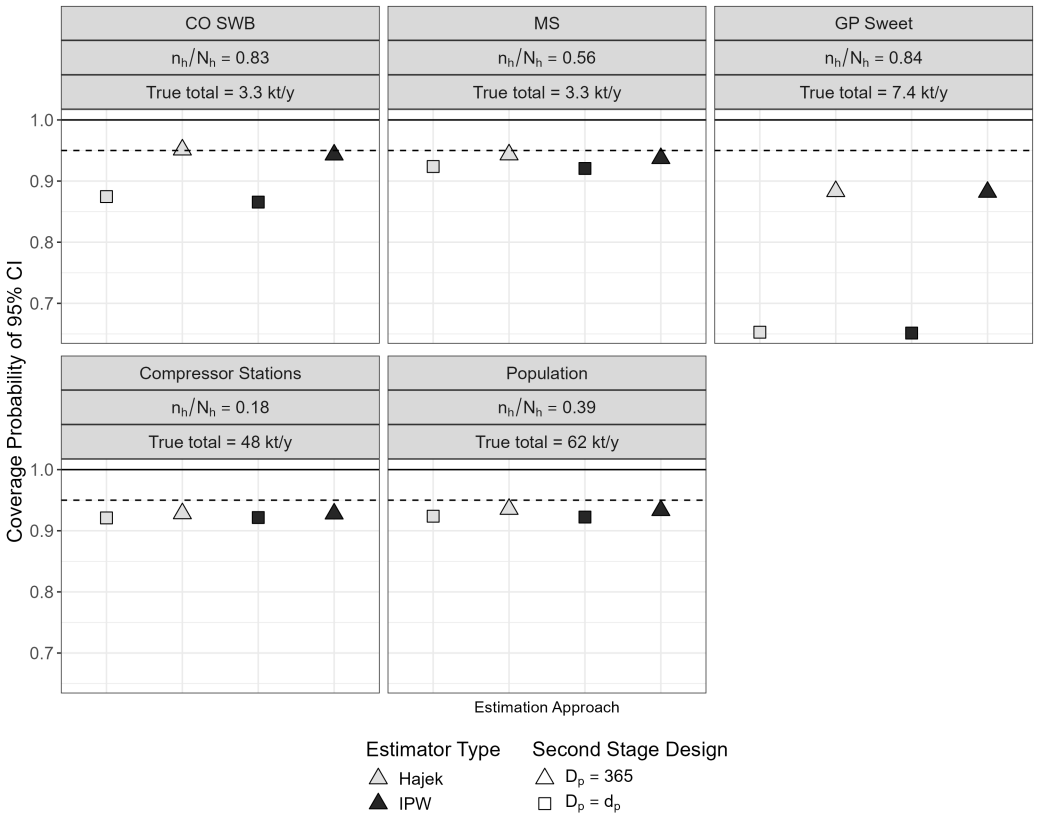


Figure 4.4: Coverage probability of 95% Wald CIs for all estimation approaches by stratum. The dotted line represents the nominal coverage probability.

4.4.2 Data Analysis

In this section, we apply the methods described in Section 4.3 to calculate a provincial methane inventory for the data described in Section 4.2. The data analysis was performed with the R package `methaneInventory`, available on GitHub from https://github.com/augustinewigle/methaneInventory.

We analyse the BC data in eight different ways, using different approaches to estimate $\overline{Y}_{pt\cdot}$ (either IPW or Hájek estimator), different approaches to model the selection of days in stage II, and different approaches to address measurement uncertainty. We also compare the results to the published inventory compiled in Johnson et al. (2023) using the nested MC approach. The different stage II and measurement uncertainty approaches are detailed below.

We use two different values for D in the stage II design to illustrate the ability of our method to incorporate uncertainty due to the number of days sampled. First, we assume that the population of interest is represented by the d_p days on which we sampled component p. In other words, we let $D = d_p$ for component p. From this perspective, the sampling at stage II contributes nothing to the variance, since we are sampling the entire population. This is an equivalent treatment to the main analysis presented by Johnson et al. (2023). Second, we set $D = 365$, treating the year as the population of interest. Changing the D value only impacts the variance estimate and does not change the point estimate.

We also consider measurement error in two different ways. First, we investigate the possibility of avoiding any MC steps by performing a bias correction to the measured emission rates. We multiply all measured emission rate values by the bias correction factor in Equation 4.3, and then apply the methods in Section 4.3.2 without using the measurement uncertainty MC. This provides a rough estimate where we account for measurement bias, but assume that measurement uncertainty is otherwise negligible. The other approach is to consider both the bias and uncertainty coming from measurement error by applying Algorithm 1. No bias correction to the measured rates is required with this approach because the MC simultaneously corrects for bias and variance. When the measurement uncertainty algorithm is applied, we use $B = 5000$ MC iterations. Convergence was assessed by plotting $\tilde{V}^{I} + \tilde{V}^{II} + \tilde{V}^{III}$ versus the number of MC iterations for each stratum. The different approaches to measurement error can impact both the point estimate and the variance estimate.

Estimates of the total measurable methane emissions for the province of BC in 2021 for eight different multi-stage approaches and the nested MC approach of Johnson et al. (2023) are shown in Figure 4.5. The point estimates within the same estimation approach (multi-stage Hájek or multi-stage IPW) are very similar regardless of the uncertainty sources considered. The multi-stage Hájek approach gives point estimates of the provincial emissions that are slightly higher than the nested MC approach, while the multi-stage IPW approach leads to slightly lower point estimates. This is not too surprising based on the simulation study results, where we observed upwards bias with the Hájek estimator for some strata. As expected, considering additional sources of uncertainty (stage II, measurement) increases the variance estimate for both multi-stage approaches.

Based on the simulation results, we expect the multi-stage IPW approach to be unbiased with respect to the *ptq*-design and we should account for all possible sources of uncertainty. Thus, we expect the most reliable multi-stage approach to be the multi-stage IPW estimator which accounts for stage II and measurement uncertainty. The 95% CI for provincial emissions using this approach is (85.2 kt/y, 132.4 kt/y), whereas the nested MC

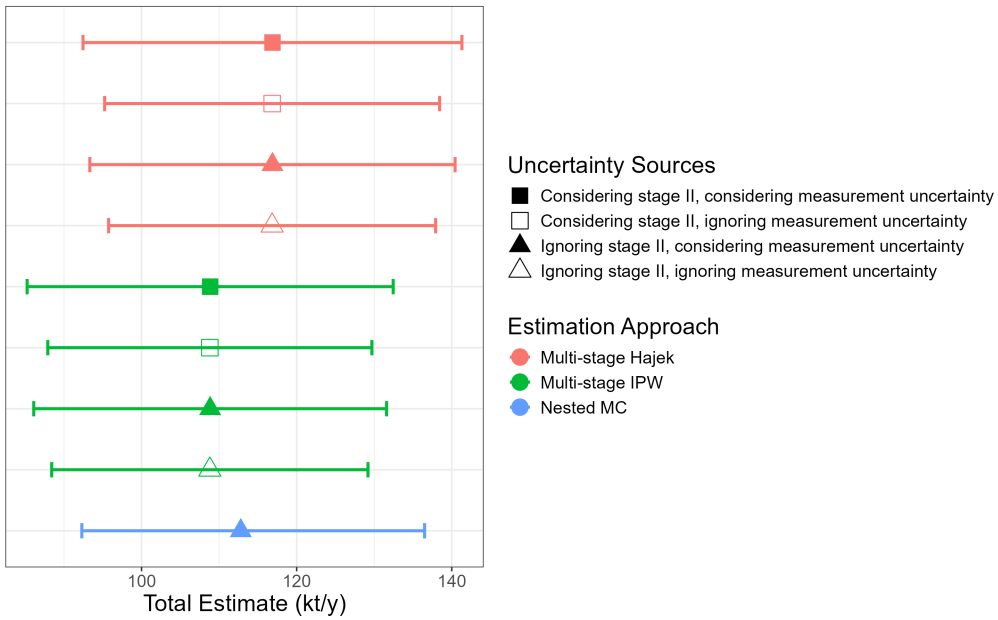


Figure 4.5: Point estimates and 95% CIs for all methods compared on the BC data. In the legend, considering stage II indicates that $D = 365$ whereas ignoring stage II indicates that $D = d_p$. Similarly, considering measurement uncertainty means the measurement uncertainty MC algorithm was used while ignoring measurement uncertainty means that the bias-correction method without the algorithm was used.

approach gives a 95% CI of (92.3 kt/y, 136.5 kt/y), which is slightly narrower (a length of 47.2 kt/y versus 44.2 kt/y). While the CIs overlap significantly, the multi-stage IPW approach suggests slightly lower emissions. This could be due to the use of the symmetric Wald CI in this approach, whereas the nested MC method uses quantile-based CIs that allow for asymmetry.

In Table 4.2, we show the total emissions estimate and variance estimates by source for each stratum, using the multi-stage IPW approach considering both stage II and measurement uncertainty. There is important variation between strata. Two strata which are not shown (enhanced recovery schemes, water and waste) had no non-zero methane measurements in the survey data, and so we conservatively estimate that they emit zero methane over the year. Apart from those two strata, the lowest emitting stratum is tank farms and natural gas liquid storage, which is estimated to emit 0.026 kt/y. This contrasts with the highest emitting stratum, compressor stations, which is estimated to emit 44.937 kt/y. The top four highest emitting strata account for over 80% of the total estimated

Table 4.2: Estimates of the total methane emissions in BC for 2021, variance estimates by uncertainty source, and total variance estimates by stratum using the multi-stage approach with IPW considering all three stages and measurement uncertainty. Due to rounding, small non-zero values are shown as 0.00. Acronyms: GP = gas plant, MWB = multi-well battery, LNGP = liquid natural gas plant, CO = crude oil, CTF = custom treating facility, GGS = gas gathering system, SWB = single-well battery, MS = meter station, TF = tank farm, NGL = natural gas liquids.

Stratum	Total Estimate (kt/y)	$\tilde{V}^{I}$	$\tilde{V}^{II}$	$\tilde{V}^{III}$	$\tilde{V}^{m}$	$\tilde{V}^{Total}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.001	0.000	0.000	0.001
Water Hub Battery	0.097	0.000	0.000	0.004	0.002	0.007
MS	0.162	0.003	0.001	0.000	0.001	0.006
CO SWB	0.194	0.001	0.003	0.003	0.002	0.008
Gas SWB	0.227	0.012	0.008	0.000	0.006	0.027
Disposal/storage	0.385	0.038	0.023	0.000	0.007	0.068
Mixed OG Battery	0.391	0.000	0.014	0.000	0.005	0.019
CO MWB	0.764	0.002	0.101	0.001	0.016	0.119
CTF and GGS	0.801	0.108	0.032	0.007	0.028	0.174
GP Acid Gas Low	3.805	0.267	0.184	0.003	0.072	0.525
Gas MWB (group)	4.754	1.022	0.771	0.001	0.432	2.227
Other GP and LNGP	8.473	0.446	2.056	0.005	0.472	2.978
Wells	12.364	9.136	0.680	0.151	1.704	11.670
Gas MWB (effluent)	14.227	1.398	1.613	0.010	0.426	3.448
GP Sweet	17.240	4.021	3.481	0.201	2.476	10.178
Compressor Stations	44.937	80.099	23.502	0.015	9.816	113.432
Population	108.847	96.552	32.470	0.401	15.561	144.984

emissions for BC. Compressor stations accounted for an estimated 41.3% of total provincial emissions. The next highest-emitting strata are sweet gas plants, wells, and multi-well gas batteries (effluent), contributing 15.8, 13.1, and 11.4% of emissions respectively. The remaining strata account for an estimated 18.4% of the total emissions. These findings align well with those obtained using the nested MC method, for example, they find the same top four emitting strata.

Turning to the variance decomposition, the variance estimate for the provincial total and the contribution from each uncertainty source using the multi-stage IPW estimator with and without measurement error are shown in Figure 4.6. With both approaches,

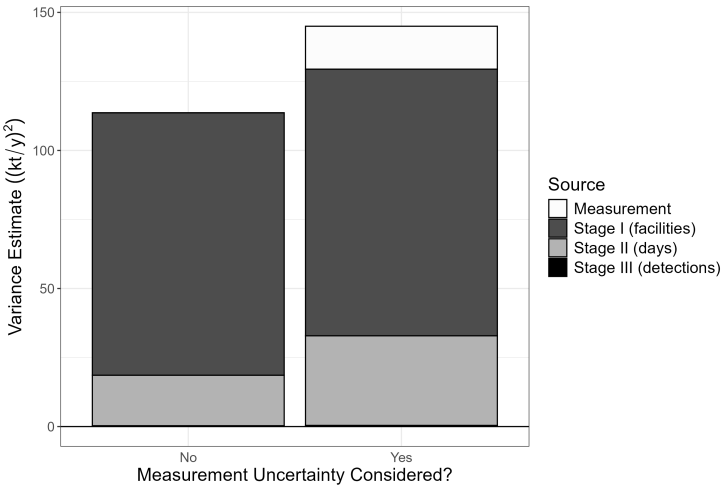


Figure 4.6: The estimated variance contributions for the provincial total from different uncertainty sources, with and without considering measurement error through the MC algorithm. Results are from the multi-stage IPW approach where stage II uncertainty was considered.

the biggest contributor to uncertainty is sampling error due to the facilities, followed by sampling error due to the days. Addressing these sources of uncertainty in future survey designs could lead to more precise estimates. A key difference between the approaches with and without measurement error is that the estimated variance contribution from each stage increases when measurement error is considered. We expect this is due to the fact that the underlying distribution for each measurement is right-skewed, and the bias-correction approach only addresses the bias, but does not adequately capture the variance. On the provincial level, when all four uncertainty sources are considered, detection uncertainty is the smallest contributor to the variance.

We can investigate the variance breakdowns within individual strata. Figure 4.7 shows the percentage of the total variance from each source for every stratum. Uncertainty in higher emitting strata tends to be dominated by stage I or stage II. Lower emitting strata are more likely to have measurement uncertainty and detection uncertainty contribute an important amount to the variance. This suggests that reducing the variance in a particular stratum total estimate will require a tailored, case-by-case approach.

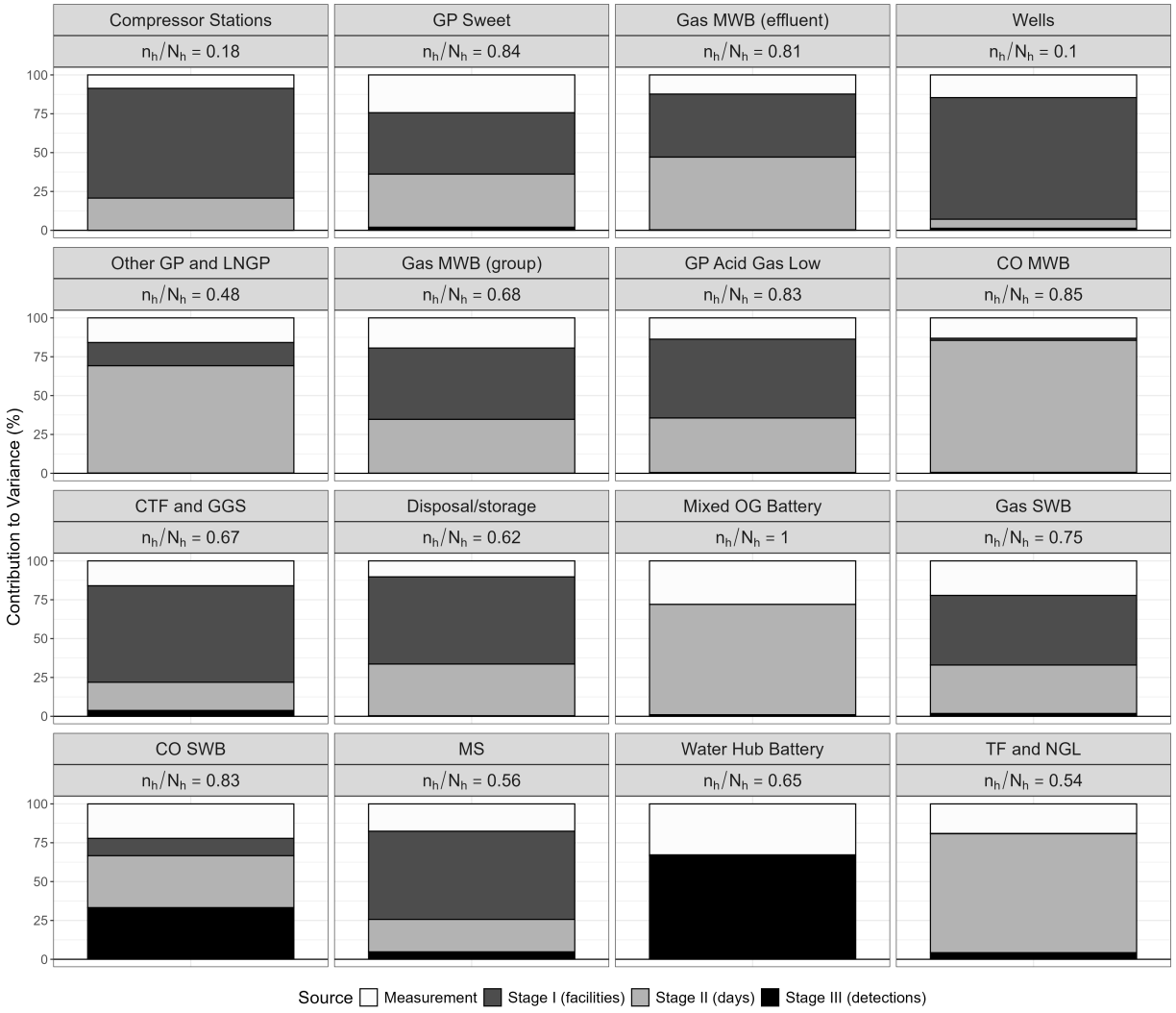


Figure 4.7: Percent contribution to the estimated total variance by uncertainty source for every stratum in the BC data. Strata are organized from highest (top left) to lowest (bottom right) emitting. The results are from the multi-stage IPW approach which includes stage II and measurement uncertainty. Acronyms: GP = gas plant, MWB = multi-well battery, LNGP = liquid natural gas plant, CO = crude oil, CTF = custom treating facility, GGS = gas gathering system, SWB = single-well battery, MS = meter station, TF = tank farm, NGL = natural gas liquids.

4.5 Discussion and Future Work

We showed how a multi-stage framework can be leveraged to quantify measurable methane emissions given data from a complex aerial survey. We also proposed a MC algorithm which leverages knowledge about the survey technology to incorporate measurement uncertainty into the multi-stage approach. Overall, the methods account for the four key sources of uncertainty in aerial surveys of methane emissions from OG production facilities: sampling components, sampling days, detection uncertainty, and measurement uncertainty. The results yield similar insights to the nested MC approach that is the current standard for developing measurement-based inventories, with the additional ability to explicitly decompose the variance contribution from different uncertainty sources.

An advantage of the multi-stage approach is its relatively low computational burden compared to the currently used approach. Our approach has reduced computational needs in two ways. First, the currently used approach relies on a complicated MC approach to account for measurement and detection error where numerical integration and root-finding methods are needed at every iteration (Johnson et al., 2023). By addressing detection error using IPW or Hájek estimation, we can significantly simplify the MC algorithm needed to quantify measurement uncertainty. Second, by deriving analytical formulas for estimators and CIs, we remove the need for a bootstrap or other MC approach to be nested within the measurement error MC. The existing nested MC approach requires a total of $B^{MC} \times B^{BS}$ MC iterations, where B^{MC} is the number of iterations in their initial MC and B^{BS} is the number of bootstrap samples taken in the mirror-match bootstrap. Alternatively, since our approach does not rely on a bootstrap procedure, our approach requires only B MC iterations in total, where B is the number of iterations in the measurement uncertainty algorithm. Johnson et al. (2023) require a total of 10^8 MC iterations to produce their inventory for BC, while we used only 5000. Additionally, the code used to implement the nested MC approach is not publicly available, while we provide an open-source R package, methaneInventory, which makes the methods described herein accessible, available at https://github.com/augustinewigle/methaneInventory.

It is unclear whether the use of the Hájek estimator provides benefits over IPW in the context of estimating methane emissions from OG. Although it is known that the Hájek estimator is more efficient than IPW in many contexts, the variance estimates for the Hájek estimator were larger than those for IPW in the data analysis for the province and for a majority of strata (when stage II and measurement uncertainty were considered). It also gives larger point estimates for the totals. Another downside to the Hájek estimator is that it requires manipulation of the stage II and III designs in order to analyse the data, which interferes with the interpretation of the variance decompositions for those stages. Overall,

more investigation is needed to understand which stratum and population characteristics lead to efficiency gains from using the Hájek estimator. Our proposed Hájek approach may be useful in other survey contexts where detection error or non-response contributes significantly to the variance and also leads to empty samples.

In the data analysis, we assume the design at stage II is a SRS of d_p out of D days. In reality, the sampling process was more complex. A second day of sampling for a component p was only carried out if there were any detections at the entire site containing p on the initial survey day. This means that the number of days p is sampled, d_p, is actually random, which is not considered in the current variance estimate. The probability that two surveys are conducted at component p is related to the probability of having any detections at the site containing p on day t_1, where t_1 is the day where the initial survey was conducted for component p. This probability is given by

$$\gamma_p = 1 - \prod_{u \in \text{site} \supset p} \prod_{q=1}^{Q_{pt_1}} (1 - \phi_{pt_1 q}).$$

If γ_p is approximately zero or one, then d_p is effectively fixed. We calculated conservative estimates of γ_p, where the facility was used in place of the site. The first quartile of the estimated γ_p's was 1.00, showing that the assumption that d_p is fixed is reasonable.

Another way in which the data departs from the assumed SRS of days is that sampling was only possible over a relatively short time period. This limits our ability to capture any potential long-term temporal patterns, such as seasonal variation in emission rates. In the SRS, we assume all 365 days have an equal chance of being selected. In reality, only days that are relatively close together have a non-zero probability of being selected, which requires complex modelling to describe. We chose to use a SRS because it is a simple way to incorporate the fact that we want to estimate a quantity that relates to 365 days using only one or two days of data. Johnson et al. (2023) performed a sensitivity analysis where the day-to-day variability was exaggerated via a bootstrap procedure to investigate its impact on the estimate and its uncertainty. In future surveys, sampling could be carried out in a way which more closely resembles SRS and covers a wider temporal range. However, feasibility considerations may prevent the days from being selected using a true probability design. The temporal variations that can be expected at OG facilities are still an open question in the literature. More complex and realistic sampling designs could be implemented in the multi-stage approach as new information on temporal variations in methane emissions from OG becomes available.

The sampling at stage I is also impacted by practical constraints which make following any probability design difficult. Analysing the data as if it were a probability sample is

in line with the established method where a stratified sampling scheme is assumed in the mirror-match bootstrap (Johnson et al., 2023). It is important to be aware of the potential implications of the simplifying assumptions made when translating the true data collection process to a statistical probability design to be used in the analysis. In particular, any facilities which are deemed inaccessible will have a true inclusion probability of zero. If the part of the population which has zero probability of being sampled differs from the population with non-zero probability of being sampled, the estimates could be biased. For example, if the northern facilities in BC which could not be sampled due to their remoteness tend to emit more than the sampled southern facilities, the results of the survey will underestimate total emissions since the surveyed facilities do not represent the population. Efforts should be made to include as many facilities as possible in the list of candidate sites for surveys. If it is necessary to exclude a group of facilities, any possible systematic differences between the surveyed facilities and the excluded facilities should be investigated and reported with the inventory. Additionally, results can be presented for the population (involving extrapolation for the portion of the population which had no chance to be included) alongside an estimate specifically for the surveyed facilities (that is, the facilities in the population which had a non-zero inclusion probability) which may not represent the population of interest, but will be unbiased. In the case of the BC data, we would report the inventory estimated for the southern facilities alongside the province-wide inventory. This is as simple as estimating an inventory where the facility population sizes are equal to the sample sizes. The results for this analysis using the multi-stage IPW estimator, the measurement uncertainty MC and considering stage II uncertainty are available in Appendix C.8.8. If data about the systematic differences between included and excluded facilities are available, an area of future work would be to develop weights which account for the sampling bias.

A key strength of our approach is that it can be leveraged in the planning of future surveys. Uncertainty due to the sampling of facilities and days contribute significantly to the uncertainty of population estimates. These also happen to be the two sources of uncertainty over which we have the most control, assuming we are not in a position to improve the technical abilities of the technology used to make measurements. The variance contribution from stages I and II can easily be predicted by plugging different sample structures into the true variance formulas for $\hat{T}$ given in Appendix C.4, where different representative values for $\overline{Y}_{p\cdot\cdot}$ based on previous data can be used. This gives a simple approach to identify efficient sampling designs in terms of sampling error, something that is not possible in the existing approach (Johnson et al., 2023). A rigourous approach to determining an optimal survey design is a possible avenue of future work.

