Cohort Analysis using the Multiphase Method:

Application to Oregon Viral Hepatitis Mortality, 1995–2010

by

Miriam R. Elman

A thesis presented to the Department of Public Health & Preventive Medicine
and the Oregon Health & Science University School of Medicine
in partial fulfillment of the requirements for the degree of
Master of Public Health
June 2012

School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Miriam R. Elman

has been approved

Jodi Lapidus, Ph.D. – Mentor/advisor

Amy Sullivan, Ph.D., M.P.H. – Member

Atif Zaman, M.D., M.P.H. – Member

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

For my parents

ACKNOWLEDGEMENTS

I am honored to have brilliant and generous individuals mentor me on this project and my adventures in public health. I would like to express my gratitude to my thesis committee members Dr. Jodi Lapidus, Dr. Amy Sullivan, and Dr. Atif Zaman for their guidance and insight on this research project. Dr. Lapidus went beyond the duty of a thesis advisor by actively engaging me, encouraging independent thinking, and piquing my curiosity about statistics at every opportunity. I want to especially thank Dr. Sullivan who not only played a pivotal role in this project but also my career in public health, both of which she initiated. I would like to convey my deep appreciation to Dr. Clyde Dent who participated unofficially on my committee, contributed to my understanding of possible extensions for this method, and significantly enhanced my knowledge of and passion for biostatistics.

I would like to recognize the insightful advice of Dr. Jessina McGregor who tirelessly helped me. I am also grateful for the assistance I received from Alicia Parkman, for assistance in obtaining the study data; Linda Daniels-Cuevas, for providing surveillance estimates on hepatitis morbidity; and Claire Smith, for advice on population estimates.

I would like to thank the Public Health and Preventive Medicine Department and my colleagues at Multnomah County Health Department for their support and contribution to my education. Finally, this research could not have been done without my friends and family particularly Billy Galaxy, Aaron Elman, Yuko Shiroma, and Dr. Harleen Singh.

Abstract

Background: Evaluating surveillance data – such as notifiable disease reports and vital statistics records – over time using direct or indirect rates is a common public health tool for discerning and predicting disease trends. In turn, these assessments may help explain the etiology of health outcome and inform prevention and planning efforts. While rates by time period and age are easily calculated, they may obscure factors that influence disease or mortality risk. Age, period, and cohort (APC) analyses seek to uncover these influences by partitioning trends into components associated with changes over time within a given age group (age effects), time period (period effects), and birth cohort (cohort effects). Because available surveillance data for viral hepatitis predominantly represent prevalent disease, it is difficult to track changes in incidence over time or anticipate the magnitude of the disease burden. Recent studies of viral hepatitis infection in the United States have proposed that generational differences exist in morbidity and mortality of this disease.[1-3] Consequently, an APC analysis of viral hepatitis mortality may contribute information about factors perpetuating infection, suggest whether current trends are likely to be sustained, and guide public health planning efforts.

Objective: To evaluate Oregon viral hepatitis mortality for the presence of cohort effects in the baby boomer generation – individuals born between 1950 and 1965 – by applying the multiphase method, a novel method of APC analysis[4,5]

Methods: Deaths related to viral hepatitis were abstracted from multiple-cause mortality variables from Oregon death certificates for 1995 to 2010 using International Classification of Diseases 9th and 10th revision codes (ICD-9: 070; ICD-10: B15-19, B94.2). These data were evaluated for the presence of cohort effects using the three stage multiphase method: (1) data were assessed using pairwise graphical inspection; (2) log-additive components of age and period were removed using a median polish;[6,7] (3) the remaining cohort effect was separated from error using a linear regression model and its relative magnitude estimated.

Results: Qualitative evidence from the first two steps of the multiphase method suggested the presence of age and cohort effects in viral hepatitis deaths. After removal of the log-additive effects of age and period, estimated viral hepatitis mortality rates remained significantly higher for individuals born 1950–1965 than preceding birth cohorts for 1910–1949 (p=0.03) or subsequent birth cohorts for 1966–1985 (p=0.003). No significant period effects occurred at the population level.

Conclusion: We demonstrate a computationally straightforward method for assessing temporal trends using aggregate data. By applying this method to mortality rates for viral hepatitis, a clear pattern of increase in deaths is discernible between 1950 and1965 compared with the cohorts before and after. These findings may contribute to public health surveillance and planning efforts.

INTRODUCTION

Evaluating disease and mortality patterns over time are common tools in public health. Trend assessments are valuable because they can help predict future increases of disease, understand disease etiology, and inform prevention efforts. Summarizing mortality by rates based on calendar year of death is a frequently employed method of assessing mortality trends. Trend analyses using mortality rates is especially popular among local or state health departments that seek to plan for and address population-level needs, but may lack the resources to conduct more complex analyses. While rates are easily calculated and can provide public health practitioners with information about disease behavior over time, crucial details about the trends may be lost in the averaging process inherent to rates.[8] Because rates are population-level summary measures, factors that may help explain observed patterns over time are often omitted. For instance, mortality rates analyzed over time do not take into account that decedents were born at different times. Generational differences, for instance, may arise when individuals born at the same time are exposed to different levels of a particular risk factor than those born at other times. These generational differences based on differential exposure to a risk factor are known as cohort effects. Understanding whether such cohort effects exist and the extent of their influence using age-period cohort (APC) analysis may provide insight for certain diseases, helping to inform surveillance and program planning.

Historically, analyses that assessed the influence of cohort effects relied on qualitative, graphical approaches.[9-13] Beginning in the second half of the 20th century, more quantitative approaches were developed. These analyses, however, have been criticized for making *a priori*, possibly invalid assumptions about the relationship between age, period, and cohort as well as difficult to interpret. [8,14,15] In this thesis, I propose to explore the multiphase method of cohort analysis. This approach, recently developed by Keyes and colleagues in 2010, provides a straightforward and computational simple method to both assess and quantify cohort effects in trend data.[16] In reviewing and critically examining conceptual and statistical details of the multiphase method, this thesis makes both theoretical and applied contributions to analytic tools used in public health. To accomplish this, I detail the steps of the multiphase method and apply it to 1995–2010 viral hepatitis mortality data for Oregon residents. To facilitate use of this approach in an applied public health setting, I also present resources to support local and state public health departments seeking to employ this method.

