Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yuzi Zhang Date

Statistical Methods for Disease Surveillance Based on Multiple Data Streams

By

Yuzi Zhang
Doctor of Philosophy

Biostatistics

Howard H. Chang, Ph.D. and Robert H. Lyles Ph.D.
Advisor

Lance A. Waller, Ph.D.
Committee Member

Sarita Shah, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D., M.P.H
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Disease Surveillance Based on Multiple Data Streams

By

Yuzi Zhang
B.E., China Pharmaceutical University, China, 2016
M.S.P.H., Emory University, GA, 2018

Advisor: Howard H. Chang, Ph.D. and Robert H. Lyles Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2023

Abstract

Statistical Methods for Disease Surveillance Based on Multiple Data Streams

By Yuzi Zhang

Disease surveillance systems are widely implemented to monitor diseases distribution and detect outbreaks. An important task of disease surveillance is to infer the number of prevalent or cumulative incident cases. When there are multiple disease surveillance systems in operation for monitoring the same disease among essentially closed populations, the capture-recapture (CRC) approach is an appealing tool used for integrating information across the systems to estimate the total number of diseased cases. We first develop a hierarchical modeling framework for analyzing individual-level surveillance data collected from multiple surveillance systems at multiple surveillance sites that allows for individual-level heterogeneity in capture probabilities, and borrows information across surveillance sites to improve the estimation of disease case counts. Second, we propose an accessible sensitivity and uncertainty analysis using a multinomial distribution-based maximum likelihood estimation (MLE) procedure that hinges on a key inestimable parameter for two-catch CRC experiments. Under this multinomial model, we also derive bias-corrected estimators which allow for any user-specified level of the dependency between two systems. We next clarify some crucial pitfalls of the popular log-linear model-based approach to CRC estimation. Finally, motivated by those pitfalls, we develop an alternative framework again under the multinomial distribution-based model, and hinging on the choice of a key parameter that reflects dependences among surveillance systems. This alternative framework leverages generalizations of the closed-form estimator derived in the sensitivity and uncertainty analysis framework, and extends the associated bias correction procedures to incorporate CRC studies involving an arbitrary number of systems. Under the alternative framework, we show how expert opinion can be incorporated in the spirit of prior information to guide estimation in an appealing and transparent way, and how an adapted credible interval approach can be used to facilitate inference exhibiting favorable frequentist properties. By generalizing the idea in the proposed uncertainty analysis targeting for two-catch cases, the proposed framework permits principled uncertainty analyses via which a user can acknowledge his/her level of confidence in assumptions made about the key dependency parameter.

Statistical Methods for Disease Surveillance Based on Multiple Data Streams

By

Yuzi Zhang
B.E., China Pharmaceutical University, China, 2016
M.S.P.H., Emory University, GA, 2018

Advisor: Howard H. Chang, Ph.D. and Robert H. Lyles Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2023

Acknowledgments

I would like first express my sincere gratitude to my two advisors, Dr. Robert Lyles and Dr. Howard Chang, for their unwavering support and encouragement over past years. Thank you, Dr. Robert Lyles, for introducing me to capture-recapture methods; the dissertation would not have been completed without your invaluable guidance and profound insights. The guidance provided by you has been invaluable in my academic and personal growth.

I also would like to thank Dr. Howard Chang for his mentorship and continuous support, not only during the dissertation but also through various other research projects. His expertise in statistics and environmental epidemiology has greatly enhanced the quality of my work in every aspect. His support has made my PhD journey truly enjoyable. It is my fortunate to have him as my advisor.

I extend my gratitude to my committee members, Dr. Lance Waller and Dr. Sarita Shah, for their time and helpful discussions. I also want to thank Dr. Lance Waller for the support during my job search.

Many thanks to all faculty and staff in BIOS department. The inclusive environment they have cultivated plays an important role in my academic journey.

Lastly, huge thanks to my parents for their unconditional love and company. Without them, I would not have reached this point.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to Disease Surveillance

Disease surveillance systems that monitor disease distribution and detect outbreaks are critical for developing, implementing and evaluating public health programs. Disease surveillance systems are typically categorized as passive or active. So-called "passive" surveillance systems obtain incident cases information through reports filed by healthcare facilities, including hospitals, clinics, and laboratories (Hadorn and Stärk, 2008). In a passive system, no deliberate efforts are made for identifying incident cases from the targeted populations. In contrast, active surveillance systems are designed to periodically screen for disease in target populations, often utilizing highly sensitive, standardized diagnostic methods (Jamison et al., 2006). Hence, passive surveillance systems are generally more cost-effective and can be established over large geographic areas (Hadorn and Stärk, 2008). Data from passive surveillance systems have become more readily available in many parts of the world due to increased capabilities of electronic reporting (Simonsen et al., 2016; Huff et al., 2017; Wang et al., 2019). Despite their potentials as an important data source, passive surveillance systems are known to suffer from under-ascertainment (Gibbons et al., 2014). Specifically, passive surveillance relies on encounters with health care facilities and some individuals with the disease of interest may not seek care.

Disease surveillance systems are generally associated with under-ascertainment, and the capture probability of disease surveillance systems could be influenced by various factors, such as age, sex, race, economic status, access to health care, disease symptoms, and disease severity (Gibbons et al., 2014; Peixoto et al., 2020). Identifying factors that affect the case ascertainment in a surveillance system is an important research area. This knowledge can be used to infer the total number of incident cases in the study region, leveraging the large amount of data from passive surveillance systems and the overlap between multiple surveillance systems.

1.2 Capture-Recapture Methods in Disease Surveillance

When multiple overlapping disease surveillance systems are implemented to survey a disease across a study period among closed populations, capture-recapture (CRC) methods can be applied to estimate the number of prevalent or incident cases (N). CRC methods are originally developed for the purpose of estimating the abundance of animal populations based on experiments recording the animal encounter history over multiple trapping occasions. More recently, the use of the CRC methods has also become common in epidemiological studies used for studying human populations. Specially, CRC methods are widely used in various disease surveillance projects for estimating the number of diseased cases, such as cancer (McClish and Penberthy, 2004), HIV infections (Abeni et al., 1994; Bernillon et al., 2000; Héraud-Bousquet et al., 2012), COVID-19 infections (Böhning et al., 2020), and other diseases (Van Hest et al., 2008). Different surveillance systems can be viewed in the same way as the trapping occasions considered in the ecological studies.

1.2.1 Data structure

To implement the CRC methods, surveillance data collected from multiple overlapping surveillance systems can be summarized at two different levels, i.e., individual and population levels. Specifically, the observed individual-level CRC data contain binary capture indicators of separate surveillance systems for individuals who have been identified by surveillance systems. In contrast, the population-level CRC data only provide the frequency of each possible capture history, except for the number of cases never identified by any system. For instance, Table 1.1 and 1.2 present individual and population levels data under the case where two surveillance systems are implemented. Let n_{ij} be the observed number of individuals having capture history

(i, j), subscripts of 1 indicates captured and 0 not captured by a given system. y_{1q} and y_{2q} denote capture indicators (captured/not captured) of individual q for system 1 and 2, respectively. Table 1.1 displays three possible capture histories for an individual; the individual who has capture history $(0, 0)$ cannot be observed. Aggregating the individual-level data across all captured individuals results in the population-level data as shown in Table 1.2. Since individuals with capture history $(0, 0)$ are unobserved, the cell count n_{00} is unknown. We are interested in estimating the total incident cases N, which is equivalent to estimating N_{00}.

It worth noting that, due to the unobserved cell count (e.g., n_{00} for the two-catch case), at least one untestable assumption regarding the dependency between systems is required to enable the estimation of N (Lyles et al., 2021a). One classical assumption is that of independence, which states that capture efforts (i.e., trapping occasions in ecology, surveillance systems in epidemiology) operate independently at the population level. For the case where two capture efforts are implemented, several estimators have been developed under the independence assumption. The well-known Lincoln-Petersen (LP) estimator is the earliest example (Lincoln, 1930; Petersen, 1896), while an approximately unbiased estimator was subsequently proposed by Chapman (Chapman, 1951) to overcome positive bias inherent in the LP estimator.

Table 1.1: Individual-level data of two-catch CRC methods

Captured in system 1	Captured in system 2	Capture history (y_{1q}, y_{2q})
Yes	Yes	(1, 1)
Yes	No	(1, 0)
No	Yes	(0, 1)
No	No	(0, 0)

Table 1.2: Population-level data of two-catch CRC methods

	Captured in system 2		
Captured in system 1	Yes	No	
Yes	n_{11}	n_{10}	
No	n_{01}	n_{00} =?	
			N = ?

1.2.2 Multinomial models

In CRC literature, the multinomial model is a popular choice for analyzing population-level CRC data (Darroch, 1958; Wittes, 1974; Seber et al., 1982). Using the two-catch population-level data as an example, this model assumes $(N_{11}, N_{10}, N_{01}, N_{00}) \sim$ Multinomial$(N, p_{11}, p_{10}, p_{01}, p_{00})$, where p_{ij} is the population-level probability of having capture history (i, j). These probabilities of different capture histories are often characterized using parameters of interest based on an assumed dependency structure between data streams. For example, under the classical independence assumption, those probabilities can be expressed as the function of two marginal probabilities, p_1 and p_2, where p_1 and p_2 are proportions identified by the first and the second streams, respectively. The population-level multinomial model also allows one to introduce dependencies between data streams by representing probabilities of possible capture histories based on a different set of parameters. In this dissertation, we factorize probabilities of possible capture histories using conditional probabilities to allow the consideration of various dependency structures across data streams.

Note that parameters used to express probabilities of different capture histories are defined at the population-level. In other words, the population-level multinomial model does not make assumptions on the individual-level capture probabilities (Lyles et al., 2021a). Consider the individual-level multinomial model, and

let $H_{ij,q}$ denote the binary indicator of whether the q-th individual has capture history (i, j). Specifically, $h_{11,q} = 1$ indicates the individual q was captured by both systems, equivalently, $y_{1q} = 1$ and $y_{2q} = 1$ for individual q. Then we have $(H_{11,q}, H_{10,q}, H_{01,q}, H_{00,q}) \sim \text{Multinomial}(1, p_{11,q}, p_{10,q}, p_{01,q}, p_{00,q})$, where $p_{ij,q}$ is the individual-specific probability of having capture history (i, j). By drawing N samples from this individual-level multinomial model, four cell counts (i.e., the number of cases having capture histories (i, j) for $i, j \in \{0, 1\}$) can be obtained. Prior authors have shown that the population-level multinomial model can be obtained by repeating this data generation mechanism for an arbitrary number of times while introducing a joint distribution $f(\boldsymbol{p})$ to model capture probabilities, $(p_{11,q}, p_{10,q}, p_{01,q}, p_{00,q})$ (Lyles et al., 2021a).

1.2.3 Poisson models

The Poisson log-linear model is another popular choice for modeling the population-level CRC data (Cormack, 1989; Hook and Regal, 1995; Chao et al., 2001). The model assumes the observed cell counts follow Poisson distributions with means dependent on their capture histories (Cormack, 1989). The dependency structure between systems is introduced by including interaction terms between binary indicators for data systems. By varying the combinations of interaction terms, different dependency structures can be obtained. The Poisson model is closely related to the population-level multinomial model. For example, Cormack and Jupp (1991) showed that maximum likelihood estimators (MLEs) of N derived using the population-level multinomial model are equivalent to estimates yielded by Poisson log-linear models when equivalent assumptions are imposed.

1.2.4 Models allowing individual heterogeneity in capture probabilities

To estimate the total number of cases, one can focus on the individual-level capture indicators (e.g., y_{1q} and y_{2q} for $q = 1, \ldots, N$ under the two-catch case). Models focusing on the individual-level capture indicators permit the consideration of the heterogeneity in capture probabilities. For example, $P(y_{tq} = 1) = p_{tq}$ can be modeled via a logit model to allow individual heterogeneity in capture probabilities, where t is the index for data stream and q is the index for individual (Huggins, 1989; Coull and Agresti, 1999; Tounkara and Rivest, 2015). Specifically, the model in general form can be written as:

$$\text{logit}(p_{tq}) = \alpha_t + \boldsymbol{X}_q^T \boldsymbol{\beta}_t + \epsilon_q, \tag{1.1}$$

where α_t is the system-specific intercept, $\boldsymbol{X}_q$ is a vector of individual-level covariates included for the system t, $\boldsymbol{\beta}_t$ are system-specific model coefficients, and ϵ_q denotes the random effect which is typically assumed to follow a mean-zero normal distribution. The inclusion of random effects is a way to incorporate unobserved heterogeneity capture probabilities. The critical untestable assumption imposed in this model is the conditional independence, which implies that individual-level capture indicators are independent conditional on covariates included in the model and the random effect. With the estimated capture probabilities $\hat{p}_{tq}$, the estimate of N is obtained by applying the Horvitz-Thompson estimator that is $\sum_{q=1}^{n_c} 1/\hat{p}_{tq}$, where n_c is the number of uniquely identified cases (Huggins, 1989).

1.3 Specific Aims

This dissertation focuses on the development of CRC methods for estimating the number of diseased cases and/or the disease prevalence based on surveillance data

collected from multiple overlapping systems. The specific aims of this dissertation are:

- **Aim 1.** Develop a Bayesian hierarchical model to improve the estimation of case counts while incorporating individual heterogeneity in capture probabilities and borrowing information across surveillance locations based on spatial-referenced surveillance data collected from multiple overlapping surveillance systems.

- **Aim 2.** Propose an accessible and unified sensitivity and uncertainty analysis framework based on the population-level multinomial model focusing on a key inestimable parameter.

- **Aim 3.** Clarify pitfalls associated with the commonly used log-linear modeling paradigm when analyzing CRC data in epidemiological studies.

- **Aim 4.** Develop a transparent modeling framework relying on closed-form estimators derived from the modeling idea adopted in Aim 2, while leveraging expert opinion to guide the estimation of case counts to address pitfalls clarified in Aim 3.

Chapter 2

A Hierarchical Model for Analyzing Multi-Site Individual-Level Disease Surveillance Data

2.1 Background

Geo-reference CRC data and observed spatial patterns in incident cases have motivated the development of methods accounting for data collected at multiple surveillance sites. For example, the N-mixture model proposed by Royle (2004) aims to estimate size of animal populations from data collected at multiple trapping sites by multiple trapping efforts. This model used a Poisson distribution to model the unobserved true population size at each site, which are treated as nuisance parameters and marginalized out of the likelihood. Estimators for site-specific population size at sites with multiple trapping efforts have been derived using estimated site-specific case rates. Recently, Li et al. (2020) extended the N-mixture model under a Bayesian spatial hierarchical framework to estimate disease incident cases. Compared to the N-mixture models, the extended model allows to estimate site-specific incident cases at sites where overlapping occurs, as well as sites where only one system is available. An important application of the extended N-mixture models is to bias-correct a passive disease surveillance system that covers large areas. While the existing methods allow for capture probabilities to vary across sites and systems, they cannot account for individual-level heterogeneity in capture probability. As discussed by Otis et al. (1978) and many others, fail to account for individual heterogeneous capture probabilities may result in severely biased estimator of the number of incident cases. Thus, methods that account for individual heterogeneity in capture probability and incorporate spatial patterns in cases are needed for analyzing individual-level surveillance data collected over multiple sites by multiple systems.

In this chapter, we propose an approach for analyzing multi-site individual-level CRC data obtained by linking a passive and an active surveillance system. Specifically, in a unified modeling framework, we aim to (1) examine factors associated with individual-level capture probabilities, (2) examine factors associated with true incident cases across sites, and (3) estimate true incident cases by bias-correcting data

from a passive surveillance system. Our approach combines an individual-level binomial mixture (BM) model, which allows for covariate-dependent capture probability, with a spatial Poisson regression model for site-specific true incident cases. Unlike previous N-mixture models for multi-site CRC data, our approach treats the true number of incident cases as parameters of interest. Inference and estimation are carried out via a Bayesian two-stage procedure to address missing covariate information on individuals not captured by any system.

We introduce the motivating pulmonary tuberculosis (PTB) surveillance data from a major center of ongoing transmission in China in Section 2.2. In Section 2.3, we describe the proposed statistical framework and the corresponding two-stage Bayesian estimation procedure. In Section 2.4, we present simulation studies for evaluating the performance of the proposed framework, particularly highlighting the benefits of supplementing the BM model with a spatial process model for estimating true incident cases. In Section 2.5, the proposed approach is applied to the PTB surveillance data, and we report estimated total PTB incident cases, and factors associated with system-specific capture probabilities and PTB rates. Finally, we conclude with a discussion of future methodological directions in Section 2.6.

2.2 Motivating Data

In China, the National Infectious Disease Reporting System (NIDRS) is a passive, country-wide, web-based infectious disease reporting network that collects case data on reportable diseases. Complementary to NIDRS, cross-sectional prevalence surveys are conducted periodically (e.g., every five years) for select infections, and provinces implement annual sentinel surveillance surveys in select locations. Hereafter, we refer to these three surveillance systems as S1, S2, and S3, respectively.

We analyzed linked S1, S2 and S3 PTB data in Sichuan, a province with a high

burden of PTB and a population of more than 80 million people (Yang et al., 2008). The passive NIDRS (S1) has been operating in all 181 counties of Sichuan since 2004, and more than 71,000 PTB cases were reported in 2010 in the province. We supplemented S1 data with data from a cross-sectional PTB prevalence survey (S2) conducted in 2010 for surveying PTB among populations over 15 years of age at the community-level, as well as from sentinel surveillance (S3) conducted in nine select communities each year from 2012 to 2016 targeting populations over 15 years of age. The total number of PTB cases reported by these three surveillance systems represents an underestimation of the true number in Sichuan because these three surveillance systems cannot enumerate all PTB cases. Potential factors associated with underestimation including that S1 only captures PTB cases who had encounter at health facilities covered by S1, and S2 and S3 (active systems) only capture cases at sub-regions of Sichuan. One goal of the current analysis is to examine individual and area-level covariates that may be associated with system-specific capture probability.

To estimate the total number of PTB cases in the study region, we assume the population is closed during the study period. We obtained de-identified two two-catch CRC datasets by performing record linkage between S1 and S2, and between S1 and S3 uusing patient name, sex, date of birth, and residence address identifiers. Briefly, S1 data from 2009 to 2010 were linked to S2 data, and S1 data from 2011 to 2016 were linked to S3 data. The aggregated ascertainment histories of the S1-S2 and S1-S3 linked data are summarized in Table 2.1. Figure 2.1 shows the community sites where both S1 and S2, or S1 and S3, were linked. We note that S2 and S3 do not temporally overlap during the study period. Our study included a total of 58 reporting sites with linked data: 24 sites that yielded S1-S2 linked data, and 34 sites that yielded S1-S3 linked data.

Individual-level case characteristics (e.g., age, sex) were recorded for each individual by all three systems. Population-level risk factors of PTB and geographical

information collected at each surveillance site included annual gross domestic product per capita (GDP) from 2009 to 2016; prefecture-level average area per capita in 2010, a measurement of crowding (average area); county-level proportions of elderly ($>$ 65 years old) in 2010 (percent of elderly); longitude; latitude; and elevation in meter. Summary statistics of individual-level and county/prefecture-level variables are presented in Table 2.2 and Table 2.3.

Table 2.1: PTB cases ascertainment by S1 and S2 in 2009-2010, and by S1 and S3 in 2011-2016, in the study region

	2009 - 2010		2011 - 2016	
	Captured in S2		Captured in S3	
Captured in S1	Yes	No	Yes	No
Yes	84	37	85	155
No	117	?	172	?

Table 2.2: Summary statistics of individual-level variables for motivating PTB data

Covariates	S1-S2 (Total = 238)	S1-S3 (Total = 412)	S1 (Total = 71844)
Age Group			
< 40	18.49%	23.30%	42.07%
$[40, 50)$	18.49%	16.50%	18.37%
$[50, 60)$	17.23%	19.17%	16.66%
$[60, 70)$	29.41%	26.21%	14.12%
≥ 70	16.38%	14.80%	8.79 %
Gender			
females	27.31%	30.83%	29.65%
males	72.69%	69.17%	70.35%

2.3 Methods

2.3.1 Model Specification

Consider a passive surveillance system that operates at S sites. Let N_s denote the unobserved true number of incident cases at site s, and the total number of incident

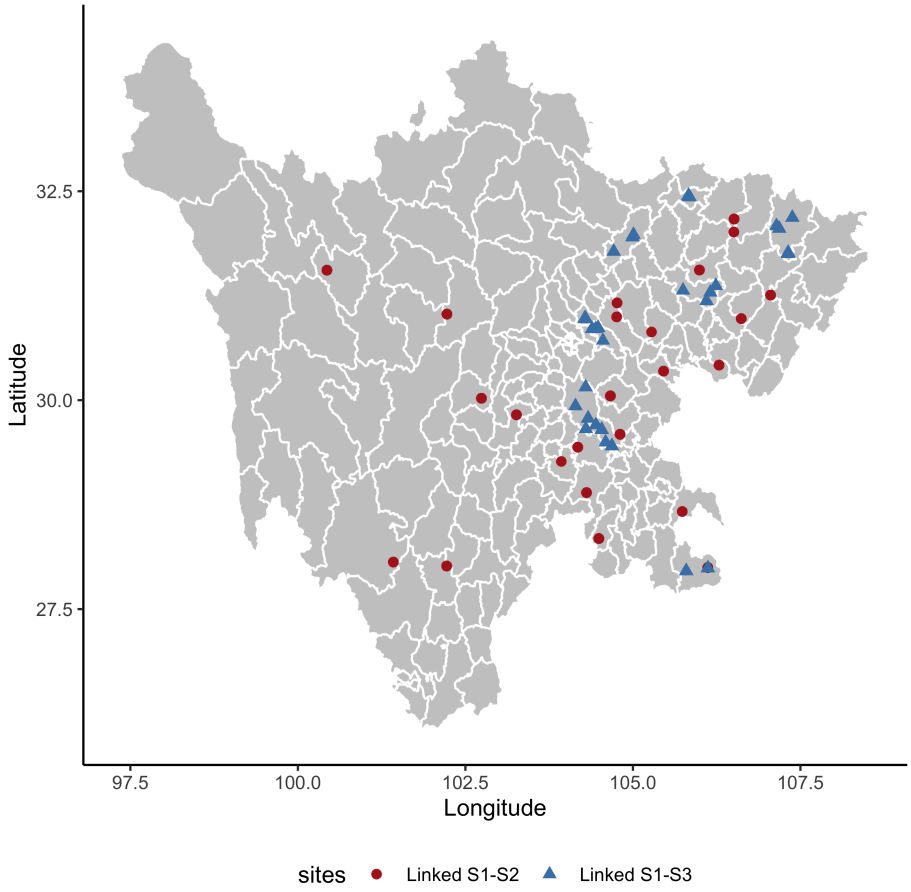


Figure 2.1: Counties within Sichuan Province where S1 and S2, or S1 and S3 surveillance are active.

Table 2.3: Summary statistics of county/prefecture level covariates over all 181 counties for motivating PTB data

Covariates	Mean (SD)	Median [Min, Max]
Longitude	103.96 (1.93)	104.04 [98.20, 107.99]
Latitude	30.17 (1.52)	30.39 [26.45, 33.66]
GDP	19537.31 (12293.93)	15511 [4785, 66912]
Elevation	1433.75 (1304.15)	714.98 [298.09, 4499.40]
Average area	34.24 (7.38)	34.74 [15.27, 58.57]
Percent of elderly	10.12 (2.66)	10.66 [4.60, 16.39]

cases in the study region is given by $N = \sum_{s=1}^{S} N_s$. Among these S sites, a supplementary active surveillance system only occurs at K sites with $K < S$. Without loss of generality, we assume the first K sites have both passive and active surveillance data. We refer to these first K sites as the linked sites, and the other $S - K$ sites as the unlinked sites. For $i = 1, \ldots, N_s$, $s = 1 \ldots, S$, let y_{1is} denote the capture indicator of i-th individual for the passive system. Similarly, for $i = 1, \ldots, N_s$, $s = 1 \ldots, K$, let y_{2is} denote the capture indicator for the active system.

We assume $Y_{tis} \sim \text{Bin}(1, p_{tis})$ for $t = 1, 2$ with density function $p_{tis}^{y_{tis}}(1-p_{tis})^{(1-y_{tis})}$, and p_{tis} is interpreted as the capture probability of system t for individual i at site s. At the K linked sites, individual-level capture history is modeled using a binomial mixture (BM) model (Dorazio and Andrew Royle, 2003; Royle, 2004; Tounkara and Rivest, 2015). Assume Y_{1is} and Y_{2is} are independent given p_{1is} and p_{2is}, and because individuals with capture history $y_{1is} = 0$ and $y_{2is} = 0$ are not observed, the data likelihood of the capture history for individual i at the linked site s has a conditional joint density (Finney, 1947) given by

$$P(Y_{1is} = y_{1is}, Y_{2is} = y_{2is} | y_{1is} = 1 \text{ or } y_{2is} = 1) = \frac{p_{1is}^{y_{1is}}(1-p_{1is})^{(1-y_{1is})} p_{2is}^{y_{2is}}(1-p_{2is})^{(1-y_{2is})}}{1-(1-p_{1is})(1-p_{2is})}.$$

To allow for individual heterogeneity in capture probabilities p_{1is} and p_{2is}, logit models are used to incorporate individual-level covariates. In addition to linking active surveillance data and passive surveillance data, we further model the unobserved incident cases N_s as a Poisson process for all sites where any surveillance system operates, $s = 1, \ldots, K, K+1, \ldots, S$. Our final proposed model is given by following hierarchical model that includes logistic regressions for individual-level capture indicators and a spatial Poisson regression for site-specific incident cases:

For linked sites, $s = 1, \ldots, K$:

$$\begin{gathered} P\left(Y_{1is} = y_{1is}, Y_{2is} = y_{2is} | y_{1is} = 1 \text{ or } y_{2is} = 1\right) \\ = \frac{p_{1is}^{y_{1is}}(1-p_{1is})^{(1-y_{1is})} p_{2is}^{y_{2is}}(1-p_{2is})^{(1-y_{2is})}}{1-(1-p_{1is})(1-p_{2is})}, \\ \log\left(\frac{p_{1is}(\boldsymbol{\beta}_1)}{1-p_{1is}(\boldsymbol{\beta}_1)}\right) = \boldsymbol{x}_{1is}^T\boldsymbol{\beta}_1, \quad \log\left(\frac{p_{2is}(\boldsymbol{\beta}_2)}{1-p_{2is}(\boldsymbol{\beta}_2)}\right) = \boldsymbol{x}_{2is}^T\boldsymbol{\beta}_2, i = 1, \ldots, n_s, \\ n_s = \sum_i^{N_s} I(y_{1is} = 1 \text{ or } y_{2is} = 1), \end{gathered} \tag{2.1}$$

For unlinked sites, $s = K+1, \ldots, S$:

$$n_s = \sum_{i=1}^{N_s} I(y_{1is} = 1), \tag{2.2}$$

For all sites, $s = 1, \ldots, S$:

$$\begin{gathered} N_s | \lambda_s(\boldsymbol{\alpha}, \epsilon_s) \sim \text{Poisson}(\lambda_s(\boldsymbol{\alpha}, \epsilon_s)), \\ \log(\lambda_s(\boldsymbol{\alpha}, \epsilon_s)) = \boldsymbol{z}_s^T\boldsymbol{\alpha} + \epsilon_s, \end{gathered} \tag{2.3}$$

where n_s is number of disease cases identified by surveillance systems at site s, $\boldsymbol{x}_{1is}$ and $\boldsymbol{x}_{2is}$ denote vectors of individual and location-specific covariates, such as demographic and residence information with corresponding system-specific coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, $I(\cdot)$ denotes the indicator function, $\lambda_s(\boldsymbol{\alpha}, \epsilon_s)$ is the expected number of incident cases at site s, $\boldsymbol{z}_s$ is a vector of risk factors for the disease of interest with regression coefficient $\boldsymbol{\alpha}$, and spatial random effects ϵ_s follow a mean-zero Gaussian process with an exponential covariance function. Specifically, $(\epsilon_1, \ldots, \epsilon_S)^T \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}_{ij} = \sigma^2 \exp(-\frac{D_{s,s'}}{\phi})$ where $D_{s,s'}$ denotes the valid distance metric between site s and s', and σ^2 and ϕ represent marginal variance and range parameter in exponential covariance function respectively.

2.3.2 A Two-Stage Bayesian Procedure for Inference

Our goal is to obtain estimates of total cases N and site-specific cases N_s in the study region correcting for under-ascertainment of the surveillance system. The main difficulty in estimation is that site-specific cases N_s do not appear in the individual-level capture history likelihood. Hence, posterior samples of N_s cannot be directly obtained from specifying a full conditional distribution that includes contributions from both the BM model and the spatial Poisson model. We propose a Bayesian two-stage estimation procedure using Markov Chain Monte Carlo (MCMC) algorithm for the hierarchical model. In the first stage, model coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are estimated with the CRC data at the K linked sites. Given estimates of capture probabilities of observed individuals, posterior distributions of N_s are obtained using the Horvitz-Thompson estimator. In the second stage, the estimated N_s and their associated uncertainties are treated as the observed data likelihood for the spatial Poisson regression. Furthermore, a simulation-based procedure based on posterior predictive distributions of the capture history is developed to account for using the Horvitz-Thompson (H-T) estimator in first-stage estimation. Details of the proposed two-stage estimation procedure are presented in subsequent sections.

2.3.3 First-Stage Estimation

First-stage estimation is based on the BM model with CRC data formed by augmenting the passive system with the active system at K linked sites. The likelihood of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be written as

$$L\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \{\boldsymbol{y}_{1,s}, \boldsymbol{y}_{2,s}\}_{s=1}^{K}\right) = \prod_{s=1}^{K}\prod_{i=1}^{n_s} \frac{p_{1is}^{y_{1is}}(1-p_{1is})^{(1-y_{1is})} p_{2is}^{y_{2is}}(1-p_{2is})^{(1-y_{2is})}}{1-(1-p_{1is})(1-p_{2is})}, \quad (2.4)$$

where $\boldsymbol{y}_{t,s} = \{y_{tis}, i = 1\ldots, n_s\}$ for $t = 1, 2$ denotes individual capture history at site s. The posterior distributions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are sampled by Metropolis-Hastings

(M-H) algorithms. With posterior samples of model coefficients, the posterior distributions of individual-specific capture probabilities can be obtained by applying inverse logit transformation to linear combinations of posterior samples of model coefficients. Specifically, for the active system, the posterior distribution of p_{2is} is given by $\text{logit}^{-1}(\boldsymbol{x}_{2is}^T\boldsymbol{\beta}_2)$, $i = 1, \ldots, n_s, s = 1, \ldots, K$. For the passive system, capture probabilities can be obtained for both linked and unlinked sites. In other words, p_{1is} is computed for each individual who has been identified by the passive system, $i = 1, \ldots, n_s, s = 1, \ldots, S$.

The Horvitz-Thompson estimator of N_s is given by

$$N_s^* = \sum_{i=1}^{n_s} 1/q_{is}, \quad q_{is} := \begin{cases} 1 - (1 - p_{1is})(1 - p_{2is}) & \text{for } s = 1, \ldots, K \\ p_{1is} & \text{for } \ s = K+1, \ldots, S. \end{cases} \tag{2.5}$$

Simply applying posterior samples of p_{1is} and p_{2is} to the Equation (2.5) will provide conditional posterior samples of N_s^* assuming known individual capture history, p_{1is}, and p_{2is}. We denote the desired first-stage marginal posterior distribution as $[N_s^*|\boldsymbol{h}_s]$, where $\boldsymbol{h}_s = \{\boldsymbol{y}_{1,s}, \boldsymbol{y}_{2,s}\}$ for linked sites and $\boldsymbol{h}_s = \boldsymbol{y}_{1,s}$ for unlinked sites. To appropriately propagate statistical uncertainties in estimating coefficients in the BM model to the H-T estimator (i.e., generate samples from $[N_s^*|\boldsymbol{h}_s]$), we draw a random sample from the limiting distribution of the H-T estimator $N(\frac{1}{q_{is}}, \frac{1-q_{is}}{q_{is}^2})$ for each posterior sample of q_{is} (Huggins, 1989). This additional sampling step approximates the uncertainty in N_s^* with known capture probabilities. As a result of this simulation-based procedure, we obtain posterior samples which appropriately quantify uncertainties associated with using the H-T estimator. Finally, a point estimate of N_s denoted as $\widehat{N}_s$ is obtained by computing the posterior mean; 95% credible intervals are obtained by taking the 2.5% and 97.5% percentile of the resulting posterior samples. The posterior samples of N are obtained by summing up posterior samples of N_s. In summary, first-stage estimation yields preliminary estimates of total and site-specific

incident cases solely based on CRC data at linked sites.

2.3.4 Second-Stage Estimation

Let $\hat{\boldsymbol{N}} = (\hat{N}_1, \ldots, \hat{N}_S)^T$ denote the vector of preliminary estimates of site-specific incident cases obtained from first-stage estimation and $\boldsymbol{N} = (N_1, \ldots, N_S)^T$ denote the vector of true site-specific incident cases. To connect results from first-stage estimation to the Poisson log-linear model, we first note that the target joint posterior distribution of $\boldsymbol{\Theta} = \{\boldsymbol{N}, \boldsymbol{\alpha}, \sigma^2, \phi\}$ is

$$[\boldsymbol{\Theta}|\hat{\boldsymbol{N}}] \propto \int [\hat{\boldsymbol{N}}|\boldsymbol{N}, \{\boldsymbol{h}_s\}_{s=1}^S] \times [\boldsymbol{N}|\boldsymbol{\lambda}] \times [\boldsymbol{\lambda}|\boldsymbol{\alpha}, \sigma^2, \phi] \times [\boldsymbol{\alpha}] \times [\sigma^2] \times [\phi] d\boldsymbol{h}_1 \ldots d\boldsymbol{h}_S, \quad (2.6)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)^T$. To obtain posterior samples of $\boldsymbol{\Theta}$ with M-H algorithms based on the Equation (2.6) using results obtained from first-stage estimation, we apply two approximations. Firstly, to approximate $[\hat{\boldsymbol{N}}|\boldsymbol{N}, \{\boldsymbol{h}_s\}_{s=1}^S]$, we use a multivariate normal distribution with mean $\boldsymbol{N}$ and variance-covariance matrix computed as the empirical variance of $\hat{\boldsymbol{N}}$ from first-stage posterior samples. The second approximation replaces the outer integration of capture history via an additional sampling step.

We begin by generating posterior samples of $\boldsymbol{\Theta}$ using Equation (2.6) while ignoring the outer integration (i.e., assuming capture history $\boldsymbol{h}_s$ is known from observed CRC data). We refer to these posterior samples as naive posterior samples, which will underestimate uncertainties associated with posterior distribution of $\boldsymbol{N}$ because uncertainties of capture histories have not been incorporated. To approximate the outer integration of Equation (2.6), $\hat{N}_s$ and estimated coefficients of BM models are used to impute multiple datasets of individual-level capture history. In other words, we repeatedly draw observations $\{y_{1is}, y_{2is}\}$ from their posterior predictive distributions. Each imputed dataset is analyzed separately using the two-stage estimation

procedure to generate naive posterior samples of $\boldsymbol{N}$. Finally, naive posterior samples of all model parameters are pooled across imputed datasets to approximate samples generated from the desired distribution presented in Equation (2.6). Full details of generating imputed dataset are given in Appendix A.1.

2.3.5 Prior Distributions

We assign weakly informative priors for model parameters. The prior distribution for each element in model parameter vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_s$ and $\boldsymbol{\alpha}$ are normal distributions. For spatial random effect σ^2 and range parameter ϕ, the prior distributions are Inverse-Gamma and Gamma distributions respectively. All distribution parameters are specified at values that represent weakly informative priors.

2.4 Simulation Studies

We conducted simulation studies to evaluate the performance of the proposed modeling framework in estimating the number of incident cases. We focused on bias, root mean square error (RMSE) and 95% credible interval coverage for the total incident cases N. A variety of statistical models have been developed for analyzing CRC data. However, none of them are directly comparable to our proposed model, which allows for individual-level covariates in capture probabilities and borrows information across surveillance sites within a unified framework. Hence, our simulation studies aim to illustrate potential advantages of the proposed model to the most competitive method for individual-level CRC data, i.e., the BM model. Implementation of the BM model can be viewed as the first-stage estimation of our model. We note that the BM model does not borrow information across sites, nor does it utilize site-level covariates for case rate because the likelihood does not involve the unobserved true number of cases. We also examined the impacts of ignoring the spatial dependency

by specifying exchangeable location-specific random effects (i.e., non-spatial model) and including multiple active systems.

2.4.1 Simulation Design

We generated site-specific incident cases N_s from a spatial Poisson process with log mean α and an exponential covariance function for 50 sites. For each simulated individual, capture history was generated from a Binomial distribution with capture probabilities determined by logit models for the passive and active systems. The logit models contained individual-level continuous covariates $X_1 \sim \text{Uniform}(-1, 1)$ and $X_2 \sim \text{Uniform}(-2, 2)$ respectively with both regression coefficients set to 1. Since the logit models include two different covariates, two systems work independently at the population level. Finally, the CRC data were obtained at 10 randomly sampled sites by keeping individuals who were captured by at least one system. For the other 40 unlinked sites, only the capture history provided by the passive system was recorded. We examined the following simulation scenarios: (1) high versus low mean capture probabilities of the two systems by varying intercepts of the logit models (β_{10} and β_{20}), (2) strong versus weak spatial dependence in log case rate via parameter ϕ, (3) high versus low case rate controlled by log mean α. Smaller values for β_{10}, β_{20}, α and ϕ correspond to smaller mean capture probabilities for both systems, smaller overall case rate, and weaker spatial correlation between site-specific incident cases. A total of eight simulation scenarios were designed and 100 datasets were generated for each scenario. To mimic the real-world situation, N and N_s were fixed for each scenario, but individual-level capture history changed across replications. Details of data generation procedure for the simulation studies can be found in Appendix A.2.1.