We performed a simulation study evaluating the performance of the multi-stage esti-

mators in the absence of measurement error, and found that they have good performance when the daily design is properly accounted for. An important next step is to conduct a more extensive simulation which considers measurement error. This will enable the validation of the measurement uncertainty MC approach described in Section 4.3.5, as well as a formal comparison to the bias-correction approach used in the data analysis.

It is important to note that the approach proposed herein only estimates measurable emissions; that is, emissions which have a non-zero detection probability. Estimating emissions which are too small to be detected using GML require alternative data collection and estimation techniques. Methods for estimating unmeasurable emissions are covered in Johnson et al. (2023). Due to the good performance of the GML technology, measurable emissions have been found to account for a majority of total methane emissions from upstream OG in multiple studies of different oil-producing regions (Johnson et al., 2023; Conrad et al., 2023c,a). The presented methods are therefore an important contribution to advancing the ease and accuracy of estimating OG methane emissions as a whole. The methods can also be adapted to any MDQT, provided that the probability of detection function and measurement error distribution are available. Accurate statistical models for measurement error from a variety of MDQTs are therefore important for enabling the broader use of these methods.

Chapter 5

A Bayesian Approach for Estimating Uncertainty in Measurements from Methane Emission Quantification Technologies

5.1 Introduction

Due to the OG industry's significant contribution to methane emissions, many countries have enacted policy interventions focused on reducing emissions from this sector. Interventions include setting targets for methane emissions reductions and mandating LDAR programs (Climate and Clean Air Coalition, 2021; Environment and Climate Change Canada, 2016). These data-driven initiatives rely on measurements of methane emissions from a wide range of technologies, and are thus subject to quantification uncertainty. It is important to characterize the measurement uncertainty associated with a jurisdictional methane inventory, such as one compiled using the methods from Chapter 4, as year-to-year trends in inventories can inform future climate policy decisions and only considering point estimates may obscure or exaggerate important changes (Solazzo et al., 2021; MacKay et al., 2021; Conrad et al., 2023c). LDAR programs should be optimized with quantification uncertainty in mind to give the best trade-off between cost and emissions reductions, as high quantification uncertainty can impact the cost-effectiveness (Fox et al., 2021b). In order for methane inventories to be reported with appropriate uncertainty and LDAR programs to be designed optimally, accurate and accessible models for the uncertainty associated

with measurements from MDQTs are needed.

A wide range of technologies have been deployed to quantify methane emissions from the OG industry, including quantitative OGI (QOGI) cameras (Wilson, 2016; Mitchell et al., 2015; Kang et al., 2022); stationary (Heltzel et al., 2022) and mobile (Brantley et al., 2014; Atherton et al., 2017; Caulton et al., 2018) methane concentration sensors; and aerial (Conrad et al., 2023c; Sherwin et al., 2021) and satellite-based (Varon et al., 2019) measurements. All of these systems use a physics-based measurement model to produce an emission rate.

The physics-based measurement models can provide insights into the uncertainty of methane emission estimates, and how specific technologies should be deployed to minimize these uncertainties. Measurement models may be very complex or unavailable to researchers due to proprietary aspects of the technology. Further, results of a purely physics-based uncertainty estimate may not agree with what is observed in real-world scenarios due to missing or inadequate modelling of external uncertainty sources. Moreover, existing physics-based uncertainty analyses do not typically include guidance for how the results should be applied in practice, and there is a lack of consistency in reporting of the results (Caulton et al., 2018). For example, most published physics-based approaches do not show how their results should be used to derive a 95% confidence interval based on a given measurement from the technology. This presents a barrier to the application of accurate measurement uncertainty models in the calculation of inventories or the modelling of LDAR programs.

Data-driven approaches are an alternative to physics-based approaches which rely on controlled release (CR) data, where methane is released at known rates, and technology providers take measurements. The measurements can then be compared to the true release rates to learn about the expected error of the technology. Many measurement campaigns have been conducted with the goal of assessing the performance of MDQTs (Edie et al., 2020; Singh et al., 2021; Johnson et al., 2021; Bell et al., 2022; Sherwin et al., 2021; Conrad et al., 2023b). However, data-driven approaches employed on these measurements have mainly been limited to linear regression approaches (Edie et al., 2020; Singh et al., 2021; Johnson et al., 2021; Bell et al., 2022; Sherwin et al., 2021), with the exception of Conrad et al. (2023b), who provide an approach to derive the distribution of the true emission rate given a measurement from aerial GML technology, which was deployed in Chapter 4 to incorporate measurement uncertainty into a methane inventory. A limitation of existing approaches is that the models have a constant relationship between uncertainty and the true emission rate (either strictly additive or multiplicative). Also, approaches often do not account for the possibility that a technology may detect methane when there is none being released. Finally, there is a lack of reporting on how the methods should be applied

in different applications, such as in performing field measurements in LDAR programs or as part of measurement-based surveys.

To address the lack of flexible and widely applicable approaches for quantifying uncertainty in methane emissions measurements, in this chapter we propose a highly customizable data-driven approach for characterising measurement uncertainty in methane emission rates and provide guidelines for how the uncertainty results can be used in practice. In Section 5.2.1 we describe a flexible Bayesian model for the measured emission rate given a true emission rate which is fit to CR data. We then present an algorithm to find the posterior distribution of the true emission rate given a new measurement made in the field, such as part of a LDAR program, in Section 5.2.2. We describe a set of CRs for a variety of MDQTs in Section 5.3. In Section 5.4.1, we fit the Bayesian uncertainty model to the described CR data. We demonstrate the use of the algorithm for use on new measurement data in Section 5.4.2. Finally, we discuss our findings and possible avenues of future work in Section 5.5.

5.2 Methods

5.2.1 A Flexible Model for the Measured Emission Rate Given the True Rate

In this section, we describe a Bayesian model that answers the following question: For a given true emission rate, what range of measurements could be expected, given the observed CR data? The models described in this section are conditional on the data used to fit the models, so their prediction accuracy can only be guaranteed for measurements made under similar conditions. Additionally, the model is fit only to data where methane was detected, and therefore does not account for missed detections.

Let Y_i be the true emission rate associated with the ith observation in the measurement campaign, and M_i be the emission rate estimated by the technology for the ith observation, $i = 1, \ldots, n$ where n is the total number of observations for the given technology. Here, we describe a likelihood that models the relationship between Y_i and M_i, given that methane was detected. The relationship between Y_i and the bias and variability of M_i can be complicated, since both may change over the range of Y_i. Additionally, the relationship between Y_i and M_i may not be strictly linear, such as the case in Figure 5.1. The model must also account for the fact that all technologies may report a "false positive", that is, estimating a non-zero M_i when $Y_i = 0$.

A flexible likelihood that fulfills these requirements is given by

$$\log(M_i) = \log(\psi_i) + \epsilon_i, \tag{5.1}$$

where

$$\epsilon_i \sim \mathrm{N}(0, \mathbf{Var}(\epsilon_i)) \tag{5.2}$$

and

$$\psi_i = \mathrm{median}(M_i \mid Y_i).$$

This likelihood is normal on the log scale, which corresponds to a log-normal likelihood on the measurement scale. The median of M_i is modelled by a continuous piece-wise function of Y_i:

$$\psi_i = \begin{cases} \alpha_0 + \alpha_1 Y_i + \alpha_2 Y_i^2 & Y_i \leq \gamma; \\ \alpha_0 + \beta_0 + (\alpha_1 + \beta_1) Y_i & Y_i > \gamma, \end{cases}$$

which is quadratic for values of Y_i below a threshold γ and linear above γ. To ensure that the function is continuous at $Y_i = \gamma$, we impose the restriction that $\beta_0 = \alpha_2\gamma^2 - \beta_1\gamma$.

The specification of ψ_i can be modified to give the best prediction results and fit to the data, depending on the detection technology. For example, the threshold parameter γ, β_0 and β_1 could be removed, which would give a quadratic relationship characterized by α_0, α_1, and α_2 over the whole range of Y_i.

The model can be rewritten to facilitate interpretation by exponentiating both sides of Equation (5.1):

$$M_i = \psi_i \times e^{\epsilon_i}, \tag{5.3}$$

where ψ_i is the median measurement for a true emission rate of Y_i and e^{ϵ_i} is equal to one on average. That is, the function ψ_i adjusts for systematic bias and e^{ϵ_i} is a multiplicative error term that accounts for remaining variation.

The likelihood in Equation (5.1) is an extension of the scheme proposed by Conrad et al. (2023b). Their model is a special case of our likelihood where $\alpha_0 = 0$, $\alpha_2 = 0$, $\beta_0 = 0$, $\beta_1 = 0$, and $\mathbf{Var}(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$. In this version of the model, the median value of M_i has a multiplicative relationship with Y_i, that is, $\psi_i = \alpha_1 \times Y_i$, and the multiplicative error term has a constant variance over the range of Y_i. Our model expands on this in three ways. First, we allow for piecewise linear and quadratic relationships between the median measurement and Y_i and accommodate false positives rather than using a strictly linear model which does not allow false positives. Second, the inclusion of the threshold parameter γ allows more flexibility in modelling the relationship between the median of M_i over the range of Y_i rather than assuming a common median function for all Y_i. Third,

we investigate different variance structures for ϵ_i which can allow the variance to change with Y_i to more accurately model the patterns observed in CR data for some technologies. Another difference is that Conrad et al. (2023b) investigate different distributions for the error term, whereas we restrict ourselves to the log-normal distribution, but investigate different forms for the median and variance that are motivated by the data. Finally, we take a fully Bayesian approach to estimation and inference, whereas Conrad et al. (2023b) use a (frequentist) maximum likelihood approach.

It is important to note that in this model the errors are additive on the log scale, which implies multiplicative errors on the raw measurement scale as shown in Equation (5.3). The simplest way to model the variation is to set the variance of ϵ_i to a constant, $\mathbf{Var}(\epsilon_i) = \sigma^2$ for all i. Multiplicative errors may be suitable for lower and moderately-sized emission rates, but for large values of Y_i purely multiplicative errors may overestimate variability for some technologies. To accommodate this, we also propose using $\mathbf{Var}(\epsilon_i) = (\sigma^{-2} + Y_i/\eta)^{-1}$ as an alternative variance structure for ϵ_i which allows the variability of the error terms to decrease with increasing Y_i while also ensuring that the variance is continuous along the range of Y_i. With this variance form, when $Y_i = 0$, the variance on the log scale is equal to σ^2, which decreases as Y_i increases. The parameter η controls how quickly the variance decreases with Y_i. The approach to choosing an appropriate likelihood, including the variance specification, is described below.

Table 5.1 summarizes the parameters in the likelihood, their units, scenarios in which they may be included or removed, and interpretations of their roles in the model. As with any Bayesian model, an appropriate prior distribution depends on the context of the problem at hand, including pre-existing knowledge and the scale of the data. We specify prior distributions with the units in mind, as well as considering the role of each parameter in the model. A detailed discussion of the prior distribution specification is given in Appendix D.1.

Model Selection

The Deviance Information Criterion (DIC) was used to guide model selection as it combines goodness-of-fit while penalizing for model complexity to avoid over-fitting (Spiegelhalter et al., 2002). We are also guided by plotting prediction bands derived from the posterior predictive distribution of measurements for different values of Y_i over a scatterplot of the data used to fit the model. If the prediction bands show a much wider or narrower spread than the data used to fit the model, this indicates that the variance is not modelled well. The generalizability of the model may also be assessed by comparing the prediction bands to additional data that was excluded from the model fit ("external data"). If the model

Table 5.1: Summary of the units for each parameter in the likelihood, important to keep in mind when specifying prior distributions. Exclusion criteria: * = may be removed if technology does not have false positives in CR data, † = may be removed if its removal leads to a simpler model which has adequate fit and predictive performance.

Parameter	Units	Exclusion criteria	Interpretation
α_0	$[M_i]$	*	Median measurement when $Y_i = 0$
α_1	Unitless	None	Coefficient for Y_i in the quadratic relationship between ψ_i and Y_i (when $Y_i < \gamma$ if applicable)
α_2	$[M_i]^{-1}$	†	Coefficient for Y_i^2 in the quadratic relationship between ψ_i and Y_i (when $Y_i < \gamma$ if applicable)
β_0	$[M_i]$	*,†	Change in intercept of linear relationship when $Y_i > 0$
β_1	Unitless	†	Change in slope of linear relationship when $Y_i > 0$
γ	$[M_i]$	†	Allows relationship between Y_i and ψ_i to change when $Y_i > \gamma$
σ^2	Unitless	None	Variance parameter
η	$[M_i]$	†	Allows variance to decrease as Y_i increases - larger value means milder reduction in variance

predictions are consistent with the external validation data, this is an encouraging sign that either the CR data was collected in diverse conditions, or the bias and variance of the measurements do not depend on factors which differ between the training and external data. In either of these scenarios, the results are generalizable to different conditions. Investigating the residuals, defined as the difference between the model-predicted value ($\hat{M}_i$) and the observed data point (M_i), provides further insight into areas of improvement for the model.

The DIC, prediction bands, and residual plots were used for model selection as follows: First, the simplest model possible with constant variance was fit to the data (a multiplicative model with α_1 if there are no false positives in the data or a linear model with α_0 and α_1 otherwise). The DIC was calculated using JAGS. Prediction bands were compared to the data used to fit the model and residual plots were inspected. If the prediction bands were much wider or narrower than the spread of the data, this indicated that the variance model should be explored. If the residual plots show structured and/or large residuals, this indicated that the median parameterization may need to be modified. Models were then

augmented as suggested by the diagnostic plots, refit, and DIC was re-calculated. This process was repeated until the diagnostic plots were satisfactory and the DIC was at least three less than that of the previous model.

5.2.2 Algorithm for Quantifying Measurement Uncertainty in New Measurements

A key application of the models in Section 5.2.1 is calculating a credible interval for the true emission rate based on methane measurements made in the field. Let M^{new} be the new non-zero measurement made in the field and Y^{new} represent the unknown true emission rate associated with M^{new}. We wish to know the distribution of Y^{new} given M^{new} and our uncertainty model derived from CR data. Let $\underline{M} = (M_1, Y_1, ..., M_n, Y_n)'$ be a vector of the CR data for the technology under consideration. The distribution of interest is $p(Y^{new} \mid M^{new}, \underline{M})$. Using Bayes theorem, we can say that

$$\begin{aligned} p(Y^{new} \mid M^{new}, \underline{M}) &\propto p(M^{new} \mid Y^{new}, \underline{M}) p(Y^{new} \mid \underline{M}) \\ &= p(M^{new} \mid Y^{new}, \underline{M}) p(Y^{new}), \end{aligned} \tag{5.4}$$

where the new measurement and true emission rate are modelled as independent of the CR data. That is, we assume the distribution of true emission rates investigated in the CR trials does not provide any information about the true emission rate observed in the field. The distribution $p(Y^{new})$ can be thought of as a prior distribution representing our beliefs about the emission rate distribution we expect to see. The distribution $p(M^{new} \mid Y^{new}, \underline{M})$ is the likelihood of observing M^{new} given Y^{new} and $\underline{M}$. If we assume that the new measurements obey the same uncertainty model as the CR data, then this distribution is the likelihood selected in Section 5.2.1 for the given technology, marginalised over the posterior distribution of the statistical model parameters.

We propose the use of a weighted bootstrap technique (also called sampling/importance resampling) (Gelman et al., 2013; Smith and Gelfand, 1992) to obtain a sample from the posterior in Equation (5.4). This requires

1. Drawing a sample from the prior $p(Y^{new})$.
2. Computing normalized weights proportional to $p(M^{new} \mid Y^{new}, \underline{M})$ for each element of the sample from Step 1.
3. Resampling from the Step 1 sample using the normalized weights from Step 2.

This algorithm is implemented in R functions included in the repository https://github.com/augustinewigle/methaneUQ. A detailed description of each step of the algorithm is available in Appendix D.2.

5.3 Data

Several methane emission quantification technologies from different providers were evaluated through two CR measurement campaigns executed at Carbon Management Canada (CMC)'s Newall County Research Station near Brooks, Alberta, the first during April 20 - 26, 2022 and the second during September 25 - October 1, 2022. Providers conducted measurements and analysed data in the same way they would deploy in a typical survey scenario, for example, as part of an LDAR program. Providers compiled their own estimates and provided them to the academic team. In a small minority of the cases, the technology providers provided a null measurement due to sensor failures, or inappropriate meteorological conditions such as excessive cloud cover. These null measurements were excluded from the CR data for model fitting purposes.

The technologies which were evaluated in the CR campaigns were QOGI (3 different providers), truck-mounted Tunable Diode Laser-Absorption Spectroscopy (truck TDLAS); and aerial Near-InfraRed HyperSpectral Imaging (aerial NIR HSI). For a description of the physical basis for each technology, please refer to Wigle et al. (2024). A summary of the number of controlled releases completed for each technology provider and campaign is given in Table 5.2. Anonymized release and emissions estimates data from both campaigns are available at https://github.com/augustinewigle/methaneUQ.

To maximise the observed meteorological conditions in our CR data, we conducted two campaigns at different times of year. We consider a range of release rates that correspond to realistic scenarios from 0 to 80 kg/h. Finally, we performed releases from a variety of industrially relevant scenarios, including 1.7, 3.4, and 4.8 m tall stacks, an unlit flare, and a storage tank. Due to the location of the measurement campaigns and modality of several technologies, it was not possible to carry out fully blinded CRs where the technology providers were unaware of the existence, location, and release rate of CRs. Instead, providers were aware of the existence and location of releases, but not the rate.

In addition to the data from the two measurement campaigns, CR data reported by Conrad et al. (2023b) on the Bridger GML technology described in Chapter 4 was used to demonstrate the applicability of the methods. This technology is a type of TDLAS, and in this chapter, we refer to this technology as the aerial TDLAS technology. The range of

emission rates investigated in this CR data is from 0.4 to 66 kg/h. Some of the releases were conducted where the provider was aware that releases were being conducted, but not the location or amount being released, and some were conducted alongside a contracted survey so the provider was completely unaware that controlled releases were being performed, representing a superior degree of blinding. A total of 404 non-zero measurements were available. For more details, see Conrad et al. (2023b).

Table 5.2: Summary of technologies, providers, and number of non-zero measurements from the CR campaigns conducted at the CMC facility in Alberta. N_1 and N_2 refer to the number of observations collected for a given technology during the first and second campaigns, respectively.

Technology	N_1	N_2
QOGI Provider A	117	0
QOGI Provider B	0	71
QOGI Provider C	14	106
Truck TDLAS	142	125
Aerial NIR HSI	46	37

5.4 Results

5.4.1 Uncertainty Model Results

The model described in Section 5.2.1 was fit to the available CR data for the six technology/provider combinations. The likelihoods selected for each type are summarized in Table 5.3. The results are conditional on the CR data used to fit the models, therefore we comment on the generalizability of the models to an arbitrary release scenario by comparing to external data, if available.

QOGI

Prediction bands and posterior median predictions are shown for QOGI Providers A, B, and C in Figures 5.1, 5.2 and 5.3, respectively. All QOGI technologies underestimate emissions on average. The likelihoods from QOGI Providers B and C are best modelled using a quadratic function for the median below a small threshold, and a linear function at larger release rates. QOGI Provider A is best modelled using a quadratic function for the

Table 5.3: Summary of selected models for each methane quantification technology provider.

Technology	Selected likelihood	
	ψ_i (Median of M_i)	$\mathbf{Var}(\epsilon_i)$
QOGI A	$\alpha_0 + \alpha_1 Y_i + \alpha_2 Y_i^2$	$(\sigma^{-2} + Y_i/\eta)^{-1}$
QOGI B	$\begin{cases} \alpha_0 + \alpha_1 Y_i + \alpha_2 Y_i^2 & Y_i \leq \gamma; \\ \alpha_0 + \beta_0 + (\alpha_1 + \beta_1) Y_i & Y_i > \gamma. \end{cases}$	σ^2
QOGI C	$\begin{cases} \alpha_0 + \alpha_1 Y_i + \alpha_2 Y_i^2 & Y_i \leq \gamma; \\ \alpha_0 + \beta_0 + (\alpha_1 + \beta_1) Y_i & Y_i > \gamma. \end{cases}$	$(\sigma^{-2} + Y_i/\eta)^{-1}$
Truck TDLAS	$\alpha_0 + \alpha_1 Y_i$	$(\sigma^{-2} + Y_i/\eta)^{-1}$
Aerial TDLAS	$\alpha_1 Y_i$	$(\sigma^{-2} + Y_i/\eta)^{-1}$
Aerial NIR HSI	$\alpha_0 + \alpha_1 Y_i$	σ^2

median. The releases measured by Provider A have a more limited range than the other QOGI technologies with a max Y_i value of 30 kg/h, compared to 80 kg/h for Provider C and 50 kg/h for Provider B. For QOGI technologies in general, the likelihood has more curvature in the lower range of Y_i while a linear relationship on the log scale is suitable for higher release rates.

QOGI Provider C was present for both campaigns. Only 14 measurements were made for this technology at the first campaign, which we use as external data. These data points fall within the 95% prediction band, suggesting that the model is generalizable. The other technology providers were only present for one campaign and thus no external data are available for these providers.

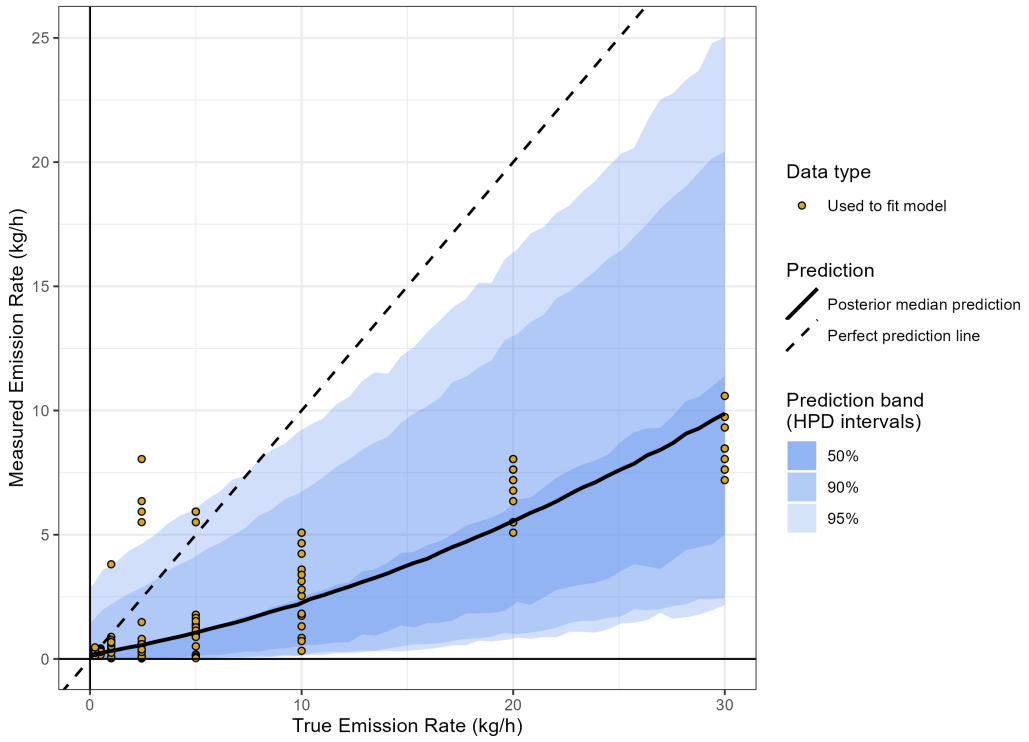


Figure 5.1: Uncertainty quantification model results for QOGI Provider A. The model was fit to data from the first campaign. No external data were available.

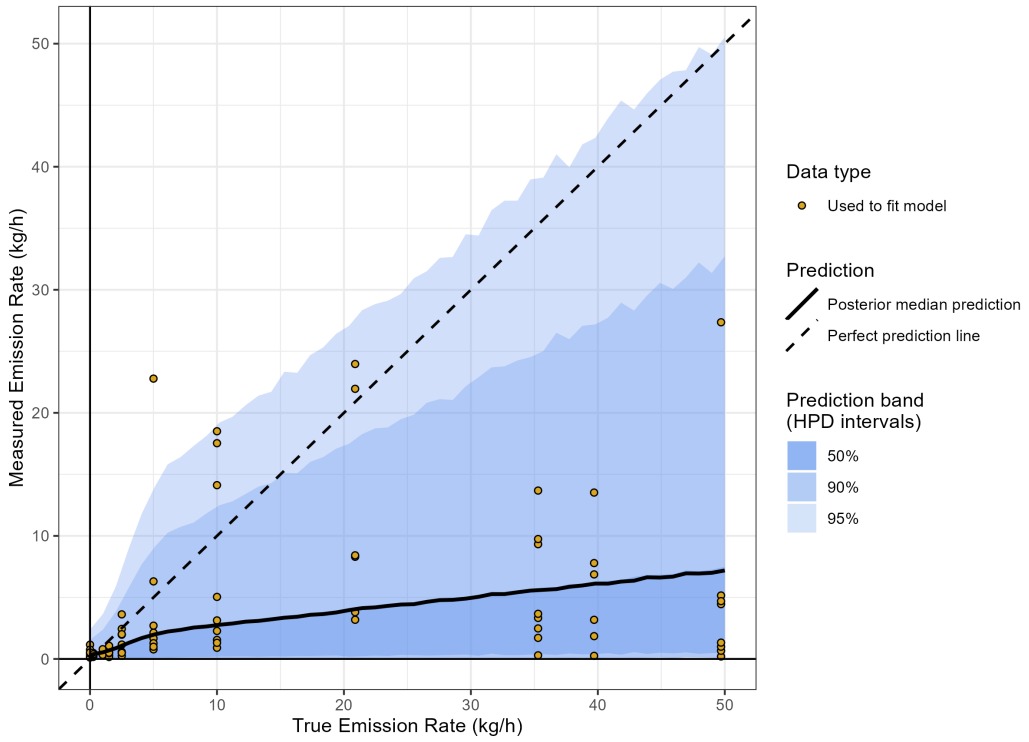


Figure 5.2: Uncertainty quantification model results for QOGI Provider B. The model was fit to data from the second campaign. No external data were available.

