

General Bayesian Calibration Framework for Model Contamination and Measurement Error

Siquan Wang

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy  
under the Executive Committee  
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

© 2023

Siquan Wang

All Rights Reserved

## **Abstract**

General Bayesian Calibration Framework for Model Contamination and Measurement Error

Siquan Wang

Many applied statistical applications face the potential problem of model contamination and measurement error. The form and degree of contamination as well as the measurement error are usually unknown and sample-specific, which brings additional challenges for researchers. In this thesis, we have proposed several Bayesian inference models to address these issues, with the application to one type of special data for allergen concentration measurement, which is called serial dilution data and is self-calibrated. In our first chapter, we address the problem of model contamination by using a multilevel model to simultaneously flag problematic observations and estimate unknown concentrations in serial dilution data, a problem where the current approach can lead to noisy estimates and difficulty in estimating very low or high concentrations. In our second chapter, we propose the Bayesian joint contamination model for modeling multiple measurement units at the same time while adjusting for differences between experiments using the idea of global calibration, and it could account for uncertainty in both predictors and response variables in Bayesian regression. We are able to get efficacy gain by analyzing multiple experiments together while maintaining robustness with the use of hierarchical models. In our third chapter, we develop a Bayesian two-step inference model to account for measurement uncertainty propagation in regression analysis when the joint inference model is infeasible. We aim to increase model inference reliability while providing flexibility to users by not restricting the type of inference model used in the first step. For each of the proposed methods, We also demonstrate how to integrate multiple

model building blocks through the idea of Bayesian workflow. In extensive simulation studies, we show that our proposed methods outperform other commonly used approaches. For the data applications, we apply the proposed new methods to the New York City Neighborhood Asthma and Allergy Study (NYC NAAS) data to estimate indoor allergen concentrations more accurately as well as reveal the underlying associations between dust mite allergen concentrations and the exhaled nitric oxide (NO) measurement for asthmatic children. The methods and tools developed here have a wide range of applications and can be used to improve lab analyses, which are crucial for quantifying exposures to assess disease risk and evaluating interventions.

## Table of Contents

List of Figures . . . . .	viii
List of Tables . . . . .	x
Acknowledgments . . . . .	1
Chapter 1: Introduction and Overview . . . . .	1
Chapter 2: Bayesian Framework for Model Contamination with Application to Serial Dilution Assay . . . . .	6
2.1 Statistical analysis of contaminated data . . . . .	6
2.1.1 Background . . . . .	6
2.1.2 Contamination of immunoassays . . . . .	6
2.1.3 Motivating application . . . . .	9
2.1.4 Statistical modeling of assay data and contamination . . . . .	12
2.1.5 Our contributions . . . . .	13
2.2 Bayesian contamination model for dilution assays . . . . .	13
2.2.1 Notation and setup . . . . .	13
2.2.2 Bayesian contamination model . . . . .	14
2.2.3 Prior specification . . . . .	15
2.2.4 Alternative model specification . . . . .	16

2.2.5	Computation . . . . .	16
2.3	Bayesian workflow for contamination model . . . . .	17
2.4	Analysis of a dust mite allergen immunoassay plate . . . . .	19
2.4.1	Initial model construction . . . . .	19
2.4.2	Model fitting while addressing computation issues . . . . .	20
2.4.3	Model evaluation and specific aims for immunoassays data . . . . .	21
2.4.4	Standard data concentration recovery ratio . . . . .	23
2.4.5	Four-parameter logistic model versus five-parameter logistic model . . . . .	25
2.4.6	Improved procedure for traditional inference method . . . . .	27
2.4.7	Sensitivity analysis . . . . .	28
2.4.8	Model modification and extension . . . . .	29
2.5	Simulation study . . . . .	30
2.5.1	Evaluation of the proposed method under various contamination settings . . . . .	30
2.5.2	Extreme case . . . . .	33
2.6	Discussion . . . . .	35
Chapter 3:	Bayesian Joint Modeling of Exposure and Outcome with Uncertainty in Exposure Measure . . . . .	37
3.1	Introduction and background . . . . .	37
3.1.1	Local and global statistical calibration . . . . .	37
3.1.2	Background to the multiple Multiplex plates experiment setting . . . . .	38
3.1.3	Introduction to the NYC NAAS dataset . . . . .	39
3.1.4	Potential pitfalls of current methods for allergen estimation and regression-based association studies . . . . .	40

3.1.5	Real-life evidence for Bayesian joint modeling for multiple measurement units . . . . .	43
3.1.6	Our contributions . . . . .	44
3.2	Methods . . . . .	44
3.2.1	Problem setup . . . . .	44
3.2.2	Bayesian joint contamination model for multiple Multiplex plates in epidemiological association studies . . . . .	46
3.2.3	Prior specification . . . . .	49
3.2.4	Sensitivity analysis and computation . . . . .	50
3.3	Bayesian workflow for joint inference model . . . . .	51
3.4	Application studies . . . . .	53
3.4.1	Childhood asthma studies . . . . .	53
3.4.2	Summary statistics of NYCNAAS data and research question of interest . . . . .	55
3.4.3	Bayesian linear regression on asthmatic children . . . . .	57
3.4.4	Insights from this study . . . . .	64
3.5	Simulation studies . . . . .	65
3.5.1	Model comparison and evaluation in severe contamination setting . . . . .	65
3.5.2	Efficient multi-plate experiment design . . . . .	68
3.6	Discussion . . . . .	70
Chapter 4: Bayesian Two-step Model for Measurement Uncertainty Adjustment . . . . .		73
4.1	Introduction and background . . . . .	73
4.1.1	Insights from multiphase studies . . . . .	73
4.1.2	Challenges in epidemiologic association studies . . . . .	73

4.1.3	Statistical models in handling uncertainty from multi-step modeling . . . . .	74
4.1.4	Measurement uncertainty as a source of model contamination . . . . .	76
4.1.5	Our contributions . . . . .	77
4.2	Methods . . . . .	78
4.2.1	Notations and background . . . . .	78
4.2.2	Two-step Bayesian measurement error model . . . . .	78
4.2.3	Measurement uncertainty propagation . . . . .	81
4.2.4	Prior specification and Bayesian computation . . . . .	82
4.2.5	Connection between measurement uncertainty and missing data problem . . . . .	83
4.2.6	Robust model specification for measurement uncertainty . . . . .	84
4.2.7	Comparison with competing methods . . . . .	85
4.3	Bayesian workflow for two-step inference model . . . . .	86
4.4	Application studies . . . . .	88
4.5	Simulation studies . . . . .	91
4.5.1	Model comparison and evaluation in the setting without contamination . . . . .	92
4.5.2	Model comparison and evaluation in the setting with sample contamination . . . . .	96
4.6	Discussion . . . . .	99
	Conclusion or Epilogue . . . . .	109
	References . . . . .	116
	Appendix A: Appendices to Chapter 2 . . . . .	117
A.1	Simulation Study: Prior predictive check . . . . .	117

Appendix B: Appendices to Chapter 3 . . . . . 119

    B.1 Simulation Study: More efficient plate design . . . . . 119

Appendix C: Appendices to Chapter 4 . . . . . 121

    C.1 Simulation Study: Alternative robust model for measurement uncertainty . . . . . 121

## List of Figures

2.1	Rate response (mOD U/min) measured in an immunoassay for a range of dust mite allergen (Der p 1) concentrations, where samples are contaminated with 8 different concentrations of a known contaminant (household laundry detergent) and mOD U/min stands for millioptical density units per minute. . . . .	8
2.2	Examples of sample contamination from the New York City Neighborhood Asthma and Allergy Study. Curves show data from standards (calibration data) and three unknown samples on a single microtiter plate. . . . .	11
2.3	Demonstration of sequential model improvement under Bayesian workflow . . . . .	21
2.4	Posterior median and 95% probability interval of the mean function dose-response curves for the standards and each new sample estimated using our proposed model. The posterior median and corresponding 50% probability interval for the probability of contamination are listed at the bottom of each plot. . . . .	22
2.5	Concentration estimation of Der f 1 for all new samples in a single multiplex plate using the classical calibration and the new Bayesian method with 95% probability interval. Samples with * are estimated to have high posterior probability of contamination. . . . .	23
2.6	Comparison of the standard calibration data concentration recovery ratio between the classical method and the Bayesian contamination model. . . . .	26
3.1	Comparison of multiple independent fitting standard data calibration curves for each measurement unit. . . . .	43

3.2	Bayesian regression coefficients summary for NYCNAAS data, where the posterior median and 80% uncertainty interval are presented. . . . .	58
3.3	Comparison of indoor allergen concentration estimation by the classical calibration method and the Bayesian joint model. . . . .	60
3.4	Comparison of indoor allergen concentration estimation by the Bayesian joint model and the naive two-step model using local calibration. . . . .	61
3.5	Comparison of indoor allergen concentration estimation by the classical calibration method and the naive two-step model using local calibration. . . . .	62
3.6	Example of one specific unknown sample having larger classical concentration estimation compared to Bayesian joint concentration estimation. . . . .	62
4.1	Bayesian regression coefficients summary for NYCNAAS data, where the posterior median and 80% uncertainty interval are presented. . . . .	90
4.2	Comparison of indoor allergen concentration estimation by the naive two-step method and Bayesian two-step method. . . . .	91
4.3	Comparison of true concentration, the posterior median of concentration estimation in the first-step model ( $W$ ) and posterior draw of concentration in the Bayesian two-step model ( $X$ ) in simulation scenario 1 (with a line with slope 1 and intercept 0 as reference). . . . .	96
4.4	Comparison of true concentration, the posterior median of concentration estimation in the first-step model ( $W$ ) and posterior draw of concentration in the Bayesian two-step model ( $X$ ) in simulation scenario 2 (with a line with slope 1 and intercept 0 as reference). . . . .	99

A.1 Posterior median and 95% probability interval of the mean function dose-response curves for the standards and each new sample estimated using our proposed model using stronger exponential prior for unknown concentrations. The posterior median and corresponding 50% probability interval for the probability of contamination are listed at the bottom of each plot. . . . . 118

## List of Tables

2.1	Map for standards and new samples (dilutions) in a multiplex plate with 96 wells. . . . .	9
2.2	Standards and selected unknown samples from a multiplex plate with the concentrations of the new samples estimated using the classical calibration method. . . . .	10
2.3	Comparison between different methods in terms of root mean squared error (RMSE) and coverage probability (CP) of 95% probability interval for each unknown sample in two different contamination settings. Unk1-15 are uncontaminated samples and Unk16-20 are contaminated samples. . . . .	32
2.4	Root mean squared error (RMSE) and coverage probability (CP) of 95% probability interval of our proposed method stratified by the estimated contamination status in posterior draws for each unknown sample. Unk1-15 are uncontaminated samples and Unk16-20 are contaminated samples. Ratio: proportion of posterior draws being classified as contaminated samples. . . . .	33
2.5	Summary of root mean squared error (RMSE) across 15 uncontaminated samples (Unk1-15) and 5 contaminated samples (Unk16-20) for severe contamination case in the extreme case simulation scenario. Ratio: proportion of posterior draws being classified as contaminated samples. . . . .	34
3.1	Summary of patient characteristics in NYC NAAS study. . . . .	56
3.2	Raw serial dilution assay data for unknown sample 6 on 012711 plate one, OOR< means below the detection limit. . . . .	63

3.3	Comparison of four methods' performance on regression coefficient of dust mite concentration (true value is 0.4), now the contamination level is 1 (severe contamination), RMSE: root mean squared error, CP: coverage probability, IW: interval width. . . . .	68
3.4	Illustration for a more efficient design for standards and new samples (dilutions) in a multiplex plate with 96 wells. . . . .	69
3.5	Comparison of the original design with 26 observations of calibration sample and 23 unknown samples per plate versus new design with 14 observations of calibration sample and 27 unknown samples per plate, RMSE: root mean squared error, CP: coverage probability, IW: interval width. . . . .	69
4.1	Comparison of the naive two-step model and Bayesian two-step model performance with the benchmark model on regression coefficient of the key predictor of interest (true value is 0.4) under no contamination with 500 iterations, RMSE: root mean squared error, CP: coverage probability. . . . .	95
4.2	Comparison of three methods' performance on regression coefficient of the key predictor of interest (true value is 0.4) under contamination with 500 iterations, RMSE: root mean squared error, CP: coverage probability. . . . .	98
B.1	A potentially more efficient design for standards and new samples (dilutions) in a multiplex plate with 96 wells. . . . .	120
B.2	Supplementary section an even more efficient design with 8 observations of calibration sample and 29 unknown samples per plate. . . . .	120

## **Acknowledgements**

First and foremost, I would like to thank my advisors Dr. Andrew Gelman and Dr. Qixuan Chen. During the past years of my doctoral studies, I have learned many things from them that I cannot imagine ever learning from others. Dr. Gelman and Dr. Chen have always been encouraging and supportive, and I gradually developed my research tastes under their supervision. Whenever I have had any questions, Dr. Gelman and Dr. Chen have always been willing to help with great patience and consideration. They genuinely care about their students and always hope that they succeed in their field of choice.

I would also like to thank all of the faculty members for their willingness to be on my dissertation committee. I deeply appreciate the helpful comments and suggestions from Dr. Todd Odgen, Dr. Matthew Perzanowski, and Dr. Xiao Wu, which helped me to improve my dissertation dramatically. I am extremely grateful for their time and support.

In addition, I treasured my time in the Biostatistics Department, where I met many close friends and received much more support than I expected. I enjoyed the department events greatly and felt involved during my time at there. I would also like to express my deepest thanks to all my fellow students over the years for being there together with me.

Lastly, I would like to express my deepest thanks to my family. I would not have achieved as much as I have done without their unconditional support, care, and love.

## Chapter 1: Introduction and Overview

In this thesis, we seek to develop multiple Bayesian statistical models for solving a series of problems related to model contamination, global calibration, and measurement error in the field of applied statistics, with a special focus on serial dilution assay data and public health datasets. Motivated by some current problems and challenges researchers and technicians are facing in lab data measurement and model construction, we develop several Bayesian inference methods and models to address not only the specific problems we have encountered but also propose general Bayesian framework ideas that could be adapted to more general applied statistics problem and thus benefit broader user groups.

For the application examples, we work on developing reliable and robust Bayesian inference procedures to investigate the potential association between exposure to indoor allergens and children's asthma. Asthma is one of the most common diseases that affect young people. Therefore, it is crucial to conduct large-scale epidemiologic studies to investigate the potential causes of asthma in children. The ultimate applied goals of this series of studies are first to estimate indoor allergen concentrations in wet lab experimental settings accurately and then to use appropriate statistical models to investigate the potential associations between exposure to indoor allergens and asthma morbidity among children. We propose several Bayesian inference methods that could significantly improve the current measuring technology commonly used in immunoassay labs to measure allergen concentrations and the current estimation method used in epidemiology research. Furthermore, by proposing the idea of a Bayesian workflow, we provide users with a toolbox that they can use when they face similar problems.

In our first project, we are motivated by the analysis of raw immunoassay data of indoor allergens, where researchers have sought to obtain an accurate estimation of specific indoor allergen concentrations using lab machines and algorithms. Our methods focus on data generated by im-

immunoassays, which are widely used in modern biological science to detect the concentration of a particular analyte. In the standard serial dilution process, multiple dilutions are applied to the sample to generate repeated measurements and thus give more chance for better estimation accuracy. The primary estimation process involves fitting statistical models to the dataset generated through a serial dilution assay, where multiple dilutions are usually applied to each sample to increase the chance of obtaining an accurate estimation. Commonly used techniques in the field of serial dilution assay include Enzyme-Linked Immunosorbent Assays (ELISA), which are based on the interaction of antibodies and antigens where the antibodies attach to specific proteins or nucleic acids and Multiplex-array for indoor allergens (MARIA). Compared with ELISA which could only estimate one type of allergen at a time, MARIA could simultaneously estimate multiple allergens and also provides more precise and convenient measurements. Thus in our project, we have focused on serial dilution assay data generated by MARIA. In a serial dilution assay in this estimation process, a standard sample is usually used for calibration analysis, and by using the estimated calibration curve, technicians obtain corresponding estimates for the unknown sample concentrations. However, researchers have found that this conventional estimation method cannot provide satisfying estimation results due to challenges related to the detection limits and the potential existence of sample contamination. To solve these two challenges, our first project developed a Bayesian mixture model for estimating unknown sample concentrations in a serial dilution assay. This not only allows each unknown sample the flexibility to be contaminated but also borrows information from the standard calibration sample for more accurate parameter estimation using hierarchical models. Moreover, by treating the unknown sample concentration itself as a random variable, we were even able to estimate very low concentrations accurately and to directly overcome the problem of being below the detection limit, which is commonly encountered in serial dilution assays. Furthermore, our model was able to output the sample-level posterior probability of contamination for each unknown sample, which could be used as a flag for lab technicians to investigate the potential reasons behind sample contamination further. By performing various simulation settings, we demonstrated the improved performance of our proposed method compared

with several competing methods, including the conventional method that is used in current lab environments. Furthermore, we applied our proposed method to a real serial dilution assay plate and found improved estimation accuracy compared with current lab results. In addition to the proposed method, we developed the general idea of a Bayesian workflow in this model contamination problem setup. This workflow enables users to conduct reliable statistical inferences – from the initial construction of the model to its evaluation and modification.

After developing the Bayesian mixture model for sample contamination, we sought to extend our proposed method to more general settings. We faced two remaining challenges in the field. First, each experimental unit's capacity is typically limited, and multiple experimental units are usually measured under different lab environments, such as time, temperature, and humidity. A natural question is whether we could extend our proposed methods to model multiple experimental units simultaneously. Second, the estimation of allergen concentration is usually not the end of the story; rather, it mainly provides some intermediate results for further large-scale epidemiologic studies. The Bayesian model fitting process not only provides a single-point estimation but also contains all necessary information contained in the posterior distribution; therefore, the question is whether we could take advantage of the uncertainty measurement and consider it a necessary part of modeling in epidemiologic association studies. In the second project, we started by comparing local calibration with global calibration. We developed a joint Bayesian inference model that integrated all of the aforementioned concepts and had two main parts. First, we applied hierarchical models to allow different experimental units to have their own calibration curve parameters while simultaneously using partial pooling to borrow information from each other for more efficient statistical inference. Second, we directly propagated the measurement uncertainty of the unknown sample concentrations to follow-up epidemiologic association studies by switching from the classical fixed-X regression design to the random-X design. This provided additional information for concentration estimations and resulted in more reliable inference. In the simulation studies, we compared this joint modeling approach with competing methods and found that it could recover the true regression coefficients more accurately than other methods. Furthermore, we explored

the potential of a more efficient study design by putting more unknown samples and reducing the number of calibration standard samples on each experimental unit by taking advantage of global calibration in the joint Bayesian model. In real-life studies, we applied our proposed models to the NYC Neighborhood Asthma and Allergy Study (NAAS) and observed improvements in indoor allergen exposure measurement accuracy and the new methods led to important findings regarding the associations between indoor allergen concentrations and asthma morbidity among children.

The major motivation for our third project was that although Bayesian joint modeling is compelling and can estimate the underlying associations between indoor allergen concentrations and asthma morbidity accurately, it requires the access to both the raw immunoassay data for indoor allergens and the epidemiologic data. However, data access can sometimes be limited; researchers might not have access to both raw lab measurement data and individual-level epidemiologic data. Therefore, it is important to develop multi-step Bayesian inference models that allow the modeling of different datasets in multiple steps. In the third project, we aimed to develop a two-step model that considered measurement uncertainty in the follow-up epidemiologic studies while allowing the epidemiologic model to be fitted separately from the models for estimating the concentrations, which is easier to implement than the joint model developed in the second project. The final model that we proposed was a Bayesian two-step inference model. In the first step, we did not use any parametric assumptions, and the only requirement was that the model should output a point estimation and the associated measurement uncertainty. In the second step, we used a measurement error model to propagate the measurement uncertainty in the epidemiological regression model. The ultimate goal of our Bayesian two-step inference model was not to lose much estimation efficacy and accuracy when considering measurement uncertainty compared with the Bayesian joint model while providing a more user-friendly modeling strategy. The said strategy should allow users to conduct research projects even with partial information efficiently. We further investigated multiple simulation scenarios and found that, compared with competing methods, our proposed Bayesian two-step model achieved satisfactory results. We applied the two-step model to the NYC NAAS and found the results aligned well with the findings of our second project. The application

results from the real-life studies have also indicated that our Bayesian two-step method had similar measurement accuracy as the Bayesian joint model in our second project. Furthermore, the idea of the Bayesian workflow was also applicable to the second and third projects. Overall, we considered the existence of measurement errors to be another source of model contamination, which corresponds to the general idea of applying Bayesian inference models to solve the problem of model contamination in our first project.

## **Chapter 2: Bayesian Framework for Model Contamination with Application to Serial Dilution Assay**

### **2.1 Statistical analysis of contaminated data**

#### 2.1.1 Background

The Bayesian approach to contaminated data is to use a mixture model so that each data point has a large probability of coming from the uncontaminated model and a small probability of coming from a contaminated distribution (Box and Tiao, 1968). The probability of contamination can be estimated from the data. Two general challenges arise: contamination can happen at different levels of an experiment (not just for individual data points), and the model for the contaminated observations is typically much more speculative than the model for uncontaminated data.

We propose to address the challenges of Bayesian contamination models using several steps. First, we use a hierarchical model so that groups of data as well as individual data points, can be contaminated. Second, we use the estimated mixture probabilities as a way to flag possibly-contaminated measurements while also aiming for the contaminated model to provide rough inference for underlying parameters of interest. Third, we embed the process within a Bayesian workflow so that deviations in posterior predictive checks can be used to improve the contamination model.

#### 2.1.2 Contamination of immunoassays

We develop our mixture model approach in the context of literal contamination that arises in laboratory measurement. Epidemiologic investigations of environmental-disease associations rely on accurate measurements of environmental exposures. To determine the level of environmental ex-

posure, samples are collected, and concentrations are measured in labs using immunoassays with serial dilutions.

An immunoassay is a biochemical test that measures the presence or concentration of an analyte depending on the reaction to an antibody or an antigen (Vashist and Luong, 2018). Typically, the test is conducted with serial dilutions of both a standard sample and multiple unknown samples in a microtiter plate to detect and quantify proteins and other substances that can be bound to antibodies or antigens (Van Weemen and Schuurs, 1971; Engvall and Perlmann, 1972). In the standard serial dilution process, multiple dilutions are applied to both the standard and unknown samples, and signal response to the antibody-antigen interaction is measured at the dilutions of each sample. The serial dilutions of the standard sample are used to make a calibration curve relating the signal response to the known concentrations of the diluted standard sample. The calibration curve is then used for estimating the concentrations of the unknown samples at different dilutions. The serial dilutions of the unknown samples allow the measure of uncertainty and improve estimation accuracy.

In classical calibration inference, the concentrations of the diluted unknown samples are determined by reading off from the calibration curve of the same plate. The estimated concentrations are then scaled back by dividing by the corresponding dilution levels and averaged to obtain a final estimate of the concentration for each unknown sample (Gelman et al., 2004). Although it is easy to implement, the classical calibration approach has several limitations. First, estimating the concentration by inverting the calibration curve does not account for calibration uncertainty and can lead to significant measurement error (Lee and Whitmore, 1999; Giltinan and Davidian, 1994). Second, variances in measurements at the extreme ends of the calibration curve can place sample measurements below limits of detection (LOD), and these measurements are typically replaced with a constant such as 0, LOD, LOD/2, or  $\text{LOD}/\sqrt{2}$ , but any of these options can lead to serious bias in statistical inference (Guo et al., 2010). Third, averaging the estimates of diluted samples is inefficient since measurements of highly diluted samples could have greater variance by scaling the measurements by dilution levels.

Another major challenge associated with the analysis of immunoassays is sample contamination, where samples contain not only the analytes of interest but also other substances that can interfere with the performance of an assay. Figure 2.1 shows the results of an experiment evaluating the impact of sample contamination in measuring a protein from dust mites, Der p 1, using immunoassays. In this experiment, a commercially available laundry detergent was added as a contaminant to 16 dust samples at 8 different contamination concentrations, ranging from 1 in 10 to 1 in 100 million. Serial dilution was then applied to each of the 16 contaminated samples. Figure 2.1 shows that the signal response curves look very different across different contamination concentrations, with more flat response curves associated with more contaminated samples. Thus, when contamination exists in the unknown samples, the classical calibration inference method does not apply anymore, because signal response curves for the contaminated samples can be very different from the calibration curve. Since the concentration or even presence of these substances is often unknown, methods that can detect and account for potential contamination in the samples are needed when analyzing immunoassay data.

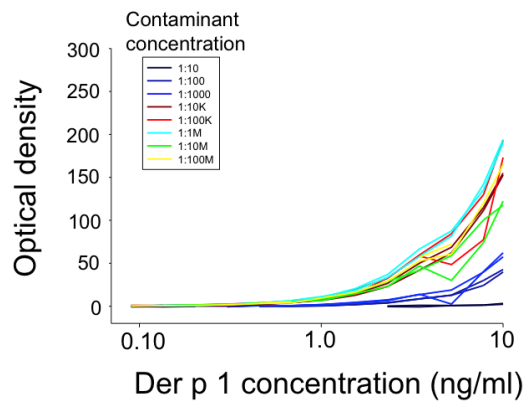


Figure 2.1: Rate response (mOD U/min) measured in an immunoassay for a range of dust mite allergen (Der p 1) concentrations, where samples are contaminated with 8 different concentrations of a known contaminant (household laundry detergent) and mOD U/min stands for millioptical density units per minute.

### 2.1.3 Motivating application

Our study focuses on data generated by immunoassays for measuring indoor allergen levels from the New York City Neighborhood Asthma and Allergy Study (NYC NAAS). The NYC NAAS is a study of 7 and 8-year-old children with and without asthma (Perzanowski et al., 2008, 2013; Chen et al., 2016). Homes of participating families were visited, and a dust sample was collected from each child’s bed by vacuuming the fitted sheet on the upper half of the bed and both sides of the pillows. The bed dust samples were extracted with PBS 0.05% Tween, pH 7.4 at a concentration of 50 mg/ml and stored at –20 degrees Celsius until analysis. Allergens from cockroach, cat, dog, rat, mouse, and dust mites were measured by immunoassays using microtiter plates.

	1	2	3	4	5	6	7	8	9	10	11	12
	Standard Sample			Unknown Samples								
				1/10	1/100	1/10000	1/10	1/100	1/10000	1/10	1/100	1/10000
A	1	1/16	1/256	Unk 1	Unk 1	Unk 1	Unk 9	Unk 9	Unk 9	Unk 17	Unk 17	Unk 17
B	1	1/16	1/256	Unk 2	Unk 2	Unk 2	Unk 10	Unk 10	Unk 10	Unk 18	Unk 18	Unk 18
C	1/2	1/32	1/512	Unk 3	Unk 3	Unk 3	Unk 11	Unk 11	Unk 11	Unk 19	Unk 19	Unk 19
D	1/2	1/32	1/512	Unk 4	Unk 4	Unk 4	Unk 12	Unk 12	Unk 12	Unk 20	Unk 20	Unk 20
E	1/4	1/64	1/1024	Unk 5	Unk 5	Unk 5	Unk 13	Unk 13	Unk 13	Unk 21	Unk 21	Unk 21
F	1/4	1/64	1/1024	Unk 6	Unk 6	Unk 6	Unk 14	Unk 14	Unk 14	Unk 22	Unk 22	Unk 22
G	1/8	1/128	1/2048	Unk 7	Unk 7	Unk 7	Unk 15	Unk 15	Unk 15	Unk 23	Unk 23	Unk 23
H	1/8	1/128	1/2048	Unk 8	Unk 8	Unk 8	Unk 16	Unk 16	Unk 16	blank	blank	HC Control

Table 2.1: Map for standards and new samples (dilutions) in a multiplex plate with 96 wells.

Each microtiter plate contains 96 wells (12 columns and 8 rows). Table 2.1 shows a map of standards and unknown samples. This plate contains 2 replicates of standards (columns 1-3) and 23 unknown samples (columns 4-12). Each of the standard sample was prepared at the same known and fixed concentration and then diluted using 12 two-fold dilutions ranging from 1 to 1/2048. Each of the 23 unknown samples was analyzed using 3 dilutions at 1/10, 1/100, and 1/10000. This big variation of dilution levels in the unknown samples was designed to allow the estimation of a wide range of concentrations. Two blank samples were included in the plate to capture background noise.

Table 2.2 shows standard data for another allergen from dust mites, Der f 1, and the estimated concentrations of 6 selected unknown samples using the classical calibration method in a microtiter

Sample ID	Dilution	Signal Response	Con. (ng/ml)	Sample ID	Dilution	Signal Response	Con. (ng/ml)
Standard	1	16418	125	Standard	1/128	906	0.98
	1	18977	125		1/128	1141	0.98
	1/2	16350	62.5		1/256	397	0.49
	1/2	17960	62.5		1/256	450	0.49
	1/4	14573	31.25		1/512	164	0.24
	1/4	14625	31.25		1/512	166	0.24
	1/8	12380	15.63		1/1024	72.17	0.12
	1/8	13310	15.63		1/1024	77.83	0.12
	1/16	8728	7.81		1/2048	34.33	0.06
	1/16	9175	7.81		1/2048	39.27	0.06
	1/32	5152	3.91		0	4	0
	1/32	5313	3.91		0	4	0
	1/64	2353	1.95				
	1/64	2424	1.95				
Unk 10	1/10	4259.5	32.47	Unk 23	1/10	1904	15.87
	1/100	241	32.21		1/100	783	78.78
	1/10000	6	OOB<		1/10000	39	647.51
Unk 3	1/10	660	6.92	Unk 22	1/10	51	0.87
	1/100	646	68.04		1/100	7	OOB<
	1/10000	74	1233.87		1/10000	3	OOB<
Unk 9	1/10	475	5.39	Unk 16	1/10	9	OOB<
	1/100	251	33.23		1/100	5	OOB<
	1/10000	6	OOB<		1/10000	8	OOB<

Table 2.2: Standards and selected unknown samples from a multiplex plate with the concentrations of the new samples estimated using the classical calibration method.

plate. The standard data are presented with signal responses and true concentrations on the top, with an initial concentration of 125 ng/ml. The bottom of Table 2.2 shows the signal responses and estimated concentrations of the 6 unknown samples on the same plate. From this plate, we can see the limitations of the classical calibration approach. First, the estimated concentrations of the three dilutions of the same sample can be very different. For example, with unknown sample 3, the estimated Der f 1 concentration ranges from 6.92 ng/ml at a dilution of 1/10 to 1233.87 ng/ml at a dilution of 1/10000. The very large estimate of 1233.87 ng/ml was obtained by dividing the estimated concentration of 0.123 (obtained by reading off the calibration curve with a signal response of 74) by the dilution level 1/10000. Averaging the estimated concentrations of these three diluted samples to obtain the final estimated concentration can be biased and inefficient. Second, the concentrations are not estimable for samples with the signal response at the extreme ends of the calibration curve. For example, with unknown sample 16, all three dilutions lead to signal responses that are smaller than those of the lowest concentrations in the standards (0.06

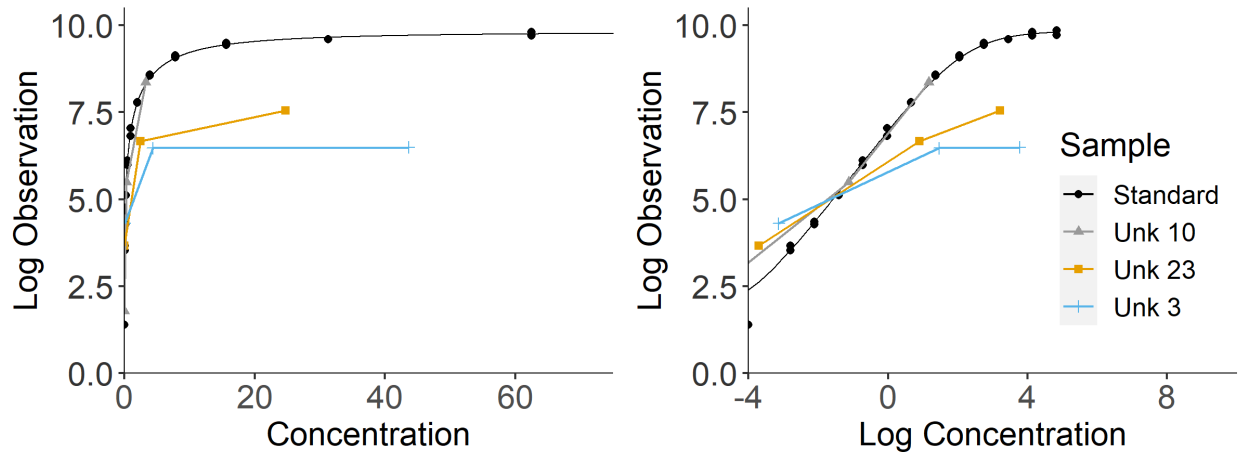


Figure 2.2: Examples of sample contamination from the New York City Neighborhood Asthma and Allergy Study. Curves show data from standards (calibration data) and three unknown samples on a single microtiter plate.

ng/ml), and thus all three dilutions resulted in a concentration estimate of OOR< (out of range).