2.4.2 Comparison with One-Stage BM Model

The estimated mean bias, RMSE and coverage of 95% credible intervals of N for different simulation scenarios from the BM model, the proposed model, and the proposed model using exchangeable random effects in the second-stage estimation are summarized in Table 2.4. In all scenarios, our proposed model outperforms the BM model in terms of mean bias and RMSE when the spatial dependency is either incorporated or not. For the BM model, a positive bias is observed across all scenarios, especially when capture probabilities are small. This is likely because the Horvitz-Thompson estimator may result in estimates which are biased upward when estimated capture probabilities are supplied, the positive bias may be more evident when number of captured cases is small (i.e., capture probabilities are small) (McDonald and Amstrup, 2001; Tounkara and Rivest, 2015). The bias reduction associated with the proposed model is most remarkable when capture probabilities are small for both systems. For example, the mean bias decreased from 139 to 11 under Scenario 1 when average capture probabilities for both systems are around 27%. Although our proposed model has larger mean bias in Scenario 6, the difference is negligible compared to the true N. This could be due to the large number of observed cases to sufficiently estimate N_s based on the BM model.

The relative RMSE using BM model as reference indicates that our proposed model with or without correct specification of the spatial dependency reduced RMSE under all scenarios, particularly when spatial correlation is strong. Under most of scenarios, the proposed model correctly incorporating spatial dependency has better performance in terms of RMSE compared to the proposed model using exchangeable random effects.

In terms of the coverage, the 95% credible intervals constructed by all models achieved or were close to the nominal level under most of scenarios. These results support the validity of the proposed two-stage estimation procedure in conducting

inference of N. Without implementing the simulation-based procedure used for approximating the integral in Equation (2.6), the 95% CI resulted in under-coverage as suggested by the simulation results presented in Table 2.5.

2.4.3 Positive Dependence between Systems

The simulations above assumed the two surveillance systems have independent capture probabilities at the population level. We also examined the performance of the proposed model when two surveillance systems are positively correlated at the population level. The positive dependency between two systems are introduced by including a common covariate to the two logit capture probability models; details about this simulation study can be found in Appendix A.2.2. The mean bias and RMSE summarized from this simulation study suggest that the proposed model still outperforms the BM model under all simulation scenarios as shown in Table 2.6.

2.4.4 Benefits of Multiple Active Systems

Our motivating application involves three surveillance systems, where a passive system is linked to two active systems, resulting in two independent CRC datasets. To illustrate the benefits of additional linkage between passive and active surveillance systems, we extended our simulation study to the scenario where two active systems that do not have overlap are available. The data generation procedure is similar to the procedure outlined for the previous simulation study with following differences: (1) two sets of linked sites without spatial overlap were simulated; (2) a total of three logit models including an intercept and a common covariate were specified for each system, separately. The simulated data resulted in 100 cases on average at each site, and averaged capture probability of systems 1 (passive), 2 (active), and 3 (active) are 0.27, 0.35, and 0.33, respectively. We simulated a total of 100 sites, 10 sites were randomly selected as sites where S1 and S2 are linked, and 20 other sites were randomly

selected as sites where S1 and S3 are linked. Details about modifying the proposed model to analyze such data and the full parameters set-up can be found in Appendix A.3 and A.2.3.

Table 2.7 presents the averaged estimates, averaged posterior standard deviation, and RMSE of regression coefficients of the BM model from using two datasets and one of them. With both active systems being linked to the passive system, coefficients associated with those three systems can be estimated simultaneously. As we expect, estimates of coefficients are empirically unbiased regardless of the degree of the system linkage. However, RMSE was reduced when two active systems were analyzed together compared to analyzing datasets separately (S1-S2 and S1-S3), particularly for coefficients related to capture probabilities of S1, the system is linked to both active systems. For example, comparing analysis using only S1-S2 linked data to using all data, the RMSE reduced from 0.32 to 0.19 for the intercept. This can be explained by the fact that both S2 and S3 contribute to the data likelihood used for estimating coefficients associated with S1. We also observed that RMSE of coefficients associated with systems 2 and 3 decreased when two datasets were used simultaneously. This reduction is likely because the precision of estimated coefficients associated with S1 improved, resulting in more accurate estimation of capture probability associated with the two other active systems. Finally, as a result of improved estimation for coefficients of the BM model, the estimation of total number of cases also improved in the second-stage estimation when additional active systems are available (Table 2.8).

Table 2.4: Mean bias, RMSE, coverage of 95% credible intervals of N estimated from BM model, the proposed model, and the proposed model using exchangeable random effcets when two systems are independent at the population level over 100 simulated data

		BM model			Proposed model				Proposed model (exchangeable random effects)			
Scenario	True N	Mean bias	Coverage	RMSE	Mean bias	Coverage	RMSE	Relative[a] RMSE	Mean bias	Coverage	RMSE	Relative[a] RMSE
Low prevalence ($\alpha = 3.3$), weak spatial dependence ($\phi = 10$)												
1 Low capture rate ($\beta_{10} = \beta_{20} = -1$)	1550	139	0.90	362	11	0.91	336	0.93	-17	0.88	333	0.92
2 High capture rate ($\beta_{10} = \beta_{20} = -0.5$)	1896	70	0.98	241	-13	0.98	199	0.83	-17	0.99	207	0.86
Low prevalence ($\alpha = 3.3$), strong spatial dependence ($\phi = 30$)												
3 Low capture rate ($\beta_{10} = \beta_{20} = -1$)	1643	153	0.95	354	-29	0.93	261	0.74	57	0.96	284	0.80
4 High capture rate ($\beta_{10} = \beta_{20} = -0.5$)	1265	78	0.95	199	-27	0.98	149	0.75	2	0.97	157	0.79
High prevalence ($\alpha = 4.3$), weak spatial dependence ($\phi = 10$)												
5 Low capture rate ($\beta_{10} = \beta_{20} = -1$)	3659	116	0.96	435	35	0.95	409	0.94	-11	0.97	379	0.87
6 High capture rate ($\beta_{10} = \beta_{20} = -0.5$)	4180	5	0.96	301	-16	0.98	282	0.94	-21	0.97	280	0.93
High prevalence ($\alpha = 4.3$), strong spatial dependence ($\phi = 30$)												
7 Low capture rate ($\beta_{10} = \beta_{20} = -1$)	4616	108	0.94	462	-24	0.95	426	0.92	41	0.96	446	0.97
8 High capture rate ($\beta_{10} = \beta_{20} = -0.5$)	3612	82	0.96	288	-11	0.98	257	0.89	43	0.99	264	0.92

[a] Relative RMSE is computed using BM model as reference

Table 2.5: Mean bias, RMSE, coverage of 95% credible intervals of N estimated from the proposed model without implementing the simulation-based procedure used for approximating the integral in Equation (2.6) over 100 simulated data when two systems are independent at the population level

Scenario	Mean bias	Coverage	RMSE
1	32	0.65	341
2	-19	0.88	225
3	74	0.68	314
4	25	0.87	176
5	24	0.75	502
6	-49	0.80	315
7	7	0.73	433
8	2	0.80	306

Table 2.6: Mean bias, RMSE, coverage of 95% credible intervals of N estimated from BM model and the proposed model over 100 simulated data when two systems are positively correlated at the population level

		BM model			Proposed model			
Scenario	True N	Mean bias	Coverage	RMSE	Mean bias	Coverage	RMSE	Relative[a] RMSE
Low prevalence ($\alpha = 3.9$), weak spatial dependence ($\phi = 10$)								
1 Low capture rate ($\beta_{10} = \beta_{20} = -1.75$)	2470	268	0.93	647	90	0.94	474	0.73
2 High capture rate ($\beta_{10} = \beta_{20} = -1.5$)	2720	143	0.98	427	-6	0.97	309	0.72
Low prevalence ($\alpha = 3.9$), strong spatial dependence ($\phi = 30$)								
3 Low capture rate ($\beta_{10} = \beta_{20} = -1.75$)	2910	136	0.97	448	-60	0.92	368	0.82
4 High capture rate ($\beta_{10} = \beta_{20} = -1.5$)	3109	138	0.95	403	56	0.97	315	0.78
High prevalence ($\alpha = 4.6$), weak spatial dependence ($\phi = 10$)								
5 Low capture rate ($\beta_{10} = \beta_{20} = -1.75$)	6355	105	0.99	521	4	0.97	472	0.91
6 High capture rate ($\beta_{10} = \beta_{20} = -1.5$)	6342	167	0.95	542	81	0.95	474	0.87
High prevalence ($\alpha = 4.6$), strong spatial dependence ($\phi = 30$)								
7 Low capture rate ($\beta_{10} = \beta_{20} = -1.75$)	6918	78	0.97	571	-29	0.96	493	0.85
8 High capture rate ($\beta_{10} = \beta_{20} = -1.5$)	5224	185	0.95	538	107	0.92	476	0.89

[a] Relative RMSE is computed using the BM model as reference

Table 2.7: Mean bias, averaged posterior standard deviation (SD), RMSE of regression coefficients in the BM model over 100 simulated data under situation where the passive system is linked to two active systems

		S1-S2 and S1-S3 linked[a]			S1-S2 linked[b]			S1-S3 linked[c]		
	True value	Mean bias	Posterior SD	RMSE	Mean bias	Posterior SD	RMSE	Mean bias	Posterior SD	RMSE
S1 intercept	-1.75	-0.05	0.19	0.19	-0.07	0.33	0.32	-0.07	0.23	0.22
slope	1	0.04	0.18	0.17	0.04	0.31	0.28	0.06	0.22	0.20
S2 intercept	-1.75	-0.02	0.26	0.25	-0.04	0.34	0.32	-	-	-
slope	1.5	0.04	0.27	0.26	0.04	0.33	0.30	-	-	-
S3 intercept	-1.85	0.05	0.21	0.23	-	-	-	0.04	0.24	0.24
slope	1.5	0.05	0.21	0.24	-	-	-	0.06	0.23	0.25

* S1, S2, and S3 refer to surveillance systems 1, 2, and 3, respectively.
[a] Using both S1-S2 and S1-S3 linked data to estimate coefficients; all coefficients can be estimated.
[b] Using only S1-S2 data to estimate coefficients; only coefficients associated with capture probabilities of S1 and S2 can be estimated.
[c] Using only S1-S3 data to estimate coefficients; only coefficients associated with capture probabilities of S1 and S3 can be estimated.

Table 2.8: Mean bias, RMSE, coverage of 95% credible intervals of N estimated from the proposed model over 100 simulations under situation where the passive system is linked to two active systems

		S1-S2 and S1-S3 linked[a]			S1-S2 linked[b]			S1-S3 linked[c]		
	True N	Mean bias	RMSE	Coverage	Mean bias	RMSE	Coverage	Mean bias	RMSE	Coverage
BM model	9712	234	696	0.97	572	1404	0.92	297	878	0.95
Proposed model	9712	150	634	0.96	182	1129	0.90	65	828	0.87

[a] Using both S1-S2 and S1-S3 linked data.
[b] Using only S1-S2 data.
[c] Using only S1-S3 data.

2.5 Application

The proposed framework was applied to bias-correct reported PTB cases from the widely-distributed passive system (S1) in Sichuan in 2010. We modified the model to allow for two independent CRC datasets, i.e., the S1-S2 and S1-S3 linked datasets. In addition to estimating PTB cases, operating characteristics of the three surveillance systems and risk factors of PTB were also explored using the proposed model. The final logit models and spatial Poisson log-linear model including at-risk populations as an offset were selected by retaining covariates with 95% CI excluding 0. In this data analysis, age group with age < 40 and female were used as reference group; all continuous risk factors of PTB and geographical information (longitude, latitude and elevation) were centered at their means computed across all sites.

Goodness of fit of the model was assessed by comparing several quantities summarizing observed data to their corresponding posterior predictive distributions. We used the following quantities to summarize observed PTB data collected in Sichuan. For the S1-S2 linked data, we compute the number of PTB cases captured by S1 and S2 alone at linked sites, denoted as n_{S_1}, n_{S_2}, respectively. Similarly, we compute the number of PTB cases captured by S1 and S3 alone and denoted as $n^*_{S_1}$, n_{S_3}. From Table 2.1, we have $n_{S_1} = 121$, $n_{S_2} = 201$, $n^*_{S_1} = 240$, and $n_{S_3} = 257$. We also computed the number of PTB cases captured by S1 alone for each of 181 counties. The posterior predictive distributions of those quantities were obtained by simulation. Specifically, we drew observed outcomes (e.g., capture indicators of S1 at S1-S2 linked sites) from its predictive distribution and then computed posterior median and 95% posterior intervals. 10,000 samples were generated and details of simulating from the corresponding posterior predictive distributions are provided in Appendix A.4.

Figure 2.2 displays observed and posterior predictive distributions for summary statistics of observed linked and unlinked data to assess model fit. We observed that posterior predictive distributions are centered around observed values, which suggest

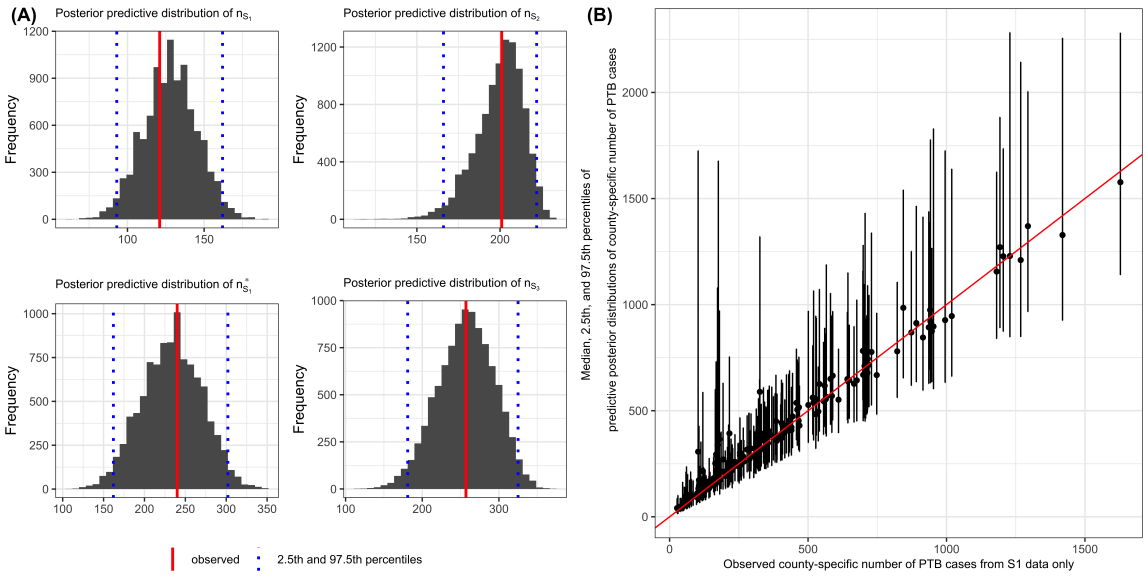


Figure 2.2: Observed and posterior predictive distributions for quantities summarizing observed data.

that the final model fitted PTB data well.

The estimated odds ratios for system-specific capture probability are presented in Table 2.9. Capture probability of S1 increased with latitude, decreased over time and was higher for male cases. Unlike S1, the capture probabilities of the cross-sectional active system (S2) tended to have higher capture of cases in older age groups. Similar age effects were also observed in the sentinel system (S3). Additionally, S3 capture probability was strongly impacted by elevation. We also found that S3 had higher capture probability among patients who resided at locations with high GDP. The overall capture probabilities for each system were estimated to be 0.36 (95% CI: 0.32, 0.41) for S1, 0.73 (95% CI: 0.66, 0.79) for S2 and 0.37 (95% CI: 0.31, 0.43) for S3 respectively.

Table 2.10 presents the associations between several risk factors and PTB prevalence from the Poisson log-linear model. The effect of elevation on PTB prevalence (1.48 95%CI: 1.17, 1.85) indicates that areas with higher elevation tend to have higher PTB prevalence. Higher elevation areas in the study region are generally rural (most major urban areas reside in the low elevation plains) and exhibit lower socioeconomic

Table 2.9: Results from the first-stage estimation applied to analyzing individual capture probabilities of PTB data

Systems	Covariates	Estimated ORs (posterior SD)	95% Credible Intervals
S1	Baseline	0.76 (0.16)	(0.50, 1.12)
	Longitude	1.18 (0.11)	(0.98, 1.41)
	Latitude	0.80 (0.06)	(0.68, 0.92)
	Year	0.84 (0.04)	(0.77, 0.92)
	Gender		
	male *versus* female	1.37 (0.25)	(0.95, 1.92)
S2	Baseline	0.82 (0.30)	(0.41, 1.55)
	Age (years)		
	$[40, 50)$ *versus* < 40	3.14 (1.78)	(0.99, 7.65)
	$[50, 60)$ *versus* < 40	5.65 (3.98)	(1.53, 15.99)
	$[60, 70)$ *versus* < 40	10.03 (7.17)	(2.67, 28.17)
	≥ 70 *versus* < 40	10.89 (13.18)	(2.06, 36.65)
S3	Baseline	0.45 (0.13)	(0.26, 0.75)
	GDP	1.33 (0.18)	(1.02, 1.73)
	Elevation	10.41 (6.83)	(2.84, 27.92)
	Age (years)		
	$[40, 50)$ *versus* < 40	1.77 (0.59)	(0.88, 3.17)
	$[50, 60)$ *versus* < 40	2.35 (0.77)	(1.19, 4.24)
	$[60, 70)$ *versus* < 40	2.75 (0.84)	(1.46, 4.75)
	≥ 70 *versus* < 40	5.89 (2.52)	(2.50, 12.04)

status and poorer quality housing (Liu et al., 2005; Wanyeki et al., 2006). Proportion elderly and average area per capita (a measure of crowding) were found to be significantly associated with PTB prevalence; and in fact they are known risk factors of PTB (Wanyeki et al., 2006; Lönnroth et al., 2009; Li et al., 2019).

From our hierarchical model, the total number of PTB cases in Sichuan in 2010 is estimated to be 170,579 (95% CI: 138,180, 209,971), which is more than twice the number of cases observed in S1 alone. Figure 2.3 displays the maps of observed versus adjusted incident PTB cases from the passive system (S1) and the corresponding prevalence (per 1000 population). There is high spatial heterogeneity in PTB cases across the study region, and S1 suffers from substantial under-ascertainment. The prevalence estimates in Figure 2.3 help identify areas with high PTB prevalence that may be neglected if only unadjusted S1 data are considered. To better visualize differences spatially, Figure A.1 and Figure A.2 present the estimated bias for the number of PTB cases and PTB prevalence associated with S1.

Table 2.10: Results from the second-stage estimation applied to analyzing PTB prevalence of PTB data

Covariates	**Estimated RRs (posterior SD)**	**Posterior 95% CIs**
Elevation	1.48 (0.17)	(1.17, 1.85)
Average area	0.94 (0.02)	(0.89, 0.99)
Percent of elderly	1.16 (0.05)	(1.06, 1.27)
σ^2	0.32 (0.08)	(0.21, 0.52)
ϕ	55.75 (15.49)	(32.88, 92.40)

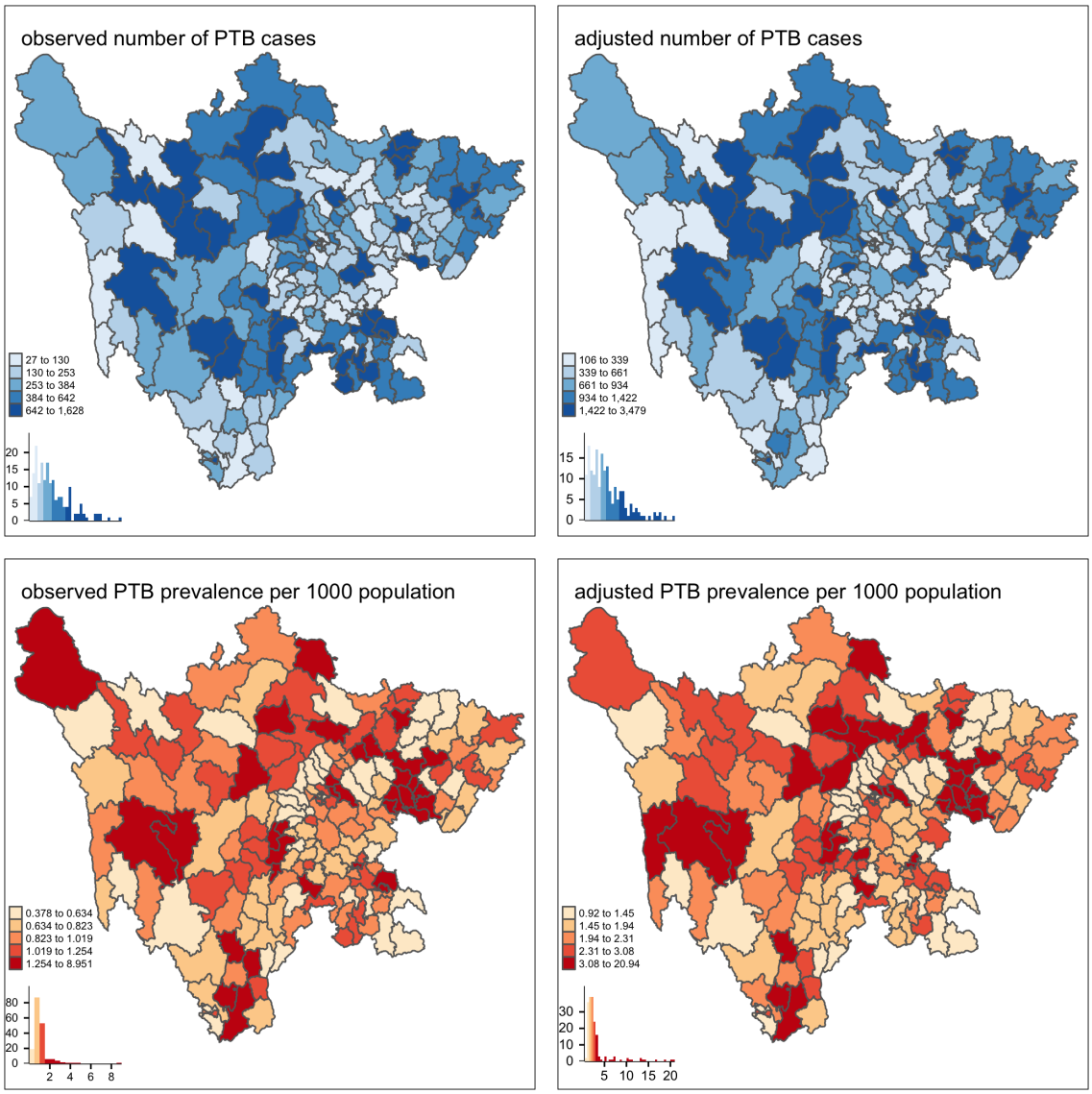


Figure 2.3: A map of observed/adjusted number of PTB cases, and observed/adjusted PTB prevalence per 1000 population in Sichuan in 2010 with county borders.

2.6 Discussion

In this chapter, we propose a statistical framework that integrates a BM model with a spatial Poisson model to address under-ascertainment of a passive surveillance system when active surveillance systems data are also available to construct a CRC dataset in sub-regions. To our best knowledge, this is the first time the commonly used BM

model for individual-level CRC data and Poisson model for aggregate-level incident cases are employed jointly. Doing so allows us to incorporate individual-level covariates for capture probabilities, covariates for disease rates, and spatial dependence. A two-stage Bayesian estimation procedure is developed for conducting inference of the incident cases, while addressing missing covariate information for individuals not captured by any system. Compared to the methods which solely rely on CRC data, the additional latent process on N_s allows us to utilize data from locations without active surveillance.

Applying the proposed model to the study of PTB in the study region, we identified factors that affect the capture probabilities of three surveillance systems that monitors PTB, risk factors and spatial pattern of PTB cases. Passive surveillance (S1) exhibited the lowest average capture probability and was impacted by geographic location, time period and case sex. When appropriate, estimates of parameters that govern S1 capture probabilities can be used to bias-correct S1 PTB cases at locations and years without active surveillance. Surprisingly, on average, the capture probability of sentinel surveillance (S3) was low (0.37) and was impacted by economic conditions, geographic location and age. This low capture probability may indicate that this active surveillance effort was not well implemented. The factors identified as influencing capture probabilities may be considered as strategies are sought to strengthen public health surveillance in the study region.

Our proposed model can be viewed as a general hierarchical framework for correcting raw cases reported by passive surveillance systems when other samples of the epidemiological process of interest are available. Specifically, the first-stage model does not have to be restricted to the BM model presented in Section 2.3. Any model that provides site-specific estimation of incident cases based on CRC data, such as models that belong to the classes of mixture models, can be utilized. Our model can also be easily extended to dealing with the situation in which the targeted passive

system can be augmented with multiple active surveillance systems, since the BM model used in the first-stage estimation is able to accommodate CRC data formed by multiple surveillance efforts (Tounkara and Rivest, 2015). In light of this generality, one important future extension of the proposed model is to include consideration of unmeasured heterogeneity in capture probability. For example, to account for both measured and unmeasured heterogeneity, the logit-normal model proposed by Coull and Agresti (1999) is an appropriate model considered in the first-stage estimation.

The Horvitz-Thompson estimator may result in unrealistically large estimation of site-specific incident cases when estimated individual-specific capture probabilities are close to zero. This is because individuals who have very low probabilities of being captured will contribute greatly to the estimation of N. These unrealistically large estimates may affect our second-stage estimation. This warrants a further direction to develop or incorporate an improved estimator that is robust to extremely small capture probabilities as a substitute for the Horvitz-Thompson estimator. For example, one convention adopted in estimating average treatment effect with inverse probability weighting, is to restrict propensity scores within the range [0.1, 0.9] (Crump et al., 2009).

Chapter 3

Sensitivity and Uncertainty Analysis for Two-Stream CRC Methods

3.1 Background

It is important to note that essentially all CRC methods rely upon untestable assumptions to enable estimation of the total number of diseased cases N. Here, untestable implies that observed data contain no information about justifying imposed assumptions. For example, the classical independence assumption discussed in Section 1.2 cannot be justified solely based on the observed data. This point has been widely discussed in the literature. For instance, under two-catch CRC, Darroch et al. (1993) observed that the cross-product ratio measuring the dependency between two surveillance efforts cannot be estimated directly based on the observed data alone, and that assumptions must be made in order to identify an estimate of this ratio to permit estimation of N. Additionally, Coull and Agresti (1999) and Dorazio and Andrew Royle (2003) suggested that regularly used model selection metrics such as Akaike's information criterion (AIC; Akaike (1974)) are not adequate tools for defending untestable model-based assumptions to which estimators of N can be extremely sensitive. This sensitivity has been demonstrated in empirical and applied settings, such as a study to quantify workplace injuries and illnesses via CRC which showed that the estimate is sensitive to assumptions about both dependence of surveillance efforts and the heterogeneity in individual capture probabilities (Boden, 2014). A recent effort to estimate the number of healthcare workers who have died fighting COVID-19 based on two overlapping surveillance efforts similarly illustrated that the estimated number of diseased cases are sensitive to the specification of the coefficient of interaction term in the adapted log-linear model (Zhang and Small, 2020). The exponentiation of this coefficient is equivalent to the cross-product ratio discussed in Darroch et al. (1993).

In epidemiologic research, it is now commonplace to study the effect of untestable assumptions through sensitivity analysis, such as assessing the effect of misclassification and unmeasured confounders (Fox et al., 2005; Groenwold et al., 2010; Lyles and Lin, 2010). In CRC settings, a few authors have considered sensitivity analysis

to evaluate the effect of untestable assumptions on estimation of N (Boden, 2014; Gerritse et al., 2015; Zhang and Small, 2020). Specifically, the study conducted in Boden (2014) to quantify workplace injuries and illnesses via CRC has showed that the estimate is sensitive to assumptions about both dependence of surveillance efforts and the heterogeneity in individual capture probabilities. Recently, Zhang and Small (2020) illustrated that the estimated number of healthcare workers who have died fighting COVID-19 obtained from the adapted log-linear model is sensitive to the specification of the coefficient of interaction term in the model when analyzing a two-catch CRC data.

In this chapter, we focus on CRC experiments with two streams. We develop a sensitivity analysis providing a novel visualization based on a key inestimable parameter reflecting the assumed level of association between the two data streams under a population-level multinomial model (Darroch, 1958; Seber et al., 1982). In addition to exploring the sensitivity of the N estimation with respect to the key inestimable parameter, we further propose a simulation-based uncertainty analysis for quantifying uncertainty in the estimate of N associated with the variation in the key inestimable parameter, together with statistical uncertainties. Relevant prior work in the case of measurement error modeling suggests that incorporating variation in a sensitivity parameter by introducing a subjective prior can be preferable to assuming a fixed value for the sensitivity parameter, in terms of inferential properties such as the tradeoff of interval coverage and width (Gustafson, 2005).

The proposed uncertainty analysis can be conceptualized as a model averaging approach based on a simulation-based procedure by viewing the estimate of N with a given value of the key parameter as one fitted model. While model averaging approaches have been proposed in CRC contexts in a frequentist framework where model selection metrics are used for estimating weights for each model (Stanley and Burnham, 1998; Cameron et al., 2012), the uncertainty analysis advocated here does

not rely on model selection metrics (which are insufficient for defending CRC model-based assumptions (Coull and Agresti, 1999; Darroch et al., 1993)) and is similar in spirit to complex fully Bayesian analyses (Chatterjee and Mukherjee, 2016; Lee et al., 2003). Here, we target those epidemiologists practicing surveillance by crystallizing specification of assumptions about the key parameter, offering a clear and accessible approach to both sensitivity and uncertainty analyses, and demonstrating that covariates explaining heterogeneity in the population can be incorporated by stratifying. We generalize the classical recommendations for incorporating covariates by stratification (Sekar and Deming, 1949) or through regression models with categorical covariates (Huggins, 1989), in which independence of the capture efforts is assumed within each stratum. Instead, the proposed framework allows the epidemiologist to contemplate and implement stratum-specific assumptions about the dependency between data streams. Finally, we demonstrate that special cases of the proposed uncertainty analysis can serve as a general CRC interval estimation approach.

3.2 Methods

3.2.1 Maximum Likelihood Estimators

We follow the same notations and adapt the population-level multinomial model described in Section 1.2.2 for deriving MLEs of N under the two-catch case. Typical population-level two-catch CRC data are summarized in Table 1.2. Due to the unobserved cell count n_{00}, the population-level multinomial model is not directly applicable in practice. The conditional model $(N_{11}, N_{10}, N_{01}|N_c = n_c) \sim \text{Multinomial}(n_c, p^*_{11}, p^*_{10}, p^*_{01}, p^*_{00})$ is considered, where n_c is the number of cases caught at least once, p^*_{ij} is defined as p_{ij}/p_c and $p_c = p_{11} + p_{10} + p_{01}$ is the probability of being caught at least once. To allow for modeling the dependency between two streams, we introduce the parameters $(p_1, p_{2|1}, p_{2|\bar{1}})$, where p_1 is the marginal probability of

identification in stream 1, and $p_{2|1}$ and $p_{2|\bar{1}}$ are the probability of a case is identified by stream 2 given identified or not identified by stream 1, respectively. Among these three parameters, only $p_{2|1}$ can be estimated from the observed data. The other two parameters are inestimable, since the cell count n_{00} is not observed. As a result of this non-identifiability, the familiar CRC modeling challenge is that at least one unverifiable assumption is required for estimating N.

Consider a situation where a researcher assumes a specific value for the inestimable parameter $\psi = p_{2|\bar{1}}$. Given a valid ψ (i.e., $0 < \psi \leq 1$), the maximum likelihood estimator (MLE) of N and its variance estimator are given by (Chen, 2020; Lyles et al., 2021a):

$$\hat{N}_{\psi} = n_{11} + n_{10} + \frac{n_{01}}{\psi}, \tag{3.1}$$

$$\hat{Var}(\hat{N}_{\psi}) = \frac{(1-\psi)}{\psi^2} n_{01}. \tag{3.2}$$

Alternatively, the unverifiable assumption can be imposed less directly via a ratio $\phi = \frac{p_{2|1}}{p_{2|\bar{1}}}$. This ratio is a population-level measurement of the dependency between the two streams. The commonly assumed Lincoln-Petersen (LP) condition (i.e., independence assumption) corresponds to $\phi = 1$, whereas $\phi > 1$ suggests an overall positive association between the two streams (cases that are identified by stream 1 tend to be more likely to also be identified by stream 2), and < 1 indicates negative association. In the CRC literature, the case of $\phi > 1$ is usually referred to as "trap happiness", while $\phi < 1$ is known as "trap aversion" (Seber et al., 1982). With a known population-level ratio ϕ, the MLE of N and its variance estimator become (Chen, 2020; Lyles et al., 2021a):

$$\hat{N}_{\phi} = n_{11} + n_{10} + \frac{n_{01}(n_{11} + n_{10})}{n_{11}} \phi, \tag{3.3}$$

$$\hat{Var}(\hat{N}_{\phi}) = w_1(w_1 - 1)n_{11} + w_2(w_2 - 1)n_{10} + w_3(w_3 - 1)n_{01}, \tag{3.4}$$

where $w_1 = 1 - \frac{n_{01}n_{10}}{n_{11}^2}\phi$, $w_2 = 1 + \frac{n_{01}}{n_{11}}\phi$, and $w_3 = (1 + \frac{n_{10}}{n_{11}})\phi$. Here, Equation (3.4) is simpler and more generalizable expression of a result given in Chen (2020).

The odds ratio $\theta = \frac{p_{10}p_{01}}{p_{11}p_{00}}$ is another population-level measure of dependency that has been a focal point in previous literature (Darroch et al., 1993; Boden, 2014; Wolter, 1990). The parameters ϕ and θ are largely interchangeable in that the value 1 is a benchmark for both, if θ indicates the two streams are positively correlated, so does ϕ. The MLE of N based on known θ and its variance estimator are:

$$\hat{N}_\theta = n_{11} + n_{10} + \frac{n_{10}n_{01}}{n_{11}}\theta, \tag{3.5}$$

$$\hat{Var}(\hat{N}_\theta) = w_1(w_1 - 1)n_{11} + w_2(w_2 - 1)n_{10} + w_3(w_3 - 1)n_{01}, \tag{3.6}$$

where $w_1 = 1 - \frac{n_{01}n_{10}}{n_{11}^2}\theta$, $w_2 = 1 + \frac{n_{01}}{n_{11}}\theta$, and $w_3 = 1 + \frac{n_{10}}{n_{11}}\theta$. The detailed derivations for variance estimators of MLEs in Equations (3.1), (3.3), and (3.5) appear in Appendix B.3.

It is important to emphasize that assumptions are imposed at the population level. For example, the assumption $\phi = 1$ or $\theta = 1$ technically allows for a mixture of individuals characterized by trap-happiness and trap-aversion, if this mixture happens to result in $\phi = 1$ or $\theta = 1$ at the population level (Lyles et al., 2021a). Similarly, a constant value assumed for ψ does not necessarily imply that this value is the same for each individual. The parameters ψ and ϕ depend on the labeling of the data streams. Specifically, $p_{2|1}/p_{2|\bar{1}}$ is not necessarily equal to $p_{1|2}/p_{1|\bar{2}}$ and $p_{2|\bar{1}}$ is generally different from $p_{1|\bar{2}}$. When basing sensitivity analysis on ϕ or ψ, we suggest starting with selecting the most comfortable labeling, meaning that researchers are more confident in making assumptions about the key parameter under that selected labeling. In contrast, the parameter θ is invariant to different labeling.

The MLE in Equation (3.1) is strictly unbiased, while the MLEs in Equations (3.3) and (3.5) are biased. Their bias is asymptotically negligible, since those MLEs are consistent estimators (Sanathanan, 1972). To reduce mean bias associated with

$\hat{N}_\phi$ under small to moderate sample size conditions, we generalize a Taylor-series expansion approach taken by Lyles et al. (2021a) to allow the incorporation of any value of ϕ and derive two bias-corrected estimators referred to as the BC and BC2 estimators. These two bias-corrected estimators and their variance estimators are:

$$\hat{N}_\phi^{BC} = \hat{N}_\phi - \frac{n_{10}n_{01}}{n_{11}^2}\phi, \tag{3.7}$$

$$\hat{Var}\left(\hat{N}_\phi^{BC}\right) = w_1(w_1 - 1 - C)n_{11} + w_2(w_2 - 1 - C)n_{10} + w_3(w_3 - 1 - C)n_{01}, \tag{3.8}$$

where $C = \frac{n_{10}\hat{p}_{01}}{n_{11}^2}\phi$, $w_1 = 1 - \frac{n_{10}n_{01}}{n_{11}^2}\phi$, $w_2 = 1 + \frac{n_{01}}{n_{11}\phi}$, and $w_3 = \phi + \frac{n_{10}}{n_{11}}\phi - \frac{n_{10}}{n_{11}^2}\phi$, and

$$\hat{N}_\phi^{BC2} = \hat{N}_\phi - \frac{n_{10}n_{01}}{(n_{11} + 0.5)^2}\phi, \tag{3.9}$$

$$\hat{Var}(\hat{N}_\phi^{BC2}) = w_1(w_1 - 1 - C)n_{11} + w_2(w_2 - 1 - C)n_{10} + w_3(w_3 - 1 - C)n_{01}, \tag{3.10}$$

where $C = \frac{2n_{11}n_{01}\hat{p}_{01}}{(n_{11}+0.5)^3} - \frac{n_{10}\hat{p}_{01}}{(n_{11}+0.5)^2}\phi$, $w_1 = 1 - \frac{n_{10}n_{01}}{n_{11}^2}\phi + \frac{2n_{11}n_{01}}{(n_{11}+0.5)^2}$, $w_2 = 1 + \frac{n_{01}}{n_{11}}\phi - \frac{n_{01}}{(n_{11}+0.5)^2}\phi$, and $w_3 = \phi + \frac{n_{10}}{n_{11}}\phi - \frac{n_{10}}{(n_{11}+0.5)^2}\phi$. These same bias correction procedures can be applied to $\hat{N}_\theta$.

Under the Lincoln-Petersen conditions, we found the BC2 estimator to be nearly identical to the bias-corrected estimator of Chapman (Chapman, 1951; Lyles et al., 2021a). Thus, one can also consider a simple direct generalization, by introducing the parameter ϕ into the original form of the Chapman estimator to obtain another bias-corrected estimator; we refer to this estimator as a generalized Chapman estimator.