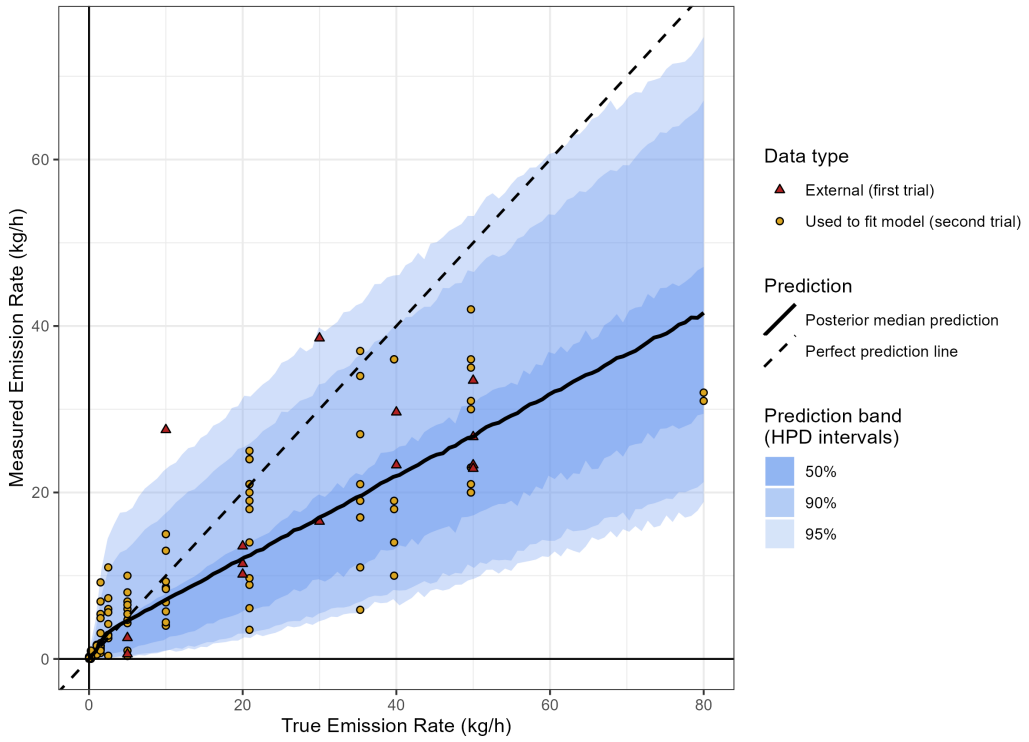


Figure 5.3: Uncertainty quantification model results from QOGI Provider C. The model was fit to data from the second campaign. CR data from the first campaign was used as external data.

TDLAS

Results from the selected models for truck and aerial TDLAS systems are shown in Figure 5.4 and the left panel of Figure 5.5 respectively. The truck-based TDLAS tends to underestimate emissions, while the aerial-based TDLAS overestimates on average.

For truck-based TDLAS, the model was fit using data from the second campaign, while data from the first campaign were used as external data to assess the model's generalizability. Most of the external data points fall within the prediction bands. However, the median trend appears different for the external data. A possible explanation for this is that weather conditions may have differed between the two campaigns. The generalizability would be improved if more CR data were collected and included in the model fitting.

Figure 5.5 compares the predictions resulting from the Bayesian uncertainty model derived for the aerial TDLAS and the one presented by Conrad et al. (2023b). The Bayesian model gives narrower prediction bands than the other model, which is particularly noticeable in the upper range of Y_i. This is likely due to the different variance specifications used in the models; the model by Conrad et al. (2023b) uses a constant variance whereas the present model allows the variance to change with Y_i. The median predictions are very similar between the two models.

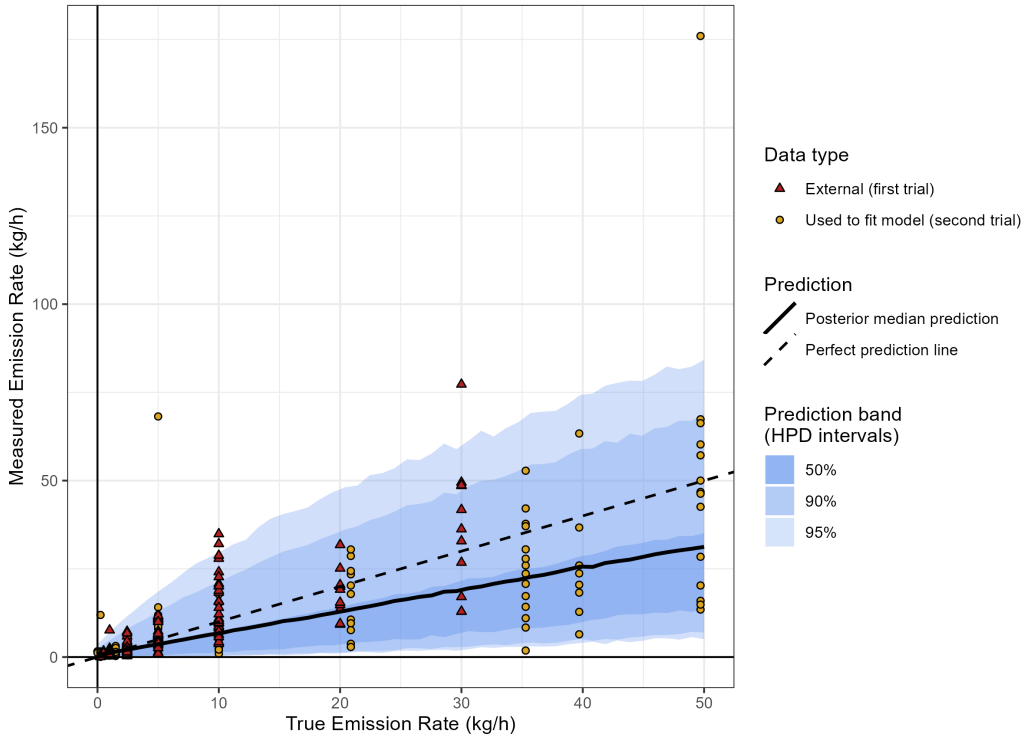


Figure 5.4: Uncertainty quantification model results for truck-based TDLAS data. The model was fit to data from the second campaign. CR data from the first campaign was used as external data.

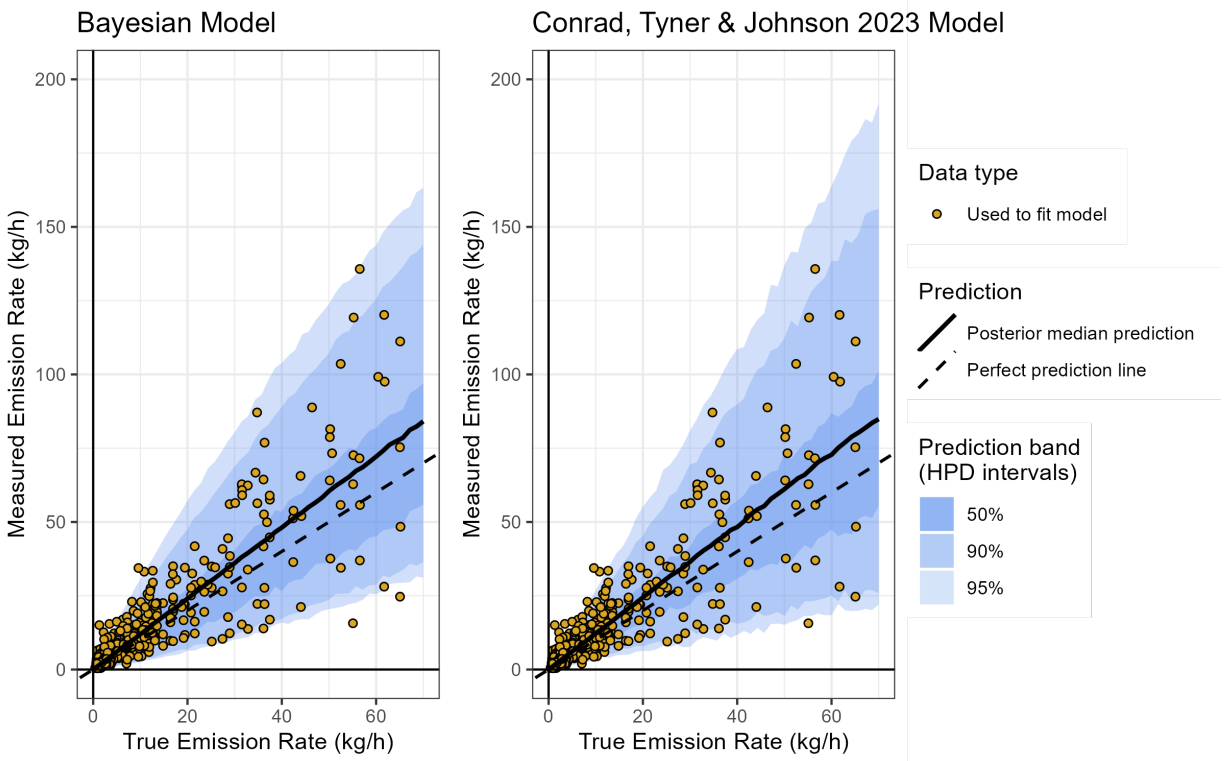


Figure 5.5: Side-by-side comparison of Bayesian model proposed in Table 5.3 and that presented in Conrad et al. (2023b) for aerial TDLAS.

Aerial NIR HSI

The prediction bands from the selected model for the aerial NIR HSI technology are shown in Figure 5.6. The technology tends to overestimate emissions. The model was fit to data from the second campaign, while data from the first campaign were used as external data. The technology tended to underestimate the true releases during the first campaign and overestimate the true releases during the second campaign. Conditions during the first campaign were considered marginal due to excessive cloud cover and atypical of those under which commercial measurements were conducted, while those of the second campaign were nearly ideal. Given the different conditions between the campaigns, it is not surprising that the model derived from data taken from the second campaign consistently overestimates the observed data from the first campaign. Therefore, the model is not generalizable.

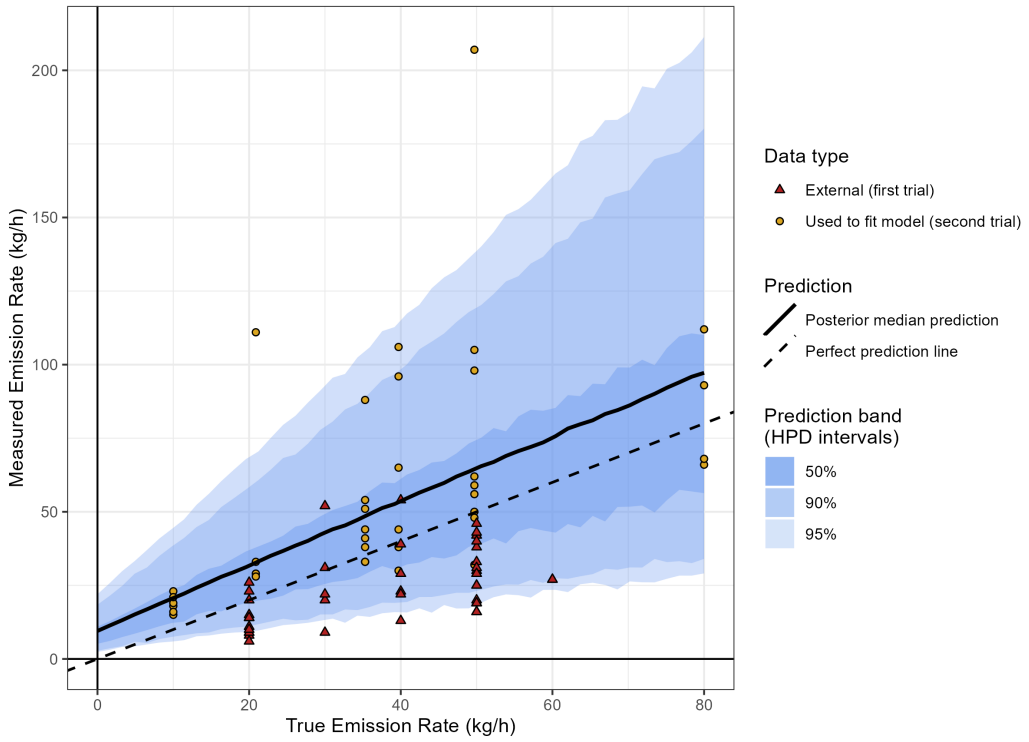


Figure 5.6: Uncertainty quantification model results for aerial NIR HSI data. The model was fit to data from the second campaign. CR data from the first campaign were used as external data.

5.4.2 Example of Quantifying Uncertainty in New Measurements

We now demonstrate how the models developed in the previous section may be used to estimate uncertainty in a new measurement using the QOGI Provider C model as an example. The weighted bootstrap algorithm described in Section 5.2.2 and Appendix D.2 was followed with $L = 5000$, $J = 10,000$, and $K = 4000$.

When conducting QOGI measurements, a key operational consideration is the number of independent measurements that should be made, given the well-known variability of this technology. Accordingly, we consider two measurement scenarios: one where a single measurement of a source is made, and one where five independent measurements of the same source are made. We also investigate two different prior distributions, $p(Y^{new})$, to show how information flows from the prior to the posterior distribution.

We simulate the process of performing measurements in the field as follows: First, we choose a hypothetical true value for the source we will measure, $Y_i = 25$ kg/h. Then, we simulate measurements by drawing from the posterior predictive distribution, $p(\tilde{M}_i \mid M_i, Y_i = 25 \text{ kg/hr})$. To demonstrate how prior information impacts the estimates, we consider two different prior distributions, reflecting different degrees of prior information that may be available at the time of the measurement. We investigate four different scenarios:

1. Only one measurement, $M_1^{new} = 19.7$ kg/h, is drawn, and $p(Y^{new})$ is a uniform distribution from 0 to 200 kg/h.

2. Five measurements are drawn, $M_1^{new} = 19.7$ kg/h, $M_2^{new} = 11.6$ kg/h, $M_3^{new} = 8.4$ kg/h, $M_4^{new} = 18.1$ kg/h, and $M_5^{new} = 17.0$ kg/h, and $p(Y^{new})$ is the same uniform distribution as Scenario 1.

3. Only one measurement, M_1^{new}, is used, and $p(Y^{new})$ is a log-normal distribution, which is often used to model leak rate distributions (Zavala-Araiza et al., 2018). We set the shape parameter to 1, location parameter equal to zero, and scale parameter equal to 2.6 (NIST, 2023).

4. All of $M_1^{new}, \ldots, M_5^{new}$ are used, and $p(Y^{new})$ is the same log-normal distribution as Scenario 3.

The different information expressed in the two priors is shown in the grey histograms in Figure 5.7. The uniform prior expresses that all values between 0 and 200 kg/h are equally likely. This may be a naive choice, because surveys have shown that extremely high emitters are much less likely than lower emitters (Brandt et al., 2016; Zavala-Araiza et al., 2018). The log-normal prior expresses that there is about a 50% chance that the emission rate is less than 13.6 kg/h and 75% chance that the emission rate is less than 26.2 kg/h, with very large values being rare, but possible.

The posterior distributions for different numbers of measurements are shown along with the corresponding priors in Figure 5.7. The posterior distributions are not centred around the measured values, which were all less than the true value of 25 kg/h. This is a reflection of the results shown in Figure 5.3, where it is clear that the technology systematically underestimates the true emission rate. It is also clear to see the influence of the prior distribution - the narrower prior distribution provides more information and narrows the posterior distribution. It is therefore crucial that the prior distribution specified for Y^{new} is a true reflection of the state of knowledge about the plausible emission rates. The 90% CIs

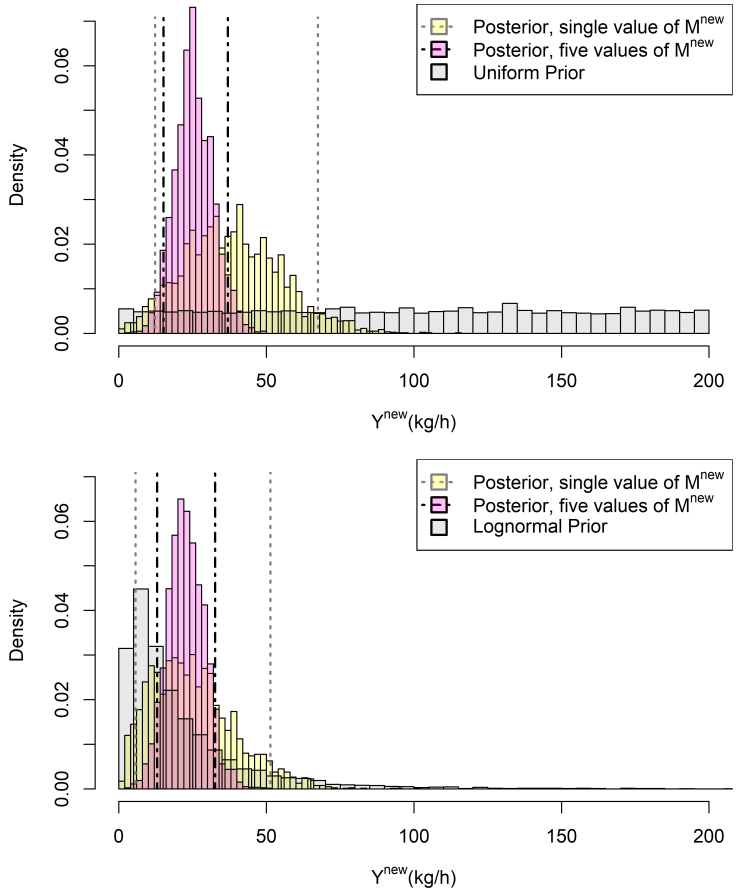


Figure 5.7: Prior and posterior distributions for the true value of the emission rate given the observed measurement(s), prior distribution, and CR data. Posterior 90% CIs are shown with dashed lines.

for Y^{new} given the measurement(s) and their lengths for the different settings are shown in Table 5.4. The number of measurements made seems to have a bigger impact on the length of the CI than the prior distribution used.

An important consideration when applying this algorithm is that the error model will only reflect the true measurement uncertainty if the new measurement was obtained under conditions similar to those under which the CR data were obtained. In particular, we caution against fitting error models to CR data collected only in ideal operating conditions, because the resulting CIs will be too narrow.

Table 5.4: Summary of 90% credible intervals (CI) resulting from weighted bootstrap algorithm for different priors and numbers of new measurements made.

Prior	Number of Measurements	90% CI	Length of CI
Unif(0,200)	1	(13.1, 71.7)	58.6
Unif(0,200)	5	(15.0, 37.2)	22.2
LogNormal(2.6, 1)	1	(2.1, 47.3)	45.2
LogNormal(2.6, 1)	5	(13.0, 34.1)	21.1

5.5 Discussion and Future Work

As highlighted by Figures 5.1 - 5.6, relying solely on measurements without considering their uncertainty can lead to significant misinterpretations of the underlying emission rate. In this chapter, we present a Bayesian approach to summarising the variability and bias of a technology using CR data, and provide an algorithm for deriving the posterior distribution of a true methane emission rate, given the CR data and new measurement(s). These methods can be useful as inputs to probabilistic simulations to assess the effectiveness of LDAR programs or in calculating the uncertainty attached to estimated methane inventories.

While not the focus of this work, a key finding of this study is that increasing the number of measurements reduces the width of the posterior; that is, as we collect more data, we can be more certain about the true value of the emission rate from a given source. Table 5.4 shows that the uncertainty in estimating a true emission rate is reduced when we have five measurements compared to only one measurement. The implications of this result are immediately of significance for regulations regarding LDAR programs. Currently, regulations on OG producers tend to specify the annual frequencies of LDAR surveys on production sites and facilities. We have shown that the number of measurements made in a given survey could have a significant impact on the precision of the estimated methane emission rate. Therefore, in addition to specifying the annual number of surveys, regulators would be advised to specify the minimum number of observations at any emission source in each survey. Alternatively, and perhaps more appropriately due to the relationship between uncertainty and true emission rate, a desired uncertainty range per emission source should be specified, rather than the number of measurements. The models presented here could be used to determine how many measurements should be performed for a certain technology and emission rate to attain the desired uncertainty range.

Although extensive meteorological data were collected during the campaigns, we refrain from incorporating them in the statistical models. This is because the goal of the models

is to summarise the performance of a technology over a variety of conditions so as to be as generally applicable as possible. To this end, the campaigns were conducted at different times of year, involved different release structures, and each trial was carried out over multiple days, so that the results could be used to assess the performance of the technologies over a variety of conditions. In addition, we did not want our models to rely on inputs that may not be available during regular operating conditions. Nonetheless, it may be possible to improve the predictive performance of the models by incorporating meteorological data into the likelihood.

The uncertainty contributed by a specific technology includes measurement uncertainty as well as detection uncertainty. The methods in this chapter address measurement uncertainty only. These components may be considered separately, and this has been implemented by Conrad et al. (2023b). They found that the detection probability of Bridger's GML is related to different factors than the measurement uncertainty, emphasizing that the two sources of uncertainty are distinct. An advantage of the Bayesian approach taken in this chapter is that it lends itself well to model extension. A possible avenue of future work is the modelling of detection probability and measurement uncertainty simultaneously using a hierarchical Bayesian model.

The approach to creating a methane inventory presented in Chapter 4 utilises a distribution for the true emission rate given a measurement from the technology, which coincides with the distribution derived by the weighted bootstrap algorithm, $p(Y^{new} \mid M^{new}, \underline{M})$. The error models derived in this chapter could therefore be integrated into the measurement error MC portion of the inventory approach. Improvements to the efficiency of the algorithm would be beneficial in this application, and are an important area of future work.

Chapter 6

Conclusion

In this thesis, we have examined several problems in NMA and the quantification of methane emissions. We proposed new methods in NMA to address knowledge gaps in CNMA and uncertainty in treatment hierarchies. We proposed novel applications of statistical models to improve our understanding of uncertainty in methane inventories and quantification technology measurements. We conclude by summarizing the contributions of each chapter, and avenues for future work.

In Chapter 2, we identified an important difference in the underlying assumptions of Bayesian and frequentist additive CNMA models. We showed that the original Bayesian approach (an anchored approach) is sensitive to misspecification and proposed a more robust Bayesian alternative (unanchored). We also discussed possible interpretations of "placebo" or "no treatment" as the anchor, if assumptions are met. In a simulation study, we demonstrated that when misspecified, the additive anchored Bayesian CNMA model may result in biased relative effect estimates, incorrect CI coverage probability, and an incorrect treatment hierarchy. On the other hand, our novel unanchored Bayesian methods, along with the unanchored frequentist approach, are robust. We applied the anchored and unanchored models to a published network. This research is an important step in connecting the fields of Bayesian and frequentist NMA. A possible avenue of future work is to relax the additivity assumption by including interaction terms in the unanchored Bayesian CNMA model and investigate the interpretations of the relative effect parameters in these models. Another approach to relaxing additivity in the Bayesian framework would be to incorporate structure in the prior distributions on the relative effects of multi-component treatments.

Chapter 3 focused on the challenge of quantifying uncertainty in a treatment hierarchy

produced from an NMA. We proposed the Precision Of Treatment Hierarchy, or POTH, a metric which quantifies the level of certainty in a treatment hierarchy from a Bayesian or frequentist NMA. The metric is less subjective and easier to interpret than existing approaches, even when the number of treatments in the network is large. We offer guidance on its interpretation, and provide an R function that can be used to obtain a relevant reference distribution for any given network. We also introduced leave-one-out POTH residuals for quantifying the relative contribution of each treatment to the level of certainty in the hierarchy, and best-k-treatments POTH as a way to understand how the certainty changes when we consider only the top k treatments. In an analysis of 267 published networks, we found an empirical distribution of POTH and investigated the association between POTH and the number of treatments in the network, the proportion of significantly different treatment effects, and the estimated heterogeneity. Finally, we provide worked examples to demonstrate the use of POTH with published datasets. Updating the database to include more recent networks and re-analysing it to obtain a more current distribution of POTH is of interest. An important area of future work is a more detailed simulation study which explores more factors and how they affect POTH, including increasing the number of studies per comparison, network geometry, and different relative treatment effects.