Viral hepatitis is a disease for which an age-period-cohort analysis may provide information that could benefit public health efforts not available through traditional analyses of rates. The Centers for Disease Control and Prevention (CDC) estimates that nationally 3.5–5.3 million persons are chronically infected with viral hepatitis and 15,000 persons die annually as a result of chronic liver disease associated with these infections.[17] In Oregon, the extent of new reports of chronic infection – over 6,000 reports annually for the last 5 years – persists despite declines in reports of acute infection (unpublished Oregon surveillance data from ORPHEUS database, Oregon Health Division). Public health efforts to address this large burden of disease are compounded by poor estimates of incidence. True incidence of viral hepatitis is difficult to determine because both acute and chronic infections are asymptomatic. Chronic infections are of particular interest because they lead to serious complications such as chronic liver disease, cirrhosis, and liver cancer – if left untreated. However, most chronically infected individuals typically do not know they are ill and remain undiagnosed for decades post-exposure until sequelae manifest. By the time these individuals are diagnosed, they have been asymptomatic for a long time and, thus, represent prevalent disease. Incidence estimates, however, are needed to determine the risk of contracting the disease, how rapidly that risk is changing, and whether observed trends are likely to be sustained. While an APC analysis of Oregon viral hepatitis mortality will not yield incidence rates, it may contribute information about factors perpetuating the persistence of chronic infection, suggest whether this trend is likely to continue, and provide information to public health programs to guide planning efforts. Basing this analysis on mortality data rather than disease reports will help minimize underreporting of subclinical infection and decrease bias in diagnosis due to access to healthcare. Further, the multiphase method requires many years of data for analysis, which is not available for notifiable disease reports of viral hepatitis in Oregon.

In my evaluation of the multiphase method, I anticipate finding an approach to APC analysis that has limitations, but provides an accessible alternative method to analyze temporal trends. Previous studies in the United States have found that both viral hepatitis infection and mortality disproportionately affect persons aged 45 to 64 year and implicate the baby boomer generation, individuals born around 1946 to 1965.[1-3] Yet these studies have not conducted a formal cohort analysis to assess the cohort effects, their contribution relative to age and period effects, or magnitude. Thus, cohort effects may be present and the baby boomer generation more impacted by viral hepatitis mortality than other birth cohorts. Since surveillance studies have demonstrated that viral hepatitis infections occur most among young adults 20–39 years old and health outcomes tend to develop during middle-age after a long subclinical phase,[18] age effects are also possible. In particular, mortality rates are expected to increase with age and peak around 60 years old. Whether period effects are detectable in the data is less clear. I suspect there may be a slight rise in rates after 1998 due to changes in clinical practices. Secular changes due to risk prevention strategies (e.g., screening the national blood supply, policies to prevent healthcare exposures) and improvements in screening tests may also impact period effects.

Following this introduction, section 2 provides a framework to understand the multiphase method in context with other APC approaches. It reviews the historical development of APC models and supplies an overview of how the multiphase method compares to other analytic techniques. Section 2 also provides background for viral hepatitis morbidity and mortality in the United States by describing acute and chronic hepatitis, related sequelae and the ensuing burden of disease, and limitations of the surveillance. In this section, I discuss why I anticipate finding cohort effects in the baby boomer generation and summarize what previous studies have found. The third section specifies each step of the multiphase method and teases out its methodological details. Additionally, this section describes the mortality data I use for my illustration of the multiphase method including the source of the datasets, data elements used, preparations for conducting the analysis, and statistical tests applied. Section 4 focuses on the results of applying the multiphase method to Oregon viral hepatitis mortality data. The final section concludes by assessing findings from the analysis and performance of the multiphase method.

BACKGROUND

Framework for Age-Period-Cohort Analysis

Concepts and Definitions

Age-period-cohort analysis is used in epidemiology to analyze temporal trends in disease incidence and mortality. Before giving further detail, it is necessary to describe the components integral to this type of analysis: age, period and cohort effects.

Age effects. Individuals' risk of a health outcome often varies through the aging process. Age effects are differences in risk associated with different age groups regardless of the time period or generation to which an individual belongs. These differences may result from accumulated exposure and/or physiologic changes that are experienced differentially at stages in an individuals' life course. For example, bedwetting is highest in children 54-36 months old and its incidence decreases with age.[19] Figure 1 depicts a hypothetical disease rate in which only age effects are present. In this diagram, the disease rate increases linearly across age for each birth cohort.

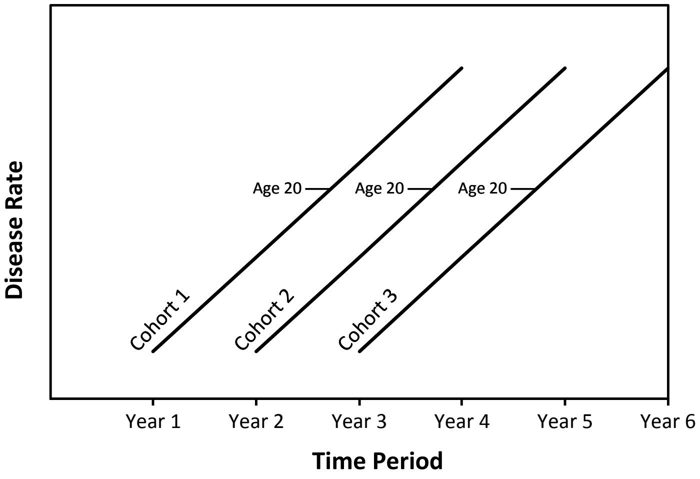


Figure 1. Hypothetical rate of disease in three birth cohorts over time with only age effects operative.
(Reproduced from Keyes and Li, 2011)

Period effects. Period effects result from changes in population-wide exposures to a health outcome that begin at a specific point in time and affect all age groups simultaneously. Increase in cancer incidence among survivors of the nuclear bombing of Japanese cities Hiroshima and Nagasaki in 1945 is an example of a period effect.[20] These effects may also arise artificially – particularly in surveillance and other forms of aggregated data – due to changes in medical technology; data collection; and disease screening, definition, or classification. Figure 2 shows a hypothetical disease rate influenced only by period effects. As depicted, the entire population experiences an increase in disease incidence in year 2 regardless of cohort or age.