With given ϕ, the generalized estimator and its variance estimator are given by

$$\hat{N}^*_{Chap} = \frac{(n_{11}+n_{10}+1)(n_{11}+n_{01}\phi+1)}{(n_{11}+1} - 1, \tag{3.11}$$

$$\hat{Var}(\hat{N}^*_{Chap}) = w_1(w_1-1-C)n_{11} + w_2(w_2-1-C)n_{10} + w_3(w_3-1-C)n_{01}, \tag{3.12}$$

where $C = \frac{n_{10}\hat{p}_{01}}{(n_{11}+1)^2}\phi$, $w_1 = 1 - \frac{n_{10}n_{01}}{(n_{11}+1)^2}\phi$, $w_2 = 1 + \frac{n_{01}}{(n_{11}+1)}\phi$, and $w_3 = \phi + \frac{n_{10}}{(n_{11}+1)}\phi$. We provide derivations and algebraic forms for all these bias-corrected estimators and their variances in Appendix B.2 - B.3.

3.2.2 Sensitivity Analysis

Unverifiable assumptions about the population-level dependency can be couched in terms of different lynchpin parameters, which can be the basis for sensitivity analysis. For example, prior studies (Gerritse et al., 2015; Zhang and Small, 2020) chose the coefficient of the two-way interaction term in a log-linear model fitted to two-stream CRC data as the key parameter. Although this parameter is interpretable and the log-linear model is closely related to the multinomial model adopted here (Cormack, 1989; Lum and Ball, 2015; Sandland and Cormack, 1984), propagating the uncertainty of the interaction term could be difficult given that the profile likelihood approach is a highly recommended basis for obtaining log-linear model-based interval estimates (Sanathanan, 1972). In contrast, closed-form MLEs of N hinging on the parameters ψ, ϕ and θ can greatly facilitate the propagation of uncertainty in the approach proposed herein. Given that ψ and ϕ are more easily generalizable to incorporate multiple streams, here we focus on those two parameters. We illustrate this approach with publicly available HIV surveillance data, noting that a similar sensitivity analysis could also be applied using θ as the key parameter. Motivating data were collected from four data streams in Lazio, Italy, during 1990. For the purpose of illustration,

we selected one pair of these data streams (centers I and II); the data are presented in Table 3.1.

Table 3.1: Cell counts for two-stream CRC HIV data analyzed in Abeni et al. (1994)

	Captured in stream 2	
Captured in stream 1	Yes	No
Yes	14	222
No	679	?

Figure 3.1 shows how the MLE of N varies with the assumed values of ψ and ϕ, with point-wise Wald-type 95% confidence intervals (CIs) assuming known ψ or ϕ. Although the range of variation in ψ and ϕ cannot be compared directly, we note that the estimated N ranges from the 7,000 to 30,000 for the ψ within the range (0.02, 0.1), and varying ϕ from 0.75 to 3 corresponds to the estimated N varying from 8,820 to 34,574. The Lincoln–Petersen estimator and the estimator proposed by Chao (1987) are labeled. While the Lincoln–Petersen estimator assumes $\phi = 1$, the Chao estimator assumes a level of positive dependency estimated based on a model developed assuming conditions that would typically be unrealistic in two-stream epidemiologic surveillance (Lyles et al., 2021a). The Lincoln–Petersen estimator yields an estimate of 11,682 (95% CI: 5,807, 17,557); the corresponding estimated ψ is 0.059. The Chao estimator yields an estimate of 29,908 (95% CI: 14,252, 45,564), with corresponding estimated ψ and ϕ taking the values of 0.023 and 2.59, respectively. The huge difference between these two estimates of N stems from the vastly different projected values of ϕ.

The blue line with smaller slope in Figure 3.1(B) shows the sensitivity plot that would have resulted if 50 patients in the n_{01} cell had instead appeared in the n_{11} cell, with n_{10} remaining the same ($n_{11} = 64$, $n_{10} = 222$, $n_{01} = 679$). It is clear that based on this altered data, the estimation of N is far less sensitive to the assumption imposed on ϕ over the range depicted. This finding is consistent with prior observations (Gerritse et al., 2015) that estimation is less sensitive when the implied coverage (measured by

$\frac{n_{11}}{n_{11}+n_{01}}$) of the two streams is high.

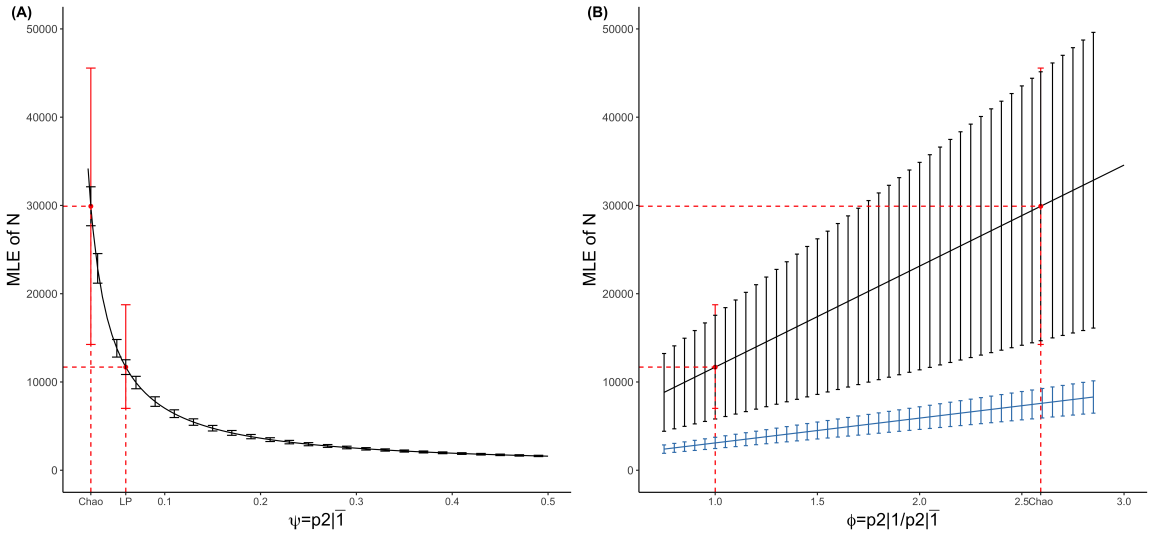


Figure 3.1: Sensitivity plots based on data from Table 3.1. The black error bars represent point-wise Wald-type 95% CIs assuming known ψ and ϕ. Red solid points and error bars mark the Lincoln–Petersen estimator and the estimator of Chao (1987) along with their 95% CIs; note that in Figure 3.1(B), $\phi = 1$ corresponds to the Lincoln–Petersen estimator. The blue line denotes the sensitivity plot based on the data where 50 patients in the n_{01} cell has moved to the n_{11} cell while the n_{10} cell remains the same. MLE=maximum likelihood estimator.

We emphasize that all MLEs on the sensitivity plots (Figure 3.1) yield the exact same maximized value of the multinomial likelihood (Lyles et al., 2021a), implying that the observed data provide no information about the dependence assumption. In fact, $\phi = 1$ is the only specific assumed value that could be potentially defended in practice, for example, if one data stream is implemented as post-enumeration random sample or otherwise as a random sample taken agnostically with respect to the other (Chao et al., 2008; Lyles et al., 2021a,b). Data-driven assumptions about ψ or ϕ obtained through metrics such as the Chao (1987) estimator or a log-linear model as applied to two streams are seldom clear to practitioners, and can also never be verified based only on the observed data (Lyles et al., 2021a). For this reason, the readily accessible sensitivity analysis embodied in Figure 3.1 provides a tool for researchers to visualize how the estimates of N respond to different unverifiable assumptions

about these key parameters. This further motivates us to propose an extension of the sensitivity analysis geared toward quantifying the uncertainty when estimating N subsequent to embedding variation about ψ or ϕ by assumption.

3.2.3 Uncertainty Analysis

As Figure 3.1 illustrates, specifying a known value for ψ leads to an exceedingly precise estimator. However, this estimator can very seldom be unlocked for direct use outside of sensitivity analysis, except under a unique study design (Lyles et al., 2021b). In the typical scenario, we propose an uncertainty analysis anchored on the intuitively accessible ratio parameter ϕ. Specifically, we encourage that epidemiologist to reflect on competing forces (e.g., temporal and geographical coverage of the two surveillance systems), in order to postulate which state of nature ($\phi > 1$ or $\phi < 1$) is more likely at the population level. Further, the analysis relies on specification of an assumed distribution (akin to a prior) for ϕ, centered at the epidemiologist's best guess and reflecting his or her level of confidence in that guess and the anticipated feasible range.

First, we reiterate that the data likelihood contains no information for updating ϕ. As such, the distribution to be postulated for ϕ is generally to be specified based on expert opinion. For the estimable parameter $p_{2|1}$, we apply a weakly informative prior to obtain its posterior distribution as in standard Bayesian analysis.

To implement the proposed uncertainty analysis, two options are available. Option (1) is to propagate the variation in ϕ along with statistical uncertainties into the appropriate level of uncertainty about ψ, and to leverage the unbiased estimator $\hat{N}_{\psi}$. Option (2) is to accommodate the variation in ϕ directly using a bias-corrected estimator. Detailed procedures for obtaining 95% percentile intervals for N under Options (1) and (2) are presented in Appendix B.4.

As illustrated previously (Lyles et al., 2021a), under the Lincoln–Petersen conditions in which $\psi = p_{2|1}$, adopting a $Beta(1, 0)$ prior for $p_{2|1}$ and replacing ψ in

Equation (3.1) by the resulting posterior mean yields the Chapman estimator. This connection motivates us to use the $Beta(1, 0)$ prior for $p_{2|1}$ when implementing Option (1). If implementing Option (2), we recommend use of the BC2 estimator or the generalized Chapman estimator, since the BC estimator is downward biased and suffers from instability when the probability associated with capture history (1,1) is small (Table 3.5).

While Option (2) can be implemented more directly, Option (1) is more easily generalizable. This is because derivations of bias-corrected estimators require extra effort and become specific to the chosen definition of the ratio parameter ϕ as the number of streams increases. For example, a total of three definitions of ϕ can be considered under three stream case (see Appendix B.7). In contrast, the direct analogue to $\hat{N}_{\psi}$ is readily available in the multiple (> 2) stream case.

We conducted uncertainty analysis using Option (1), exploring different assumed distributions for ϕ for the HIV data (Table 3.1). To promote some commonality with prior work (Chatterjee and Mukherjee, 2016), we considered three distributions for illustration: (a) Uniform(0.75, 1.25), (b) Normal(1, 0.07^2), and (c) Uniform(1, 2). The first two distributions reflect the same best guess, since they are both centered at 1; however, both reflect a lack of complete faith in that assumption. Compared to the first, the second distribution covers essentially the identical range but with smaller variation and greater confidence in the best guess of 1. The third distribution centered at 1.5 represents a case in which we assume an expert has reason to believe the two streams are positively correlated at the population level.

As shown in Figure 3.2, compared to assuming ϕ equal to a specific value, allowing variation in ϕ naturally results in a more conservative interval. Also as expected, assuming less variation (i.e., distribution (a) vs. (b)) yields a narrower interval. We obtained CIs based on the assumption of ϕ equal to a fixed value by applying the proposed uncertainty analysis for interval estimation with the ϕ distribution degenerating

to that fixed value, a practice that we study empirically in a subsequent section.

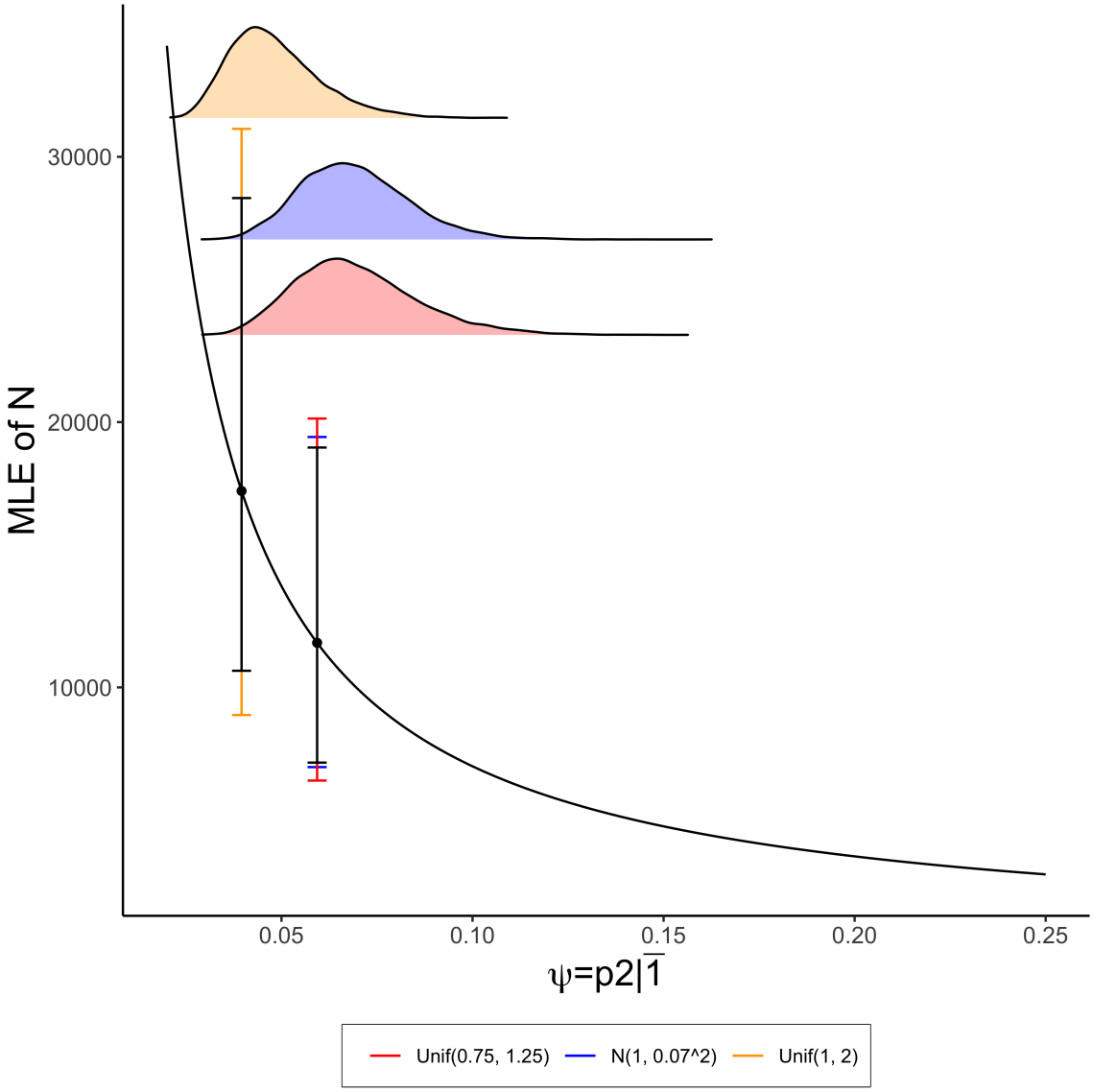


Figure 3.2: Uncertainty analysis for data from Table 3.1. The density plots in red, blue, and yellow reflect the posterior distributions of ψ when the assumed distributions of ϕ are Unif(0.75, 1.25), N(1, 0.07^2), and Unif(1, 2), respectively. The black error bars represent the 95% CIs obtained from the proposed uncertainty analysis assuming $\phi = 1$ (7,165, 19,042), and $\phi = 1.5$ (10,627, 28,446). The error bars in red, blue, and yellow denote 95% CIs obtained from the uncertainty analysis with assumed distributions of ϕ are Unif(0.75, 1.25), N(1, 0.07^2), and Unif(1, 2), respectively.

We note that the Lincoln–Petersen conditions represent a natural "middle ground" for uncertainty analysis, with the LP or Chapman estimator reported with its usual measure of statistical uncertainty representing a special case in which the assumed

distribution for ϕ degenerates to 1. This provides additional motivation to examine the proposed uncertainty analysis as a general approach for interval estimation of N. For example, when the distributional assumption for ϕ degenerates to a specific value, the 95% CI constructed based on the proposed uncertainty analysis can be viewed as an empirical approach to obtain interval estimation under the assumption that ϕ equals that value.

In practice, postulating the center of the assumed distribution for ϕ is typically a difficult task. However, expert opinion in this regard is arguably more defensible than reliance on a specific statistical model to elucidate ϕ, particularly in the case of two streams. The task may be more readily tackled within strata formed by variables deemed to be associated with likelihood of capture at the population level. Within such strata, it may be easier for the epidemiologist to postulate where the center of the ϕ distribution should be, and in particular whether it should be > 1, $= 1$, or < 1. The choice of both the center and spread of the assumed distribution can clearly be stratum specific. The proposed uncertainty analysis thus offers a principled way to account for covariates, acknowledging the fact that the true ϕ is unknown within each stratum and yielding point and interval estimates for N obtained by summing over strata.

3.2.4 Sensitivity Analysis with A Known Case Ratio

In practice, one strategy to aid with postulating the range of key parameters is to leverage external information. For example, Wolter (1990) assumed a known ratio of cases across the sexes together with the same odds ratio (i.e., θ) in the two groups to obtain estimates of N and θ. Motivated by this idea, we apply the proposed sensitivity analysis under the assumption that the case ratio is known and that key parameters ϕ or θ are the same across strata defined by a binary covariate. Specifically, when the proposed sensitivity analysis is applied within each stratum, the crossing point

of the two sensitivity plots provides a visual estimate of N and the key parameter (ϕ or θ). The corresponding derived estimators are presented in Appendix B.5 - B.6. It can be shown that these estimators coincide with the MLEs of those parameters.

To demonstrate its power as a visualization tool, we applied the sensitivity analysis anchored on ϕ and θ within each stratum to the data explored in Wolter (1990) (Table B.1). As shown in Figure 3.3, estimates of ϕ or θ under Wolter's assumptions can be directly read from the sensitivity plots and were used to estimate N (numerical values of estimated ϕ, θ, and N are given in Table B.2). We note that the estimated N here is the same regardless of whether the equivalence assumption was imposed in terms of ϕ or θ; however, this equivalence is not guaranteed in general. We also note that when the sensitivity plots for the two strata are parallel or overlap, the crossing point cannot be identified. In other words, the estimators of ϕ and θ are not well defined in such instances, adding a visual clarification of the potential instability in Wolter's proposed estimator (Mallet et al., 1994).

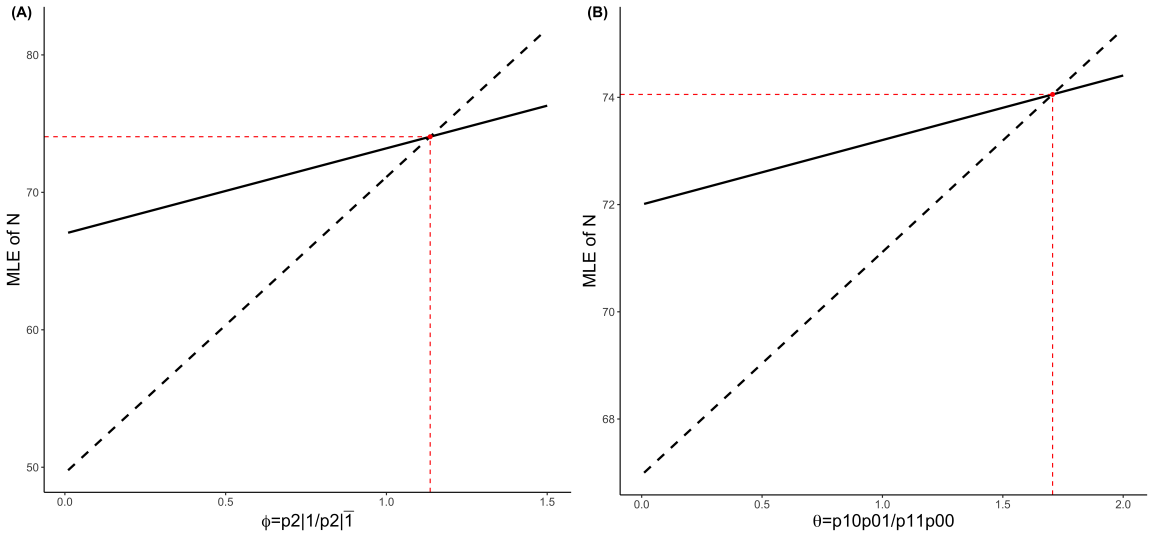


Figure 3.3: Sensitivity plot with known case ratio based on data from Table B.1 (Wolter, 1990). The red solid points mark the crossing points of sensitivity plots; the black solid lines represent sensitivity plots for females and the black dashed lines represent scaled sensitivity plots (i.e., divided by the known sex ratio $r = 1.15$) for males.

3.3 Simulation Studies

In this section, we summarize simulations designed to evaluate the performance of the proposed uncertainty analysis. The simulation study was implemented using Option (1); results obtained under different scenarios are presented in Table 3.2. For simulation without stratification, the data were generated assuming ϕ follows a Uniform(0.75, 1.25) distribution and the assumed prior distribution of ϕ is correctly specified. As suggested by the coverage summary, the proposed uncertainty analysis is a valid approach to fully and accurately quantified the uncertainties in estimating N. In contrast, failing to account for that variation by assuming a degenerate distribution for ϕ at 1 resulted in severe under-coverage. For scenarios reflecting the actual application need where uncertainties are assumed for a degenerate true ϕ, simulation results are summarized in Table 3.3. We found that incorporating uncertainties markedly improved the coverage compared to assuming ϕ degenerates to the wrong value. The performance of the proposed uncertainty analysis was also evaluated when using Option (2) under scenarios where the prior distribution of ϕ was correctly specified. Compared to Option (1), using Option (2) resulted in the reduction in median width of intervals (Table 3.4).

We further conducted a simulation study under the situation where a binary covariate is incorporated by stratification. We explored two types of mixture, the first a mixture of independence and negative association and the second a mixture of independence and positive association. Simulation results demonstrate that the proposed uncertainty analysis is easily adapted to incorporating stratification and providing excellent coverage when uncertainty in ϕ is correctly imposed (Table 3.2). Under-coverage is again demonstrated when variation in ϕ is ignored.

Simulation results for evaluating the proposed bias-corrected estimators are presented in Table 3.5. We found that the BC2 and generalized Chapman estimators are virtually unbiased, and almost identical as N increases. In cases where both

were downward biased, the BC2 estimator was slightly less biased compared to the generalized Chapman estimator and notably less biased than the BC estimator.

Interval estimation under the Lincoln–Petersen conditions based on the proposed uncertainty analysis achieved comparable performance with a previously proposed transformed logit CI (Sadinle, 2009), both in terms of coverage and median interval width (Table 3.6). The transformed logit CI was selected for comparison due to its reliable performance in coverage (Lyles et al., 2021a; Sadinle, 2009).

3.4 Discussion

We have developed a sensitivity and uncertainty analysis framework focused upon a key inestimable parameter for CRC experiments with two surveillance efforts. We have discussed three definitions based on intuitive interpretations of that inestimable parameter, i.e., ψ, ϕ, and θ. We have derived closed-form MLEs hinging on these key parameters, where the one based on a known value of $\psi = p_{2|\bar{1}}$ (Equation 3.1) was previously confirmed as unbiased (Lyles et al., 2021a). We have provided bias-corrected estimators as alternatives to the MLEs based on known ϕ and θ. The proposed sensitivity analysis can be anchored on either one of those parameters, to graphically study the impact of the key parameters on the estimation of N. With this novel data visualization tool, epidemiologists can gauge sensitivity of the estimate N to critical assumptions at the beginning of the analysis. In addition, this visualization allows the direct identification of the MLE of N under the case with a known case ratio (Wolter, 1990). Applying the sensitivity analysis to motivating HIV data (Abeni et al., 1994), we have emphasized the importance of incorporating expert opinion in terms of directionality (< 1, $= 1$, > 1) and where to center the best guess for the parameter ϕ or θ, given that the observed data carry no information about those parameters.

The proposed uncertainty analysis approach is further designed for quantifying uncertainties in the estimation of N while accounting for assumed variation in ϕ. While this aspect of our work is similar in spirit to fully Bayesian analyses previously developed (Chatterjee and Mukherjee, 2016), our uncertainty analysis is more intuitive and practical, providing general access for epidemiologic applications. Two options are available for incorporating the assumed variation of ϕ. Option (1) leverages the unbiased MLE $\hat{N}_{\psi}$, while Option (2) directly uses bias-corrected MLEs which can be used to improve the efficiency of the uncertainty analysis (Table 3.4). We recommend the use of Option (1), due to its direct extension to incorporate multiple streams. First, the MLE assuming a known value of the parameter analogous to $\psi = p_{2|\bar{1}}$ can be easily derived under the multiple stream case. Second, we find that different assumptions can be related back to a parameter analogous to ψ. For example, the inestimable parameter $p_{3|\bar{1}\bar{2}}$ is the analogue to $\psi = p_{2|\bar{1}}$ in the three-stream case; we provide the MLE for N assuming a known $p_{3|\bar{1}\bar{2}}$ (the analogue to Equation 3.1) in Appendix B.7. We aspire to fully develop the extension to multiple streams, and to use the methods provided here to motivate a transparent general modeling framework for analyzing CRC data in epidemiologic surveillance studies (discussed in Chapter 5). In this chapter, we have also illustrated that the proposed uncertainty analysis can serve as a general interval estimation approach with reliable performance under various scenarios.

Table 3.2: Simulation results for evaluating uncertainty quantification with and without stratification

Scenario	Assuming variation in ϕ		Ignoring variation in ϕ	
No stratification				
$(p_1, p_2)^a, N$	median width	coverage (%)	median width	coverage (%)
(0.1, 0.25), 1000	822	94.5	652	86.1
(0.1, 0.25), 500	534	94.7	468	91.2
(0.1, 0.5), 1000	607	94.5	374	71.8
(0.1, 0.5), 500	363	94.6	265	82.4
(0.25, 0.25), 1000	551	95.1	374	79
(0.25, 0.25), 500	337	94.8	265	86.3
(0.5, 0.5), 1000	282	94.9	124	49.4
(0.5, 0.5), 500	157	94.9	88	66.2
Stratified by a binary variable				
$(p_1, p_{2\|1})^b, (\phi_1, \phi_2)^c, N$	median width	coverage (%)	median width	coverage (%)
(0.1, 0.25), (1, 0.8), 2200	1295	95.3	1128	76.6
(0.1, 0.25), (1, 0.8), 1100	863	95.3	841	85.8
(0.1, 0.25), (1, 1.2), 2200	969	94.8	632	55.4
(0.1, 0.25), (1, 1.2), 1100	572	95.4	459	68.6
(0.1, 0.5), (1, 0.8), 2200	866	95.2	627	63.8
(0.1, 0.5), (1, 0.8), 1100	524	95.2	450	76.7
(0.1, 0.5), (1, 1.2), 2200	464	94.9	201	37.3
(0.1, 0.5), (1, 1.2), 1100	250	94.9	143	50.1
(0.25, 0.25), (1, 0.8), 2200	1303	95.2	943	83.9
(0.25, 0.25), (1, 0.8), 1100	872	95.6	703	90.1
(0.25, 0.25), (1, 1.2), 2200	975	95.1	528	65.6
(0.25, 0.25), (1, 1.2), 1100	578	94.9	382	79.5
(0.5, 0.5), (1, 0.8), 2200	875	95.0	529	72.9
(0.5, 0.5), (1, 0.8), 1100	532	95.1	380	82.9
(0.5, 0.5), (1, 1.2), 2200	471	94.8	174	43.4
(0.5, 0.5), (1, 1.2), 1100	257	94.9	124	58.1

[a] p_1 and p_2 are marginal probabilities of capture in stream 1 and 2.
[b] p_1 and $p_{2|1}$ are marginal capture probability of stream 1 and probability of being identified by stream 2 given identified by stream 1, with these values the same within each stratum.
[c] ϕ_1 and ϕ_2 denote the center of distribution of ϕ within each stratum.

Table 3.3: Simulation results for evaluating uncertainty quantification when the true $\phi = 0.9$ or 1.1

Scenario	The proposed uncertainty analysis assumes $\phi = 1$		The proposed uncertainty analysis assumes $\phi \sim$Uniform(0.8, 1.2)	
Data was generated by assuming $\phi = 0.9$				
$(p_1, p_{2\|1}, \phi)^a, N$	median width	coverage (%)	median width	coverage (%)
(0.1, 0.25, 0.9), 1000	712	92.5	840	96.7
(0.1, 0.25, 0.9), 500	513	94.1	562	96.2
(0.1, 0.5, 0.9), 1000	411	83.2	593	97.4
(0.1, 0.5, 0.9), 500	291	89.7	366	96.9
(0.25, 0.25, 0.9), 1000	408	87.6	544	97.2
(0.25, 0.25, 0.9), 500	288	91.8	343	96.5
(0.5, 0.5, 0.9), 1000	134	59.1	264	98.9
(0.5, 0.5, 0.9), 500	95.0	79.0	150	98.0
Data was generated by assuming $\phi = 1.1$				
$(p_1, p_{2\|1}, \phi)^a, N$	median width	coverage (%)	median width	coverage (%)
(0.1, 0.25, 1.1), 1000	592	89.2	696	93.6
(0.1, 0.25, 1.1), 500	424	90.7	463	92.9
(0.1, 0.5, 1.1), 1000	339	83.3	488	95.3
(0.1, 0.5, 1.1), 500	241	87.1	301	93.8
(0.25, 0.25, 1.1), 1000	342	86.2	452	94.8
(0.25, 0.25, 1.1), 500	244	89.4	286	93.9
(0.5, 0.5, 1.1), 1000	115	66.8	219	97.6
(0.5, 0.5, 1.1), 500	82	80.1	126	96.2

[a] p_1 and $p_{2|1}$ are marginal capture probability of stream 1 and probability of being identified by stream 2 given identified by stream 1, with these values the same within each stratum.

Table 3.4: Simulation results for evaluating uncertainty quantification using Option 2 with different bias-corrected estimators

Scenario	Option (1)	Bias-corrected estimators used in Option (2)					
		$\hat{N}^*_{chap}$		$\hat{N}^{BC}_{\phi}$		$\hat{N}^{BC2}_{\phi}$	
$(p_1, p_2)^a, N$	median width	coverage (%)	median width	coverage (%)	median width	coverage (%)	median width	coverage (%)
(0.1, 0.25), 1000	822	94.3	806	94.7	770	93.8	772	94
(0.1, 0.25), 500	534	94.7	520	94.3	464	92.7	470	93
(0.1, 0.5), 1000	608	94.7	600	94.9	592	94.6	593	94.5
(0.1, 0.5), 500	363	94.6	356	95.1	344	94.3	345	94.3
(0.25, 0.25), 1000	552	95.1	546	95.4	538	95	539	95
(0.25, 0.25), 500	336	94.8	332	94.9	321	94.3	322	94.3
(0.5, 0.5), 1000	282	95.1	281	95.1	281	95	281	95.1
(0.5, 0.5), 500	157	94.9	156	95	156	94.9	156	94.9

[a] p_1 and p_2 are marginal probabilities of capture in stream 1 and 2.

Table 3.5: Simulation results for evaluating bias-corrected estimators of N with a known ϕ

Scenario	Mean (SD) [a]			
$(p_1, \psi, \phi)^b, N$	$\hat{N}_\phi$	$\hat{N}_\phi^{BC}$	$\hat{N}_\phi^{BC2}$	$\hat{N}_{chap}^*$
(0.1, 0.2, 0.75), 50	-	27 (18)	33 (16)	32 (16)
(0.1, 0.3, 0.75), 50	-	31 (22)	40 (19)	39 (19)
(0.1, 0.2, 1), 50	-	30 (21)	38 (19)	37 (19)
(0.1, 0.3, 1), 50	64 (44)	34 (23)	44 (22)	43 (22)
(0.1, 0.2, 1.5), 50	64 (46)	34 (24)	44 (24)	43 (23)
(0.1, 0.3, 1.5), 50	64 (45)	39 (22)	48 (23)	47 (23)
(0.1, 0.2, 0.75), 100	134 (88)	61 (49)	85 (43)	83 (43)
(0.1, 0.3, 0.75), 100	138 (99)	69 (49)	95 (49)	92 (47)
(0.1, 0.2, 1), 100	139 (100)	66 (51)	93 (49)	90 (48)
(0.1, 0.3, 1), 100	134 (99)	77 (45)	99 (50)	97 (48)
(0.1, 0.2, 1.5), 100	134 (101)	78 (47)	99 (52)	97 (50)
(0.1, 0.3, 1.5), 100	120 (76)	90 (34)	101 (42)	100 (40)
(0.1, 0.2, 0.75), 200	284 (219)	145 (100)	198 (107)	193 (102)
(0.1, 0.3, 0.75), 200	256 (183)	171 (76)	201 (96)	199 (91)
(0.1, 0.2, 1), 200	265 (204)	163 (86)	201 (104)	198 (99)
(0.1, 0.3, 1), 200	236 (143)	187 (65)	201 (82)	200 (79)
(0.1, 0.2, 1.5), 200	235 (143)	187 (66)	201 (84)	200 (81)
(0.1, 0.3, 1.5), 200	215 (72)	197 (48)	200 (54)	200 (54)
(0.1, 0.2, 0.75), 350	437 (302)	311 (126)	352 (163)	349 (156)
(0.1, 0.3, 0.75), 350	396 (188)	338 (95)	352 (122)	351 (118)
(0.1, 0.2, 1), 350	406 (226)	332 (106)	351 (137)	350 (132)
(0.1, 0.3, 1), 350	378 (121)	347 (87)	351 (95)	351 (94)
(0.1, 0.2, 1.5), 350	377 (123)	346 (90)	350 (97)	350 (97)
(0.1, 0.3, 1.5), 350	363 (76)	349 (67)	350 (68)	350 (68)

[a] Mean and SD denote averaged bias and standard deviation of point estimates across 10,000 simulations for each estimator, respectively; $\hat{N}_\phi$ is not reported if more than 30% of simulated datasets had $n_{11} = 0$.

[b] $p_1, \psi = p_{2|\bar{1}}$, and $\phi = p_{2|1}/p_{2|\bar{1}}$ are probabilities used for generating data based on multinomial distribution.

Table 3.6: Simulation results for evaluating interval estimation under Lincoln-Petersen conditions

Scenario	Uncertainty analysis			Transformed logit		
$(p_1, p_2)^a, N$	median width	coverage (%)	(% missed high, % missed low)	median width	coverage (%)	(% missed high, % missed low)
(0.05, 0.05), 10000	7592	95.0	(1.91, 3.09)	7499	95.3	(1.36, 3.38)
(0.05, 0.05), 5000	5506	95.1	(1.65, 3.25)	5382	95.6	(0.90, 3.49)
(0.05, 0.05), 2500	4042	95.2	(1.12, 3.65)	3869	95.9	(0.11, 4.03)
(0.05, 0.1), 10000	5161	95.1	(2.00, 2.86)	5146	95.2	(1.81, 3.00)
(0.05, 0.1), 5000	3703	94.9	(1.79, 3.28)	3678	95.2	(1.48, 3.31)
(0.05, 0.1), 2500	2638	95.2	(1.52, 3.25)	2606	95.7	(0.90, 3.38)
(0.05, 0.2), 10000	3430	94.7	(2.25, 3.03)	3434	94.8	(2.23, 3.00)
(0.05, 0.2), 5000	2439	94.8	(2.27, 2.94)	2437	95.1	(2.11, 2.80)
(0.05, 0.2), 2500	1744	95.0	(1.76, 3.26)	1742	95.3	(1.61, 3.13)
(0.1, 0.1), 10000	3534	95.2	(1.99, 2.83)	3535	95.1	(1.93, 2.93)
(0.1, 0.1), 5000	2519	94.8	(2.22, 3.02)	2514	95.1	(1.99, 2.91)
(0.1, 0.1), 2500	1797	95.1	(1.83, 3.08)	1788	95.4	(1.58, 2.98)
(0.1, 0.2), 10000	2356	95.0	(2.15, 2.88)	2358	95.1	(2.03, 2.89)
(0.1, 0.2), 5000	1669	94.8	(2.31, 2.85)	1669	95.1	(2.24, 2.68)
(0.1, 0.2), 2500	1190	94.9	(2.19, 2.95)	1190	95.0	(2.12, 2.84)

[a] p_1 and p_2 are marginal probabilities of capture in stream 1 and 2.

Chapter 4

Pitfalls of the Log-linear Modeling Framework for CRC Studies

4.1 Background

Log-linear models are widely adopted CRC methods to estimate disease risk or prevalence in human populations, which assumes independent Poisson distributions for the observed cell counts (Fienberg, 1972; Cormack, 1989). To estimate the total number of cases, the log-linear model regresses observed cell counts against indicators of the capture history and their interactions. The popularity of log-linear models is partially due to their allowance for dependency between systems via interaction terms. In addition, the accessibility of log-linear models also greatly promotes their applications. For example, the `GENMOD` procedure in `SAS` and the `glm` function in `R` can be directly used for fitting log-linear models to CRC data (Institute, 1985; R Core Team, 2023).

The application of log-linear models has been extended to incorporate covariates. Discrete covariates can be incorporated by stratification of the cell counts, or by directly including the covariates and their interactions with indicators of the capture history in the log-linear model (Fienberg, 1972; Hook and Regal, 1995). The log-linear model is a discrete cell-count model, and hence can only incorporate continuous covariates by first categorizing them. However, logistic models allow one to include both individual-level discrete and continuous covariates to model the capture probability of each data stream by assuming independence conditional on covariates (Huggins, 1989; Alho, 1990). Further extensions of this model permit consideration of the dependency structure between data streams while also modeling the individual-level capture probabilities using the logistic model (Zwane and van der Heijden, 2005). In this model, the dependency between data streams is induced by allowing capture probabilities to depend on covariates and responses to data streams.