In Chapter 4, we presented a novel application of a multi-stage sampling framework to estimate methane emissions using aerial survey data, inspired by a real dataset collected in BC, Canada. First, we showed how the data collection process of a OG survey corresponds to a multi-stage survey, accounting for detection error, daily variation, and sampling error. We presented estimators for the stratum and population totals and their variances, an interpretable breakdown of the variance by source, and showed how the Hájek estimator can be used in this setting, even when there are empty samples. We also provided a MC approach to combine measurement uncertainty with the multi-stage framework. We conducted a simulation study based on the BC data to validate the properties of the multi-stage estimators in the absence of measurement error. Finally, we applied the methods to the BC dataset and compared our findings with the existing approach to analysing the data. Our approach has similar findings, while providing additional insights into the sources of variability and being less computationally intensive. This work can be extended by incorporating more complicated sampling designs. It can readily be used to compare the efficiency of sampling designs for future surveys. An approach to finding an optimal sampling design is also of interest. Finally, as uncertainty models for more MDQTs become available, the methods can be applied to a wider variety of datasets. This can potentially have a broad impact as measurement-based surveys using various MDQTs become more common (Johnson et al., 2023; Conrad et al., 2023c,a; Varon et al., 2019).

In Chapter 5, we proposed a data-driven approach to quantify uncertainty in measure-

ments from different MDQTs. We provided a flexible Bayesian model for the measured emission rate given a true emission rate, which is fit to CR data. We showed how a weighted bootstrap algorithm can be used to incorporate new measurements and prior information to obtain the posterior distribution of a true emission rate. We fit the flexible Bayesian model to CR data for six different technologies, and showed that the methods capture complex non-linear behaviour. We demonstrated the use of the weighted bootstrap algorithm for different numbers of new measurements and prior distributions, and showed how uncertainty could be reduced by making repeated measurements. An area of future work for this project would be to use a hierarchical Bayesian model to account for both detection uncertainty and measurement uncertainty. Another area of future work would be to incorporate these models into the measurement error MC of the multi-stage approach to estimating methane inventories proposed in Chapter 4. Improvements to the efficiency of the algorithm will be important in this application.

References

Atherton, E., Risk, D., Fougère, C., Lavoie, M., Marshall, A., Werring, J., Williams, J. P., and Minions, C. (2017). Mobile measurement of methane emissions from natural gas developments in northeastern British Columbia, Canada. *Atmospheric Chemistry and Physics*, 17(20):12405–12420.

Balduzzi, S., Rücker, G., Nikolakopoulou, A., Papakonstantinou, T., Salanti, G., Efthimiou, O., and Schwarzer, G. (2023). Netmeta: An R Package for Network Meta-Analysis Using Frequentist Methods. *Journal of Statistical Software*, 106:1–40.

Béliveau, A., Boyne, D. J., Slater, J., Brenner, D., and Arora, P. (2019). BUGSnet: An R package to facilitate the conduct and reporting of Bayesian network Meta-analyses. *BMC Medical Research Methodology*, 19(1):196.

Bell, C., Rutherford, J., Brandt, A., Sherwin, E., Vaughn, T., and Zimmerle, D. (2022). Single-blind determination of methane detection limits and quantification accuracy using aircraft-based LiDAR. *Elementa: Science of the Anthropocene*, 10(1):00080.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2011). *Introduction to Meta-Analysis*. John Wiley & Sons.

Boussageon, R., Howick, J., Baron, R., Naudet, F., Falissard, B., Harika-Germaneau, G., Wassouf, I., Gueyffier, F., Jaafari, N., and Blanchard, C. (2022). How do they add up? The interaction between the placebo and treatment effect: A systematic review. *British Journal of Clinical Pharmacology*, 88(8):3638–3656.

Brandt, A. R., Heath, G. A., and Cooley, D. (2016). Methane Leaks from Natural Gas Systems Follow Extreme Distributions. *Environmental Science & Technology*, 50(22):12512–12520.

Brantley, H. L., Thoma, E. D., Squier, W. C., Guven, B. B., and Lyon, D. (2014). Assessment of Methane Emissions from Oil and Gas Production Pads using Mobile Measurements. *Environmental Science & Technology*, 48(24):14508–14515.

Bulluck, H., Kwok, C. S., Ryding, A. D., and Loke, Y. K. (2015). Safety of short-term dual antiplatelet therapy after drug-eluting stents: An updated meta-analysis with direct and adjusted indirect comparison of randomized control trials. *International Journal of Cardiology*, 181:331–339.

Caulton, D. R., Li, Q., Bou-Zeid, E., Fitts, J. P., Golston, L. M., Pan, D., Lu, J., Lane, H. M., Buchholz, B., Guo, X., McSpiritt, J., Wendt, L., and Zondlo, M. A. (2018). Quantifying uncertainties from mobile-laboratory-derived emissions of well pads using inverse Gaussian methods. *Atmospheric Chemistry and Physics*, 18(20):15145–15168.

Chiocchia, V., Nikolakopoulou, A., Papakonstantinou, T., Egger, M., and Salanti, G. (2020). Agreement between ranking metrics in network meta-analysis: An empirical study. *BMJ Open*, 10(8):e037744.

Climate and Clean Air Coalition (2021). Global Methane Pledge. `https://www.globalmethanepledge.org/`.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons.

Conrad, B. M., Tyner, D. R., and Johnson, M. R. (2023a). The Futility of Relative Methane Reduction Targets in the Absence of Measurement-Based Inventories. *Environmental Science & Technology*, 57(50):21092–21103.

Conrad, B. M., Tyner, D. R., and Johnson, M. R. (2023b). Robust probabilities of detection and quantification uncertainty for aerial methane detection: Examples for three airborne technologies. *Remote Sensing of Environment*, 288:113499.

Conrad, B. M., Tyner, D. R., Li, H. Z., Xie, D., and Johnson, M. R. (2023c). A measurement-based upstream oil and gas methane inventory for Alberta, Canada reveals higher emissions and different sources than official estimates. *Communications Earth & Environment*, 4(1):1–10.

Curteis, T., Wigle, A., Michaels, C. J., and Nikolakopoulou, A. (2025). Ranking of treatments in network meta-analysis: Incorporating minimally important differences. *BMC Medical Research Methodology*, 25(1):67.

Davies, A. L. and Galla, T. (2021). Degree irregularity and rank probability bias in network meta-analysis. *Research Synthesis Methods*, 12(3):316–332.

Delignette-Muller, M. L. and Dutang, C. (2015). Fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, 64:1–34.

Denwood, M. J. (2016). Runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software*, 71:1–25.

Dias, S., Ades, A. E., Welton, N. J., Jansen, J. P., and Sutton, A. J. (2018). *Network Meta-Analysis for Decision-Making*. John Wiley & Sons.

Edie, R., Robertson, A. M., Field, R. A., Soltis, J., Snare, D. A., Zimmerle, D., Bell, C. S., Vaughn, T. L., and Murphy, S. M. (2020). Constraining the accuracy of flux estimates using OTM 33A. *Atmospheric Measurement Techniques*, 13(1):341–353.

Efthimiou, O. (2018). Practical guide to the meta-analysis of rare events. *Evidence Based Mental Health*, 21(2).

Efthimiou, O., Rücker, G., Schwarzer, G., Higgins, J. P., Egger, M., and Salanti, G. (2019). Network meta-analysis of rare events using the Mantel-Haenszel method. *Statistics in Medicine*, 38(16):2992–3012.

Efthimiou, O., Seo, M., Karyotaki, E., Cuijpers, P., Furukawa, T. A., Schwarzer, G., Rücker, G., and Mavridis, D. (2022). Bayesian models for aggregate and individual patient data component network meta-analysis. *Statistics in Medicine*, 41(14):2586–2601.

Elliott, W. J. and Meyer, P. M. (2007). Incident diabetes in clinical trials of antihypertensive drugs: A network meta-analysis. *The Lancet*, 369(9557):201–207.

Environment and Climate Change Canada (2016). Pan-Canadian Framework on Clean Growth and Climate Change: Canada's plan to address climate change and grow the economy. `https://publications.gc.ca/site/eng/9.828774/publication.html`.

Fox, T. A., Gao, M., Barchyn, T. E., Jamin, Y. L., and Hugenholtz, C. H. (2021a). An agent-based model for estimating emissions reduction equivalence among leak detection and repair programs. *Journal of Cleaner Production*, 282:125237.

Fox, T. A., Hugenholtz, C. H., Barchyn, T. E., Gough, T. R., Gao, M., and Staples, M. (2021b). Can new mobile technologies enable fugitive methane reductions from the oil and gas industry? *Environmental Research Letters*, 16(6):064077.

Freeman, S. C., Scott, N. W., Powell, R., Johnston, M., Sutton, A. J., and Cooper, N. J. (2018). Component network meta-analysis identifies the most effective components of psychological preparation for adults undergoing surgery under general anesthesia. *Journal of Clinical Epidemiology*, 98:105–116.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., deBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., and Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64(4):383–394.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926.

Hájek, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. *Foundations of statistical inference*, page 326.

Heltzel, R. S., Johnson, D. R., Zaki, M. T., Gebreslase, A. K., and Abdul-Aziz, O. I. (2022). Machine learning techniques to increase the performance of indirect methane quantification from a single, stationary sensor. *Heliyon*, 8(12):e11962.

Higgins, J. P. T., López-López, J. A., Becker, B. J., Davies, S. R., Dawson, S., Grimshaw, J. M., McGuinness, L. A., Moore, T. H. M., Rehfuess, E. A., Thomas, J., and Caldwell, D. M. (2019). Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Global Health*, 4(Suppl 1).

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685.

IEA (2023). Global Methane Tracker 2023 – Analysis. `https://www.iea.org/reports/global-methane-tracker-2023`.

Intergovernmental Panel On Climate Change (2023). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 1 edition.

Johnson, M. R., Conrad, B. M., and Tyner, D. R. (2023). Creating measurement-based oil and gas sector methane inventories using source-resolved aerial surveys. *Communications Earth & Environment*, 4(1):1–9.

Johnson, M. R., Tyner, D. R., and Szekeres, A. J. (2021). Blinded evaluation of airborne methane source detection using Bridger Photonics LiDAR. *Remote Sensing of Environment*, 259:112418.

Kang, R., Liatsis, P., and Kyritsis, D. C. (2022). Emission Quantification via Passive Infrared Optical Gas Imaging: A Review. *Energies*, 15(9):3304.

Kemp, C. E. and Ravikumar, A. P. (2021). New Technologies Can Cost Effectively Reduce Oil and Gas Methane Emissions, but Policies Will Require Careful Design to Establish Mitigation Equivalence. *Environmental Science & Technology*, 55(13):9140–9149.

Kish, L. (1965). *Survey Sampling*. Wiley.

Kriston, L., von Wolff, A., Westphal, A., Hölzel, L. P., and Härter, M. (2014). Efficacy and Acceptability of Acute Treatments for Persistent Depressive Disorder: A Network Meta-Analysis. *Depression and Anxiety*, 31(8):621–630.

Lavoie, M., Risk, D., O'Connell, E., Atherton, E., Gorski, J., and Johnson, J. (2021). Evaluating the benefits of alternative leak detection programs. *EarthArXiv pre-print*. https://doi.org/10.31223/X5F628.

Lu, G. and Ades, A. (2009). Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805.

Lu, G., Welton, N. J., Higgins, J. P. T., White, I. R., and Ades, A. E. (2011). Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. *Research Synthesis Methods*, 2(1):43–60.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16):2313–2324.

MacKay, K., Lavoie, M., Bourlon, E., Atherton, E., O'Connell, E., Baillie, J., Fougère, C., and Risk, D. (2021). Methane emissions from upstream oil and gas production in Canada are underestimated. *Scientific Reports*, 11(1):8041.

Madan, J., Chen, Y.-F., Aveyard, P., Wang, D., Yahaya, I., Munafo, M., Bauld, L., and Welton, N. (2014). Synthesis of Evidence On Heterogeneous Interventions with Multiple Outcomes Recorded Over Multiple Follow-Up Times Reported Inconsistently: A Smoking Cessation Case-Study. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 177(1):295–314.

Mills, E. J., Druyts, E., Ghement, I., and Puhan, M. A. (2011). Pharmacotherapies for chronic obstructive pulmonary disease: A multiple treatment comparison meta-analysis. *Clinical Epidemiology*, 3:107–129.

Mills, E. J., Thorlund, K., and Ioannidis, J. P. A. (2012). Calculating additive treatment effects from multiple randomized trials provides useful estimates of combination therapies. *Journal of Clinical Epidemiology*, 65(12):1282–1288.

Mitchell, A. L., Tkacik, D. S., Roscioli, J. R., Herndon, S. C., Yacovitch, T. I., Martinez, D. M., Vaughn, T. L., Williams, L. L., Sullivan, M. R., Floerchinger, C., Omara, M., Subramanian, R., Zimmerle, D., Marchese, A. J., and Robinson, A. L. (2015). Measurements of Methane Emissions from Natural Gas Gathering Facilities and Processing Plants: Measurement Results. *Environmental Science & Technology*, 49(5):3219–3227.

Nikolakopoulou, A., Chaimani, A., Furukawa, T. A., Papakonstantinou, T., Rücker, G., and Schwarzer, G. (2024). When does the placebo effect have an impact on network meta-analysis results? *BMJ Evidence-Based Medicine*, 29(2):127–134.

Nikolakopoulou, A., Higgins, J. P. T., Papakonstantinou, T., Chaimani, A., Giovane, C. D., Egger, M., and Salanti, G. (2020). CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine*, 17(4):e1003082.

NIST (2023). 1.3.6.6.9. Lognormal Distribution. `https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm`.

Papakonstantinou, T. (2019). Nmadb: Network Meta-Analysis Database API.

Papakonstantinou, T., Salanti, G., Mavridis, D., Rücker, G., Schwarzer, G., and Nikolakopoulou, A. (2022). Answering complex hierarchy questions in network meta-analysis. *BMC Medical Research Methodology*, 22(1):47.

Petropoulou, M., Efthimiou, O., Rücker, G., Schwarzer, G., Furukawa, T. A., Pompoli, A., Koek, H. L., Del Giovane, C., Rodondi, N., and Mavridis, D. (2021). A review of methods for addressing components of interventions in meta-analysis. *PLoS One*, 16(2):e0246631.

Petropoulou, M., Nikolakopoulou, A., Veroniki, A.-A., Rios, P., Vafaei, A., Zarin, W., Giannatsi, M., Sullivan, S., Tricco, A. C., Chaimani, A., Egger, M., and Salanti, G. (2017). Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of Clinical Epidemiology*, 82:20–28.

Pirani, A., Alegria, A., Khourdajie, A. A., Gunawan, W., Gutiérrez, J. M., Holsman, K., Huard, D., Juckes, M., Kawamiya, M., Klutse, N., Krey, V., Matthews, R., Milward, A., Pascoe, C., van der Shrier, G., Spinuso, A., Stockhause, M., and Xing, X. (2022). The implementation of FAIR data principles in the IPCC AR6 assessment process. *Zenodo*.

Playford, E. G., Webster, A. C., Craig, J. C., and Sorrell, T. C. (2004). Antifungal agents for preventing fungal infections in solid organ transplant recipients. *Cochrane Database of Systematic Reviews*, (3).

Pompoli, A., Furukawa, T. A., Efthimiou, O., Imai, H., Tajika, A., and Salanti, G. (2018). Dismantling cognitive-behaviour therapy for panic disorder: A systematic review and component network meta-analysis. *Psychological Medicine*, 48(12):1945–1953.

Psaty, B. M., Lumley, T., Furberg, C. D., Schellenbaum, G., Pahor, M., Alderman, M. H., and Weiss, N. S. (2003). Health Outcomes Associated With Various Antihypertensive Therapies Used as First-Line Agents: A Network Meta-analysis. *JAMA*, 289(19):2534–2544.

Ribeiro, R. A., Ziegelmann, P. K., Duncan, B. B., Stella, S. F., da Costa Vieira, J. L., Restelatto, L. M. F., Moriguchi, E. H., and Polanczyk, C. A. (2013). Impact of statin dose on major cardiovascular events: A mixed treatment comparison meta-analysis involving more than 175,000 patients. *International Journal of Cardiology*, 166(2):431–439.

Ripple, W. J., Wolf, C., Gregg, J. W., Rockström, J., Mann, M. E., Oreskes, N., Lenton, T. M., Rahmstorf, S., Newsome, T. M., Xu, C., Svenning, J.-C., Pereira, C. C., Law, B. E., and Crowther, T. W. (2024). The 2024 state of the climate report: Perilous times on planet Earth. *BioScience*, 74(12):812–824.

Rosenberger, K. J., Duan, R., Chen, Y., and Lin, L. (2021). Predictive P-score for treatment ranking in Bayesian network meta-analysis. *BMC Medical Research Methodology*, 21(1):213.

Rücker, G., Petropoulou, M., and Schwarzer, G. (2019). Network meta-analysis of multicomponent interventions. *Biometrical Journal*, 62(3):808–821.

Rücker, G., Schmitz, S., and Schwarzer, G. (2021). Component network meta-analysis compared to a matching method in a disconnected network: A case study. *Biometrical Journal*, 63(2):447–461.

Rücker, G. and Schwarzer, G. (2014). Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. *Statistics in Medicine*, 33(25):4353–4369.

Rücker, G. and Schwarzer, G. (2015). Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Medical Research Methodology*, 15(1):58.

Salanti, G., Giovane, C. D., Chaimani, A., Caldwell, D. M., and Higgins, J. P. T. (2014). Evaluating the Quality of Evidence from a Network Meta-Analysis. *PLOS ONE*, 9(7):e99682.

Salanti, G., Nikolakopoulou, A., Efthimiou, O., Mavridis, D., Egger, M., and White, I. R. (2022). Introducing the Treatment Hierarchy Question in Network Meta-Analysis. *American Journal of Epidemiology*, 191(5):930–938.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Sherwin, E. D., Chen, Y., Ravikumar, A. P., and Brandt, A. R. (2021). Single-blind test of airplane-based hyperspectral methane detection via controlled releases. *Elementa: Science of the Anthropocene*, 9(1):00063.

Singh, D., Barlow, B., Hugenholtz, C., Funk, W., Robinson, C., and Ravikumar, A. P. (2021). Field Performance of New Methane Detection Technologies: Results from the Alberta Methane Field Challenge. *EarthArXiv pre-print*. `https://doi.org/10.31223/X5GS46`.

Sitter, R. R. (1992). A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association*, 87(419):755–765.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian Statistics without Tears: A Sampling-Resampling Perspective. *The American Statistician*, 46(2):84–88.

Smith, T. M. F. (1991). Post-Stratification. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(3):315–323.

Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., and Janssens-Maenhout, G. (2021). Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases. *Atmospheric Chemistry and Physics*, 21(7):5655–5683.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J R Stat Soc Series B Stat Methodology*, 64(4):583–639.

Thorlund, K. and Mills, E. (2012). Stability of additive treatment effects in multiple treatment comparison meta-analysis: A simulation study. *Clinical Epidemiology*, 4:75–85.

Tyner, D. R. and Johnson, M. R. (2021). Where the Methane Is—Insights from Novel Airborne LiDAR Measurements Combined with Ground Survey Data. *Environmental Science & Technology*, 55(14):9773–9783.

Varon, D. J., McKeever, J., Jervis, D., Maasakkers, J. D., Pandey, S., Houweling, S., Aben, I., Scarpelli, T., and Jacob, D. J. (2019). Satellite Discovery of Anomalously Large Methane Point Sources From Oil/Gas Production. *Geophysical Research Letters*, 46(22):13507–13516.

Veroniki, A. A., Straus, S. E., Rücker, G., and Tricco, A. C. (2018). Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *Journal of Clinical Epidemiology*, 100:122–129.

Vollrath, C., Hugenholtz, C. H., and Barchyn, T. E. (2024). Onshore methane emissions measurements from the oil and gas industry: A scoping review. *Environmental Research Communications*, 6(3):032001.

Walter, S. D. and Balakrishnan, N. (2022). A method was developed for correcting the bias in the usual study weights in meta-analyses. *Journal of Clinical Epidemiology*, 152:23–29.

Welton, N. J., Caldwell, D. M., Adamopoulos, E., and Vedhara, K. (2009). Mixed Treatment Comparison Meta-Analysis of Complex Interventions: Psychological Interventions in Coronary Heart Disease. *American Journal of Epidemiology*, 169(9):1158–1165.

Westfall, J. A., Patterson, P. L., and Coulston, J. W. (2011). Post-stratified estimation: Within-strata and total sample size recommendations. *Canadian Journal of Forest Research*, 41(5):1130–1139.

White, I. R. (2011). Multivariate Random-effects Meta-regression: Updates to Mvmeta. *The Stata Journal*, 11(2):255–270.

Wigle, A. and Béliveau, A. (2022). Bayesian unanchored additive models for component network meta-analysis. *Statistics in Medicine*, 41(22):4444–4466.

Wigle, A. and Béliveau, A. (2024). Estimating Methane Emissions from the Upstream Oil and Gas Industry Using a Multi-Stage Framework. *ArXiv pre-print*. `https://doi.org/10.48550/arXiv.2407.08827`.

Wigle, A., Béliveau, A., Blackmore, D., Lapeyre, P., Osadetz, K., Lemieux, C., and Daun, K. J. (2024). Estimation and Applications of Uncertainty in Methane Emissions Quantification Technologies: A Bayesian Approach. *ACS ES&T Air*.

Wigle, A., Béliveau, A., Salanti, G., Rücker, G., Schwarzer, G., Mavridis, D., and Nikolakopoulou, A. (2025). Precision of Treatment Hierarchy: A Metric for Quantifying Certainty in Treatment Hierarchies from Network Meta-Analysis. *ArXiv pre-print*. `https://doi.org/10.48550/arXiv.2501.11596`.

Wigle, A. and Schwarzer, G. (2024). POTH: Precision of Treatment Hierarchy (POTH).

Wilson, A. (2016). New Optical Gas-Imaging Technology for Quantifying Fugitive-Emission Rates. *Journal of Petroleum Technology*, 68(08):78–79.

Wu, C. and Thompson, M. E. (2020). *Sampling Theory and Practice*. Springer.

Wu, Y.-C., Shih, M.-C., and Tu, Y.-K. (2021). Using Normalized Entropy to Measure Uncertainty of Rankings for Network Meta-analyses. *Medical Decision Making*, 41(6):706–713.

Wu, Y.-C. and Tu, Y.-K. (2022). High robustness does not always imply low uncertainty of treatment rankings: An empirical study of 60 network meta-analyses. *Research Methods in Medicine & Health Sciences*, 3(4):116–120.

Xu, C., Chen, Y.-P., Du, X.-J., Liu, J.-Q., Huang, C.-L., Chen, L., Zhou, G.-Q., Li, W.-F., Mao, Y.-P., Hsu, C., Liu, Q., Lin, A.-H., Tang, L.-L., Sun, Y., and Ma, J. (2018). Comparative safety of immune checkpoint inhibitors in cancer: Systematic review and network meta-analysis. *BMJ*, 363:k4226.

Zavala-Araiza, D., Herndon, S. C., Roscioli, J. R., Yacovitch, T. I., Johnson, M. R., Tyner, D. R., Omara, M., and Knighton, B. (2018). Methane emissions from oil and gas production sites in Alberta, Canada. *Elementa: Science of the Anthropocene*, 6:27.

APPENDICES

Appendix A

Chapter 2 Appendix

A.1 Implications of the Anchor

Consider a network where there are four different treatments, A, B, C, and D, and where B and C are sometimes combined into a multicomponent treatment, $B+C$. We will suppose our outcome of interest is the number of cured patients, and so a large positive response represents a more effective treatment. Assume that the magnitude of the treatment effects are ordered as follows,

$$D < C < B < A,$$

that is, A is the most effective and D is the least effective relative to the other treatments.

Now we will consider two different models; one in which treatment D is the anchor, and one in which treatment A is the anchor. In the first model, treatment D is the anchor, so we obtain the relative effect of $B + C$ working on the scale relative to D, that is,

$$d_{D,B+C} = d_{D,B} + d_{D,C}. \tag{A.1}$$

We can write the single component relative effects relative to D as follows:

$$d_{D,D} = 0 < d_{D,C} < d_{D,B} < d_{D,A}.$$

Notice that because the relative effects are all positive relative to D, this implies that $d_{D,B+C}$ is greater than $d_{D,B}$ or $d_{D,C}$ and so in this model, combining treatments B and C results in a more effective treatment.

On the other hand, let us consider a model where A is the anchor. Then we have

$$d_{A,B+C} = d_{A,B} + d_{A,C}, \tag{A.2}$$

and the relative effects are

$$d_{A,D} < d_{A,C} < d_{A,B} < d_{A,A} = 0.$$

Now all of the effects relative to the anchor are negative, and so Equation (A.2) implies that $d_{A,B+C}$ is less than either of $d_{A,B}$ or $d_{A,C}$, meaning in this model, we expect combining treatments B and C to be worse than either B or C alone.