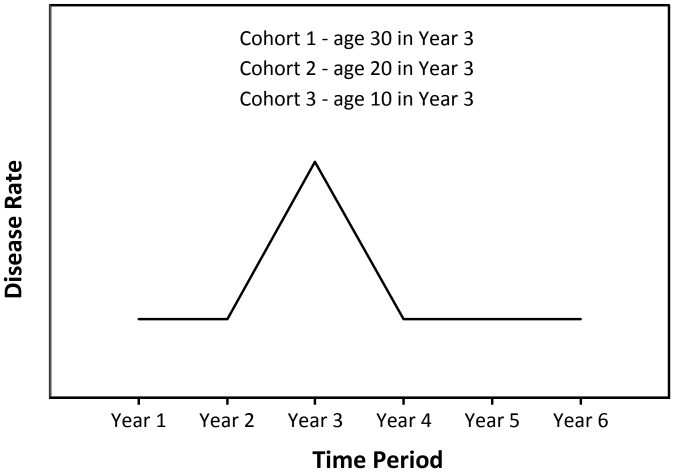


Figure 2. Hypothetical rate of disease in three birth cohorts over time with only period effects operative.
(Reproduced from Keyes and Li, 2011)

Cohort effects. Individuals born at the same time may be exposed to different levels of a particular risk factor than those born at other times. Cohort effects represent changes in the risk of a health outcome based on birth year. For example, the risk of paralysis from poliomyelitis was higher in cohorts born between 1916 and 1955 than those born after 1972 when wild polio was eliminated in the United States following widespread vaccination.[21] If cohort effects exist, it is likely that they will show up when analyzing trends by period (as indirect effects).[8] Figure 3 demonstrates a hypothetical disease rate where only cohort effects exist. In this diagram, each cohort has a disease rate that is constant over age and period; however, the rate for Cohort 1 is lower than either Cohort 2 or 3.

While age, period, and cohort effects are distinct concepts, they are difficult to formally separate because of the relationship between terms. That is, an individual's birth cohort is determined by age within a fixed period. Once a time period has been defined, every age corresponds to a specific birth cohort. For example, say the time period is 2000–2004. Then individuals that are 25–29 years old are born in the 1975–1979 birth cohort, individuals that are 30–34 are born in the 1970–1974 birth cohort, individuals that are 35–39 are born in the 1965–1969 birth cohort, and so forth. Thus, subtracting age from period yields cohort (Cohort = Period – Age). Consequently, period and age will be perfectly correlated for any cohort (i.e., knowing an individual's birth cohort and age completely determines the value of period). Any analysis that seeks to estimate these effects separately must contend with the collinearity of these variables.

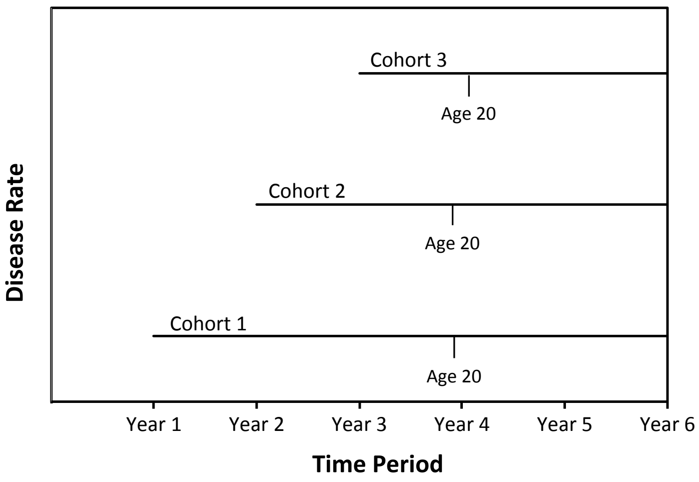


Figure 3. Hypothetical rate of disease in three birth cohorts over time with only cohort effects operative.
(Reproduced from Keyes and Li, 2011)

Historical Use and Statistical Approaches

Age-period-cohort analysis began as a descriptive tool to assess temporal trends in health outcomes and mortality. Early epidemiologic approaches used graphs to follow patterns of disease and mortality over generations.[9,11,12] Especially influential among this early work was the posthumously published paper by Frost (1939) that studied tuberculosis mortality in Massachusetts from 1880–1930. Frost found differences in mortality rates by age with successive birth cohorts, a pattern that was masked when data were examined by age and period. Other seminal epidemiologic studies of the 20th century have identified cohort effects in all-cause mortality,[10] cancer,[13] and peptic ulcer mortality[22] using two-dimensional graphical approaches. While graphs are a useful first step to identify the presence of age, period, and cohort effects, they remain qualitative and limited in their ability to distinguish and quantify the magnitude of each effect. Moreover, only two of the three variables can be examined simultaneously with graphs due to the interdependence of the variables; one effect will always remain uncontrolled.[5,14] Due to limitations in graphical approaches and the desire to estimate the relative contribution of each effect, researchers have sought quantitative approaches to cohort analysis.