In epidemiological studies, practitioners usually undertake the fitting of all possible log-linear models after excluding the highest-order interaction term. This exclusion is by common convention as a means of ensuring identifiability, since the number of cases not captured by any system is unobserved. Subsequently, the estimated

N and associated 95% CI are typically reported based on the model identified by a model selection metric, e.g., Akaike's information criterion (AIC), likelihood ratio test statistics, or the Bayesian Information Criterion (BIC) (Akaike, 1974; Schwarz, 1978; Hook and Regal, 1997; Héraud-Bousquet et al., 2012; Barocas et al., 2018). Generally, the log-linear model with the lowest AIC will be selected as the "best" model to obtain the estimate of N (Hook and Regal, 1997; Héraud-Bousquet et al., 2012). However, it has been noted in various applied studies that different log-linear models can result in widely different estimated N (Poorolajal et al., 2017; Ramos et al., 2020; Zhang and Small, 2020). In other words, the estimation of N can be very sensitive to model assumptions that cannot be verified using only the observed data (Fienberg, 1972; Hook and Regal, 1997; Lyles et al., 2021a; Zhang et al., 2022). However, those commonly applying log-linear models for CRC-based estimation in current practice are often unaware of the extent to which regularly-used model selection metrics based on the observed data are inadequate.

In this chapter, we illustrate two major pitfalls associated with the log-linear model paradigm for CRC, primarily stemming from the fact that it dissuades users from carefully considering or attempting to vet the assumptions (e.g., regarding population-level dependencies among data streams) required to enable the estimation of N. First, we show that the log-linear model framework is highly "exclusionary". Specifically, the log-linear model generally excludes many possible estimates of N by design (as illustrated in Section 4.4). Second, we illustrate how model selection metrics can be deceptive and inadequate tools for CRC model selection. We clarify how such metrics choose models based on certain observable constraints in the data, despite the fact that these constraints actually provide no information about true non-identifiable dependency parameters characterizing the data streams that are central to the valid estimation of N. The goal is to caution that application of the log-linear model framework to analyze CRC data requires a clear understanding of these pitfalls, with

careful consideration of what key dependency assumptions are imposed (sometimes implicitly) by the adopted model.

4.2 Motivating Data

We used two publicly available epidemiological CRC datasets for quantifying HIV infections, together with one dataset simulated in a setting where referrals between data streams occur (Jones et al., 2014). The data demonstrate how the log-linear model framework excludes many possible estimates of N by design. The two real data examples involve three and four data streams, respectively, and both were analyzed previously using the log-linear model (Abeni et al., 1994; Poorolajal et al., 2017). Specifically, the three-stream HIV CRC data were collected in Iran in 2016 and the four-stream HIV data were collected in Lazio, Italy during 1990.

4.3 MLEs of N with a Given Key Dependency Parameter

We extend the population-level multinomial model and the parametrization adopted in Chapter 3 to incorporate multiple (>) data streams (Darroch, 1958; Zhang et al., 2022). We have discussed the connection between the population-level multinomial model and the individual-level multinomial model in Section 1.2.2. We assume a closed population with N diseased cases, with $K \geq 2$ surveillance streams implemented. Under this situation, a total of 2^K unique capture histories exist. Let the set $\mathcal{O} = \{(11\ldots1), (101\ldots1), \ldots, (0\ldots0)\}$ containing 2^K sequences denote the collection of all possible capture histories, where capture histories are arranged in lexicographic order and 1 denotes captured and 0 not captured. Let h_i denote the i-th sequence in the set $\mathcal{O}$ with $i = 1, \ldots, 2^K$, N_{h_i} and n_{h_i} denote the true and observed

number of cases having capture history h_i. Because the number of cases not captured by any stream is unobserved, $n_{0\ldots0}$ is missing.

In Chapter 3, we used the parameter vector $(p_1, p_{2|1}, p_{2|\bar{1}})$ to model the 3 observed cell (Table 1.2) under the population-level multinomial model while allowing an arbitrary level of dependency between two streams (Lyles et al., 2021a). For a given value of the inestimable parameter $\psi = p_{2|\bar{1}}$, the closed form MLE of N is shown in Equation (3.1). When using the ratio parameter $\phi = p_{2|1}/p_{2|\bar{1}}$ to measure the dependency between two streams, the corresponding MLE of N with known ϕ is provided in Equation (3.3). We emphasize again that any valid value of ψ or ϕ yields exactly the same maximized log-likelihood value.

For the case of $K > 2$ streams, the extended parameter vector containing conditional probabilities $(p_1, p_{2|1}, p_{2|\bar{1}}, p_{3|12}, p_{3|\bar{1}2}, p_{3|1\bar{2}}, \ldots, p_{K|\bar{1}\bar{2}\ldots\bar{K-1}})$ models the observed $2^K - 1$ cell counts, where p_1 denotes the marginal probability of identification by the first stream and the remaining parameters are population-level conditional probabilities of identification in the k-th stream given all other capture histories arranged in lexicographic order ($\bar{\cdot}$ indicates no, otherwise yes). For example, $p_{3|1\bar{2}}$ represents the proportion captured by the third stream given captured by the first stream but not captured by the second. As noted above, in the two-stream case we refer to $p_{2|\bar{1}}$ as the key inestimable parameter and use ψ to denote it. Similarly, we denote $p_{K|\bar{1}\bar{2}\ldots\bar{K-1}}$ as ψ and treat it as the key inestimable parameter when $K > 2$.

Given $K > 2$ streams, we derive the MLE of N for a known value of ψ and its variance estimator under the population-level conditional multinomial model as follows:

$$\hat{N}_\psi = (n_c - n_{00\ldots1}) + \frac{n_{00\ldots1}}{\psi}, \tag{4.1}$$

$$\hat{Var}(\hat{N}_\psi) = \frac{1-\psi}{\psi^2} n_{00\ldots1}, \tag{4.2}$$

where n_c is the observed number of cases caught at least once, and $n_{00\ldots1}$ is the observed number of cases caught by the last stream but not by any other stream. This MLE and its variance estimator under the three-stream case have been proposed previously (see Appendix B.7), but Equations (4.1) - (4.2) extend the result to incorporate an arbitrary number of streams. As in the two-stream case, we emphasize that $\hat{N}_\psi$ is equally consistent with the observed cell counts for any valid value of ψ (i.e., $0 < \psi \leq 1$) that one supplies. That is, the parameter ψ is non-identifiable based on the observed data alone.

4.4 The Exclusionary Property of CRC Log-linear Models

Unlike the population-level multinomial model adopted in Chapter 3, the log-linear model assumes that cell counts are independent Poisson variables conditional on model coefficients that determine their corresponding means. For the simplest CRC setting with two-stream data and no covariates, the corresponding fully specified log-linear model is:

$$\log\left[E\left(N_{h_i}\right)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \delta X_1 X_2, \tag{4.3}$$

where $X_k = 1$ if cases are identified by the k-th stream under capture history h_i, and 0 otherwise, for $k = 1, 2$. The estimated N is then computed as $n_c + \exp\hat{\alpha}$, where $\hat{\alpha}$ is the MLE of the intercept. With 3 observed cell counts, it is impossible to estimate all 4 model coefficients. In the above example, the standard CRC log-linear modeling convention of dropping the K-way interaction term implies setting $\delta = 0$. In addition, a hierarchy principle is also typically applied, whereby any log-linear model considered must include all possible lower level interaction terms before the inclusion

of any higher level interaction term. When $K > 2$, a valid candidate model also must include all first-order terms. Applying the first convention for the basic two-stream case, only 3 models would therefore be valid candidates; however, note that there are a total of 7 possible models that can be fitted if these conventions are not applied. Each of these has been closely examined by prior authors, who showed that the full set of 7 possible models allows for at most 4 unique estimates of the ratio parameter $\phi = p_{2|1}/p_{2|\bar{1}}$ (see Table 2 in Lyles et al. (2021a)). Yet, as noted earlier, there are in fact an infinite number of values of the key parameter (ψ or ϕ) that yield unique MLEs of N and are equally consistent with the observed data. This simple special case exhibits the exclusionary nature of CRC log-linear models, in the sense that they exclude many possible estimates by design. As we will see, this important pitfall is not avoided simply by including additional streams and/or covariates.

4.4.1 A Toy Example

To illustrate the exclusionary nature of the log-linear model, we apply the novel visualization tool proposed in Chapter 3 to the toy set of cell counts presented in Table 4.1. In this example, we deliberately set n_{11} equal to n_{01} in order to demonstrate other pitfalls discussed in later sections.

Table 4.1: Toy two-stream CRC data

	Captured in stream 2		
Captured in stream 1	Yes	No	
Yes	$n_{11} = 250$	$n_{10} = 500$	$n_{1\cdot} = 750$
No	$n_{01} = 250$	$n_{00} =?$	
	$n_{\cdot 1} = 500$		$N =?$

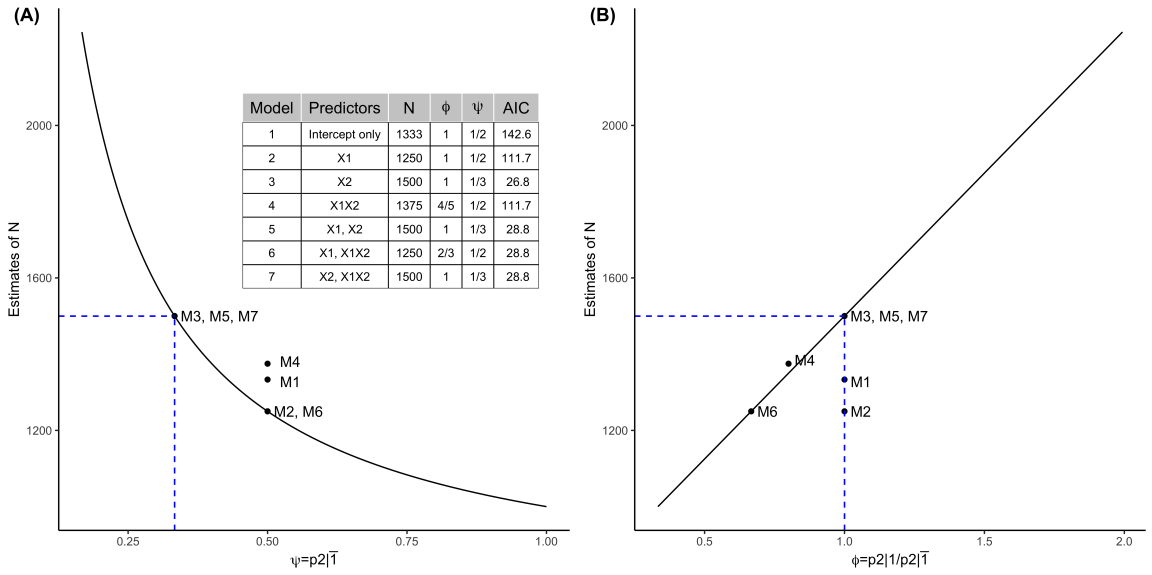
† $n_{1\cdot} = n_{11} + n_{10}$ and $n_{\cdot 1} = n_{11} + n_{10}$.

Figure 4.1 shows estimates from all 7 possible log-linear models as well as MLEs of N in Equations (3.1) and (3.3) with various assumed ψ and ϕ values based on the example data. The figure demonstrates that the log-linear model framework excludes

large swaths of possible estimates of N. In this example, considering all possible log-linear models only results in 4 unique estimates of N, despite the fact that a continuum of estimated N values is consistent with the observed data. As shown in Figure 4.1(B), the 7 possible log-linear models only permit ϕ to take estimated values of 2/3, 4/5, and 1. Yet, any value of ϕ within the range $[1/3, \infty)$ would result in an equally valid estimate of N based on the observed data. The lower bound of 1/3 is derived based on the natural lower bound of N, i.e., N is no smaller than n_c the number of cases captured at least once.

It is also noteworthy that, in this example, none of the possible log-linear models allows positive dependency between data streams (i.e., $\phi > 1$). However, there is no evidence in the observed data to exclude this possibility, in light of the unobserved cell count n_{00}. Moreover, two positively correlated streams are often seen in practice (Hook and Regal, 2000). For instance, both streams may tend to identify cases with similar characteristics.

Although it has been shown that MLEs of N derived using the population-level multinomial model are equivalent to estimates yielded by log-linear models when equivalent assumptions are imposed (Cormack and Jupp, 1991), we observe that there are estimates from unsaturated (i.e., the number of model coefficients is less than the number of observed capture histories) log-linear models (e.g., M1, M4) that do not appear on the curve continuum in Figure 4.1. This is because unsaturated log-linear models also impose assumptions on observed cell counts, while the MLEs of N assuming known ψ or ϕ do not since the observed cell counts were used. That is, when the log-linear model imposes constraints on observed cell counts, the two models indeed impose different sets of assumptions. For example, model 1 constrains the three fitted cell counts to be equal (i.e., $p_{11} = p_{10} = p_{01}$), so that the fitted counts differ from the observed in this example. Since fitted cell counts always coincide with observed cell counts when fitting saturated log-linear models, estimates from those



Model	Predictors	N	ϕ	ψ	AIC
1	Intercept only	1333	1	1/2	142.6
2	X1	1250	1	1/2	111.7
3	X2	1500	1	1/3	26.8
4	X1X2	1375	4/5	1/2	111.7
5	X1, X2	1500	1	1/3	28.8
6	X1, X1X2	1250	2/3	1/2	28.8
7	X2, X1X2	1500	1	1/3	28.8

Figure 4.1: Estimates of N from all possible log-linear models based on two-stream toy example CRC data presented in Table 4.1. Black solid line denotes the MLE of N in Equation (3.1) with assumed $\psi = p_{2|\bar{1}}$ in panel (A); black solid line denotes the MLE of N in Equation (3.3) in panel (B); black solid points represent estimates from all possible 7 log-linear models; blue dashed lines mark the estimates from the log-linear model with the lowest AIC (i.e., Model 3).

models are consistent with the MLEs in Equations (3.1) and (3.3) when fixing ψ or ϕ at the corresponding log-linear model-based estimates. In this example where $n_{11} = n_{01}$, the unsaturated model 3 happens to result in an estimated N equal to an MLE based on Equation (3.3), as shown in Figure 4.1(B). However, when n_{11} is not equal to n_{01}, estimates from model 3 will deviate from the MLE. The estimate from the unsaturated model 2 is also on the curve because it forces the sum of observed n_{11} and n_{10} is equal to the sum of fitted n_{11} and n_{10}.

Importantly, the log-linear model framework continues to exclude many valid estimates even though the number of possible log-linear models increases exponentially with the number of data streams. For the three-stream case, the fully specified log-

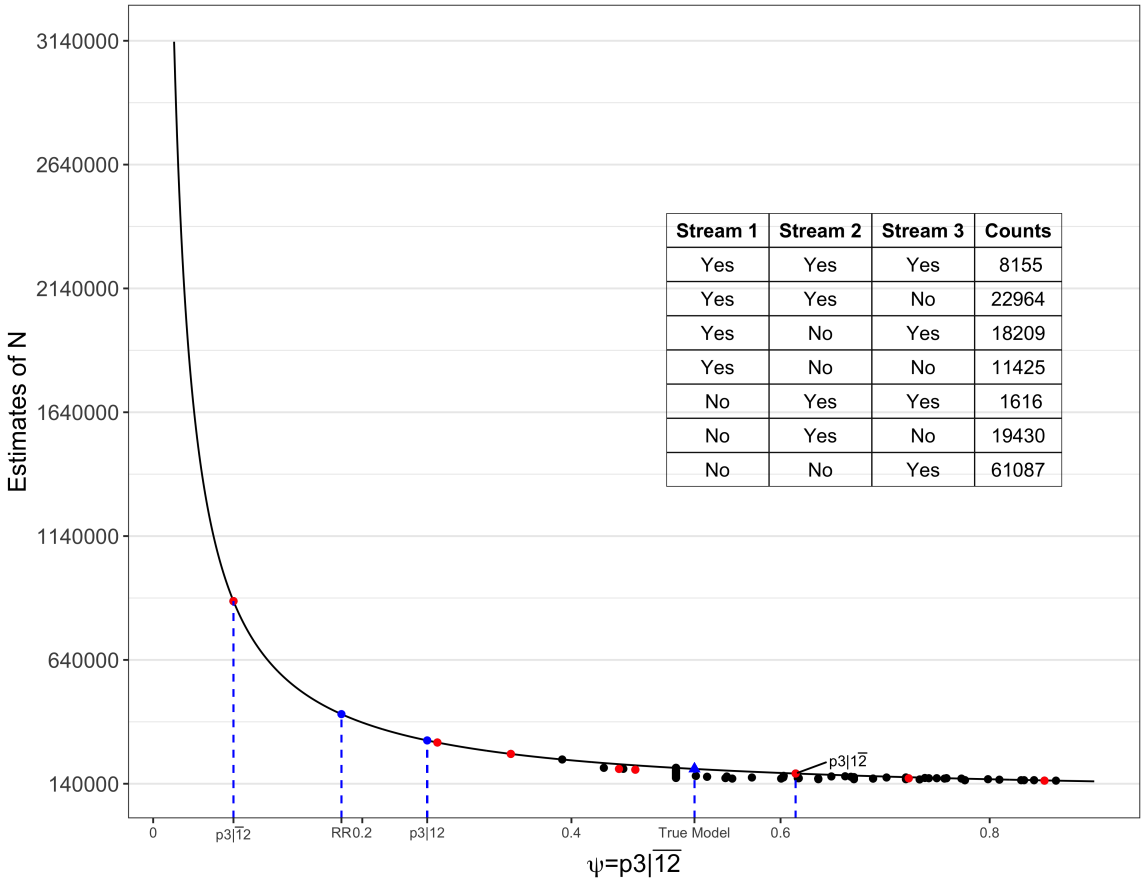
linear model is given by

$$\log\left[E\left(N_{h_i}\right)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 X_1 X_2 + \gamma_2 X_1 X_3 + \gamma_3 X_2 X_3 + \delta X_1 X_2 X_3. \tag{4.4}$$

With 7 observed cell counts, a total of 127 log-linear models are possible, of which only 8 are valid candidates when applying both standard conventions discussed previously. While model (4.4) allows both positive and negative associations between data streams by including pairwise interaction terms, studies have pointed out that the log-linear model framework is incapable of incorporating certain dependency scenarios (Jones et al., 2014).

To expand on this point, Figure 4.2 displays a variety of estimated N values obtained under different assumptions using log-linear models and the MLE in Equation (4.1), based on simulated three-stream data analyzed in Jones et al. (2014). These data were simulated from a population-level multinomial distribution by assuming $N = 200,000$, the first stream (S1) and the third stream (S3) are independent conditional on the second stream (S2), and that 20% of cases captured by S1 are referred to S3. In Figure 4.2, while we cut ψ at the values of 0.02 and 0.9, it is possible that ψ can go beyond 0.9 (in which case the estimated N would remain flat) or below 0.02 (in which case the estimated N could theoretically blow up). It is clear that estimates of N from log-linear models only occupy a small part of the curve. For example, no log-linear model places the estimated N within the range 310,000 to 860,000, despite the fact that all values in that range (and all other values depicted on the curve) are equally consistent with the observed data. A key point here (explored further below) is that the log-linear modeling framework leads the user to believe that such exclusions are based on relevant information available in the data, but this is not the case. In addition, as prior authors (Jones et al., 2014) concluded, no log-linear model captures the true associations among the three streams under which these data were

generated.



Stream 1	Stream 2	Stream 3	Counts
Yes	Yes	Yes	8155
Yes	Yes	No	22964
Yes	No	Yes	18209
Yes	No	No	11425
No	Yes	Yes	1616
No	Yes	No	19430
No	No	Yes	61087

Figure 4.2: Estimates of N from all possible log-linear models based on simulated data from the last column of Table 4 of Jones et al. (2014). Black solid line denotes the MLE of N in Equation (4.1) as assumed $\psi = p_{3|\bar{1}\bar{2}}$ varies; red solid points denote estimates from the 8 possible log-linear models when imposing the usual conventions (i.e., no 3-way interaction and following the hierarchy principle); the union of red and black solid points denote estimates from all 127 possible log-linear models; blue solid points/dashed lines denote MLEs in Equation (4.1) under four different assumptions. On x-axis, RR denotes the assumption $\psi = \frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$, $p_{3|12}$ denotes the assumption $\psi = p_{3|12}$. The text $p_{3|1\bar{2}}$ denotes the assumption $\psi = p_{3|1\bar{2}}$, and the text $p_{3|\bar{1}2}$ denotes the assumption $\psi = p_{3|\bar{1}2}$. The blue triangle/dashed line denotes the estimated N by imposing correct assumptions, which assume S1 and S3 are independent conditional on S2 and 20% referral of individuals from S1 to S3.

We also explored actual three-stream CRC data analyzed in a previous study (Poorolajal et al., 2017). In Figure 4.3, we observe for example that there is no log-linear model implying the potentially reasonable assumption that $psi = \frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$,

i.e., the population-level association between S2 and S3 is the same regardless of identification (or not) by S1. Similarly, no model reflects the assumption $\psi = p_{3|12}$, i.e., the proportion captured by S3 among those not captured by S1 and S2 is equal to the proportion captured by S3 among those captured by both S1 and S2. As shown in Table 4.3, the 8 possible models only yield at most 8 unique estimates of N when applying the usual CRC conventions. Again, however, a continuum of estimates is consistent with the observed data and a significant swath of these estimates are unachievable when fitting all possible log-linear models.

We note that the number of possible log-linear models increases exponentially even when the standard conventions are applied. With four data streams, there are 113 possible log-linear models under the standard conventions (Hook and Regal, 1995), while there are a total of 32,767 models if we consider all possible combinations of predictors. However, with this many possible models, the log-linear model framework still excludes many possible estimates as illustrated using data from Abeni et al. (1994) (Figure 4.4). For example, no log-linear model projects $\psi = p_{4|\bar{1}\bar{2}\bar{3}}$ into the range from 0.06 to 0.17. However, this range in fact contains many plausible assumptions, including for example that the key parameter ψ is equal to $p_{4|\bar{1}23}$ or $p_{4|1\bar{2}3}$.

For cases involving more than two streams (Figures 4.2 - 4.4), note again that not all estimates from log-linear models are on the black curve. As discussed in the two-stream case, log-linear models that impose the same assumption about the key parameter ψ do not necessarily impose the same assumptions about observed cell counts due to constraints built into unsaturated models.

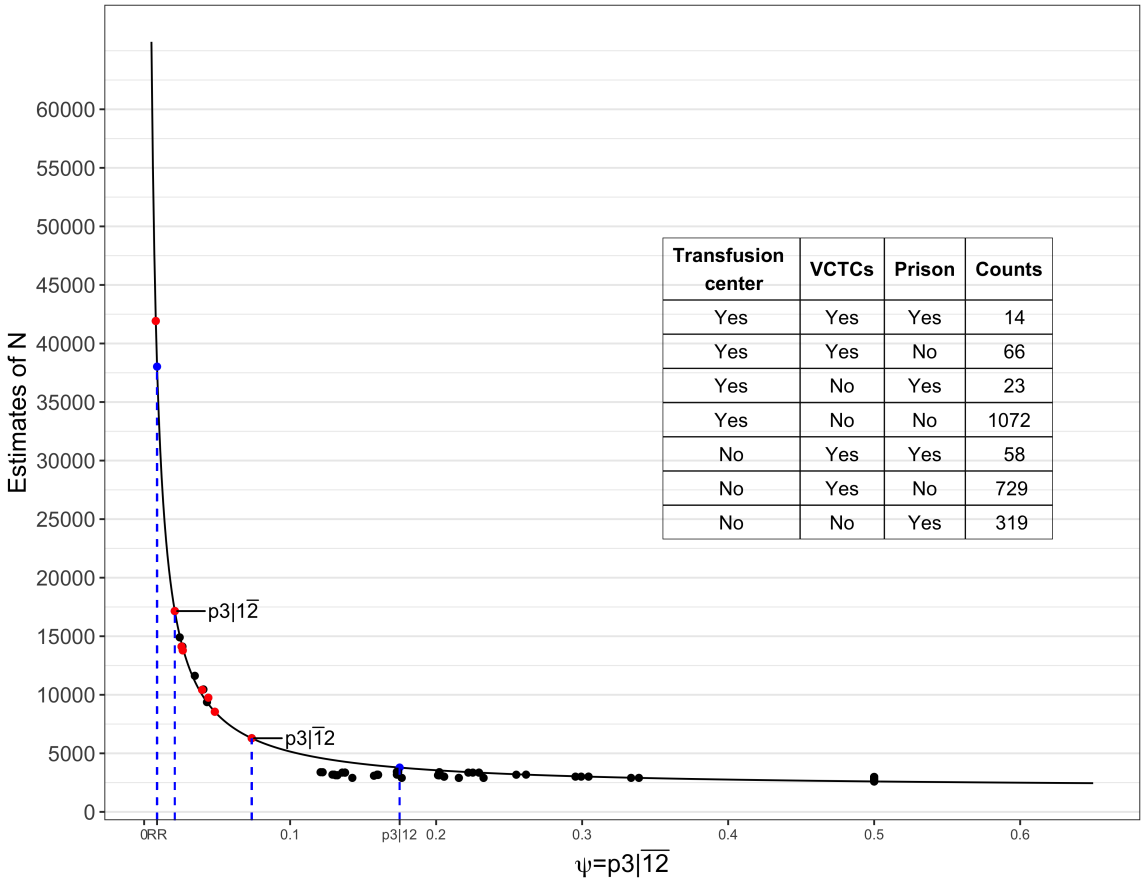
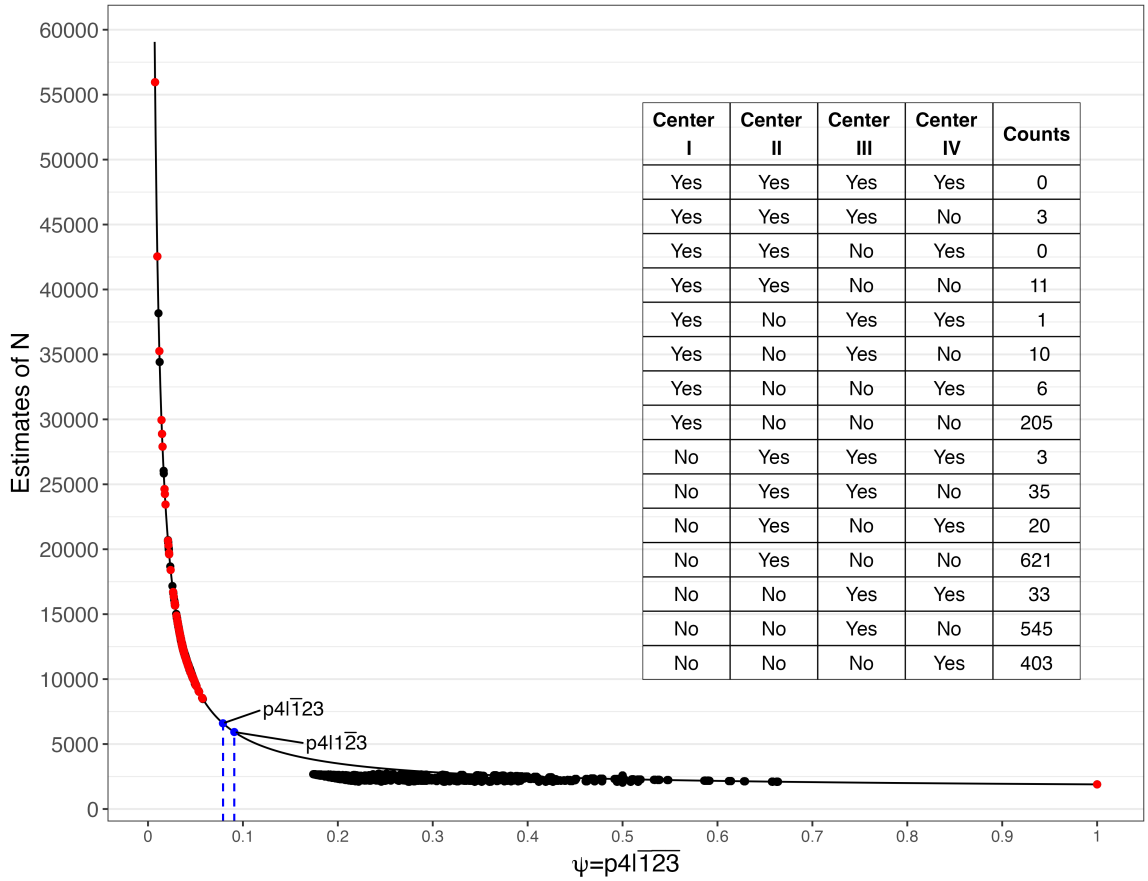


Figure 4.3: Estimates of the total HIV case count from all possible log-linear models based on data from Figure 1 of Poorolajal et al. (2017), with "Transfusion center", "VCTCs", and "Prison" comprising Streams 1, 2, and 3, respectively. Black solid line denotes the MLE of N in Equation (4.1) as assumed $\psi = p_{3|\bar{1}\bar{2}}$ varies; red solid points denote estimates from the 8 possible log-linear models when imposing the usual conventions (i.e., no 3-way interaction and following the hierarchy principle); the union of red and black solid points denote estimates from all 127 possible log-linear models; blue solid points/dashed lines denote MLEs in Equation (4.1) under four different assumptions. On x-axis, RR denotes the assumption $\psi = \frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$, $p_{3|12}$ denotes the assumption $\psi = p_{3|12}$. The text $p_{3|1\bar{2}}$ denotes the assumption $\psi = p_{3|1\bar{2}}$, and the text $p_{3|\bar{1}2}$ denotes the assumption $\psi = p_{3|\bar{1}2}$.



Center I	Center II	Center III	Center IV	Counts
Yes	Yes	Yes	Yes	0
Yes	Yes	Yes	No	3
Yes	Yes	No	Yes	0
Yes	Yes	No	No	11
Yes	No	Yes	Yes	1
Yes	No	Yes	No	10
Yes	No	No	Yes	6
Yes	No	No	No	205
No	Yes	Yes	Yes	3
No	Yes	Yes	No	35
No	Yes	No	Yes	20
No	Yes	No	No	621
No	No	Yes	Yes	33
No	No	Yes	No	545
No	No	No	Yes	403

Figure 4.4: Estimates of the total number HIV cases from all possible log-linear models based on four-stream CRC data from Table 2 of Abeni et al. (1994), with "Center I", "Center II" and "Center III", and "Center IV" comprising Streams 1, 2, 3, and 4, respectively. Black solid line denotes the MLE of N in Equation (4.1) as assumed $\psi = p_{4|\bar{1}\bar{2}\bar{3}}$ varies; red solid points denote estimates from the 113 possible log-linear models when imposing the usual conventions (i.e., no 4-way interaction and following the hierarchy principle); the union of red and black solid points denote estimates from all possible 32,767 log-linear models. The text $p_{4|\bar{1}23}$ denotes the assumption $\psi = p_{4|\bar{1}23}$, and the text $p_{4|1\bar{2}3}$ denotes the assumption $\psi = p_{4|1\bar{2}3}$.

4.5 AIC is Deceiving as a Metric for CRC Model Selection

To settle upon a final model for estimating N under the log-linear model framework, practitioners often rely on common metrics such as AIC, BIC, or likelihood ratio test statistics. However, numerous sources have raised concerns about the adequacy of such metrics in CRC modeling for epidemiologic surveillance (Fienberg, 1972; Hook and Regal, 1997; Coull and Agresti, 1999; Lyles et al., 2021a).

The main issue associated with typical model selection metrics in CRC settings is that they attempt to identify a "best-fitting" model based solely on the observed data, which in themselves contain no information about any key dependency parameter upon which assuredly valid estimation and inference could be based. That is, certain untestable assumptions (e.g., regarding the dependency structure between streams) are necessary to identify an estimate of N, but these metrics are incapable of assessing untestable assumptions. Nonetheless, a metric like AIC will happily select a log-linear model on the basis of certain testable assumptions. The problem is that such a model then projects an assumption about a non-identifiable dependency parameter that fundamentally determines the estimate of N. Unfortunately, the unsuspecting user is typically unaware of the exclusionary property, or the fact that this projection is little more than a mathematical construct. Indeed, the log-linear model that fits the observed data best might project an assumption about dependencies among data streams which differs greatly from the underlying truth. That underlying truth could be much more consistent with the estimated N from a poorly fitting log-linear model, or, as we have seen, might not be attainable by any log-linear model.

Table 4.2 presents estimates along with the AIC for each model fitted to the toy example data in Table 4.1. Note that model 3 yields the lowest AIC; this occurs because $n_{11} = n_{01}$ in the observed data. The corresponding mathematical construct,

for which there is no basis, then projects the unobserved cell count n_{00} to be the same as the observed cell count n_{10}. The key point is that the testable constraint $E(N_{11}) = E(N_{01})$ producing the selected model can be defended based on the data. However, the assumption $E(N_{10}) = E(N_{00})$ is untestable in the observed data. Suppose the true value of ϕ were 1.5, which indicates $E(N_{00}) = 875$ and $E(N) = 1,875$. This truth differs greatly from the model-projected estimate of $E(N_{00})$, which equals 500, from applying model 3.

Importantly, the problem associated with relying on testable constraints to project the unobserved cell count remains even when the number of streams increases. Using model 2 in Table 4.3 as an example, the estimated unobserved cell count N_{000} takes its form only because the testable constraints $E(N_{111})/E(N_{110}) = E(N_{101})/E(N_{100}) = E(N_{011})/E(N_{010})$ are assumed under model 2. The computation of N_{000} is a purely mathematical construct, and the untestable assumption used for inferring N_{000} cannot be justified using the observed data. Yet, a practitioner selecting model 2 based on a metric like AIC would be led to believe otherwise.

Table 4.2: Possible log-linear models for two-stream toy example data in Table 4.1

		Fitted cell counts [b]				MLE of key parameters [c]		Results from the example data presented in Table 4.1			
Model [a]	Predictors	$\hat{N}_{11}$	$\hat{N}_{10}$	$\hat{N}_{10}$	$\hat{N}_{00}$	$\hat{\psi}$	$\hat{\phi}$	$\hat{\psi}$	$\hat{\phi}$	$\hat{N}$	AIC
1	Intercept only	$n_c/3$	$n_c/3$	$n_c/3$	$n_c/3$	$1/2$	1	$1/2$	1	1333	142.6
2	X_1	$n_{1\cdot}/2$	$n_{1\cdot}/2$	n_{01}	n_{01}	$1/2$	1	$1/2$	1	1250	111.7
3	X_2	$n_{\cdot 1}/2$	n_{10}	$n_{\cdot 1}/2$	n_{10}	$\frac{n_{\cdot 1}}{2n_{10}+n_{\cdot 1}}$	1	$1/3$	1	1500	26.8
4	X_1X_2	n_{11}	$\frac{n_{10}+n_{01}}{2}$	$\frac{n_{10}+n_{01}}{2}$	$\frac{n_{10}+n_{01}}{2}$	$1/2$	$\frac{4n_{11}}{2n_{11}+n_{10}+n_{01}}$	$1/2$	0.8	1375	111.7
5	X_1, X_2	n_{11}	n_{10}	n_{01}	$n_{10}n_{01}/n_{11}$	$n_{11}/(n_{11}+n_{10})$	1	$1/3$	1	1500	28.8
6	X_1, X_1X_2	n_{11}	n_{10}	n_{01}	n_{01}	$1/2$	$2n_{11}/n_{1\cdot}$	$1/2$	$2/3$	1250	28.8
7	X_2, X_1X_2	n_{11}	n_{10}	n_{01}	n_{10}	$\frac{n_{01}}{n_{01}+n_{10}}$	$\frac{n_{11}(n_{10}+n_{01})}{n_{01}n_{1\cdot}}$	$1/3$	1	1500	28.8

[a] The intercept (α) is included in all models.
[b] $n_c = n_{11}+n_{10}+n_{01}$, $n_{1\cdot} = n_{11}+n_{10}$, $n_{\cdot 1} = n_{11}+n_{01}$; analytic results in columns 3 through 8 reproduced from Lyles et al. (2021a).
[c] $\hat{\psi} = \hat{N}_{10}/(\hat{N}_{10}+\hat{N}_{00})$ and $\hat{\phi} = \hat{N}_{11}(\hat{N}_{01}+\hat{N}_{00})/(\hat{N}_{11}+\hat{N}_{10})\hat{N}_{01}$.
[d] $\hat{N} = n_c + \exp\hat{\alpha}$, where $\hat{\alpha}$ is the estimated intercept from fitting the log-linear model based on the toy example data presented in Table 4.1.