These two cases show how models with different anchors make different assumptions about the effect of combining treatments. The model which is "best" will depend on the data. In the example in this section, the two models have opposite implications for the multicomponent treatment $B+C$. If the data shows that $B+C$ is a better treatment than B or C alone, than the first model will probably perform well, while the second model will likely fit the data poorly. Overall, it is important to understand what is being used as the anchor in the additivity equation, as some anchors may give a poor fit to the data.

A.2 Equivalence of Novel Bayesian Models

Theorem 2. *The arm-based model in section 2.4.2 is equivalent to the contrast-based model in section 2.4.1, with the exception that the contrast-based method requires a normality assumption when calculating the standard errors in the matrix* $\boldsymbol{S}_i^*$.

Proof. Applying $\boldsymbol{U}_i^*$ to a vector of observed log-odds ratios in each arm of trial i will give the data vector $\boldsymbol{y}_i^*$. More specifically, $U_{i,jk}^* = 1$ if the k^{th} arm is the minuend of the j^{th} contrast, -1 if the k^{th} arm is the subtrahend of the j^{th} contrast, and 0 otherwise. The rows of $\boldsymbol{U}_i^*$ sum to zero and the first column of $\boldsymbol{U}_i^*$ contains all -1s. We can show that multiplying $\boldsymbol{U}_i^*$ by (2.16) gives equation (2.15):

$$\begin{aligned}
\boldsymbol{U}_i^* \text{logit}(\boldsymbol{\pi}_i) &= \alpha_i \boldsymbol{U}_i^* \mathbf{1}_{1\times a_i} + \boldsymbol{U}_i^* \boldsymbol{V}_i \boldsymbol{d} + \boldsymbol{U}_i^* \boldsymbol{\epsilon}_i \\
&= \boldsymbol{U}_i^* \boldsymbol{V}_i \boldsymbol{d} + \boldsymbol{\epsilon}_i^* \\
&= \boldsymbol{\mu}_i^*,
\end{aligned}$$

noting that $\boldsymbol{U}_i^*\mathbf{1}_{1\times a_i} = \mathbf{0}_{(a_i-1)\times 1}$, and $\boldsymbol{U}_i^*\boldsymbol{\epsilon}_i$ has the same probability distribution as $\boldsymbol{\epsilon}_i^*$. The latter follows from $\boldsymbol{U}_i^*\boldsymbol{\epsilon}_i \sim \text{MVN}_{a_i-1}(\mathbf{0}, \boldsymbol{U}_i^*\boldsymbol{\Sigma}_{a_i-1\times a_i-1}\boldsymbol{U}_i^{*\prime})$, where

$$\begin{aligned}\boldsymbol{U}_i^*\boldsymbol{\Sigma}_{a_i-1\times a_i-1}\boldsymbol{U}_i^{*\prime} &= 0.5\sigma^2\boldsymbol{U}_i^*(\boldsymbol{J}+\boldsymbol{I})\boldsymbol{U}_i^{*\prime}\\ &= 0.5\sigma^2(\boldsymbol{O}+\boldsymbol{U}_i^*\boldsymbol{I}\boldsymbol{U}_i^{*\prime})\\ &= 0.5\sigma^2(\boldsymbol{U}_i^*\boldsymbol{U}_i^{*\prime})\\ &= \boldsymbol{\Sigma}_{a_i-1\times a_i-1},\end{aligned}$$

where $\boldsymbol{J}$ is an $a_i \times a_i$ square matrix of ones, $\boldsymbol{O}$ is an $a_i \times a_i$ matrix of zeroes, and $\boldsymbol{I}$ is the $a_i \times a_i$ identity matrix. □

A.3 Simulation Data Generation Mechanism

We adopt a data generation mechanism which produces arm-level data, and then transform the resulting arm-level data to contrast-based summaries. See Appendix A.4 for details on how the relative effect parameters were set. After setting the parameter values and the network structure, datasets are generated from the model in Section 2.2.1. First, $\text{logit}(\boldsymbol{\pi}_i)$, $i = 1, \ldots, I$ is drawn from Equation (2.2) and each binomial probability π_{ij} is determined via the inverse logit function. Then binomial counts for each arm R_{ij} are drawn from $\text{Bin}(n_{ij}, \pi_{ij})$.

Once outcomes are generated for all trial arms, the data are transformed to contrast-level summaries. The appropriate summary for the contrast between arm 2 and arm 1 in trial i is the log-odds ratio

$$y_{i2}^* = \log\left(\frac{R_{i2}/(n_{i2}-R_{i2})}{R_{i1}/(n_{i1}-R_{i1})}\right) \tag{A.3}$$

with corresponding standard error $SE_{i2} = \sqrt{\frac{1}{R_{i2}} + \frac{1}{n_{i2}-R_{i2}} + \frac{1}{R_{i1}} + \frac{1}{n_{i1}-R_{i1}}}$ (Borenstein et al., 2011).

A.4 Parameter Calculation in Simulations

Here we describe the process by which we set the relative effect parameters. In each network and sample size setting, we simulate two datasets corresponding to two different anchors.

Let z and z' represent the first and second anchor treatments which are to be used in a given network respectively. First, the relative effects of the single component treatments relative to the first anchor treatment z, $d_{z,j}$ are set. Then the relative effects for single components with the second anchor as the comparator, $d_{z',k}$, are determined according to the consistency equations, that is, $d_{z',k} = d_{z,k} - d_{z,z'}$. The relative effect parameter vector corresponding to z as the anchor is $\boldsymbol{d}^{(z)}$, a $(C-1)$-length vector containing the relative effects for all components (except z) compared to treatment z, and similarly for $\boldsymbol{d}^{(z')}$.

The true multicomponent relative effects are needed to compute bias in the simulation. The multicomponent effects relative to z when z is the anchor are given by the additivity equation (2.1), that is, the sum of its component effects relative to z, and similarly for when z' is the anchor. For comparison purposes, we transform the multicomponent relative effects relative to z' using the consistency equations so we can compare the multicomponent effects relative to z when z is the anchor to the multicomponent effects relative to z when z' is the anchor (these are shown in Table 2.3). That is, we compute $d_{z,k}$ and $d_{z',k}$ for all multicomponent treatments according to $d_{z,k} = \sum_{x \in k} d_{z,x}$ and $d_{z',k} = \sum_{x \in k} d_{z',x}$. Then we calculate the multicomponent effects relative to z *when* z' *is the anchor* according to the consistency equations, that is,

$$\begin{aligned} d_{z,k}(\text{where } z' \text{ is the anchor}) &= d_{z',k}(\text{where } z' \text{ is the anchor}) - d_{z',z} \\ &= \sum_{x \in k} d_{z',x} - d_{z',z}. \end{aligned}$$

Appendix B

Chapter 3 Appendix

B.1 Ranking Probabilities in the Minimum-Certainty Hierarchy Scenario

In Section 3.2, we defined the scenario where there is minimum certainty in a ranking to be the case when all SUCRA values are equal to 0.5. We illustrate this scenario where all ranking distributions are flat, that is, the probability that treatment i takes rank u is equal to $1/N$ for all k and u from 1 to N. However, other configurations of ranking probabilities are possible.

In Table B.1, we show two illustrative examples where the SUCRA values of all $N = 4$ treatments are equal to 0.5, but the ranking probabilities are not all equal to $1/N$. In the first scenario, the ranking distribution for treatment 3 is no longer flat. The probabilities are now higher for ranks 2 and 3 and lower for ranks 1 and 4. This corresponds to a reduction in the variance of treatment 3's ranking distribution compared to the variance of a flat ranking distribution. However, the ranking distribution of treatment 4 must also now be non-flat to compensate for this and maintain all SUCRA values equal to 0.5. The ranking distribution of treatment 4 now has higher probabilities at the highest and lowest ranks, corresponding to an increase in its variance compared to a flat distribution. In general, if the variance of one treatment's ranking distribution is less than that of a flat ranking distribution, the variance of the ranking distribution of another treatment (or treatments) must be greater than that of a flat ranking distribution. Recall that in Section 3.2.4, we argued that the variance of a treatment's ranking distribution is a measure of the certainty in that treatment's ranking. Thus, when all SUCRAs are 0.5, any treatment

with more certainty in its place in the hierarchy will be balanced by a reduction in the certainty of the place of one or more other treatments.

Scenario 2 shows another example where the ranking distribution of treatments 3 and 4 are not flat and all SUCRA values are equal to 0.5. In this case, the variance of the ranking distributions for treatments 3 and 4 are equal to the variance of a flat ranking distribution. These distributions can be described as multimodal, since each treatment has two different ranks with increased probabilities, and those ranks are not adjacent (for treatment 3, the ranks are 1 and 3). This shows that there is still very little certainty about which rank the treatments are likely to take in the hierarchy. Although the distributions are not flat, they have the same variance as a flat ranking distribution, and they provide similar information about which ranks the treatments are likely to take.

In summary, multiple configurations of probabilities are possible which give the minimum certainty in hierarchy scenario. If all distributions are not flat, any increase in certainty (reduction in variance) must be compensated by a decrease in certainty (increase in variance) in the ranking distributions of one or more treatments. Multimodality in the ranking distributions which does not change the certainty in a treatment's rank is also possible. Thus, all possible configurations of ranking probabilities where $\text{SUCRA}(k) = 0.5$ for all k provide the same amount of certainty about the treatment hierarchy overall.

Table B.1: Two scenarios illustrating different joint ranking distributions where all SUCRAs are equal to 0.5. In Scenario 1, the certainty in the rank of treatment 3 is increased, with an equal and opposite decrease in the certainty of the rank of treatment 4. In Scenario 2, both treatments 3 and 4 have multimodal ranking distributions which have equal certainty to those treatments with flat ranking distributions.

	Scenario 1				Scenario 2			
Rank (u)	Treatment (k)				Treatment (k)			
	1	2	3	4	1	2	3	4
1	0.25	0.25	0.10	0.40	0.25	0.25	0.30	0.20
2	0.25	0.25	0.40	0.10	0.25	0.25	0.10	0.40
3	0.25	0.25	0.40	0.10	0.25	0.25	0.40	0.10
4	0.25	0.25	0.10	0.40	0.25	0.25	0.20	0.30
SUCRA	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$\mathbf{E}\left[\text{rank}(k)\right]$	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5

B.2 Obtaining A Simulation-Based Reference Distribution for POTH

Suppose we have calculated POTH for a treatment hierarchy generated by an NMA and we wish to obtain a relevant reference distribution. Our network comprises M pairwise comparisons connecting N treatments. The network structure is described by the $M \times N$ design matrix $\boldsymbol{B}$, as described in Section 1.1.2. Also important to distinguish our network is a vector of the variances (squared standard errors) corresponding to the M observed pairwise comparisons, $\boldsymbol{SE}^2 = (SE_1^2, SE_2^2, ..., SE_M^2)'$ and the estimated heterogeneity parameter $\hat{\sigma}^2$, which equals zero if a common-effect model was used. Let $\boldsymbol{\theta} = (\theta_{1k}, \theta_{2k}, ..., \theta_{Nk})'$ be a vector of relative effects of the N treatments relative to an arbitrary reference treatment k which describes our setting of interest. The reference distribution is that generated by repeated sampling from a network with structure described by $\boldsymbol{B}$, precision in the data determined by $\boldsymbol{SE}^2$, true heterogeneity equal to $\hat{\sigma}^2$, and with true relative effects $\boldsymbol{\theta}$. The methods described in this section are implemented in the 'refdist' function of the R package **poth**.

The relative effect vector $\boldsymbol{\theta}$ should be set to represent a relevant reference point. A common setting would be $\boldsymbol{\theta} = (0, 0, ..., 0)'$ which corresponds to no true treatment differences in the network. This setting allows us to understand the distribution of POTH in the network described by $\boldsymbol{B}$, $\boldsymbol{SE}$, and $\hat{\sigma}^2$ if all observed treatment differences are solely due to chance. The use of no treatment differences as a reference point may be less relevant in networks which include placebo and several active treatments if it is commonly understood that placebo is inferior to any active treatment for the research question under consideration. An alternative in this case is to set $\boldsymbol{\theta} = (d, 0, ..., 0)'$, where treatment 1 corresponds to the placebo or inactive treatment. This corresponds to the assumption that placebo is inferior to all active treatments by the value of d, while observed treatment differences between *active* treatments are solely due to chance.

Once a suitable $\boldsymbol{\theta}$ is chosen, study-level relative effects are sampled according to

$$\boldsymbol{y} \sim \text{MVN}_M(\boldsymbol{B\theta}, \boldsymbol{W}_{adj}^{-1}), \tag{B.1}$$

where $\boldsymbol{W}_{adj}$ as defined in Section 1.1.2. If there are multi-arm studies present, the simulated relative effects from this model will not be transitive. Thus, for each a-arm study, we use the first $a-1$ simulated study-level relative effects to compute the remaining $\binom{a}{2} - (a-1)$ study-level comparisons to ensure the relative effects are transitive within the study.

A NMA model is then fit to the simulated data $\boldsymbol{y}$ with assumed standard errors equal to $\boldsymbol{SE}$. If a common-effects model was used to simulate the data, then a common-effects

model is used to analyse the data, and vice versa. POTH is then calculated from the resulting effect estimates and their standard errors as described in Section 3.2.3. The data simulation and POTH calculation process is repeated L times to give a sample of L POTH values, representing the reference distribution.

B.3 Proof that $V(N)$ is Uniquely Minimized Under Maximum Certainty in the Hierarchy

We claim that there is a unique minimum for $V(N)$ when $p_{ku} = 1$ for $u = \mathbf{E}[\text{rank}(k)]$ and $p_{kv} = 0$ for all $v \neq u$, for $k = 1, \ldots, N$. We will show that this is a unique minimum by showing that $V(N) = 0$ implies $p_{ku} = 1$ for $u = \mathbf{E}[\text{rank}(k)]$ and $p_{kv} = 0$ for all $v \neq u$, for $k = 1, \ldots, N$.

If $V(N) = 0$, then we have

$$\begin{aligned} 0 &= \frac{1}{N} \sum_{k=1}^{N} \sum_{u=1}^{N} \left(u - \mathbf{E}[\text{rank}(k)]\right)^2 p_{ku} \\ \implies 0 &= \left(u - \mathbf{E}[\text{rank}(k)]\right)^2 p_{ku} \quad \forall \quad k, u = 1, \ldots, N. \end{aligned} \tag{B.2}$$

For (B.2) to hold, we have one of $u = \mathbf{E}[\text{rank}(k)]$ or $p_{ku} = 0$ for every u. Considering the additional constraint that $\sum_{k=1}^{N} p_{ku} = 1$ and $\sum_{u=1}^{N} p_{ku} = 1$ for all $k, u = 1, \ldots, N$, this can only be satisfied where $p_{ku} = 1$ for $u = \mathbf{E}[\text{rank}(k)]$ and 0 otherwise. Thus, $V(N)$ is uniquely minimized under the maximum certainty in hierarchy scenario. Further, this shows that POTH is uniquely maximized under maximum certainty in the hierarchy.

B.4 Networks Producing Smallest and Largest POTH in NMA Database

The network which produced the smallest POTH observed in the database of NMAs investigated the risk of death for different terms of dual antiplatelet therapy after drug-eluting stents (Bulluck et al., 2015). There were four different time periods (treatments) investigated and data were taken from seven studies. The record ID in **nmadb** is 482231. The POTH calculated for this network was 0.096. In the simulation-based reference distribution where there were no true treatment differences, there was a 90.1% chance of having

a larger POTH value. This indicates a very low level of certainty in the hierarchy, and it would be likely to observe a similar or higher level of certainty even if there were no true treatment differences.

The network which produced the largest POTH observed in the database of NMAs regarded the relative risk of major cardiovascular events at different statin doses (Ribeiro et al., 2013). Statins were classified as low, intermediate, and high doses and compared to placebo, resulting in four treatments total. There were 39 studies in the network. The record ID in **nmadb** is 501379. The POTH value is over 0.999. In a reference setting where there were no true treatment differences, there was a 0% chance of observing a larger POTH. We therefore have a high extent of certainty in the hierarchy and it would be extremely unlikely to observe this level of certainty if there were no true treatment differences.

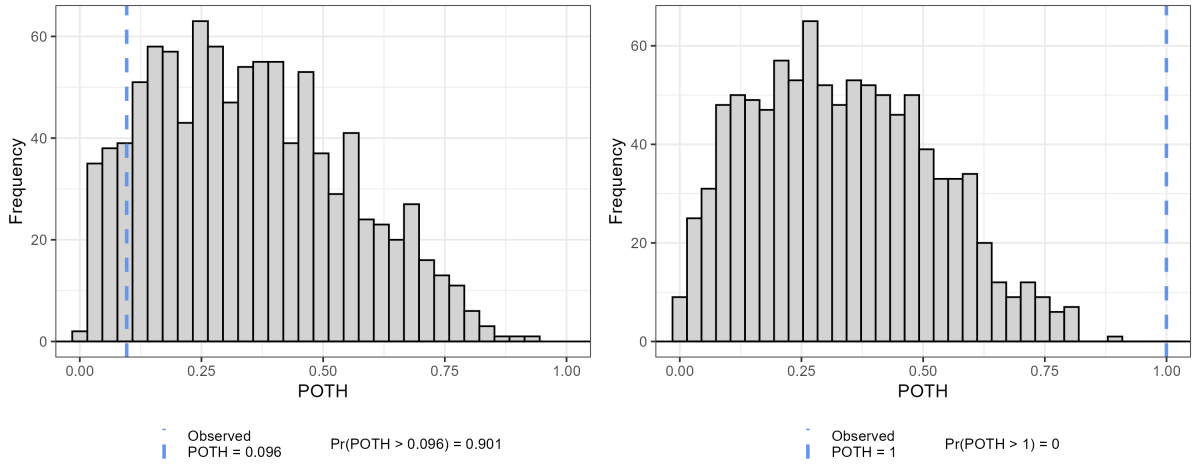


Figure B.1: Reference distributions for POTH where there are no true treatment differences in the network with the lowest (left) and highest (right) POTH observed in the database analysis.

The estimated between-study variance σ^2, the I^2, and POTH for both networks are shown in Table B.2. Forest plots for both networks are shown in Figure B.2.

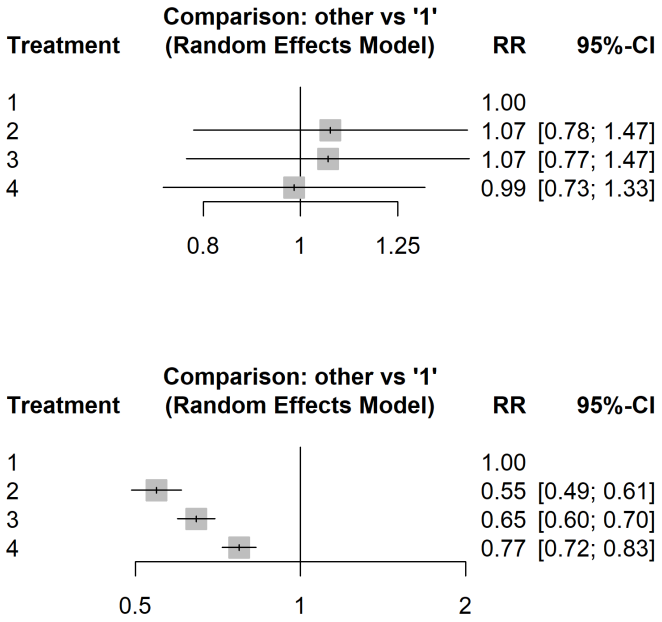


Figure B.2: Forest plots showing estimates and 95% CIs for relative effects of network producing the lowest (top) and highest (bottom) POTH in the database of NMAs.

Table B.2: Between-study heterogeneity (σ^2), I^2, and POTH for the two extreme networks in the database.

	Lowest POTH Network	Highest POTH Network
σ^2	0.00	0.00
I^2 (95% CI)	0% (0-79%)	0% (0-37%)
POTH	0.096	0.999762

B.5 Extra Data Examples

Here we include two supplementary worked examples to further demonstrate the use and interpretation of POTH. In each example, a random-effects NMA model was fit to the data using the REML method to estimate heterogeneity.

B.5.1 Antifungal Treatments to Prevent Mortality for Solid Organ Transplant Recipients

A meta-analysis of treatments for preventing fungal infection in solid organ transplant recipients was conducted by Playford et al. (2004). They report data from ten studies. There are five treatments to compare; four anti-fungal treatments and a control. The data is available in **nmadb** using record ID 501370 (Papakonstantinou, 2019). The outcome of interest is mortality.

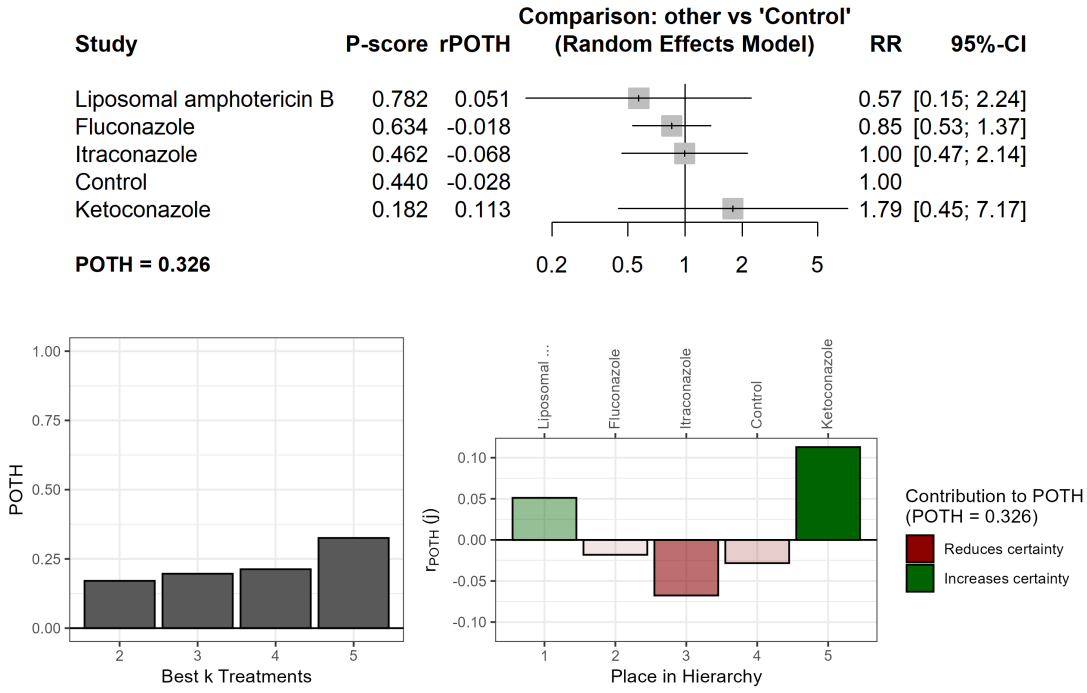


Figure B.3: Above: Estimated confidence intervals (CI) for the relative risk of mortality for anti-fungal treatments versus control in the organ transplant network. P-scores and residual POTH are shown to the left of CIs. RRs less than one favour the intervention and RRs greater than one favour control. The treatments are organized from largest (highest ranking in hierarchy) to smallest (lowest ranking in hierarchy) P-scores. A positive residual indicates that the treatment contributes to the certainty in the hierarchy, while a negative residual indicates that the treatment reduces the certainty in the hierarchy.
Below, left: Best k treatments POTH for $k = 2, ..., N$ in the organ transplant network.
Below, right: POTH residuals $r_{\text{POTH}}(j)$ for all treatments in the network.

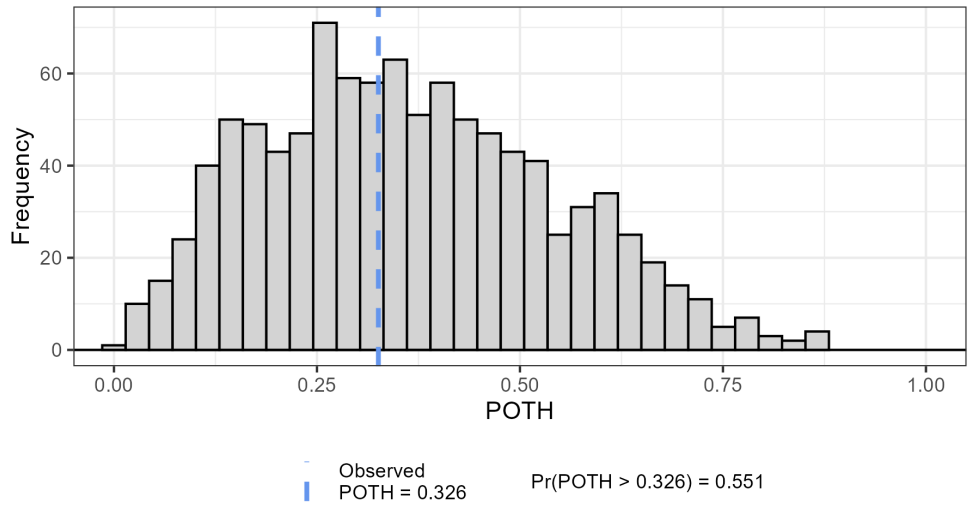


Figure B.4: Reference distribution for antifungals example where no treatment differences exist.