In the second half of the 20th century, the sociologist Norman Ryder (1965) proposed that birth cohorts might have structural properties which emerge as a result of the conditions, barriers, and resources that each cohort is born into.[23] These circumstances distinctly shape the patterns and experiences of individuals, potentially impacting health outcomes for that cohort like race or social class. Ryder's publication spurred an interest in assessing the unique influence of cohort membership, independent of age and period effects. In this conceptualization, age and period are seen as confounders of the cohort effect.[24] Thus, statistical models were developed to estimate cohort effects by controlling for age and period effects. In these models, age, period, and cohort are assumed to have a linear relationship with the outcome of interest and each slope is calculated taking into account the additive influence of the other two effects.[16] Statistical models attempting to estimate the three effects simultaneously, however, stumble into an identifiability problem because of the exact linear dependency of the three variables. Recall, cohort = period – age. This collinearity complicates the regression models which are generally used to estimate each effect. Explicitly, the design matrix for regression models will not have full rank, its inverse will not exist, and it will yield a non-invertible estimator when ordinary least squares estimation is attempted due to the perfect correlation of the variables.[14]

Prominence of Ryder's conceptualization of cohort effects and interest in age-period-cohort analysis has engendered a variety of methods to address the identifiability problem. One of the first approaches to be developed as well as the most common is constraint-based regression. This method places one or more restrictions on the regression model to concurrently estimate the effects of age, period, and cohort.[24] A simple constraint may consider only two of the three variables.[25] For instance, a model might assume that there are no changes over time that affect the model and exclude the period term. An alternative constraint to exclusion of an effect altogether is to assume that its slope is zero.[15,25] In this scenario, still using period effects as the constrained term, it is assumed that no linear changes occur in observed rates vis-à-vis time period. As a result, any operative trends are forced to be related to cohort or age terms.[15] Another possible constraint is to equate two of the effects in the model or explicitly define slope parameters.[25] For example, if it is believed that there is no change over a specific interval of time, the slope parameters for two adjacent periods may be set equal to each other.

Constraint-based approaches have been criticized, however, because parameter estimates are sensitive to constraint choice, which is difficult to validate.[8,14] Another widely used approach characterizes trends by their linear components and deviations from linearity.[25,26] These methods are widely used in epidemiology, particularly in cancer research, but have been critiqued as having limited interpretative value. While these two approaches are the most popular, others have been developed. These methods share the conceptualization of cohort effects as obscured by the influences of period and age.[16] In this conceptualization, most common in sociology, cohort membership itself represents an exposure as the experiences of a particular cohort shape its members' patterns of health.[16] Thus, cohort effects represent the totality of environmental influences unique to individuals born during a particular time.[16] Such analyses conceive of the interdependency variables as confounding. That is, age and period effects need to be controlled to assess cohort effects, the real exposure of interest.

An alternate interpretation of the relationship of these variables – one popular in epidemiology – is that age and period interact to produce unique generational experiences.[16] The conceptualization of cohort effects as an interaction of age and period effects was first proposed by Greenberg et al. (1950) in their analysis of syphilis rates of the 1940s.[27] Although this conceptualization still defines the effects to have an exact dependency – that is, cohort = period – age, exposures are not intrinsic to birth cohorts. Rather, a cohort effect "occurs when different distributions of disease arise from a changing or new environmental cause affecting age groups differently. A cohort effect, therefore, is conceptualized as a period effect that is differentially experienced through age-specific exposure or susceptibility to that event or cause (i.e., interaction or effect modification)."[16] In this interpretation, there is no need to address the identifiability problem because cohort effects are not conceived as independent from age and period. The median polish is a statistical method that has been used to explicitly estimate cohort effects using this interpretation.[4,7]

Tukey (1977) developed the median polish to describe data in a two-way contingency table and remove the additive influence of both the row and column variables.[6] This technique makes no assumptions about the distribution or structure of the data in the table. Consequently, it can be used for any data type contained in a table including rates, logarithms of rates, proportions and counts. The method works by alternately subtracting column and row medians from each cell of the table. After several iterations, the cell values stabilize near zero leaving residual values that contain the non-additive components. The residual values measure the deviation of each cell from an additive model. Thus, cells with large residual values indicate potentially important "joint" effects of row and column variables.

This method was first applied to age-period-cohort analysis by Selvin.[7] Looking at mortality data, he conceptualized age and period effects as additive increases to a background mortality rate. That is, the death rate (R_{ij}) for the ith age category and jth year is modeled as

$$R_{ij} = \mu + \alpha_i + \tau_j \qquad (1)$$

where μ is the underlying mortality rate, which occurs at some constant level in the population regardless of other model variables; α_i is the age effects; and τ_j is the period effects. In this model, both age and period independently influence R_{ij}. Age is the only factor that influences mortality when period effects are constant and vice versa. For most scenarios in which both time and age influence mortality, each effect would be added to the background rate. Equation (1) does not account for cohort effects. Because the influence of age is not the same for all time periods and the influence of time is not the same for all ages when a cohort effect exists, Selvin regards this effect as a time/age interaction. Consequently, he modifies the original equation to

$$R_{ij} = \mu + \alpha_i + \tau_j + \gamma_{ij} \tag{2}$$

where γ_{ij} is a cohort effect, an interaction between age and period effects. Then, a cohort effect does not exist independently of age and period effects. With this conceptual model, Selvin applies the median polish to two-way contingency tables for age and calendar year. The polish removes the additive influences of age and time; any remaining residuals represent the possible presence of a non-additive cohort effect. In absence of a cohort effect, the residuals should average to zero.

In his model descriptions and application of the median polish, Selvin does not include random error. To account for this oversight (ϵ_{ij}), equations (1) and (2) are rewritten

$$R_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij} \tag{3}$$

$$R_{ij} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \epsilon_{ij} \tag{4}$$

These revised equations indicate that the residuals resulting from the polished contingency tables actually contain random error in addition to the postulated cohort effect ($\gamma_{ij} + \epsilon_{ij}$). The multiphase method proposed by Keyes and Li (2010) extends Selvin's work by separating the cohort effect and error terms.[16] By applying regression after the median polish, the residuals from the median polish are partitioned into systematic and non-systematic components. The systematic element is considered the cohort effect, the remaining variance the random error. Additionally, the multiphase method parametrically quantifies the relative magnitude of each cohort effect compared to a referent.