The existence of contamination is a concern, as environmental samples were collected from bed and floor samples in settled dust, and unknown contaminants can be present in the samples that could affect the assay. Figure 2.2 plots log-transformed signal responses read from the machine versus true or estimated concentrations of Der f 1 for the standards and three unknown samples from Table 2.2. To better visualize the lower concentrations, log transformation is applied to the  $x$ -axis in the plot on the right. The dots represent the data from the standards, overlaid with the calibration curve. The values on the  $x$ -axis for the three unknown samples were derived from their final estimated concentrations using the classical calibration method times the corresponding dilution levels. Figure 2.2 shows that the signal response curves of unknown samples 3 and 23, plotted using plus sign and square, respectively, are quite different from the calibration curve estimated from the standards. The crossing of the lines implies that no estimated concentration would allow the two curves to line up. For comparison, we also plot the data from unknown sample 10 (using triangles), which has a similar response curve shape to the calibration curve. This may suggest potential contamination in unknown samples 3 and 23 but not in unknown samples 10.

#### 2.1.4 Statistical modeling of assay data and contamination

Contamination brings extra challenges to the estimation problem in serial dilution assays. Models that can not only overcome the limitations of classical calibration inference but also detect and handle unknown contaminated samples are in need of immunoassays. Previous work in the field of concentration estimation in serial dilution assay includes the estimation of parameters of the dose-count curve using a model assuming proportional variance with respect to the mean count in radioligand assay (Finney, 1976), summary statistics including the mean, median, standard deviation of serial dilution data and estimation of threshold concentrations accounting for grouping of observations (Hamilton and Rinaldi, 1988), concentration-response modeling that assumes additive sequential dilution error model on log-scale of concentrations between the reference preparation and test preparation in serial dilution assay (Racine-Poon et al., 1991), serial dilution error model accounting for the propagation of random measurement error in the dilution process (Higgins et al., 1998), and a lognormal measurement error model for serial dilution assays (Lee and Whitmore, 1999). Nonlinear multilevel models for repeated measurement data and calibration inference have been shown to have potential advantages (Giltinan and Davidian, 1994; Davidian and Giltinan, 2017). Bayesian methods for serial dilution assays have been shown to yield better estimates of unknown concentrations. Bayesian calibration framework using hierarchical models, including multiple sources of variation has been proposed to solve the problem (Gelman et al., 2004). The Bayesian calibration framework has been extended by adding an informative prior to the calibration model (Klauenberg et al., 2015). Bayesian lognormal measurement error model with constant variances has been shown to outperform the classical calibration curve inversion method (Feng et al., 2011). Some robust Bayesian models for serial dilution assays have been proposed, including Bayesian model averaging weighted by posterior model probabilities (Morales et al., 2006) and a robust Bayesian multilevel model that separates the experimental noise into detection sensor noise and stochastic natural noise (Fong et al., 2012). On the other hand, some studies show that model misspecification in generalized linear mixed models could lead to significant errors in both the inference of fixed effects and random effects (Hui et al., 2021). Also, hierarchical Bayesian

models and Bayesian mixture models have been applied to the detection of sample contamination in water samples and micro-bio samples (Busschaert et al., 2011; Liu et al., 2021). However, none of them consider the particular problem of modeling contamination in serial dilution assay data.

### 2.1.5 Our contributions

In this project, we develop a novel general Bayesian framework for concentration estimation in serial dilution assays allowing contamination in the unknown samples. The goal is for the user to be able to fit a model that estimates the underlying concentration while also identifying the samples that are likely to be subject to contamination. This involves a general challenge, familiar from other areas of statistics, of evaluating a hypothesis that is imperfectly specified: in this case, the size of the contamination and the form of the contamination model are unknown. Because of this misspecification, we will expect the inference under contamination to have a systematic error, hence in practice the identification of contamination can motivate further investigation.

## 2.2 Bayesian contamination model for dilution assays

### 2.2.1 Notation and setup

We consider data of a single allergen in a single microtiter plate. Let  $x$  denote allergen concentration, and  $y$  denote signal response to the antibody-antigen interaction, measured as color intensities read from the machine. Define  $\theta_0$  to be the known concentration of the undiluted standard sample,  $\theta_j$  to be the unknown true concentration for new sample  $j$  on the same plate, and  $d_i$  to be the dilution level of observation  $i$ . For serial dilution assays, we have:

$$x_{ij} = x_{init}d_i, \tag{2.1}$$

where  $x_{init}$  equals  $\theta_0$  for standard sample and  $\theta_j$  for unknown samples.

Studies have shown that log transformation of the measurements has the effect of regularizing the variance in the measurement error model (Feng et al., 2011). Therefore, we model the log-

transformed  $y$  instead of  $y$  as a function of  $x$  in the serial dilution assays. To allow a smooth and increasing dose-response relationship between  $\log(y)$  and  $x$ , we consider a four-parameter logistic curve:

$$E(\log(y)|x, \boldsymbol{\beta}) = \log(g(x, \boldsymbol{\beta})) = \log\left(\beta_1 + \frac{\beta_2}{1 + (x/\beta_3)^{-\beta_4}}\right), \quad (2.2)$$

where  $\beta_1$  represents the color intensity at zero concentration,  $\beta_2$  represents the increase to saturation,  $\beta_3$  represents the concentration where the curve turns, and  $\beta_4$  represents the rate at which saturation occurs (Higgins et al., 1998).

For observations  $y_{i0}$  from the standards sample with known initial concentration  $\theta_0$  and known dilution level  $d_i$ , we assume a lognormal measurement error model:

$$\log(y_{i0}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_0 d_i, \boldsymbol{\beta})), \sigma_y). \quad (2.3)$$

### 2.2.2 Bayesian contamination model

Contamination in serial dilution assay is generally defined as the situation where the dose-response curves of some of the unknown samples are different from the dose-response curve of the standard calibration sample. With the assumption that the contaminate has different effects on the assay results at different concentrations of the contaminant that does not directly correlate with changes in measured analyte, our general inference goal is to ensure that the concentration estimates of uncontaminated samples are not affected by contaminated samples, and we can flag and correctly estimate the concentrations of contaminated samples even with contamination. Thus we need a robust and flexible model to allow potential contamination. With each unknown sample, we model it using a mixture of two normal measurement error models in the log scale. For observations  $y_{ij}$  from sample  $j$  with known dilution level  $d_i$  but unknown concentration  $\theta_j$ , we have

$$\log(y_{ij}) \stackrel{ind}{\sim} (1 - \lambda) * \text{normal}(\log(g(\theta_j d_{ij}, \boldsymbol{\beta})), \sigma_y) + \lambda * \text{normal}(\log(g(\theta_j d_{ij}, \boldsymbol{\beta}_j)), \sigma_{y_j}), \quad (2.4)$$

where

$$\boldsymbol{\beta}_j = (\beta_1, \beta_{2j}, \beta_3, \beta_4), \beta_{2j} = \beta_2 e^{\delta_{\beta_{2j}}}, \sigma_{y_j} = \sigma_y e^{\delta_{\sigma_j}}. \quad (2.5)$$

The first mixture component models the unknown sample using the same measurement error model as the standard sample. The second mixture component allows the unknown sample to have different  $\beta_2$  and  $\sigma_y$  in the measurement error model than that for the standard sample. The intuition for only allowing  $\beta_2$  to be different from the standard sample is that  $\beta_2$  controls the increase to saturation, which has been shown to explain the most difference between the standard calibration curve and the signal response curve of contaminated samples based on real data. Also,  $y_{ij}$  tends to show more variation among contaminated samples than the standard and uncontaminated samples. We let the variance of the second mixture component,  $\sigma_{y_j}$ , to be larger than the variance of the standard sample,  $\sigma_y$ . If we believe that most of the unknown samples are uncontaminated, we can assign an informative prior to  $\lambda$  with a mean close to 0, for example,  $\lambda \sim \text{beta}(1, 10)$ . We can modify the values of the shape parameters in the beta distribution for plates with a low to high proportion of contaminated samples.

### 2.2.3 Prior specification

In our setting, all four of the  $\boldsymbol{\beta}$  parameters must be positive, so we consider weakly informative priors as follows:

$$\log(\boldsymbol{\beta}) \sim \text{normal}(0, 10), \sigma_y \sim \text{normal}^+(0, 10). \quad (2.6)$$

To model the contamination, we assign a weakly informative normal prior for  $\delta_{\beta_{2j}} \sim \text{normal}(0, 1)$  and weakly informative exponential prior for  $\delta_{\sigma_j} \sim \text{exponential}(1)$ . The latter indicates that the potentially contaminated sample should have a larger variance than the uncontaminated samples.

We assign hierarchical exponential priors for the concentrations  $\theta$  across different unknown samples  $j$  and let the model estimate the hyperparameter from the data. The exponential prior is consistent with the possibility of zero or effectively zero concentrations for some samples while

allowing high concentrations for others:

$$\theta_j \sim \text{exponential}(\mu), \mu \sim \text{normal}^+(0, 0.1). \quad (2.7)$$

#### 2.2.4 Alternative model specification

We assume weakly informative hierarchical exponential priors for  $\theta_j$  so that we can model both large concentrations and the lower positive end well. Alternative prior distributions for  $\theta_j$  include less diffuse exponential distribution, t-distribution on the log-scale, horseshoe distribution, and skewed-normal distributions. We will further discuss the performance of these alternative priors in the simulation studies. For the mixture proportion parameter  $\lambda$ , we assign a beta prior with a mean close to zero to generally represent the situation where most of the unknown samples are uncontaminated. Alternatively, we can modify such prior by changing the values of the shape parameters in the beta distribution to accommodate situations where the contamination proportions are high. Also, remember that the main goal of our contamination model is not to accurately model the underlying unknown contamination mechanism, but to generate potential signals for lab technicians to consider the causes behind the existence of contamination. So it is natural to adapt our contamination model based on specific lab experimental settings and needs. For example, the model could be extended by allowing all the  $\beta$  parameters to be different between the standard sample and unknown samples.

#### 2.2.5 Computation

We fit the model using the NUTS algorithm in Stan as called from `cmdstanr` (Homan and Gelman, 2014; Carpenter et al., 2017). NUTS is a variation of Hamiltonian Monte Carlo that combines the ideas of Markov chain Monte Carlo and deterministic simulation methods by using the derivatives of the density function being sampled, thus allowing the random walk behavior to move more efficiently to the target distribution. We track the mixing of the simulated chains using effective sample size and the  $\widehat{R}$  diagnostic (Vehtari et al., 2021).

Based on the posterior inference summary, we need to make sure that all the parameters in the model have a large enough effective sample size, which measures the autocorrelation within MCMC chains for reliable posterior inference results. In Bayesian computation, the role of effective sample size in the Markov chain Monte Carlo central limit theorem is similar to the concept of the number of independent observations used in the classical central limit theorem. Also, all the parameters in the model have  $\widehat{R}$  values very close to 1 both in simulation studies and real-life examples, which indicates that the chains are mixing well. We further visualize the mixture of chains using R package bayesplot and find all the chains are mixing very quickly and well both in simulation studies and real-life examples. Finally, the computation time for our proposed model is speedy and can be measured in seconds both in simulation studies and real-life examples.

### **2.3 Bayesian workflow for contamination model**

When facing the possibility of model contamination, we typically do not know what models we will end up fitting, and it becomes necessary to compare different models and their implications, following the general principles of workflow for Bayesian model evaluation and computing (Gelman et al., 2020).

We need to choose our initial model based on our understanding of the applied problem and the data. At this step, we could treat the components of the Bayesian model using the idea of modular construction, for example, starting from something simple and then generalizing and expanding the components when needed. Sometimes we might also need to scale or transform the model parameters to make them interpretable. Posterior predictive checking is helpful for checking the model fit, and prior predictive checking is useful for checking the implications of a generative model. Robustness analysis considers nearby alternative models. For the contamination problem, we could study the behavior of the fit under different degrees of contamination level across different parts of the parameter spaces. By repeating the simulation many times and fitting our model to each simulated data set, we could end up with a more comprehensive understanding of our model.

For model evaluation and modification, we need to find a proper summary of results after

propagating uncertainty in Bayesian inference and think about whether our inference results make sense given the data and the research question of interest. Especially for modeling contamination, usually the form of contamination is unknown, and thus we usually start with flexible models and make necessary adaptations based on the model evaluation results. We could also perform sensitivity analysis to study the prior influence. Finally, we could extend or modify our model as necessary when there is new data or a change of constraints in the prior distribution.

Specific aims for immunoassays data include getting the posterior probability of each sample being contaminated, inference for the underlying concentration given the contamination model as well as gathering additional information for the samples or new data for further inference. Furthermore, there are two immunoassays-specific practical aims we would like to address in the simulation section. First, besides the classical prior predictive checking, we are also interested in how robust our method is to the potentially misspecified underlying data generation process. For example, currently we are assigning an exponential prior for the unknown concentrations, but in real-life, maybe the true underlying concentration distribution could follow some other distributions, for example, the lognormal distribution. Second, we are interested in how well our model would fit a lot of exactly zero or very small concentrations. Although exact 0 allergen concentration is not common in real life, it is possible in theory. This scenario brings us extra challenges in the following two aspects: Firstly, such observed concentrations are kind of unlikely generated from the exponential distribution as we specify as our prior; Secondly, the relative extreme large concentrations could be either classified as “contaminated samples” with inaccurate concentration estimation (since they are different from the majority of the remaining unknown samples) or “uncontaminated samples” with accurate concentration estimation based on the degree of flexibility of our model. To be more general, there is always a trade-off between the ability to accurately pick up those indeed contaminated samples and being too sensitive to extreme values that even classify those uncontaminated samples with unusually large concentrations as contaminated samples since the general definition of contamination is to be different from others in the sample plate.

In the following sections, we will use both the applied example as well as simulation studies to

demonstrate our proposed general Bayesian workflow for model contamination.

## 2.4 Analysis of a dust mite allergen immunoassay plate

As discussed above and shown in Table 2.2, many of the observations in the NYC NAAS are either below the detection limit or have abnormal uncertainty across the three diluted measurements of the same unknown sample, which might indicate contamination. In this section, we applied our method to explore whether it could accurately estimate the underlying concentration and identify potentially contaminated samples. We would also like to illustrate several key workflow steps we have applied in the NYC NAAS data analysis.

### 2.4.1 Initial model construction

The first step in our model-building process is to start with something simple. The starting point of our model building is the following Bayesian normal measurement error model for analyzing serial dilution assay data with  $y$  as the raw observation from the machine,  $x$  as the concentration and  $\boldsymbol{\beta}$  as the dose-response curve parameters (Gelman et al., 2004):

$$E(y|x, \boldsymbol{\beta}) = g(x, \boldsymbol{\beta}) = \beta_1 + \frac{\beta_2}{1 + (x/\beta_3)^{-\beta_4}}, \quad (2.8)$$

$$y \stackrel{ind}{\sim} \text{normal} \left( g(x, \boldsymbol{\beta}), \left( \frac{g(x, \boldsymbol{\beta})}{A} \right)^\alpha \sigma_y \right), \quad (2.9)$$

Here in (2.9) both  $A$  and  $\alpha$  are regularizing constants. And we sequentially adapt and improve our model based on that framework by making several adaptations and extensions to (2.9). First, we have observed data skewness, a log transformation to the data would make the inference more interpretable and have the effect of regularizing the variance in the measurement error model (Feng et al., 2011). In order to make our parameters interpretable in the sense of shape and rate for the dose-response curve, we perform the corresponding log transformations for the curve parameters. Also, since we switch to the log scale, the original multiplicative variance component has now

become linear and simpler, and we name it the intermediate model:

$$E(\log(y)|x, \boldsymbol{\beta}) = \log(g(x, \boldsymbol{\beta})) = \log\left(\beta_1 + \frac{\beta_2}{1 + (x/\beta_3)^{-\beta_4}}\right), \quad (2.10)$$

$$\log(y) \stackrel{ind}{\sim} \text{normal}(\log(g(x, \boldsymbol{\beta})), \sigma_y), \quad (2.11)$$

Then we extend the intermediate model by adding another mixture component with different mean and variance components to allow for potential contamination in the samples.

$$\log(y) \stackrel{ind}{\sim} (1 - \lambda) * \text{normal}(\log(g(x, \boldsymbol{\beta})), \sigma_y) + \lambda * \text{normal}(\log(g(x, \boldsymbol{\beta}^*)), \sigma_y^*). \quad (2.12)$$

To follow the idea of modular construction, we put flexible models and weakly informative priors as placeholders for future modification if we gain some insights later in the workflow. In Figure 2.3, we can see that the model fitting results are getting much better as we gradually improve our model. Sample 23 is highly suspicious to be contaminated based on our discussion in Table 2.2, and the fitting results are much better as we go from the initial model to the final model. Even for the standard calibration sample, we find that our initial model underestimates the variance component while the intermediate model overestimates the variance component.

#### 2.4.2 Model fitting while addressing computation issues

We fit our Bayesian models using HMC in Stan and validate the computation of our models. We check both the convergence diagnostics of the algorithm as well as whether the computation time is reasonable. We used the default computational settings in Stan for the adaptation and warmup and found the  $\widehat{R}$  diagnostic statistic is close to 1, and the effective sample size is sufficient without divergent transitions. The computational time frame is also reasonably short, which enables us to leverage our proposed model to a much larger data set if necessary. In practice, sometimes computational issues would bother the users much, and (Gelman et al., 2020) presents a comprehensive summary of methods addressing the potential computational problems.

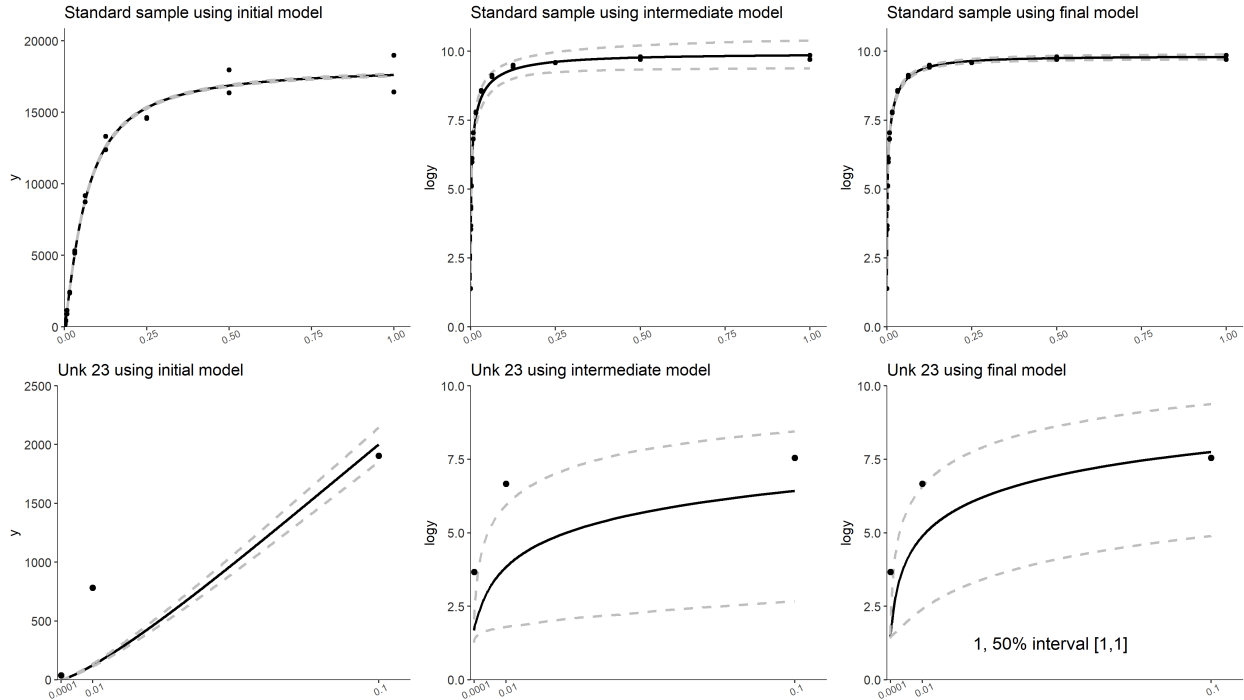


Figure 2.3: Demonstration of sequential model improvement under Bayesian workflow

### 2.4.3 Model evaluation and specific aims for immunoassays data

For model evaluation, we first develop graphical representations for how our model fits the data in Figure 2.4 and Figure 2.5 to summarize inference and propagate uncertainty. Figure 2.4 shows our model-fitting results to one of the microtiter plates measured in the NYC NAAS for Der f 1. The top three plots represent the zoomed-in estimated dose-response curve for standard data with the y-axis for the log-transformed  $y$  and the x-axis for the dilution level. We ordered the unknown samples by their upper bound of the corresponding uncertainty level. For each unknown sample, we plotted the three observed measurements on the log scale, the fitted mean curve of the four-parameter logistic regression based on the posterior median of the  $\beta$  parameters, and the Bayesian 95% posterior interval of the fitted curves. For each new sample, we also output the estimated probability of contamination, showing the posterior median and the associated 50% probability intervals. Compared to the classical calibration method, which reads off the estimation using the same standard data for all new samples, our Bayesian model allows each new sample to have its

own dose-response curve if they are identified by the model to be contaminated samples. Instead of yielding an unstable estimate of concentration by averaging the highly variable estimates from the three dilutions, our proposed method could fit the observed data well with enough uncertainty to cover the potential variation and contamination. In this plate, the model estimated the probability of contamination to be 1, with the upper and lower bounds of the 50% probability interval both being 1 in the unknown samples 17, 23, 9, 3, and 16. This may suggest contamination in these new samples. Figure 2.4 also shows that some of the dilution measurements do not align well with the corresponding estimated response curve. Our model estimates a wider 95% probability interval to allow all the observed measurements to fall inside the estimated lower and upper bounds of the curve.

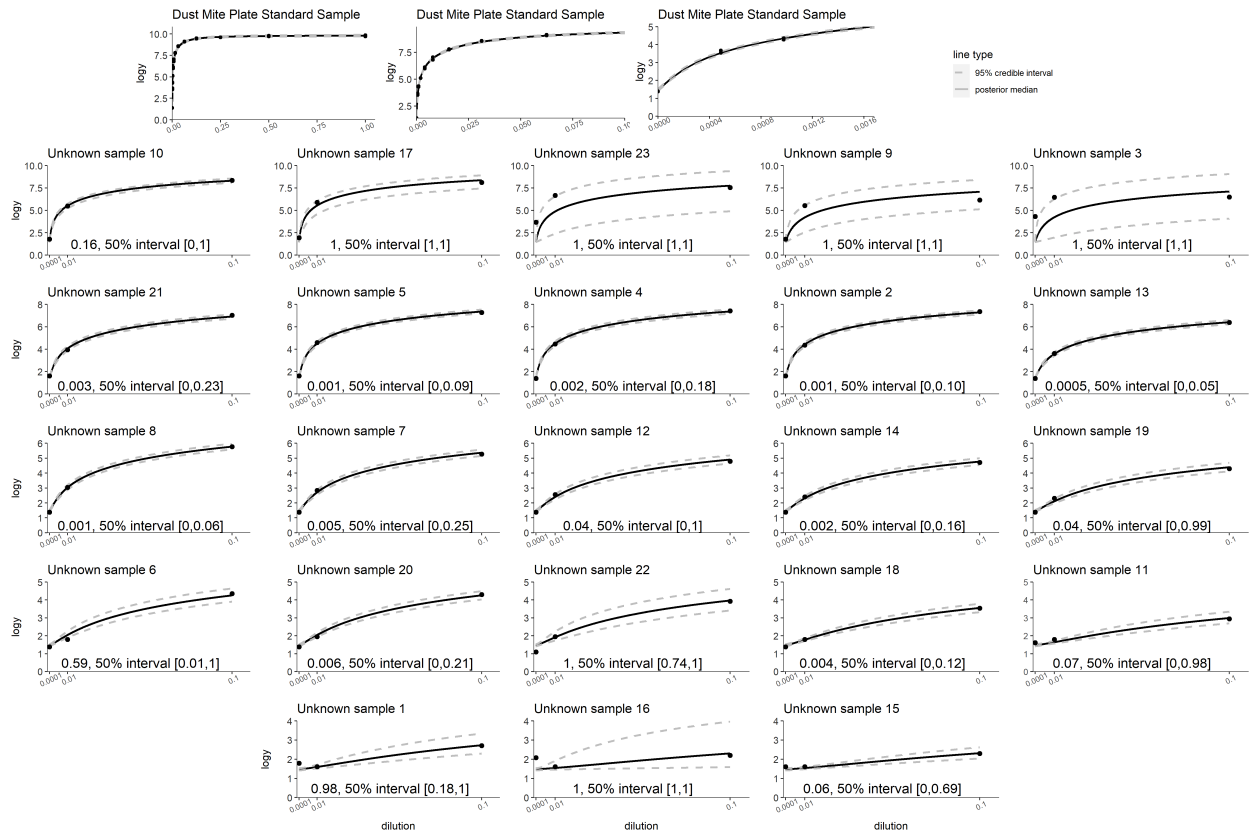


Figure 2.4: Posterior median and 95% probability interval of the mean function dose-response curves for the standards and each new sample estimated using our proposed model. The posterior median and corresponding 50% probability interval for the probability of contamination are listed at the bottom of each plot.

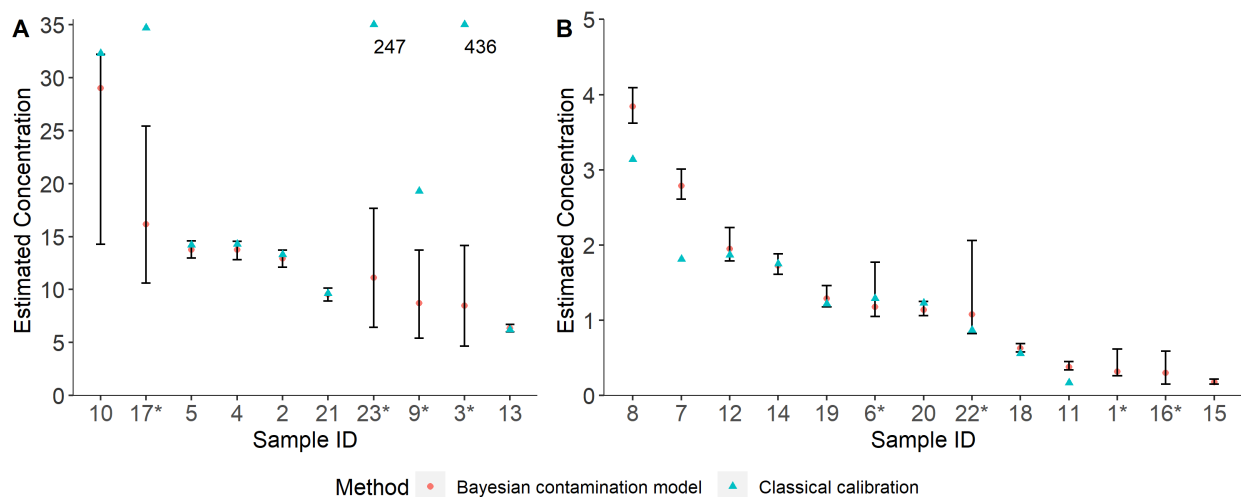


Figure 2.5: Concentration estimation of Der f 1 for all new samples in a single multiplex plate using the classical calibration and the new Bayesian method with 95% probability interval. Samples with \* are estimated to have high posterior probability of contamination.

Figure 2.5 shows the estimated concentrations of the 23 unknown samples, with the posterior median (denoted using dots) and 95% probability interval using the new Bayesian approach as well as the point estimate using the classical calibration method (denoted using triangle). For unknown samples 17, 23, 9, and 3, which have high probabilities of contamination, the Bayesian model estimated a very different concentration than using the classical calibration approach and a wide 95% probability interval associated with the Bayesian estimate. Unknown samples 1, 16, and 15 are classified as below the limit of detection using classical calibration, but our Bayesian model provides estimates with very low concentrations.

#### 2.4.4 Standard data concentration recovery ratio

In real-life lab measurements, experienced lab technicians may use some subjective decision rule to check the sample plate quality. For example, although the concentrations of unknown samples are unknown, on each plate, we have multiple measurements for the standard calibration data with known initial concentration and known dilution factors. When lab technicians apply the traditional method to such plates, they usually generate the observed concentration for each dilution level of the standard calibration sample and compare it to the theoretically expected concentration value at

that dilution level. If the observed concentration has too much deviation from the expected concentration, that observation of the standard calibration sample will be discarded. Following this logic, we compare such deviation between our proposed method and the traditional method and find that the traditional method could not estimate the first several dilutions accurately, mainly due to ignoring the measurement error and the non-linearity of the calibration curve, but our proposed method could estimate it very well for all the dilutions of the standard sample. This further demonstrates the advantage of our method compared to the traditional estimation method commonly used in serial dilution lab experiments.

To be more specific, since the traditional method does not take measurement error into consideration, and given the nonlinear S-shape of the calibration curve generated by the standard sample data only, the measurement error would be amplified greater in the nonlinear parts of the curve, which are the lower and upper parts. So to make a more accurate estimation, the lab technicians would like to use only the middle part of the standard sample calibration curve, which has higher precision of estimation. A direct evaluation metric of such estimation precision in the classical method is whether directly inverting the estimated standard sample calibration curve results in a concentration estimation that is very close to the expected true concentration at that dilution. The reason for only choosing the standard calibration sample recovery ratio is that we only know the concentration of the standard sample, and a usually used range of reasonable recovery ratio is between 0.7 and 1.3, which corresponds to a 30% relative deviation. Although in our proposed Bayesian mixture model, we treat the standard calibration sample concentration as a known value instead of a parameter to be estimated, we could still use the posterior median of the four curve-specific  $\beta$  parameters to reconstruct the mean function of the calibration curve. Then we could directly invert this mean function, and although doing this will ignore the measurement error, it is the most similar way to make the inference results comparable to the recovery ratio defined in the classical method. As shown in Figure 2.6, the classical method, the sample recovery ratio is 1.43, 1.33, 0.79, 1.04, 1.01, 1.03, 0.98, 0.99, 1.00, 0.98, 0.98, 0.98, respectively, corresponding to each of the 2-fold dilution factors. For our Bayesian mixture model, the sample recovery ratio is 0.78,

1.01, 0.78, 1.06, 1.07, 1.10, 1.03, 1.00, 0.98, 0.92, 0.97, 1.06, respectively, corresponding to each of the 2-fold dilution factors. Here we do not consider the zero dilution observation since it will not reflect any useful information regarding the standard calibration sample concentration. Based on the above results and the 30% relative deviation criteria, the classical method will discard the first two observations of the standard calibration sample, which will result in potential information loss, especially for the  $\beta_2$  parameter since it represents the increase to saturation. In comparison, our proposed method will not discard any of the standard calibration sample observations since they all fall within the range. In later dilutions, there might be some subtle differences between the classical method and our method, but they are all in the relatively same scales, so it is not a concern. Also, we should keep in mind that we do not take the measurement error into consideration yet in this recovery ratio calculation, and our proposed Bayesian mixture model already generates results better than the classical method. And in our main analysis part, the measurement error around the calibration curve has always been taken into consideration, which indicates that our proposed Bayesian mixture model actually performs even better in the real-life example. Although we have not used this recovery ratio as a main evaluation criterion, this analysis further increases our confidence in the model performance and also convinces the lab technicians that our proposed method is reliable.