The 95% CIs, P-scores, and POTH residuals from the random-effects NMA for the transplant data are shown in Figure B.3. The POTH calculated using the P-scores is 0.326, reflecting low certainty in the treatment hierarchy. We obtained a reference distribution for this network where no treatment differences exist and found that 55.1% of the simulated datasets had a POTH value greater than 0.326. This means there is a 55.1% chance that a treatment hierarchy obtained under the reference scenario will be more certain than the treatment hierarchy we observed. This POTH represents a low extent of certainty in the hierarchy and it would be reasonable to see a similar level of certainty in a network with no true treatment differences. Correspondingly, there is no evidence of treatment differences from control in the forest plot in Figure B.3.

Turning to the POTH residuals, we note that Ketoconazole, the lowest-ranked treatment in the hierarchy, has the largest positive residual equal to 0.113. Despite its wide CI, the estimated relative effect of Ketoconazole is noticeably different from those of the other treatments, so overall it adds certainty to the hierarchy. The best k treatments POTH, $k = 2, \ldots, 5$ is illustrated in the bottom panel of Figure B.3. There is a steady increase in POTH as more treatments are considered, with the largest jump occurring when k increases from 4 to 5, corresponding to the addition of Ketoconazole to the set of competing treatments.

Suppose we wish to produce a hierarchy of all non-control treatments in addition to

the hierarchy of all treatments. We use $\text{POTH}_{\boldsymbol{T}}$ to quantify the certainty in the hierarchy of all non-control treatments by defining $\boldsymbol{T}$ = {Liposomal amphotericin B, Fluconazole, Itraconazole, and Ketoconazole}. The $\text{POTH}_{\boldsymbol{T}}$ for the non-control treatments is 0.354. In fact, this corresponds to the POTH residual for control. Removing control from the set of treatments leads to an increase in POTH of 0.028. There is slightly more certainty in the hierarchy of non-control treatments compared to the hierarchy of all treatments, but the certainty is still low.

B.5.2 Efficacy of Pharmacological Treatments for Persistent Depressive Disorder

Kriston et al. conducted an NMA assessing the efficacy and acceptability of treatments for persistent depressive disorder (Kriston et al., 2014). Data from 45 studies comparing 29 treatments (28 pharmacological treatments, and placebo) were accessed in **nmadb** from record ID 480666 (Papakonstantinou, 2019). The binary outcome of interest is an improvement of 50% or more on a symptom severity scale, so larger values are desirable.

A forest plot in Figure B.6 shows the results of the random-effects NMA. The treatments are organized by decreasing P-score. The POTH is calculated as 0.559, corresponding to a moderate extent of certainty. In FigureB.5, we show the reference distribution for a similar network where there were no true treatment differences. We found a 15.3% chance of obtaining a POTH value greater than 0.559 in the reference scenario. The hierarchy is thus moderately certain, and it would be unlikely to observe a more certain hierarchy in a network with no treatment differences.

The POTH residuals for all 29 treatments and the best k treatments POTH in the depression network for $k = 2, \ldots, 29$ is illustrated in Figure B.7. Based on the $r_{\text{POTH}}(j)$, the treatments contributing the most to POTH are the four worst treatments in the hierarchy, and the top treatment in the hierarchy, phenelzine. Dothiepine is the treatment with the lowest POTH residual equal to -0.022, indicating that it reduces the certainty in the hierarchy the most out of all treatments in the network. This is no surprise as it has a very wide CI and similar point estimates to neighbouring treatments in the hierarchy. As more treatments are considered, the POTH rises relatively steadily. This is reflective of the lack of clustering of estimated treatment effects as seen in the forest plot. The best k treatments POTH is the lowest when considering only the top 2 treatments, phenelzine and duloxetine.

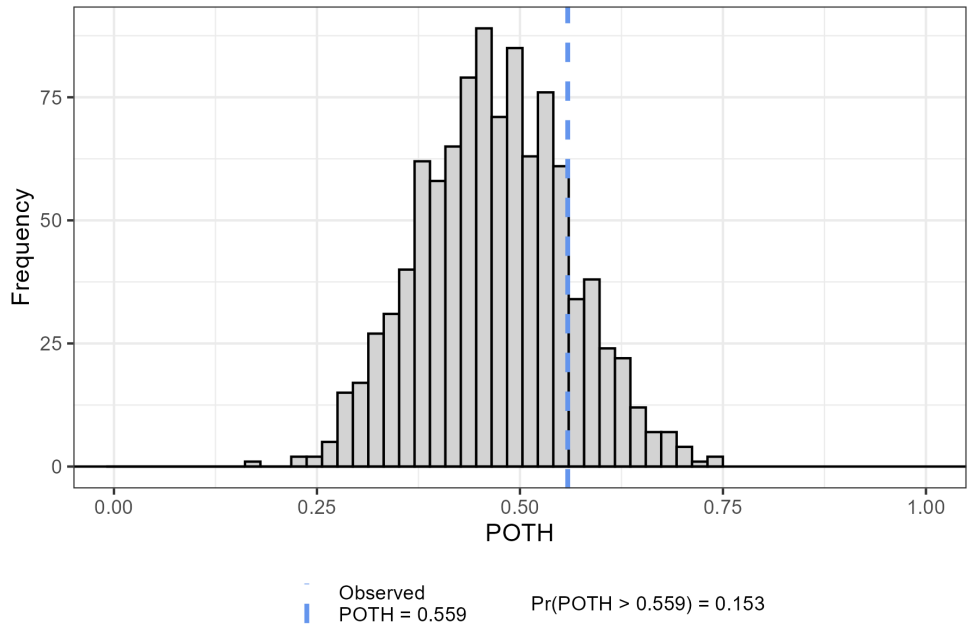


Figure B.5: Reference distribution for POTH in the depression example, where there are no true treatment differences.

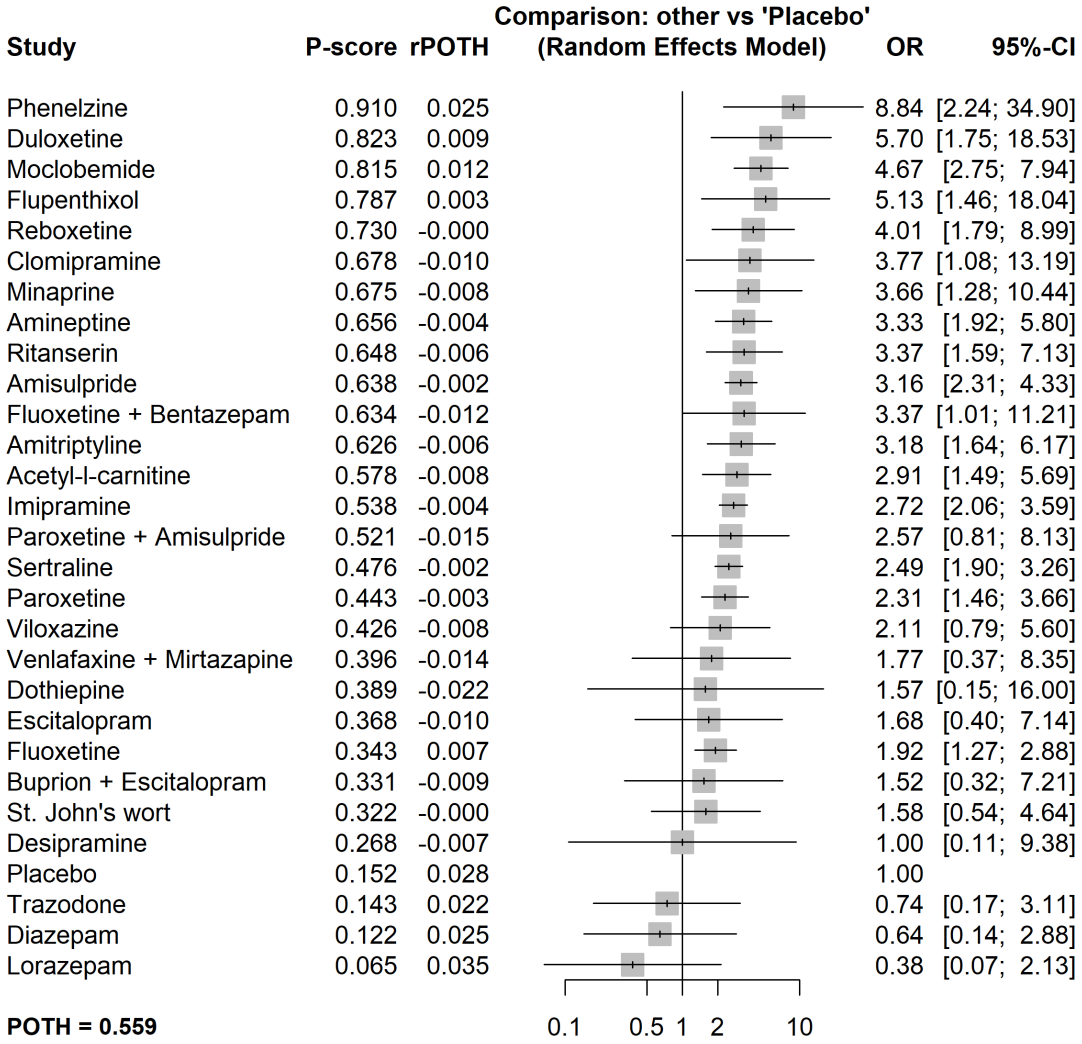


Figure B.6: Estimated CIs for the odds ratio of 50% improvement of symptom severity for pharmacological interventions versus placebo in the depression network. P-scores and residual POTH are shown to the left of CIs. ORs greater than one favour the intervention and ORs less than one favour placebo. The treatments are organized from largest to smallest P-scores. A positive residual indicates that the treatment contributes to the certainty in the hierarchy, while a negative residual indicates that the treatment reduces the certainty in the hierarchy.

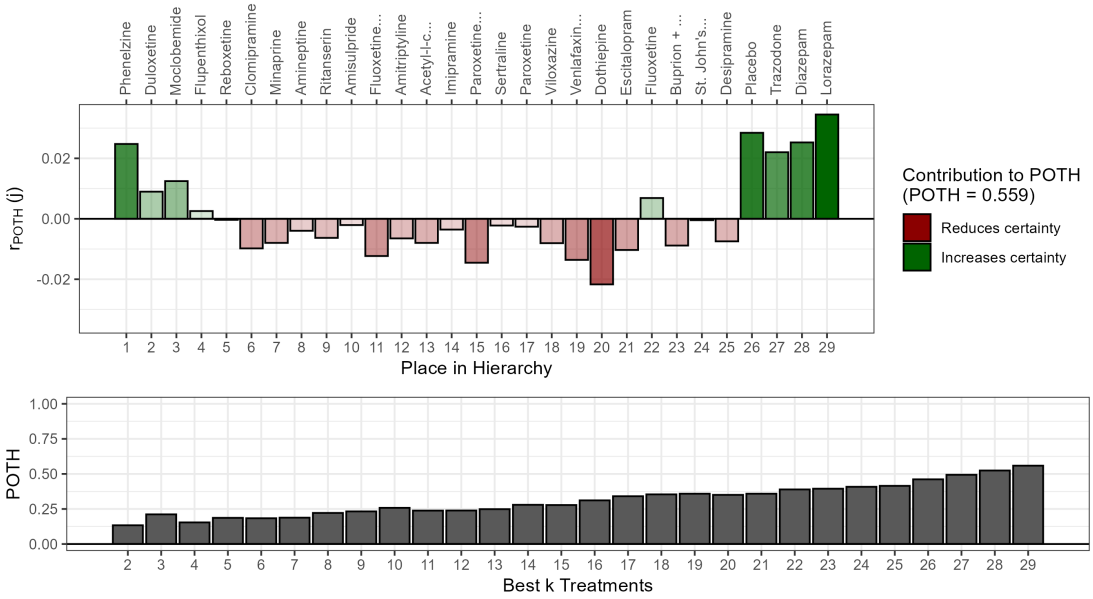


Figure B.7: Left: Best k treatments POTH for $k = 2$ to N treatments in the depression network.
Right: POTH residuals $r_{\text{POTH}}(j)$ for all treatments in the network.

B.6 What is the Distribution of Precision of Treatment Hierarchies in Networks with No Treatment Differences? A Simulation Study

B.6.1 Set-up

The aim of this simulation study is to understand the distribution of POTH when there are no treatment differences present. We also investigate the impact of network geometry, precision of the data, and heterogeneity on the POTH. We simulated data from a random-effects NMA model where all relative treatment differences are equal to zero, the between-study heterogeneity is equal to σ, the observed data in study i is a mean difference y_i, and the standard error is given by $SE(y_i)$.

$SE(y_i)$ was set to 0.12, 1.10, or 2.07. Between-study heterogeneity σ was set to 0 (representing no heterogeneity), 0.12, and 2.07. We investigate three possible network

geometries when there are 5 treatments in the network: star, line, and fully connected. For each network geometry, we assume there are two studies per treatment comparison.

For a given network geometry and settings for $SE(y_i)$ and σ, data was simulated from a random-effects NMA model. The vector of relative effects is set to $(0, \ldots, 0)'$, so all five treatments are equally effective. Study-specific relative treatment effects are simulated according to

$$\delta_i \sim N(0, \sigma^2) \tag{B.3}$$

for $i = 1, \ldots, I$ where I is the number of studies in the network. Observed data is simulated according to

$$y_i \sim N(\delta_i, SE(y_i)^2). \tag{B.4}$$

For each scenario, we simulated 8000 NMA datasets. Each NMA dataset was analysed using a frequentist random-effects NMA model using the R package **netmeta** using the REML method for estimating σ (Balduzzi et al., 2023). The POTH for each simulated NMA dataset was calculated using the **poth** R package (Wigle and Schwarzer, 2024).

B.6.2 Results

The mean of POTH observed in the 8000 simulated NMA datasets for each simulation scenario is shown in Figure B.8. The mean of POTH for the different settings range from 0.327 to 0.462. For a given level of heterogeneity and network type, as the precision of the data increases, the mean POTH increases. This makes sense, as more precise data will correspond to tighter confidence intervals on the relative effects and therefore higher certainty in the rank probability distributions. In the absence of between-study heterogeneity (where $\sigma = 0$), the mean POTH is the smallest for the line network and the largest for the full network. In other words, the more a network relies on indirect evidence, the lower the POTH will be, on average. Finally, we see an increase in the mean POTH as the heterogeneity increases. At first glance this may seem counter-intuitive, however, we speculate that the reason for this trend is that the data are generated assuming no treatment differences. Because of this, as heterogeneity increases, the magnitude of the observed effect sizes in the simulated datasets increases, which can lead to more pronounced differences in P-scores, and thus, higher POTH.

We show the observed distribution of POTH for the combinations of low and high standard error and heterogeneity for all three network types in Figure B.9. The fully connected network has the most peaked distributions while the line network has the flattest distributions. Interestingly, in all scenarios, a relatively wide range of POTHs were observed.

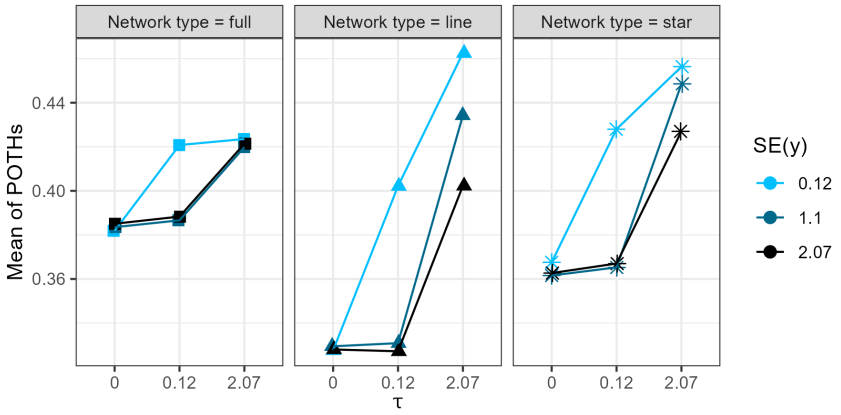


Figure B.8: Mean of POTH over the simulated datasets for different combinations of heterogeneity (represented by τ), network type, and standard error in the data.

For all scenarios, there is over a 55% chance that the POTH will be less than 0.5. On the other hand, POTH values over 0.8 are very rare - less than a 17% chance of being observed, and less than 1% for some scenarios. Accordingly, we tentatively assign an observed POTH value less than 0.5 to convey low certainty and an observed POTH greater than 0.8 to convey high certainty. POTHs close to 1 were still observed for a small subset of simulated datasets, which illustrates that POTH can only summarise the certainty in the treatment hierarchy based on the observed data, and even when no true underlying treatment differences exist, we may still observe data which suggest a clear hierarchy of treatments.

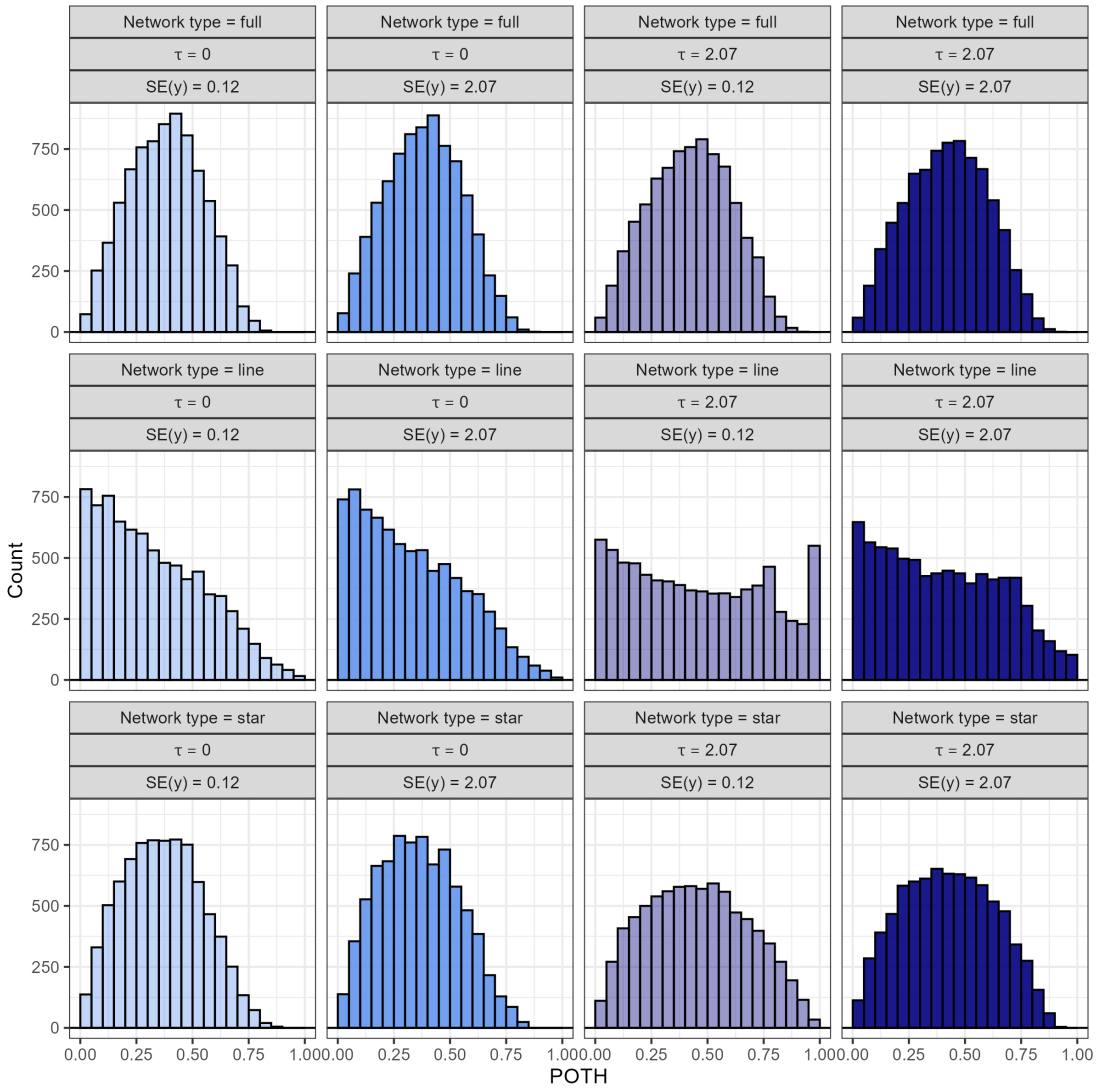


Figure B.9: Histograms showing the distribution of POTH observed for different settings in the simulation study.

Appendix C

Chapter 4 Appendix

C.1 Note on Sample and Population Sizes

The information on number of facilities and facility type at each site comes from confidential, anonymized data provided by the MERC and BC OGRIS. This data includes all facilities and wells that had 'Active' status in Petrinex, including shut-in, in the province of BC as of June 2021. We use the June 2021 sample sizes in this thesis because the information on the number of active facilities of each type at each site was only available to us for June 2021. The survey was conducted in September and October of 2021 and so the sample and population sizes for each facility type reported in the supplementary information of Johnson et al. (2023) were obtained from Petrinex in September and October 2021 respectively. The number of active facilities and wells in the Petrinex database decreased from June 2021 to September 2021, so the sample sizes used to analyse the data in this thesis are equal to or greater than those reported in Johnson et al. (2023). It is worth noting that the decrease in the database does not necessarily mean a decrease in the number of active facilities in the field. For facilities besides wells, the population size was determined using the number of each facility type with Active status in Petrinex in September 2021, when most of the surveys were conducted. For wells, the population and sample sizes were determined using the number of wells with active status in Petrinex as of June 28, 2021 because this represents the breakdown of wells at each site available to us.

C.2 Treatment of Wells

Johnson et al. (2023) differentiate between on-site wells (wells which share a location with other facilities) and off-site wells (wells at completely separate locations from other facility types). They attribute emissions from on-site wells to other facilities at the site as it may not be possible to distinguish between components associated with the well versus the other facilities, so only off-site wells have emissions attributed to them. We take a different approach because the data from the MERC and BC OGRIS attributed emissions to both on- and off-site wells and does not differentiate between on- and off-site wells.

Unlike other facility types, multiple wells may share the same pieces of equipment. For example, four wells may be connected to a single tank, and thus, if the tank is emitting methane, it cannot be attributed to a single well. For every leaking component that is associated with a well, we sum the estimated average emission rate for that survey occasion for all detected well components at that site and divide by the number of wells at the site, and set

$$\hat{\overline{Y}}_{pt\cdot} = \frac{\sum_{p \in S^{\mathrm{I}}(\text{Wells})} \hat{\overline{Y}}_{pt\cdot}}{\text{Number of wells at site}} \tag{C.1}$$

for all the wells at that site for all t. The variance is given by

$$\sum_{p \in S^{\mathrm{I}}(\text{Wells})} \hat{\mathrm{V}}_{pt\cdot} / (\text{Number of wells at site})^2.$$

This is similar to the approach in Johnson et al. (2023), where emissions are divided equally among wells at the same site.

C.3 Summary of Notation Used in Chapter 4

Table C.1: A summary of the notation and terms used in Chapter 4.