Multiphase method of Age-Period-Cohort Analysis

The multiphase method provides a simple, robust, and easily interpretable technique for estimating cohort effects. Compared to other types of APC analysis, this nonparametric method makes minimal assumptions about data structure, data distribution, or the relationship between age, period, and cohort. Because the median polish does not rely on a specific distribution or structure, it is extremely versatile and can be applied to a wide variety of data including rates, log rates, proportions, and counts. Although conceptually similar to the linear contrast method of Holford, the multiphase method provides a quantitative estimate of the age/period interaction.[16] Further, this method explicitly defines cohort effect as an interaction between period and age which facilitates interpretation of the results.[16]

The multiphase method does not address limitations – such as overlapping cohorts and missing data – inherent to all APC analyses that use contingency tables. The problem of overlapping cohorts arises because the multiphase method follows the convention of labeling birth cohorts by subtracting the youngest age from the earliest and latest year in the period interval. For instance, the 1954–1957 birth cohort is formed from the 41–44 age group and 1995–1998 period by subtracting 41 from 1995 and 1998. In this example, some of the individuals categorized into the 1954–1957 birth cohort will actually be born between 1950–1953. Thus, this convention incorrectly classifies some individuals into erroneous birth cohorts. Overlapping cohorts are formed due to this misclassification and mutually exclusive cohort risks cannot be estimated. Missing data is another issue that results from the way cohorts are constructed in the multiphase method. As with other APC methods that employ data from contingency tables, fewer data points are available for the youngest and oldest age categories when cohorts are formed. Because this method calculates cohort effects by averaging the outcome of interest over each birth cohort and age, estimates for cohort categories with sparse data may be more influenced by age effects and, thus, less reliable.[5] These estimates should be interpreted with caution. Like other age-period-cohort analyses, the primary objective of the multiphase method is to assess the component effects of age, period, and cohort that lead to changes in trends over time rather than test causation. Moreover, these factors are likely distant proxies for the true constructs that mediate disease and mortality trends. If more proximal variables can be identified, directly measured, and tested, the resulting analyses may be more methodologically sound and meaningful for public health applications.[5]

Despite limitations, the multiphase method may be useful to identify influences on trends not otherwise apparent in disease or mortality rates. Due to characteristics of the natural history of viral hepatitis, discussed in the next section, incidence rates are difficult to estimate. In this circumstance, analyzing data using the multiphase method may aid in understanding trends. As mentioned, viral hepatitis has been hypothesized to differentially impact the baby boomer generation more than proceeding or subsequent birth cohorts. If mortality increases are demonstrated to be primarily limited to the baby boomers, these findings might suggest that incidence rates are decreasing in other birth cohorts along with the risk of becoming infected. Such information is not obtainable through traditional analyses of rates from surveillance data currently available for this disease.

Motivating Example – Viral Hepatitis Mortality

Hepatitis C is a bloodborne infection caused by an RNA virus and principally transmitted through percutaneous but also mucosal exposure to infected blood.[28,29] The greatest risk factor for hepatitis C is injection drug use[2,30] but another common risk is blood transfusion before 1992 when plasma-derived products began to be screened for the virus. [2,30] Other documented routes of transmission include occupational needle stick exposures,[31 32] inadequate infection control in healthcare settings,[33-35] and high-risk sexual behavior.[2,36-38] Surveillance reports of individuals living with chronic Hepatitis C infection from 2009 from CDC's Emerging Infections Program (EIP) indicate that males and non-Hispanic Whites are disproportionately affected in the United States – 66.3% and 24.7% of reported cases, respectively.[18] The majority of these infections occur in adults 40–54 years old.[18]

Like hepatitis C, hepatitis B virus is transmitted by percutaneous and mucosal exposure. While blood has the highest concentrations of virus, other body fluids such as semen and saliva are also infectious.[39] The primary routes of transmission for the virus are sexual contact and percutaneous exposure to body fluids such as through injection drug use, occupational exposure via needle stick injuries, perinatal exposure to an infected mother, and prolonged close personal contact with an infected person.[18] EIP surveillance of individuals living with chronic hepatitis B infection from 2009 reports an equal proportion of men and women affected while Asians/Pacific Islanders and non-Hispanic blacks represent the majority of cases for whom race/ethnic information was available – 24.8% and 10.3%, respectively.[18] Current literature suggests that most infected Asian/Pacific Islander and non-Hispanic black cases may be immigrants and refugees from endemic areas who likely acquired infection outside of the United

States.[40,41] In contrast to hepatitis C, the age of infected individuals tends to be younger with ages between 25 and 54 years predominating.[18]

Hepatitis B and Hepatitis C virus (referred to together as “viral hepatitis”) can cause both acute and chronic infection. The acute form of viral hepatitis usually occurs within 6 months of exposure and manifests as mild illness that lasts a of couple weeks. Acute infection may be asymptomatic but may also present with nonspecific symptoms such as anorexia, nausea, vomiting, fatigue, abdominal pain, and jaundice.[42] In some individuals, the initial infection resolves spontaneously – usually within the first year. Chronic hepatitis occurs in individuals whose immune system fails to clear the acute infection. The proportion of newly infected individuals that will develop chronic infection is 75–85% for hepatitis C and differs by age for hepatitis B (>90% of infants, 25–50% of children ages 1–5 years old, and 6–10% of adults).[43]

Chronic viral hepatitis is the leading cause of cirrhosis and hepatocellular carcinoma (HCC) in the United States,[44] producing 78% of HCC cases and 57% of cirrhosis cases.[45] Chronic liver disease and cirrhosis have been among the 15 leading causes of death in the United States for over 10 years.[46-55] HCC accounts for approximately 90% of primary liver cancer and this type of cancer has a 5-year survival rate below 12%.[56] Viral hepatitis has been dubbed the “silent killer” because most chronically infected people are asymptomatic until symptoms of advanced liver disease appear many years later after exposure.[39,43] An Institute of Medicine report estimates that 3.5 –5.3 million people in the United States are living with chronic hepatitis B or C infections.[43]