#### 2.4.5 Four-parameter logistic model versus five-parameter logistic model

In this project, for the mean function of the standard data calibration curve, we use the four-parameter logistic model where each of the model parameters has clear scientific interpretations. However, this is not the only possible parametric model that researchers could use in their own research projects. Actually, we do not impose any restrictions on the mean function of the standard data calibration curve, and thus users could choose whatever model they believe is reasonable and valuable. A closely related parametric specification of the mean function is the five-parameter logistic model in which one additional parameter is introduced to account for the symmetry between the lower end and upper end of the curve (Cumberland et al., 2015). Currently, the classical

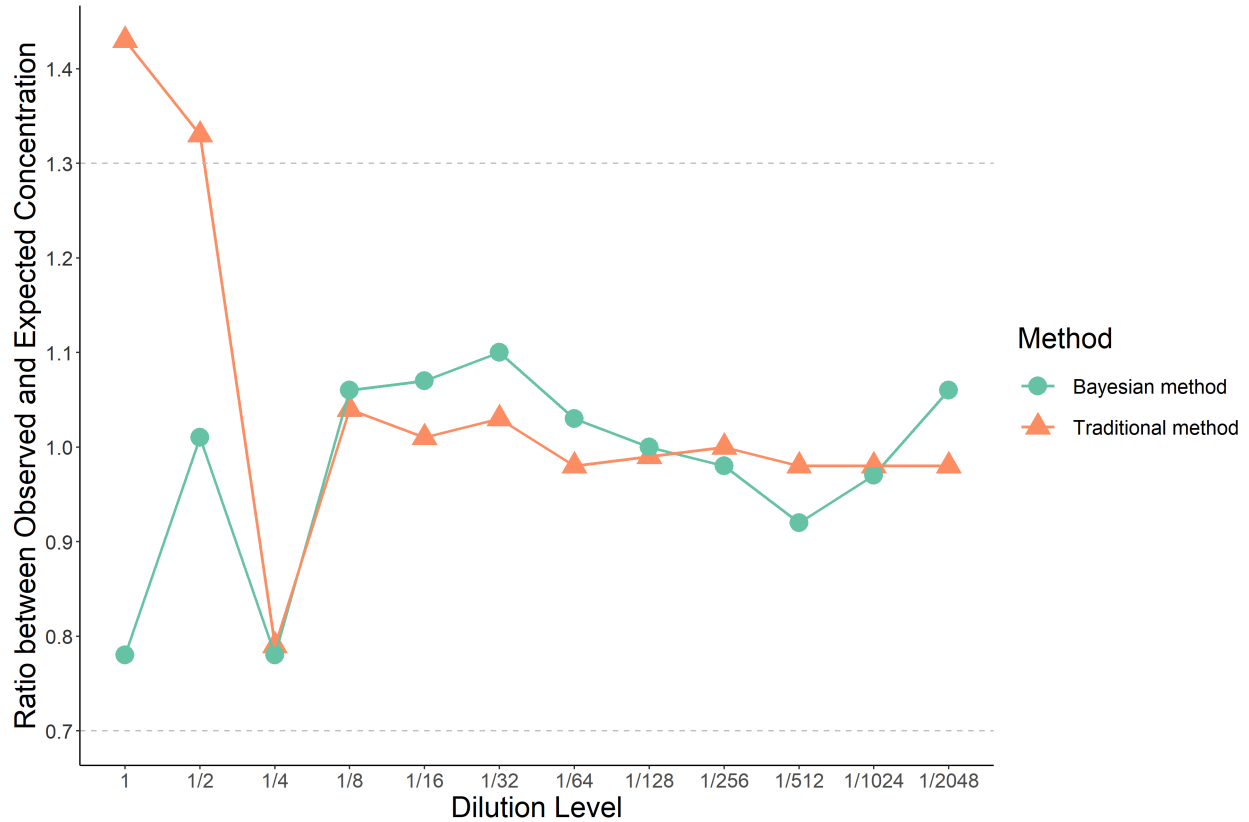


Figure 2.6: Comparison of the standard calibration data concentration recovery ratio between the classical method and the Bayesian contamination model.

estimation method used in the serial dilution assay lab generates the calibration curve parameters estimations and unknown concentrations estimation based on this five-parameter logistic model. But we still use the four-parameter logistic model based on the following reasons: first, we do not directly observe such unsymmetry pattern of the calibration curve for the standard calibration data in the study, especially given the high resolution of the raw measurement based on the new multiplex assay technology; second, even the traditional method uses this five-parameter logistic model, it will only use the inner part of the calibration, which is approximately linear to make the inference, so actually, the advantages of this five-parameter logistic model are not fully reflected even in the traditional method; third, what will bias the unknown concentration estimation most is not the unsymmetry at the lower and upper end of the calibration curve, but instead the main reasons for inaccurate estimation are because of ignoring the measurement error and sample con-

tamination, which have all been addressed in our proposed Bayesian mixture model; finally, we also compare the estimation results based on our method and the estimation results based on the traditional method, and for those samples that are not classified as being contaminated samples, these two estimation methods yield pretty similar results and thus indicates the four-parameter logistic model could also do a good job. Combining these reasons above, the choice of the four-parameter logistic model is a reasonable choice for estimating unknown sample concentrations in the context of serial dilution assay. But also, we should keep in mind that there is no limit on the preferred model, and there is no single model that always has the best performance compared to competing models across various settings, and users have their own flexibility to design whatever inference model they would like to use.

#### 2.4.6 Improved procedure for traditional inference method

The traditional method usually includes two steps for estimating concentrations. The first step is to use the standard sample to estimate the calibration curve, while the second step is to use the inverse of this curve to obtain the unknown sample concentration. Faced with the problems of measurement error being ignored, errors being induced when inverting a nonlinear curve, and contamination occurring, some efforts have been made to improve this traditional method. First, for the standard sample, studies have calculated a ratio of estimated concentrations across different dilutions, only selecting the subset of dilutions that matches this ratio; typically, for the standard sample, this is in the range of 0.7–1.3 of the corresponding dilution. More specifically, although the calibration curve is usually nonlinear, only the intermediate linear part inside is taken to ensure more reliable estimation results. Second, studies have performed a similar procedure for unknown samples, where the 1/10, 1/100, and 1/10,000 dilutions are also matched with the estimation concentrations  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$ . The corresponding diluted observations are only retained if  $\theta_1/\theta_2 = 10$  and  $\theta_2/\theta_3 = 100$ ; otherwise, they are discarded. Third, studies have distinguished between being below the detection limit and sample contamination. For example, if three observations of one unknown sample have some problems, the lab technicians will base on their domain experience

to decide whether this problem is caused by sample contamination (e.g., perhaps a lab technician forgot to apply the correct dilution to certain samples) or being below the detection limit. Fourth, studies have argued that using a five-parameter curve provides more reliable results since it can capture the asymmetries of the calibration curve's two ends.

However, these improved procedures are far from perfect. First, although only using the internal linear part of the calibration curve partially considers the high measurement error at both ends, it discards many observations and thus loses experimental efficacy. Furthermore, for unknown samples, there would typically only be three observations, and simply discarding some of these would be highly inefficient. Moreover, the range of 0.7–1.3 is only based on prior lab experience and might not generalize well to other settings. Second, no clear definition exists for distinguishing between sample contamination and being below the detection limit, and moreover, the results are usually based purely on subjective judgment. Furthermore, despite this distinguishment, the classical method does not provide solutions to the problems of sample contamination and being below the detection limit. Third, a debate persists regarding the four- and five-parameter models, and sometimes a more complicated model might not always perform better. Moreover, since this procedure only uses the inner linear part of the calibration curve to perform the estimation, the advantage of the asymmetries of the two ends of the calibration curve in the five-parameter model might not be particularly obvious.

#### 2.4.7 Sensitivity analysis

To assess the influence of prior information, we have run sensitivity analysis by refitting the model with multiple priors mentioned in Section 2.2.3. To be more specific, we have tried different priors for the unknown concentrations and different priors for the mixture proportion. The model fitting results did not change much, which indicates our model is robust, and even weakly informative priors could contribute to the model inference. In appendix, we show the model fitting results like Figure 2.4 but use stronger exponential priors on the underlying concentration  $\theta_j$ . Except for several unknown samples that are now identified as potentially contaminated, given this stronger

prior, the overall fitting results are robust. In Section 2.6, we propose future working directions to incorporate the idea of the relative importance of each dilution observation to evaluate how much information each data point provides and whether there are subgroups of data that are difficult to fit in the model for a more efficient experimental design.

#### 2.4.8 Model modification and extension

For model modification, we could easily modify our model with more informative priors justified by experienced lab technicians and also have the ability to incorporate more lab data as well as additional information associated with each observation. For example, if we have another layer of information regarding the sample collection process, such as the relative geographical information of each sample, we could extend our model by modeling the concentration of each sample given that additional information. Besides, our contamination model specification is also pretty flexible, and it could be generalized to allow more complicated contamination models if necessary. If in some other studies, there is clear evidence showing that the remaining parameters in the dose-response curve should also vary for the contaminated samples or a different parametric form of contamination exists, we could easily adapt and extend our proposed contamination model to reflect such new information.

We might also present the model estimation result stratified by the contamination factor. In our model output, we combine the two mixture components together based on the mixture proportion and output a final aggregated concentration estimation. However, in some situations, if needed, we could output two stratified concentration estimations, which are concentration under the contamination case and concentration under the uncontaminated case, respectively. This could provide more accurate estimation results if the users genuinely believe that the model contamination classification status results are trustworthy and thus could gain more estimation efficiency. For example, suppose the output posterior median probability of being contaminated for a particular unknown sample is 0.8 with a very tight associated uncertainty interval. In that case, it might make more sense to only take the concentration estimation under the contamination case as the final estimation

result.

In order to better understand our proposed model and address the two immunoassays-specific practical aims, in the following section, we use constructed data to compare the performance of our model with other methods in various scenarios using the ideas of fake-data simulation and simulation-based calibration.

## 2.5 Simulation study

We conduct extensive simulation studies, mimicking the real data in the NYC NAAS plate we examined in Section 2.4. Specifically, we use the same initial concentration for the standards and consider the same levels of dilutions for both the standard and unknown samples, with the  $\beta$  parameters in the four-parameter logistic model taking similar values as those estimated from the real plate:  $\beta_1 = 4$ ,  $\beta_2 = 18000$ ,  $\beta_3 = 8$ ,  $\beta_4 = 1.5$  and  $\sigma_y = 0.1$ . In the simulations presented in Sections 2.5.1 and 2.5.2, we compare the estimates of  $\theta_j$  using the proposed Bayesian contamination model, the base model shown in equation (2.9) which uses Bayesian hierarchical models to handle serial dilution assay estimations (Gelman et al., 2004), and the classical calibration method. We use the root mean square error (RMSE) and coverage probability of the 95% uncertainty interval as the evaluation metrics. Because the classical calibration method only generates point estimate without the associated uncertainty, only RMSE is presented. In addition, we conduct two prior predictive checking simulations to validate our data generative model and to check the robustness of our Bayesian contamination model to potential model misspecification. Both simulations show that our proposed method performs very well. The simulation designs and results are in the supplement.

### 2.5.1 Evaluation of the proposed method under various contamination settings

To evaluate the performance of our proposed method under various contamination settings, we simulate data for new samples considering a wide range of true concentrations and various degrees of deviation from the standard curve among the contaminated samples. In each replicate

of the simulation, we simulate a plate of dilution assay data with 2 replicates of standard sample and 20 unknown samples. For the standard calibration data, we set  $\theta^{\text{standard}} = 125$  and generate 26 observations with 13 dilution levels for each replicate,  $d_i = 0, 1, 1/2, 1/4, \dots, 1/2048$ , based on  $\log(y_i) \sim \text{normal}(\log(g(\theta^{\text{standard}}d_i, \beta)), \sigma_y)$ . Then we fix the initial concentration  $\theta_j$  of 15 uncontaminated unknown samples to be (2.14, 2.17, 3.07, 5.30, 5.72, 5.76, 6.89, 8.14, 10.64, 12.28, 19.37, 20.04, 23.41, 28.62, 32.09), which are generated from the exponential distribution with rate parameter equals to 0.1 and represent reasonable ranges of real-life allergen concentrations. For each uncontaminated unknown sample, we generate three observations with dilution levels  $d_i = 1/10, 1/100, 1/10000$  based on  $\log(y_{ij}) \sim \text{normal}(\log(g(\theta_j d_i, \beta)), \sigma_y)$ . To generate two different degrees of contamination setting, we fix the initial concentration  $\theta_j^{\text{contaminated}}$  of five contaminated unknown samples to be (3.37, 11.70, 12.56, 14.03, 14.79), which are also generated from the exponential distribution with rate parameter equals to 0.1. For each contaminated unknown sample, we generate 3 observations with dilution levels  $d_i = 1/10, 1/100, 1/10000$  based on  $\log(y_{ij}) \sim \text{normal}(\log(g(\theta_j^{\text{contaminated}} d_i, \beta_j)), \sigma_{y_j})$ , where  $\beta_j = (\beta_1, \beta_{2j}, \beta_3, \beta_4)$ . For mild contamination setting, we set  $\beta_{2j} = \beta_2 e^{0.1}$ ,  $\sigma_{y_j} = \sigma_y e^{0.1}$ ; and for more contamination setting, we set  $\beta_{2j} = \beta_2 e^1$ ,  $\sigma_{y_j} = \sigma_y e^1$ . We repeat the simulation 500 times.

Table 2.3 shows that our model outperforms the two competing methods with smaller RMSE and better CP, especially for the case with more contamination. Table 2.4 provides a deeper analysis of our proposed method separated by predicted contamination status, where in each replicate of simulation, the posterior median and the 95% probability interval are calculated among the draws of  $\theta_j$  stratified by the values of corresponding draws of contamination status. The proportion of draws classifying a sample as contaminated is calculated in each replicate of simulation, and is averaged across all replicates and presented as the contamination classification ratio in Table 2.4. By looking at the contamination classification ratio, we could make our estimation even better. The contamination classification ratio is low among all the uncontaminated samples no matter in the less or more contamination cases, where we also observe lower RMSE and closer to the nominal level CP among the draws classified as “uncontaminated” than those classified as “contaminated.”

	Less Contamination Case					More Contamination Case				
	Final Model		Base Model		Classical Calibration RMSE	Final Model		Base Model		Classical Calibration RMSE
	RMSE	CP	RMSE	CP		RMSE	CP	RMSE	CP	
Unk1 (2.14)	0.15	0.97	0.17	0.62	8.44	0.17	0.99	0.19	0.76	8.35
Unk2 (2.17)	0.15	0.98	0.16	0.60	8.63	0.16	0.99	0.19	0.73	8.54
Unk3 (3.07)	0.21	0.98	0.23	0.49	8.66	0.22	0.99	0.26	0.63	8.71
Unk4 (5.30)	0.30	0.98	0.34	0.51	7.50	0.31	1.00	0.36	0.60	7.43
Unk5 (5.72)	0.34	0.98	0.39	0.50	8.09	0.34	0.99	0.41	0.60	8.25
Unk6 (5.76)	0.34	0.99	0.39	0.44	7.86	0.35	1.00	0.43	0.55	7.78
Unk7(6.89)	0.37	0.99	0.45	0.40	8.51	0.37	1.00	0.49	0.51	8.45
Unk8 (8.14)	0.45	0.97	0.56	0.43	8.54	0.46	0.99	0.60	0.51	8.55
Unk9 (10.64)	0.61	0.97	0.73	0.37	7.42	0.62	0.99	0.77	0.46	7.40
Unk10 (12.28)	0.64	0.98	0.81	0.36	7.73	0.64	1.00	0.87	0.41	7.64
Unk11 (19.37)	1.09	0.97	1.34	0.29	7.02	1.09	0.99	1.41	0.33	7.17
Unk12 (20.04)	1.06	0.98	1.34	0.31	8.06	1.05	1.00	1.43	0.34	8.13
Unk13 (23.41)	1.32	0.98	1.71	0.28	6.60	1.33	0.99	1.81	0.29	6.62
Unk14 (28.62)	1.57	0.97	2.13	0.24	7.32	1.57	0.99	2.26	0.26	7.47
Unk15 (32.09)	1.81	0.98	2.43	0.23	7.26	1.82	0.99	2.60	0.24	7.23
<b>Unk16 (3.37)</b>	0.35	0.88	0.36	0.37	9.67	3.87	0.57	3.99	0	20.65
<b>Unk17 (11.70)</b>	1.13	0.86	1.24	0.21	8.96	11.15	0.71	14.44	0	26.56
<b>Unk18 (12.56)</b>	1.24	0.85	1.40	0.22	9.12	11.82	0.71	14.92	0	29.86
<b>Unk19 (14.03)</b>	1.31	0.85	1.46	0.21	8.51	12.35	0.75	16.69	0	29.62
<b>Unk20 (14.79)</b>	1.40	0.88	1.58	0.23	8.85	13.24	0.74	18.81	0	31.79

Table 2.3: Comparison between different methods in terms of root mean squared error (RMSE) and coverage probability (CP) of 95% probability interval for each unknown sample in two different contamination settings. Unk1-15 are uncontaminated samples and Unk16-20 are contaminated samples.

In contrast, the contamination classification ratio is much higher among the contaminated samples in the more contamination case, where the estimate of  $\theta_j$  and its 95% uncertainty interval has smaller RMSE and close to the nominal level CP using draws classified as “contaminated” but has large RMSE and poor CP among draws classified as “uncontaminated.” The contamination classification ratio is low, and the estimates of  $\theta_j$  based on draws classified as “uncontaminated” yield a smaller RMSE than those classified as “contaminated” among the contaminated samples in the less contamination case. This suggests that it could be more beneficial to classify a contaminated sample as uncontaminated if the contamination level is mild and our model is flexible enough to accommodate that. This also demonstrates that our robust Bayesian mixture model performs well in the existence of potential contamination.

We have three observations from this simulation. First, our final model can effectively flag contaminated samples with moderate to high level of contamination. Second, when the contamination ratio is small, the estimates based on draws classified as “uncontaminated” perform similarly as

	Less Contamination Case					More Contamination Case				
	Uncontaminated		Contaminated		Ratio	Uncontaminated		Contaminated		Ratio
	RMSE	CP	RMSE	CP		RMSE	CP	RMSE	CP	
Unk1 (2.14)	0.15	0.95	1.24	1.00	0.04	0.16	0.97	1.04	1.00	0.09
Unk2 (2.17)	0.15	0.96	1.25	0.99	0.04	0.16	0.97	1.06	1.00	0.09
Unk3 (3.07)	0.21	0.95	1.53	0.99	0.04	0.22	0.97	1.38	1.00	0.09
Unk4 (5.30)	0.30	0.97	2.18	1.00	0.03	0.31	0.99	1.81	1.00	0.08
Unk5 (5.72)	0.34	0.96	2.20	1.00	0.04	0.34	0.97	2.00	1.00	0.08
Unk6 (5.76)	0.34	0.96	2.39	1.00	0.04	0.34	0.98	1.92	1.00	0.08
Unk7(6.89)	0.37	0.97	2.52	0.99	0.03	0.37	0.98	2.04	1.00	0.08
Unk8 (8.14)	0.45	0.96	2.85	1.00	0.03	0.46	0.97	2.17	1.00	0.08
Unk9 (10.64)	0.61	0.95	3.04	1.00	0.03	0.61	0.97	2.48	1.00	0.08
Unk10 (12.28)	0.64	0.97	3.64	0.99	0.03	0.64	0.98	2.70	1.00	0.08
Unk11 (19.37)	1.09	0.96	5.20	0.99	0.03	1.09	0.97	3.95	1.00	0.07
Unk12 (20.04)	1.06	0.96	5.28	0.99	0.03	1.05	0.98	4.14	1.00	0.07
Unk13 (23.41)	1.30	0.96	6.30	0.99	0.03	1.31	0.98	5.65	1.00	0.07
Unk14 (28.62)	1.57	0.95	8.59	0.99	0.03	1.58	0.98	7.71	1.00	0.07
Unk15 (32.09)	1.81	0.97	9.73	0.97	0.03	1.80	0.98	9.17	1.00	0.07
<b>Unk16 (3.37)</b>	0.35	0.82	1.90	0.99	0.05	3.93	0.13	5.08	0.92	0.41
<b>Unk17 (11.70)</b>	1.13	0.77	3.63	0.98	0.04	12.41	0.13	8.52	0.97	0.39
<b>Unk18 (12.56)</b>	1.24	0.78	3.74	0.99	0.04	12.24	0.13	8.90	0.98	0.38
<b>Unk19 (14.03)</b>	1.32	0.79	4.17	0.98	0.04	14.39	0.13	9.12	0.98	0.44
<b>Unk20 (14.79)</b>	1.40	0.80	4.30	0.99	0.04	15.70	0.13	8.13	0.98	0.43

Table 2.4: Root mean squared error (RMSE) and coverage probability (CP) of 95% probability interval of our proposed method stratified by the estimated contamination status in posterior draws for each unknown sample. Unk1-15 are uncontaminated samples and Unk16-20 are contaminated samples. Ratio: proportion of posterior draws being classified as contaminated samples.

those based on all draws in Table 2.3, thus it’s safe to report the overall estimate as in Table 2.3 when the contamination classification ratio is relatively small. Third, when the contamination ratio is relatively large, the estimates based on draws classified as “contaminated” have much improved performance compared to the corresponding estimates using all draws, and thus it is beneficial to report the stratified estimates using the subset of draws classified as “contaminated.”

### 2.5.2 Extreme case

To test the performance of our method in a more extreme case, we follow the same simulation setup in the more contamination case in Section 2.5.1, but fix the initial concentration  $\theta_j$  of the 15 uncontaminated unknown samples to be (0, 0, 0, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16) and the initial concentration  $\theta_j^{\text{contaminated}}$  of the 5 contaminated unknown samples to be (0, 1/16, 1/4, 2, 8). The values of these initial concentrations represent an extreme case in which most of the allergen concentrations are quite small except for one or two unknown samples having

relatively large allergen concentrations.

	All	Proposed Bayesian Model		Ratio	Base model	Classical method
		Uncontaminated	Contaminated		RMSE	RMSE
Unk1 (0)	0.059	0.056	0.091	0.133	0.020	OOOR<
Unk2 (0)	0.059	0.057	0.091	0.136	0.019	OOOR<
Unk3 (0)	0.058	0.056	0.090	0.137	0.017	OOOR<
Unk4 (1/128)	0.052	0.050	0.085	0.126	0.017	OOOR<
Unk5 (1/64)	0.047	0.045	0.078	0.129	0.016	OOOR<
Unk6 (1/32)	0.041	0.039	0.074	0.127	0.024	OOOR<
Unk7(1/16)	0.035	0.034	0.066	0.118	0.047	OOOR<
Unk8 (1/8)	0.041	0.040	0.085	0.122	0.072	12.847
Unk9 (1/4)	0.044	0.043	0.137	0.128	0.058	11.616
Unk10 (1/2)	0.053	0.052	0.233	0.120	0.061	11.300
Unk11 (1)	0.085	0.084	0.377	0.112	0.100	10.108
Unk12 (2)	0.145	0.144	0.478	0.104	0.165	10.872
Unk13 (4)	0.250	0.249	0.611	0.095	0.284	8.692
Unk14 (8)	0.448	0.442	2.274	0.101	0.562	9.361
Unk15 (16)	1.055	0.856	7.769	0.182	1.097	8.474
<b>Unk16 (0)</b>	0.135	0.097	0.162	0.562	0.070	OOOR<
<b>Unk17 (1/16)</b>	0.159	0.124	0.202	0.473	0.110	OOOR<
<b>Unk18 (1/4)</b>	0.355	0.288	0.486	0.458	0.270	24.597
<b>Unk19 (2)</b>	1.874	2.006	1.759	0.465	2.028	20.198
<b>Unk20 (8)</b>	5.600	8.092	0.871	0.540	8.895	26.330

Table 2.5: Summary of root mean squared error (RMSE) across 15 uncontaminated samples (Unk1-15) and 5 contaminated samples (Unk16-20) for severe contamination case in the extreme case simulation scenario. Ratio: proportion of posterior draws being classified as contaminated samples.

We start the modeling by assuming  $\theta_j \sim \text{exponential}(\mu)$  and  $\mu \sim \text{normal}^+(0, 0.1)$  as in equation (2.7). We then apply several alternative priors for  $\theta_j$ , including modeling the exponential rate parameter for  $\theta_j$  as  $\mu \sim \text{normal}^+(0, 2.5)$ ; using t distribution  $\log(\theta_j) \sim t_\nu(\mu, \sigma)$  with diffuse hyper-parameters; using skewed-normal distribution  $\theta_j \sim \text{skewed-normal}(\xi, \omega, \alpha)$  with diffuse hyperparameters; and using the continuous Horseshoe model for  $\theta_j$ . After comparison, we find that the performance of the exponential prior with  $\mu \sim \text{normal}^+(0, 0.1)$  is most robust in RMSE in this extreme scenario. Table 2.5 shows the RMSE of our proposed model (using all draws and stratified by the classified contaminated status), the base model proposed by (Gelman et al., 2004), and the classical calibration method. Overall, our final model outperforms its competitors in RMSE except for the samples with very low concentrations. The less ideal performance among these extremely low concentration samples is expected because our model uses diffuse and flexible priors and lets the data tell the model what to do. If there is one sample with a much higher concentration than the others, the model will naturally put more weight on that observation and

thus could perform poorly at the extremely low end. If the research really cares about the samples with very low concentrations, alternative prior could be used. The classical calibration method fails to estimate those low concentrations and claims them as below the limit of detection.

## **2.6 Discussion**

Statistical inference under potential model contamination is an essential topic in both the studies of immunoassays and other general applied statistics problems. This work proposes a general Bayesian workflow for inference under contamination and applies it to serial dilution assay data analysis. After applying our proposed model, we could produce reasonable estimates for those potentially contaminated samples accounting for high variability across different dilutions due to contamination as well as undercover those unknown samples' concentration, which cannot be measured by the classical calibration approach. Also, our framework could generate signals for researchers to further investigate those potential contamination samples for better scientific analysis results. Empirical and simulation studies demonstrate the advantages of our Bayesian contamination model among other competing methods.

Another contribution is that it introduces Bayesian workflow for inference under contamination. We first state the general steps for Bayesian workflow for model contamination and immunoassay data-specific aims for the workflow. Then we use both the applied example and simulation studies to show how each step would be carried out in an applied statistics problem. This highlights essential steps in conducting Bayesian inference in broader applied topics and would facilitate users with different backgrounds to quickly adapt our work to their studies.

Our studies face a few limitations. First, our contamination model only assumes a subset of the dose-mean response curve parameters to be different between the calibration sample and the potentially contaminated samples based on real lab data insights. In our data, the contaminated samples are different from the calibration sample most in terms of the saturation level and rate of saturation occurrence. This might not be comprehensive enough for all types of contamination happening in the sample collection and measurement process. Second, although our general workflow for model

contamination could identify the potentially contaminated samples in both real data and various simulation settings well, it will be more interesting to provide further insights into the reasons for contamination. For example, in serial dilution assay data, we could incorporate the relative spatial information and the experimental conditions for collected dust samples to gain insights into the potential causes of contamination. In addition, we could develop and refine our user-specific model using the feedback from lab technicians to better model the contamination in serial dilution assay.

## **Chapter 3: Bayesian Joint Modeling of Exposure and Outcome with Uncertainty in Exposure Measure**

### **3.1 Introduction and background**

#### 3.1.1 Local and global statistical calibration

In most statistical calibration problems, researchers are interested in measuring unknown quantities based on evidence from calibration experiments. This process typically involves the inverse prediction problem. In the first step, researchers estimate the functional form of the calibration curve, while in the second step, they infer estimates of the unknowns by directly inverting the calibration curve (Osborne, 1991). As in the bias–variance trade-off in statistics, a similar trade-off always exists in a lab measurement setting between local and global calibration. Local calibration adjusts for more factors and does not require additional efforts to build models, but it is also costly. Global calibration is usually more statistically efficient but requires more effort in modeling. Furthermore, global calibration can be beneficial since it provides additional information. In this chapter, we study this general topic in the context of applying Bayesian local and global calibration models for serial dilution assay data.

A serial dilution assay is one of the technologies commonly used to estimate allergen concentration in a lab setting. In most applied datasets, the researcher would usually conduct multiple experimental units to account for the relatively large sample size. The local calibration method means that the researcher treats each of the experimental units as independent samples and conducts unit-level modeling and calibration based only on the data available on that specific experimental unit. This approach is relatively straightforward since many models for single-unit serial dilution assay estimation and calibration models have already been developed; thus, researchers only need

to apply those methods many times repeatedly before finally bringing all of the estimation results together. However, drawbacks exist to this approach, namely that the repeated work might introduce additional errors while ignoring information from other available experimental units might lead to losses of estimation efficacy and statistical power. By contrast, global calibration simultaneously pools together all available experimental units and generates estimation results through a single joint model. The main advantage of this is that it can solve the problem in one go, rendering repetitive work unnecessary. Furthermore, it can make the most effective use of the available information in the data, thus making the information more efficient. However, researchers also face challenges regarding a more complicated model structure as well as the potential risk of model misspecification and overfitting.

We used the following steps to study and address the challenges of Bayesian global calibration models. First, we used a multilayer hierarchical model so that each primary subgroup level (e.g., plate-level data) would have different calibration parameters and each secondary subgroup level (e.g., sample-level data) could be contaminated. Second, we jointly applied the global calibration model together with the epidemiology association model; thus, we were able to more effectively plug estimation uncertainty into our key predictor of interest and derive more reliable results. Third, we extended the global calibration inference results for a more efficient lab experiment design using Bayesian partial pooling.

### 3.1.2 Background to the multiple Multiplex plates experiment setting

Allergen concentration estimation is usually the first step in indoor epidemiology and environmental science studies focused on the association between specific allergens and long-term diseases. The Multiplex Array for Indoor Allergens (MARIA) is a commonly used tool in large-scale epidemiological association studies on indoor allergens, which typically involve measurements of multiple multiplex plates using a serial dilution assay. Compared with another commonly used technology in this field, namely the enzyme-linked immunosorbent assay (ELISA), MARIA is much less labor-intensive and time-consuming (Earle et al., 2007). Compared with ELISA,

MARIA has the following two major advantages: First, it provides a much higher resolution, and second, it can estimate multiple allergens of interest within a single experimental plate, whereas ELISA can only estimate one specific allergen of interest within a single experimental plate (Dize et al., 2018).

Although MARIA can estimate multiple allergens simultaneously, the allowable sample size in each plate is still constrained. In many epidemiology and environmental science cohort studies, the sample size is usually large, and multiple MARIA plates are usually required to be performed for the complete estimation of corresponding allergen concentrations of all participants. This creates additional challenges for researchers due to the heteroskedasticity among plates across different experimental conditions. Commonly, some subtle changes in lab conditions, such as temperature and humidity, result in large changes in plate-specific calibration parameters and estimation accuracy (Tandon et al., 2009; King et al., 2013). This complexity of plate heteroskedasticity causes additional challenges for reproducible lab measurements. In traditional calibration methods, researchers first use local calibration to estimate the plate-specific calibration curve plate by plate, and for each plate; then, the calibration curve is inverted to obtain the estimated concentration of unknown samples. During this local calibration process, additional information is sometimes associated with each plate to reflect the relative quality of the measurements. However, this process is not only time-consuming but could also amplify human operational error, and furthermore, it usually cannot provide valid inference results for poor-quality plates.

### 3.1.3 Introduction to the NYC NAAS dataset

Recent studies have provided evidence that exposure to indoor allergens in early life is an important predictor of asthma development later in life. Given that the majority of children diagnosed with asthma are sensitized to at least one indoor allergen, it is crucial to reduce or even avoid exposure to indoor allergens to reduce asthma morbidity among children (Sheehan and Phipatanakul, 2016).