Table 4.3: Possible log-linear models for three-stream CRC data when applying the usual conventions

Model [a]	Predictors	MLE of the key parameter $\psi = p_{3\|\bar{1}\bar{2}}$ [b]	Fitted N_{000} [c]
1	X_1, X_2, X_3	$\frac{\left(\hat{N}_{111}\right)^{3/8}\left(\hat{N}_{101}\right)^{1/4}\left(\hat{N}_{011}\right)^{1/4}\left(\hat{N}_{001}\right)^{1/8}}{\left(\hat{N}_{111}\right)^{3/8}\left(\hat{N}_{101}\right)^{1/4}\left(\hat{N}_{011}\right)^{1/4}\left(\hat{N}_{001}\right)^{1/8}+\left(\hat{N_{110}}\right)^{1/4}\left(\hat{N}_{100}\right)^{3/8}\left(\hat{N}_{010}\right)^{3/8}}$	$\frac{\left(\hat{N}_{100}\right)^{1/2}\left(\hat{N}_{010}\right)^{1/2}\left(\hat{N}_{001}\right)^{1/2}}{\left(\hat{N}_{111}\right)^{1/2}}$
2	X_1, X_2, X_3 X_1X_2	$\frac{\left(\hat{N}_{111}\right)^{1/3}\left(\hat{N}_{101}\right)^{1/3}\left(\hat{N}_{011}\right)^{1/3}}{\left(\hat{N}_{111}\right)^{1/3}\left(\hat{N}_{101}\right)^{1/3}\left(\hat{N}_{011}\right)^{1/3}+\left(\hat{N}_{110}\right)^{1/3}\left(\hat{N}_{100}\right)^{1/3}\left(\hat{N}_{010}\right)^{1/3}}$	$\frac{\left(\hat{N}_{110}\right)^{1/3}\left(\hat{N}_{100}\right)^{1/3}\left(\hat{N}_{010}\right)^{1/3}}{\left(\hat{N}_{111}\right)^{1/3}\left(\hat{N}_{101}\right)^{1/3}\left(\hat{N}_{011}\right)^{1/3}}\hat{N}_{001}$
3	X_1, X_2, X_3 X_1X_3	$\frac{\left(\hat{N}_{111}\right)^{1/6}\left(\hat{N}_{110}\right)^{1/6}\left(\hat{N}_{011}\right)^{2/3}\left(\hat{N}_{001}\right)^{1/3}}{\left(\hat{N}_{111}\right)^{1/6}\left(\hat{N}_{110}\right)^{1/6}\left(\hat{N}_{011}\right)^{2/3}\left(\hat{N}_{001}\right)^{1/3}+\left(\hat{N}_{101}\right)^{1/6}\left(\hat{N}_{100}\right)^{1/6}\hat{N}_{010}}$	$\frac{\left(\hat{N}_{101}\right)^{1/3}\left(\hat{N}_{100}\right)^{1/3}\left(\hat{N}_{001}\right)^{1/3}}{\left(\hat{N}_{111}\right)^{1/3}\left(\hat{N}_{110}\right)^{1/3}\left(\hat{N}_{011}\right)^{1/3}}\hat{N}_{010}$
4	X_1, X_2, X_3 X_2X_3	$\frac{\left(\hat{N}_{111}\right)^{1/6}\left(\hat{N}_{110}\right)^{1/6}\left(\hat{N}_{101}\right)^{2/3}\left(\hat{N}_{001}\right)^{1/3}}{\left(\hat{N}_{111}\right)^{1/6}\left(\hat{N}_{110}\right)^{1/6}\left(\hat{N}_{101}\right)^{2/3}\left(\hat{N}_{001}\right)^{1/3}+\hat{N}_{100}\left(\hat{N}_{011}\right)^{1/6}\left(\hat{N}_{010}\right)^{1/6}}$	$\frac{\left(\hat{N}_{011}\right)^{1/3}\left(\hat{N}_{010}\right)^{1/3}\left(\hat{N}_{001}\right)^{1/3}}{\left(\hat{N}_{111}\right)^{1/3}\left(\hat{N}_{110}\right)^{1/3}\left(\hat{N}_{101}\right)^{1/3}}\hat{N}_{100}$
5	X_1, X_2, X_3 $X_1X_2,\ X_1X_3$	$\frac{\hat{N}_{011}}{\hat{N}_{011}+\hat{N}_{010}}$	$\frac{\hat{N}_{010}\hat{N}_{001}}{\hat{N}_{011}}$
6	X_1, X_2, X_3 $X_1X_2,\ X_2X_3$	$\frac{\hat{N}_{101}}{\hat{N}_{101}+\hat{N}_{100}}$	$\frac{\hat{N}_{100}\hat{N}_{001}}{\hat{N}_{101}}$
7	X_1, X_2, X_3 $X_1X_3,\ X_2X_3$	$\frac{\hat{N}_{110}\left(\hat{N}_{101}\right)^{1/4}\left(\hat{N}_{011}\right)^{1/4}\left(\hat{N}_{001}\right)^{3/4}}{\hat{N}_{110}\left(\hat{N}_{101}\right)^{1/4}\left(\hat{N}_{011}\right)^{1/4}\left(\hat{N}_{001}\right)^{3/4}+\left(\hat{N}_{111}\right)^{1/4}\hat{N}_{100}\hat{N}_{010}}$	$\frac{\hat{N}_{100}\hat{N}_{010}}{\hat{N}_{110}}$
8	X_1, X_2, X_3 $X_1X_3,\ X_2X_3,\ X_2X_3$	$\frac{\hat{N}_{110}\hat{N}_{101}\hat{N}_{011}}{\hat{N}_{110}\hat{N}_{101}\hat{N}_{011}+\hat{N}_{111}\hat{N}_{100}\hat{N}_{010}}$	

[a] The intercept (α) is included in all models.

[b] $\hat{N}_{ijk}$ denotes the fitted cell count with capture history (ijk) and is obtained by computing estimated $E(N_{ijk})$ from the fitted log-linear model, where $i, j, k \in \{0, 1\}$; Under models 1- 4, and 7 (which are unsaturated models), fitted cell counts do not have closed form and can be computed by numerically maximize the Poisson log-likelihood; while fitted cell counts $\hat{N}_{011} = n_{011}$ and $\hat{N}_{010} = n_{010}$ under the model 5, and fitted cell counts $\hat{N}_{101} = n_{101}$ and $\hat{N}_{100} = n_{100}$ under the model 6; the saturated model 8 yields each of fitted cell counts equal to its corresponding observed cell count.

[c] Fitted N_{000} is computed as $\exp(\hat{\alpha})$, where $\hat{\alpha}$ is the MLE of α.

For the three-stream CRC data reflected in Figures 4.2 and 4.3, Tables 4.4 and 4.5 present results from the candidate log-linear models following the standard conventions, together with those models (among all possible) that yield the most favorable AIC. Table 4.4 illustrates that all saturated models fit the data equivalently, despite the fact that the resulting estimates range from 152,000 to 877,000. No unsaturated model "beats" those saturated models in terms of AIC. When following the standard conventions, model 8 (resulting in an estimate far from the true $N = 200,000$) would be selected. We note again that these data were simulated under a referral scenario that no log-linear model is able to incorporate; thus, researchers have no chance of approaching the true dependency structure even if they carefully follow the recommended log-linear model paradigm. Similarly, we observe that model 9 in Table 4.5 is an unsaturated model which yields lower AIC than the saturated models. This model assumes the testable constraint, $E(N_{110})/E(N_{101}) = E(N_{010})/E(N_{001})$, and projects the unobserved count based on the subsequent assumption $E(N_{000}) = E(N_{100})E(N_{001})/E(N_{101})$. While the observable constraint might be supported based on the observed data, the assumption about the unobserved count is a mathematical construct that could not actually be justified without external knowledge about operating characteristics of the surveillance streams.

We conducted additional simulation studies to further demonstrate that AIC is not a reliable tool for CRC model selection. These simulations focus on the three-stream case, which is a common scenario in epidemiological studies. We simulated 1,000 datasets from a population-level multinomial model under two different scenarios assuming $N = 5,000$. In the first scenario, one testable assumption and one untestable assumption were imposed, while in the second scenario, two testable assumptions and one untestable assumption were imposed. Details of these simulation settings are given in Appendix C.

Under scenario 1, it is easy to show that the model that was most frequently

Table 4.4: Log-linear models for three-stream data simulated by assuming $N = 200,000$ and analyzed by (Jones et al., 2014) under the usual conventions and with most favorable AIC

Fitted models under the usual conventions										
Model	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	$\hat{N}$	$\hat{\psi}$	AIC
1	✓	✓	✓					199,910	0.446	51,619
2	✓	✓	✓	✓				260,386	0.342	18,469
3	✓	✓	✓		✓			196,734	0.461	51,544
4	✓	✓	✓			✓		161,909	0.723	21,365
5	✓	✓	✓	✓	✓			877,366	0.077	7,936
6	✓	✓	✓	✓		✓		181,214	0.614	3,210
7	✓	✓	✓		✓	✓		152,553	0.852	14,078
8	✓	✓	✓	✓	✓	✓		306,540	0.272	93
Fitted models with most favorable AIC										
Model	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	$\hat{N}$	$\hat{\psi}$	AIC
9	✓	✓	✓	✓	✓		✓	877,366	0.077	93
10	✓	✓	✓	✓		✓	✓	181,214	0.614	93
11	✓	✓	✓		✓	✓	✓	152,553	0.863	93
12	✓	✓		✓	✓	✓	✓	203,973	0.5	93
13	✓		✓	✓	✓	✓	✓	162,316	0.759	93
14		✓	✓	✓	✓	✓	✓	154,311	0.842	93

[†]✓ indicates the predictor is included.

Table 4.5: Log-linear models for three-stream analyzed by Poorolajal et al. (2017) under the usual conventions and with most favorable AIC

Fitted models under the usual conventions										
Model	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	$\hat{N}$	$\hat{\psi}$	AIC
1	✓	✓	✓					10,421	0.04	112
2	✓	✓	✓	✓				8,550	0.048	105
3	✓	✓	✓		✓			9,751	0.044	111
4	✓	✓	✓			✓		13,792	0.026	66
5	✓	✓	✓	✓	✓			6,290	0.074	90
6	✓	✓	✓	✓		✓		17,149	0.021	66
7	✓	✓	✓		✓	✓		14,122	0.026	68
8	✓	✓	✓	✓	✓	✓		419,22	0.008	60
Fitted models with most favorable AIC										
Model	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	$\hat{N}$	$\hat{\psi}$	AIC
9	✓	✓	✓			✓	✓	14,904	0.024	59
10	✓	✓	✓	✓	✓		✓	6,291	0.074	60
11	✓	✓	✓	✓		✓	✓	17,149	0.021	60
12	✓	✓	✓		✓	✓	✓	14,122	0.026	60
13	✓	✓		✓	✓	✓	✓	2,600	0.5	60
14	✓		✓	✓	✓	✓	✓	3,010	0.304	60
15		✓	✓	✓	✓	✓	✓	3,353	0.229	60

[†]✓ indicates the predictor is included.

(80%) selected by the AIC implies the testable assumption used for data generation (Table 4). The untestable assumption projected by this model suggests that the key parameter $\psi = p_{3|\bar{1}\bar{2}} = 0.5$. However, the data were generated assuming this key parameter equals 0.1. Hence, it is not surprising that the average estimate obtained from the selected model $(3,040)$ is far afield from the true value of $5,000$. The other two models that were selected by the AIC under some simulations also project very different estimates of N. In similar fashion, the two most frequently selected models under scenario 2 permit the two prespecified testable assumptions when certain constraints are applied to model coefficients. Again, the average estimates obtained from the two most frequently selected models (6,848 and 4,407) fall far from the truth because neither model projects the correct untestable assumption.

When applying the standard conventions under scenario 1, note that the saturated model was almost always selected as "best". However, the corresponding averaged estimated N is 8,053, greatly overestimating the true value of 5,000. All of these simulation results highlight the fact that using AIC as a CRC model selection metric is misleading, since models fitting the observed data "best" cannot in fact be assumed to reliably project the correct untestable assumption which is critical for estimating N.

4.6 Discussion

In this chapter, we first demonstrated that the CRC log-linear model framework is highly exclusionary, in the sense that applying that framework can exclude, by design, broad ranges of estimates of N that are in fact as consistent with the observed data as any others. In practice, we believe any CRC model framework that ignores potentially valid estimates before they can even be considered should be used with great caution. For example, prior authors suggest that epidemiological CRC data streams often tend

Table 4.6: Frequency of log-linear models selected by the AIC and averaged estimates from AIC-favored log-linear models across 1,000 simulations among 127 possible log-linear models and 8 possible log-linear models under the usual conventions

All possilbe 127 models									
	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	Frequency [a]	Averaged Estimates of $N = 5,000$
Scenario 1 [b]	✓	✓		✓	✓		✓	798	3,040
		✓	✓	✓		✓	✓	52	3,804
	✓	✓	✓	✓			✓	150	8,052
Scenario 2 [c]	✓	✓	✓	✓	✓			490	6,848
	✓	✓	✓	✓			✓	1	6,232
	✓	✓			✓	✓	✓	1	4,646
	✓		✓		✓	✓	✓	453	4,407
	✓	✓	✓	✓	✓	✓		55	6,871
8 possible models under the usual conventions									
	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	Frequency [a]	Averaged Estimates of $N = 5,000$
Scenario 1 [b]	✓	✓	✓	✓		✓		4	4,089
	✓	✓	✓	✓	✓	✓		996	8,053
Scenario 2 [c]	✓	✓	✓	✓	✓			846	6,848
	✓	✓	✓	✓		✓		1	4,409
	✓	✓	✓	✓	✓	✓		153	6,871

[†] ✓ indicates the predictor is included.
[a] The number of simulations that the model was selected by the AIC across 1,000 simulations.
[b] Data were generated from population-level multinomial model with $N = 5,000$, $p_1 = 0.3$, $p_{2|1} = 0.2$, $p_{2|\bar{1}} = 0.3$, $p_{3|12} = 0.8$, $p_{3|1\bar{2}} = 0.16$, $p_{3|\bar{1}2} = 0.5$, $\psi = 0.1$. Those parameters imply $E(N_{011}) = E(N_{010}) = 525$ and $p_{3|12}/p_{3|\bar{1}2} = p_{3|1\bar{2}}/\psi = 1.6$.
[c] Data were generated from population-level multinomial model with $N = 5,000$, $p_1 = 0.3$, $p_{2|1} = 0.5$, $p_{2|\bar{1}} = 0.3$, $p_{3|12} = 0.35$, $p_{3|1\bar{2}} = 0.35$, $p_{3|\bar{1}2} = 0.25$, $\psi = 0.4375$. Those parameters imply $E(N_{111}) = E(N_{101}) = 262.5$, $E(N_{110}) = E(N_{100}) = 487.5$, $p_{3|1\bar{2}}/\psi = 0.8$.

to reflect a net positive dependence (Hook and Regal, 2000). However, as we have seen in Figures 4.1 - 4.4, log-linear models might exclude many feasible estimates corresponding to the situation in which data streams are positively correlated.

As stated in Coull and Agresti (1999), the dependency between data streams cannot reliably be uncovered using only the observed cell counts. We have further emphasized this point via simulation studies demonstrating that the model selection procedure relying on the metric AIC is fundamentally deceiving. Specifically, the model that fits the observed data "best" cannot be assumed to project the correct untestable assumption about the unobserved cell count. Other model selection metrics (e.g., BIC) would suffer from the same problem. It is also worth noting that adding extra data streams cannot necessarily provide information for reliably inferring the dependency among existing data streams. Along these lines, it can be tempting to infer pairwise dependences using three-stream data (Lum and Ball, 2015). However, no matter how many data streams are included, the number of cases not captured by any streams is still unobserved. Thus, dependencies among data streams typically remain unclear. Because of the exclusionary property and the fact that model selection in the CRC log-linear model framework is fraught with the pitfalls exhibited herein, we encourage concerted efforts toward a departure from the current reality in which that framework is the centerpiece of standard practice for CRC-based epidemiological surveillance.

As one step toward such a departure, we note that the novel visualization plots based on the MLE in Equation (4.1) have been demonstrated to permit a continuum of estimates of N. This closed-form MLE potentially sheds light on an alternative modeling framework (discussed in details in Chapter 5) that would avoid the "exclusionary" problem of the log-linear model. In terms of model selection, an accessible alternative framework focused on leveraging an epidemiologist's knowledge about operating characteristics of surveillance streams would be attractive, since (despite what

metrics like AIC might suggest) no information about the true dependencies among streams is available in the observed data alone. A key benefit of such an approach would be the ability to foster careful consideration and transparency with regard to the untestable assumption(s) upon which estimation is based, as opposed to having them dictated as murky mathematical constructs based on the modeling framework that one adopts. Nevertheless, we believe the nature of CRC surveillance will almost always suggest a role for sensitivity and/or uncertainty analysis (Zhang and Small, 2020; Zhang et al., 2022).

Chapter 5

A CRC Modeling Framework for Disease Surveillance Emphasizing Expert Opinion in the Spirit of Prior Information

5.1 Background

Motivated by the two main pitfalls associated with the log-linear model illustrated in Chapter 4, we propose a modeling framework based on a population-level multinomial model and hinging on a key parameter to characterize dependencies among data streams. In the proposed framework, all possible estimates are theoretically attainable. In addition, the framework incorporates expert opinion about the level of dependency between data streams in a transparent way to drive estimation rather than purely relying on mathematical constructs to project dependency assumptions which could be hidden and misleading (Zhang et al., 2023a).

In the multinomial model, dependencies between data streams are introduced by representing probabilities of possible capture histories based on a set of parameters (Fienberg, 1972). In the proposed model, we consider a set of conditional probabilities to characterize probabilities of different capture histories. This parametrization is in the same spirit as one introduced previously by Farcomeni (2011). Specifically, in Farcomeni (2011), a general class of models was developed to enable the estimation of total case counts by introducing linear constraints on those conditional probabilities. This idea of utilizing conditional probabilities was further extended to accommodate observed and unobserved heterogeneities in capture probabilities by including covariates and random effects when analyzing individual-level CRC data (Farcomeni, 2016).

Although our proposed modeling framework uses an analogous parameterization, it significantly differs from previous work by allowing the user to impose both linear and non-linear constraints on conditional probabilities. Additionally, the proposed model treats one conditional probability as the key parameter (i.e., the probability of identification by the last data stream given lack of identification by any other data stream) which is the basis to incorporate expert understanding of how data streams interact with each other. The introduction of this key parameter enables a closed-

form estimator under various constraints based on the multinomial distribution-based likelihood. We also propose bias-corrected estimators to reduce mean bias associated with the closed-form estimator under certain constraints. As a demonstrably preferable alternative to confidence intervals based on the asymptotic normality of the estimator, we provide credible intervals with favorable frequentist properties by extending a Dirichlet-multinomial-based procedure developed by Lyles et al. (2021a). This Dirichlet-multinomial-based approach further permits uncertainty analyses to produce credible intervals that acknowledge the user's level of confidence on the dependency assumption by assigning a distribution to a selected key dependence parameter.

5.2 Methods

5.2.1 Preliminaries

Consider a CRC study where $K \geq 2$ surveillance systems are implemented for monitoring a disease among a closed population in which N diseased cases exist. Denote $\mathcal{O} = \{(11\ldots11), (11\ldots10), \ldots, (00\ldots00)\}$ the set that includes 2^K sequences, where each sequence of length K consists of 0s and 1s and represents a possible capture history, 1 indicates captured by the system and 0 indicates not captured. In the set $\mathcal{O}$, capture histories are arranged in the lexicographic order. Let N_{h_i} denote the true number of cases having capture history h_i that is the i-th element in $\mathcal{O}$, for $i = 1, \ldots, 2^K$. Note that $\sum_{i=1}^{2^K} N_{h_i} = N$. Due to the nature of CRC studies, we cannot observe cases that are not captured by any system, i.e., $N_{h_{2^K}}$ is unobserved. In other words, only $n_c = \sum_{i=1}^{2^K-1} n_{h_i}$ cases are observed. We model those counts N_{h_i} using a population-level multinomial model (Darroch, 1958),

$$\{N_{h_i}\}_{i=1}^{2^K} \sim \text{Multinomial}\left(N, \{p_{h_i}(\boldsymbol{\theta})\}_{i=1}^{2^K}\right), \tag{5.1}$$

where $p_{h_i}(\boldsymbol{\theta})$ is the probability of having capture history h_i and $\boldsymbol{\theta}$ is the vector of parameters used for characterizing multinomial probabilities. In this dissertation, we choose a marginal probability and $2^K - 1$ conditional probabilities as parameters to make this characterization. Specifically, $\boldsymbol{\theta} = (p_1, p_{2|1}, p_{2|\bar{1}}, p_{3|12}, p_{3|\bar{1}2}, \ldots, p_{K|\overline{12\ldots K-1}})$, where the marginal probability p_1 is the probability of being captured by the first system and the conditional probabilities again are arranged in the lexicographic order where $\bar{k}$ indicates not captured by the k-th system, otherwise yes. For example, $p_{3|\bar{1}2}$ represents the proportion of being captured by the third system conditional on identified by the second system but not identified by the first system. Compared to conditional probabilities used in Farcomeni (2011) that describe capture probabilities for each individual, the conditional probabilities included in $\boldsymbol{\theta}$ are used to characterize the probability of having certain capture history at the population-level. Focusing on population-level parameters, the model in Equation (5.1) does not require capture probabilities are homogeneous across individuals (Lyles et al., 2021a). We note that the labeling of the data streams (i.e., determining which data stream should be labeled as the last data stream) matters in this framework, and should be considered carefully as part of the process of exerting expert opinion (see Section 5.5). Any capture probability $p_{h_i}(\boldsymbol{\theta})$ can then be factorized using those probabilities. As an example, $p_{011}(\boldsymbol{\theta}) = (1 - p_1)p_{2|\bar{1}}p_{3|\bar{1}2}$ when $K = 3$.

Using the model given in (5.1), the likelihood of N and $\boldsymbol{\theta}$ can be written as follows:

$$\begin{aligned} L(N, \boldsymbol{\theta}) &= L_b(N, \boldsymbol{\theta}) \times L_c(\boldsymbol{\theta}) \\ &= \frac{N!}{n_c!(N - n_c)!} p_c(\boldsymbol{\theta})^{n_c} \left[1 - p_c(\boldsymbol{\theta})\right]^{(N - n_c)} \times \frac{n_c!}{\prod_{i=1}^{2^K-1} n_{h_i}!} \prod_{i=1}^{2^K-1} \left[\frac{p_{h_i}(\boldsymbol{\theta})}{p_c(\boldsymbol{\theta})}\right]^{n_{h_i}}, \end{aligned} \tag{5.2}$$

where $p_c(\boldsymbol{\theta}) = \sum_{i=1}^{2^K-1} p_{h_i}(\boldsymbol{\theta})$ is the probability of being captured at least once. As shown in Equation (5.2), the likelihood $L(N, \boldsymbol{\theta})$ is decomposed into two parts: the

binomial likelihood $L_b(N, \boldsymbol{\theta})$, and the conditional likelihood $L_c(\boldsymbol{\theta})$ which does not include N. With this likelihood decomposition, a well-known estimator is given by

$$\hat{N} = \frac{n_c}{p_c(\hat{\boldsymbol{\theta}})}, \tag{5.3}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE) derived by maximizing $L_c(\boldsymbol{\theta})$ (Fienberg, 1972; Sanathanan, 1972). At least one constraint that is unverifiable based only on the observed data is required to estimate $\boldsymbol{\theta}$ and N, since only $2^K - 1$ counts are observed. Our approach hinges on the notion that couching this unverifiable constraint in terms of an interpretable parameter, and directing expert opinion toward it, opens up a CRC framework with significant advantages over existing paradigms that attempt to identify that parameter solely through mathematical constructs.

Let ψ denote the last conditional probability $p_{K|\overline{12\ldots K-1}}$ in $\boldsymbol{\theta}$, which is the proportion of being identified by the last system but not by any other system. When ψ is known, a closed-form MLE of N based on incorporating an arbitrary number (K) of data streams using the model in Equation (5.1) has been introduced in Chapter 4, that is (Zhang et al., 2023b):

$$\hat{N}_\psi = (n_c - n_{0\ldots1}) + \frac{n_{0\ldots1}}{\psi}, , \tag{5.4}$$

where $n_{0\ldots1}$ is the observed number of cases captured by the last system but not by any other system. By varying the value of ψ, the MLE in Equation (5.4) allows one to provide estimates of N corresponding to all feasible dependency structures across systems. For example, Figure D.1 shows possible estimates of N under three- and four-catch cases when valid ψ values are supplied into Equation (5.4). This MLE acknowledges the continuum of estimated N. Again, we emphasize that labeling matters when applying this MLE. As shown in Figure D.1, assigning different labels to data streams results in different estimates of N even if the same value of ψ is

assumed.

The assumption about the dependency structure between systems is key to the estimation of N (Chao et al., 2001; Lyles et al., 2021a). This MLE has been shown to be an unbiased estimator if the assumed ψ is correct, although inserting any valid value of $\psi \in (0, 1]$ leads to an estimate that is equally consistent with the observed data (Lyles et al., 2021a). The MLE in Equation (5.4) is also the MLE in Equation (5.3) when assuming ψ is known.

5.2.2 Proposed modeling framework

Point estimation

The MLE in Equation (5.4) provides a precise closed-form estimator of N; however, specifying an exact value of the non-identifiable parameter ψ is rarely possible except under unique study designs (e.g., Lyles et al. (2021b)). To exploit this closed-form MLE more generally in practice, we consider a class of estimators which is obtained by relating ψ to other conditional probabilities in the parameter vector $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^*$ denote the rest of the parameters in $\boldsymbol{\theta}$ with ψ removed and define $\psi = g(\boldsymbol{\theta}^*)$, where $g(\cdot)$ is a function that relates ψ to other parameters. When the constraint $\psi = g(\boldsymbol{\theta}^*)$ is imposed, all parameters in $\boldsymbol{\theta}^*$ are estimable using the conditional likelihood in Equation (5.2). The estimator of N obtained by plugging the estimate $\hat{\psi} = g(\widehat{\boldsymbol{\theta}^*})$, where $\widehat{\boldsymbol{\theta}^*}$ denotes the estimated $\boldsymbol{\theta}^*$, into Equation (5.4) is:

$$\hat{N}_{\psi}\Big|_{\psi=g(\widehat{\boldsymbol{\theta}^*})} = (n_c - n_{0\ldots1}) + \frac{n_{0\ldots1}}{g(\widehat{\boldsymbol{\theta}^*})}. \tag{5.5}$$

The estimator in Equation (5.5) theoretically permits all possible estimates of N that are consistent with the observed data. For example, defining $g(\boldsymbol{\theta}^*) = p_{2|1}$ for $K = 2$ means that the resulting estimator corresponds to the estimator derived in the prior study under the independence assumption (Lincoln, 1930; Lyles et al., 2021a).

In epidemiological studies, three-catch CRC data ($K = 3$) are commonly encountered. To provide an overview of constraints that can be incorporated using the estimator in Equation (5.5), we present possible constraints having practical interpretations under the three-catch case in Table 5.1.

Table 5.1: Possible constraints that can be incorporated under the three-catch case

	Constraint [a]	Implications of the constraint when $r = 1$	Comments on the labeling of data streams under the constraint
1	$\psi = \frac{p_{3\|12}}{r}$	The proportion of cases identified by stream 3 among cases that are NOT captured by streams 1 and 2 is equal to the proportion of cases identified by stream 3 among cases that are captured by both streams 1 and 2.	The estimation of N is sensitive to the labeling.
2	$\psi = \frac{p_{3\|\bar{1}2}}{r}$	Streams 2 and 3 operate independently among cases that are NOT captured by stream 1.	The estimation of N is invariant to the switch labels of stream 2 and stream 3.
3	$\psi = \frac{p_{3\|1\bar{2}}}{r}$	Streams 1 and 3 operate independently among cases that are NOT captured by stream 2.	The estimation of N is invariant to the switch labels of stream 1 and stream 3.
4	$\psi = r\frac{p_{3\|\bar{1}2}p_{3\|1\bar{2}}}{p_{3\|12}}$	(a) The conditional association between two data streams is measured by a ratio of two conditional probabilities in $\boldsymbol{\theta}$. (b) The association between streams 1 and 3 among cases captured by stream 2 is the same as the association between stream 1 and 3 among cases NOT captured by stream 2. (c) The association between streams 2 and 3 among cases captured by stream 1 is the same as the association between stream 2 and 3 among cases NOT captured by stream 1.	The estimation of N is sensitive to the labeling.
5	$\frac{\psi}{1-\psi} = r\frac{\left[\frac{p_{3\|\bar{1}2}}{1-p_{3\|\bar{1}2}}\right]\left[\frac{p_{3\|1\bar{2}}}{1-p_{3\|1\bar{2}}}\right]}{\left[\frac{p_{3\|12}}{1-p_{3\|12}}\right]}$	(a) The conditional association between two data streams is measured by an odds ratio defined based on conditional probabilities in $\boldsymbol{\theta}$. (b) The association between streams 1 and 3 among cases captured by stream 2 is the same as the association between stream 1 and 3 among cases NOT captured by stream 2. (c) The association between streams 2 and 3 among cases captured by stream 1 is the same as the association between stream 2 and 3 among cases NOT captured by stream 1.	The estimation of N is sensitive to the labeling.

[a] r can be a fixed value or follow a pre-parametric distribution in the uncertainty analysis; $\psi = p_{3|\bar{1}\bar{2}}$) under the three-catch case.

When relating ψ to $\boldsymbol{\theta}^*$, the estimator in Equation (5.5) coincides with the MLE in Equation (5.3) if only one equality constraint [via $g(\boldsymbol{\theta}^*)$] is imposed. We note that when using ψ obtained from maximizing the conditional likelihood while assuming additional equality constraints involving estimable parameters, the estimator in

Equation (5.5) is less efficient compared to the MLE in Equation (5.3). The efficiency gained from imposing the extra constraints is not leveraged, since observed counts (rather than fitted ones based on such constraints) are used in Equation (5.5). Consider for example a case where $K = 3$, let $g(\boldsymbol{\theta}^*) = p_{3|1\bar{2}}$ (i.e.,constraint 3 in Table 5.1 when $r = 1$) and assume an additional testable equality constraint $p_{3|12} = p_{3|\bar{1}2}$. This additional constraint implies that $E(N_{111})E(N_{010}) = E(N_{011})E(N_{110})$. However, whereas both constraints could be incorporated via Equation (5.3), using Equation (5.5) to estimate N does not accommodate this additional constraint because the observed counts are used. In the CRC context, testable constraints can be defensed using common model selection metrics (e.g., AIC) based on the observed data alone. As a result of imposing extra testable constraints, an improvement in the precision of the estimation of N can be achieved. However, in fact, in the proposed modeling framework, we persuade researchers to not take advantage of the efficiency gain by imposing additional testable equality constraints. This is because the untestable constraint hinging on ψ is crucial to the estimation of N, while testable constraints are less relevant to the validity of the estimation. Compared to gain statistical precision, it is more important to focus attention on the untestable constraint that directly relates to the validity of the estimation.

In our proposed modeling framework, we recommend the estimator in Equation (5.5) for general use for several reasons. First, this closed-form estimator with given $g(\widehat{\boldsymbol{\theta}}^*)$ greatly facilitates bias-corrections, inference, and the incorporation of uncertainty in assumptions imposed via $g(\boldsymbol{\theta}^*)$ (as discussed in later sections) under various situations which are potentially encountered in practice. Specifically, these advantages occur when ψ is related to estimable parameters through only one equality constraint defined by $g(\boldsymbol{\theta}^*)$, where estimable parameters are referred to parameters in $\boldsymbol{\theta}$ that are not directly included when computing $p_c(\boldsymbol{\theta}) = 1 - (1 - p_1)\prod_{k=2}^{K}(1 - p_{k|\overline{1\ldots k-1}})$. For example, $p_{2|1}$, $p_{3|12}$, $p_{3|\bar{1}2}$, and $p_{3|1\bar{2}}$, are the estimable parameters in $\boldsymbol{\theta}$ under the

three-catch case, and each can be readily estimated in closed form. For instance, the estimator of $p_{3|12}$ is $\frac{n_{111}}{n_{111}+n_{110}}$, and estimators of the other two estimable parameters (i.e., $p_{3|\bar{1}2}$ and $p_{3|1\bar{2}}$) are obtained similarly. When ψ is defined as a function of estimable parameters, using Equation (5.5) guarantees a closed form estimator for N and avoids the complication of inestimable parameters (such as $p_{2|\bar{1}}$) required in the computation of $p_c(\boldsymbol{\theta})$ that may necessitate numerical maximization of the conditional likelihood in order to use Equation (5.3). This could be especially laborious when a non-linear constraint is assumed in $g(\boldsymbol{\theta}^*)$.

In addition to the aforementioned statistical advantages, the introduction of $g(\boldsymbol{\theta}^*)$ allows the practitioner to accommodate external knowledge about the dependency between systems. In other words, an expert's understanding and opinion about how the systems interact with each other can be directly incorporated through careful consideration of how the key parameter ψ relates to the estimable parameters in $\boldsymbol{\theta}$. Expert opinion can be the most valuable tool for guiding the estimation of N, since the assumption about the dependency cannot be verified through any model selection metric based on the observed data alone.

Bias corrections

As discussed in Chapter 3, under the two-catch case ($K = 2$), previous developments led to an estimator (i.e., Equation (3.3)) allowing any level of dependency between two systems by introducing a ratio parameter $\phi = p_{2|1}/\psi$, where ϕ measures the population-level dependency level ($= 1$ indicates independence, > 1 indicates positively associated, and < 1 indicates negatively associated) (Zhang et al., 2023a). This estimator can be obtained by defining $g(\boldsymbol{\theta}^*) = p_{2|1}/\phi$ using Equation (5.5), assuming ϕ is known. We have showed that this estimator is biased in Chapter 3, and developed two bias-corrected estimators using a Taylor-series expansion approach (Lyles et al., 2021a; Zhang et al., 2023a). In this Chapter, we exploit the fact that the analogue

to ϕ can be defined similarly for any $K > 2$, yielding a corresponding estimator of N via Equation (5.5). As in the two-stream case, these resulting estimators with given ratio parameters are biased; however, the Taylor expansion approach can again be applied to derive bias-corrected estimators. Here, we specifically extend this approach to the three-catch case and derive bias-corrected estimators based on three different definitions of the ratio parameter. Those estimators are:

$$\hat{N}_{\phi_1} = (n_c - n_{001}) + \frac{n_{001}(n_{111} + n_{110})}{n_{111}}\phi_1 - \frac{n_{110}n_{001}}{(n_{111} + 0.5)^2}, \tag{5.6}$$

$$\hat{N}_{\phi_2} = (n_c - n_{001}) + \frac{n_{001}(n_{011} + n_{010})}{n_{011}}\phi_2 - \frac{n_{010}n_{001}}{(n_{011} + 0.5)^2}, \tag{5.7}$$

$$\hat{N}_{\phi_3} = (n_c - n_{001}) + \frac{n_{001}(n_{101} + n_{100})}{n_{101}}\phi_3 - \frac{n_{100}n_{001}}{(n_{101} + 0.5)^2}, \tag{5.8}$$

where $\phi_1 = p_{3|12}/\psi$, $\phi_2 = p_{3|\bar{1}2}/\psi$, and $\phi_3 = p_{3|1\bar{2}}/\psi$. These three ratio parameters are directly interpretable, note that they are the same as the r defined in Table 5.1 for constraints 1-3. For example, $\phi_2 = 1$ indicates that the second system and the third system operate independently among cases that are not captured by the first system, which implies $g(\boldsymbol{\theta}^*) = p_{3|\bar{1}2}$. We note that the number of analogues of ϕ increases rapidly as K increases, as there are a total of $\sum_{j=0}^{K-2}\binom{K-1}{j}$ analogues of ϕ for a given K. In addition, when $g(\boldsymbol{\theta}^*)$ imposes a non-linear relationship between ψ and other estimable parameters (e.g., $g(\boldsymbol{\theta}^*) = \frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$), the derivation of the bias correction can become laborious because the Taylor expansion approach requires one to compute the Hessian matrix of a function of multiple conditional capture probabilities. Specifically, the function of interest includes $2^K - 1$ conditional capture probabilities which are defined as probabilities of having observable capture histories conditional on being identified while incorporating constraints specified via $g(\boldsymbol{\theta}^*)$. The Taylor expansion approach is not easily generalizable to such cases, as the derivation is specific to the definition of that g function.

Under the independence assumption for $K = 2$, there is a well-known classical

bias-corrected estimator (Chapman, 1951). Lyles et al. (2021a) demonstrated that inserting the $Beta(1,0)$ prior-based posterior mean of $p_{2|1}$ (i.e., $\frac{n_{11}+1}{n_{11}+n_{10}+1}$) as ψ into the estimator in Equation (5.4) replicates the Chapman estimator. Motivated by this fact, we propose a straightforward and generalizable bias-correction strategy which inserts $Beta(1,0)$ posterior means into Equation (5.5) for any estimable parameters appearing in $g(\boldsymbol{\theta}^*)$. For example, given $g(\boldsymbol{\theta}^*) = \frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$ for $K = 3$, this strategy yields the following convenient bias-corrected estimator:

$$(n_c - n_{001}) + \frac{n_{001}(n_{111}+1)(n_{011}+n_{010}+1)(n_{101}+n_{100}+1)}{(n_{101}+1)(n_{011}+1)(n_{111}+n_{110}+1)}$$

Compared to the Taylor expansion approach, this strategy can be easily applied to incorporate any form of $g(\boldsymbol{\theta}^*)$ that links ψ to estimable parameters, for any given K. In Section 5.3, we compare the two bias correction approaches in scenarios where both are accessible.

Inference via Bayesian credible intervals

To flexibly incorporate various forms of $g(\boldsymbol{\theta}^*)$ and achieve favorable coverage, we extend the Dirichlet-multinomial-based approach proposed in Lyles et al. (2021a) to provide credible intervals to accompany the estimator of N in Equation (5.5) while incorporating constraints imposed via $g(\boldsymbol{\theta}^*)$ for any given K. First, note that the conditional likelihood given in Equation (5.2) implies the observed counts conditional on n_c follow a multinomial distribution with probabilities $p^*_{h_i} = \frac{p_{h_i}(\boldsymbol{\theta})}{p_c(\boldsymbol{\theta})}$ for $i = 1, \ldots, 2^K - 1$. To obtain 95% credible intervals for N under various assumptions which are introduced via different constraints defined by $g(\boldsymbol{\theta}^*)$, we begin by generating multiple simulated datasets by assigning a Dirichlet prior to $\boldsymbol{p}^* = (p^*_{h_1}, \ldots, p^*_{h_{2^K-1}})$. Specifically, simulated datasets are obtained from multiplying n_c by posterior samples of $\boldsymbol{p}^*$. For the l-th simulated dataset, an estimate $\hat{N}_l$ is computed using Equation

(5.5) with the corresponding pre-specified $g(\boldsymbol{\theta}^*)$. However, these estimates of N are conditional on n_c. To account for the uncertainty in n_c, we implement an additional step. We draw a sample from a binomial distribution with probability $p_{c,l} = n_{c,l}/\hat{N}_l$ and size $n_c/p_{c,l}$ (this value is rounded to the nearest integer) which we denote as $n^*_{c,l}$, where $n_{c,l}$ is the sum of the generated observed cell counts based on the l-th simulated dataset. Finally, we take the 2.5^{th} and 97.5^{th} percentiles of samples computed as $n^*_{c,l}/p_{c,l}$ to form the 95% credible interval of N under the assumption specified via $g(\boldsymbol{\theta}^*)$. A detailed summary of the steps for constructing the proposed credible intervals can be found in Appendix D.1.

Uncertainty analysis

In our proposed modeling framework, assumptions about the dependency structure across systems are introduced by focusing investigator attention upon a key non-identifiable but interpretable population-level parameter ψ, and defining $g(\boldsymbol{\theta}^*)$ as a way to formulate a well-considered assumption about ψ in terms of estimable parameters. However, it is not possible to be certain about this dependency level assumption in most epidemiological studies. We propose a simulation-based approach to propagate uncertainty about the parameter characterizing assumptions about the dependency level. A parametric distribution (e.g., uniform or normal) is introduced to the parameter that characterizes the dependency strength under a given dependency structure which is specified through $g(\boldsymbol{\theta}^*)$. In the proposed uncertainty analysis, we draw multiple samples from the assumed distribution for that parameter. For each simulated value of that parameter, we then apply the Dirichlet-based approach to obtain samples of the estimated N under the specified $g(\boldsymbol{\theta}^*)$. We pool all samples across different values of that parameter together to construct the credible interval that incorporates both statistical uncertainties and the uncertainty about that parameter. Details about the proposed uncertainty analysis, which generalizes past work (e.g.,

Chatterjee and Mukherjee (2016), Zhang et al. (2023a)) in the two-stream case to the setting of arbitrary K, can be found in Appendix D.2.