Public health interventions in the past 20 years have transformed the epidemiology of acute viral hepatitis in the United States. Hepatitis B incidence has dropped 82% since 1991 following implementation of national measures to reduce transmission including universal vaccination of infants, prevention of perinatal infection through routine screening of pregnant women and immunoprophylaxis of infants born to infected mothers, routine vaccination of children and adolescents, and vaccination of adults at risk for infection.[28,40] Because no vaccine exists to prevent hepatitis C, public health measures for this disease have focused primarily on prevention efforts such as testing blood donors, treating plasma-derived products, providing risk-reduction counseling and screening to at-risk individuals, and adopting universal infection control procedures.[29] These efforts have led to a 91% decline nationally in the number of estimated new acute hepatitis C infections between 1990 and 2009.[17] Acute hepatitis B rates in Oregon exhibit similar decreases, dropping 94% in the past 20 years (Figure 4A). By contrast, Oregon rates of acute hepatitis C have fluctuated since reporting began in 1994 (Figure 4B) and remained above national estimates.[57] Despite national declines in the incidence of both acute HBV and HCV over the last two decades, the size of the population infected with chronic viral hepatitis persists. Additionally, the prevalence of cirrhosis and HCC continue to grow as do deaths from viral hepatitis.[3,58,59] In Oregon, the rate of new reports of chronic hepatitis B infection has declined only slightly since 2000 from 14.7 to 11.6 per 100,000 (Figure 4A). As for chronic hepatitis C, mandatory reporting of this infection as a notifiable disease began in Oregon in July 2005.[i] Due to the recent addition of this disease to routine surveillance, reported cases may reflect an influx of previously known cases, artificially inflating rates. Consequently, a stable trend is difficult to gauge. State-level estimates for Oregon were not available for cirrhosis and HCC prevalence but deaths related to viral hepatitis have climbed since 1995 (Figure 5).

[i] Some Oregon counties voluntarily reported hepatitis C cases prior to 2005.

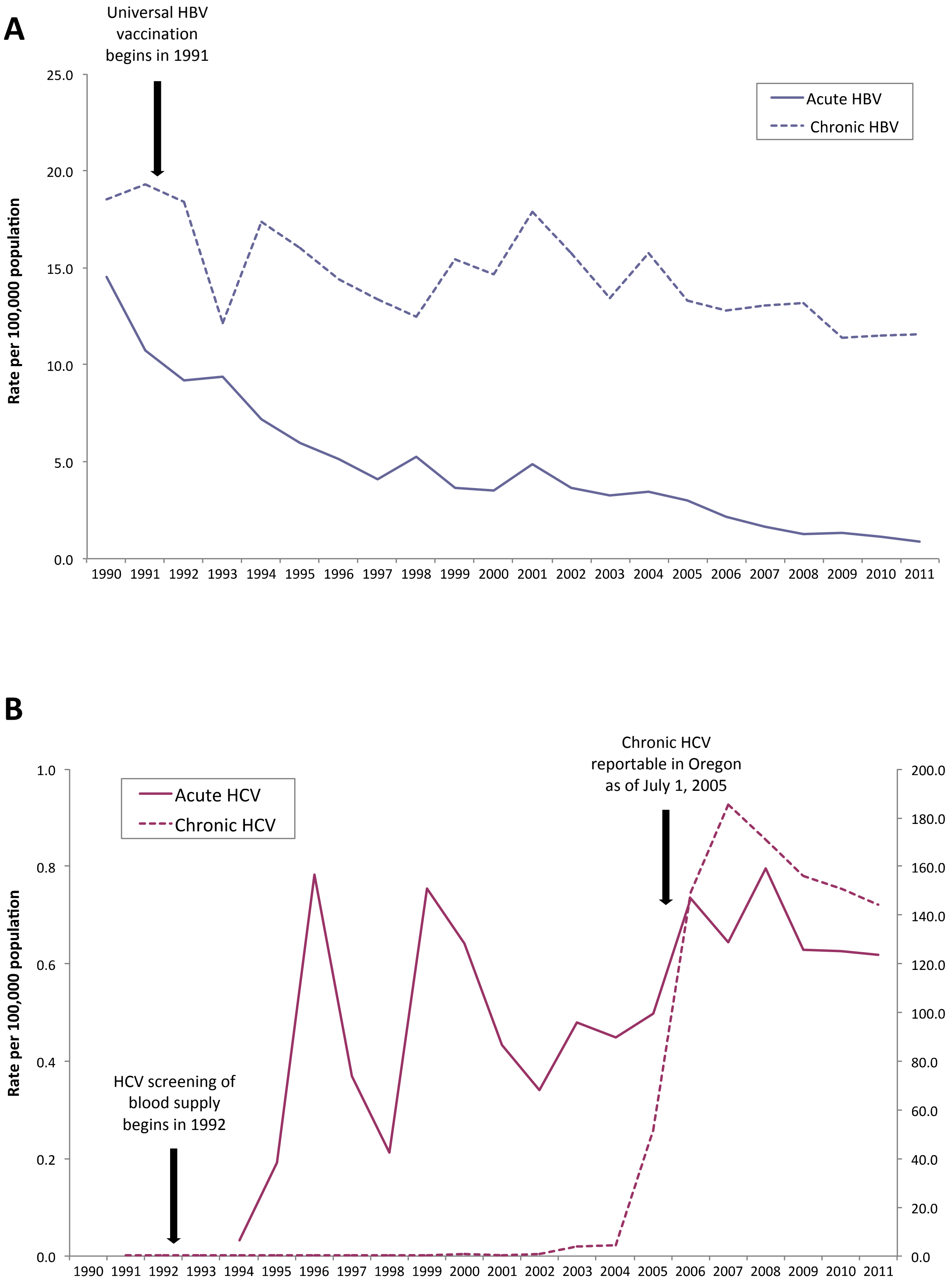


Figure 4.A., New hepatitis B (HBV) reports in Oregon residents, 1990-2011. **B.,** New hepatitis C (HCV) reports in Oregon residents, 1990-2011.