In this project, we used data from the New York City (NYC) Neighborhood Asthma and Allergy Study (NAAS), which is a study of 7- and 8-year-old children with and without asthma in NYC.

The homes of participating families were visited, during which a detailed questionnaire on the child's health history, environmental exposures, and socioeconomic and demographic information was administered. Children were classified as having asthma on the basis of whether the parent reported at least one of the following for the child in the 12 months before the questionnaire was administered: (1) wheezing; (2) being woken at night by cough without having a cold; (3) wheezing with exercise; or (4) medication use for asthma. Details of the study have been published elsewhere (Perzanowski et al., 2008; Chen et al., 2016).

During the home visit, a dust sample was collected from the child's bed by vacuuming the fitted sheet on the upper half of the bed and both sides of the pillows. The bed dust samples were extracted with phosphate buffered saline with 0.05% Tween (pH 7.4) at a concentration of 50 mg/mL and stored at  $-20$  degrees until analysis. The cockroach allergen (Bla g 2), cat allergen (Fel d 1), household dust mite allergen (Der f 1 and Der p 1; mite group 2), dog allergen (Can f 1), mouse allergen (Mus m 1), and rat allergen (Rat n 1) were measured using immunoassays with multiplex plates. In this study, we analyzed the data across 17 multiplex plates used in the NAAS to examine the associations between indoor allergens concentration, allergy sensitization, and asthma morbidity.

#### 3.1.4 Potential pitfalls of current methods for allergen estimation and regression-based association studies

Currently, the most commonly used traditional calibration inference method comprises three separate steps. First, the lab technician picks up one MARIA plate and estimates the plate-specific calibration curves using the standard samples only. Second, the calibration curve is inverted to match certain observed levels of unknown sample concentrations without accounting for measurement errors. Lastly, the first two steps are repeated for all plates available in the study. This method is simple and makes sense since one knows that, across different lab experiment conditions, there might be some plate-specific effects, and thus, the calibration curves would not be the same across plates.

However, this method is far from perfect for the following reasons: First, the calibration curve is typically nonlinear, and directly inverting it would inevitably introduce numerical errors. Second, ignoring measurement errors, especially at the very low and high ends of the calibration curve, would lead to the detection limit problem, which typically results in throwing away a certain proportion of samples. Third, this method does not consider the potential existence of contamination in unknown samples, where the calibration curve for the unknown contaminated sample might be different from that for the standard sample; thus, the basis of such a calibration process would be violated. The situation becomes even more complicated when potentially different degrees of contamination exist across multiple plates due to different experimental conditions. Fourth, the traditional method can only output a point estimation without any corresponding uncertainty measurement. Recent studies have addressed the first three pitfalls with the help of the Bayesian hierarchical calibration framework, including the method we have developed in our first project and the base model (Gelman et al., 2004).

In epidemiologic association studies, regression-based methods are commonly used because of their simplicity and interpretability. However, in most cases, researchers usually assume a fixed-x regression design, which means that the predictors in the regression model are fixed and known and only the response variable is random. However, this is not true in the current setting. The reason for this is that the key predictor in the regression model, namely indoor allergen concentration, is itself unknown and random. The existence of random covariates brings additional sources of prediction errors, where randomness in the covariates contributes to both bias and variance (Rosset and Tibshirani, 2020). Submodel selection in such a random covariate regression design requires additional care, and numerical methods such as cross-validation and bootstrapping are typically recommended (Breiman and Spector, 1992). Furthermore, evidence suggests that simply ignoring such a random design will lead to false confidence results and discoveries in some applied statistical problems (Behney, 2020; Bartlett and Keogh, 2018). Therefore, we needed to keep the estimation uncertainty and measurement error in the covariates in our epidemiology regression model in mind. This also highlights the possible advantage of using global calibration to plug such measurement

uncertainties into the key predictor in the regression model.

In the statistics literature, the idea of building a joint model to incorporate multiple data sources and achieve more complicated estimation goals simultaneously has been well-studied. One of the most studied fields of joint modeling involves the combination of longitudinal data with time-to-event data, where the distribution of the longitudinal data and event time is assumed to depend on a common set of latent random effects (Tsiatis and Davidian, 2004; Hickey et al., 2016). Given the latent random variable features, the Bayesian method has been demonstrated to perform relatively well on the joint modeling of survival and longitudinal data, which is explained by its flexibility in specifying the joint distribution, association structure, and latent variable distribution (Baghfalaki et al., 2014). Both Bayesian univariate and multivariate joint models based on linear mixed-effect models with proportional hazard assumptions are more data-driven and could decrease estimation bias as well as increase inference efficacy (Alsefri et al., 2020). Furthermore, Bayesian model averaging was demonstrated to be able to increase dynamic prediction accuracy through the consideration of joint models with different association structures through subject- and time-dependent weight adjustment (Rizopoulos et al., 2014). On the other hand, Bayesian regression methods have become a powerful and popular analytical tool in public health research, especially in the field of asthma prevention and intervention studies. One study found that Bayesian logistic regression with priors obtained through a meta-analysis resulted in shorter credible intervals and identified a positive association between NO<sub>2</sub> and lower respiratory symptoms (van Zoest et al., 2020). Moreover, Bayesian regression models with temporal random effects revealed a significant association between asthma hospitalization rate and several common air pollutants (Delamater et al., 2012). Furthermore, the application of Bayesian spatial regression models could assist in investigating and identifying patterns of asthma outcomes and their relationship with the physical environment (Ouédraogo et al., 2018). Given the success of the Bayesian approach in joint modeling with different types of data but similar goals, including in the field of asthma-related research, we sought to investigate whether joint modeling could also be helpful in estimating disease–exposure associations based on original lab measurement data and individual-level public health data.

### 3.1.5 Real-life evidence for Bayesian joint modeling for multiple measurement units

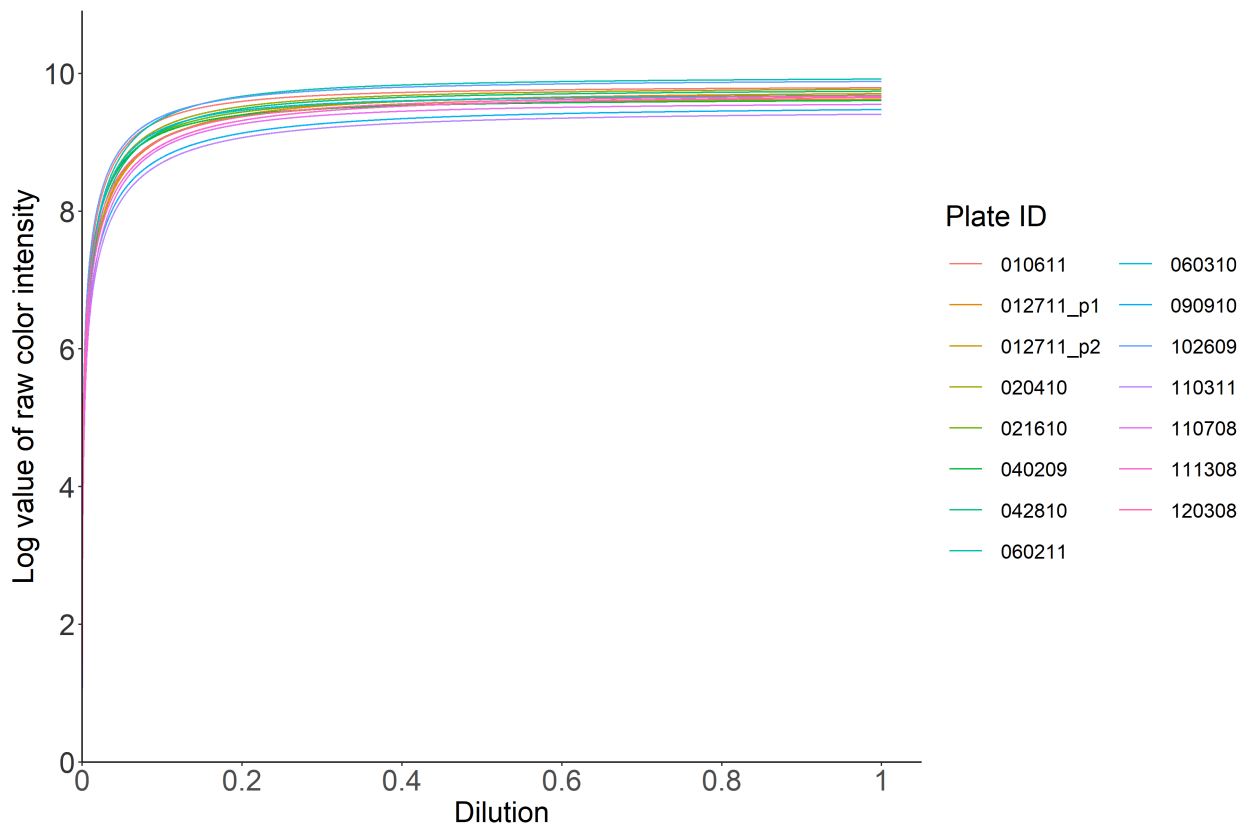


Figure 3.1: Comparison of multiple independent fitting standard data calibration curves for each measurement unit.

In this subsection, we will use a real-life example to demonstrate the fact that jointly modeling multiple measurement units is feasible and has some advantages in some real-world applications. Here we focus on the raw measurement units collected in the NYC NAAS study. Since the capacity of each measurement unit is limited and the NYC NAAS study is a large-scale epidemiological study involving a large cohort of participants, multiple measurement units have been applied to get allergen concentration estimation across multiple time points and under slightly different lab experiment background conditions. We have applied the Bayesian contamination model that is proposed in our first project and fitted the model independently to each of the measurement units

and aggregated together all the standard data calibration curves together in Figure 3.1. The unit-level standard data calibration curve will reflect some of the unit-specific characteristics such as experiment time, temperature, and humidity. From Figure 3.1 we have the following observations: First, there is some difference between these calibration curve parameters, majorly due to different lab experiment conditions and random measurement noise; Second, there is no huge deviance between the calibration curves across those measurement units in the NYC NAAS study. These findings motivate us to develop some Bayesian joint modeling techniques that could pool together all the measurement units in the study to gain more data estimation efficacy while also allowing each measurement unit to have its subject-specific parameters.

### 3.1.6 Our contributions

In this study, we proposed a Bayesian joint hierarchical model by incorporating global calibration across multiple experimental units and embedding the estimation together with epidemiological association studies within the Bayesian framework. The goal was to use the user's time more efficiently by fitting one model for all available data as well as incorporating estimation uncertainty into the key predictor in epidemiologic association studies using Bayesian regression. Furthermore, the traditional local and proposed global calibration inference methods would provide insights into future optimized experimental designs.

## 3.2 Methods

### 3.2.1 Problem setup

We model the observed data from multiple experiments across multiple MARIA plates. In this setting, assume that we have access to the raw experimental data, and let  $x$  represent the single target of interest (e.g., a single indoor allergen concentration) and  $w$  be the output read from the lab machine, which is usually measured in terms of color intensity for the antibody–antigen interaction. On each MARIA plate  $k$ , we have two types of samples. The first set of samples is called the standard calibration sample, which has known concentration  $\theta_{0k}$  and is used to generate

the estimated calibration curve between  $w$  and  $x$ . The serial dilution procedure for the standard calibration sample involves applying multiple known and fixed dilutions to obtain multiple measurements for more accurate curve estimation. The second set of samples is called the unknown sample, which involves multiple groups of collected samples (assume we have  $j$  groups in total on each of the MARIA plates) with unknown true concentration  $\theta_{jk}$ . Similarly, multiple known and fixed dilutions are applied to the unknown samples for multiple measurement chances. The process of serial dilution is summarized as follows:

$$x_{ijk} = x_{init,k} d_{ik}, \quad (3.1)$$

where  $x_{init,k}$  equals  $\theta_{0k}$  for standard sample and  $\theta_{jk}$  for unknown sample  $j$ , and  $d_{ik}$  is the dilution level of observation  $i$  on plate  $k$ .

Next, we wish to model the mean dose–response calibration curve to measure the functional relationship between the observed measurement  $w$  and the underlying target of interest  $x$ . Some studies have demonstrated that after log transformation for the observed measurement  $w$ , the resulting model would have the effect of variance regularization in the corresponding measurement error model (Feng et al., 2011). Therefore, we model the log-transformed  $w$  instead of  $w$  as a function of  $x$  in the serial dilution assays. To allow a smooth and increasing dose–response relationship between  $\log(w)$  and  $x$ , we consider the following four-parameter logistic curve:

$$E(\log(w)|x, \boldsymbol{\beta}) = \log(g(x, \boldsymbol{\beta})) = \log\left(\beta_1 + \frac{\beta_2}{1 + (x/\beta_3)^{-\beta_4}}\right), \quad (3.2)$$

There are two reasons for the abovementioned model choice. First, it has a relatively simple parametric form and provides enough smoothness and flexibility regarding the lower and upper ends of the curve. Second, it ensures a clear interpretation of each model parameter; specifically, in the model,  $\beta_1$  represents the color intensity at zero concentration,  $\beta_2$  represents the increase to saturation,  $\beta_3$  represents the concentration where the curve turns, and  $\beta_4$  represents the rate at which saturation occurs (Higgins et al., 1998).

In the following, we present the models for standard calibration samples and unknown samples separately on each plate. Since the standard calibration samples are usually of a high quality with a known concentration, the model is fairly similar to well-studied forms in the literature and what we have used in the first project. Suppose that  $w_{i0k}$  is the observation from the standard calibration sample with a known initial concentration  $\theta_{0k}$  and a known dilution level  $d_{ik}$  on plate  $k$ ; thus, we assume the following lognormal measurement error model:

$$\log(w_{i0k}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_{0k}d_{ik}, \boldsymbol{\beta}_k)), \sigma_{wk}). \quad (3.3)$$

where  $\boldsymbol{\beta}_k$  and  $\sigma_{wk}$  represent plate-specific calibration curve parameters and measurement error scales, which align well with the plate heteroskedasticity.

### 3.2.2 Bayesian joint contamination model for multiple Multiplex plates in epidemiological association studies

In this subsection, we present the Bayesian joint model for unknown samples, which comprises the Bayesian hierarchical contamination model for jointly modeling multiple plate data as well as the Bayesian regression model for epidemiological studies. This joint model enables us to directly plug the uncertainty of estimating the allergen concentrations into the Bayesian regression model.

Contamination in a serial dilution assay is generally defined as the dose–response curves or the calibration curves of some of the unknown samples differing from the dose–response curve of the standard calibration sample. To model with multiple plates, an additional layer must be added to the hierarchical model to allow some differences in parameters across plates while also allowing them to share some common hyperparameters. We model each unknown sample on each plate  $k$  using a mixture of two normal measurement error models in the log scale with plate-specific parameters. For observations  $w_{ijk}$  from sample  $j$  with known dilution level  $d_{ik}$  but unknown

concentration  $\theta_{jk}$  on plate  $k$ , we have the following:

$$\log(w_{ijk}) \stackrel{ind}{\sim} (1-\lambda_k)*\text{normal}(\log(g(\theta_{jk}d_{ijk}, \boldsymbol{\beta}_k)), \sigma_{wk})+\lambda_k*\text{normal}(\log(g(\theta_{jk}d_{ijk}, \boldsymbol{\beta}_{jk})), \sigma_{wjk}), \quad (3.4)$$

where

$$\boldsymbol{\beta}_{jk} = (\beta_{1k}, \beta_{2jk}, \beta_{3k}, \beta_{4k}), \beta_{2jk} = \beta_{2k}e^{\delta\beta_{2jk}}, \sigma_{wjk} = \sigma_w e^{\delta\sigma_{wk}}. \quad (3.5)$$

The first mixture component models the unknown sample using the same measurement error model as the standard sample on plate  $k$ . The second mixture component allows the unknown sample to have different parameters  $\beta_{2k}$  and  $\sigma_{wk}$  in the measurement error model from those for the standard sample on plate  $k$ . This form of the model specification is based on empirical lab findings that can capture the real data highly accurately.

For the hierarchical modeling of multiple plates, based on prior lab experience, the four parameter logistic curves usually differ across plates because of different lab experimental conditions, such as temperature, humidity, and experiment time. To account for such differences, we use the following independence models for  $\boldsymbol{\beta}_k$ :

$$\beta_{1k} \sim \text{normal}^+(\beta_1, \sigma_1). \quad (3.6)$$

$$\beta_{2k} \sim \text{normal}^+(\beta_2, \sigma_2). \quad (3.7)$$

$$\beta_{3k} \sim \text{normal}^+(\beta_3, \sigma_3). \quad (3.8)$$

$$\beta_{4k} \sim \text{normal}^+(\beta_4, \sigma_4). \quad (3.9)$$

Here, we let each of the beta coefficients vary across plates independently with the remaining beta coefficients. Each of them follows a normal distribution with a fixed mean and variance corresponding to their own scales. We discuss more prior choices and selections in later subsections.

Following the building of the Bayesian hierarchical contamination model, the second component of the joint modeling is the incorporation of the Bayesian regression model into the epidemi-

ologic studies. The major advantage of joint modeling is that we can directly take advantage of the estimated uncertainty of concentration estimation and plug it into the Bayesian regression model. By doing so, we can account for uncertainties in both the key predictors and the response variable in the model.

Suppose that  $y$  is our response of interest in the epidemiologic study; for example, it could be a continuous variable that reflects the health-related outcome of children or a binary variable that reflects their asthma severity. To explore the association between our response of interest  $y$  and the main predictor  $\theta$ , namely the indoor allergen concentration, we could use Bayesian regression models. Suppose that  $\mathbf{Z} = (z_1, z_2, \dots, z_m)$  represents the vectors of measured background variables (or potential confounders) associated with each participant in the epidemiologic study; then, the following Bayesian regression models could be applied:

$$y_n = \alpha + \gamma\theta_n + \boldsymbol{\eta}\mathbf{Z}_n + \epsilon_n, \quad (3.10)$$

if  $y$  is a continuous variable; and

$$Pr(y_n = 1) = \text{logit}^{-1}(\alpha + \gamma\theta_n + \boldsymbol{\eta}\mathbf{Z}_n). \quad (3.11)$$

if  $y$  is a binary variable.

There are several advantages to applying Bayesian regression here. First, with the help of the Bayesian computation software Stan, we could organically integrate the allergen concentration estimation modeling and regression modeling together and take full advantage of measurement uncertainty – both in the response variable and the key predictor – to more accurately estimate the association. Second, compared with classical regression, Bayesian regression is more flexible and could be easily generalized to incorporate prior beliefs on certain regression parameters or internal hierarchical structures of the data.

### 3.2.3 Prior specification

For the prior specification of allergen concentration, since based on their experiment-related interpretations all four of the  $\beta_k$  parameters must be positive, we consider the following weakly informative priors for their common mean vector  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ :

$$\log(\boldsymbol{\beta}) \sim \text{normal}(0, 10), \quad (3.12)$$

For the variance vector  $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ , we assign constant values proportional to the relative scale of roughly raw estimated values of  $\boldsymbol{\beta}$  from the calibration sample. For the measurement error model of the standard calibration sample on each plate, we have

$$\sigma_{wk} \sim \text{normal}^+(0, 10), \quad (3.13)$$

To model the contamination, we assign a weakly informative normal prior for  $\delta_{\beta_{2jk}} \sim \text{normal}(0, 1)$  and a weakly informative exponential prior for  $\delta_{\sigma_{jk}} \sim \text{exponential}(1)$ . This prior reflects our beliefs that each plate could have different degrees of contamination, and that the contaminated samples should have an amplified measurement error in the variance component. For mixture proportion  $\lambda_k$ , we assign a weakly informative Beta prior with a mean close to zero, which is  $\lambda_k \sim \text{beta}(1, 10)$  to reflect our prior belief that the proportion of contamination across plates is independent of each other, and also that for each plate it should be relatively small unless it reveals a poor quality-control design.

For the unknown allergen concentration of unknown samples, on each plate  $k$ , we assign hierarchical exponential priors for the concentrations  $\theta$  across different unknown samples  $j$  and let the model estimate the hyperparameter from the data. The exponential prior is consistent with the possibility of zero or effectively zero concentrations for some samples while allowing high concentrations for others:

$$\theta_{jk} \sim \text{exponential}(\mu), \mu \sim \text{normal}^+(0, 0.1). \quad (3.14)$$

For the Bayesian regression model, our default setting uses the noninformative uniform prior to support parameters. For example, for the Bayesian linear regression model, we use a uniform prior over the positive line for the scale of the normal error as well as a uniform prior over the real line for the regression coefficients. Furthermore, this prior specification for the Bayesian regression model could be easily modified to have more specific parametric forms for some of the covariates if we have strong prior beliefs.

### 3.2.4 Sensitivity analysis and computation

To validate the robustness of our proposed model in both our simulation studies and real-life data analysis, we performed the following sensitivity analysis. We tried other parametric forms to model the interdependence between the calibration curve parameters, including t distribution with a fixed degree of freedom as well as treating the degree of freedom as a hyperparameter. Furthermore, we tried different values for the prior mean of the contamination mixture probability. None of these model modifications and sensitivity analyses changed the overall model inference results, which indicated that our model was fairly robust.

The Bayesian model fitting and evaluation in this study were based on the NUTS algorithm in Stan as called from CmdStanR (Homan and Gelman, 2014; Carpenter et al., 2017). NUTS is a Hamiltonian Monte Carlo (HMC) variation that combines Markov chain Monte Carlo and deterministic simulation methods to achieve faster convergence. CmdStanR is a lightweight interface for Stan for R users that provides modularity for downstream analysis. To monitor the convergence status of the algorithm, we checked whether the model running time was reasonable and used commonly used diagnostic statistics in Bayesian computation, such as the effective sample size (ESS) and the  $\widehat{R}$  diagnostic (Gelman and Rubin, 1992).

Since the joint modeling approach involved estimating the plate-level unknown samples' concentration as well as the individual-level disease model association, we expected the computation time to be slightly longer. The actual computation time for the joint modeling approach was measured in minutes in both the simulation studies and the real-life examples, which was acceptable

given the relative complexity of its internal structures. Furthermore, the effective sample size for all parameters in the model was sufficiently large for valid posterior inference results to be obtained; moreover, all associated  $\widehat{R}$  diagnosis values were relatively close to 1 in simulation studies and real-life examples. We also checked the HMC chains' mixing status using visualization tools and found all of them to mix fairly well in the simulation studies and the real-life examples.

### **3.3 Bayesian workflow for joint inference model**

This section describes the multiple steps that we considered to be involved in the Bayesian workflow. Here, we decomposed the main research question into the following two subquestions:

1. How can a global calibration model be constructed?
2. How can the inference results from the global calibration model be used in follow-up disease-exposure association studies?

When one deals with global calibration, one would typically always aim to consider the tradeoffs between accuracy gain and model complexity. One might not know the degree of model complexity or the degree of information sharing among multiple experimental units when building a global calibration model; therefore, one should always consider a broad category of models and then carefully compare their performance following the general guidelines for Bayesian model comparison and evaluation. For our global calibration model's construction, the fundamental building block was the Bayesian contamination model that we developed in our first project. However, to pool information from multiple experimental units together, we needed to consider some additional inter-unit dependence structures. Based on real-life lab evidence, we are convinced that not all experimental units have the same properties and, due to differences in lab processing conditions, we believe that each experimental unit typically has its own calibration curve parameters. However, some common properties are shared among different experimental units' calibration curve parameters; thus, an initial natural model was used to separately consider the four calibration curve parameters on each experimental unit and model the interdependence of each one's curve

parameters across units. We then extended and modeled this interdependence using multivariate normal distributions, but we did not find much improvement based on real-life data; thus, we returned to the original model. Then, we considered adding components to the global calibration model, such as how we should allow the form as well as the degree of contamination to vary across multiple samples across multiple experimental units. We started by directly allowing each of them to have its own degree of contamination, but the contamination forms were restricted to being fixed on only a proportion of the curve parameters. If there was any new evidence or insights from the data, we could further generalize our global calibration model to allow the contamination forms to vary across the curve parameters. In the abovementioned model-building process, we needed to consider all of the necessary Bayesian computation and modeling evaluation tools mentioned in our first project for the estimation of individual experimental units' contamination. In addition, we checked whether the model could successfully achieve the more specific goals of immunoassay data, such as obtaining the posterior probability of each sample being contaminated. Since we now had multiple experimental units together, if we had found that one or several experimental units had a very high contamination ratio, then it would be helpful to return and check where the experimental procedures or conditions for those units were significantly different from others. This could also provide insights for identifying sources of contamination as well as contamination prevention in applied serial dilution assay studies.

In the disease–exposure association model, the starting point should always be regression or modified regression models. Furthermore, to ensure clearer interpretations of the model parameters, one might sometimes need to either scale the data or transform the model parameters for purposes of practical interpretation. Then, we could further extend and generalize the disease–exposure association model to more general model classes, such as stratified models, mixed-effects models, or even nonlinear models. In this project, we did not extend the regression model that far because no strong evidence exists for more complicated models. The evaluation metrics for the disease–exposure association model include both the necessary Bayesian inference evaluation tools, such as posterior predictive checking, prior predictive checking, and sensitivity analysis, but

also whether the final output has any real-life public-health interpretations and whether it could serve to promote the health of the general public. Since we typically would not know the true associations in real-life datasets, different simulation scenarios could help us greatly in understanding the performance of our method.

Here, we also took advantage of the idea of modular Bayesian model construction. First, the global calibration model started from the most straightforward and simplest form; then, we evaluated whether we needed to expand or add necessary parts. Second, the global calibration model's specification and the disease–exposure association model's specification were independent of each other. If the global calibration model could reveal information from the observed data and generate accurate exposure estimations, no matter what the form is, it could always be incorporated in the follow-up disease–exposure association models. Third, the method we proposed in the first project naturally fits as a module in the joint inference model developed in the second project; therefore, the modular idea is not only limited to parameters or a small model but could also be extended to a situation where a whole inference model is integrated within another more complicated inference model.

### **3.4 Application studies**

#### **3.4.1 Childhood asthma studies**

Asthma is recognized as the most common chronic disease in children, and over the past 40 years, its prevalence rate has increased significantly (Serebrisky and Wiznia, 2019). Moreover, the increasing hospitalization rate for asthma-related diseases and the increasing burden of asthma indicates the necessity of conducting public health research projects on asthma prevention. Researchers have hypothesized that potential causal factors in the asthma epidemic majorly consist of environmental factors, such as exposure to tobacco smoke, air pollution, exposure to indoor allergens, obesity, infections, and microbial components (Eder et al., 2006). However, the nature of asthma is highly complex, and the origins of asthma remain an open question. Among the potential risk factors for asthma development, the role of indoor allergens is of great research in-

terest. Evidence suggests that allergic sensitization is strongly associated with asthma (Plattsmills et al., 1997). Moreover, studies found that children exposed to cockroach allergens and with a cockroach allergy have not only more frequent asthma symptoms (e.g., wheezing and nights with lost sleep) but also higher hospitalization rates (Rosenstreich et al., 1997). Furthermore, in-house mouse allergen has been demonstrated to be an essential factor in asthma in children (Phipatanakul et al., 2000; Matsui et al., 2005). In addition, within the same city in large major urban areas (e.g., NYC), exposure and sensitization to such indoor allergens might differ between neighborhoods, thus potentially leading to different asthma prevalence rates (Olmedo et al., 2011). These findings and insights illustrate the importance of accurate indoor allergen assessment, measurement, and evaluation in the field of asthma development and prevention studies.

On the other hand, many clinical studies have indicated the promise of asthma prevention by studying the reduction of indoor allergens. Some clinical intervention evidence indicates support for asthma prevention by interrupting the pathway from indoor allergen exposure to allergen sensitization to atopic asthma (Gaffin and Phipatanakul, 2009). Randomized controlled trials have been conducted to assess the treatment effect of indoor allergen removal by cockroach extermination and air cleaner on asthmatic children and found a significant decrease in daytime asthma symptoms in the treatment group (Eggleston et al., 2005). Moreover, the mouse allergen has been identified as an independent risk factor for asthma morbidity, and corresponding allergen reduction interventions would assist in reducing sleep disorders related to asthma (Pongracic et al., 2008). The use of mite-impermeable bedcovers could help and lead to improvements in reducing hospital visits for mite-sensitized asthma children, thus reducing the corresponding health burden (Murray et al., 2017). In addition to the aforementioned indoor allergens, multiple clinical intervention studies have covered other common indoor allergens, including pet allergens and fungi, obtaining promising results (Ahluwalia and Matsui, 2018). However, some clinical interventions have also obtained mixed effects and results. Evidence indicates that successful interventions usually target multiple allergens and include individually tailored plans (Ahluwalia and Matsui, 2018). While multicomponent intervention strategies on indoor allergens targeting asthma improvement lead to

improved results, no clear evidence exists for certain combinations of interventions being more effective and beneficial, which is due to a lack of studies (Leas et al., 2018). This need for further research on similar topics motivates the need to develop a methodology for more accurately estimating indoor allergen concentrations and measuring asthma symptoms. Furthermore, a sufficient population size is usually required to discover the potential underlying associations and derive reliable statistical conclusions.

### 3.4.2 Summary statistics of NYCNAAS data and research question of interest

As introduced in the previous section, the full NYCNAAS data set belongs to a large-scale case-control study for children with asthma conducted in New York City. The participants are recruited through the Health Insurance Plan of New York across all four districts of New York City. Within this dataset, we have majorly four different types of available data and information recorded for each participating child in the study. The first category includes information related to the child's asthma status and asthma symptoms, such as asthma diagnosis status, asthma symptoms severity, and asthma persistence. The second category includes information related to indoor allergen concentration estimations and other allergen-related data such as dust mite allergen concentration, cat allergen concentration, mouse allergen concentration, and IgE antibody level associated with those allergens. The third category includes information related to some baseline covariates, which is basically acquired by the screening questionnaire for the child's parents, such as home income, sex of the child, race of the child, family education level, and family members smoking status. The final category includes information related to other clinical measurements related to asthma, such as exhaled nitric oxide and ambient NO. In summary, the NYCNAAS dataset is a pretty comprehensive dataset that provides researchers with much available information to explore their own research questions of interest related to children's asthma and indoor allergen concentration. A more detailed summary of descriptive statistics for the NYCNAAS dataset can be found in Table 3.1.

For our research question of interest, we are more interested in the subgroup of children with

Features	All Participants (n = 350)	Asthmatic Participants (n = 206)	Nonasthmatic Participants (n = 144)
<b>Natural log of exhaled nitric oxide</b>			
Mean	2.38	2.43	2.30
Min	0.75	0.75	0.97
Max	4.94	4.15	4.94
Missing	20	15	5
<b>Race</b>			
Black	164 (47%)	101 (49%)	63 (44%)
Others	186 (53%)	105 (51%)	81 (56%)
<b>Sex</b>			
Male	195 (56%)	118 (57%)	77 (53%)
Female	155 (44%)	88 (43%)	67 (47%)
<b>Family Smoker</b>			
Yes	73 (21%)	42 (21%)	31 (22%)
No	271 (79%)	162 (79%)	109 (78%)
Missing	6	3	4
<b>Maternal Asthma</b>			
Yes	69 (20%)	53 (26%)	16 (12%)
No	274 (80%)	151 (74%)	123 (88%)
Missing	7	2	5
<b>Maternal Education</b>			
High School or higher	317 (92%)	182 (90%)	135 (94%)
No high school	29 (8%)	21 (10%)	8 (6%)
Missing	4	3	1

Table 3.1: Summary of patient characteristics in NYC NAAS study.

asthma diagnosis and would like to investigate the potential associations between indoor allergen concentration and public health related outcomes after adjusting for some baseline covariates. For the public health-related outcome, currently among those asthmatic children, we are interested in several major potential asthma-related evaluation outcomes. The first potential outcome of interest is the measurement of exhaled nitric oxide (NO) for those asthma children. The measurement of exhaled nitric oxide (NO) and its role related to asthma development, asthma mechanism, and asthma diagnosis has been well studied in the literature (Yates, 2001; Wang et al., 2020). The second potential outcome of interest is a binary indicator regarding whether the asthma symptoms are frequent or not, which is also a measurement of the severity of the asthma status for the children. The third potential outcome of interest is related to asthma persistence, which is defined as whether the child still has asthma at a three-year-later follow-up visit.