Notation	Interpretation
ptq	Represents the three-stage sampling design
m	Represents the measurement model
$\mathrm{I}, \mathrm{II}, \mathrm{III}$	Represents stage one, stage two, and stage three of the sampling design, respectively
$\hat{}$	Represents an estimator with respect to the ptq-design

Notation	Interpretation
$\tilde{}$	Represents an estimator with respect to the measurement model
(h)	Indicates that the quantity is a property of stratum h
$T; \hat{T}$	The target of estimation from a design-based perspective; its ptq-estimator, respectively
$Y_{ptq}; u_{ptq}; a_{ptq}$	The true emission rate in kg/h, wind speed in m/s, and altitude of the plane in m from component p on day t in pass q
M_{ptq}	The measured emission rate from component p on day t in pass q
$\overline{Y}_{pt\cdot}$	The true average emission rate from component p on day t, averaged over the passes
$\hat{\overline{Y}}_{pt\cdot}; \hat{\overline{Y}}^{\text{IPW}}_{pt\cdot}; \hat{\overline{Y}}^{\text{Hájek}}_{pt\cdot}$	A generic ptq-estimator of $\overline{Y}_{pt\cdot}$, the IPW estimator of $\overline{Y}_{pt\cdot}$, and the Hájek estimator of $\overline{Y}_{pt\cdot}$
$\overline{Y}_{p\cdot\cdot}; \hat{\overline{Y}}_{p\cdot\cdot}$	The true average emission rate from component p, averaged over all days and passes; its generic ptq-estimator
$\mathrm{V}^{\text{3 stage,}}; \hat{\mathrm{V}}^{\text{3 stage}}$	The true variance of $\hat{T}$ with respect to the ptq-design; its estimator
$\mathrm{V}^j; \hat{\mathrm{V}}^j$	The true variance contribution from stage j for $j = \mathrm{I}, \mathrm{II}, \mathrm{III}$; their ptq-estimators
$\mathrm{V}_{p\cdot\cdot}; \hat{\mathrm{V}}_{p\cdot\cdot}$	The true ptq-variance of $\hat{\overline{Y}}_{p\cdot\cdot}$; its ptq-estimator
$\mathrm{V}_{pt\cdot}; \hat{\mathrm{V}}_{pt\cdot}$	The true ptq-variance of $\hat{\overline{Y}}_{pt\cdot}$; its ptq-estimator
$U^{\mathrm{I}}; U^{\mathrm{II}}_p; U^{\mathrm{III}}_{pt}$	The population of PSUs (components); The population of SSUs (days) for component p; The population of TSUs (passes) for component p and day t
$S^{\mathrm{I}}; S^{\mathrm{II}}_p; S^{\mathrm{III}}_{pt}$	The stage I sample of components; The stage II sample of days for component p; the stage III sample of passes for component p on day t
$S^{\mathrm{II}*}_p$	The stage II sample of days for which methane was detected at least once; replaces S^{II}_p when $\hat{\overline{Y}}^{\text{Hájek}}_{pt\cdot}$ is used
$\pi^{\mathrm{I}}_p; \pi^{\mathrm{I}}_{pl}$	The probability that p is selected into S^{I}; the probability that components p and l were selected into S^{I}
$\pi^{\mathrm{II}}_{t\|p}; \pi^{\mathrm{II}}_{tu\|p}$	The probability that t is selected into S^{II}_p given that p is in S^{I}; the probability that t and u are selected into S^{II}_p given that p is in S^{I}
$\pi^{\mathrm{III}}_{q\|pt}; \pi^{\mathrm{III}}_{qr\|pt}$	The probability that pass q is included in S^{III}_{pt}, that is, the probability that methane is detected in pass q, given that p is in S^{I} and t is in S^{II}_p; the probability that methane is detected in passes q and r given that p is in S^{I} and t is in S^{II}_p

Notation	Interpretation
$\pi^{\text{II}\ *}_{t\|p}; \pi^{\text{II}\ \ *}_{tu\|p}$	The probability that t is selected into S^{II}_p given that p is in S^{I} AND methane is detected in at least one pass on day t; the probability that t and u are selected into S^{II}_p given that p is in S^{I} AND methane is detected at least once on days t and u
$\pi^{\text{III}\ *}_{q\|pt}; \pi^{\text{III}\ \ *}_{qr\|pt}$	The probability that methane was detected in pass q given that t is in $S^{\text{II}*}_p$ and p is in S^{I}; the probability that methane was detected in passes q and r given that t is in $S^{\text{II}*}_p$ and p is in S^{I}
ϕ_{ptq}	The probability of detecting methane from component p on day t in pass q. Calculated using Equation (1)
$\phi_{pt\cdot}; \hat{\phi}_{pt\cdot}$	The probability of detecting methane at least once from component p on day t; its estimate using MI
$N_{h(p)}, n_{h(p)}$	The number of facilities in U^{I} belonging to the stratum containing component p; the number of facilities in S^{I} belonging to the stratum containing component p
D; d_p	The number of days in the population; the number of days on which component p is sampled
Q_{pt}	The number of passes made over component p on day t
$\tau; \tilde{\tau}$	The target of estimation considering the ptq-design and the measurement model; its estimator
$\text{V}^{\text{Total}}; \tilde{\text{V}}^{\text{Total}}$	The true variance of $\tilde{\tau}$ with respect to ptq and m; its estimator
$\text{V}^{j}; \tilde{\text{V}}^{j}$	The true variance contribution from j with respect to ptq and m; its estimator, for $j = m, \text{I}, \text{II}, \text{III}$

C.4 Variance Formulas for Multi-Stage Approach

The true variance of $\hat{T}$ with respect to the ptq-design for generic sampling designs at stages I and II is given by

$$\sum_{p\in U^{\text{I}}}\sum_{p\in U^{\text{I}}}(\pi^{\text{I}}_{pl}-\pi^{\text{I}}_{p}\pi^{\text{I}}_{l})\frac{\overline{Y}_{p\cdot\cdot}}{\pi^{\text{I}}_{p}}\frac{\overline{Y}_{m\cdot\cdot}}{\pi^{\text{I}}_{l}}+\sum_{p\in U^{\text{I}}}\frac{\text{V}_{p\cdot\cdot}}{\pi^{\text{I}}_{p}}, \tag{C.2}$$

where $\mathrm{V}_{p\cdot\cdot}$ is is the design-based variance of $\hat{\overline{Y}}_{p\cdot\cdot}^{\mathrm{IPW}}$ or $\hat{\overline{Y}}_{p\cdot\cdot}^{\mathrm{Hájek}}$ depending on the approach used. $\mathrm{V}_{p\cdot\cdot}^{\mathrm{IPW}}$ is given by

$$\frac{1}{D^2}\left\{\sum_{t\in U_p^{\mathrm{II}}}\sum_{u\in U_p^{\mathrm{II}}}(\pi_{tu|p}^{\mathrm{II}}-\pi_{t|p}^{\mathrm{II}}\pi_{u|p}^{\mathrm{II}})\frac{\overline{Y}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}}}\frac{\overline{Y}_{pu\cdot}}{\pi_{u|p}^{\mathrm{II}}}+\sum_{t\in U_p^{\mathrm{II}}}\frac{\mathrm{V}_{pt\cdot}^{IPW}}{pit}\right\}, \tag{C.3}$$

where

$$\mathrm{V}_{pt\cdot}^{\mathrm{IPW}}=\frac{1}{Q_{pt}^2}\sum_{q\in U_{pt}^{\mathrm{III}}}\frac{(1-\phi_{pt\cdot})}{\phi_{pt\cdot}}Y_{ptq}^2 \tag{C.4}$$

when the non-detection follows Poisson sampling. For the Hájek estimator, $\mathrm{V}_{p\cdot\cdot}^{\mathrm{Hájek}}$ is given by

$$\frac{1}{D^2}\left\{\sum_{t\in U_p^{\mathrm{II}}}\sum_{u\in U_p^{\mathrm{II}}}(\pi_{tu|p}^{\mathrm{II}\ *}-\pi_{t|p}^{\mathrm{II}\ *}\pi_{u|p}^{\mathrm{II}\ *})\frac{\overline{Y}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}\ *}}\frac{\overline{Y}_{pu\cdot}}{\pi_{u|p}^{\mathrm{II}\ *}}+\sum_{t\in U_p^{\mathrm{II}}}\frac{\mathrm{V}_{pt\cdot}^{\mathrm{Hájek}}}{\pi_{t|p}^{\mathrm{II}\ *}}\right\} \tag{C.5}$$

where

$$\mathrm{V}_{pt\cdot}^{\mathrm{Hájek}}\approx\frac{1}{Q_{pt}^2}\left\{\sum_{q\in U_{pt}^{\mathrm{III}}}\phi_{pt\cdot}\frac{(1-\phi_{ptq})}{\phi_{ptq}}(Y_{ptq}-\overline{Y}_{pt\cdot})^2+(\phi_{pt\cdot}-1)\left(\sum_{q\in U_{pt}^{\mathrm{III}}}(Y_{ptq}-\overline{Y}_{pt\cdot})\right)^2\right\}, \tag{C.6}$$

is a linear approximation of the variance using a Taylor series expansion.

C.5 Proof of Equivalency of Sampling Designs

We refer to the original design as the design where the stage II sample is governed by inclusion probabilities $\pi^{\text{II}}_{t|p}$ and $\pi^{\text{II}}_{tu|p}$ and the stage III sample is governed by $\pi^{\text{III}}_{q|pt}$ and $\pi^{\text{III}}_{qr|pt}$. On the other hand, the modified design is the design where the stage II and III samples are governed by $\pi^{\text{II}\,*}_{t|p}$, $\pi^{\text{III}\,*}_{qr|pt}$, $\pi^{\text{III}\,*}_{q|pt}$, and $\pi^{\text{III}\,*}_{qr|pt}$.

Let s^{I} be a fixed but arbitrary sample of PSUs. Similarly, let s^{II}_p be a fixed but arbitrary sample of SSUs for all $p \in s^{\text{I}}$, and let s^{III}_{pt} be a fixed but arbitrary sample of TSUs for all $t \in s^{\text{II}}_p$ and $p \in s^{\text{I}}$. Together, let $s = \{p \in s^{\text{I}}, t \in s^{\text{II}}_p, q \in s^{\text{III}}_{pt}\}$ represent a fixed but arbitrary sample of PSUs, SSUs and TSUs. To show equivalence between the original and modified design, we will show that the probability of observing s is equal under both designs.

Under the original design, the probability of observing s is

$$\begin{aligned}
&\Pr(S^{\text{I}} = s^{\text{I}}, S^{\text{II}}_p = s^{\text{II}}_p \text{ for all } p \in s^{\text{I}}, S^{\text{III}}_{pt} = s^{\text{III}}_{pt} \text{ for all } p \in s^{\text{I}} \text{ and } t \in s^{\text{II}}_p) \\
&\qquad = \Pr(S^{\text{I}} = s^{\text{I}}) \left[\prod_{p \in s^{\text{I}}} \Pr(S^{\text{II}}_p = s^{\text{II}}_p \mid S^{\text{I}} = s^{\text{I}}) \left\{ \prod_{t \in s^{\text{II}}_p} \Pr(S^{\text{III}}_{pt} = s^{\text{III}}_{pt} \mid S^{\text{I}} = s^{\text{I}}, S^{\text{II}}_p = s^{\text{II}}_p) \right\} \right].
\end{aligned}$$

Under the modified design, let $s^{\text{II}*}_p = \{t \in s^{\text{II}}_p \text{ s.t. } Z_{pt} = 1\}$ where $Z_{pt} = 1$ if there are any detections at component p on day t and zero otherwise. Then the probability of s in the modified design is

$$\begin{aligned}
&\Pr(S^{\text{I}} = s^{\text{I}}, S^{\text{II}*}_p = s^{\text{II}*}_p \text{ for all } p \in s^{\text{I}}, S^{\text{III}}_{pt} = s^{\text{III}}_{pt} \text{ for all } p \in s^{\text{I}} \text{ and } t \in s^{\text{II}*}_p) \\
&\qquad = \Pr(S^{\text{I}} = s^{\text{I}}) \left[\prod_{p \in s^{\text{I}}} \Pr(S^{\text{II}*}_p = s^{\text{II}*}_p \mid S^{\text{I}} = s^{\text{I}}) \left\{ \prod_{t \in s^{\text{II}*}_p} \Pr(S^{\text{III}}_{pt} = s^{\text{III}}_{pt} \mid S^{\text{I}} = s^{\text{I}}, S^{\text{II}*}_p = s^{\text{II}*}_p) \right\} \right] \\
&\qquad = \Pr(S^{\text{I}} = s^{\text{I}}) \left[\prod_{p \in s^{\text{I}}} \Pr(S^{\text{II}}_p = s^{\text{II}}_p \mid S^{\text{I}} = s^{\text{I}}) \left\{ \prod_{t \in s^{\text{II}*}_p} \phi_{pt\cdot} \right\} \left\{ \prod_{t \in s^{\text{II}}_p \setminus s^{\text{II}*}_p} (1 - \phi_{pt\cdot}) \right\} \right. \\
&\qquad\qquad \left. \left\{ \prod_{t \in s^{\text{II}*}_p} \Pr(S^{\text{III}}_{pt} = s^{\text{III}}_{pt} \mid S^{\text{I}} = s^{\text{I}}, S^{\text{II}*}_p = s^{\text{II}*}_p) \right\} \right]
\end{aligned}$$

$$= \Pr(S^{\mathrm{I}} = s^{\mathrm{I}}) \left[\prod_{p \in s^{\mathrm{I}}} \Pr(S_p^{\mathrm{II}} = s_p^{\mathrm{II}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}) \left\{ \prod_{t \in s_p^{\mathrm{II}*}} \phi_{pt\cdot} \right\} \left\{ \prod_{t \in s_p^{\mathrm{II}} \setminus s_p^{\mathrm{II}*}} (1 - \phi_{pt\cdot}) \right\} \right.$$
$$\left. \left\{ \prod_{t \in s_p^{\mathrm{II}*}} \frac{\Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}})}{\phi_{pt\cdot}} \right\} \right]$$
$$= \Pr(S^{\mathrm{I}} = s^{\mathrm{I}}) \left[\prod_{p \in s^{\mathrm{I}}} \Pr(S_p^{\mathrm{II}} = s_p^{\mathrm{II}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}) \left\{ \prod_{t \in s_p^{\mathrm{II}} \setminus s_p^{\mathrm{II}*}} (1 - \phi_{pt\cdot}) \right\} \right.$$
$$\left. \left\{ \prod_{t \in s_p^{\mathrm{II}*}} \Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}}) \right\} \right]$$
$$= \Pr(S^{\mathrm{I}} = s^{\mathrm{I}}) \left[\prod_{p \in s^{\mathrm{I}}} \Pr(S_p^{\mathrm{II}} = s_p^{\mathrm{II}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}) \left\{ \prod_{t \in s_p^{\mathrm{II}}} \Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}}) \right\} \right],$$

since for $t \in s_p^{\mathrm{II}*}$,

$$\begin{aligned} \Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}*} = s_p^{\mathrm{II}*}) &= \Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}}, Z_{pt} = 1) \\ &= \frac{\Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \cap Z_{pt} = 1 \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}})}{\Pr(Z_{pt} = 1 \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}})} \\ &= \frac{\Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}})}{\phi_{pt\cdot}}, \end{aligned}$$

and for $t \notin s_p^{\mathrm{II}*}$,

$$\Pr(S_{pt}^{\mathrm{III}} = s_{pt}^{\mathrm{III}} \mid S^{\mathrm{I}} = s^{\mathrm{I}}, S_p^{\mathrm{II}} = s_p^{\mathrm{II}}) = (1 - \phi_{pt\cdot}).$$

Thus, the original design is equivalent to the modified design.

C.6 Derivation of Variance Estimators

C.6.1 Variance Estimator for the IPW Approach with SRS

$$
\begin{aligned}
\hat{\mathrm{V}}_{p\cdot\cdot}^{\mathrm{IPW}} &= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\mathrm{II}}}\sum_{u\in S_p^{\mathrm{II}}}\frac{(\pi_{tu|p}^{\mathrm{II}}-\pi_{t|p}^{\mathrm{II}}\pi_{u|p}^{\mathrm{II}})}{\pi_{tu|p}^{\mathrm{II}}}\frac{\hat{\overline{Y}}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}}}\frac{\hat{\overline{Y}}_{pu\cdot}}{\pi_{u|p}^{\mathrm{II}}}+\sum_{t\in S_p^{\mathrm{II}}}\frac{\hat{\mathrm{V}}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}}}\right\} \\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\mathrm{II}}}\frac{(1-\pi_{t|p}^{\mathrm{II}})}{(\pi_{t|p}^{\mathrm{II}})^2}(\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}})^2+\sum_{t\in S_p^{\mathrm{II}}}\sum_{u\neq t}\frac{\pi_{tu|p}^{\mathrm{II}}-\pi_{t|p}^{\mathrm{II}}\pi_{u|p}^{\mathrm{II}}}{\pi_{tu|p}^{\mathrm{II}}}\frac{\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}}}{\pi_{t|p}^{\mathrm{II}}}\frac{\hat{\overline{Y}}_{pu\cdot}^{\mathrm{IPW}}}{\pi_{u|p}^{\mathrm{II}}}+\sum_{t\in S_p^{\mathrm{II}}}\frac{\hat{\mathrm{V}}_{pt\cdot}}{\pi_{t|p}^{\mathrm{II}}}\right\} \\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\mathrm{II}}}\left(\frac{D}{d_p}\right)^2\left(1-\frac{d_p}{D}\right)(\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}})^2+\sum_{t\in S_p^{\mathrm{II}}}\sum_{u\neq t}\frac{\frac{d_p}{D}\left(\frac{d_p-1}{D-1}\right)-\left(\frac{d_p}{D}\right)^2}{\frac{d_p}{D}\left(\frac{d_p-1}{D-1}\right)}\left(\frac{D}{d_p}\right)^2\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}}\hat{\overline{Y}}_{pu\cdot}^{\mathrm{IPW}}+\sum_{t\in S_p^{\mathrm{II}}}\frac{D}{d_p}\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{IPW}}\right\} \\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\mathrm{II}}}\frac{D(D-d_p)}{d_p^2}(\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}})^2+\sum_{t\in S_p^{\mathrm{II}}}\sum_{u\neq t}\frac{D(d_p-D)}{d_p^2(d_p-1)}\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}}\hat{\overline{Y}}_{pu\cdot}^{\mathrm{IPW}}+\sum_{t\in S_p^{\mathrm{II}}}\frac{D}{d_p}\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{IPW}}\right\} \\
&= \frac{1}{D^2}\left[\sum_{t\in S_p^{\mathrm{II}}}\left\{\frac{D(D-d_p)}{d_p^2}-\frac{D(d_p-D)}{d_p^2(d_p-1)}\right\}(\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}})^2+\sum_{t\in S_p^{\mathrm{II}}}\sum_{u\in S_p^{\mathrm{II}}}\frac{D(d_p-D)}{d_p^2(d_p-1)}\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}}\hat{\overline{Y}}_{pu\cdot}^{\mathrm{IPW}}+\sum_{t\in S_p^{\mathrm{II}}}\frac{D}{d_p}\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{IPW}}\right] \\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\mathrm{II}}}\frac{D(D-d_p)}{d_p(d_p-1)}(\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}})^2+\frac{D(d_p-D)}{d_p^2(d_p-1)}\sum_{t\in S_p^{\mathrm{II}}}\sum_{u\in S_p^{\mathrm{II}}}\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}}\hat{\overline{Y}}_{pu\cdot}^{\mathrm{IPW}}\sum_{t\in S_p^{\mathrm{II}}}\frac{D}{d_p}\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{IPW}}\right\} \\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\mathrm{II}}}\frac{(D-d_p)D}{d_p(d_p-1)}(\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}})^2+\frac{D(d_p-D)}{d_p^2(d_p-1)}\left(\sum_{t\in S_p^{\mathrm{II}}}\hat{\overline{Y}}_{pt\cdot}^{\mathrm{IPW}}\right)^2+\sum_{t\in S_p^{\mathrm{II}}}\frac{D}{d_p}\hat{\mathrm{V}}_{pt\cdot}^{\mathrm{IPW}}\right\}.
\end{aligned}
$$

C.6.2 Variance Estimator for the Hájek Approach with SRS

$$
\begin{aligned}
\hat{\mathrm{V}}_{p\cdot\cdot}^{\text{Hájek}} &= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\text{II}*}}\sum_{u\in S_p^{\text{II}*}}\frac{(\pi_{tu|p}^{\text{II}\,*}-\pi_{t|p}^{\text{II}\,*}\pi_{u|p}^{\text{II}\,*})}{\pi_{tu|p}^{\text{II}\,*}}\frac{\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}}{\pi_{t|p}^{\text{II}\,*}}\frac{\hat{\overline{Y}}_{pu\cdot}^{\text{Hájek}}}{\pi_{u|p}^{\text{II}\,*}}+\sum_{t\in S_p^{\text{II}}}\frac{\hat{\mathrm{V}}_{pt\cdot}^{\text{Hájek}}}{\pi_{t|p}^{\text{II}\,*}}\right\}\\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\text{II}*}}\frac{1-\pi_{t|p}^{\text{II}\,*}}{(\pi_{t|p}^{\text{II}\,*})^2}(\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}})^2+\sum_{t\in S_p^{\text{II}*}}\sum_{u\neq t}\frac{\pi_{tu|p}^{\text{II}\,*}-\pi_{t|p}^{\text{II}\,*}\pi_{u|p}^{\text{II}\,*}}{\pi_{tu|p}^{\text{II}\,*}}\frac{\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}}{\pi_{t|p}^{\text{II}\,*}}\frac{\hat{\overline{Y}}_{pu\cdot}^{\text{Hájek}}}{\pi_{u|p}^{\text{II}\,*}}+\sum_{t\in S_p^{\text{II}}}\frac{\hat{\mathrm{V}}_{pt\cdot}^{\text{Hájek}}}{\pi_{t|p}^{\text{II}\,*}}\right\}\\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\text{II}*}}\frac{1-\pi_{t|p}^{\text{II}}\phi_{pt\cdot}}{(\pi_{t|p}^{\text{II}}\phi_{pt\cdot})^2}(\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}})^2+\sum_{t\in S_p^{\text{II}*}}\sum_{u\neq t}\frac{(\pi_{t|p}^{\text{II}}\phi_{pt\cdot}\phi_{pu\cdot}-\pi_{t|p}^{\text{II}}\phi_{pt\cdot}\pi_{u|p}^{\text{II}}\phi_{pu\cdot})}{\pi_{tu|p}^{\text{II}}\phi_{pt\cdot}\phi_{pu\cdot}}\frac{\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}}{\pi_{t|p}^{\text{II}}\phi_{pt\cdot}}\frac{\hat{\overline{Y}}_{pu\cdot}^{\text{Hájek}}}{\pi_{u|p}^{\text{II}}\phi_{pu\cdot}}\right.\\
&\qquad\left.+\sum_{t\in S_p^{\text{II}*}}\frac{\hat{\mathrm{V}}_{pt\cdot}^{\text{Hájek}}}{\pi_{t|p}^{\text{II}}\phi_{pt\cdot}}\right\}\\
&= \frac{1}{D^2}\left\{\sum_{t\in S_p^{\text{II}*}}\frac{\left(1-\frac{d_p}{D}\phi_{pt\cdot}\right)}{\left(\frac{d_p}{D}\phi_{pt\cdot}\right)^2}(\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}})^2+\sum_{t\in S_p^{\text{II}*}}\sum_{u\neq t}\frac{\frac{d_p}{D}\frac{(d_p-1)}{(D-1)}-\left(\frac{d_p}{D}\right)^2}{\left(\frac{d_p}{D}\right)^3\frac{(d_p-1)}{(D-1)}}\frac{\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}}{\phi_{pt\cdot}}\frac{\hat{\overline{Y}}_{pu\cdot}^{\text{Hájek}}}{\phi_{pu\cdot}}+\sum_{t\in S_p^{\text{II}*}}\frac{D}{d_p\phi_{pt\cdot}}\right\}\\
&= \frac{1}{D^2}\left[\sum_{t\in S_p^{\text{II}*}}\frac{D\{D-1-\hat{\phi}_{pt\cdot}(d_p-1)\}}{d_p(d_p-1)}\left(\frac{\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}}{\phi_{pt\cdot}}\right)^2+\frac{D(d_p-D)}{d_p^2(d_p-1)}\left(\sum_{t\in S_p^{\text{II}*}}\frac{\hat{\overline{Y}}_{pt\cdot}^{\text{Hájek}}}{\phi_{pt\cdot}}\right)^2+\sum_{t\in S_p^{\text{II}*}}\frac{D}{d_p\phi_{pt\cdot}}\hat{\mathrm{V}}_{pt\cdot}^{\text{Hájek}}\right]
\end{aligned}
$$

C.7 Derivation of Total Variance of $\tilde{\tau}$

We will use the following properties:

1. $\mathbf{E}(\hat{T}^b) = \tau$ for all $b = 1, \ldots, B$,
2. $\mathbf{E}_m(\hat{T}^b) = \hat{T}$ for all $b = 1, \ldots, B$,
3. $\hat{T}^b$ is independent of $\hat{T}^c$ with respect to measurement error for $b \neq c$,
4. $\mathbf{E}_m(\hat{T}^b\hat{T}^c) = \mathbf{E}_m((\hat{T}^b)^2)$ for $b \neq c$,

where properties 2 and 3 give property 4.