5.3 Simulations

We conducted simulation studies to evaluate the performance of the proposed bias-corrected estimators, the Dirichlet-multinomial-based credible intervals, and the uncertainty analysis under various scenarios. For each scenario, 1000 replicates were simulated. We first evaluated the performance of the proposed bias-corrected estimators for estimating N, and the Dirichlet-multinomial-based credible intervals. We also implemented the log-linear model, (perhaps the most commonly used CRC method in epidemiological studies), for comparison whenever applicable (Cormack, 1989; Jones et al., 2014). Details of all of our simulation settings are given in Appendix D.3.

We consider here two constraints as examples under the three-catch case: (A) $p_{3|\bar{1}2} = \psi$ and (B) the odds ratio constraint $\frac{p_{3|12}/(1-p_{3|12})}{p_{3|\bar{1}2}/(1-p_{3|\bar{1}2})} = \frac{p_{3|1\bar{2}}/(1-p_{3|1\bar{2}})}{\psi/(1-\psi)}$. These two constraints are constraints 2 and 5 showed in Table 5.1 when fixing r at 1, both of them imply assumptions that are reasonable in practical applications. Under constraint A, two bias-corrected estimators are available. One is derived using the Taylor expansion approach as given in Equation (5.7), and the other is obtained by leveraging the $Beta(1, 0)$ prior-based posterior mean of $p_{3|\bar{1}2}$. For constraint B, the Taylor expansion approach is less straightforward since ψ is a non-linear function of other estimable parameters. However, the bias-correction based on $Beta(1, 0)$ priors for the estimable parameters is straightforward and easily accessible. We note that there are saturated log-linear models that imply these two constraints, i.e., log-linear models including a number of predictors equal to the number of observed cell counts. These two saturated log-linear model can be found in Equations (D.1) and (D.2). For the log-linear model, the profile likelihood approach is often advocated in statistical lit-

erature for computing the confidence interval for N (Cormack, 1992; Gimenez et al., 2005). However, in practical CRC analyses for epidemiological studies, the confidence interval is commonly obtained by exponentiating a simple Wald-based interval for the intercept (Zhang and Small, 2020).

Table 5.2 presents the relative bias, the coverage probability, and the average width of 95% intervals for estimating N under constraint 1. It is clear that the log-linear model and the estimator in Equation (5.5) produced the same point estimates across the different simulation scenarios. Note that estimates obtained from both methods are biased upwards, with the bias decreasing as the sample size and/or the probability of being caught at least once increases. Compared with the biased estimator in Equation (5.5), the two bias-corrected estimators derived using the Taylor expansion approach and leveraging the $Beta(1, 0)$ prior significantly reduced the bias especially for moderate sample sizes (e.g., $N = 500$). In addition, these two bias correction approaches are almost identical in all cases.

From Table 5.2, we also observe that the coverage of the 95% credible intervals obtained from the proposed Dirichlet-based approach achieved or was close to the nominal level regardless of what estimators were used. In addition, the two common interval estimation approaches applied to the log-linear model also provided satisfactory coverage across all scenarios. However, the Dirichlet-based approach using bias-corrected estimators resulted in a reduction in the interval width under most scenarios.

Simulation results based on constraint 2 are shown in Table 5.3, with the equivalence between the estimator of N in Equation (5.5) and the log-linear model-based estimate again demonstrated. The bias-corrected estimator obtained by inserting posterior means of estimable parameters into Equation (5.5) demonstrably reduced the positive bias of the uncorrected estimator in Equation (5.5) across all scenarios. Credible intervals obtained in conjunction with this bias-corrected estimator yielded

reduced interval width, while maintaining appealing coverage compared to the other intervals.

To demonstrate the flexibility of the proposed modeling framework, we next carried out a simulation study under scenarios where no log-linear model is applicable. Specifically, we considered a scenario where referral exists in the three-catch case (Jones et al., 2014). The proposed modeling framework was implemented to accommodate this scenario by introducing a parameter to characterize the referral proportion and specifying a constraint that correctly implies the dependency structure between the three systems. For the purpose of comparison, we applied the log-linear modeling paradigm in which the AIC was used for selecting the model among candidate models. As expected, our proposed modeling framework provided valid estimates and excellent coverage, while the log-linear model failed to incorporate such scenarios as illustrated in the previous study (Table 5.4) (Jones et al., 2014).

We further conducted simulations to illustrate the benefits of the proposed strategy of imposing a well-considered assumption via linking estimable parameters to the inestimable parameter ψ. First, we generated data under the independence assumption for the three-catch case and fitted the proposed model assuming a less restrictive constraint where $\psi = p_{3|\bar{1}2}$ (note that while independence implies this constraint, the converse is not true). Under this latter constraint, we only focused on the ψ parameter and made no assumptions about the association between system 1 and system 2. As shown in Table 5.5, the coverage of credible intervals obtained under both the independence assumption and the assumption $\psi = p_{3|\bar{1}2}$ is satisfactory. As expected, wider intervals were observed when the less restrictive constraint was assumed. However, estimates obtained via this less restrictive constrain were empirically unbiased when the proposed bias-correction was implemented. On the other hand, when data were generated under the less restrictive constraint $\psi = p_{3|\bar{1}2}$, assuming independence yielded biased estimates and a severe under-coverage problem (Table 5.5). Thus, ap-

plication of the targeted constraint offers an accessible robust alternative to the full independence assumption.

To evaluate the performance of the uncertainty analysis, we considered a setting in which the untestable assumption about the dependency between system 1 and system 3 conditional on lack of capture by system 2 can be benchmarked by the testable assumption about the dependency between system 1 and system 3 conditional on capture by system 2. The corresponding constraint can be expressed as $\frac{p_{3|12}}{p_{3|\bar{1}2}} = r\frac{p_{3|1\bar{2}}}{\psi}$, where r is referred to as the key parameter determining the dependency across systems, and it can be a fixed value or follow a pre-specified parametric distribution to allow for uncertainty. Table 5.6 shows coverage and average widths of 95% credible intervals obtained from the proposed uncertainty analysis across different forms of r. We first demonstrated that the uncertainty analysis successfully incorporates both statistical uncertainties and the uncertainty of the assumption about dependency, as suggested by the coverage of 95% credible intervals obtained under scenarios where the distribution of r is correctly specified. When data were generated by allowing r to vary within a range (i.e., $r \sim \text{Uniform}(0.8, 1.2)$) mistakenly assuming a fixed value of r resulted in an expected under-coverage problem that is aggravated as N increases. However, the under-coverage problem alleviates when a distribution which is less variant (i.e., $N(1, 0.06^2)$ essentially covers the same range as the Uniform(0.8, 1.2) but is less variant) compared to the one used for generating data was introduced for r. We also showed that incorporating uncertainty for the key parameter r is beneficial relative to assuming an incorrect fixed value for r, with respect to providing credible intervals for N with better coverage. This is the key point of such uncertainty analyses, as previously outlined in the simpler two-catch setting discussed in Chapter 3 (Zhang et al., 2023a).

Table 5.2: Simulation results for estimating N under constraint 1, $p_{3|\bar{1}2} = \psi$: relative bias, 95% coverage, and width of intervals averaged across 1000 simulated datasets where p_c is the probability of being caught at least once.

		$p_c = 0.53$			$p_c = 0.64$		
N	Estimator [a]	Relative bias [b]	Coverage	Average width	Relative bias [b]	Coverage	Average width
500	Proposed (no BC)	0.061	93.5	406.5	0.014	93.7	175.1
	Proposed (BC)	0.004	92.1	314.6	0.002	93.8	165.7
	Proposed (BC beta prior)	0.003	91.7	311	0.002	93.5	165.3
	Log-linear (exp)	0.061	94.7	471.1	0.014	93.1	174.3
	Log-linear (prof)	-	95.2	416.0	-	93.6	179.7
1000	Proposed (no BC)	0.022	95.7	504.8	0.009	95.2	241.2
	Proposed (BC)	-0.001	94.4	454.0	0.003	95.5	234.7
	Proposed (BC beta prior)	-0.002	94.8	453.5	0.003	95.4	234.7
	Log-linear (exp)	0.022	95.4	519.2	0.009	94.6	235.3
	Log-linear (prof)	-	95.4	526.9	-	94.9	243.7
2000	Proposed (no BC)	0.011	94.4	676.1	0.004	94.3	334.5
	Proposed (BC)	0	94.5	647.1	0.002	94.4	330.7
	Proposed (BC beta prior)	0	94.1	646.1	0.002	94.1	330.2
	Log-linear (exp)	0.011	94.6	681.7	0.004	93.6	323.6
	Log-linear (prof)	-	94.6	688.2	-	94.3	336.4
5000	Proposed (no BC)	0.003	96.0	1035.8	0.003	94.3	524.5
	Proposed (BC)	-0.001	95.8	1017.2	0.002	94.4	521.8
	Proposed (BC beta prior)	-0.001	95.8	1018.1	0.002	94.2	522.3
	Log-linear (exp)	0.003	95.7	1031.4	0.003	93.1	504.2
	Log-linear (prof)	-	95.7	1038.6	-	93.6	519.3

[a] Proposed (no BC) = estimator in Equatio (5.5), Proposed (BC) = estimator in Equation (5.7), Proposed (BC beta prior) = estimator derived by leveraging Beta posterior for $p_{3|\bar{1}2}$, Log-linear (exp) = 95% Wald-type confidence interval obtained based on the estimated intercept, Log-linear (prof) = 95% confidence interval obtained using the profile likelihood approach. [b] Relative bias is computed as $\sum_i (\hat{N}_i - N)/1000$, where i is the simulation index.

Table 5.3: Simulation results for estimating N under constraint 2, $\frac{p_{3|12}/(1-p_{3|12})}{p_{3|\bar{1}2}/(1-p_{3|\bar{1}2})} = \frac{p_{3|1\bar{2}}/(1-p_{3|1\bar{2}})}{\psi/(1-\psi)}$: relative bias, 95% coverage, and width of intervals averaged across 1000 simulated datasets, where p_c is the probability of being caught at least once.

		$p_c = 0.55$			$p_c = 0.63$		
N	Estimator [a]	Relative bias [b]	Coverage	Average width	Relative bias [b]	Coverage	Average width
500	Proposed (no BC)	0.061	93.4	585.8	0.045	94.5	426.3
	Proposed (BC beta prior)	0.02	92.6	496.6	0.026	94.0	381.1
	Log-linear (exp)	0.061	93.8	630.0	0.045	94.8	444.3
	Log-linear (prof)	-	93.1	606.5	-	93.5	429.2
1000	Proposed (no BC)	0.03	94.4	730.7	0.016	94.8	529.3
	Proposed (BC beta prior)	0.012	94.2	679.1	0.009	94.5	505.4
	Log-linear (exp)	0.030	94.4	750.5	0.016	95.0	537.7
	Log-linear (prof)	-	94.2	738.9	-	94.5	529.2
2000	Proposed (no BC)	0.015	94.6	963.9	0.015	94.4	724.6
	Proposed (BC beta prior)	0.007	94.3	931.8	0.012	94.2	710.6
	Log-linear (exp)	0.015	94.9	974.2	0.015	94.5	726.8
	Log-linear (prof)	-	94.2	968.1	-	94.7	722.2
5000	Proposed (no BC)	0.008	94.2	1471.3	0.008	93.7	1101.5
	Proposed (BC beta prior)	0.005	94.3	1452.7	0.006	94.0	1095.0
	Log-linear (exp)	0.008	94.5	1472.2	0.008	93.4	1099.2
	Log-linear (prof)	-	94.6	1467.7	-	93.6	1097.4

[a] Proposed (no BC) = estimator in Equatio (5.5), Proposed (BC beta prior) = estimator derived by leveraging Beta posterior for $p_{3|12}$, $p_{3|\bar{1}2}$, and $p_{3|1\bar{2}}$, Log-linear (exp) = 95% Wald-type confidence interval obtained based on the estimated intercept, Log-linear (prof) = 95% confidence interval obtained using the profile likelihood approach. [b] Relative bias is computed as $\sum_i(\hat{N}_i - N)/1000$, where i is the simulation index.

Table 5.4: Simulation results for estimating N under referral scenarios where a proportion q of cases are referred from stream 1 to stream 3, and streams 1 and 3 are independent given capture status in stream 2: relative bias, 95% coverage, and width of intervals averaged across 1000 simulated datasets, where q is the proportion of cases that are referred from stream 1 to stream 3.

		$q = 0.1$			$q = 0.3$		
N	Estimator [a]	Relative bias [b]	Coverage	Average width	Relative bias [b]	Coverage	Average width
500	Proposed (no BC)	0.046	95.0	603.2	0.135	95.1	1286.9
	Log-linear (exp)	0.248	55.1	581.0	0.634	48.9	1022.3
	Log-linear (prof)	-	58.8	490.5	-	51.8	869.5
1000	Proposed (no BC)	0.020	95.1	471.9	0.037	95.4	1025.9
	Log-linear (exp)	0.203	58.8	769.7	0.466	35.1	1104.0
	Log-linear (prof)	-	62.1	765.4	-	38.0	1100.7
2000	Proposed (no BC)	0.008	95.5	518.0	0.017	94.9	873.7
	Log-linear (exp)	0.190	58.6	1064.0	0.404	10.6	1414.7
	Log-linear (prof)	-	61.4	1058.2	-	11.9	1412.1
5000	Proposed (no BC)	0.004	95.5	760.1	0.004	95.6	1066.4
	Log-linear (exp)	0.180	24.0	1614.5	0.382	0	2115.2
	Log-linear (prof)	-	25.3	1609.7	-	0	2110.3

† The log-linear model was selected based on the AIC for each simulation, candidate log-linear models follow the usual conventions that are no three-way interactions, and include all capture indicators for streams.

[a] Proposed (no BC) = estimator in Equatio (5.5), Log-linear (exp) = 95% Wald-type confidence interval obtained based on the estimated intercept, Log-linear (prof) = 95% confidence interval obtained using the profile likelihood approach.

[b] Relative bias is computed as $\sum_i (\hat{N}_i - N)/1000$, where i is the simulation index.

Table 5.5: Simulation results for estimating N when the dependency structure is misspecified: relative bias, 95% coverage, and width of intervals averaged across 1000 simulated datasets.

		Data generated under the independence assumption			Data generated under the assumption		
N	Assumption specified [a] in the fitted model	Relative [b] bias	Coverage	Average width	Relative [b] bias	Coverage	Average width
500	independence	0.013	94.4	179.0	-0.204	6.2	92.5
	$\psi = p_{3\|\bar{1}2}$	0.005	92.9	346.1	0.004	92.1	314.6
1000	independence	0.006	94.6	261.5	-0.206	0.2	132.6
	$\psi = p_{3\|\bar{1}2}$	-0.001	95.8	511.9	-0.001	94.4	454.0
2000	independence	0.003	93.7	374.4	-0.206	0.0	189.5
	$\psi = p_{3\|\bar{1}2}$	0.003	94.4	738.9	0.000	94.5	647.1
5000	independence	0.001	95.0	596.3	-0.207	0.0	300.1
	$\psi = p_{3\|\bar{1}2}$	-0.001	93.7	1153.5	-0.001	95.8	1017.2

[a] The bias-corrected estimator given in Equation (5.7) is used when the model assumes $\psi = p_{3|\bar{1}2}$.
[b] Relative bias is computed as $\sum_i (\hat{N}_i - N)/1000$, where i is the simulation index.

Table 5.6: Coverage and width of 95% credible intervals obtained from uncertainty analysis imposing the constraint $\frac{p_{3|12}}{p_{3|\bar{1}2}} = r\frac{p_{3|1\bar{2}}}{\psi}$; results were averaged across 1000 simulated datasets and bias-corrections leveraging $Beta(1,0)$ priors were implemented.

		Data generated assuming $r \sim \text{Uniform}(0.8, 1.2)$		Data generated assuming $r = 0.9$		Data generated assuming $r = 1.1$	
N	Assumption	Coverage	Average width	Coverage	Average width	Coverage	Average width
1000	$r = 1$	90.9	835.8	92.8	514.0	93.8	625.5
	$r \sim N(1, 0.06^2)$	92.0	861.9	92.9	532.2	93.4	643.8
	$r \sim \text{Uniform}(0.8, 1.2)$	95.1	952.0	96.1	602.7	95.9	731.1
2000	$r = 1$	88.8	1163.0	88.5	712.4	93.5	873.0
	$r \sim N(1, 0.06^2)$	88.5	1197.7	90.1	751.1	94.6	922.0
	$r \sim \text{Uniform}(0.8, 1.2)$	94.4	1409.4	96.3	912.1	97.5	1122.6
5000	$r = 1$	81.2	1813.1	83.8	1124.4	90.8	1357.0
	$r \sim N(1, 0.06^2)$	84.7	1970.3	89.3	1271.6	93.0	1515.5
	$r \sim \text{Uniform}(0.8, 1.2)$	94.2	2620.3	97.9	1774.4	97.5	2127.2
10000	$r = 1$	68.9	2547.7	75.5	1591.7	79.9	1932.7
	$r \sim N(1, 0.06^2)$	75.2	2984.5	87.1	1944.3	89.1	2376.2
	$r \sim \text{Uniform}(0.8, 1.2)$	93.7	4421.7	99.3	3030.1	98.7	3722.5
20000	$r = 1$	55.2	3643.2	60.0	2240.8	67.0	2707.1
	$r \sim N(1, 0.06^2)$	71.7	4769.1	84.9	3174.7	88.2	3847.7
	$r \sim \text{Uniform}(0.8, 1.2)$	95.2	7906.7	100	5512.1	99.7	6673.2

5.4 Real Data Applications

We applied our proposed modeling framework to two real CRC datasets which have been analyzed before using the log-linear modeling paradigm (Abeni et al., 1994; Poorolajal et al., 2017). Both datasets were collected for surveying the number of HIV infections in a particular area; the first real data application involves three data streams, while the second contains four data streams.

5.4.1 Three-stream HIV CRC data

The CRC data for this example relate to HIV-positive patients identified by three data streams: stream 1 (transfusion center), stream 2 (volunteer counseling and testing centers), and stream 3 (prison) in Iran in 2016. In the original analysis, a final estimate was obtained from a log-linear model including all three capture indicators for separate data streams, a two-way interaction term between capture indicators for streams 1 and 2, and a two-way interaction term between capture indicators for streams 2 and 3 (Poorolajal et al., 2017). As a result, the total number of HIV positive cases (N) was estimated to be 17,149 with 95% Wald-type confidence interval (CI) (11,979 - 25,078). One can show that the log-linear model selected in the previous study imposes two constraints expressed as $p_{3|12} = p_{3|\bar{1}2}$ and $p_{3|1\bar{2}} = \psi$ using the parameterization adopted in our proposed modeling framework. This model was selected based on AIC considerations, following common practice. However, prior work clearly demonstrates that AIC is not a defensible tool for model selection in the CRC context. A fact hidden from the typical user is that the log-linear model framework led to selecting this model because the testable constraint (i.e., $p_{3|12} = p_{3|\bar{1}2}$) was largely consistent with the observed data. A key issue is that the crucial untestable constraint ($p_{3|1\bar{2}} = \psi$) does not follow logically from the testable constraint but rather is projected as a mathematical construct under the exclusionary log-linear

framework (Zhang et al., 2023b).

In the proposed modeling framework, we recommend the use of the estimator in Equation (5.5) that utilizes observed counts and incorporates the constraint which relates the inestimable parameter ψ to other estimable parameters. Assuming the same untestable constraint as the selected log-linear model (i.e., $g(\boldsymbol{\theta}^*) = p_{3|1\bar{2}}/r$ and $r = 1$) and applying the bias-correction, the proposed model resulted in the estimate of 16,530 (95% CI: 11,644 – 22,933). Since the constraint imposed to enable the estimation of N cannot be verified by the observed data alone, we also applied the proposed uncertainty analysis in which r was assumed to follow the Uniform(0.8, 1.2), the resulting 95% CI was (11,016 – 25,385). We note that application of the proposed modeling framework with an investigator choosing the above constraint with $r = 1$ yields interval estimates similar to those reported by the original authors.

Besides the dependency specified via the selected log-linear model, we explored other possible dependency structures between data streams that the investigator could assume using the proposed model. The resulting estimates are shown in Figure 5.1. We note that all estimates reported under the proposed model framework are bias-corrected using the procedure relying on $Beta(1, 0)$ priors. We first considered the dependency structure specified by defining $g(\boldsymbol{\theta}^*) = p_{3|\bar{1}2}/r$. Specifically, we assumed r equal to 2, which implies a positive association between streams 2 and 3 among cases not identified by stream 1. Such a positivity assumption could be reasonable in this setting, as both streams 2 and 3 are volunteer-based (Poorolajal et al., 2017). To allow for uncertainty in this assumption, the Uniform(1.6, 2.4) distribution was also introduced to r. It is clear that the estimates of N obtained under these assumptions are much smaller than the estimate from the selected log-linear model.

We also explored other constraints which imply positive associations between streams 2 and 3. Specifically, we anchored the association between streams 2 and 3 among cases that were not identified by stream 1 using the association between

streams 2 and 3 among cases that were identified by stream 1. The strength of the latter association can be reflected using different quantities that can be estimated using the observed data. For example, we used relative risks (RRs) $p_{3|12}/p_{3|1\bar{2}}$ and $p_{3|\bar{1}2}/\psi$ to measure those two associations. In this real data example, the observed data indicated that there is a positive association between streams 2 and 3 among cases captured by stream 1. The constraint to enable the estimation is expressed as $\frac{p_{3|12}}{p_{3|1\bar{2}}} = r\frac{p_{3|\bar{1}2}}{\psi}$, where r determines the strength of the association between streams 2 and 3 among cases not identified by stream 1. The resulting estimates are given in Figure 5.1 and labeled as RR = 1 or 2 and $RR \sim$ Uniform(0.8, 1.2) or Uniform(1.6, 2.4), indicating $r = 1$ or 2, and $r \sim$ Uniform(0.8, 1.2) or Uniform(1.6, 2.4). Lastly, we considered the constraint $\frac{p_{3|12}/(1-p_{3|12})}{p_{3|\bar{1}2}/(1-p_{3|\bar{1}2})} = r\frac{p_{3|1\bar{2}}/(1-p_{3|1\bar{2}})}{\psi/(1-\psi)}$, which is referred to as an odd-ratio (OR) type constraint. Under this constraint, the association between streams 2 and 3 among cases identified by stream 1 was measured using the odd ratio and was estimated from the real data. The estimates along with 95% CIs are provided in Figure 5.1 when assuming $r = 2$ and r follows Uniform(1.6, 2.4). This point estimate (i.e., assuming OR=2) is similar to the one obtained using the constraint defined via the relative risk (i.e., RR=2).

5.4.2 Four-stream HIV CRC data

This four-stream HIV data were collected from four sites implemented for testing HIV-1 infected patients in Lazio, Italy during 1990 (Abeni et al., 1994). The prior authors reported an estimate using a log-linear model containing four capture indicators for each data source and the two-way interaction between indicators for streams 3 and 4. One can show that this log-linear model implies the following constraints on the inestimable parameter ψ: $p_{4|12\bar{3}} = p_{4|\bar{1}2\bar{3}} = p_{4|1\bar{2}\bar{3}} = \psi$. We computed point estimates, corresponding 95% CIs, and 95% CIs allowing uncertainties around the assumptions under these three separate constraints (Figure 5.2). As recommended under the

proposed framework, we only focused on the constraint associated with ψ and did not impose testable constraints (i.e., $p_{4|12\bar{3}} = p_{4|\bar{1}2\bar{3}} = p_{4|1\bar{2}\bar{3}}$). The two constraints ($p_{4|\bar{1}2\bar{3}} = \psi$ and $p_{4|1\bar{2}\bar{3}} = \psi$) resulted in estimates which fall into the same range as the estimate obtained from the log-linear model selected by the original authors. However, note for example that the $p_{4|12\bar{3}} = \psi$ assumption produces an estimate that is much smaller. The previous authors stated that there are cases referred from stream 3 to stream 4. Thus, it is reasonable to speculate that streams 3 and 4 were positively associated. To characterize this positive association, we considered three constraints: (a) $\psi = p_{4|1\bar{2}3}/r$, (b) $\frac{p_{4|\bar{1}23}}{p_{4|\bar{1}2\bar{3}}} = r_1 \frac{p_{4|\bar{1}\bar{2}3}}{\psi}$, and (c) $\frac{p_{4|1\bar{2}3}}{p_{4|1\bar{2}\bar{3}}} = r_2 \frac{p_{4|\bar{1}\bar{2}3}}{\psi}$. Constraint (a) relates the inestimable ψ to the estimable parameter $p_{4|1\bar{2}3}$ yields an estimated N of 6329 (95% CI: 3412 – 13,786) when assuming $r = 2$. As shown in Figure 5.2, the estimate under constraint (b) specifying r_1 equal to 2 coincided with the estimate from the selected log-linear model. However, they imply different dependency structures between data streams. When assuming $r_2 = 2$, constraint (c) resulted in an estimate much larger than the others.

5.5 Discussion

This Chapter presents an accessible modeling framework based on the population-level multinomial model, with interpretable conditional probabilities introduced to characterize the probability of different capture histories. By focusing on the key inestimable parameter ψ (i.e., the probability of being captured by the last data stream given not identified by any other data stream), the proposed framework allows the user to flexibly incorporate expert opinion about dependencies among data streams to guide the estimation of the number of diseased cases. Unless under a careful design to ensure the independence between data streams (Lyles et al., 2021b), we believe relying on expert opinion is in fact the most promising way to specify the depen-

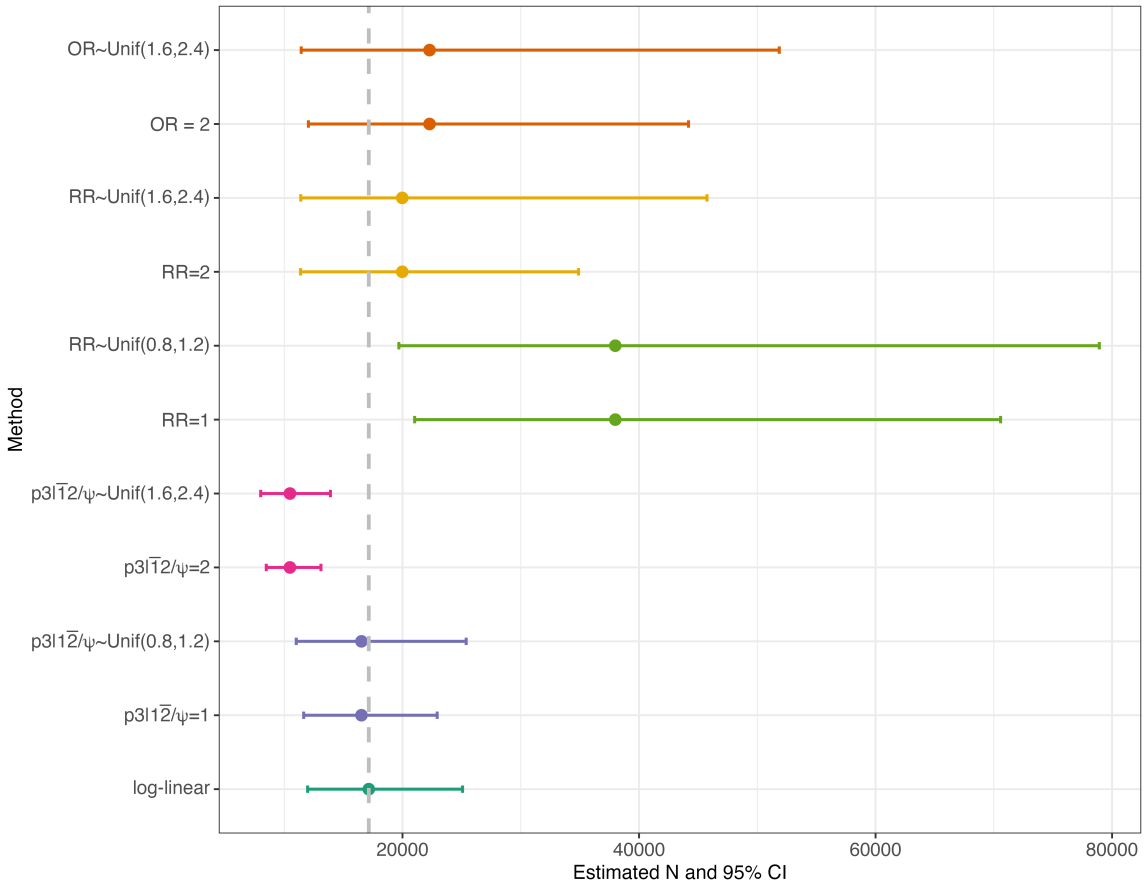


Figure 5.1: Estimates and 95% credible intervals of N obtained from applying the proposed modeling framework while imposing different constraints to three-catch HIV CRC data collected in Iran in 2016. The dashed grey line marks the point estimate reported by Poorolajal et al. (2017), who analyzed the same based on a log-linear model. RR denotes the relative risk-type constraint $\frac{p_{3|12}}{p_{3|1\bar{2}}} = r\frac{p_{3|1\bar{2}}}{\psi}$, e.g., RR=2 indicates $r = 2$. OR denotes the odd ratio-type constraint $\frac{p_{3|12}/(1-p_{3|12})}{p_{3|\bar{1}2}/(1-p_{3|\bar{1}2})} = r\frac{p_{3|1\bar{2}}/(1-p_{3|1\bar{2}})}{\psi/(1-\psi)}$, e.g., OR $\sim$ Unif(1.6, 2.4) indicates $r \sim$ Uniform(1.6, 2.4).

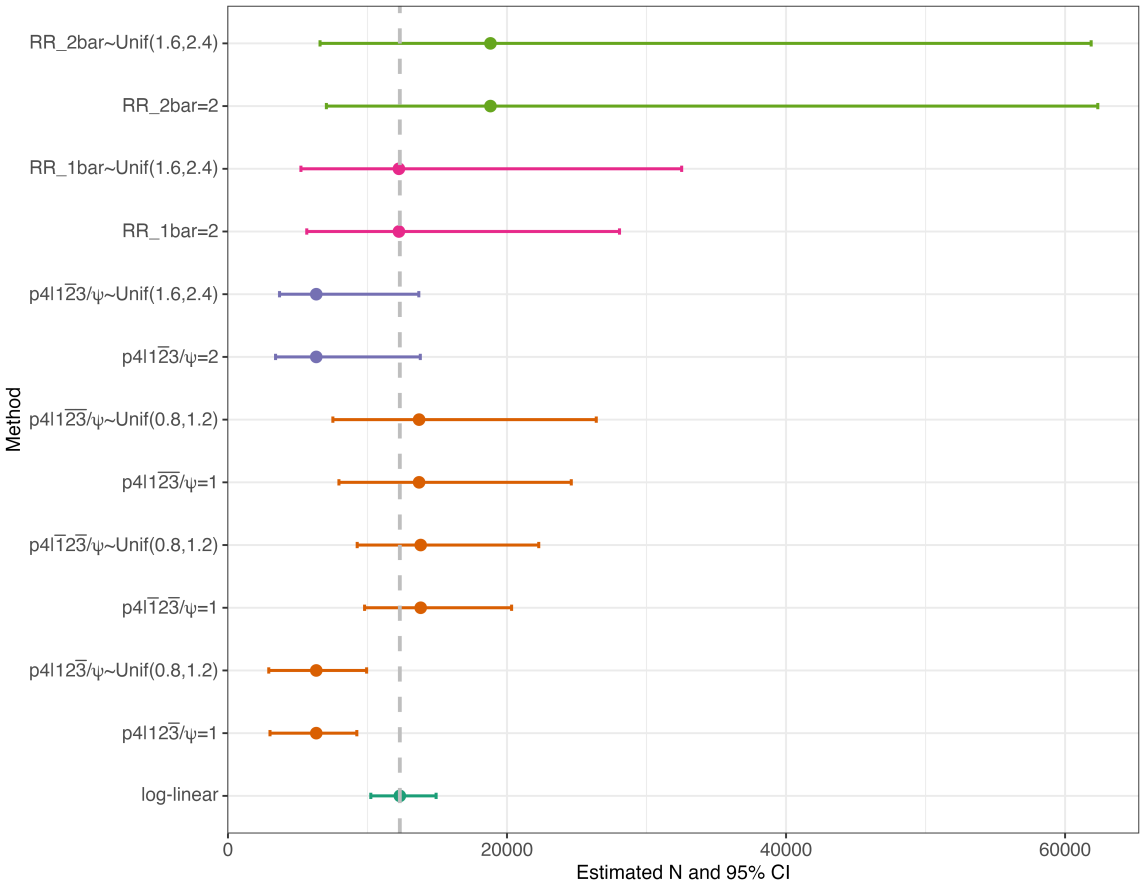


Figure 5.2: Estimates and 95% credible intervals of N obtained from applying the proposed modeling framework while imposing different constraints to four-catch HIV CRC data collected in Lazio, Italy during 1990. The dashed grey line marks the point estimate reported by Abeni et al. (1994), who analyzed the same data using a log-linear model. RR_1bar denotes the relative risk-type constraint $\frac{p_{4|1\bar{2}3}}{p_{4|\bar{1}2\bar{3}}} = r_1 \frac{p_{4|\bar{1}\bar{2}3}}{\psi}$, e.g., RR_1bar=2 indicates $r_1 = 2$. RR_2bar denotes the relative risk-type constraint $\frac{p_{4|1\bar{2}3}}{p_{4|1\bar{2}\bar{3}}} = r_2 \frac{p_{4|\bar{1}\bar{2}3}}{\psi}$, e.g., RR_2bar$\sim$Unif(1.6, 2.4) indicates $r_2 \sim$ Uniform(1.6, 2.4).

dency assumption which drives the estimation. Since we recommend imposing the dependency assumption by relating the key parameter to other estimable conditional probabilities, carefully determining how to label data streams can greatly facilitate the estimation. For example, when considering a scenario where data streams A, B, and C are implemented, it is better to label stream A or B as the last data stream when one has confidence to speculate that streams A and B are positively associated.

Besides assigning labels to data streams, choosing constraints (i.e., determining $g(\boldsymbol{\theta}^*)$) is also a critical step in the proposed modeling framework. We provide an overview of practical constraints in Table 5.1 for three-catch cases, and implemented those constraints in simulation studies and the real data application. As described in Table 5.1, the odd ratio-type constraint (i.e., constraint 5 in Table 5.1) as well as the relative risk-type constraint (i.e., constraint 4 in Table 5.1) anchor the inestimable association using estimable associations, while two constraints adopt different measurements of associations. We observed that the odd ratio-type constraint yields the same estimation regardless of the labeling when the value of r is specified, this conclusion can be easily demonstrated theoretically. Compared to the use of odds ratio-type constraint, using the relative risk-type constraint more flexibly incorporates the dependency assumption since this constraint allows leverage expert opinions to determine which stream should be labeled as the last data stream. Additionally, in the real data application using three-catch data, we observed that estimates obtained from imposing the odds ratio type constraint were generally associated with larger standard errors compared to estimates based on the relative risk-type constraint (e.g., Figure 5.1). Thus, we recommend the use of relative risk-type constraint over the odds ratio-type constraint when analyzing three-catch data.

We proposed a straightforward bias-correction procedure leveraging $Beta(1, 0)$ posterior means of the estimable conditional probabilities to reduce the mean bias. This can be viewed in spirit as a significant generalization of classical bias corrections

under stream independence assumptions (e.g., Chapman (1951)). In conjunction with bias-corrected estimators, we find that our proposed adapted credible interval approach produces intervals of smaller length, while maintaining favorable coverage compared to direct use of the unadjusted estimator. The uncertainty analysis accomplished based on the Dirichlet-multinomial-based approach allows one to characterize and readily accommodate uncertainties in the dependency assumption. Through simulation studies, we have shown that such uncertainty analysis augments robustness under the proposed modeling framework, in light of virtually unavoidable mis-specification of the true dependency assumption.

Compared to the log-linear modeling framework which has been demonstrated to exclude potentially large swaths of feasible estimates by design (Zhang et al., 2023b), the proposed model anchoring on the estimator in Equation (5.5) makes all possible estimates attainable. Stratified CRC analysis are often desired to provide the stratum-specific estimate of the diseased counts (Héraud-Bousquet et al., 2012). The proposed modeling framework can be easily extended to incorporate categorical covariates by applying the modeling framework separately within each stratum formed by those categorical covariates. The estimated total number of cases is then obtained by summing up estimates across strata. Compared to incorporating categorical covariates in the log-linear modeling framework, the natural of the credible intervals approach greatly facilitates the interval estimation for the total number of cases. Specifically, the corresponding interval is computed by pooling posterior samples of the case counts for each stratum. Since the proposed modeling framework focusing on population-level CRC data, it is less straightforward to allow the inclusion of continuous covariates. One possible approach may be to model the estimable probabilities used for specifying the constraint using logit models, and then apply the estimator in Equation (5.5) while integrating over possible values of ψ.

Chapter 6

Summary and Future Work

6.1 Summary

This dissertation focuses on developing statistical methods for analyzing disease surveillance data collected from multiple disease surveillance systems to improve the estimation of the number of diseased cases.

In Chapter 2, a hierarchical model has been proposed under the full Bayesian modeling framework to analyze individual-level surveillance data collected from multiple surveillance systems over multiple surveillance sites. The proposed model permits the consideration of individual-specific heterogeneous capture probabilities and borrows information across surveillance sites in an unified modeling framework. Compared to methods which solely rely on individual-level surveillance data collected at surveillance sites where overlaps between surveillance systems exist, the proposed model also allows one to utilize surveillance data from locations where only one surveillance system operates. Simulation studies have demonstrated the improvement in the estimation when applying the proposed method compared to models cannot incorporate the additional information provided at locations where only one system is in operation.