For the covariates to put in the regression model, we would definitely include the key predictor of our interest, which is the concentration of indoor allergens. At the same time, we should also include some baseline covariates for the patients involved in the study to remove any potential

confounding effect. Based on previous studies of the NYCNAAS data (Perzanowski et al., 2008; Chen et al., 2016), we include the following baseline covariates in the epidemiologic association studies: the sex of the child; race of the child (Black vs Others); whether there was any smokers at home; whether the child’s mom has high school or more education; whether the child has maternal asthma.

### 3.4.3 Bayesian linear regression on asthmatic children

Since we currently only focus on children that have been diagnosed with asthma, we first have done some inclusion-exclusion criteria on the NYCNAAS dataset. Based on the clinical definition of asthma case and control status, we only select the sub-population that has a clear and definite asthma diagnosis. After this inclusion-exclusion screening procedure, we select the continuous public-health outcome of interest, which is the exhaled nitric oxide (NO) measurement, as the response variable  $Y$  in the following epidemiological association regression model:

$$Y_n \stackrel{ind}{\sim} \text{normal}(\alpha + \gamma\theta_n + \eta_1 Z_{1n} + \eta_2 Z_{2n} + \eta_3 Z_{3n} + \eta_4 Z_{4n} + \eta_5 Z_{5n}, \epsilon_n). \quad (3.15)$$

where  $\theta_n$  represents the indoor dust mite concentration estimation for patient  $n$ , which is our main predictor of interest.  $Z_{1n}$  represents the sex for patient  $n$ ,  $Z_{2n}$  represents the race indicator for patient  $n$ ,  $Z_{3n}$  represents the baseline family smoker existence status for patient  $n$ ,  $Z_{4n}$  represents the baseline parental education level for patient  $n$ , and  $Z_{5n}$  represents the baseline parental asthma status for patient  $n$ .

Then we have fitted three different Bayesian regression models: The first regression model directly uses the allergen concentration estimation obtained by the classical calibration method used in labs and treats the point estimation of allergen concentration as a known and fixed value in the Bayesian regression model; The second regression model takes the approach we have proposed in the Bayesian joint model framework, where the raw serial dilution assay data measurement units are combined together using the Bayesian global calibration framework and the Bayesian regression model is also embedded within the joint model as a second component, which could

allow the randomness of the allergen concentration estimation to be taken into consideration. The third method is a naive two-step model, where we first apply the Bayesian contamination model proposed in Chapter 2 to each of the measurement units in the first-step analysis. And then in the second step of the analysis, we use the posterior median of the allergen concentration estimation as the true and fixed value in the Bayesian regression model. To make these three methods of indoor allergen concentration estimation directly comparable, we have also applied appropriate measurement unit transformations. The regression coefficients point estimates, where we have used the posterior median in Bayesian methods, together with the Bayesian uncertainty interval, have been presented in Figure 3.2:

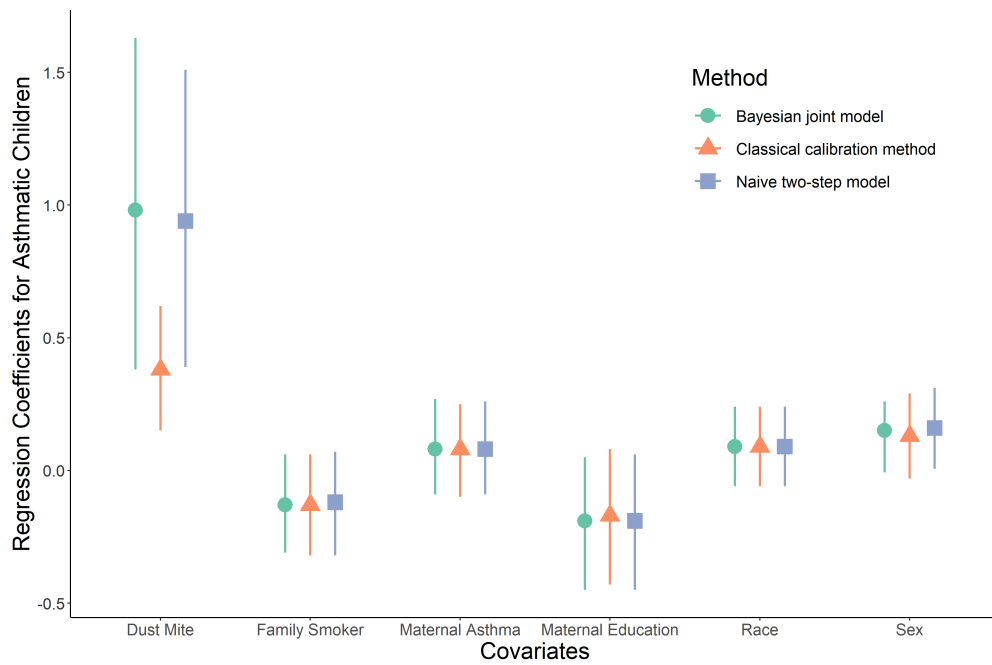


Figure 3.2: Bayesian regression coefficients summary for NYCNAAS data, where the posterior median and 80% uncertainty interval are presented.

From Figure 3.2, we have several observations: First, the traditional calibration method, the two-step model using local calibration, and the Bayesian joint model all have identified the indoor dust mite allergen concentration as a significant and important factor for the exhaled nitric oxide (NO) measurement for asthmatic children; Second, although these models have similar findings on the effect of indoor dust mite allergen concentration on the exhaled nitric oxide (NO) measurement

for asthmatic children, the scale of effects is pretty different, and our Bayesian joint model which takes the potential existence of sample contamination and measurement uncertainty into consideration has identified a much larger effect than the classical calibration method. The two-step method estimated a slightly smaller effect than the joint modeling but yielded a narrower probability interval; Third, all the remaining baseline covariates that are included in the epidemiologic studies are not identified as significant predictors by these three models. To further investigate the reasons behind the difference of coefficient scales in the Bayesian linear regression model using the classical calibration estimation method and the Bayesian joint model, we have generated the following plot to compare the point estimation by classical calibration method and Bayesian joint model:

Compared to the point estimation of the Bayesian joint model, some observations of the classical calibration methods are highly inflated in their values and thus result in a decrease in the regression coefficient estimate associated with the allergen concentration. There are several potential explanations behind that: First, since the classical calibration method does not take the measurement uncertainty into consideration, it may lead to such attenuation in the estimated values; Second, when there is contamination, the classical calibration method still assume the contaminated sample shares the same calibration curve parameters as the standard calibration sample, which is not true and ignores the existence of contamination could also lead to biased concentration estimates since contaminated samples usually tend to have larger measurement uncertainty and behave differently than what the classical calibration method assumes. To be more specific, ignoring potential sample contamination and ignoring measurement uncertainty could generally have a larger effect on the unknown sample observations with a dilution factor of 1/10000, and after transferring the scales back, this usually will lead to an inflated concentration estimation. Third, the problem of below detection limit, which limits the use of the classical calibration method, might also partially explain this difference since the Bayesian joint model could overcome the problem of below detection limit easily by treating corresponding parameters as random variables.

In Figure 3.3, we further investigate the reason behind the observation that some points have higher classical calibration method estimation results than the Bayesian joint model estimation

results. The solid line represents the reference line with slope 1 and intercept 0, and the dashed line represents the boundary line with slope 1 and intercept 0.3 that separate the points into the group that has higher classical calibration method calibration method estimation results (inconsistent group) and the remaining group (consistent group). For those points in inconsistent group, we check their corresponding raw serial dilution measurement samples and find all of them are identified as contaminated samples by the Bayesian joint model. This partially explains such deviance between the points in inconsistent group and the solid reference line since the classical calibration method estimation does not take sample contamination into consideration, which will result in biased estimation results.

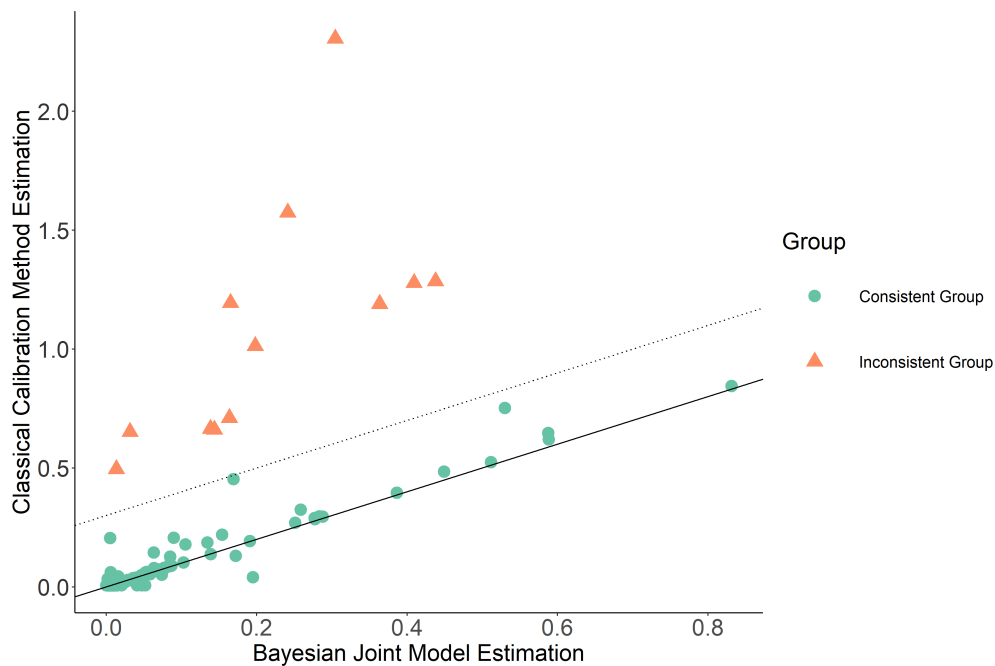


Figure 3.3: Comparison of indoor allergen concentration estimation by the classical calibration method and the Bayesian joint model.

For a more comprehensive comparison, we have also generated Figure 3.4 and Figure 3.5, which are pretty similar to the central idea in Figure 3.3. In Figure 3.4, we compare the estimation results between the naive two-step model that uses local calibration by the method proposed in our first project with the Bayesian joint model that uses global calibration. We have observed that the overall trend between the Bayesian joint model estimation and the naive two-step estimation

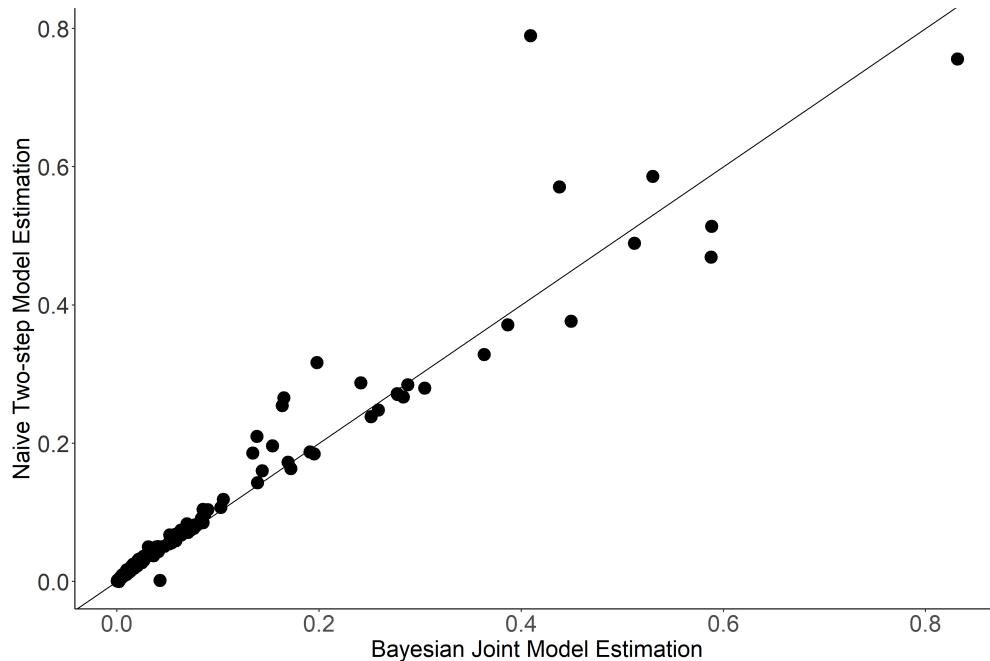


Figure 3.4: Comparison of indoor allergen concentration estimation by the Bayesian joint model and the naive two-step model using local calibration.

aligns well, but as the underlying values increase, there is some subtle deviance between these two measurements. In Figure 3.5, we also compare the naive two-step model that uses local calibration by the method proposed in our first project with the traditional calibration method used currently in serial dilution assay wet labs and find similar patterns as we see in Figure 3.3. Compared to the naive two-step estimation, some estimations by the classical calibration method are having higher values, and after further investigation, we find that this inflation is majorly due to ignorance of sample contamination and measurement error.

To be more specific, in Figure 3.6, we have picked up one specific unknown sample that has much higher values of concentrations from the classical calibration method than what we have from the Bayesian estimate, from the traditional method we found that on that plate this sample has three observations but the observation with dilution factor 1/10000 has been identified as <OOR, and the two other observations with 1/10 and 1/100 have similar but relatively large observations. For the classical calibration method, on that plate, nearly all the other samples are estimated as < OOR, so it is kind of awkward for only one sample to have a very high concentration, instead, it might

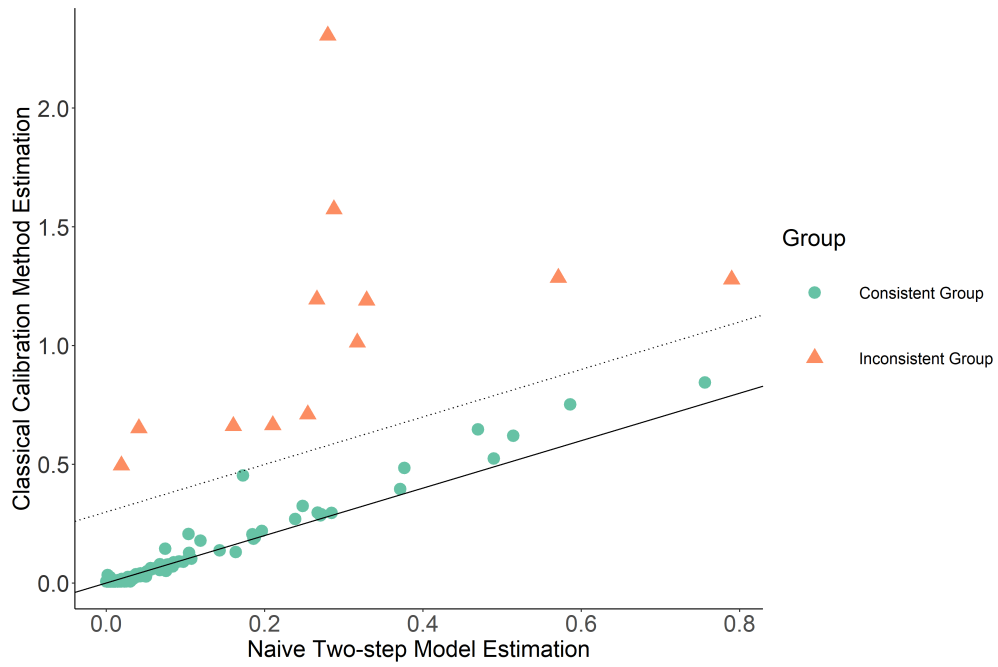


Figure 3.5: Comparison of indoor allergen concentration estimation by the classical calibration method and the naive two-step model using local calibration.

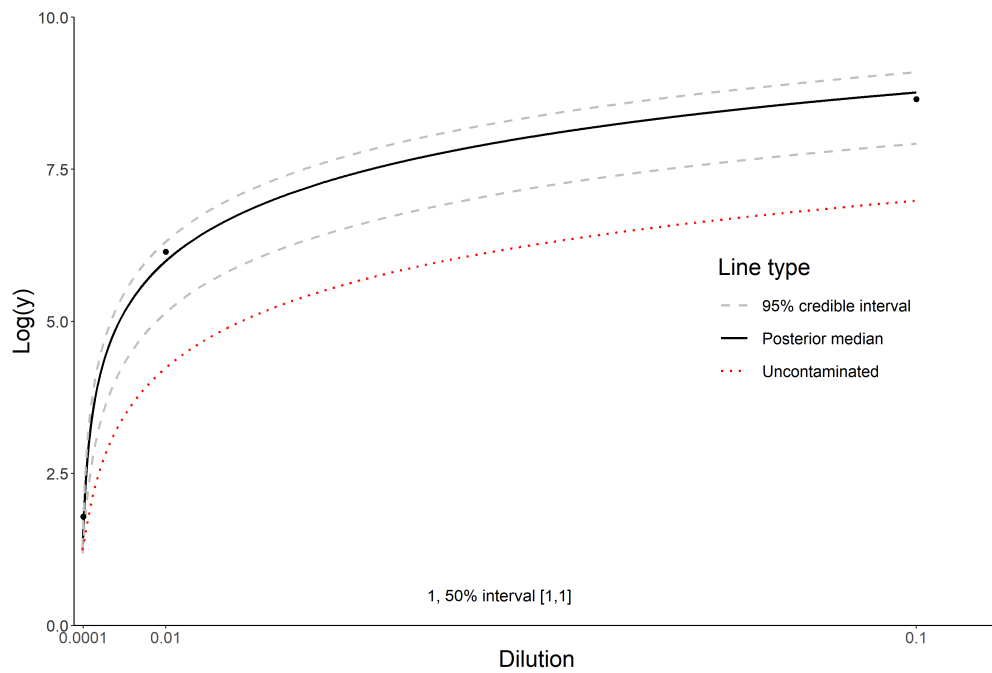


Figure 3.6: Example of one specific unknown sample having larger classical concentration estimation compared to Bayesian joint concentration estimation.

be more reasonable to believe that this unknown sample is contaminated and thus it should share a different set of calibration curves parameters compared to the standard calibration data. For better comparison, in Figure 3.6, we have plotted the estimation result by Bayesian joint model and classical calibration method separately, together with the raw data observation points. The 'Uncontaminated' line represents the corresponding mean dose-response curve for the situation if this unknown sample shares exactly the same set of calibration parameters as the set of standard calibration data and we found that this generated curve could not fit the three observed points well. For the Bayesian joint model, we have presented both the mean dose-response curve as well as the corresponding uncertainty interval after considering the contamination situation and letting the unknown sample have its own curve parameters, and found that the fitting results had improved a lot. In addition, we have also checked how the method we proposed in our first project fits this unknown sample and found very similar results to the Bayesian joint model and the mean dose-response curve overlaps with the one generated by the Bayesian joint model and thus we did not include that in Figure 3.6.

Table 3.2 summarizes the observed values on the selected unknown sample in Figure 3.6, which has a larger classical calibration method concentration estimation compared to that by the Bayesian joint model. We could find that the 1/10000 dilution observation by the classical calibration method falls below the detection limit set by the classical calibration method and thus could not generate any valid estimates, which is not optimal and also makes the large estimation of the 1/10 and 1/100 dilutions suspicious.

	Observed Value	Classical Calibration Method Concentration
1/10 Dilution	5726	67.4
1/100 Dilution	466	71.08
1/10000 Dilution	6	OOR <

Table 3.2: Raw serial dilution assay data for unknown sample 6 on 012711 plate one, OOR< means below the detection limit.

#### 3.4.4 Insights from this study

In this applied study, we have found that there are subtle differences between the indoor dust mite concentration estimation generated by the classical calibration method commonly used in labs and the Bayesian joint model we have proposed. Given the potential pitfalls of the classical calibration method due to the below detection limit issue, potential sample contamination, and ignorance of measurement error in both the lab concentration estimation process and the epidemiological model, we are confident that our proposed Bayesian joint model could generate more accurate, robust and efficient estimations in both the raw concentration estimation as well as the regression model coefficients estimation. The evidence from the NYCNAAS dataset might suggest that the classical calibration method might overestimate the indoor allergen concentration, and thus the identified associations found in the study might be underestimated with a downward bias, as we have demonstrated from our analysis.

Furthermore, our model has identified a strong relationship between indoor dust mite allergen concentration and the exhaled nitric oxide (NO) measurement for asthmatic children. This finding might be helpful for clinical experts and researchers to design corresponding asthma prevention and treatment plans for the early development of children's asthma. Also, it motivates similar research projects to investigate further other potentially important factors related to indoor allergen concentration that might be helpful in early asthma detection, diagnosis, and prevention. This applied study has demonstrated that Bayesian methods could help epidemiological researchers investigate potential underlying associations between public health-related responses and key predictors of interest in general. And more specifically, incorporating flexible mixture models for potential model contamination and using global calibration to jointly model multiple measurement units together with the combination of concentration estimation and epidemiological regression model in the field of serial dilution assay could not only generate more accurate indoor allergen concentration estimation but also could discover the underlying associations more accurately.

## 3.5 Simulation studies

### 3.5.1 Model comparison and evaluation in severe contamination setting

In this simulation setting, we were interested in the following two major research questions:

1. How do different competing methods perform in this multi-plate contamination scenario?
2. What are the potential gains of incorporating uncertainty in both the key predictor and outcome in Bayesian regression?

We compared the performance of the following four models, which could be further divided into two broad categories. The category of the two-step model meant that we first obtained the point estimation of allergen concentration either using a direct curve-inversion approach or the posterior median from Bayesian models; then, we forwarded this point estimation in a regression setting as a fixed-X design. The category of the joint model means the model we proposed in this paper, which forwards the estimation uncertainty into a regression setting as a random-X design. Specifically, we sought to compare the two-step classical calibration estimation model, which is currently commonly used in labs (model 1); the two-step base model proposed in (Gelman et al., 2004) (model 2); the naive two-step model using the Bayesian contamination model proposed in our first project (model 3); and the Bayesian joint contamination model proposed in this chapter (model 4). The two-step base estimation model (model 2), which considers measurement error but ignores sample contamination, serves as an intermediate comparison between the two-step classical calibration model (model 1) which ignores the measurement error as well as sample contamination, and the naive two-step model using the Bayesian contamination model proposed in our first project (model 3) which considers both measurement error and sample contamination. The motivation for comparing the Bayesian joint model performance with these two-step models is that sometimes in practice, epidemiologists might find the Bayesian joint model overly complicated for their applications. Moreover, in some studies, environmental health researchers only have access to the individual level of epidemiological data, including the disease outcome, key predictor of

interest (e.g., indoor allergen concentration), and patients' characteristics data. In these cases, the joint modeling approach is no longer applicable. In such settings, the two-step model could serve as an alternative modeling approach. The simulation process was divided into four steps, which are described as follows.

In step 1, we aimed to simulate the calibration curve parameters across plates, and we assumed that we had 10 plates in total. Since we were now using independent Beta assumptions, we simulated  $\beta_k$  across the 10 plates as follows:

$$\beta_{1k} \sim \text{normal}^+(\beta_1, 2). \quad (3.16)$$

$$\beta_{2k} \sim \text{normal}^+(\beta_2, 1000). \quad (3.17)$$

$$\beta_{3k} \sim \text{normal}^+(\beta_3, 5). \quad (3.18)$$

$$\beta_{4k} \sim \text{normal}^+(\beta_4, 1). \quad (3.19)$$

where  $k$  represents the plate ID 1, 2, ..., 10. Furthermore, given the relatively different scales of calibration curve parameters values observed from real-life studies, we set  $\beta_1 = 1$ ,  $\beta_2 = 15000$ ,  $\beta_3 = 8$ , and  $\beta_4 = 1.5$ , and we assigned different standard deviations for these four different normal distributions.

In step 2, we aimed to simulate observed measurements  $w_{ik}$  ( $i$  for dilution and  $k$  for the plate) for the standard calibration sample on each plate. After simulating the  $\beta_k$  for each plate, we simulated standard calibration samples with a known initial concentration  $\theta_0 = 125$  and applied two-fold dilutions (two replicates at each dilution level ranging from 1, 1/2, ..., 1/2048 and 0). For plate  $k$ , we had the following:

$$\log(w_{ik}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_0 d_{ik}, \beta_k)), \sigma_w), \quad (3.20)$$

For simplicity, we assume the variance scale for standard calibration samples are same across plates and set  $\sigma_w = 0.1$ , and  $d_{ik} = 0, 0, 1, 1, 1/2, 1/2, \dots, 1/2048, 1/2048$ .

In step 3, we aimed to simulate observed measurements  $W_{ijk}$  ( $i$  for dilution,  $j$  for unknown sample ID, and  $k$  for the plate) for 23 unknown samples on each plate. Here, we assigned 23 unknown samples on each plate to restore the real-life plate setting. For each of the unknown sample concentrations  $\theta_{jk}$  ( $j = 1, 2, \dots, 23, k = 1, 2, \dots, 10$ ), we randomly sampled it from Exponential (0.1) and simulated the observed values of unknown samples using 1/10, 1/100, and 1/10,000 dilutions. Within the 23 unknown samples on each plate, we let five of them be contaminated with contamination factor  $\delta_{\beta_{2jk}} = 1$  and  $\delta_{\sigma_{jk}} = 1$ ; thus, their observed values were sampled from the following:

$$\log(w_{ijk}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_{jk}d_{ijk}, \boldsymbol{\beta}_{jk})), \sigma_w e^{\delta_{\sigma_{wk}}}), \quad (3.21)$$

where  $\boldsymbol{\beta}_{jk} = (\beta_{1k}, \beta_{2k} e^{\delta_{\beta_{2jk}}}, \beta_{3k}, \beta_{4k})$

For the remaining 18 uncontaminated unknown samples on each plate, we use the same data generation process as generating the calibration standard samples:

$$\log(w_{ijk}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_{jk}d_{ijk}, \boldsymbol{\beta}_k)), \sigma_w), \quad (3.22)$$

where  $\sigma_w = 0.1$  and  $d_{ijk} = 1/10, 1/100, 1/10000$  for each of the unknown sample on each plate.

In step 4, we aimed to simulate the epidemiological response variable  $Y$ . Within each simulation iteration, we had 230 patients. For each patient, we assumed them to have a continuous health-related response variable  $Y$  that follows a normal distribution with the mean depending on covariates such as age ( $x_1$ ), sex ( $x_2$ ), race ( $x_3$ ), and allergen concentration ( $x_4$ , which is equivalent to  $\theta_{jk}$ ). Among these covariates, we assumed the age distribution to be uniform (8,16), the sex ratio to be 0.5:0.5, and the race ratio to be 0.45, 0.2, and 0.35, respectively. Then, we simulated the  $Y$  response variable using the following expression:

$$Y \stackrel{ind}{\sim} \text{normal}(0.5 - 0.05 * x_1 + 0.1 * x_2 + 0.1 * x_3 + 0.4 * \log(x_4), 1). \quad (3.23)$$

We repeated the simulation 500 times and summarized the estimation results for the target of

interest, namely the regression coefficient for allergen concentration in the regression model among all four competing methods. We found that the proposed Bayesian joint model had the smallest bias and root mean square error. The naive two-step model also worked well in terms of coverage probability with a larger RMSE compared to the Bayesian joint model. The other two models did not perform well in the situation, for which potential reasons could be sample contamination and measurement uncertainty for the calibration curve estimation.

	Bias	RMSE	80% CP	80% IW	95% CP	95% IW
Traditional Model (model 1)	-0.2065	0.492	1%	0.10	5%	0.16
Base Model (model 2)	-0.1507	0.449	31%	0.11	44%	0.16
Naive two-step Model (model 3)	-0.0019	0.129	80%	0.15	94%	0.23
Bayesian joint Model (model 4)	0.0007	0.055	81%	0.15	96%	0.23
Benchmark Model	-0.0003	0.051	80%	0.13	95%	0.21

Table 3.3: Comparison of four methods' performance on regression coefficient of dust mite concentration (true value is 0.4), now the contamination level is 1 (severe contamination), RMSE: root mean squared error, CP: coverage probability, IW: interval width.

### 3.5.2 Efficient multi-plate experiment design

Compared with the single plate experiment, the Bayesian joint model could pool information from the calibration standard samples across multiple plates; thus, we sought to determine whether we could reduce the number of calibration standard sample observations within each plate and allow for more unknown samples to be placed on each one for a more efficient experiment design while maintaining estimation accuracy. Since the space available on each plate is limited, if we could reduce the number of calibration standard sample observations, we could have estimates for more unknown samples with the same number of experiments. In this simulation setting, we used the setting in simulation 1 as our benchmark, where each plate had 26 observations of the calibration standard sample (with two-fold dilutions: 0, 0, 1, 1, 1/2, 1/2, 1/4, 1/4, 1/8, 1/8, 1/16, 1/16, 1/32, 1/32, 1/64, 1/64, 1/128, 1/128, 1/256, 1/256, 1/512, 1/512, 1/1,024, 1/1,024, 1/2,048, 1/2,048) and 69 observations for 23 unknown samples (each with dilutions 1/10, 1/100, 1/10,000). We maintained everything the same as in simulation 1 except now, on each plate, we had 14

observations of standard calibration samples (with two-fold dilutions: 0, 0, 1, 1, 1/4, 1/4, 1/8, 1/8, 1/128, 1/128, 1/512, 1/512, 1/2,048, 1/2,048) and 81 observations for 27 unknown samples (each with dilutions 1/10, 1/100, 1/10,000). Then, we sought to compare the estimation results of the regression coefficient of allergen concentration using our Bayesian joint model in these two settings and we repeated the simulation 500 times.

We found that, compared with the original design, the bias and RMSE of the new experimental design had decreased slightly compared with the original design, but this might be because we increased the sample size in the Bayesian linear regression from 230 to 270 in the new experimental design. This simulation study indicated that we perhaps do not need that many observations of the standard calibration sample on each plate, and that we could more effectively use those spaces for more new samples. This might provide some innovative ideas for designing more efficient MARIA plates.

	1	2	3	4	5	6	7	8	9	10	11	12
	Standard Sample			Unknown Samples								
				1/10	1/100	1/10000	1/10	1/100	1/10000	1/10	1/100	1/10000
A	1	1/4	1/8	Unk 1	Unk 1	Unk 1	Unk 9	Unk 9	Unk 9	Unk 17	Unk 17	Unk 17
B	1	1/4	1/8	Unk 2	Unk 2	Unk 2	Unk 10	Unk 10	Unk 10	Unk 18	Unk 18	Unk 18
C	1/128	1/512	1/2048	Unk 3	Unk 3	Unk 3	Unk 11	Unk 11	Unk 11	Unk 19	Unk 19	Unk 19
D	1/128	1/512	1/2048	Unk 4	Unk 4	Unk 4	Unk 12	Unk 12	Unk 12	Unk 20	Unk 20	Unk 20
E	Unk 24	Unk 24	Unk 24	Unk 5	Unk 5	Unk 5	Unk 13	Unk 13	Unk 13	Unk 21	Unk 21	Unk 21
F	Unk 25	Unk 25	Unk 25	Unk 6	Unk 6	Unk 6	Unk 14	Unk 14	Unk 14	Unk 22	Unk 22	Unk 22
G	Unk 26	Unk 26	Unk 26	Unk 7	Unk 7	Unk 7	Unk 15	Unk 15	Unk 15	Unk 23	Unk 23	Unk 23
H	Unk 27	Unk 27	Unk 27	Unk 8	Unk 8	Unk 8	Unk 16	Unk 16	Unk 16	blank	blank	HC Control

Table 3.4: Illustration for a more efficient design for standards and new samples (dilutions) in a multiplex plate with 96 wells.

	Bias	RMSE	80% CP	80% IW	95% CP	95% IW
Joint Model Original Design	0.0007	0.055	81%	0.15	96%	0.23
Joint Model New Design	0.0004	0.053	78%	0.14	95 %	0.21

Table 3.5: Comparison of the original design with 26 observations of calibration sample and 23 unknown samples per plate versus new design with 14 observations of calibration sample and 27 unknown samples per plate, RMSE: root mean squared error, CP: coverage probability, IW: interval width.