The variance of $\tilde{\tau}$ with respect to measurement error and the design is given by

$$
\begin{aligned}
\mathbf{Var}(\tilde{\tau}) &= \mathbf{Var}\left(\frac{\sum_{b=1}^{B}\hat{T}^b}{B}\right) \\
&= \frac{1}{B^2}\left(\sum_{b=1}^{B}\mathbf{Var}(\hat{T}^b) + \sum_{b=1}^{B}\sum_{c\neq b}\mathbf{Cov}(\hat{T}^b, \hat{T}^c)\right) \\
&= \frac{1}{B^2}\left(\sum_{b=1}^{B}\mathbf{Var}(\hat{T}^b) + \sum_{b=1}^{B}\sum_{c\neq b}\mathbf{E}(\hat{T}^b\hat{T}^c) - \mathbf{E}(\hat{T}^b)\mathbf{E}(\hat{T}^c)\right) \\
&= \frac{1}{B^2}\left(\sum_{b=1}^{B}\mathbf{Var}(\hat{T}^b) + \sum_{b=1}^{B}\sum_{c\neq b}\mathbf{E}_{ptq}\mathbf{E}_m(\hat{T}^b\hat{T}^c) - \mathbf{E}(\hat{T}^b)^2\right) \quad \text{using property 1,} \\
&= \frac{1}{B^2}\left(\sum_{b=1}^{B}\mathbf{Var}(\hat{T}^b) + \sum_{b=1}^{B}\sum_{c\neq b}\mathbf{E}[(\hat{T}^b)^2] - \mathbf{E}(\hat{T}^b)^2\right) \quad \text{using property 4,} \\
&= \frac{1}{B^2}\left\{B\mathbf{Var}(\hat{T}^b) + B(B-1)\mathbf{Var}(\hat{T}^b)\right\} \\
&= \mathbf{Var}(\hat{T}^b) \\
&= \mathbf{Var}_m\mathbf{E}_{ptq}(\hat{T}^b) + \mathbf{E}_m\mathbf{Var}_{ptq}(\hat{T}^b) \\
&= \mathbf{Var}_m(T^b) + \mathbf{E}_m(\mathrm{V}^b).
\end{aligned}
$$

We can therefore estimate $\mathbf{Var}(\tilde{\tau})$ with

$$\frac{\sum_{b=1}^{B}(\hat{T}^b - \tilde{\tau})^2}{B-1} + \frac{\sum_{b=1}^{B}\hat{\mathrm{V}}^{3\text{ stage}}{}_b}{B}$$

or using the fact that the estimator $\hat{\mathrm{V}}^{3\text{ stage}}$ is the sum of variance contributions from each stage,

$$\frac{\sum_{b=1}^{B}(\hat{T}^b - \tilde{\tau})^2}{B-1} + \frac{\sum_{b=1}^{B}\hat{\mathrm{V}}^{\mathrm{I}}{}_b}{B} + \frac{\sum_{b=1}^{B}\hat{\mathrm{V}}^{\mathrm{II}}{}_b}{B} + \frac{\sum_{b=1}^{B}\hat{\mathrm{V}}^{\mathrm{III}}{}_b}{B}.$$

C.8 Additional Data Analysis Results

C.8.1 Results for Multi-Stage Hájek with Stage II Uncertainty and Measurement Error

Table C.2: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\tilde{V}^{I}$	$\tilde{V}^{II}$	$\tilde{V}^{III}$	$\tilde{V}^{m}$	$\tilde{V}^{Total}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.001	0.000	0.000	0.001
Water Hub Battery	0.097	0.003	0.000	0.000	0.002	0.006
MS	0.199	0.004	0.003	0.000	0.002	0.008
Gas SWB	0.227	0.012	0.009	0.000	0.006	0.027
CO SWB	0.251	0.001	0.011	0.000	0.003	0.016
Mixed OG Battery	0.389	0.000	0.014	0.000	0.005	0.019
Disposal/storage	0.451	0.056	0.026	0.000	0.009	0.092
CTF and GGS	0.837	0.110	0.056	0.000	0.038	0.204
CO MWB	0.886	0.003	0.106	0.000	0.018	0.127
GP Acid Gas Low	4.181	0.297	0.199	0.000	0.080	0.576
Gas MWB (group)	5.709	1.785	0.845	0.000	0.663	3.293
Other GP and LNGP	9.138	0.777	2.027	0.000	0.521	3.326
Wells	13.402	10.355	0.868	0.007	1.871	13.101
Gas MWB (effluent)	14.761	1.491	1.651	0.001	0.446	3.589
GP Sweet	19.438	5.150	4.128	0.000	2.685	11.963
Compressor Stations	46.888	83.104	25.311	0.002	10.429	118.846
Population	116.881	103.149	35.255	0.011	16.926	155.342

C.8.2 Results for Multi-stage IPW with No Stage II Uncertainty and Measurement Error

Table C.3: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\tilde{V}^{I}$	$\tilde{V}^{II}$	$\tilde{V}^{III}$	$\tilde{V}^{m}$	$\tilde{V}^{Total}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.000	0.000	0.000	0.001
Water Hub Battery	0.097	0.000	0.000	0.004	0.002	0.007
MS	0.162	0.004	0.000	0.000	0.001	0.005
CO SWB	0.194	0.001	0.001	0.003	0.002	0.006
Gas SWB	0.227	0.014	0.000	0.000	0.006	0.021
Disposal/storage	0.385	0.047	0.000	0.000	0.007	0.054
Mixed OG Battery	0.391	0.000	0.000	0.000	0.005	0.006
CO MWB	0.764	0.016	0.000	0.001	0.016	0.033
CTF and GGS	0.801	0.117	0.003	0.007	0.028	0.155
GP Acid Gas Low	3.805	0.298	0.000	0.003	0.072	0.373
Gas MWB (group)	4.754	1.266	0.001	0.001	0.432	1.700
Other GP and LNGP	8.473	1.459	0.003	0.005	0.472	1.939
Wells	12.364	9.736	0.000	0.151	1.704	11.591
Gas MWB (effluent)	14.227	1.709	0.001	0.010	0.426	2.146
GP Sweet	17.240	4.571	0.043	0.201	2.476	7.290
Compressor Stations	44.937	99.380	0.069	0.015	9.816	109.280
Population	108.847	118.619	0.122	0.401	15.561	134.703

C.8.3 Results for Multi-stage Hájek with No Stage II Uncertainty and Measurement Error

Table C.4: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\tilde{V}^{I}$	$\tilde{V}^{II}$	$\tilde{V}^{III}$	$\tilde{V}^{m}$	$\tilde{V}^{Total}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.000	0.000	0.000	0.001
Water Hub Battery	0.097	0.003	0.000	0.000	0.002	0.006
MS	0.199	0.005	0.001	0.000	0.002	0.007
Gas SWB	0.227	0.014	0.001	0.000	0.006	0.021
CO SWB	0.251	0.002	0.003	0.000	0.003	0.009
Mixed OG Battery	0.389	0.000	0.000	0.000	0.005	0.006
Disposal/storage	0.451	0.066	0.001	0.000	0.009	0.076
CTF and GGS	0.837	0.120	0.027	0.000	0.038	0.185
CO MWB	0.886	0.019	0.001	0.000	0.018	0.038
GP Acid Gas Low	4.181	0.331	0.003	0.000	0.080	0.415
Gas MWB (group)	5.709	2.052	0.002	0.000	0.663	2.717
Other GP and LNGP	9.138	1.806	0.008	0.000	0.521	2.336
Wells	13.402	11.078	0.064	0.007	1.871	13.020
Gas MWB (effluent)	14.761	1.807	0.010	0.001	0.446	2.264
GP Sweet	19.438	5.771	0.244	0.000	2.685	8.701
Compressor Stations	46.888	103.745	0.226	0.002	10.429	114.402
Population	116.881	126.820	0.592	0.011	16.926	144.350

C.8.4 Results for Multi-stage IPW with Stage II Uncertainty and No Measurement Error

Table C.5: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\hat{V}^{I}$	$\hat{V}^{II}$	$\hat{V}^{III}$	$\tilde{V}^{m}$	$\hat{V}^{3\text{ stage}}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.001	0.000	0.000	0.001
Water Hub Battery	0.099	0.000	0.000	0.004	0.000	0.004
MS	0.162	0.003	0.000	0.000	0.000	0.004
CO SWB	0.194	0.001	0.001	0.002	0.000	0.004
Gas SWB	0.226	0.012	0.003	0.000	0.000	0.015
Disposal/storage	0.385	0.038	0.016	0.000	0.000	0.054
Mixed OG Battery	0.388	0.000	0.008	0.000	0.000	0.008
CO MWB	0.765	0.002	0.085	0.001	0.000	0.087
CTF and GGS	0.810	0.109	0.008	0.004	0.000	0.121
GP Acid Gas Low	3.807	0.268	0.113	0.002	0.000	0.383
Gas MWB (group)	4.755	1.027	0.353	0.001	0.000	1.380
Other GP and LNGP	8.491	0.401	1.543	0.004	0.000	1.947
Wells	12.360	7.845	0.437	0.116	0.000	8.398
Gas MWB (effluent)	14.230	1.400	1.178	0.008	0.000	2.586
GP Sweet	17.212	4.019	1.028	0.154	0.000	5.201
Compressor Stations	44.897	79.927	13.517	0.011	0.000	93.456
Population	108.807	95.051	18.292	0.306	0.000	113.649

C.8.5 Results for Multi-stage Hájek with Stage II Uncertainty and No Measurement Error

Table C.6: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\hat{V}^{I}$	$\hat{V}^{II}$	$\hat{V}^{III}$	$\tilde{V}^{m}$	$\hat{V}^{3\text{ stage}}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.001	0.000	0.000	0.001
Water Hub Battery	0.099	0.002	0.000	0.000	0.000	0.003
MS	0.199	0.004	0.001	0.000	0.000	0.005
Gas SWB	0.226	0.012	0.003	0.000	0.000	0.015
CO SWB	0.250	0.001	0.008	0.000	0.000	0.009
Mixed OG Battery	0.387	0.000	0.008	0.000	0.000	0.008
Disposal/storage	0.451	0.056	0.017	0.000	0.000	0.074
CTF and GGS	0.849	0.111	0.021	0.000	0.000	0.132
CO MWB	0.887	0.003	0.088	0.000	0.000	0.091
GP Acid Gas Low	4.183	0.299	0.119	0.000	0.000	0.418
Gas MWB (group)	5.710	1.795	0.186	0.000	0.000	1.981
Other GP and LNGP	9.157	0.756	1.483	0.000	0.000	2.239
Wells	13.398	8.923	0.577	0.001	0.000	9.502
Gas MWB (effluent)	14.762	1.491	1.198	0.000	0.000	2.689
GP Sweet	19.416	5.158	1.434	0.000	0.000	6.592
Compressor Stations	46.838	82.911	14.705	0.001	0.000	97.617
Population	116.839	101.523	19.848	0.003	0.000	121.375

C.8.6 Results for Multi-stage IPW with No Stage II Uncertainty and No Measurement Error

Table C.7: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\hat{V}^{I}$	$\hat{V}^{II}$	$\hat{V}^{III}$	$\tilde{V}^{m}$	$\hat{V}^{3\text{ stage}}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.000	0.000	0.000	0.000
Water Hub Battery	0.099	0.000	0.000	0.004	0.000	0.004
MS	0.162	0.003	0.000	0.000	0.000	0.004
CO SWB	0.194	0.001	0.000	0.002	0.000	0.004
Gas SWB	0.226	0.013	0.000	0.000	0.000	0.013
Disposal/storage	0.385	0.044	0.000	0.000	0.000	0.044
Mixed OG Battery	0.388	0.000	0.000	0.000	0.000	0.000
CO MWB	0.765	0.014	0.000	0.001	0.000	0.015
CTF and GGS	0.810	0.111	0.002	0.004	0.000	0.116
GP Acid Gas Low	3.807	0.287	0.000	0.002	0.000	0.289
Gas MWB (group)	4.755	1.138	0.000	0.001	0.000	1.140
Other GP and LNGP	8.491	1.207	0.003	0.004	0.000	1.213
Wells	12.360	8.230	0.000	0.116	0.000	8.346
Gas MWB (effluent)	14.230	1.627	0.000	0.008	0.000	1.635
GP Sweet	17.212	4.179	0.033	0.154	0.000	4.365
Compressor Stations	44.897	91.006	0.053	0.011	0.000	91.070
Population	108.807	107.861	0.092	0.306	0.000	108.259

C.8.7 Results for Multi-stage Hájek with No Stage II Uncertainty and No Measurement Error

Table C.8: Estimates of the total methane emissions for 2021, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\hat{V}^{I}$	$\hat{V}^{II}$	$\hat{V}^{III}$	$\tilde{V}^{m}$	$\hat{V}^{3\text{ stage}}$
		([kt/y]^2)				([kt/y]^2)
TF and NGL	0.026	0.000	0.000	0.000	0.000	0.000
Water Hub Battery	0.099	0.003	0.000	0.000	0.000	0.003
MS	0.199	0.004	0.001	0.000	0.000	0.005
Gas SWB	0.226	0.013	0.000	0.000	0.000	0.013
CO SWB	0.250	0.002	0.002	0.000	0.000	0.004
Mixed OG Battery	0.387	0.000	0.000	0.000	0.000	0.000
Disposal/storage	0.451	0.063	0.000	0.000	0.000	0.063
CTF and GGS	0.849	0.114	0.014	0.000	0.000	0.127
CO MWB	0.887	0.016	0.001	0.000	0.000	0.017
GP Acid Gas Low	4.183	0.319	0.003	0.000	0.000	0.322
Gas MWB (group)	5.710	1.854	0.002	0.000	0.000	1.855
Other GP and LNGP	9.157	1.529	0.006	0.000	0.000	1.536
Wells	13.398	9.396	0.051	0.001	0.000	9.449
Gas MWB (effluent)	14.762	1.721	0.008	0.000	0.000	1.729
GP Sweet	19.416	5.358	0.187	0.000	0.000	5.545
Compressor Stations	46.838	94.870	0.170	0.001	0.000	95.042
Population	116.839	115.261	0.446	0.003	0.000	115.710

C.8.8 Results for Multi-Stage IPW with Measurement and Stage II Uncertainty for Only Southern Facilities

Table C.9: Estimates of the total methane emissions for 2021 from facilities in BC south of the 58 degree north latitude line, variance estimates by uncertainty source, and total variance estimates by stratum

Stratum	Total Estimate (kt/y)	$\tilde{V}^{I}$	$\tilde{V}^{II}$	$\tilde{V}^{III}$	$\tilde{V}^{m}$	$\tilde{V}^{Total}$
		([kt/y]^2)				([kt/y]^2)
CO SWB	0.161	0.000	0.001	0.002	0.001	0.005
Gas SWB	0.171	0.000	0.005	0.000	0.003	0.008
Gas MWB (group)	3.249	0.000	0.360	0.001	0.202	0.563
Gas MWB (effluent)	11.484	0.000	1.050	0.008	0.277	1.335
Mixed OG Battery	0.391	0.000	0.014	0.000	0.005	0.019
Water Hub Battery	0.063	0.000	0.000	0.003	0.001	0.004
GP Sweet	14.482	0.000	2.429	0.169	1.747	4.345
GP Acid Gas Low	3.143	0.000	0.125	0.002	0.049	0.177
Compressor Stations	7.961	0.000	0.735	0.003	0.308	1.046
CO MWB	0.650	0.000	0.073	0.001	0.011	0.085
CTF and GGS	0.539	0.000	0.013	0.004	0.013	0.030
Disposal/storage	0.240	0.000	0.009	0.000	0.003	0.012
MS	0.091	0.000	0.000	0.000	0.000	0.001
Other GP and LNGP	4.035	0.000	0.465	0.002	0.107	0.574
TF and NGL	0.014	0.000	0.000	0.000	0.000	0.000
Wells	1.244	0.000	0.001	0.015	0.017	0.033
Population	47.916	0.000	5.281	0.211	2.743	8.235

Appendix D

Chapter 5 Appendix

D.1 Prior Distributions

The parameters of the likelihood to be estimated are α_0, α_1, σ, along with optional parameters α_2, γ, β_1, and/or η. Note that β_0 is a function of other parameters to ensure that the piece-wise function is continuous, so a prior is not specified for this parameter. Let the vector of statistical model parameters be represented by $\boldsymbol{\theta}$. All parameters are assumed to be independent, that is, the prior distribution can be written as

$$p(\boldsymbol{\theta}) = p(\alpha_0, \alpha_1, \sigma, \alpha_2, \gamma, \beta_1, \eta) = p(\alpha_0)p(\alpha_1)p(\sigma)p(\alpha_2)p(\gamma)p(\beta_1)p(\eta), \tag{D.1}$$

so we can specify individual prior distributions for each parameter.

Since α_0 is the median measurement when the true emission rate is zero, it should be small and non-negative. It is parameterized by a gamma distribution, where the shape and rate parameters can be chosen so the mean of the distribution is similar to the mean false positives observed in the data. For example, for the data reported by QOGI Operator A, the average false positive is 0.27 kg/h so the prior is a gamma distribution with shape parameter = 0.5 and rate parameter = 2, which has a mean of 0.25 and variance of 0.125.

In a simple linear model, α_1 is the slope of Y_i so for every one unit increase in Y_i, the median measurement increases by α_1 units. A perfect technology would have $\alpha_1 = 1$. Thus we use a prior for α_1 which has a median of 1 and is non-negative. Further, we seek a distribution with the property that for any constant $k > 1$, the probability that $\alpha_1 > k$ should be the same as the probability that $\alpha_1 < 1/k$, or in other words, the probability that

the technology over-estimates by a factor of k is the same as the probability that it under-estimates by a factor of $1/k$. This property is desirable for the prior because information about under- or over-estimation should only come from the data. Thus we use a standard log-normal distribution (shape parameter equal to one, location parameter equal to zero, and scale parameter equal to zero (NIST, 2023)) because it fulfills this requirement. For example, if α_1 follows the standard log-normal distribution the probability that $\Pr(\alpha_1 < 1/2) = \Pr(\alpha_1 > 2) = 0.244$.

Coefficient α_2 is associated with Y_i^2 when $Y_i \leq \gamma$. Like α_0 and α_1, it must be non-negative. Therefore, we use a half-normal distribution with scale parameter equal to one. This gives a prior variance of 0.60 kg/h^{-2}. In units of kg/h, this is equivalent to a variance of 1.29.

Parameter β_1 (unitless, non-negative) represents the change in α_1 when $Y_i > \gamma$. As with α_1, we use a standard log-normal distribution as a prior for β_1.

The threshold parameter γ represents the value of Y_i for which the relationship with M_i changes from quadratic to linear. We must constrain the parameter γ to the range of Y_i in the CR data, otherwise it cannot be estimated. We use a uniform prior distribution on $(0, Y_{max})$ where Y_{max} is the largest value of Y_i observed in the CR data for a given technology.

Parameter σ^2 represents either the variance of measurements on the log scale in a constant variance model, or the variance of measurements when $Y_i = 0$ on the log scale. We use a vague non-negative prior of a half-normal with variance parameter set to 100 on σ, as suggested in (Gelman, 2006).

Finally, if the more complicated variance model is used, a prior must be chosen for η. Little external information is known about η except it must be non-negative. We use a half-normal with variance set to 100.

The sensitivity of results to prior specification was checked for all models. Results were obtained for the stated priors. Next, the model was refit with priors where the range and/or variance was changed for some parameters. The posterior distribution of each parameter was then compared between the two models. The resulting 90% prediction bands were also compared between the models. Unless otherwise stated in the results section, the model results were insensitive to the prior specification.

D.2 Detailed Algorithm

The weighted bootstrap algorithm for obtaining a sample from the posterior of interest outlined in Section 5.2.2 has the following steps:

1. Drawing a sample from the prior $p(Y^{new})$.
2. Computing normalized weights proportional to $p(M^{new} \mid Y^{new}, \underline{M})$ for each element of the sample from Step 1
3. Resampling from the Step 1 sample using the normalized weights from Step 2.

The sample can then be used to compute quantile-based or highest posterior density (HPD) CIs for the true emission rate, given a measurement or set of measurements. This algorithm is implemented in R functions included in the repository `https://github.com/augustinewigle/methaneUQ`. In the following sections, we describe the details of each step of the algorithm.

D.2.1 Step 1

A prior distribution should be chosen based on the best available data. For example, a continuous distribution such as the log-normal could be fit to data from a large sample of survey data from the region of interest. Then a sample from the fitted log-normal distribution can be drawn. Let L be the size of the sample drawn from the prior distribution. This step of the algorithm is described in pseudocode in Algorithm 2.

Algorithm 2 Pseudocode for Step 1 of the weighted bootstrap algorithm

```
Define L
Define prior_distribution
Define sample_from_prior ← vector of length L
for l in 1:L do
    sample_from_prior[l] ← sample drawn from prior_distribution
end for
```

D.2.2 Step 2

We need computational techniques to determine $p(M^{new} \mid Y^{new}, \underline{M})$ in order to calculate weights. We can see that

$$\begin{aligned} p(M^{new} \mid Y^{new}, \underline{M}) &= \int p(M^{new} \mid Y^{new}, \boldsymbol{\theta}, \underline{M}) p(\boldsymbol{\theta} \mid Y^{new}, \underline{M}) d\boldsymbol{\theta} \\ &= \int p(M^{new} \mid Y^{new}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \underline{M}) d\boldsymbol{\theta}. \end{aligned} \tag{D.2}$$

The distribution $p(M^{new} \mid Y^{new}, \boldsymbol{\theta})$ is the likelihood given by the uncertainty model from the previous section where $Y_i = Y^{new}$, and $p(\boldsymbol{\theta} \mid \underline{M})$ is the posterior distribution of $\boldsymbol{\theta}$ given the controlled release data. We can approximate $p(M^{new} \mid Y^{new}, \underline{M})$ by approximating the integral on the right side of Equation (D.2). Let P be the number of parameters in the vector $\boldsymbol{\theta}$ for the given technology.

The integral in Equation (D.2) can be approximated via Monte Carlo simulation for a single value of Y^{new} as follows:

1. Sample $\boldsymbol{\theta}^j$ $j = 1, \ldots, J$ times from $p(\boldsymbol{\theta} \mid \underline{M})$
2. For $j = 1, \ldots, J$, calculate $p(M^{new} \mid Y^{new}, \boldsymbol{\theta}^j)$
3. The integral in Equation (D.2) $\approx \frac{\sum_{j=1}^{J} p(M^{new} \mid Y^{new}, \boldsymbol{\theta}^j)}{J}$.

In Algorithm 3 we describe how this can be implemented to obtain the normalized weights needed in Step 3 of the weighted bootstrap algorithm.

Algorithm 3 Pseudocode for getting normalized weights at Step 2 when there is a single new measurement.

Define `mnew` ← value(s) of new measurement(s)
Define `theta` ← matrix with `P` columns containing MCMC draws from $p(\boldsymbol{\theta} \mid \underline{M})$
Define `mnew_likelihood` ← a function of $M^{new}, Y^{new}, \boldsymbol{\theta}$
Define `J` ← Number of samples to take from the posterior of $\boldsymbol{\theta}$
Define `theta_sample` ← matrix with `J` rows and `P` columns
for `l in 1:L` **do**
  `current_y` ← `sample_from_prior[l]`
  Define `integrand` ← vector of length `J`
  for `j in 1:J` **do**
    `theta_sample[j,]` ← row sampled from `theta`
    `integrand[j]` ← `mnew_likelihood(`M^{new} `= mnew,` Y^{new} `= current_y,` $\boldsymbol{\theta}$ `= theta_sample[j,])`
  end for
  `unnormalized_weights[l]` ← `mean(integrand)`
end for
for `l in 1:L` **do**
  `normalized_weights[l]` ← `unnormalized_weights[l]/sum(unnormalized_weights)`
end for

The function `mnew_likelihood` is the mathematical coding of the likelihood function selected when fitting the measurement uncertainty model to the controlled release data. For example, for operator QOGI C, the selected likelihood is given by

$$\log(M_i) = \log(\psi_i) + \epsilon_i,$$

where

$$\epsilon_i \sim \mathrm{N}(0, (\sigma^{-2} + Y_i/\eta)^{-1})$$

and

$$\psi_i = \begin{cases} \alpha_0 + \alpha_1 Y_i + \alpha_2 Y_i^2 & Y_i \leq \gamma; \\ \alpha_0 + \beta_0 + (\alpha_1 + \beta_1) Y_i & Y_i > \gamma \end{cases}. \tag{D.3}$$

In fact, it is more accurate to write $\psi_i(Y_i, \boldsymbol{\theta})$ since it depends on the true value of Y_i and the values of the statistical model parameters. Then the function `mnew_likelihood` for a single new measurement is given by the normal density function with mean equal to $\log(\psi_i(Y_i, \boldsymbol{\theta}))$ and variance parameter $(\sigma^{-2} + Y_i/\eta)^{-1}$ evaluated at M^{new}.

The algorithm is also applicable to the case where we have repeated measurements of the same source, that is, where $\underline{M}^{new}$ is a vector, $\underline{M}^{new} = (M_1^{new}, \ldots, M_n^{new})$. Under the assumption that the uncertainty in each measurement is independent of previous measurements, conditional on Y^{new}, we have that $p(\underline{M}^{new} \mid Y^{new}, \boldsymbol{\theta}^j) = \prod_{i=1}^{n} p(M_i^{new} \mid Y^{new}, \boldsymbol{\theta}^j)$. That is, we simply take the product of the normal densities evaluated at each measurement value for the same values of Y^{new} and $\boldsymbol{\theta}^j$.

D.2.3 Step 3

At the final stage of the weighted bootstrap, we perform resampling from the sample drawn in Step 1. That is, we take a random sample from that drawn from the prior in Step 1 where the probability of sampling each element is proportional to its normalized weight calculated in Step 2. Suppose we want to have K samples from the distribution in Equation (5.4). Algorithm 4 describes the steps to achieve this.

Algorithm 4 Pseudocode for obtaining the final bootstrapped sample from the posterior distribution (Step 3).

```
Define K ← Number of bootstrap samples
Define final_sample ← vector of length K
for k in 1:K do
   final_sample[k] ← sample from sample_from_prior with weights
normalized_weights
end for
```