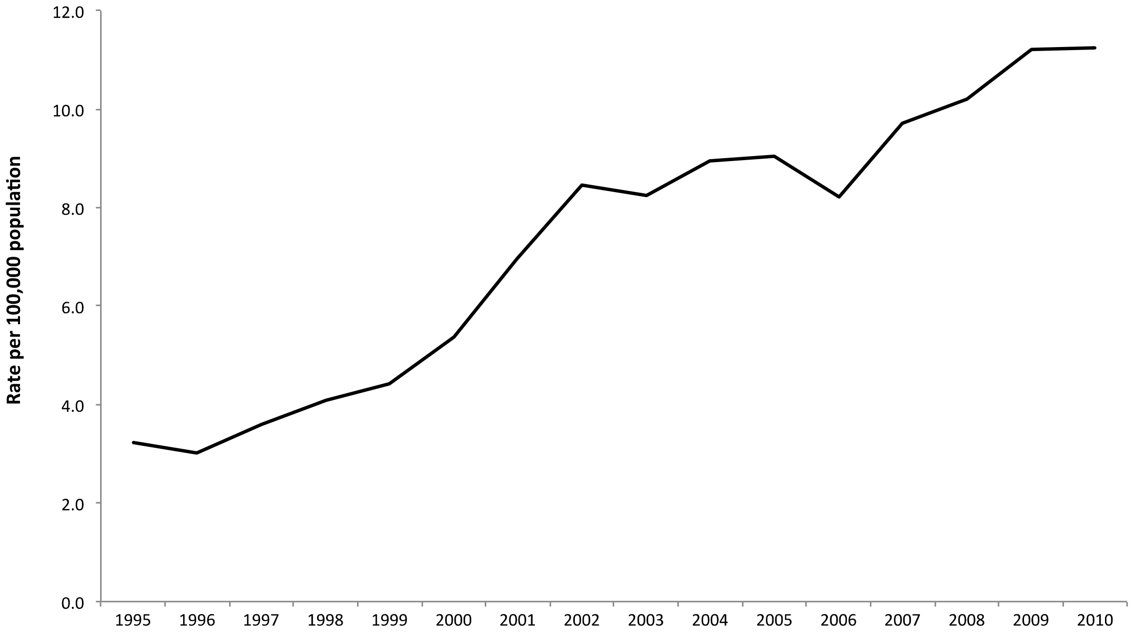


Figure 5. Mortality rates for viral hepatitis in Oregon residents by year, 1995-2010

Simulation models for the United States predict that between 2005 and 2021 the population living with hepatitis C will decrease by 24% from 3.15 million to 2.47 million in 2021.[60] However, the burden of disease is expected to intensify over the same time period. Studies estimate the latent period between viral hepatitis infection and serious complications as 2–4 decades post-exposure (2–3 decades for cirrhosis and 3–4 decades for HCC).[42,61] Because most hepatitis C cases remain asymptomatic until serious complications develop, the population currently infected with hepatitis C will experience high rates of chronic liver disease, cirrhosis, liver cancer, and mortality as it ages.[60,62] Few prediction models are available for estimating the impact of current chronic hepatitis B infections and, existing forecasts are not specific to the United States. Yet, as with hepatitis C, there remains a reservoir of approximately 800,000 chronically infected individuals that will contribute to the viral hepatitis burden as they age.[41,63] Furthermore, 90% of new case reports are immigrants from areas where hepatitis B infection is endemic and are likely exposed to the virus prior to immigrating.[64] Since the majority of these individuals are infected before arriving in the United States, the prevalence of individuals living with chronic hepatitis B infection will probably continue without additional, global public health interventions. As many of these individuals may not have access to healthcare, controlling chronic infection through antiviral treatment and mitigation of long-term consequences may not be possible. As a consequence of these factors, changes in the carcinogenic consequences of viral hepatitis will not be seen for decades after variations in viral hepatitis prevalence.

Age plays a crucial role in viral hepatitis deaths and the baby boomer generation is especially impacted. In a study of participants from the National Health and Nutrition Examination Surveys (NHANES) between 1999 and 2002, the overall prevalence of HCV in the United States was estimated to be 1.6%. The prevalence nearly tripled for individuals ages 40 –49.[2] Armstrong et al. (2000) used mathematical models to estimate the prevalence of hepatitis C infection and graphed estimates at age 60 by year of birth.[1] They found that individuals born around 1945 and 1964 were at greatest risk for acquiring hepatitis C. National serum surveys support these estimates[2] as do mortality rates for 1999–2007.[3] While again less data are available for Hepatitis B, a study of 1999–2008 NHANES data found that prevalence of chronic hepatitis B infection increased with age, peaking in persons 50–59 years old.[65] Other studies have shown that cases of chronic hepatitis B are predominant in immigrant adults in the 40–49 age group.[66,67] Moreover, hepatitis B vaccine programs are too new to impact the prevalence of chronic hepatitis B in adults.[64] While these studies have suggested that a cohort effect likely exists in the baby boomer generation, efforts to evaluate this effect remain primarily graphical, do not quantify its magnitude, or assess the role of age and period effects.[2]

METHODS

Requirements, Set up, and Phases of Multiphase Method

The next section provides a general exposition of the multiphase method. The intent is to provide public health practitioners with a technical guide for applying this approach in their work. In the following section, I will apply the methods outlined here to Oregon viral hepatitis mortality data.

Data Requirements

The multiphase method requires data aggregated into *m* age groups over *n* time periods. The intervals for the age groups and time periods must be of equal width. For example, if ages are grouped by 5-year intervals, so too are the time periods. Also, a minimum of 3 birth cohorts each with at least three cells of age-period data is recommended to conduct this analysis. Figure 6 demonstrates the structure of hypothetical contingency tables with the minimum dimensions for this condition. Counts, rates, and proportions of health-related diseases and conditions are the most common form of this aggregate data for applied public health, but this method does not necessitate a specific type of data (as long as it is consistent within the table).

	τ_1	τ_2	τ_3	τ_4	τ_5
α_1					
α_2					
α_3					

	τ_1	τ_2	τ_3	τ_4
α_1				
α_2				
α_3				
α_4				

Figure 6. Structure of hypothetical contingency tables with the minimum number of dimensions for the multiphase method

Set Up

The multiphase method begins by separating aggregated outcome data into *m* age groups within *n* periods using the same width of interval. Data should be then arranged into a two-way contingency table with *m* rows and *n* columns, a row for each age group and a column for each period (see Figure 7). Each cell designates a rate[ii] corresponding to each $i = 1,\ldots, m$ age group and $j = 1,\ldots, n$ period.