### 3.6 Discussion

Global calibration inference using multiple sources of calibration data is crucial in large cohorts of environmental health studies and other general applied statistical problems. This study proposed a general Bayesian joint model for inference on multiple experimental raw data plates in serial dilution assay studies. The proposed model allows each plate to have its own calibration curves while keeping the overall inference accurate and convenient with the help of a hierarchical model and partial pooling. The inference results could be easily matched back to each individual plate to obtain inferences regarding unknown samples' concentration and contamination status. We used both real-world NYC NAAS data and simulation studies to demonstrate the advantages of our Bayesian joint model, among other competing methods. Furthermore, the general Bayesian joint model could provide additional uncertainty measurements when applied with the Bayesian regression model for epidemiological association studies. The traditional fixed-X regression design is unsuitable for some epidemiologic associations because some critical predictors of interest might be random variables themselves. Nevertheless, our Bayesian joint calibration model could organically integrate the key predictors of the interest estimation and regression analysis parts by naturally plugging the uncertainty into the estimation under a random-X regression design. By applying the proposed joint model to NYC NAAS data, we found that accounting for uncertainty in key predictors in Bayesian regression could help researchers to identify associations more accurately.

In addition, we illustrated the general idea of a Bayesian workflow and how it guided us step-by-step in constructing the proposed Bayesian joint model. The flexibility of the joint model was demonstrated by the fact that the target of interest estimation process and association estimation submodels did not overly rely on each other. Thus, if any changes are made in the dataset or additional constraints or attention are required, one could easily adapt the corresponding parts of the submodel while not affecting the overall joint model. This endows the proposed joint model with the ability to be applied beyond the field of serial dilution assay data. In fact, in any general

applied statistical problems, if the idea of global calibration could be applied, the proposed joint model might also be beneficial. This also proves that global calibration and the partial pooling of information from multiple sources could increase the model's efficacy, make the model transferrable to other similar settings, and reduce unnecessary operations and computation greatly.

Nevertheless, our study has several limitations. First, although the independence model between the four parameters in the calibration curve proposed by our Bayesian model has both computational efficacy and fits the NYC NAAS dataset well, it still might be too ideal for some complicated real-world settings. In that case, the parameters of the calibration curves might also correlate with each other, and we could be able to model the correlation between plates in stronger functional forms if we had additional lab condition information using multivariate models. Second, the underlying causes of sample contamination and plate-specific effects are usually unknown. Therefore, our proposed global calibration model attempts to approximate the patterns that we observed based on empirical studies in the lab. However, in some cases, we might have additional information regarding the sources of contamination and the plate effect. In further studies, we could develop more complicated models that can account for sample-related information (e.g., spatial information) and plate-related information (e.g., experiment conditions) for the improved quantification of uncertainty by adding additional layers of hierarchical structures.

Finally, the idea of two-step modeling has greater flexibility and has satisfactory performance in simulation studies. Also, it allows users to refer to the results of previous studies in the literature and to apply them directly to their regression models. For example, possibly due to confidential issues, particular studies have not been able to make their original dataset available to the public, but their summary statistics could be released without any limitation. If researchers wish to conduct follow-up research projects based on such a confidential dataset, the aforementioned concept of a two-step model approach could directly enable them to “borrow” information from additional sources, thus facilitating their own research projects. Furthermore, this two-step modeling approach could be easily extended and generalized to more complicated models based on actual research needs. In our third project, we were motivated by this idea and actually attempted

to extend this two-step modeling approach to make it more suitable for application in cases where the measurement uncertainty cannot be ignored.

## **Chapter 4: Bayesian Two-step Model for Measurement Uncertainty**

### **Adjustment**

#### **4.1 Introduction and background**

##### 4.1.1 Insights from multiphase studies

In statistical modeling and inference, it is highly desirable to obtain all of the required data simultaneously and to build a complete model for conducting inference. However, some studies only provide partial data and information at each cutoff point or stage, such as multiphase studies; thus, researchers must develop methods to integrate the statistical analyses at different stages. Such multi-step models are usually easy to fit and have simpler structures compared with one-size-fits-all models, but challenges arise in aspects such as how to account for the uncertainties in different steps of the analyses in the final inference. Bayesian inference has the natural advantage of better modeling uncertainty using probabilistic modeling. In this study, we aimed to develop Bayesian inference for users who wish to account for the uncertainty in the multi-step analyses with application in environmental health studies.

##### 4.1.2 Challenges in epidemiologic association studies

Uncertainty is one of the critical parts of statistical modeling since, in real-life situations, one might never know the actual underlying data generation and collection process. By carefully designing inference models and imposing assumptions of uncertainty sources, structures, and scales, researchers aim to minimize the impact of measurement uncertainty. Specifically, in this study, we were interested in improving the inference currently in use in epidemiologic association studies with exposure measured in labs, which consists of two steps. In the first step, a wet lab procedure

is conducted to measure the exposure. Typically, this usually only entails point estimation. In the second step, a regression model is used to investigate the associations between the exposure and a health outcome adjusting for potential confounders.

However, some challenges exist in the second-step regression analysis. If the point estimation is simply plugged into the regression model, it is assumed to be a fixed and true value of the exposure. However, lab measurement is far from perfect, and there is always uncertainty associated with the estimation. Furthermore, such direct modeling might not be the optimal solution since now one faces a random-x design instead of a fixed-x design. Thus, it is critical that first, the lab raw measurement evaluation procedures should consider both point estimation as well as the associated standard error estimation, and second, regression association studies should also consider the randomness in the observed key predictor of interest.

#### 4.1.3 Statistical models in handling uncertainty from multi-step modeling

Measurement errors are one of the most fundamental problems in applied statistics since real-life data sets often contain variables measured with error. Researchers need to be careful when dealing with such situations; otherwise, the final inference results might be biased or misleading. Measurement error models usually consist of two key components: the first component is the structure and distribution assumption for the relationship between the observed measurements and the unobserved true values; the second component is the type and scale of the additional information available for evaluating the first component in the model (Carroll et al., 2006). To solve this problem, many statistical methods have been proposed to adjust for measurement errors in a regression setting, especially when measurement errors exist in the covariates. Estimation correction based on conditional moments of measurement errors on those erroneously observed covariates could have less bias and smaller errors, especially in Berkson error models (Whittemore and Keller, 1988). In logistic regression, a functional maximum likelihood estimator was proposed based on the covariates' normal measurement error model, which has superior performance (Stefanski, 1985). Furthermore, the moment reconstruction method, which involves similar ideas as

regression calibration, uses the variance-preserving empirical Bayes estimate of the true values conditional on the outcome variable. It was proposed and demonstrated to have superior performance to regression calibration in logistic regression and case-control studies when measurement errors exist in the explanatory variables (Freedman et al., 2004). In generalized linear models, an adapted version of the Expectation-Maximization algorithm has been proposed, where the M-step iteratively updates the regression coefficients based on the calculated approximate conditional moments of the true covariate values on the observed covariate values in the E-step (Schafer, 1987). For nonlinear models, the measurement error problem was well studied in (Carroll et al., 2006). The small error variance approximation of covariates that might suffer from measurement errors was derived by (Chesher, 1991) and the results indicate that the impact of measurement errors is closely related to the curvature properties of the densities of the remaining error-free covariates. Full Bayesian inference models and iterative conditional models based on smoothing splines and regression P-splines have exhibited improved performance over existing frequentist approaches in certain studies (Berry et al., 2002). Furthermore, the idea of two-stage modeling, where the first stage usually involves the exposure model and the second stage usually involves the health model, was demonstrated to be able to correct finite-sample bias and to generate accurate standard errors in the estimation (Szpiro and Paciorek, 2013).

The problem of measurement errors has also attracted researchers' attention in the field of epidemiologic studies. Measurement errors in one or more of the key predictors of interest are fairly standard in large-scale epidemiological regression studies. A study found that the measurement error's size and type both influence the health effect estimates in air pollution epidemiology (Goldman et al., 2011). While a large proportion of medical studies have reported the potential existence of measurement errors, only a very small subset have investigated or corrected them (Brakenhoff et al., 2018). Simply ignoring such errors would lead to biased estimates and the loss of power in estimating the true associations (Carroll et al., 2006). Many efforts have been made to reduce potential measurement errors in epidemiologic studies, such as standardized wet lab experimental procedures (Tworoger and Hankinson, 2006). In epidemiologic studies, the con-

ditional independence model assumes conditional independence within the disease, measurement, and exposure models with graphical structures and was demonstrated to perform well under certain circumstances (Richardson and Gilks, 1993b). Another commonly used statistical method for correcting such measurement errors is regression calibration, where the true but unobserved exposure is replaced by the conditional mean of the true exposure on the measured exposure and the other error-free covariates in the regression model (Rosner et al., 1989; Carroll and Stefanski, 1990). However, compared with traditional regression calibration, Bayesian measurement error models may have better frequentist properties than MLE procedures in small samples or sparse data situations (Greenland and Mansournia, 2015; Bartlett and Keogh, 2018)). Furthermore, since Bayesian methods provide flexibility in prior specification as well as a probabilistic method of modeling uncertainty, they are well adapted to field measurement error models. Several Bayesian approaches for handling measurement errors in predictor variables have been proposed, including Bayesian conditional independence models (Richardson and Gilks, 1993a), Bayesian logistic regression with a mixture measurement error model (Schmid and Rosner, 1993), and Bayesian methods with item response theory (Fox and Glas, 2003). Moreover, the application of Bayesian methods in epidemiologic research is highly promising. Bayesian regression methods facilitate the realistic use of prior information and provide an alternative toolkit for penalized regression (Greenland, 2007). Furthermore, posterior probabilities provide easily understandable alternatives to  $p$ -values and have much clearer clinical interpretation in the assessment of exposure–disease relations (Dunson, 2001). Additionally, in more challenging situations, such as the existence of highly correlated exposure data due to associations between measured exposures and latent variables, which is fairly standard in epidemiologic research, Bayesian hierarchical models could stabilize the estimation compared with maximum likelihood estimation (MacLehose et al., 2007).

#### 4.1.4 Measurement uncertainty as a source of model contamination

Some researchers would also define the measurement error as a source of contamination in the model. In our first project, we developed the whole Bayesian workflow to solve the problem of

model contamination, particularly with its application to serial dilution assay data. In general, model contamination means that we are not sure whether the model used for inference matches the underlying data generation process. Examining our first and third projects together, we have built a much more complete picture for handling model contamination in applied statistical problems. The first component is model contamination when performing calibration analysis, where the contamination can make the unknown samples do not have the same response curve as the calibration samples. The solution that we have provided involves giving the unknown samples relative flexibility in the parametric form of the distribution while also using the calibration data to provide guidance on the parameter estimation through mixing distributions. The second component is model contamination when using estimation results from previously conducted studies to construct new models, where the contamination is mainly from measurement uncertainty. The solution that we have provided involves using the measurement error model to propagate the measurement uncertainty derived from previous studies into the new models. By combining these considerations and modeling strategies for model contamination together, we are able to produce more reliable estimation results, and thus, we might discover the effect of interest more accurately.

#### 4.1.5 Our contributions

In this study, we proposed a two-step Bayesian inference model by considering measurement uncertainty from multiple sources. The goal was to provide users with an easy-to-use tool to fill in the gap between raw experimental measurement and subsequent analysis by accounting for measurement uncertainty from the previous modeling. Another goal was to enable users to pool together necessary information from multiple studies and to use integrated results for their further analysis. In simulation studies, we compared our proposed methods with the traditional method and the joint inference method under different contamination scenarios. Furthermore, we explored how our proposed method could help investigators study the association between allergen exposure and asthma morbidity in real-life application.

## 4.2 Methods

### 4.2.1 Notations and background

In this study, we developed a Bayesian two-step model that can take advantage of the point estimation and associated measurement uncertainty in the first-step analysis before using this information to obtain more accurate statistical inference results in the second-step analysis. We developed our method for the two-step analysis, but the central idea of measurement uncertainty propagation could easily be extended to multi-step analysis if required. We did not impose any restriction on the type of statistical models used in the first-step analysis; however, for the second-step analysis, since we were more interested in the association studies between one key predictor of interest and the health outcome, we focused on the regression-based models in the second-step analysis in this chapter.

Suppose that  $X$  represents the main predictor of interest in the study, and the first-step analysis provides us with the point estimation  $W$  for  $X$  and the measurement uncertainty  $S$  for this estimation. We define  $Y$  as the health outcome in the second-step analysis. We seek to investigate the association between  $Y$  and  $X$  based on the information provided by  $W$  and  $S$ . Furthermore, we define  $Z$  as the error-free covariates. Our main purpose is to estimate the association between  $Y$  and  $X$  given  $Z$ .

### 4.2.2 Two-step Bayesian measurement error model

In our two-step Bayesian measurement error model, we sought to integrate the measurement error model with the regression model. A natural choice of measurement error model is the following normal model:

$$X_i \sim \text{normal}(W_i, S_i), \tag{4.1}$$

where  $W_i$  and  $S_i$  are fixed and known outputs from the first-step inference model. Here we should notice an alternative way of specifying the measurement error model as follows:

$$W_i \sim \text{normal}(X_i, s_i), \quad (4.2)$$

Model (4.1) is different from model (4.2), which is often assumed in measurement error models. However, different from the standard measurement error model, our first-step Bayesian analysis provides  $W$  and  $S$  as the posterior summary statistics, then equation (4.1) is the preferred modeling approach since it reflects the fact that our second-step model is based on posterior inference of the first-step analysis. Equation (4.1) has the following interpretation: For the true but unknown concentration  $X$ , the first-step model produces the posterior draws as estimates, and  $W$  is the posterior median, which is a point estimation for summarizing the posterior distribution.  $S$  is the standard deviation of the posterior distribution, which could be approximated using the empirical standard deviation of the posterior draws in the Bayesian model fitting process and directly output from the first-step model. By contrast, Equation (4.2) has the following interpretation:  $W$  as the posterior median has followed the normal measurement error model with the mean being the true but unknown concentration  $X$ , but now  $s$  is no longer the standard deviation of the posterior distribution but rather the measurement uncertainty for  $W$ . Usually, in the setting,  $s$  is not directly observable, and we could assign a weakly informative prior to it in the model estimation. Alternatively, if there exists some calibration sample, we could use the calibration sample to estimate the measurement uncertainty for  $W$ . For example, suppose for the calibration sample we know their underlying true parameter of interest  $\theta$ , and after fitting the first-step model we could get some estimates of that parameter of interest based on the posterior summary statistics, which is  $\hat{\theta}$ . Next, we could fit a simple linear regression model of  $\hat{\theta}$  on  $\theta$  and use the residual standard deviation as an estimate for  $s$  in this setting.

In the second-step model, we sought to model the association between  $Y$  and  $X$  using the

following Bayesian regression model:

$$Y_i \sim \text{normal}(\alpha + \beta X_i + \boldsymbol{\eta} \mathbf{Z}_i, \sigma_y), \quad (4.3)$$

Here we are using a random-x regression design, where both the critical predictor of interest and the primary health outcome are random variables.

Another way to think of this is that the classical two-step method, which only takes the point estimation from the first model and treats it as a known constant, is highly similar to the idea of single imputation in the context of missing data problems. The disadvantage of single imputation is that it ignores the measurement uncertainty and introduces bias to the estimation results. By contrast, the Bayesian two-step model, which considers the associated measurement uncertainty, is similar to the idea of multiple imputations in the context of missing data problems. However, instead of taking multiple imputed values to fit the model and aggregate the results in the later stage, our proposed two-step Bayesian measurement error model extracts the summary statistics of multiple imputed values and uses these high-level summary statistics in later analyses.

For the measurement error model part, we currently consider the normal measurement error model. However, in some situations, the normal measurement error model might not be a good approximation for the posterior distribution of  $X$ ; therefore, we may consider other more robust distributions for the model specifications. A natural choice of distribution in this setting is the  $t$  distribution, which has the following model specification:

$$X_i \sim \text{t-distribution}(\nu, W_i, S_i). \quad (4.4)$$

where  $\nu$  represents the degree of freedom in the model specification. We could either use an empirical degree of freedom calculated based on the total sample size minus the number of measurement plates included in the study, or we could assign some prior distribution on  $\nu$  using the family of gamma distribution, such as the  $\text{gamma}(2, 0.1)$ , and let the model estimate it. This alternative measurement error model specification would provide more robust inference results and

might be useful in some cases. Besides the Student's  $t$  distribution, we could also use other robust distributions to model the measurement error.

After modifying this first-step measurement error model using more robust parametric models, the second-step regression model does not need to be modified. This also illustrates the flexibility of our proposed Bayesian two-step inference method. By using the idea of modular modeling, we divide the whole inference into multiple components. For each of the components, we have the flexibility to modify it slightly while not affecting the remaining parts of the inference model, as demonstrated in the above example. Similarly, we could also modify the second step of the model, such as by changing the linear regression model to a logistic regression model without modifying the first step model.

#### 4.2.3 Measurement uncertainty propagation

In most cases, we can simply plug in the estimated measurement uncertainty in the method proposed above in the position of  $S$ . However, sometimes we cannot directly use the original measurement uncertainty of previous studies in the subsequent analysis if we want to perform some transformation on the raw estimation results. For example, log or squared-root transformation on certain covariates might sometimes be a better choice for modeling. In such a setting, one must think about how to appropriately plug in the measurement uncertainty after transformation. Here, we provide two options for users to account for such transformations. The first option is relatively simple, but it requires the first-step analysis to include Bayesian inference procedures. In such a setting, we could simply apply the corresponding transformation into the posterior draws of the parameters of interest in the first-step analysis and then calculate the empirical standard deviation of the transformed posterior draws and treat it as the new measurement uncertainty estimation.

The second option applies to broader application cases with the use of the delta method. Specifically, if we know the exact functional form of the transformation and the first-order derivative exists in the point estimation of the mean of the parameter of interest, with the delta method, we could directly calculate the corresponding transformed variance estimation. This option has the

advantage of being applicable in nearly all common transformations wherever the delta method is applicable. However, since the delta method also has some parametric assumptions, and if these assumptions are not satisfied, the performance of the delta method might be affected.

#### 4.2.4 Prior specification and Bayesian computation

Since not many parameters are included in our proposed Bayesian two-step measurement error model, the prior choice is relatively straightforward. The primary logic behind the prior specification is to get easily interpretable regression coefficients in our second-step regression model. To be more specific, suppose that we have made some appropriate standardization on  $Y$  and  $Z$  such that they have mean zero and variance one. Note that this standardization could either be on an absolute scale or a relative scale to a specific subpopulation. Then, a natural choice of prior for the regression intercept is as follows:

$$\alpha \sim \text{normal}(0, 1), \tag{4.5}$$

For the regression coefficient, since the baseline covariates have been standardized, a natural choice is to use the following prior to reflect the baseline level effect:

$$\eta \sim \text{normal}(0.5, 0.5), \tag{4.6}$$

For the coefficient of the key predictor of interest, the following prior is applied:

$$\beta \sim \text{normal}(0, 1), \tag{4.7}$$

For the error term, since we do not have any additional information for it and the health outcome has been standardized, we use the following weakly informative prior:

$$\sigma_y \sim \text{normal}^+(0, 1). \tag{4.8}$$

For the Bayesian model fitting and posterior inference, we use the CmdStanR interface in R,

which provides a light and quick interface to the no U-turn sampler (NUTS) algorithm in Stan (Homan and Gelman, 2014; Carpenter et al., 2017). With the help of the NUTS, the algorithm selects an appropriate number of leapfrog steps in each iteration to allow the proposals to traverse the posterior without performing unnecessary work. We also use the ESS and the  $\widehat{R}$  diagnostic to monitor the convergence of the algorithm (Gelman and Rubin, 1992).

Then, we performed some Bayesian computation diagnoses on the computation results in simulation studies and real-life examples. The effective sample size is large enough for all of the parameters in the model, and their  $\widehat{R}$  values are around 1. Compared with the joint modeling approach that we proposed in the second project, the computation time for the current model is much shorter and can be finished in seconds, which provides users with a faster alternative option for dealing with the applied research problems.

#### 4.2.5 Connection between measurement uncertainty and missing data problem

Although we have majorly discussed the measurement error problem in this project, here we would also like to briefly discuss its connection with the missing data problem. Commonly used techniques in the missing data research field, such as the multiple imputation framework, have been proposed to solve the measurement error model. Given our two-step problem setup framework, this connection between the measurement error and missing data problems becomes clearer. The central question that we are interested in is how to apply the inference results from the first-step model to the second-step model, given that the first-step inference results are not true values. The first way to think about this problem is to treat the first-step inference results as observed measurements with measurement errors and use the corresponding measurement error models to solve the problem. This is exactly the approach that we have proposed in this project. The second way to think about this problem is to treat the unknown true values as a missing data problem. If the first-step model uses Bayesian inference methods, those posterior draws for the target of interest could partially serve as different imputation results in the multiple imputation framework. Then, we could use corresponding missing data models to solve this problem by combining inferences

from multiple posterior draws together. Although some subtle differences exist in the central ideas of these two ways of thinking, they are also fairly similar, especially in the scenario where the first-step model also uses the Bayesian inference model and the output mean and standard deviation estimates for the unobserved true values are based on posterior distribution summary statistics. If so, the difference mainly lies in whether we summarize the posterior distribution first and then input the second-step model or directly input the posterior draws into the second-step model and then perform the aggregation and summarization later.

#### 4.2.6 Robust model specification for measurement uncertainty

We could also further consider relaxing the parametric assumptions of the proposed measurement error model. In this study, we used the normal measurement error model, which performs well in most situations. We also considered more robust parametric models, such as Student's  $t$  distribution for the measurement error model. Following this robust modeling strategy, we could even further relax the parametric assumptions of the measurement error model and use Bayesian semiparametric or nonparametric models instead. The Bayesian semiparametric regression model usually includes a generalized linear model combined with specific carefully designed priors together with hierarchical structures. Furthermore, nonparametric transformations such as Bayesian P-splines, which could simultaneously estimate the smooth functions and smoothing parameters, might also be beneficial in modeling measurement errors. In this case, we would not be restricted by the parametric form of the measurement error; instead, we could only assume that the measurement error is symmetrically distributed. Furthermore, we could even further extend this idea to the situation where the measurement error is not symmetrically distributed, which is fairly common in serial dilution assay lab measurement, where the estimations are usually either upward or downward-biased with measurement errors. Then, we might use a mixture of measurement errors, where usually a symmetrically distributed measurement error model is combined with a skewed distribution to account for such additional measurement bias. In summary, there is no limit to the possible measurement errors that we could specify in our proposed Bayesian two-step inference

approach, and researchers could easily adapt the measurement error part to their actual research needs.

#### 4.2.7 Comparison with competing methods

In this subsection, we discuss the performance of our proposed Bayesian two-step inference method with the competing methods, namely joint modeling and naive two-step modeling without considering measurement uncertainty. We first need to define these two competing methods. The joint modeling method involves using a single large model to incorporate all available information at hand and model the dependence between different parameters and information sources simultaneously. An example of the joint modeling approach is the Bayesian joint model we have proposed in our second project. The joint modeling approach correctly models the underlying problem, but it usually involves a much more complicated modeling process with a lack of flexibility to be applied in a similar setting. The naive two-step approach means simply plugging in the point estimation results from any previous studies and treating them as known and fixed values in the following analysis. The naive two-step modeling is easy to implement, whereas the disadvantage is that it ignores the potential measurement uncertainty and thus might introduce unexpected bias in the following analysis.

In most applied epidemiologic studies, the naive two-step method is commonly used because of its simplicity. Furthermore, in some situations, the estimation uncertainty in the previous studies are not available, especially when the analysis is performed somewhere else. However, in the first project, since our Bayesian method provides both the point estimation as well as the corresponding estimation standard error, thus giving us the ability to apply our proposed Bayesian two-step model. We compared the performance of the three methods in various simulation settings to investigate the optimal solution to apply to real-life examples.

### 4.3 Bayesian workflow for two-step inference model

This section illustrates the idea of the Bayesian workflow and the corresponding details and efforts that we made in the project. In an applied research problem, if one is concerned about the potential existence of measurement errors in the observed values, one typically thinks about the following two general aspects regarding the measurement errors – namely the form and scale of the measurement error. This information might not be directly available to researchers most of the time, and the challenge comes with the fact that one might never know what the true values should be; therefore, it is necessary to construct different models and perform model evaluation and comparison across broader model categories. To be more specific regarding the problem of measurement errors, researchers would usually like to start with the specification of the form of the measurement error. A starting point is to either choose a parametric form of the measurement error or a semiparametric measurement error form or even a nonparametric measurement error form; here, the parametric model would be clear and straightforward but might also be too restrictive and suffer from the risk of model misspecification in some scenarios. The semiparametric and nonparametric forms have more flexibility and robustness, but they might suffer from a loss of estimation efficacy and statistical power in some scenarios. After specifying the form of measurement error, we should consider how to specify the scale of the measurement error, which is critical if choosing a strong parametric assumption for the measurement error’s distribution. A starting point is to assume that every observation has shared the same measurement error scale. However, sometimes, this homogeneous measurement error scale specification might be too ideal, and a slightly more complicated specification would allow internal hierarchical structures and clustering within the data. Within each smaller group, all the observations share the same measurement error scale. To further increase the flexibility, we could even assume that every observation has different measurement error scales. Typically, one must construct the initial model based on the data available and any additional information regarding the applied problem itself. In this project, we started with parametric measurement error models and used the most commonly used and simple model of nor-

mal measurement error. For the scale, we started with the most flexible approach to specification, where we allowed each observation to have its own measurement error scale. This specification required us to have much more additional information regarding the individual-level measurement error scale. Because of the modular construction idea, we could always expand or simplify these model specification components as needed.

The aforementioned measurement error model is usually called the exposure model, which focuses on how to relate the observed measurements with potential measurement error to the underlying unobserved true values. Although this exposure model is critical since it quantifies the degree of uncertainty for the observed data, in many applied problems another model is also required to relate the measurement error-corrected key parameter of interest to the outcome or response. Especially in the setting of epidemiologic studies, the exposure model is usually used to model some of the individual's exposure covariates, which might be measured with errors. Then, a disease model is usually constructed to investigate the association between these covariates and the health outcomes. For the disease model, one should always start with something simpler and then gradually add model complexity. The most commonly used association model is the regression model, which has a relatively simple parametric form and straightforward interpretation, and moreover, it generally performs well in a variety of scenarios. We started with a univariate regression model that involved the health outcome as the response and the covariate in the exposure model as the explanatory variable. However, this model specification ignores the potential existence of confounding variables, and thus, we needed to extend the univariate regression model to a multivariate regression model, which includes other necessary baseline covariates for the observation. In some more complicated cases, traditional regression models might not be able to explain the underlying association between the outcome and covariates; then, we could extend the regression model to nonlinear models or mixed-effect models, which are both very commonly used in epidemiologic association studies. Here, we note that the exposure model and the disease model are not dependent on each other in some sense. For example, if we switch the normal measurement error exposure model to another parametric exposure model, and if this exposure could reflect the

underlying data generation truth, it should not affect the final inference model much if we still use the same regression disease model. This also illustrates the idea of modular modeling framework building.

Since we proposed Bayesian models, the major model evaluation focused on summarizing the posterior distribution and checking whether the final outputs made any practical sense given the specific applied problem's context. Furthermore, we could use posterior predictive checking, robustness checking, and prior predictive checking to further examine the model fit. Based on the model evaluation results, we could also return to our initial model and think about whether the inference results were satisfactory and trustworthy. For the measurement error model, since for real-life datasets, one would never know what the underlying generative model for the measurement error is and what exactly the true values would be, one would usually start with simpler models and possibly use some visualization tools to compare the predicted true values with the observed values, perhaps obtaining some model modification insights. Furthermore, we could perform sensitivity analysis and robustness checks on the prior distribution and evaluate our model performance in some more varied situations. Furthermore, by carefully designing multiple simulation scenarios and repeating the simulation for many iterations, we could understand the advantages and disadvantages of our model more thoroughly. In addition, the modular modeling approach allowed us to adapt our model easily if the problem setup changed. For example, if some standardization or transformation of certain parameters was required, additional data sources were available, or new model restrictions were imposed, we could adjust them accordingly without changing all of the model parts.

#### **4.4 Application studies**

In this project, we will also apply our proposed Bayesian two-step model accounting for measurement uncertainty to the NYCNAAS dataset. We will follow the same procedure mentioned in our second project by only looking at the asthmatic children and treating the continuous response, which is the exhaled nitric oxide (NO) measurement as our main response of interest. For the co-

variates included in the regression model, our key predictor of interest is still the indoor dust mite allergen concentration, and we have also adjusted for the same set of baseline covariates as we have specified in our second project. To be more specific, we will apply two competing methods to the NYCNASS data in this project. The first method is the so-called 'naive two-step model' where we will use the Bayesian mixture model for contamination that we have developed in our first project and fit it to each of the measurement units respectively in the first step and collect the point estimation, for example, the posterior median, of allergen concentration on each unknown sample. Then in the second step, we will treat these point estimations as true and known values and directly use the following regression model

$$Y_n \stackrel{ind}{\sim} \text{normal}(\alpha + \gamma\theta_n + \eta_1 Z_{1n} + \eta_2 Z_{2n} + \eta_3 Z_{3n} + \eta_4 Z_{4n} + \eta_5 Z_{5n}, \epsilon_n). \quad (4.9)$$

where  $\theta_n$  represents the indoor dust mite concentration point estimation for patient  $n$ , which in this method is the posterior median output from the first step and it is our main predictor of interest.  $Z_{1n}$  represents the sex for patient  $n$ ,  $Z_{2n}$  represents the Black race indicator,  $Z_{3n}$  represents the baseline family smoker existence status,  $Z_{4n}$  represents the baseline parental education level, and  $Z_{5n}$  represents the baseline maternal asthma status.

In contrast, the second method is the 'Bayesian two-step model' that we have developed in this project. For the Bayesian two-step model, the first step is exactly the same as the naive two-step model, but in the second step, although we are still using the same structure of the Bayesian linear regression model mentioned above, in the key predictor  $\theta_n$  instead of using the point estimate output directly from the first step model, we introduce another layer of normal measurement uncertainty as we have proposed in the methods section.

After fitting these two methods to the NYCNAAS dataset, we have summarized their Bayesian regression coefficients' point estimation and the corresponding 80% uncertainty interval in Figure 4.1. For comparison purpose, we have also included the modeling fitting results by the Bayesian joint model. From Figure 4.1, we have several observations: First, both the naive two-step model and the Bayesian two-step model have identified that exposure to indoor dust mite concentration is

a significant factor associated with the exhaled nitric oxide (NO) measurement in asthmatic children, which is consistent to our findings by the Bayesian joint inference model both in the sign and scale of the regression coefficient. Second, although the coefficients for the allergen concentration are slightly different for these two methods due to the measurement uncertainty adjustment, the relative scale is pretty close and might not make a huge impact in real-life applications. We observed a little shrinkage effect on the allergen concentration coefficient estimates after taking the measurement uncertainty into consideration and this might be explained by the random noise in the data. Thirdly, compared to the estimation results using the traditional calibration method commonly used in labs, the Bayesian joint model, the naive two-step model, and the Bayesian two-step model all could generate a more accurate estimation of indoor allergen concentration by overcoming the potential downside bias. Finally, for the remaining baseline covariates included in the regression model, both models have shown that there is no strong evidence that they are associated with the exhaled nitric oxide (NO) measurement in asthmatic children, which is also consistent with our findings by the Bayesian joint inference model.