An accessible sensitivity and uncertainty analysis framework has been introduced in Chapter 3 with focus on the two-catch CRC data. The proposed sensitivity and uncertainty analysis anchors on a key inestimable parameter which is interpretable and measures the dependency between systems using the population-level multinomial model. By treating the key interpretable dependency parameter as the sensitivity parameter, the proposed sensitivity analysis provides an appealing data visualization to explore how the key parameter impacts the estimation of case counts. The simulation-based uncertainty analysis provides an interval estimation incorporating both statistical uncertainties in estimating the case counts and the uncertainty about the dependency assumption made to enable the estimation.

In Chapter 4, we clarified two main pitfalls associated with the commonly used

log-linear modeling paradigm to provide the estimation of case counts in CRC contexts. We showed that the log-linear model excludes many possible estimates by design, and the regularly used selection metrics are fundamentally flawed and deceiving when analyzing CRC data. To circumvent these pitfalls, in Chapter 5, we proposed a modeling framework that serves as an alternative to the log-linear model for analyzing CRC data in epidemiological studies. The proposed model extends the multinomial distribution-based likelihood approach adopted in Chapter 3 to incorporate setting where multiple surveillance systems are implemented (≥ 2), which theoretically permits all possible estimates. Instead of applying model selection metrics to determine the untestable dependency assumption, the proposed model leverages expert opinion in spirit of prior information to guide the estimation. The uncertainty analysis introduced in Chapter 3 further inspires the implementation of a principled simulation-based uncertainty analysis to acknowledge the confidence level of the dependency assumption made based on expert opinion in the proposed alternative modeling framework.

6.2 Future Work

The PTB data analyzed in Chapter 2 contain CRC data collected over different years. In the proposed Bayesian hierarchical model, we focus on the individual-level CRC data and allow the individual-level capture probabilities to vary by year. The PTB data can also be summarized to form year-specific population-level CRC data. This type of data motivates us to extend the modeling framework proposed in Chapter 5 to allow borrowing information across years when estimating the year-specific case counts. We may consider to adopt a hierarchical model which allows the use of year-specific conditional probabilities to characterize capture probabilities of different capture histories in each year while assuming the dependency structure between data

structure is the same across different years.

In this dissertation, the disease status is always assumed to be perfectly identified in all implemented surveillance systems. However, in practice, the disease status is potentially misclassified. This warrants future developments that aim to incorporate misclassification. For the Bayesian hierarchical model developed in Chapter 2, one can consider to introduce latent variables to represent the true disease status. Motivated by recent developments under specific CRC study designs (Ge et al., 2023), we can extend the modeling framework developed in Chapter 5 to account for errors in determining disease status using positive predictive value parameters.

Appendix A

Appendix for Chapter 2

A.1 Posterior predictive simulation procedure for generating imputed dataset

Each imputed data is generated based on following steps:

- Linked sites, for $s = 1, \ldots, K$

    1. For each linked site, sample $\hat{N}_s$ individuals from the individual-level capture-recapture data with replacement; the corresponding covariates are obtained by randomly shuffling the original covariates for $s = 1, \ldots, K$.
    2. Unconditional capture probability of each sampled individual for each system (passive and active) is computed as $\frac{\exp(\boldsymbol{x}_{ti}^T \hat{\boldsymbol{\beta}}_t)}{1+\exp(\boldsymbol{x}_{ti}^T \hat{\boldsymbol{\beta}}_t)}$, for $t = 1, 2$, $i = 1, \ldots, \hat{N}_s$, $s = 1, \ldots, K$, where $\boldsymbol{x}_{ti}$ denotes covariates vector.
    3. Draw two ascertainment histories for each sample individuals from Bernoulli distributions with mean equal to unconditional capture probabilities obtained in step 2.
    4. The imputed capture-recapture data is then formed by removing individuals who have capture history (0,0).

- $S - K$ unlinked sites, for $s = K + 1, \ldots, S$

    1. For each unlinked site, sample $\hat{N}_s$ individuals from data collected only by passive surveillance system with replacement; the corresponding covariates are obtained by randomly shuffling the original covariates for $s = K + 1, \ldots, S$.
    2. Unconditional capture probability of each sampled individuals for the passive system is computed as $\frac{\exp(\boldsymbol{x}_{1i}^T \hat{\boldsymbol{\beta}}_1)}{1+\exp(\boldsymbol{x}_{1i}^T \hat{\boldsymbol{\beta}}_1)}$, $i = 1, \ldots, \hat{N}_s$, $s = K + 1, \ldots, S$, where $\boldsymbol{x}_{1i}$ denotes covariates vector.
    3. Draw one ascertainment history for each sample individuals from Bernoulli distribution with mean equal to unconditional capture probability obtained in step 2.
    4. Combine the ascertainment history for the passive system of individuals sampled at all S sites and remove individuals with ascertainment history 0 to create the imputed data formed by passive system.

A total M imputed datasets are generated based on the above steps. In simulation studies, we have found that $M = 20$ is sufficient to obtain valid 95% credible interval coverage.

A.2 Data generation procedure

A.2.1 Two systems are independent at the population level

For every combination of parameters (β_{10}, β_{11}, β_{20}, β_{21}, α, and ϕ) presented in Table A.1, we generate 100 datasets and fix N and location-specific N_s. With given parameters, each dataset is simulated as:

1. 50 sites are randomly selected from a 100×100 square area. Randomly select 10 of them as linked sites.

2. Site-specific N_s is generated from a Poisson distribution with mean λ_s generated as $\log(\lambda_s) = \alpha + \epsilon_s$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_{50})^T \sim MVN(\mathbf{0}, \Sigma)$, $\Sigma = \sigma^2 \exp(\frac{-D}{\phi})$, $\sigma^2 = 0.2$ and D is the Euclidean distance between sites.

3. Capture history of each individual is generated with

$$\text{logit}(p_{1i}) = \beta_{10} + \beta_{11} X_{1i},\ \text{logit}(p_{2i}) = \beta_{20} + \beta_{21} X_{2i},$$

where $X_{1i} \sim \text{Uniform}(-1, 1)$ and $X_{2i} \sim \text{Uniform}(-2, 2)$.

A.2.2 Two systems are positively correlated at the population level

For every combination of parameters (β_{10}, β_{11}, β_{20}, β_{21}, α, and ϕ) presented in , we generate 100 datasets and fix N and location-specific N_s. With given parameters and set $\sigma^2 = 0.1$ across all scenarios, each dataset is simulated by following procedures outlined in A.2.1 with modifying the step 3 as:

3. Capture history of each individual is generated with

$$\text{logit}(p_{1i}) = \beta_{10} + \beta_{11} X_i,\ \text{logit}(p_{2i}) = \beta_{20} + \beta_{21} X_i,$$

where $X_i \sim \text{Uniform}(0, 1.5)$.

Because X_i presents in both logit models and β_{11} and β_{21} are greater than zero, two systems are positively correlated at the population-level.

A.2.3 Multiple active systems are included

Let S1 to denote the passive system, S2 and S3 to denote two active systems. Data used in the simulation study involving three systems (i.e., the passive system is linked

to two active systems) are simulated as:

1. 100 sites are randomly selected from a 200×200 square area. Randomly select 30 of them as linked sites, of which 10 sites are S1-S2 linked sites and the other 20 sites are S1-S3 linked sites.

2. Site-specific N_s is generated from a Poisson distribution with mean λ_s generated as $\log(\lambda_s) = \alpha + \epsilon_s$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_{100})^T \sim MVN(\mathbf{0}, \Sigma)$, $\Sigma = \sigma^2 \exp(\frac{-D}{\phi})$, where $\alpha = 4.6$, $\sigma^2 = 0.1$, $\phi = 30$, and D is the Euclidean distance between sites.

3. Capture history of each individual is generated with

$$\text{logit}(p_{1i}) = \beta_{10} + \beta_{11} X_i,\ \text{logit}(p_{2i}) = \beta_{20} + \beta_{21} X_i,\ \text{logit}(p_{3i}) = \beta_{30} + \beta_{31} X_i$$

   where $\beta_{10} = -1.75,\ \beta_{11} = 1,\ \beta_{20} = -1.75,\ \beta_{21} = 1.5,\ \beta_{30} = -1.85,\ \beta_{31} = 1.5,\ X_i \sim \text{Uniform}(0, 1.5)$.

4. The observed S1-S2 linked data consist capture histories of individuals captured by S1 or S2 at the 10 S1-S2 linked sites. Similarly, the observed S1-S3 linked data consist capture histories of individuals captured by S1 or S3 at the 20 S1-S3 linked sites. Finally, only capture histories from S1 are recorded for those 70 unlinked sites.

A.3 Estimation and inference for two independent two-stream CRC data

We follow notations and convention of labeling systems used in Chapter 2. Compare to the regular situation introduced in Section 2.3 in Chapter 2, we now have linked sites for S1-S2 and S1-S3 linked data separately and individual-level capture histories from all three systems. Suppose S1 operates at S sites, K_{12} sites have S1-S2 linked

Table A.1: Coefficients of models used for generating simulated datasets

	two systems are independent at the population level						two systems are positively correlated at the population level					
Scenario	β_{10}	β_{11}	β_{20}	β_{21}	α	ϕ	β_{10}	β_{11}	β_{20}	β_{21}	α	ϕ
1	-1	1	-1	1	3.3	10	-1.75	1	-1.75	1.5	3.9	10
2	-0.5	1	-0.5	1	3.3	10	-1.5	1	-1.5	1.5	3.9	10
3	-1	1	-1	1	3.3	30	-1.75	1	-1.75	1.5	3.9	30
4	-0.5	1	-0.5	1	3.3	30	-1.5	1	-1.5	1.5	3.9	30
5	-1	1	-1	1	4.3	10	-1.75	1	-1.75	1.5	4.6	10
6	-0.5	1	-0.5	1	4.3	10	-1.5	1	-1.5	1.5	4.6	10
7	-1	1	-1	1	4.3	30	-1.75	1	-1.75	1.5	4.6	30
8	-0.5	1	-0.5	1	4.3	30	-1.5	1	-1.5	1.5	4.6	30

data, and K_{13} sites have S1-S3 linked data. As the model presented in Section 2.3, system-specific capture probabilities are modeled via logit regression models taking following forms:

$$\log\left\{\frac{p_{1is}(\boldsymbol{\beta}_1)}{1-p_{1is}(\boldsymbol{\beta}_1)}\right\} = \boldsymbol{x}_{1is}^T\boldsymbol{\beta}_1,\ \log\left\{\frac{p_{1is}(\boldsymbol{\beta}_2)}{1-p_{2is}(\boldsymbol{\beta}_2)}\right\} = \boldsymbol{x}_{2is}^T\boldsymbol{\beta}_2,\ \log\left\{\frac{p_{3is}(\boldsymbol{\beta}_3)}{1-p_{3is}(\boldsymbol{\beta}_3)}\right\} = \boldsymbol{x}_{3is}^T\boldsymbol{\beta}_3,$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ are vectors of coefficients of the BM model with corresponding covariates $\boldsymbol{x}_{1is}$, $\boldsymbol{x}_{2is}$, and $\boldsymbol{x}_{3is}$. To estimate $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$, we first write out the data likelihood contribution from S1-S2 linked data at K_{12} linked sites as:

$$L_{12}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2;\{\boldsymbol{y}_{1,s},\boldsymbol{y}_{2,s}\}_{s=1}^{K_{12}}\right) = \prod_{s=1}^{K_{12}}\prod_{i=1}^{n_s}\frac{p_{1is}^{y_{1is}}(1-p_{1is})^{(1-y_{1is})}p_{2is}^{y_{2is}}(1-p_{2is})^{(1-y_{2is})}}{1-(1-p_{1is})(1-p_{2is})},$$

where $\boldsymbol{y}_{1,s} = \{y_{1is} : i = 1,\ldots,n_s\}$ and $\boldsymbol{y}_{2,s} = \{y_{2is} : i = 1,\ldots,n_s\}$ are vector of capture histories in S1-S2 linked data at site s for S1 and S2 respectively, and n_s are number of unique cases captured by S1 or S2 at site s. Similarly, the data likelihood contribution from S1-S3 linked data is:

$$L_{13}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_3;\{\boldsymbol{y}^*_{1,s},\boldsymbol{y}_{3,s}\}_{s=1}^{K_{13}}\right) = \prod_{s=1}^{K_{13}}\prod_{i=1}^{n_s}\frac{p_{1is}^{y^*_{1is}}(1-p_{1is})^{(1-y^*_{1is})}p_{3is}^{y_{3is}}(1-p_{3is})^{(1-y_{3is})}}{1-(1-p_{1is})(1-p_{3is})},$$

where $\boldsymbol{y}^*_{1,s} = \{y^*_{1is} : i = 1,\ldots,n_s\}$ and $\boldsymbol{y}_{3,s} = \{y_{3is} : i = 1,\ldots,n_s\}$ are vector of capture histories in S1-S3 linked data at site s for S1 and S3 respectively, and n_s are number of

unique cases captured by S1 or S3 at site s. Since S2 and S3 do not have overlap, the data likelihood contribution from S1-S2 and S1-S3 linked data is given by

$$L\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3; \{\boldsymbol{y}_{1,s}, \boldsymbol{y}_{2,s}\}_{s=1}^{K_{12}}, \{\boldsymbol{y}_{1,s}, \boldsymbol{y}_{3,s}\}_{s=1}^{K_{13}}\right) = L_{12} \times L_{13}.$$

These coefficients can be sampled using Metropolis-Hastings (M-H) algorithms. Analogous to Equation (2.5) in Section 2.3, the Horvitz-Thompson (H-T) estimator of N_s is given by

$$N_s^* = \sum_{i=1}^{n_s} 1/q_{is}, \quad q_{is} := \begin{cases} 1-(1-p_{1is})(1-p_{2is}) & \text{for } s \text{ belongs to S1-S2 linked sites} \\ 1-(1-p_{1is})(1-p_{3is}) & \text{for } s \text{ belongs to S1-S3 linked sites} \\ p_{1is} & \text{for } s \text{ belongs to unlinked sites} \end{cases}$$

Once we obtain estimates of N_s using the H-T estimator, the second-stage estimation can be implemented by following Section 2.3.4.

A.4 Goodness of fit of the proposed model for analyzing PTB data

The posterior predictive distributions of quantities summarising observed linked data (n_{S_1}, n_{S_2}, $n^*_{S_1}$, and n_{S_3}) are generated as:

1. Draw 10,000 samples from posterior samples of coefficients included in logit models for modeling system-specific capture probabilities.

2. For each sample of coefficients, compute individual-level capture probabilities for S1, S2, and S3 for individuals caught at least once at linked sites. These capture probabilities are unconditional capture probabilities.

3. For each individual caught at least once, compute probability of being caught at least once at S1-S2 linked sites and S1-S3 linked sites, separately.

4. Divide computed unconditional capture probabilities by the corresponding probability of being caught at least once to obtain conditional capture probabilities (i.e.,

the probability of being caught given being caught at least once). Finally, for each individual, we have 10,000 samples of the conditional probability.

5. For each computed conditional probability, we draw one capture indicator from the Bernoulli distribution with mean equal to that conditional probability. For example, for individual i captured at S1-S2 linked site s, capture indicators are simulated for S1 and S2 from $Bernoulli(p_{1is}^{*(j)})$ and $Bernoulli(p_{2i1}^{*(j)})$, for $j = 1, \ldots, 10,000$, where $p_{1is}^{*(j)}$ and $p_{2i1}^{*(j)}$ are conditional probabilities computed using j-th sample of coefficients. Specifically, $p_{1is}^{*(j)} = \frac{p_{1is}^{(j)}}{1-(1-p_{1is}^{(j)})(1-p_{2is}^{(j)})}$, $p_{1is}^{(j)}$ is unconditional probability computed by taking inverse logit transformation to sampled coefficients, similarly for $p_{2is}^{*(j)}$. Then, 10,000 sets of simulated capture histories are obtained at all linked sites. Finally, predictive posterior distributions of quantities of interest are obtained using those simulated capture histories.

The posterior predictive distributions of quantities summarising observed data from S1 alone (i.e., number of county-specific unique cases captured by S1) are generated as:

1. Draw 100 samples from posterior samples of a vector of true county-specific number of cases $\boldsymbol{N} = (N_1, \ldots, N_{181})^T$ obtained at the end of second-stage estimation.

2. For each $\boldsymbol{N}$, we impute $N_s - n_s^c$ number of individuals for $s = 1, \ldots, 181$, their covariates are randomly drew from observed covariates with replacement, where n_s^c is unique number of cases captured at site s. Then for those N_s individuals, we draw 100 samples of coefficients associated with capture probabilities of S1 for each and compute unconditional probabilities of being captured by S1. These probabilities are used to generate capture indicators of S1 for N_s individuals. Combine simulated capture indicators over 100 sampled vector of $\boldsymbol{N}$, we have in total of 10,000 sets of capture indicators. Finally, predictive posterior distributions of $\boldsymbol{N}$ are obtained using those simulated capture indicators of S1.

Posterior median and 95% posterior intervals are computed from those generated predictive distributions.

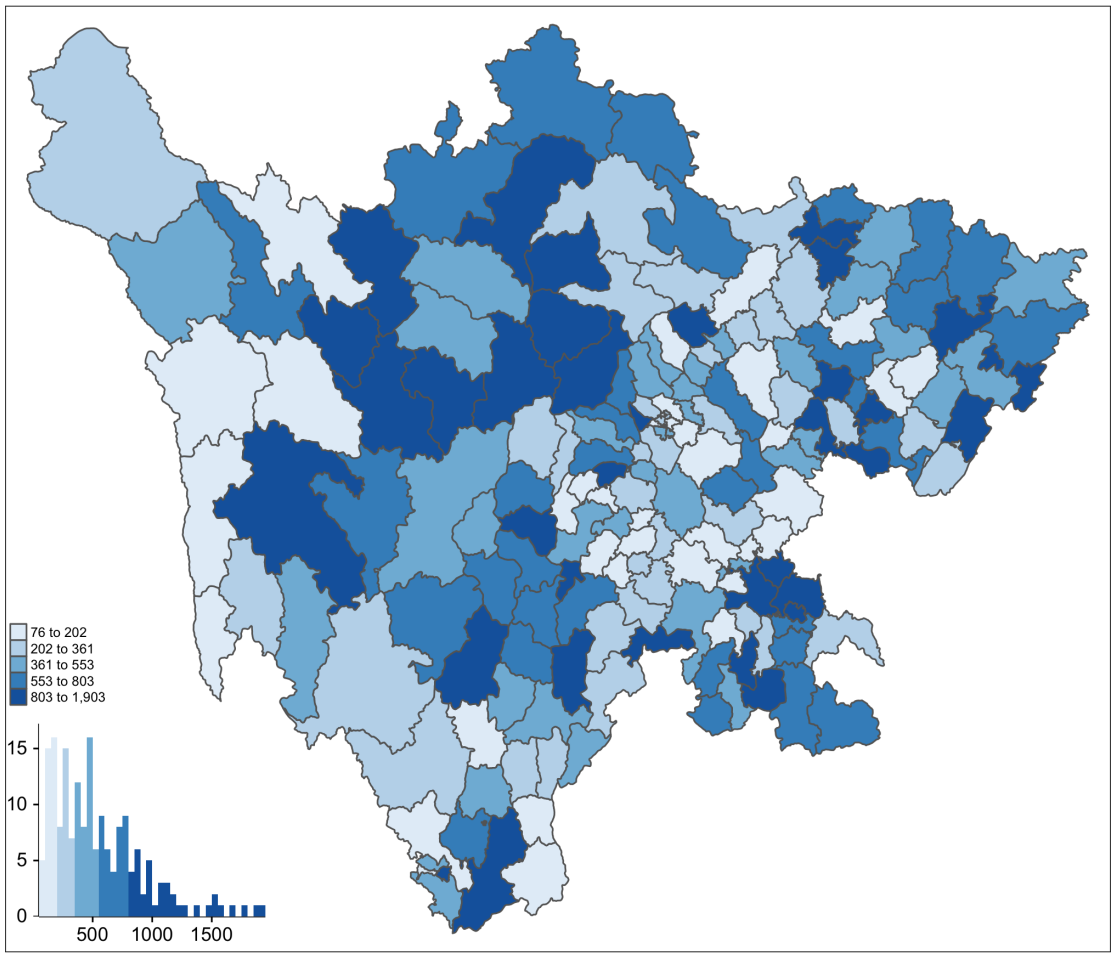


Figure A.1: A map of difference between adjusted number of PTB cases and observed PTB cases in Sichuan in 2010

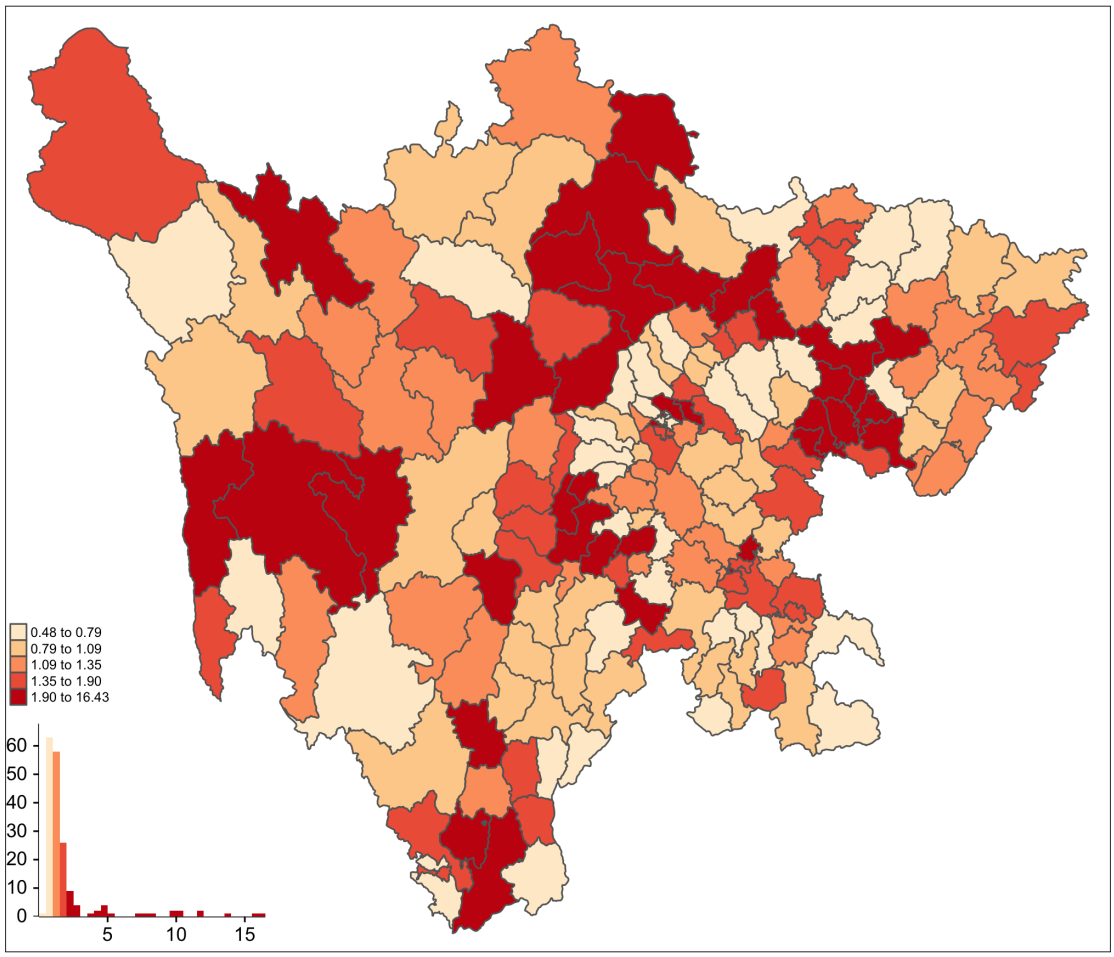


Figure A.2: A map of difference between adjusted PTB prevalence per 1000 population and observed PTB prevalence in Sichuan in 2010

Appendix B

Appendix for Chapter 3

B.1 Conditional multinomial model for population-level two-stream CRC data

$$(N_{11}, N_{10}, N_{01}|N_c = n_c) \sim \text{Multinomial}(n_c, p^*_{11}, p^*_{10}, p^*_{01}),$$

where $n_c = n_{11} + n_{10} + n_{01}$ is the number of cases caught at least once; $p^*_{ij} = p_{ij}/p_c$ in where $p_c = p_{11} + p_{10} + p_{01}$ and p_{ij} denotes the probability of having capture history $(i, j), i, j \in \{0, 1\}$.

$$L(p_c) = \frac{n_c}{n_{11}!n_{01}!n_{10}!}\left(\frac{p_{11}}{p_c}\right)^{n_{11}} \times \left(\frac{p_{10}}{p_c}\right)^{n_{10}} \times \left(\frac{p_{01}}{p_c}\right)^{n_{01}},$$
$$\hat{N} = \frac{n_c}{\hat{p}_c}.$$

B.2 Derivation of bias-corrected estimators under two-stream CRC

Given ϕ, the MLE of N is

$$\hat{N}_\phi = n_{11} + n_{10} + \frac{n_{01}(n_{11} + n_{10})}{n_{11}}\phi.$$

Define $n_c = n_{11} + n_{10} + n_{01}$, $p_c = p_{11} + p_{10} + p_{01}$, and $p^*_{ij} = \frac{p_{ij}}{p_c}$, where $i, j \in \{0, 1\}$. Dividing both sides by $\hat{N}_\phi$, we have:

$$1 = \hat{p}_{11} + \hat{p}_{10} + \frac{\hat{p}_{01}(\hat{p}_{11} + \hat{p}_{10})}{\hat{p}_{11}}\phi,$$
$$\frac{1}{\hat{p}_c} = \hat{p}^*_{11} + \hat{p}^*_{10} + \frac{\hat{p}^*_{01}(\hat{p}^*_{11} + \hat{p}^*_{10})}{\hat{p}^*_{11}}\phi.$$

Then, the $\hat{N}_\phi$ can be written as:

$$\hat{N}_\phi = n_{11} + n_{10} + \frac{n_{01}(n_{11} + n_{10})}{n_{11}}\phi = n_c\left\{\hat{p}^*_{11} + \hat{p}^*_{10} + \frac{\hat{p}^*_{01}(\hat{p}^*_{11} + \hat{p}^*_{10})}{\hat{p}^*_{11}}\phi\right\} = \frac{n_c}{\hat{p}_c}.$$

Let $\boldsymbol{p}^* = (p^*_{11}, p^*_{10}, p^*_{01})$, and define a function $f(\hat{\boldsymbol{p}}^*) = \hat{p}^*_{11} + \hat{p}^*_{10} + \frac{\hat{p}^*_{01}(\hat{p}^*_{11} + \hat{p}^*_{10})}{\hat{p}^*_{11}}\phi$. Follow the Taylor-series expansion (Jewell, 1984):

$$E\left[f(\hat{\boldsymbol{p}}^*)\right] = f\left(\hat{\boldsymbol{p}}^*\right) + g\left(\hat{\boldsymbol{p}}^*\right) + O(n_c^{-2}),$$

where $g(\boldsymbol{p}^*) = E\left[\frac{1}{2}(\hat{\boldsymbol{p}}^* - \boldsymbol{p}^*)^T \boldsymbol{D}_2(\boldsymbol{p}^*)(\hat{\boldsymbol{p}}^* - \boldsymbol{p}^*)\right]$, and $\boldsymbol{D}_2(\boldsymbol{p}^*)$ is the Hessian of f evaluated at $\boldsymbol{p}^*$, which is given by

$$\begin{bmatrix} \frac{2p^*_{10}p^*_{01}}{p^{*3}_{11}}\phi & \frac{-p^*_{01}\phi}{p^{*2}_{11}} & \frac{-p^*_{10}\phi}{p^{*2}_{11}} \\ \frac{-p^*_{01}\phi}{p^{*2}_{11}} & 0 & \frac{\phi}{p^*_{11}} \\ \frac{-p^*_{10}\phi}{p^{*2}_{11}} & \frac{\phi}{p^*_{11}} & 0 \end{bmatrix}$$

Then we have $g(\boldsymbol{p}^*)$ is:

$$g(\boldsymbol{p}^*) = \frac{1}{2} \times \frac{2p^*_{10}p^*_{01}}{p^{*3}_{11}\phi} Var(\hat{p}^*_{11}) + \frac{-p^*_{01}\phi}{p^{*2}_{11}} Cov(\hat{p}^*_{11}, \hat{p}^*_{10}) + \frac{-p^*_{10}\phi}{p^{*2}_{11}} Cov(\hat{p}^*_{11}, \hat{p}^*_{01}) + \frac{\phi}{p^*_{11}} Cov(\hat{p}^*_{10}, \hat{p}^*_{01}).$$

Under the conditional multinomial distribution model, we have:

$$Var(\hat{p}^*_{11}) = \frac{p^*_{11}(1 - p^*_{11})}{n_c},\ Cov(\hat{p}^*_{11}, \hat{p}^*_{10}) = \frac{-p^*_{11}p^*_{10}}{n_c},$$
$$Cov(\hat{p}^*_{11}, \hat{p}^*_{01}) = \frac{-p^*_{11}p^*_{01}}{n_c},\ Cov(\hat{p}^*_{10}, \hat{p}^*_{01}) = \frac{-p^*_{10}p^*_{01}}{n_c}.$$

Some algebra leads to:

$$\hat{g}(\boldsymbol{p}^*) = \frac{n_{10}n_{01}\phi}{n_{11}^2 n_c}.$$

Since $\hat{N}_\phi = n_c f(\boldsymbol{p}^*)$, the estimated bias in $\hat{N}_\phi$ is:

$$n_c\hat{g}(\hat{\boldsymbol{p}}^*) = \frac{n_{10}n_{01}}{n_{11}^2}\phi.$$

Finally, the bias-corrected estimator of N and its variation estimator are given by

$$\hat{N}_\phi^{BC} = \hat{N}_\phi - \frac{n_{10}n_{01}}{n_{11}^2}\phi \tag{B.1}$$

$$\hat{Var}\left(\hat{N}_\phi^{BC}\right) = w_1(w_1 - 1 - C)n_{11} + w_2(w_2 - 1 - C)n_{10} + w_3(w_3 - 1 - C)n_{01}, \tag{B.2}$$

where $C = \frac{n_{10}\hat{p}_{01}}{n_{11}^2}\phi$, $w_1 = 1 - \frac{n_{10}n_{01}}{n_{11}^2}\phi$, $w_2 = 1 + \frac{n_{01}}{n_{11}\phi}$, and $w_3 = \phi + \frac{n_{10}}{n_{11}}\phi - \frac{n_{10}}{n_{11}^2}\phi$.

It is clear that $\hat{N}_\phi^{BC}$ is not well-defined when $n_{11} = 0$. A small adjustment to the denominator term in the correction factor results in:

$$\hat{N}_\phi^{BC2} = \hat{N}_\phi - \frac{n_{10}n_{01}}{(n_{11} + 0.5)^2}\phi. \tag{B.3}$$

The variance estimator of the estimator $\hat{N}_\phi^{BC2}$ derived using the multivariate delta method is:

$$\hat{Var}(\hat{N}_\phi^{BC2}) = w_1(w_1 - 1 - C)n_{11} + w_2(w_2 - 1 - C)n_{10} + w_3(w_3 - 1 - C)n_{01}, \tag{B.4}$$

where $C = \frac{2n_{11}n_{01}\hat{p}_{01}}{(n_{11}+0.5)^3} - \frac{n_{10}\hat{p}_{01}}{(n_{11}+0.5)^2}\phi$, $w_1 = 1 - \frac{n_{10}n_{01}}{n_{11}^2}\phi + \frac{2n_{11}n_{01}}{(n_{11}+0.5)^2}$, $w_2 = 1 + \frac{n_{01}}{n_{11}}\phi - \frac{n_{01}}{(n_{11}+0.5)^2}\phi$, and $w_3 = \phi + \frac{n_{10}}{n_{11}}\phi - \frac{n_{10}}{(n_{11}+0.5)^2}\phi$.

The estimator $\hat{N}_\phi^{BC}$ is a direct generalization of the corrected MLE of Darroch (1958) in the two-stream case under the LP conditions, while $\hat{Var}(\hat{N}_\phi^{BC2})$ (BC2 estimator) generalizes an estimator shown by Lyles et al. (2021a) to be a direct competitor to the bias-corrected estimator of Chapman (1951). One can also consider a simple direct generalization, by

introducing the parameter ϕ into the original form of the Chapman estimator, as follows:

$$\hat{N}^*_{Chap} = \frac{(n_{11} + n_{10} + 1)(n_{11} + n_{01}\phi + 1)}{(n_{11} + 1} - 1 \tag{B.5}$$

Note that Equation (B.5) reduces to the original Chapman estimator when $\phi = 1$, and we refer it as the generalized Chapman estimator. The corresponding variance estimator is:

$$\hat{Var}(\hat{N}^*_{Chap}) = w_1(w_1 - 1 - C)n_{11} + w_2(w_2 - 1 - C)n_{10} + w_3(w_3 - 1 - C)n_{01}, \tag{B.6}$$

where $C = \frac{n_{10}\hat{p}_{01}}{(n_{11}+1)^2}\phi$, $w_1 = 1 - \frac{n_{10}n_{01}}{(n_{11}+1)^2}\phi$, $w_2 = 1 + \frac{n_{01}}{(n_{11}+1)}\phi$, and $w_3 = \phi + \frac{n_{10}}{(n_{11}+1)}\phi$.

Analogous to the recommendations of Lyles et al. (2021a), we suggest setting $\hat{Var}(\hat{N}^{BC2}_{\phi})$ in (B.3) equal to $\hat{Var}(\hat{N}^*_{Chap})$ in (**??**), in the event that $n_{11} = 0$ when using the BC2 approach.

The Taylor-series expansion approach can also be applied to bias-correct the MLE with known odds ratio θ. We first write the MLE of N in Equation (3.5) as:

$$\hat{N}_{\theta} = \frac{n_c}{\hat{p}^*_{11} + \hat{p}^*_{10} + \hat{p}^*_{01} + \frac{\hat{p}^*_{10}\hat{p}^*_{01}}{\hat{p}^*_{11}}\theta}.$$

Similarly, we define $f(\hat{\boldsymbol{p}}^*) = \hat{p}^*_{11} + \hat{p}^*_{10} + \hat{p}^*_{01} + \frac{\hat{p}^*_{10}\hat{p}^*_{01}}{\hat{p}^*_{11}}\theta$, and $\boldsymbol{D}_2(\boldsymbol{p}^*)$ is

$$\begin{bmatrix} \frac{2p^*_{10}p^*_{01}}{p^{*3}_{11}}\theta & \frac{-p^*_{01}\theta}{p^{*2}_{11}} & \frac{-p^*_{10}\theta}{p^{*2}_{11}} \\ \frac{-p^*_{01}\theta}{p^{*2}_{11}} & 0 & \frac{\theta}{p^*_{11}} \\ \frac{-p^*_{10}\theta}{p^{*2}_{11}} & \frac{\theta}{p^*_{11}} & 0 \end{bmatrix}$$

Some algebra leads to:

$$\hat{g}(\boldsymbol{p}^*) = \frac{n_{10}n_{01}\theta}{n^2_{11}n_c}$$

The bias-corrected estimator with given θ is given by

$$\hat{Var}(\hat{N}^{BC}_{\theta}) = \hat{N}_{\theta} - \frac{n_{10}n_{01}\theta}{n^2_{11}} \tag{B.7}$$

As with the BC2 estimator with given ϕ in Equation (B.3), the same adjustment can also be applied to stabilize the estimator in Equation (B.7). The corresponding BC2 estimator with given θ is

$$\hat{Var}(\hat{N}_{\theta}^{BC2}) = \hat{N}_{\theta} - \frac{n_{10}n_{01}\theta}{(n_{11}+0.5)^2} \tag{B.8}$$

B.3 Variance estimators

For notational convenience, we let $\boldsymbol{n}$ denote the vector of observed counts (n_{11}, n_{10}, n_{01}) and write the MLE of N with a given value of a key parameter (e.g., ψ, ϕ, or θ) as $\hat{N} = f(\boldsymbol{n})$. The first derivative of $\hat{N}$ with respect to $\boldsymbol{n}$ is denoted as (w_1, w_2, w_3), with a corresponding variance estimator derived using the multivariate delta method given by

$$\begin{aligned} Var(\hat{N}) = N\Big\{ & w_1 p_{11}\left[w_1(1-p_{11}) - w_2 p_{10} - w_3 p_{01}\right] \\ & + w_2 p_{10}\left[w_2(1-p_{10}) - w_1 p_{11} - w_3 p_{01}\right] \\ & + w_3 p_{01}\left[w_3(1-p_{01}) - w_1 p_{11} - w_2 p_{10}\right] \Big\} \end{aligned} \tag{B.9}$$

1. Variance estimator of $\hat{N}_{\psi}$:

   Some algebra based on the fact that $p_{11} + p_{10} + p_{01}/\psi = 1$ leads to the variance estimator in Equation (3.2).

2. Variance estimator of $\hat{N}_{\phi}$:

   Some algebra based on the fact that $p_{11} + p_{10} + p_{01}\phi + \frac{p_{10}p_{01}}{p_{11}}\phi = 1$ leads to the variance estimator in Equation (3.4).

3. Variance estimator of $\hat{N}_{\phi}^{BC}$:

   Some algebra based on the fact that $p_{11} + p_{10} + \left(p_{01}\frac{p_{10}p_{01}}{p_{11}}\right)\phi - \frac{p_{10}p_{01}}{p_{11}n_{11}}\phi = 1$ leads to the variance estimator in Equation (B.2).

4. Variance estimator of $\hat{N}_{\phi}^{BC2}$:

   Some algebra based on the fact that $p_{11} + p_{10} + \left(p_{01}\frac{p_{10}p_{01}}{p_{11}}\right)\phi - \frac{p_{10}p_{01}n_{11}}{p_{11}(n_{11}+0.5)^2}\phi = 1$ leads to the variance estimator in Equation (B.4).

5. Variance estimator of $\hat{N}^*_{Chap}$:

   Some algebra based on the fact that $p_{11} + p_{10} + p_{01}\phi + \frac{p_{10}p_{01}n_{11}}{p_{11}(n_{11}+1)}\phi = 1$ leads to the variance estimator in Equation (B.6).