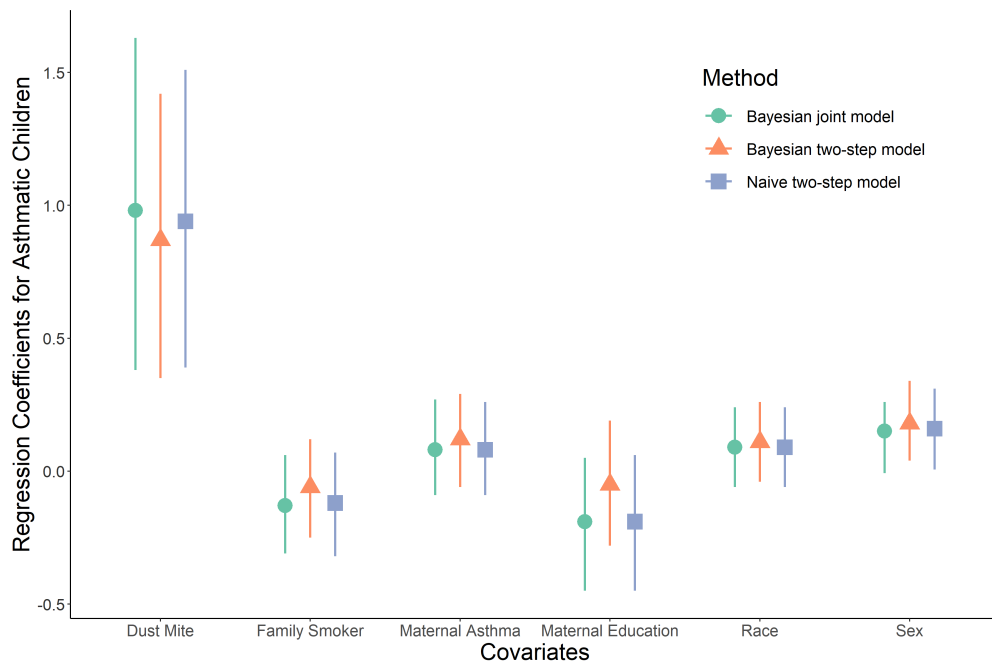


Figure 4.1: Bayesian regression coefficients summary for NYCNAAS data, where the posterior median and 80% uncertainty interval are presented.

In addition, we have also compared the point estimates of dust mite concentration estimation generated from the naive two-step model and the Bayesian two-step model in Figure 4.2. After fitting the reference line with intercept 0 and slope 1, we have found that these two model outputs are pretty similar to each other, especially near the lower end of the plot. Also, we have observed the existence of measurement uncertainty and found that the measurement error is inflating as the underlying estimates increase in value.

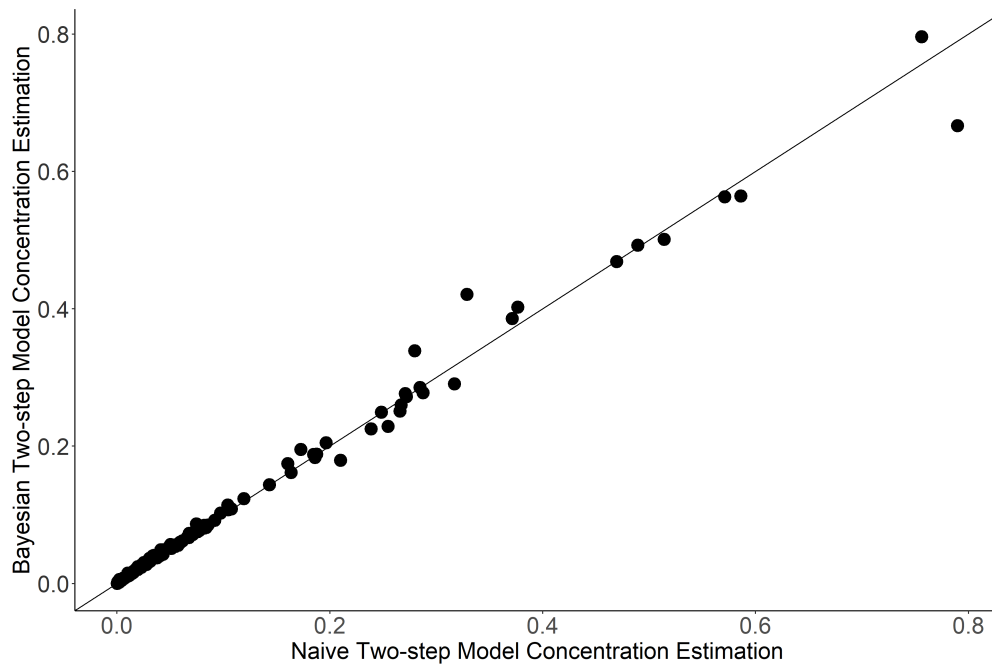


Figure 4.2: Comparison of indoor allergen concentration estimation by the naive two-step method and Bayesian two-step method.

#### 4.5 Simulation studies

Besides the real-life dataset, we also sought to perform several simulation studies to more effectively and accurately evaluate the performance of our proposed methods. Since we could control every detail in the simulation studies, we could obtain a clearer picture of the underlying data generation process and which aspects of the performance we were interested in testing. In this study, we were majorly interested in two simulation scenarios. In both simulations, we simulated raw lab

measurement data from serial dilution assays on indoor allergen and individual-level health record data from large-scale epidemiologic studies. In the first simulation scenario, we were interested in the performance of our proposed method when the first-step analysis outputs were both point estimation and the associated standard error estimation. For simplicity, we assumed that there was no sample contamination in this setting, and also that the competing method used only the point estimation output from the first-step model and treated it as a known fixed constant in the second-step analysis. In the second simulation scenario, we were interested in the more complicated cases where some sample contamination might exist in this setting. We compared the performance of our proposed two-step model to the naive two-step approach that used only the point estimation output from the first-step model and treated it as a known fixed constant in the second-step analysis, and the joint modeling combined the first-step and second-step models in a single model. The major evaluation metric was whether any efficacy gain or loss of the proposed two-step model occurred compared with the naive two-step model and the joint modeling.

#### 4.5.1 Model comparison and evaluation in the setting without contamination

In this simulation scenario, we are interested in comparing the performance between our proposed Bayesian two-step model, which takes measurement uncertainty into consideration, and the naive two-step approach, which treats the first-step point estimation results as the true value and ignores the measurement uncertainty. These two methods share the same model in the first-step analysis, which is an adapted version of the Bayesian contamination model we developed in our first project, only removing the contamination specification since we are no longer assuming the existence of contamination in this setting. We are planning to conduct the simulation in a multi-measurement-unit setting in the context of serial dilution assay data, where within each replicate of simulation, we simulate both multiple measurement-unit raw serial dilution assay measurement data as well as individual-level health outcomes and covariates. The simulation setup is similar to what we have done in our second project, where the major modification is that we are now assuming there is no contamination in the data.

In step 1, we plan to simulate the raw measurements of the standard calibration samples  $w_{ik}$  and the raw measurements of the unknown samples  $w_{ijk}$  for each dilution level  $i$  on each measurement unit  $k$  and unknown sample  $j$ . Here we choose to simulate 10 measurement units, where on each unit, there are 1 standard calibration sample and 23 unknown samples. This setup is to mimic both a typical real-life serial dilution assay unit setup as well as to match the study sample size in the epidemiologic association studies. Before simulating the raw observations, we first need to simulate the calibration curve parameters  $\beta_k$  across plates using the following independent normal model with different scales:

$$\beta_{1k} \sim \text{normal}^+(\beta_1, 2). \quad (4.10)$$

$$\beta_{2k} \sim \text{normal}^+(\beta_2, 1000). \quad (4.11)$$

$$\beta_{3k} \sim \text{normal}^+(\beta_3, 5). \quad (4.12)$$

$$\beta_{4k} \sim \text{normal}^+(\beta_4, 1). \quad (4.13)$$

Here  $k$  represents the unit indicator, and we use  $\beta_1 = 1$ ,  $\beta_2 = 15000$ ,  $\beta_3 = 8$ ,  $\beta_4 = 1.5$  based on real-life empirical evidence. And we also set the known initial concentration  $\theta_0 = 125$  and apply 2-fold dilutions based on real-life serial dilution assay measurement unit setup ( $d_{ik} = 0, 0, 1, 1, 1/2, 1/2, \dots, 1/2048, 1/2048$ ). Then for each measurement unit  $k$ , we assume a constant variance scale for the standard calibration samples, and thus we have:

$$\log(w_{ik}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_0 d_{ik}, \beta_k)), 0.1), \quad (4.14)$$

For each of the unknown sample concentrations  $\theta_{jk}$ , we randomly sample it from  $\text{normal}^+(10, 2)$  with dilution factors  $d_{ijk} = 1/10, 1/100, 1/10000$  and thus have:

$$\log(w_{ijk}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_{jk} d_{ijk}, \beta_k)), 0.1). \quad (4.15)$$

In step 2, we would like to simulate the public health response variable  $Y$ . For each of the

total 230 participants, we assume they have a continuous response variable  $Y$  that follows a normal distribution with mean depending on baseline covariates such as age ( $x_1$ ), sex ( $x_2$ ), race ( $x_3$ ), and a key predictor of interest, which is indoor allergen concentration  $\theta_{jk}$ . Among these covariates, we assume the age distribution follows uniform (8,16), the sex ratio is 0.5:0.5, and the race ratio is 0.45, 0.2, and 0.35, respectively. Then we simulate the  $Y$  response variable using the following normal distribution:

$$Y \stackrel{ind}{\sim} \text{normal}(0.5 - 0.05 * x_1 + 0.1 * x_2 + 0.1 * x_3 + 0.4 * \theta_{jk}, 1). \quad (4.16)$$

We repeated the simulation 500 times and summarized the estimation results for the target of interest, which is the regression coefficient for allergen concentration in the regression model among these two methods. Note that in the Bayesian regression model, we use two indicator variables to represent the three levels of the race variable. We find that since there is no contamination existing in the raw data, the corresponding estimation uncertainty in the first-step model is relatively small, so even the classical two-step approach, which treats the first-step point estimation results as the true values, performs pretty well. However, we could also observe that after taking the measurement uncertainty into consideration, our Bayesian two-step model is even performing better. Also, we have included the benchmark one-step model in this scenario, where a one-step Bayesian regression model is conducted using the underlying true values of allergen concentration generated in each simulation iteration as a benchmark inference for evaluating the model performance. From Table 4.1, we could find that our Bayesian two-step model has pretty satisfactory performance even compared with the benchmark one-step model.

	Bias	RMSE	80% CP	95% CP
Naive Two-step Model	-0.0078	0.0395	79%	93%
Bayesian Two-step Model	-0.0018	0.0035	81%	95%
Benchmark Model	0.0006	0.0033	80%	95%

Table 4.1: Comparison of the naive two-step model and Bayesian two-step model performance with the benchmark model on regression coefficient of the key predictor of interest (true value is 0.4) under no contamination with 500 iterations, RMSE: root mean squared error, CP: coverage probability.

Figure 4.3 provides a brief summary of what goes behind the scene in one iteration of the simulation draws, where we have compared the true values of 230 simulated allergen concentrations, the posterior median of the allergen concentration in our first-step inference model ( $W$ ) and a set of the posterior draw of the random allergen concentration generated in our Bayesian two-step model ( $X$ ) in our simulation scenario 1. From Figure 4.3, we observe that the posterior draw of the random allergen concentration generated in our Bayesian two-step model ( $X$ ) is closer to the true allergen values than the posterior median of the allergen concentration in our first-step inference model ( $W$ ).

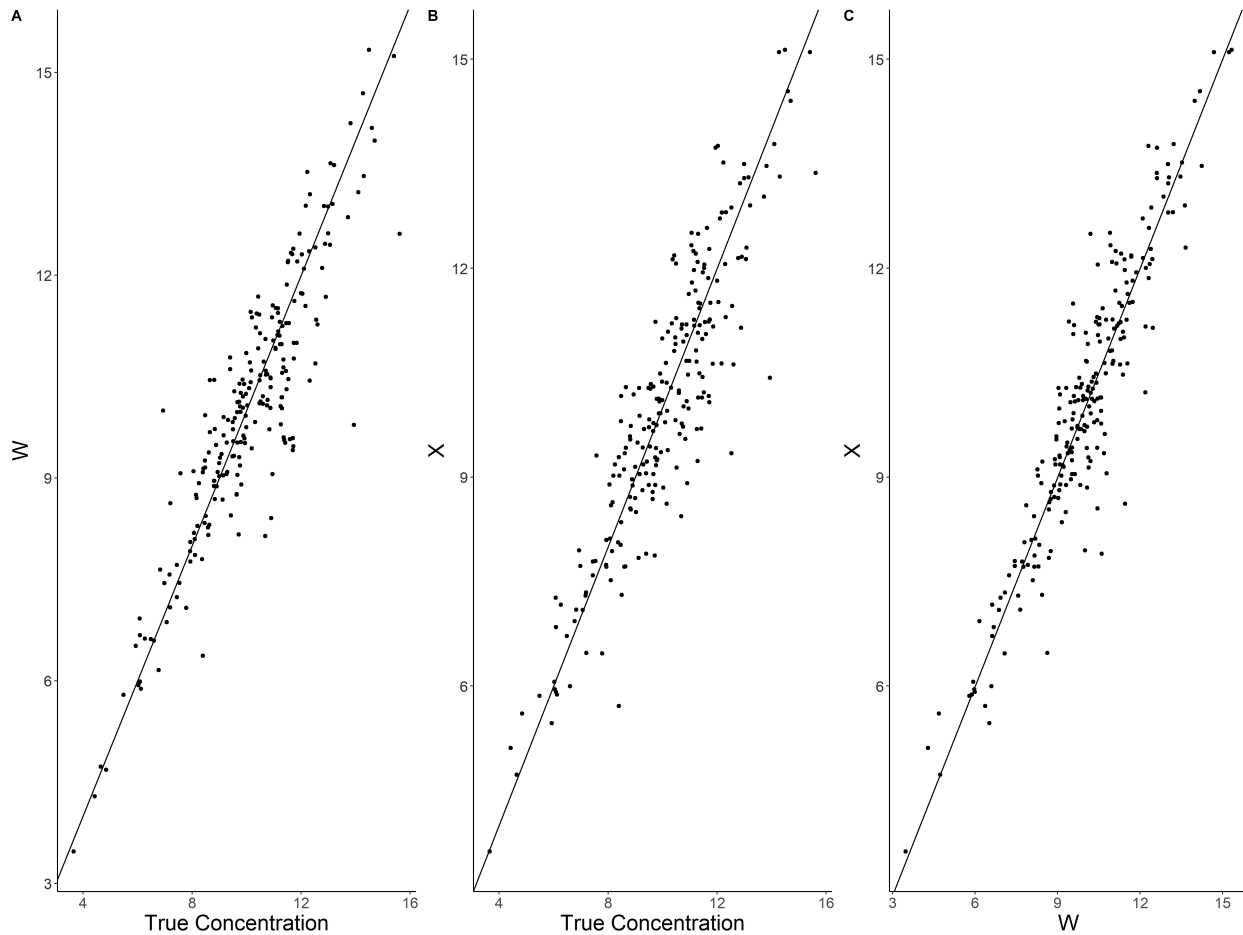


Figure 4.3: Comparison of true concentration, the posterior median of concentration estimation in the first-step model ( $W$ ) and posterior draw of concentration in the Bayesian two-step model ( $X$ ) in simulation scenario 1 (with a line with slope 1 and intercept 0 as reference).

#### 4.5.2 Model comparison and evaluation in the setting with sample contamination

Simulation setting 2 is pretty similar to what we have in the simulation scenario 1 except for the following difference: first, we allow unknown samples to have contamination in setting 2; second, since there might be some contamination in the samples, in the first-step analysis of the Bayesian two-step model and classical two-step model, we use the Bayesian contamination model developed in our project 1; third, for each of the unknown sample concentrations  $\theta_{jk}$ , we now random sample

it from  $\text{normal}^+(2, 0.5)$  with dilution factors  $d_{ijk} = 1/10, 1/100, 1/10000$ ; finally, we also include the Bayesian joint model developed in our project 2 for comparison. The simulation process is pretty similar, the only difference lies in the generation process for the raw measurements of the unknown samples.

Now for the 23 unknown samples on each measurement unit we let 5 of them be contaminated with contamination factor  $\delta_{\beta_{2jk}} = 1$  and  $\delta_{\sigma_{jk}} = 1$ , so their observed values are sampled from the following:

$$\log(w_{ijk}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_{jk}d_{ijk}, \boldsymbol{\beta}_{jk})), 0.1 * e^{\delta_{\sigma_{wk}}}), \quad (4.17)$$

where  $\boldsymbol{\beta}_{jk} = (\beta_{1k}, \beta_{2k}e^{\delta_{\beta_{2jk}}}, \beta_{3k}, \beta_{4k})$

For the remaining 18 uncontaminated unknown samples on each measurement unit, we use the same data generation as we have in simulation 1:

$$\log(w_{ijk}) \stackrel{ind}{\sim} \text{normal}(\log(g(\theta_{jk}d_{ijk}, \boldsymbol{\beta}_k)), 0.1). \quad (4.18)$$

We repeated the simulation 500 times and the performance for the above-mentioned four methods has been summarized in Table 4.2. From Table 4.2, we have the following observations: First, we have observed that both the classical two-step model and Bayesian two-step model perform worse in simulation scenario 2 compared to scenario 1, which could be mainly explained by the existence of contamination and the corresponding increased measurement uncertainty. Especially the classical two-step method shows really poor performance. Second, we have found that the Bayesian two-step model performs more like an intermediate method between the naive two-step model that totally ignores the measurement uncertainty and the Bayesian joint model which considers every possible source of uncertainty, which also makes sense. Finally, despite the model fitting complexity and longer model fitting time, the Bayesian joint model we proposed in our second project has the best performance among these three methods in terms of RMSE.

	Bias	RMSE	80% CP	95% CP
Naive Two-step Model	-0.0304	0.18	74%	92%
Bayesian Two-step Model	-0.0139	0.16	80%	96%
Bayesian joint Model	0.0135	0.15	80%	95%
Benchmark Model	0.0065	0.13	80%	95%

Table 4.2: Comparison of three methods' performance on regression coefficient of the key predictor of interest (true value is 0.4) under contamination with 500 iterations, RMSE: root mean squared error, CP: coverage probability.

Similar to what we have in the simulation scenario 1, Figure 4.4 summarizes one iteration of the simulation draws, where we have compared the true values of 230 simulated allergen concentrations, the posterior median of the allergen concentration in our first-step inference model ( $W$ ) and a set of the posterior draw of the random allergen concentration generated in our Bayesian two-step model ( $X$ ) in our simulation scenario 2. In this figure, although the posterior draw of the random allergen concentration generated in our Bayesian two-step model is of similar distance to the true allergen values compared with the posterior median of the allergen concentration in our first-step inference model, the regression coefficient estimated by the classical two-step model has larger bias compared to the Bayesian two-step model.

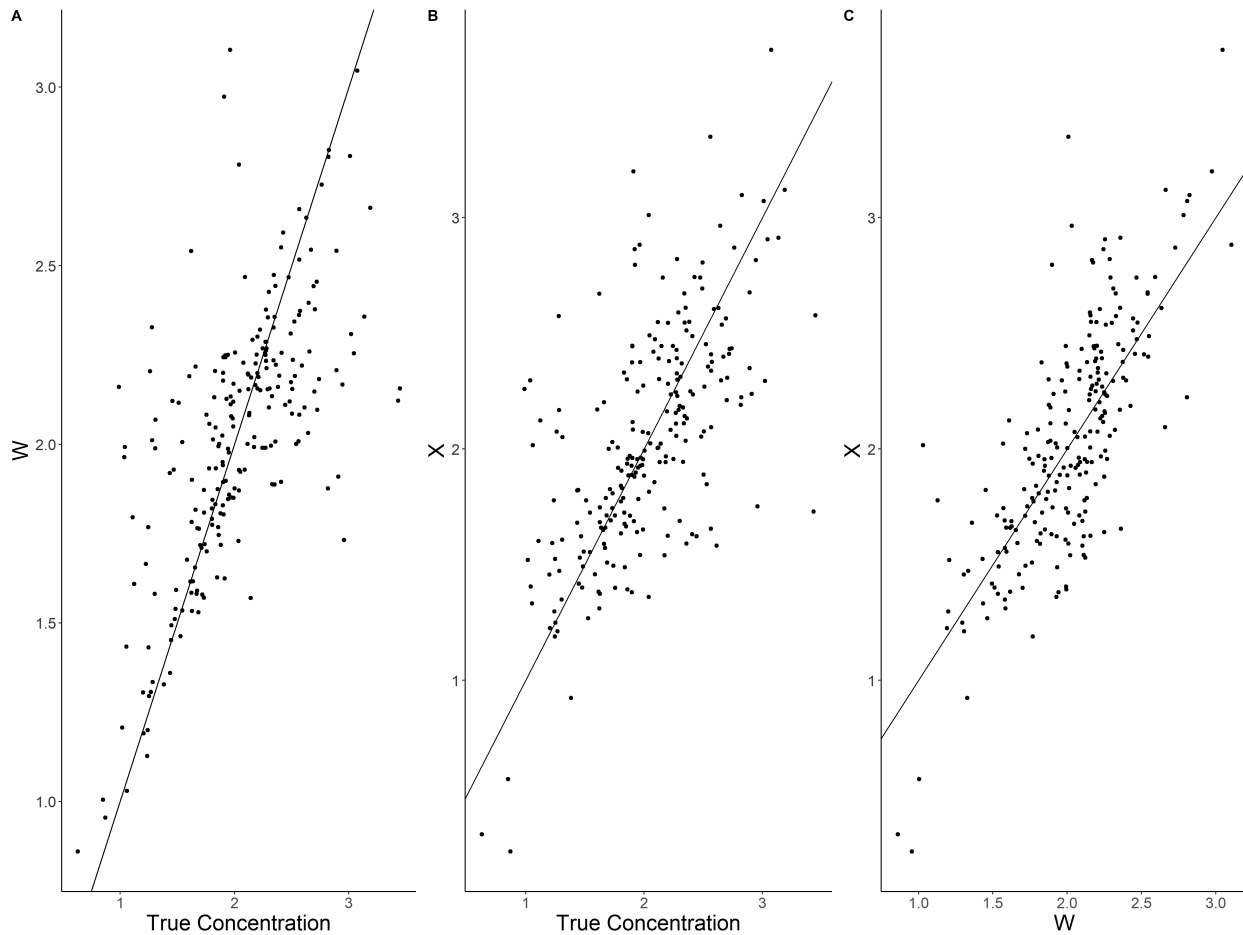


Figure 4.4: Comparison of true concentration, the posterior median of concentration estimation in the first-step model ( $W$ ) and posterior draw of concentration in the Bayesian two-step model ( $X$ ) in simulation scenario 2 (with a line with slope 1 and intercept 0 as reference).

#### 4.6 Discussion

Measurement uncertainty is a critical component when conducting statistical inference, especially in a situation where one wishes to plug some existing estimation results into a new analysis. This study proposed a general Bayesian two-step inference model and applied it in real-world children's asthma studies. By considering the measurement uncertainty, we built a bridge between raw wet lab measurement estimations and later-stage large-scale epidemiological studies. Furthermore, by

taking advantage of Bayesian regressions and considering a random-X instead of a fixed-X design, we could estimate the association between health outcomes and key predictors of interest more accurately. We used both empirical and simulation studies to demonstrate the advantages of our Bayesian two-step measurement error model, among other competing methods.

On the other hand, our study also provides users with an innovative multi-step workflow that is modular and has the high flexibility required to adapt to new applied problems. We did not impose any parametric assumptions on what kind of inference results we could use from the first-step model, which would provide users with more flexibility to seek more alternative information sources. Furthermore, each part of our proposed model can be modified accordingly to the user's specific needs while not affecting the remaining part, but they can also be integrated together as a whole to provide reliable inference results. This would allow users to easily adapt our work to their studies, and they could even extend our two-step measurement error model to a multi-step measurement error model based on their needs.

Besides, our proposed Bayesian two-step inference model is also closely related to the Bayesian contamination model we developed in our first project. Traditionally, researchers are usually satisfied with getting one point estimation for the target of interest and then either use it for upcoming analysis or draw conclusions based only on point estimation. However, the Bayesian two-step inference model illustrates the potential efficacy and estimation accuracy gain if we could get information for both the point estimation as well as the associated uncertainty measurement. This in turn shows the advantage of our first project model since it could provide such information easily to users.

Nevertheless, our study also has some limitations. First, our measurement error model assumes normality. Although this is usually the most common choice in other measurement error models, there might be some cases where this normal measurement error model is not the most appropriate choice. When the actual measurement error distribution is more complicated, perhaps the normal model would not be able to do the job very well. Second, our model requires every estimation need to have one corresponding measurement uncertainty, but this might be too ideal for some

applied problems. It would be interesting to develop some imputation models to infer the missing uncertainty measurements for some proportion of raw estimations. For example, if two studies are fairly similar to each other, we might borrow some information to impute the missing uncertainty measurements.

## **Conclusion or Epilogue**

Model contamination and measurement error bring additional challenges for researchers in many real-life application studies. Although traditional statistical methods including calibration inference and repeated measurements have been well studied, there are still some scenarios where these methods perform poorly. Bayesian methods have the natural advantage of better quantifying the uncertainty as well as more flexible modeling strategies through hierarchical models. In this thesis project, we have made several contributions to the field of Bayesian inference model development with a particular focus on the applied dataset of serial dilution assay measurement data. First, we developed Bayesian contamination models that can handle the model contamination problem in general applied statistics. Second, we developed Bayesian global calibration and joint models for multiple data sources, even with potentially different data types, which can model multiple measurement units at the same time and account for uncertainty in the regression covariates. Third, we developed a Bayesian two-step model that can serve as an easy-to-use tool in situations where global calibration and joint modeling might not be applicable by considering measurement uncertainty. Fourth, we illustrated the idea of a Bayesian inference framework building on all the aforementioned three proposed methods' development phases, where the central idea could be easily generalized to other similar research questions. By performing various simulation scenarios and applying the proposed methods to real-life datasets, we obtained an enhanced understanding of our methods and models as well as demonstrated their superior performance compared with common competing methods in a variety of settings.

We started by examining the general problem of model contamination in applied statistics,

where researchers face the challenges of the underlying data generation process being unknown and inestimable. Even though the calibration sample and data are presented, there is no clear clue or guidance about how researchers should incorporate this calibration information into a scenario where model contamination could exist. The solution that we provided uses Bayesian contamination models, where the central idea is to use the flexibility of Bayesian modeling to account for uncertainty in every aspect of the model of interest. Furthermore, by incorporating the idea of Bayesian mixture modeling, we were able to allocate the measurement uncertainty to the parts that correspond to the known calibration inference pattern as well as to the unknown contamination model. Then, we further investigated a specific model application, namely a serial dilution data assay, which usually has multiple measurements available for the sample and contains internal calibration data sources. We investigated the potential challenges of analyzing serial dilution assay data underlying the possible existence of sample contamination as well as the poor performance of commonly used statistical inference methods in the field using real-life datasets collected from the lab. Then, we proposed a general Bayesian workflow for model contamination to provide readers and users with a practical guideline for conducting initial model construction, Bayesian computation, model comparison, model evaluation, as well as model adaptation, along with examples and specific aims in the application of serial dilution assay data. The practical advantages of our proposed methods include high estimation accuracy, a flexible model structure, robust inference results, and additional information for follow-up. In terms of high estimation accuracy, our model could estimate the unknown true values accurately with shorter uncertainty intervals, especially for those small measurements traditionally not detectable due being below the detection limit. Furthermore, our model is not restrictive to the specific form of contamination pattern that we specified, and other users could easily design the most appropriate form of contamination for their own research projects either based on real data evidence or domain knowledge. Finally, our Bayesian contamination model was not only able to output the estimation for our target of interest but also provided stratified estimation results based on contamination status as well the posterior probability of being contaminated for each sample. This information could further pro-

vide lab technicians and researchers with a flag or signal to further investigate the potential causes of contamination behind the scenes. We used multiple simulation studies to demonstrate that our proposed method performed much better than other competing methods across various degrees of contamination, especially severe contamination. We also used simulations to demonstrate that our model's performance was satisfactory even under the most challenging practical setting, where all the observations were fairly close to zero; therefore, not much information was contained in the data, and the traditional inference method could not be applicable here. In real-world data examples, we demonstrated that our method could generate reasonable estimation results for samples that were inestimable by the traditional inference method, which greatly increases the data utilization efficacy for further research projects.

After observing the success of the Bayesian contamination model, we further developed a more general Bayesian joint model for model contamination and statistical association studies. Given that the Bayesian contamination model could only be applied to a single measurement unit, a natural extension of the proposed Bayesian joint model is to endow it with the ability to aggregate and analyze multiple measurement units together while also maintaining good estimation performance on each single measurement unit. By using the idea of global calibration and hierarchical models, we were able to model the within- and between-unit heteroskedasticity efficiently with relatively simple structures using partial pooling. In addition, we further extended the Bayesian joint model to allow for multiple types of data input sources as well as multiple goals of estimation outputs, including both raw exposure measurement data as well as public health-related data. The advantages of this further model extension would be that it would allow researchers to have a much more accurate estimate of measurement uncertainty of the estimation of unknown exposure. By incorporating such measurement uncertainty directly into regression-based association models, we could apply the more appropriate random-X regression design and yield more reliable results. Furthermore, we illustrated how the Bayesian workflow idea is used through the model-building and evaluation process. The advantages of our Bayesian joint model include increased estimation efficacy, reliable association study inference results, and improved generalizability. Since the

Bayesian joint model could generate the estimations for multiple measurement units at once, compared with the traditional method of fitting each measurement unit one by one, it greatly increases the estimation efficacy and reduces potential procedure errors from repetitive work. Furthermore, by directly incorporating the estimation uncertainty of the key exposure variable of interest, the Bayesian joint model could more accurately explain the variation in the observed data, and the corresponding association estimates and conclusions should be more trustworthy. In addition, the modular idea could allow users to easily replace any parts of the Bayesian joint model with another model of their own interest without affecting the overall model structures. We used extensive simulation settings to demonstrate that our Bayesian joint model leads to a greater efficacy gain than other popular methods in the context of various degrees of contamination. Furthermore, the global calibration idea in the Bayesian joint model could provide some insights into more efficient experimental designs, and we used simulation studies to validate the possibility of reducing the number of calibration samples and increasing the number of unknown samples on each measurement unit without information loss caused by the advantage of global calibration and partial pooling. Finally, we applied our method to the real-world NYC NAAS dataset to study the potential associations between indoor allergens and asthma in children. Compared with the results of traditional inference methods that ignore measurement uncertainty, we found several significant associations between the development of children's asthma symptoms and indoor allergen concentrations. These novel findings could benefit the general public's health and assist in designing asthma prevention plans.

Although the Bayesian joint model is compelling and flexible for most user cases, in some situations the computation cost and complicated model structure might still be overwhelming for researchers in other related fields. Furthermore, sometimes researchers might not have access to both the raw measurement data for the exposure of interest as well as public health data. We thus developed a Bayesian two-step inference model to provide users with an easy-to-use alternative to the Bayesian joint model, and this Bayesian two-step inference model is also applicable to situations where one only has access to the summary statistics of estimation results for the exposure of interest. Our proposed Bayesian two-step inference model decomposed public health association

studies into two separate but closely related parts. In the first part, we performed the estimation for the exposure of interest; here, we did not impose any restrictions on the type of models to be applied for broader application scenarios. In the second part, we carefully propagated the estimation uncertainty into a combination of the exposure model and the disease model. To be more specific, in the exposure model, we built parametric models that related the observed measurements with the underlying unobserved true values by considering the measurement uncertainty, and we directly plugged the underlying unobserved true values in by treating them as random parameters with Bayesian regression in the disease model. The advantages of the Bayesian two-step inference model include easy computation, high flexibility, and reliable inference results. Compared with the Bayesian joint model, the Bayesian two-step inference model is a simplified version to relieve the computational burden. Furthermore, it provides great flexibility compared with the Bayesian joint model; for example, in the first-step model, we could even allow users to use the outputs from frequentist approaches if they could accurately summarize the point estimation as well as the uncertainty measurement for the exposure of interest. Finally, since we considered the measurement uncertainty in the second-step exposure model, further benefits are improved estimation accuracy and efficiency gain compared with traditional methods, which ignore such uncertainty and have been commonly used in the field for a long time. By performing different simulation scenarios, we demonstrated a gain in estimation accuracy for the Bayesian two-step model from various settings, from cases of no contamination to severe contamination. We also applied the Bayesian two-step model to the real-world NYC NAAS dataset and found similar findings, as discovered by the Bayesian joint model, which demonstrates that the Bayesian two-step model is also trustworthy.