B.4 Procedure for obtaining 95% percentile interval for N

Under Option (1), the recommended procedure for obtaining a 95% percentile interval for N (from which we seek favorable frequentist properties) is:

(i). Specify the prior distribution of $p_{2|1}$ and the assumed distribution of ϕ.

(ii). Obtain D posterior samples of ψ by combining D posterior samples of $p_{2|1}$ and D independent random draws of ϕ generated from the assumed distribution using the definition $\psi = p_{2|1}/\phi$ (the posterior samples for ψ empirically reflect both the assumed variation in ϕ and statistical uncertainty in estimating $p_{2|1}$).

(iii). For each generated ψ, take M random draws from a Normal$(\hat{N}_\psi, \hat{Var}(\hat{N}_\psi))$ distribution (see Equations (3.1) and (3.2)).

(iv). A total of $D \times M$ posterior samples of N are obtained by combining together all random draws across each generated value of ψ.

(v). Take the 2.5^{th} and 97.5^{th} percentiles of the resulting $D \times M$ posterior samples of N to construct the 95% percentile interval.

With a Beta (1,0) prior for $p_{2|1}$, the conjugate posterior distribution of $p_{2|1}$ is a Beta distribution with mean equal to $(n_{11} + 1)/(n_{11} + n_{10} + 1)$. Lyles et al. (2021a) show that inserting this mean into Equation (3.1) in place of ψ yields the Chapman estimator. With this in mind, we recommend use of the Beta (1,0) prior for $p_{2|1}$ when implementing the proposed uncertainty analysis under Option (1).

The 95% credible interval for N obtained under Option (2) is produced as follows:

(i). Specify the prior distribution of ϕ and generate D realizations from that distribution. Importantly, accept realizations only if they yield a value greater than $n_{11}/(n_{11}+n_{10})$; this is because $\phi = p_{2|1}/\psi$ and ψ is constrained to be ≤ 1.

(ii). For each given value of ϕ, draw M random draws from a normal distribution with mean $\hat{N}_{\phi}^{BC2}$ and variance equal to the estimated variance of $\hat{N}_{\phi}^{BC2}$ (note that $\hat{N}_{\phi}^{BC2}$ could be replaced by any of the bias-corrected estimators in Equations (B.1), (B.3), and (B.5)).

(iii). Take the 2.5^{th} and 97.5^{th} percentiles of the resulting $D \times M$ posterior samples of N to construct the 95% percentile interval.

B.5 Crossing points of sensitivity plots obtained from two strata

Let $n_{11}^{s_1}$, $n_{10}^{s_1}$, and $n_{01}^{s_1}$ denote the observed cell counts in the first stratum, with the superscript s_2 used analogously for the data in the second stratum. Letting N_{s_1} and N_{s_2} represent the true number of cases for each stratum, the case ratio $r = N_{s_1}/N_{s_2}$ is assumed known (e.g., obtained from previous data on prevalence). Considering the proposed sensitivity analysis focused on ϕ, it can be shown that the x-axis and y-axis coordinates of the crossing point between the plots from stratum 2 and the scaled plots from stratum 1 (i.e., depicting N_{s_1} divided by r) are:

$$\hat{\phi} = \frac{\frac{\left(n_{11}^{s_1}+n_{10}^{s_1}\right)}{r} - \left(n_{11}^{s_2}+n_{10}^{s_2}\right)}{\frac{n_{01}^{s_2}\left(n_{11}^{s_2}+n_{10}^{s_2}\right)}{n_{11}^{s_2}} - \frac{n_{01}^{s_1}\left(n_{11}^{s_1}+n_{10}^{s_1}\right)}{r \times n_{11}^{s_1}}}, \tag{B.10}$$

$$\hat{N}_{s_2,\hat{\phi}} = n_{11}^{s_2} + n_{10}^{s_2} + \frac{n_{01}^{s_2}\left(n_{11}^{s_2}+n_{10}^{s_2}\right)}{n_{11}^{s_2}}\hat{\phi}. \tag{B.11}$$

Similarly, the crossing point of the two sensitivity analysis plots anchored on θ are given

by

$$\hat{\theta} = \frac{r(n_{11}^{s_2} + n_{10}^{s_2} + n_{01}^{s_2}) - (n_{11}^{s_1} + n_{10}^{s_1} + n_{01}^{s_1})}{\frac{n_{10}^{s_1} n_{01}^{s_1}}{n_{11}^{s_1}} - \frac{r n_{10}^{s_2} n_{01}^{s_2}}{n_{11}^{s_2}}}, \tag{B.12}$$

$$\hat{N}_{s_2,\hat{\theta}} = n_{11}^{s_2} + n_{10}^{s_2} + n_{01}^{s_2} + \frac{n_{10}^{s_2} n_{01}^{s_2}}{n_{11}^{s_2}} \hat{\theta} \tag{B.13}$$

B.6 MLEs with a known case ratio

Let $n_c^{s_1}$ and $n_c^{s_2}$ denote the number of distinct cases identified by the two data streams for each stratum. The $p_{11}^{s_1}$, $p_{10}^{s_1}$, and $p_{01}^{s_1}$ denote corresponding capture probabilities for the first stratum and the probability of being caught at least once for the first stratum is denoted as $p_c^{s_1} = p_{11}^{s_1} + p_{10}^{s_1} + p_{01}^{s_1}$. Similarly, the superscript s_2 is used for the second stratum. The conditional multinomial likelihood is given by

$$L = \frac{n_c^{s_1}!}{n_{11}^{s_1}! n_{10}^{s_1}! n_{01}^{s_1}!} \left(\frac{p_{11}^{s_1}}{p_c^{s_1}}\right)^{n_{11}^{s_1}} \left(\frac{p_{10}^{s_1}}{p_c^{s_1}}\right)^{n_{10}^{s_1}} \left(\frac{p_{01}^{s_1}}{p_c^{s_1}}\right)^{n_{01}^{s_1}} \frac{n_c^{s_2}!}{n_{11}^{s_2}! n_{10}^{s_2}! n_{01}^{s_2}!} \left(\frac{p_{11}^{s_2}}{p_c^{s_2}}\right)^{n_{11}^{s_2}} \left(\frac{p_{10}^{s_2}}{p_c^{s_2}}\right)^{n_{10}^{s_2}} \left(\frac{p_{01}^{s_2}}{p_c^{s_2}}\right)^{n_{01}^{s_2}}$$

The MLEs of $p_c^{s_1}$ and $p_c^{s_2}$ under the assumption that ϕ is the same across strata have the following forms:

$$\hat{p}_c^{s_1} = \frac{n_c^{s_1}}{n_c^{s_1} + \frac{n_{10}^{s_1} n_{01}^{s_1}}{n_{11}^{s_1}} \phi} \tag{B.14}$$

$$\hat{p}_c^{s_2} = \frac{n_c^{s_2}}{n_c^{s_2} + \frac{n_{10}^{s_2} n_{01}^{s_2}}{n_{11}^{s_2}} \phi} \tag{B.15}$$

Since we have $\frac{n_c^{s_1}}{p_c^{s_1}} = \frac{r n_c^{s_2}}{p_c^{s_2}}$, the MLE of ϕ is

$$\hat{\phi} = \frac{\frac{(n_{11}^{s_1} + n_{10}^{s_1})}{r} - (n_{11}^{s_2} + n_{10}^{s_2})}{\frac{n_{01}^{s_2}(n_{11}^{s_2} + n_{10}^{s_2})}{n_{11}^{s_2}} - \frac{n_{01}^{s_1}(n_{11}^{s_1} + n_{10}^{s_1})}{r \times n_{11}^{s_1}}} \tag{B.16}$$

Supplying the MLE of ϕ to Equation (B.15) and using the fact that $N_{s_2} = n_c^{s_2} / \hat{p}_c^{s_2}$, the MLE of N_{s_2} is

$$\hat{N}_{s_2,\hat{\phi}} = n_{11}^{s_2} + n_{10}^{s_2} + \frac{n_{01}^{s_2}(n_{11}^{s_2} + n_{10}^{s_2})}{n_{11}^{s_2}} \hat{\phi}. \tag{B.17}$$

The MLEs of ϕ and N_{s_2} coincide with estimators in Equations (B.10) and (B.11). We also verified that estimators in Equations (B.12) and (B.13) are the same as the MLEs provided in Equations 4 and 5 of Wolter (1990), derived using the multinomial model.

B.7 MLEs under three-stream CRC

When three overlapping surveillance streams are implemented, 7 observed cell counts are obtained, denoted as n_{111}, n_{110}, n_{101}, n_{100}, n_{011}, n_{010}, n_{001}. The MLE and its variance estimator with given $p_{3|\bar{1}\bar{2}}$ derived based on population-level multinomial model are given by

$$\hat{N} = n_{111} + n_{110} + n_{101} + n_{100} + n_{011} + n_{010} + n_{001}/p_{3|\bar{1}\bar{2}}, \tag{B.18}$$

$$\hat{Var}(\hat{N}) = \frac{(1 - p_{3|\bar{1}\bar{2}})}{p^2_{3|\bar{1}\bar{2}}} n_{001}. \tag{B.19}$$

Note that Equation (B.18) is a direct generalization of Equation (3.1). Three inestimable ratio parameters (akin to relative risks) are available as focal points for sensitivity analysis in this case, i.e., $\phi_1 = p_{3|12}/p_{3|\bar{1}\bar{2}}$, $\phi_2 = p_{3|\bar{1}2}/p_{3|\bar{1}\bar{2}}$, and $\phi_3 = p_{3|1\bar{2}}/p_{3|\bar{1}\bar{2}}$. Any of the three ratio parameters would have an intuitive interpretation, for example, the ratio assumption $p_{3|1\bar{2}}/p_{3|\bar{1}\bar{2}} = 1$ reflects an assumption that identification in stream 3 is not impacted by identification in stream 1, given non-identification in stream 2.

Table B.1: Cell counts for two-stream capture-recapture analyzed in Wolter (1990)

	Males		Females	
	Captured in S2		Captured in S3	
Captured in S1	Yes	No	Yes	No
Yes	46	11	54	13
No	20	?	5	?

Table B.2: Results from analyzing data presented in Table B.1 by applying the proposed sensitivity analysis under the assumption that ϕ or θ are equal across sexes and the sex ratio is known to be 1.15

Point estimate	Assume ϕ are equal across sexes	Assume θ are equal across sexes
ϕ	1.1	-
θ	-	1.7
N	159	159

Appendix C

Appendix for Chapter 4

Data were generated from the population-level multinomial:

$$(N_{111}, N_{110}, N_{101}, N_{100}, N_{011}, N_{010}, N_{001}, N_{000})$$
$$\sim Multi(N; p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000})$$

The true number of cases N is set to $5,000$ under both scenarios, and capture probabilities are computed based on parameters $(p_1, p_{2|1}, p_{2|\bar{1}}, p_{3|12}, p_{3|\bar{1}2}, p_{3|1\bar{2}}, p_{3|\bar{1}\bar{2}})$:

$$\begin{aligned}
p_{111} &= p_1 p_{2|1} p_{3|12}, \\
p_{110} &= p_1 p_{2|1}(1 - p_{3|12}), \\
p_{101} &= p_1(1 - p_{2|1}) p_{3|1\bar{2}}, \\
p_{100} &= p_1(1 - p_{2|1})(1 - p_{3|1\bar{2}}), \\
p_{011} &= (1 - p_1) p_{2|\bar{1}} p_{3|\bar{1}2}, \\
p_{010} &= (1 - p_1) p_{2|\bar{1}}(1 - p_{3|\bar{1}2}), \\
p_{001} &= (1 - p_1)(1 - p_{2|\bar{1}})\psi, \\
p_{000} &= (1 - p_1)(1 - p_{2|\bar{1}})(1 - \psi),
\end{aligned}$$

where under three-stream cases, $\psi = p_{3|\bar{1}\bar{2}}$.

Scenario 1

We assume the probability of having capture history (011) is equal to the probability of having capture history (010), and that the association between the first data stream and the third data stream is not affected by whether cases are identified by the second data stream. Converting to mathematical expressions, these stipulations correspond to setting the testable assumption $p_{011} = p_{010}$, i.e., $E(N_{011}) = E(N_{010})$, and the untestable assumption $p_{3|12}/p_{3|\bar{1}2} = p_{3|12\bar{}}/\psi$. True values of the parameters are: $p_1 = 0.3$, $p_{2|1} = 0.2$, $p_{2|\bar{1}} = 0.3$, $p_{3|12} = 0.8$, $p_{3|12\bar{}} = 0.16$, $p_{3|\bar{1}2} = 0.5$, $\psi = 0.1$.

Scenario 2

We impose two testable assumptions $E(N_{111}) = E(N_{101})$ and $E(N_{110}) = E(N_{100})$, and one untestable assumption which states that the key parameter $\psi = p_{3|12\bar{}}/0.8$. This untestable assumption implies that, among those not identified by the second stream, cases are more likely to be captured by the third stream if they are NOT captured by the first stream, i.e., the first stream and third stream are negatively correlated conditional on a lack of capture by the second stream.

Appendix D

Appendix for Chapter 5

D.1 Dirichlet-multinomial-based approach for inference

Let $\boldsymbol{n} = \{n_{h_i}\}$, for $i = 1, \ldots, 2^K - 1$, denote the vector of observed counts under the K-catch case, where $K \geq 2$. Let $\boldsymbol{p}^* = \{p_{h_i}/p_c\}$, for $i = 1, \ldots, 2^K - 1$, denote the vector of probabilities of having the observable capture histories conditional on the event of being identified at least once, where p_c is the probability of that event. Here the parameter $\boldsymbol{\theta}$ is removed for notational simplicity; however, a constraint is indeed imposed by specifying $g(\boldsymbol{\theta}^*)$ to avoid the non-identifiability issue when estimating N. As described in Section 5.2.2, $\boldsymbol{n}|n_c \sim \text{Multinomial}(n_c, \boldsymbol{p}^*)$, where n_c is the number of uniquely identified cases.

The procedure for obtaining the proposed 95% credible intervals for N under the constraint $g(\boldsymbol{\theta}^*)$ is summarized as follows:

(1). Assign a Dirichlet prior to $\boldsymbol{p}^*$ and draw L posterior samples for $\boldsymbol{p}^*$ from the conjugate Dirichlet posterior distribution, denoting these samples as $\boldsymbol{p}_l^*$, for $l = 1, \ldots, L$.

(2). Obtain L simulated CRC datasets by multiplying n_c by $\boldsymbol{p}_l^*$, then round the product to be integers, denoted as n_l.

(3). Use each simulated dataset n_l to calculate the number of uniquely identified cases $n_{c,l}$

and compute an estimate $\hat{N}_l$ via Equation (5.5), mimicking the desired estimand under the constraint $g(\boldsymbol{\theta}^*)$.

(4). For each pair of $n_{c,l}$ and $\hat{N}_l$, compute $p_{c,l} = n_{c,l}/\hat{N}_l$ and draw a sample $n^*_{c,l}$ from Binomial $\left(\frac{n_c}{p_{c,l}}, p_{c,l}\right)$. Note that $p_{c,l}$ should be no larger than 1, and is thus set to 1 in the rare event that this fails to hold.

(5). Take the 2.5^{th} and 97.5^{th} percentiles of the posterior samples of N computed as $n^*_{c,l}/p_{c,l}$ to construct the 95% credible interval.

Based on simulation studies conducted in this work, we found $L = 1,000$ to be sufficient to obtain a reliable 95% credible interval.

D.2 Uncertainty analysis

We use an example to illustrate the procedure of applying the proposed uncertainty analysis to obtain 95% credible intervals which acknowledge statistical uncertainties and propagate the uncertainty about the key parameter which controls the dependency between data streams. Let $K = 3$ and define $g(\boldsymbol{\theta}^*) = r\frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$, which is the same setting as in the simulation study conducted to assess the uncertainty analysis with results given in Table 5.6. For this example, the key dependency parameter is r and we assume it to follow a parametric distribution denoted as f. The procedure generalizes a prior proposal for the special case of $K = 2$ as discussed in Chapter 3 (Zhang et al., 2023a) and is similar to the method described in Appendix D.1 for obtaining 95% credible intervals without assuming extra uncertainty about the key dependency parameter. Specifically,

(1). Draw M samples based the assumed parametric distribution f, and denote the sample as r_m, for $m = 1, \ldots, M$.

(2). For each r_m, apply the procedure for obtaining L posterior samples of N under the constraint $g(\boldsymbol{\theta}^*) = r_m\frac{p_{3|\bar{1}2}p_{3|1\bar{2}}}{p_{3|12}}$ (i.e., follow steps (1)-(5) described in Section D.1).

(3). Take the 2.5^{th} and 97.5^{th} percentiles of the posterior samples of N by pooling $L \times M$ posterior samples together to construct the desired 95% credible interval.

D.3 Simulation settings

Data generation

Simulation studies in Section 5.3 focused on the three-catch case; thus, data were simulated using the following population-level multinomial distribution:

$$\begin{aligned}&(N_{111}, N_{110}, N_{101}, N_{100}, N_{011}, N_{010}, N_{001}, N_{000})\\ &\sim \text{Multinomial}(N; p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000}),\end{aligned}$$

where N is the true number of cases.

Using conditional probabilities, $\boldsymbol{\theta} = (p_1, p_{2|1}, p_{2|\bar{1}}, p_{3|12}, p_{3|\bar{1}2}, p_{3|1\bar{2}}, \psi)$, introduced in this work, the multinomial probabilities are computed as:

$$\begin{aligned}&p_{111} = p_1 p_{2|1} p_{3|12}, \quad p_{110} = p_1 p_{2|1}(1 - p_{3|12}),\\ &p_{101} = p_1(1 - p_{2|1}) p_{3|1\bar{2}}, \quad p_{100} = p_1(1 - p_{2|1})(1 - p_{3|1\bar{2}}),\\ &p_{011} = (1 - p_1) p_{2|\bar{1}} p_{3|\bar{1}2}, \quad p_{010} = (1 - p_1) p_{2|\bar{1}}(1 - p_{3|\bar{1}2}),\\ &p_{001} = (1 - p_1)(1 - p_{2|\bar{1}})\psi, \quad p_{000} = (1 - p_1)(1 - p_{2|\bar{1}})(1 - \psi),\end{aligned}$$

where the computation of ψ depends on the assumed constraint(s) in each simulation scenario.

Parameters used in producing simulation results included in Section 5.3

1. Parameters used in simulation scenarios presented in Table 5.2, where data were generated under the constraint $\psi = p_{3|\bar{1}2}$.

N	p_1	$p_{2\|1}$	$p_{2\|\bar{1}}$	$p_{3\|12}$	$p_{3\|\bar{1}2}$	$p_{3\|1\bar{2}}$	ψ	p_c
500	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
500	0.35	0.4	0.2	0.3	0.3	0.1	0.3	0.64
1000	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
1000	0.35	0.4	0.2	0.3	0.3	0.1	0.3	0.64
2000	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
2000	0.35	0.4	0.2	0.3	0.3	0.1	0.3	0.64
5000	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
5000	0.35	0.4	0.2	0.3	0.3	0.1	0.3	0.64

2. Parameters used in simulation scenarios presented in Table 5.3, where data were generated under the constraint $\frac{p_{3|12}/(1-p_{3|12})}{p_{3|\bar{1}2}/(1-p_{3|\bar{1}2})} = \frac{p_{3|1\bar{2}}/(1-p_{3|1\bar{2}})}{\psi/(1-\psi)} = r$. Different ψ values were realized by varying r.

N	p_1	$p_{2\|1}$	$p_{2\|\bar{1}}$	$p_{3\|12}$	$p_{3\|\bar{1}2}$	$p_{3\|1\bar{2}}$	ψ	r	p_c
500	0.35	0.4	0.2	0.25	0.142857	0.25	0.14	2	0.55
500	0.35	0.4	0.2	0.25	0.294118	0.25	0.29	0.8	0.63
1000	0.35	0.4	0.2	0.25	0.142857	0.25	0.14	2	0.55
1000	0.35	0.4	0.2	0.25	0.294118	0.25	0.29	0.8	0.63
2000	0.35	0.4	0.2	0.25	0.142857	0.25	0.14	2	0.55
2000	0.35	0.4	0.2	0.25	0.294118	0.25	0.29	0.8	0.63
5000	0.35	0.4	0.2	0.25	0.142857	0.25	0.14	2	0.55
5000	0.35	0.4	0.2	0.25	0.294118	0.25	0.29	0.8	0.63

3. Parameters used in simulation scenarios presented in Table 5.5, where data were generated under the independence assumption (i.e., $p_{2|1} = p_{2|\bar{1}}$, and $p_{3|12} = p_{3|\bar{1}2} = p_{3|1\bar{2}} = \psi$) and the assumption $\psi = p_{3|\bar{1}2}$.

Parameters used for generating data under the independence assumption are:

N	p_1	$p_{2\|1} = p_{2\|\bar{1}}$	$p_{3\|12} = p_{3\|\bar{1}2} = p_{3\|1\bar{2}} = \psi$	p_c
500	0.2	0.25	0.1	0.46
1000	0.2	0.25	0.1	0.46
2000	0.2	0.25	0.1	0.46
5000	0.2	0.25	0.1	0.46

Parameters used for generating data under the assumption $\psi = p_{3|\bar{1}2}$ are:

N	p_1	$p_{2\|1}$	$p_{2\|\bar{1}}$	$p_{3\|12}$	$p_{3\|\bar{1}2}$	$p_{3\|1\bar{2}}$	ψ	p_c
500	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
1000	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
2000	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53
5000	0.35	0.4	0.2	0.3	0.1	0.1	0.1	0.53

4. Parameters used in simulation scenarios presented in Table 5.6, where data were generated under the constraint $\frac{p_{3|12}}{p_{3|\bar{1}2}} = r\frac{p_{3|1\bar{2}}}{\psi}$.
   Parameters used for generating data are ($p_1 = 0.2 p_{2|1} = 0.4, p_{2|\bar{1}} = 0.2, p_{3|12} = 0.25, p_{3|\bar{1}2} = 0.2, p_{3|1\bar{2}} = 0.2$). When assuming $r \sim$ Uniform(0.8, 1.2), data were generated by first drawing 1000 replicates of r from Uniform(0.8, 1.2), and computing the corresponding ψ for each sampled r value. Finally, we generated data from the multinomial distribution with probabilities computed using ψ and the other parameters specified beforehand. When fixing r, the data generation procedure is the same as before.

5. Parameters used in simulation scenarios presented in Table 5.4, where data were generated under the referral scenario where a proportion q of cases are referred from stream 1 to stream 3, and streams 1 and 3 are independent given the capture status under stream 2. Specifically, $p_{3|12} = p_{3|\bar{1}2}$ and $\psi = p_{3|1\bar{2}}$.

N	q	p_1	$p_{2\|1}$	$p_{2\|\bar{1}}$	$p_{3\|12}$	$p_{3\|\bar{1}2}$	$p_{3\|1\bar{2}}$	ψ	p_c
500	0.1	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
500	0.3	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
1000	0.1	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
1000	0.3	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
2000	0.1	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
2000	0.3	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
5000	0.1	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68
5000	0.3	0.35	0.4	0.3	0.1	0.3	0.1	0.3	0.68

D.4 Log-linear models fitted in simulation studies with results presented in Tables 5.2 and 5.3

The log-linear model implies constraint $p_{3|\bar{1}2} = \psi$

$$\log\left[E\left(N_{h_i}\right)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 X_1 X_2 + \gamma_2 X_1 X_3 + \delta X_1 X_2 X_3, \tag{D.1}$$

The log-linear model implies constraint $\frac{p_{3|12}/(1-p_{3|12})}{p_{3|\bar{1}2}/(1-p_{3|\bar{1}2})} = \frac{p_{3|1\bar{2}}/(1-p_{3|1\bar{2}})}{\psi/(1-\psi)}$

$$\log\left[E\left(N_{h_i}\right)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 X_1 X_2 + \gamma_2 X_1 X_3 + \gamma_3 X_2 X_3, \tag{D.2}$$

where X_1, X_2, and X_3 are capture indicators for streams 1, 2, and 3, respectively; 1 indicates captured, and 0 otherwise. For example, under model (D.1), $\log\left[E\left(N_{h_1}\right)\right] = \alpha + \beta_1 + \beta_2 + \beta_3 + \gamma_1 + \gamma_2 + \delta$.

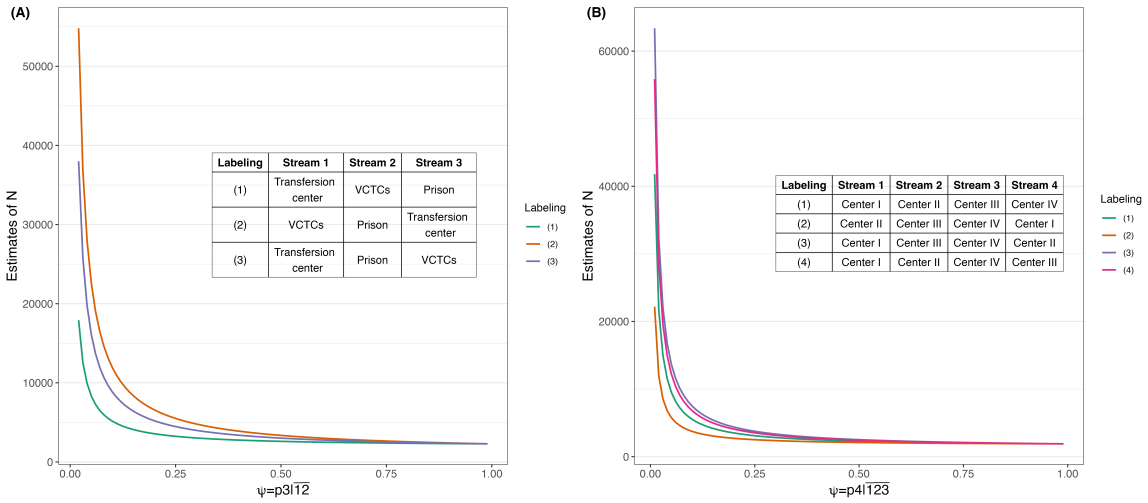


Figure D.1: Estimates of N from the closed form MLE in Equation (5.4) by varying the assumed ψ under three/four-catch case with different labeling. Figure D.1(A) uses the three-catch HIV CRC data analyzed in Poorolajal et al. (2017) and Figure D.1(B) uses the four-catch HIV CRC data analyzed in Abeni et al. (1994).

Bibliography

Abeni, D. D., Brancato, G., and Perucci, C. A. (1994). Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection. *Epidemiology*, pages 410–414.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

Barocas, J. A., White, L. F., Wang, J., Walley, A. Y., LaRochelle, M. R., Bernson, D., Land, T., Morgan, J. R., Samet, J. H., and Linas, B. P. (2018). Estimated prevalence of opioid use disorder in massachusetts, 2011–2015: a capture–recapture analysis. *American Journal of Public Health*, 108(12):1675–1681.

Bernillon, P., Lievre, L., Pillonel, J., Laporte, A., and Costagliola, D. (2000). Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of aids cases: France 1990–1993. *International journal of epidemiology*, 29(1):168–174.

Boden, L. I. (2014). Capture–recapture estimates of the undercount of workplace injuries and illnesses: Sensitivity analysis. *American journal of industrial medicine*, 57(10):1090–1099.

Böhning, D., Rocchetti, I., Maruotti, A., and Holling, H. (2020). Estimating the undetected infections in the covid-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases*, 97:197–201.

Cameron, C. M., Coppell, K. J., Fletcher, D. J., and Sharples, K. J. (2012). Capture-recapture using multiple data sources: estimating the prevalence of diabetes. *Australian and New Zealand Journal of Public Health*, 36(3):223–228.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791.

Chao, A., Pan, H.-Y., and Chiang, S.-C. (2008). The petersen–lincoln estimator and its extension to estimate the size of a shared population. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(6):957–970.

Chao, A., Tsay, P., Lin, S.-H., Shau, W.-Y., and Chao, D.-Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in medicine*, 20(20):3123–3157.

Chapman, D. G. (1951). Some properties of hyper-geometric distribution with application to zoological census. *University of California Publications Statistics*, 1:131–160.

Chatterjee, K. and Mukherjee, D. (2016). On the estimation of homogeneous population size from a complex dual-record system. *Journal of Statistical Computation and Simulation*, 86(17):3562–3581.

Chen, J. (2020). Sensitivity and uncertainty analysis for two-stream capture-recapture in epidemiological surveillance. *Master of Science in Public Health Thesis, Department of Biostatistics and Bioinformatics, The Rollins School of Public Health, Emory University.*

Cormack, R. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics*, pages 567–576.

Cormack, R. and Jupp, P. (1991). Inference for poisson and multinomial models for capture-recapture experiments. *Biometrika*, 78(4):911–916.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, pages 395–413.

Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55(1):294–301.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Darroch, J. N. (1958). The multiple-recapture census: I. estimation of a closed population. *Biometrika*, 45(3/4):343–359.

Darroch, J. N., Fienberg, S. E., Glonek, G. F., and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88(423):1137–1148.

Dorazio, R. M. and Andrew Royle, J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59(2):351–364.

Farcomeni, A. (2011). Recapture models under equality constraints for the conditional capture probabilities. *Biometrika*, 98(1):237–242.

Farcomeni, A. (2016). A general class of recapture models based on the conditional capture probabilities. *Biometrics*, 72(1):116–124.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59(3):591–603.

Finney, D. (1947). The truncated binomial distribution. *Annals of Eugenics*, 14(1):319–328.

Fox, M. P., Lash, T. L., and Greenland, S. (2005). A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International journal of epidemiology*, 34(6):1370–1376.

Ge, L., Zhang, Y., Ward, K. C., Lash, T. L., Waller, L. A., and Lyles, R. H. (2023). Tailoring capture-recapture methods to estimate registry-based case counts based on error-prone diagnostic signals. *Statistics in Medicine*.

Gerritse, S. C., van der Heijden, P. G., and Bakker, B. F. (2015). Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of official statistics*, 31(3):357–379.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., et al. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC public health*, 14(1):1–17.

Gimenez, O., Choquet, R., Lamor, L., Scofield, P., Fletcher, D., Lebreton, J.-D., and Pradel, R. (2005). Efficient profile-likelihood confidence intervals for capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 10:184–196.

Groenwold, R. H., Nelson, D. B., Nichol, K. L., Hoes, A. W., and Hak, E. (2010). Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *International journal of epidemiology*, 39(1):107–117.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables.

Hadorn, D. C. and Stärk, K. D. (2008). Evaluation and optimization of surveillance systems for rare and emerging infectious diseases. *Veterinary research*, 39(6):1.

Héraud-Bousquet, V., Lot, F., Esvan, M., Cazein, F., Laurent, C., Warszawski, J., and Gallay, A. (2012). A three-source capture-recapture estimate of the number of new hiv diagnoses in children in france from 2003–2006 with multiple imputation of a variable of heterogeneous catchability. *BMC infectious diseases*, 12(1):1–9.

Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic reviews*, 17(2):243–264.

Hook, E. B. and Regal, R. R. (1997). Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American journal of epidemiology*, 145(12):1138–1144.

Hook, E. B. and Regal, R. R. (2000). Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *American journal of epidemiology*, 152(8):771–779.

Huff, A., Allen, T., Whiting, K., Williams, F., Hunter, L., Gold, Z., Madoff, L., and Karesh, W. (2017). Biosurveillance: a systematic review of global infectious disease surveillance systems from 1900 to 2016. *Rev Sci Tech*, 36(2):513–524.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

Institute, S. (1985). *SAS user's guide: Statistics*, volume 2. Sas Inst.

Jamison, D. T., Breman, J. G., Measham, A. R., Alleyne, G., Claeson, M., Evans,

D. B., Jha, P., Mills, A., and Musgrove, P. (2006). *Disease control priorities in developing countries*. The World Bank.

Jewell, N. P. (1984). Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics*, pages 421–435.

Jones, H. E., Hickman, M., Welton, N. J., De Angelis, D., Harris, R. J., and Ades, A. (2014). Recapture or precapture? fallibility of standard capture-recapture methods in the presence of referrals between sources. *American journal of epidemiology*, 179(11):1383–1393.

Lee, S.-M., Hwang, W.-H., and Huang, L.-H. (2003). Bayes estimation of population size from capture-recapture models with time variation and behavior response. *Statistica Sinica*, pages 477–494.

Li, T., Cheng, Q., Li, C., Stokes, E., Collender, P., Ohringer, A., Li, X., Li, J., Zelner, J. L., Liang, S., et al. (2019). Evidence for heterogeneity in china's progress against pulmonary tuberculosis: uneven reductions in a major center of ongoing transmission, 2005–2017. *BMC infectious diseases*, 19(1):1–11.

Li, X., Chang, H. H., Cheng, Q., Collender, P. A., Li, T., He, J., Waller, L. A., Lopman, B. A., and Remais, J. V. (2020). A spatial hierarchical model for integrating and bias-correcting data from passive and active disease surveillance systems. *Spatial and Spatio-temporal Epidemiology*, 35:100341.

Lincoln, F. C. (1930). *Calculating waterfowl abundance on the basis of banding returns*. Number 118. US Department of Agriculture.

Liu, J., Yao, H., and Liu, E. (2005). Analysis of factors affecting the epidemiology of tuberculosis in china. *The International Journal of Tuberculosis and Lung Disease*, 9(4):450–454.

Lönnroth, K., Jaramillo, E., Williams, B. G., Dye, C., and Raviglione, M. (2009). Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Social science & medicine*, 68(12):2240–2246.

Lum, K. and Ball, P. (2015). Estimating undocumented homicides with two lists and list dependence. *Human Rights Data Analysis Group*, 1.

Lyles, R. H. and Lin, J. (2010). Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in medicine*, 29(22):2297–2309.

Lyles, R. H., Wilkinson, A. L., Williamson, J. M., Chen, J., Taylor, A. W., Jambai, A., Jalloh, M., and Kaiser, R. (2021a). *Alternative capture-recapture point and interval estimators based on two surveillance streams*. Springer (in press).

Lyles, R. H., Zhang, Y., Ge, L., England, C., Ward, K., Lash, T. L., and Waller, L. A. (2021b). Using capture-recapture to enhance precision of representative sampling-based case count estimates. submitted to Journal of Survey Statistics and Methodology.

Mallet, M., Rivest, L., and Bell, W. R. (1994). Capture-recapture estimation with known sex ratio.

McClish, D. and Penberthy, L. (2004). Using medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. *Medical care*, pages 1111–1116.

McDonald, T. L. and Amstrup, S. C. (2001). Estimation of population size using open capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6:206–220.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135.

Peixoto, V. R., Nunes, C., and Abrantes, A. (2020). Epidemic surveillance of covid-19: Considering uncertainty and under-ascertainment. *Portuguese Journal of Public Health*, 38(1):23–29.

Petersen, C. G. J. (1896). The yearly immigration of young plaice in the limfjord from the german sea. *Rept. Danish Biol. Sta.*, 6:1–48.

Poorolajal, J., Mohammadi, Y., and Farzinara, F. (2017). Using the capture-recapture method to estimate the human immunodeficiency virus-positive population. *Epidemiology and Health*, 39.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramos, P. L., Sousa, I., Santana, R., Morgan, W. H., Gordon, K., Crewe, J., Rocha-Sousa, A., and Macedo, A. F. (2020). A review of capture-recapture methods and its possibilities in ophthalmology and vision sciences. *Ophthalmic Epidemiology*, 27(4):310–324.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Sadinle, M. (2009). Transformed logit confidence intervals for small populations in single capture–recapture estimation. *Communications in Statistics-Simulation and Computation*, 38(9):1909–1924.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43(1):142–152.

Sandland, R. and Cormack, R. (1984). Statistical inference for poisson and multinomial models for capture-recapture experiments. *Biometrika*, 71(1):27–33.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Seber, G. A. F. et al. (1982). The estimation of animal abundance and related parameters.

Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115.

Simonsen, L., Gog, J. R., Olson, D., and Viboud, C. (2016). Infectious disease surveillance in the big data era: towards faster and locally relevant systems. *The Journal of infectious diseases*, 214(suppl_4):S380–S385.

Stanley, T. R. and Burnham, K. P. (1998). Information-theoretic model selection and model averaging for closed-population capture-recapture studies. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 40(4):475–494.

Tounkara, F. and Rivest, L.-P. (2015). Mixture regression models for closed population capture–recapture data. *Biometrics*, 71(3):721–730.

Van Hest, N., Story, A., Grant, A., Antoine, D., Crofts, J., and Watson, J. (2008). Record-linkage and capture–recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in england 1999–2002. *Epidemiology & Infection*, 136(12):1606–1616.

Wang, L., Ren, X., Cowling, B. J., Zeng, L., Geng, M., Wu, P., Li, Z., Yu, H., and Gao, G. (2019). Systematic review: national notifiable infectious disease surveillance system in china. *Online Journal of Public Health Informatics*, 11(1).

Wanyeki, I., Olson, S., Brassard, P., Menzies, D., Ross, N., Behr, M., and Schwartzman, K. (2006). Dwellings, crowding, and tuberculosis in montreal. *Social science & medicine*, 63(2):501–511.

Wittes, J. T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69(345):93–97.

Wolter, K. M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, pages 157–162.

Yang, X.-y., Zhang, N.-m., Diao, X., Mao, X., and Li, Y.-p. (2008). Epidemiological analysis of pulmonary tuberculosis in sichuan province, china, 2000–2006. *International Journal of Infectious Diseases*, 12(5):534–541.

Zhang, B. and Small, D. S. (2020). Number of healthcare workers who have died of covid-19. *Epidemiology*, 31(6):e46.

Zhang, Y., Chen, J., Ge, L., Williamson, J. M., Waller, L. A., and Lyles, R. H. (2022). Sensitivity and uncertainty analysis for two-stream capture-recapture methods in disease surveillance. *medRxiv*, pages 2022–09.

Zhang, Y., Chen, J., Ge, L., Williamson, J. M., Waller, L. A., and Lyles, R. H. (2023a). Sensitivity and uncertainty analysis for two-stream capture–recapture methods in disease surveillance. *Epidemiology*, 34(4):601–610.

Zhang, Y., Ge, L., Waller, L. A., and Lyles, R. H. (2023b). On some pitfalls of the log-linear modeling framework for capture-recapture studies in disease surveillance.

Zwane, E. and van der Heijden, P. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling*, 5(1):39–52.