There are several research directions that we would like to work on in the future. First, we wish to extend our proposed methods to more than one target exposure of interest. The Bayesian contamination model proposed in our first project focused on modeling contamination for one type of exposure only. To extend it to multiple exposures, we would not only need to consider whether different exposures have their own pattern of contamination and thus have an exposure-specific

parametric form of contamination – we would also need to consider whether the contamination status between different exposures is interdependent. For example, in the indoor allergen estimation problem, some types of allergens tend to have similar contamination reasons, while other types of allergens tend to have opposite contamination reasons. By considering this additional layer of exposure interdependence, we could build more reasonable models and generate more accurate estimations. Similar ideas could be applied to the Bayesian joint model proposed in our second project and the Bayesian two-step model proposed in our third project. For the Bayesian joint model, we are currently only considering one type of exposure and its measurement heteroskedasticity across multiple measurement units. If we had multiple targets of interest, we could build an additional hierarchical layer to model the between-unit heteroskedasticity for multiple exposures. Furthermore, it would be relatively straightforward to add additional exposure measurements in the Bayesian regression model. Similarly, in the Bayesian two-step model, we could think about how to link the measurement uncertainty between multiple exposures based on their internal characteristics and practical implications. After extending the proposed models to multiple targets of interest, we could then further extend the use cases of our methods as well as take advantage of the new MARIA measurement technology, which could simultaneously estimate more than one allergen exposure.

Second, we would like to build user-friendly interfaces and tools to help environmental health and public health researchers to easily apply our proposed methods to their own datasets and research questions. Specifically, we wish to create an R package that can implement all of the methods that we have developed thus far with detailed documentation, sample code, and example data. The main development goal would be to provide users with multiple options of applicable methods; thus, they could choose whichever way they want based on their preference, the available type of data at hand, and the background to the research questions. Furthermore, we would like to provide users with the flexibility that would allow them to modify whichever parts they think are more suitable for their own research projects, including prior distribution for various parameters in each of the models, the parametric form of contamination, the model for between-experimental-

unit heteroskedasticity, the form of measurement uncertainty for the observed estimation, and the type of association and regression models to be used. This will make our package applicable to many applied statistical models involving model contamination, global calibration, and measurement uncertainty. Besides the R package, we intend to create a user-friendly Shiny app for more interactive statistical analysis, model visualization, and model evaluation. We will create some online automatic data extraction, data transformation, and data loading pipelines to enable users to directly upload their dataset to our web-based computation tools. Moreover, we plan to provide various plotting options, including interactive graphs that summarize the model fitting results and a graphical summary of Bayesian model evaluation criteria. For example, users could directly visualize the HMC chain mixing patterns and the posterior distribution of the parameters they are interested in. Furthermore, users could switch between different methods easily and instantly view tables and plots that indicating how the model fitting results changed through different methods. By developing such tools, we would allow users with little or no statistical background to quickly understand our proposed methods and what they expect from the analysis results.

Besides, for the application studies based on NYC NAAS data, we could also work further on the following directions. First, we have listed several potential outcomes, and we currently only perform an epidemiologic association study based on one of the continuous health-related response variables. It will be of research interest to further investigate other remaining categorical outcomes related to asthma development and diagnosis. Also, we should always remember that there are no absolute criteria regarding which variable could be the response, and thus we could choose whatever variable as the response variable in the epidemiologic studies based on our research necessity. Similarly, we could add more baseline covariates into the Bayesian regression model. Other potential baseline covariates that might be interesting to be added to the model include the child's ethnicity, whether the child lives in an asthma neighborhood, whether the child has a paternal wheeze and whether the child has ever wheezed. Second, currently our Bayesian joint model can only account for one type of allergen, but there are multiple types of allergens measured by the MARIA plate, and in later research projects, once we have extended our Bayesian

joint model to a multi-target setting, we could add more types of allergen concentrations in our epidemiological association studies. Third, we might further perform some stratified versions of the epidemiological association studies based on the child's atopic (sIgE to w1,d2, i6, e72,e1, e5, tx8, gx2) status. The reason behind such a stratified regression model is that for those children with asthma, we would expect that any public health outcome that is related to allergens will be stronger in the subgroup of children with atopic status since it means that the children have the corresponding IgE antibody to at least one of the indoor allergens that we are testing. In addition, we could also add the interaction between atopic status and other covariates in the unstratified Bayesian regression model to investigate the potential existence of effect modification.

Finally, we wish to examine our proposed methods' practical implications for public health. For example, we developed the Bayesian contamination model and used it to generate contamination flags for possible poor-quality samples. A natural next step would be to investigate the potential reasons behind the existence of such contamination. Investigators could either check the lab processing procedures or rerun the samples to check whether the estimation values would change. To rerun the experiment, lab technicians could either still use the MARIA technology or switch to an older-generation technology – namely ELISA. It would also be of research interest to compare the estimation results for the same sample measured under different technologies. For example, researchers could investigate whether the contamination detection resolution differs between ELISA and MARIA. On the other hand, in the Bayesian joint and Bayesian two-step models, after estimating the association between indoor allergen concentrations and asthma morbidity in children, we could collaborate closely with environmental and public health researchers on developing new asthma prevention and intervention strategies based on our model fitting results.

## References

- Ahluwalia, S. K. and Matsui, E. C. (2018). Indoor Environmental Interventions for Furry Pet Allergens, Pest Allergens, and Mold: Looking to the Future. *JACI: In Practice*, 6(1):9–19.
- Alsefri, M., Sudell, M., García-Fiñana, M., and Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Med. Res. Methodol.*, 20(1):94.
- Baghfalaki, T., Ganjali, M., and Berridge, D. (2014). Joint modeling of multivariate longitudinal mixed measurements and time to event data using a Bayesian approach. *J. Appl. Stat.*, 41(9):1934–1955.
- Bartlett, J. W. and Keogh, R. H. (2018). Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. *Stat Methods Med Res*, 27(6):1695–1708.
- Behney, A. C. (2020). Ignoring uncertainty in predictor variables leads to false confidence in results: a case study of duck habitat use. *Ecosphere*, 11(10).
- Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian Smoothing and Regression Splines for Measurement Error Problems. *Journal of the American Statistical Association*, 97(457):160–169.
- Box, G. E. P. and Tiao, G. (1968). A Bayesian approach to some outlier problems. *Biometrika*, 55:119–129.
- Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., and van Smeden, M. (2018). Measurement error is often neglected in medical literature: a systematic review. *Journal of Clinical Epidemiology*, 98:89–97.
- Breiman, L. and Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, 60(3):291.
- Busschaert, P., Geeraerd, A., Uyttendaele, M., and Van Impe, J. (2011). Hierarchical Bayesian analysis of censored microbiological contamination data for use in risk assessment and mitigation. *Food Microbiology*, 28(4):712–719.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan : A Probabilistic Programming Language. *J. Stat. Softw.*, 76(1).
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, 0 edition.
- Carroll, R. J. and Stefanski, L. A. (1990). Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, 85(411):652–663.

- Chen, Q., Zhong, X., Acosta, L., Divjan, A., Rundle, A., Goldstein, I. F., Miller, R. L., and Perzanowski, M. S. (2016). Allergic sensitization patterns identified through latent class analysis among children with and without asthma. *Annals of Allergy, Asthma & Immunology*, 116(3):212–218.
- Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78(3):451–462.
- Cumberland, W. N., Fong, Y., Yu, X., Defawe, O., Frahm, N., and Rosa, S. D. (2015). Nonlinear calibration model choice between the four and five-parameter logistic models. *J. Biopharm. Stat.*, 25(5):972–983. PMID: 24918306.
- Davidian, M. and Giltinan, D. M. (2017). *Nonlinear Models for Repeated Measurement Data*. Routledge, 1 edition.
- Delamater, P. L., Finley, A. O., and Banerjee, S. (2012). An analysis of asthma hospitalizations, air pollution, and weather conditions in Los Angeles County, California. *Sci. Total Environ.*, 425:110–118.
- Dize, L., Martin, D., Gwyn, S., Perin, J., Gaydos, C., and Trent, M. (2018). Comparison of three serological assays to measure antibody response to Chlamydia antigen Pgp3 in adolescent and young adults with pelvic inflammatory disease. *International Journal of STD & AIDS*, 29(13):1324–1329.
- Dunson, D. B. (2001). Commentary: Practical Advantages of Bayesian Analysis of Epidemiologic Data. *American Journal of Epidemiology*, 153(12):1222–1226.
- Earle, C. D., King, E. M., Tsay, A., Pittman, K., Saric, B., Vailes, L., Godbout, R., Oliver, K. G., and Chapman, M. D. (2007). High-throughput fluorescent multiplex array for indoor allergen exposure assessment. *Journal of Allergy and Clinical Immunology*, 119(2):428–433.
- Eder, W., Ege, M. J., and von Mutius, E. (2006). The Asthma Epidemic. *New England Journal of Medicine*, 355(21):2226–2235.
- Eggleston, P. A., Butz, A., Rand, C., Curtin-Brosnan, J., Kanchanaraksa, S., Swartz, L., Breysse, P., Buckley, T., Diette, G., Merriman, B., and Krishnan, J. A. (2005). Home environmental intervention in inner-city asthma: a randomized controlled clinical trial. *Annals of Allergy, Asthma & Immunology*, 95(6):518–524.
- Engvall, E. and Perlmann, P. (1972). Enzyme-linked immunosorbent assay, Elisa. 3. Quantitation of specific antibodies by enzyme-labeled anti-immunoglobulin in antigen-coated tubes. *Journal of Immunology (Baltimore, Md.: 1950)*, 109(1):129–135.
- Feng, F., Sales, A. P., and Kepler, T. B. (2011). A Bayesian approach for estimating calibration curves and unknown concentrations in immunoassays. *Bioinformatics*, 27(5):707–712.
- Finney, D. J. (1976). Radioligand Assay. *Biometrics*, 32(4):721.
- Fong, Y., Wakefield, J., De Rosa, S., and Frahm, N. (2012). A Robust Bayesian Random Effects Model for Nonlinear Calibration Problems. *Biometrics*, 68(4):1103–1112.

- Fox, J.-P. and Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, 68(2):169–191.
- Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., and Carroll, R. J. (2004). A New Method for Dealing with Measurement Error in Explanatory Variables of Regression Models. *Biometrics*, 60(1):172–181.
- Gaffin, J. M. and Phipatanakul, W. (2009). The role of indoor allergens in the development of asthma. *Current Opinion in Allergy & Clinical Immunology*, 9(2):128–135.
- Gelman, A., Chew, G. L., and Shnaidman, M. (2004). Bayesian Analysis of Serial Dilution Assays. *Biometrics*, 60(2):407–417.
- Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4).
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian Workflow. Publisher: arXiv Version Number: 1.
- Giltinan, D. M. and Davidian, M. (1994). Assays for recombinant proteins: A problem in non-linear calibration. *Statistics in Medicine*, 13(11):1165–1179.
- Goldman, G. T., Mulholland, J. A., Russell, A. G., Strickland, M. J., Klein, M., Waller, L. A., and Tolbert, P. E. (2011). Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environmental Health*, 10(1):61.
- Greenland, S. (2007). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology*, 36(1):195–202.
- Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34(23):3133–3143.
- Guo, Y., Harel, O., and Little, R. J. (2010). How Well Quantified Is the Limit of Quantification? *Epidemiology*, 21(4):S10–S16.
- Hamilton, M. A. and Rinaldi, M. G. (1988). Descriptive statistical analyses of serial dilution data. *Statistics in Medicine*, 7(4):535–544.
- Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16(1):117.
- Higgins, K. M., Davidian, M., Chew, G., and Burge, H. (1998). The Effect of Serial Dilution Error on Calibration Inference in Immunoassay. *Biometrics*, 54(1):19.
- Homan, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623.

- Hui, F. K. C., Müller, S., and Welsh, A. H. (2021). Random Effects Misspecification Can Have Severe Consequences for Random Effects Inference in Linear Mixed Models. *ISR*, 89(1):186–206.
- King, E. M., Filep, S., Smith, B., Platts-Mills, T., Hamilton, R. G., Schmechel, D., Sordillo, J. E., Milton, D., van Ree, R., Krop, E. J., Heederik, D. J., Metwali, N., Thorne, P. S., Zeldin, D. C., Sever, M. L., Calatroni, A., Arbes, S. J., Mitchell, H. E., and Chapman, M. D. (2013). A multi-center ring trial of allergen analysis using fluorescent multiplex array technology. *Journal of Immunological Methods*, 387(1-2):89–95.
- Klauenberg, K., Walzel, M., Ebert, B., and Elster, C. (2015). Informative prior distributions for ELISA analyses. *Biostatistics*, 16(3):454–464.
- Leas, B. F., D’Anci, K. E., Apter, A. J., Bryant-Stephens, T., Lynch, M. P., Kaczmarek, J. L., and Umscheid, C. A. (2018). Effectiveness of indoor allergen reduction in asthma management: A systematic review. *Journal of Allergy and Clinical Immunology*, 141(5):1854–1869.
- Lee, M.-L. T. and Whitmore, G. A. (1999). Statistical Inference for Serial Dilution Assay Data. *Biometrics*, 55(4):1215–1220.
- Liu, X.-L., Han, G., Zeng, J., Liu, M., Li, X.-Q., and Boeckx, P. (2021). Identifying the sources of nitrate contamination using a combined dual isotope, chemical and Bayesian model approach in a tropical agricultural river: Case study in the Mun River, Thailand. *Sci. Total Environ.*, 760:143938.
- MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. A. (2007). Bayesian Methods for Highly Correlated Exposure Data. *Epidemiology*, 18(2):199–207.
- Matsui, E. C., Simons, E., Rand, C., Butz, A., Buckley, T. J., Breysse, P., and Eggleston, P. A. (2005). Airborne mouse allergen in the homes of inner-city children with asthma. *Journal of Allergy and Clinical Immunology*, 115(2):358–363.
- Morales, K. H., Ibrahim, J. G., Chen, C.-J., and Ryan, L. M. (2006). Bayesian Model Averaging With Applications to Benchmark Dose Estimation for Arsenic in Drinking Water. *Journal of the American Statistical Association*, 101(473):9–17.
- Murray, C. S., Foden, P., Sumner, H., Shepley, E., Custovic, A., and Simpson, A. (2017). Preventing Severe Asthma Exacerbations in Children. A Randomized Trial of Mite-Impermeable Bedcovers. *American Journal of Respiratory and Critical Care Medicine*, 196(2):150–158.
- Olmedo, O., Goldstein, I. F., Acosta, L., Divjan, A., Rundle, A. G., Chew, G. L., Mellins, R. B., Hoepner, L., Andrews, H., Lopez-Pintado, S., Quinn, J. W., Perera, F. P., Miller, R. L., Jacobson, J. S., and Perzanowski, M. S. (2011). Neighborhood differences in exposure and sensitization to cockroach, mouse, dust mite, cat, and dog allergens in New York City. *J. Allergy Clin. Immunol.*, 128(2):284–292.e7.
- Osborne, C. (1991). Statistical Calibration: A Review. *International Statistical Review*, 59(3):309.

- Ouédraogo, A. M., Crighton, E. J., Sawada, M., To, T., Brand, K., and Lavigne, E. (2018). Exploration of the spatial patterns and determinants of asthma prevalence and health services use in Ontario using a Bayesian approach. *PLOS ONE*, 13(12):e0208205.
- Perzanowski, M. S., Chew, G. L., Divjan, A., Johnson, A., Goldstein, I. F., Garfinkel, R. S., Hoepner, L. A., Platts-Mills, T. A., Perera, F. P., and Miller, R. L. (2008). Cat ownership is a risk factor for the development of anti-cat IgE but not current wheeze at age 5 years in an inner-city cohort. *J. Allergy Clin. Immunol.*, 121(4):1047–1052.
- Perzanowski, M. S., Chew, G. L., Divjan, A., Jung, K. H., Ridder, R., Tang, D., Diaz, D., Goldstein, I. F., Kinney, P. L., Rundle, A. G., Camann, D. E., Perera, F. P., and Miller, R. L. (2013). Early-life cockroach allergen and polycyclic aromatic hydrocarbon exposures predict cockroach sensitization among inner-city children. *J. Allergy Clin. Immunol.*, 131(3):886–893.e6.
- Phipatanakul, W., Eggleston, P. A., Wright, E. C., and Wood, R. A. (2000). Mouse allergen. II. The relationship of mouse allergen exposure to mouse sensitization and asthma morbidity in inner-city children with asthma. *Journal of Allergy and Clinical Immunology*, 106(6):1075–1080.
- Plattsmills, T., Vervloet, D., Thomas, W., Aalberse, R., and Chapman, M. (1997). Indoor allergens and asthma: Report of the Third International Workshop. *Journal of Allergy and Clinical Immunology*, 100(6):S2–S24.
- Pongracic, J. A., Visness, C. M., Gruchalla, R. S., Evans, R., and Mitchell, H. E. (2008). Effect of mouse allergen and rodent environmental intervention on asthma in inner-city children. *Annals of Allergy, Asthma & Immunology*, 101(1):35–41.
- Racine-Poon, A., Weihs, C., and Smith, A. F. M. (1991). Estimation of Relative Potency with Sequential Dilution Errors in Radioimmunoassay. *Biometrics*, 47(4):1235.
- Richardson, S. and Gilks, W. R. (1993a). A Bayesian Approach to Measurement Error Problems in Epidemiology Using Conditional Independence Models. *American Journal of Epidemiology*, 138(6):430–442.
- Richardson, S. and Gilks, W. R. (1993b). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, 12(18):1703–1722.
- Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. M. (2014). Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging. *Journal of the American Statistical Association*, 109(508):1385–1397.
- Rosenstreich, D. L., Eggleston, P., Kattan, M., Baker, D., Slavin, R. G., Gergen, P., Mitchell, H., McNiff-Mortimer, K., Lynn, H., Ownby, D., and Malveaux, F. (1997). The Role of Cockroach Allergy and Exposure to Cockroach Allergen in Causing Morbidity among Inner-City Children with Asthma. *NEJM*, 336(19):1356–1363.
- Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069.

- Rosset, S. and Tibshirani, R. J. (2020). From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation. *JASA*, 115(529):138–151.
- Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74(2):385–391.
- Schmid, C. H. and Rosner, B. (1993). A bayesian approach to logistic regression models having measurement error following a mixture distribution. *Statistics in Medicine*, 12(12):1141–1153.
- Serebrisky, D. and Wiznia, A. (2019). Pediatric Asthma: A Global Epidemic. *Annals of Global Health*, 85(1):6.
- Sheehan, W. J. and Phipatanakul, W. (2016). Indoor allergen exposure and asthma outcomes. *Current Opinion in Pediatrics*, 28(6):772–777.
- Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, 72(3):583–592.
- Szpiro, A. A. and Paciorek, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology: MEASUREMENT ERROR IN TWO-STAGE ANALYSES. *Environmetrics*, 24(8):501–517.
- Tandon, M., Chapman, M., and King, E. (2009). Optimization of a Multiplex Array for Indoor Allergens-Influence of Streptavidin-Phycoerythrin (SAPE) Lots on Assay Performance. *Journal of Allergy and Clinical Immunology*, 123(2):S232–S232.
- Tsiatis, A. A. and Davidian, M. (2004). Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. *Statistica Sinica*, 14(3):809–834. Publisher: Institute of Statistical Science, Academia Sinica.
- Tworoger, S. S. and Hankinson, S. E. (2006). Use of biomarkers in epidemiologic studies: minimizing the influence of measurement error in the study design and analysis. *Cancer Causes & Control*, 17(7):889–899.
- Van Weemen, B. and Schuurs, A. (1971). Immunoassay using antigen-enzyme conjugates. *FEBS Letters*, 15(3):232–236.
- van Zoest, V., Hoek, G., Osei, F., and Stein, A. (2020). Bayesian analysis of the short-term association of NO<sub>2</sub> exposure with local burden of asthmatic symptoms in children. *Sci. Total Environ.*, 720:137544.
- Vashist, S. K. and Luong, J., editors (2018). *Handbook of immunoassay technologies*. Elsevier/AP, Academic Press, an imprint of Elsevier, London. OCLC: on1001456627.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved  $\widehat{R}$  for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16:667–718.

- Wang, X., Tan, X., and Li, Q. (2020). Effectiveness of fractional exhaled nitric oxide for asthma management in children: A systematic review and meta-analysis. *Pediatric Pulmonology*, 55(8):1936–1945.
- Whittemore, A. S. and Keller, J. B. (1988). Approximations for Regression with Covariate Measurement Error. *Journal of the American Statistical Association*, 83(404):1057–1066.
- Yates, D. H. (2001). Role of exhaled nitric oxide in asthma. *Immunology & Cell Biology*, 79(2):178–190.

## Appendix A: Appendices to Chapter 2

### A.1 Simulation Study: Prior predictive check

In this simulation study, we would like to incorporate the idea of prior predictive checking into a simulation setting which is similar to our simulation setting 1 in Chapter 2. In each replicate of the simulation, we simulate a plate of dilution assay data with 2 replicates of standard samples and 10 unknown samples. For the standard calibration data, we set  $\theta^{\text{standard}} = 125$  and generate 26 observations with 13 dilution levels for each replicate,  $d_i = 0, 1, 1/2, 1/4, \dots, 1/2048$ , based on  $\log(y_i) \sim \text{normal}(\log(g(\theta^{\text{standard}} d_i, \beta)), \sigma_y)$ . Then we generate 10 unknown sample concentrations  $\theta_j$  from the hierarchical prior distribution  $\text{exponential}(\mu_j)$ , where  $\mu_j \sim \text{normal}(0, 0.1)$ , together with 10 contamination factors  $\delta_{\beta_{2j}}$  from  $\text{normal}(0, 1)$  and  $\delta_{\sigma_j}$  from  $\text{exponential}(1)$  and 10 mixture proportion  $\lambda_j^*$  from  $\text{beta}(1, 10)$ . For each unknown sample  $j$ , we generate a binary random variable  $\phi_j$  with probability of being 1 to be  $\lambda_j^*$  and generate three observations with dilution levels  $d_i = 1/10, 1/100, 1/10000$  based on the mixture model  $\log(y_{ij}) \sim (1-\phi_j)\text{normal}(\log(g(\theta_j d_i, \beta)), \sigma_y) + \phi_j\text{normal}(\log(g(\theta_j d_i, \beta_j)), \sigma_{y_j})$ , where  $\beta_j = (\beta_1, \beta_{2j}, \beta_3, \beta_4)$ ,  $\beta_{2j} = \beta_2 e^{\delta_{\beta_{2j}}}$ ,  $\sigma_{y_j} = \sigma_y e^{\delta_{\sigma_j}}$ . We repeat the simulation 500 times and calculate the coverage probability of the Bayesian 95% probability intervals for the true underlying concentration, which is 0.95 and indicates that our priors and models are reasonable.

Followed by the above prior predictive checking simulation setting, to check the robustness of our data generation model, we use the ideas developed in our Bayesian workflow for model contamination and perform one follow-up simulation scenario, which replicates every step in the above prior predictive check but instead uses the lognormal distribution to generate random unknown concentrations. We repeat the simulation 500 times and then evaluate the coverage probability of the Bayesian 95% probability intervals for the underlying true concentrations. The correspond-

ing coverage probability is 0.95 and indicates that our model is pretty robust for potential model misspecification.

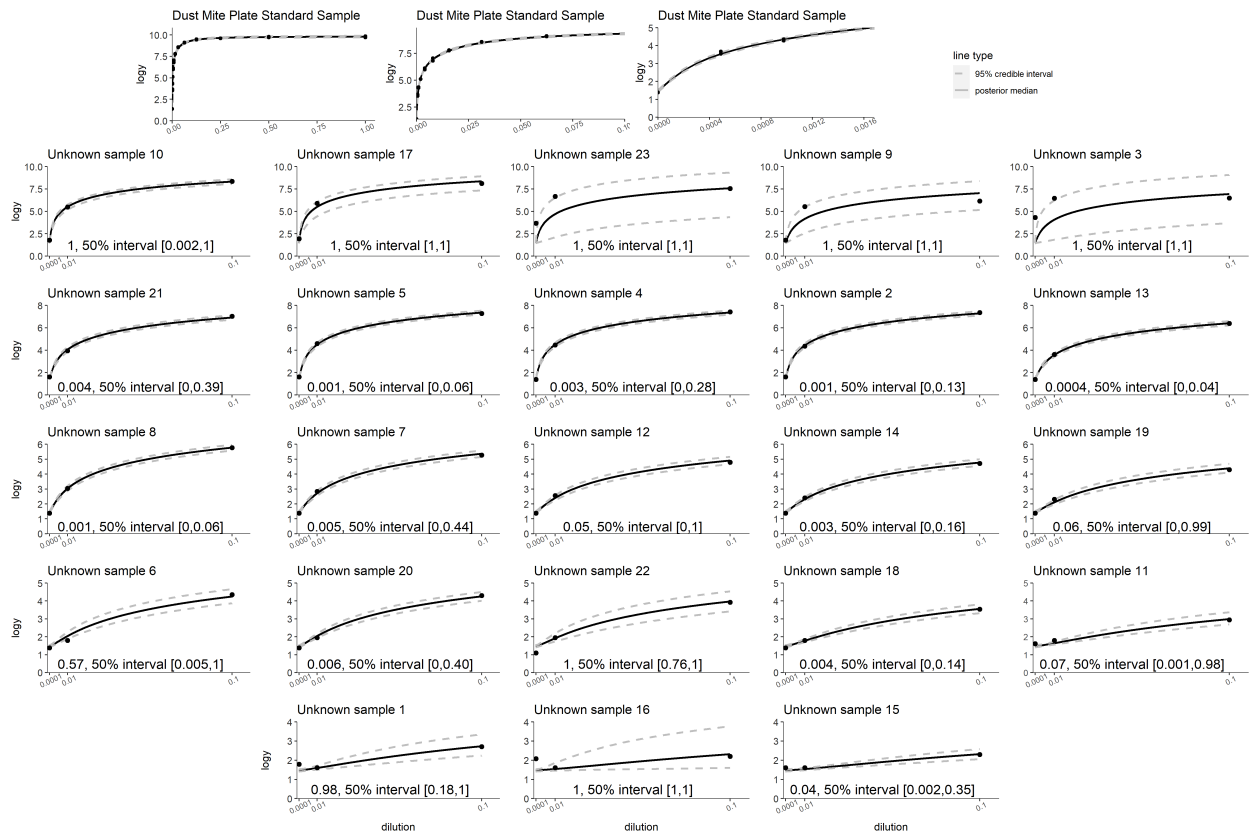


Figure A.1: Posterior median and 95% probability interval of the mean function dose-response curves for the standards and each new sample estimated using our proposed model using stronger exponential prior for unknown concentrations. The posterior median and corresponding 50% probability interval for the probability of contamination are listed at the bottom of each plot.

## Appendix B: Appendices to Chapter 3

### B.1 Simulation Study: More efficient plate design

In the main body of the project, we have shown that given the advantage of the joint modeling and partial pooling of our proposed Bayesian joint model, we could further reduce the number of dilutions applied to the standard calibration sample on each plate to allow for more rooms for the unknown samples, and thus come up with a more efficient experiment plan where we could measure more unknown samples in a single MARIA plate measurement. In this supplementary analysis, we would like to explore the potential possibility of further reducing the number of standard calibration sample observations contained on each MARIA plate to leave room for more unknown samples. The logic behind this more efficient experiment design has several supporting reasons. First, the embedded nature of simultaneously modeling multiple MARIA plates could let the model get information not only from a single MARIA plate but also some additional information from the remaining MARIA plates in the study through hierarchical modeling and partial pooling, so even less information contained in each MARIA plate will not have significant influence. Second, the initial concentration and dilution factors are both known for the standard calibration sample. Also, there is no contamination in the standard calibration sample, and thus the data quality for the standard calibration sample is very good, so fewer observations of the standard calibration sample should not lose much information. Third, each unknown sample only has three dilution factors with no replicates, which is far less than the number of standard calibration sample observations, so it is reasonable to reduce further the number of standard calibration sample observations on each plate. Now we remain everything same as in simulation 2 except now on each plate, we now have 8 observations of standard calibration samples (with 2-fold dilutions 0, 0, 1, 1, 1/8, 1/8, 1/2048, 1/2048) and 87 observations for 29 unknown sample (each with dilutions 1/10, 1/100, 1/10000).

The logic behind this design of selected dilution factors for the standard calibration sample is that first, we need the full dilution to capture the upper bound of the calibration curve and similarly need the zero dilution for modeling the lower bound of the calibration curve. Then dilution factor of 1/2048 is the smallest dilution factor for the original design, so it is also reasonable to keep it. Finally, the dilution factor of 1/8 for the standard calibration sample is very similar to the dilution factor of 1/10 for the unknown samples, and thus we also keep it. Then we have repeated the analysis procedures as simulation setting 2 with 500 iterations and found that the more efficient design will also enable the model to recover the underlying associations very well, which also demonstrates the robustness of our Bayesian model and the potential for a more efficient experiment MARIA plate design with fewer observations for the standard calibration sample and more observations for the unknown samples.

	1	2	3	4	5	6	7	8	9	10	11	12
	Standard Sample			Unknown Samples								
				1/10	1/100	1/10000	1/10	1/100	1/10000	1/10	1/100	1/10000
A	1	1/8	1/2048	Unk 1	Unk 1	Unk 1	Unk 9	Unk 9	Unk 9	Unk 17	Unk 17	Unk 17
B	1	1/8	1/2048	Unk 2	Unk 2	Unk 2	Unk 10	Unk 10	Unk 10	Unk 18	Unk 18	Unk 18
C	Unk 24	Unk 24	Unk 24	Unk 3	Unk 3	Unk 3	Unk 11	Unk 11	Unk 11	Unk 19	Unk 19	Unk 19
D	Unk 25	Unk 25	Unk 25	Unk 4	Unk 4	Unk 4	Unk 12	Unk 12	Unk 12	Unk 20	Unk 20	Unk 20
E	Unk 26	Unk 26	Unk 26	Unk 5	Unk 5	Unk 5	Unk 13	Unk 13	Unk 13	Unk 21	Unk 21	Unk 21
F	Unk 27	Unk 27	Unk 27	Unk 6	Unk 6	Unk 6	Unk 14	Unk 14	Unk 14	Unk 22	Unk 22	Unk 22
G	Unk 28	Unk 28	Unk 28	Unk 7	Unk 7	Unk 7	Unk 15	Unk 15	Unk 15	Unk 23	Unk 23	Unk 23
H	Unk 29	Unk 29	Unk 29	Unk 8	Unk 8	Unk 8	Unk 16	Unk 16	Unk 16	blank	blank	HC Control

Table B.1: A potentially more efficient design for standards and new samples (dilutions) in a multiplex plate with 96 wells.

	Bias	RMSE	80% Coverage probability	80% Interval width
Joint Model Original Design	0.0007	0.055	81%	0.15
Joint Model New Design	0.0004	0.053	78%	0.14
Joint Model A More Efficient Design	0.0005	0.047	80%	0.13

Table B.2: Supplementary section an even more efficient design with 8 observations of calibration sample and 29 unknown samples per plate.

## **Appendix C: Appendices to Chapter 4**

### **C.1 Simulation Study: Alternative robust model for measurement uncertainty**

Besides the normal model for measurement uncertainty in our proposed Bayesian two-step model, in simulation scenario 1, we also tried several alternative model specification forms and checked the robustness of our proposed model. Since we assume a normal model for measurement uncertainty, although it is relatively straightforward and simple, it still suffers from the risk of model misspecification, which is a common issue for strong forms of restrictive parametric models. To check the robustness of our model, we have also tried some other robust families of parametric distributions, including the student t-distribution with both a fixed degree of freedom and a random degree of freedom and the gamma distribution. However, we did not discover any significant changes in the corresponding Bayesian two-step model inference results. This indicates that our original model is robust enough in this simulation setting and also points out the potentially more robust model alternative specifications in more complicated settings.