Figure 7. Hypothetical two-way contingency table with *m* rows for age group and *n* columns for period

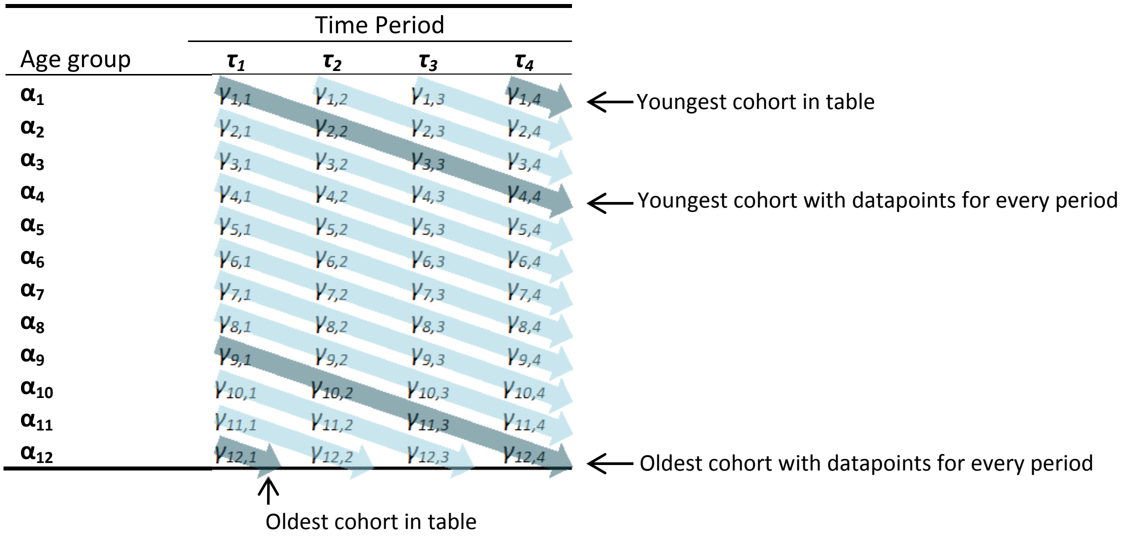
	Period				
Age group	τ_1	τ_2	**. . .**τ_j	**. . .** τ_{n-1}	τ_n
α_1	$\gamma_{1,1}$	$\gamma_{1,2}$	. . . $\gamma_{1,j}$	. . .$\gamma_{1,n-1}$	$\gamma_{1,n}$
α_2	$\gamma_{2,1}$	$\gamma_{2,2}$	. . . $\gamma_{2,j}$	. . .$\gamma_{2,n-1}$	$\gamma_{2,n}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
α_i	$\gamma_{i,1}$	$\gamma_{i,2}$	. . . $\gamma_{i,j}$	. . .$\gamma_{i, n-1}$	$\gamma_{i,n}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
α_{m-1}	$\gamma_{m-1,1}$	$\gamma_{m-1,2}$	. . . $\gamma_{m-1,j}$	. . .$\gamma_{m-1, n-1}$	$\gamma_{m-1,n}$
α_m	$\gamma_{m,1}$	$\gamma_{m,2}$	. . . $\gamma_{m,j}$	. . .$\gamma_{m,n-1}$	$\gamma_{m,n}$

For example, the rate for the first age group and first time period would be placed in the cell for the first row and column ($\gamma_{1,1}$ in Figure 7), the rate for the first age group and second time period would be placed in the cell for the first row and second column ($\gamma_{1,2}$ in Figure 7), and so on. As age and period are grouped using the same time interval, outcomes in the left-to right diagonals represent individuals of approximately the same birth cohort. There will be a maximum of $m + n - 1$ cohort categories for which outcome data exist. Further, the oldest and youngest cohorts each will have only one data point. Figure 8 demonstrates a hypothetical two-way contingency table with $m = 12$ age groups and $n = 4$ periods. As seen in this example, there are $m + n - 1 = 12 + 4 - 1 = 15$ cohorts. Cohorts are labeled by following the previously mentioned convention of subtracting the youngest age from the earliest and latest period in the

[ii] While the multiphase method can be applied to many types of outcome data (e.g., counts, rates, proportions), I anticipate that rates will be the most common and, thus, they will be the focus of this description.

cohort. Again, the 1954–1957 birth cohort is formed from the 41–44 age group and 1995–1998 period by subtracting 41 from 1995 and 1998.

Figure 8. Hypothetical two-way contingency table with age group as rows and time period as columns showing cohorts formed by diagonal cells



	Time Period			
Age group	τ_1	τ_2	τ_3	τ_4
α_1	$\gamma_{1,1}$	$\gamma_{1,2}$	$\gamma_{1,3}$	$\gamma_{1,4}$
α_2	$\gamma_{2,1}$	$\gamma_{2,2}$	$\gamma_{2,3}$	$\gamma_{2,4}$
α_3	$\gamma_{3,1}$	$\gamma_{3,2}$	$\gamma_{3,3}$	$\gamma_{3,4}$
α_4	$\gamma_{4,1}$	$\gamma_{4,2}$	$\gamma_{4,3}$	$\gamma_{4,4}$
α_5	$\gamma_{5,1}$	$\gamma_{5,2}$	$\gamma_{5,3}$	$\gamma_{5,4}$
α_6	$\gamma_{6,1}$	$\gamma_{6,2}$	$\gamma_{6,3}$	$\gamma_{6,4}$
α_7	$\gamma_{7,1}$	$\gamma_{7,2}$	$\gamma_{7,3}$	$\gamma_{7,4}$
α_8	$\gamma_{8,1}$	$\gamma_{8,2}$	$\gamma_{8,3}$	$\gamma_{8,4}$
α_9	$\gamma_{9,1}$	$\gamma_{9,2}$	$\gamma_{9,3}$	$\gamma_{9,4}$
α_{10}	$\gamma_{10,1}$	$\gamma_{10,2}$	$\gamma_{10,3}$	$\gamma_{10,4}$
α_{11}	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,3}$	$\gamma_{11,4}$
α_{12}	$\gamma_{12,1}$	$\gamma_{12,2}$	$\gamma_{12,3}$	$\gamma_{12,4}$

←Youngest cohort in table
←Youngest cohort with datapoints for every period
← Oldest cohort with datapoints for every period
↑
Oldest cohort in table

Phase I: Graphical Representation

The first step of the multiphase method is to graphically assess data from the *m* x *n* contingency table and determine whether age, period, or cohort effects exist. Because only two of the three variables can be examined simultaneously using graphs, three plots should be constructed to gauge the influence of each effect: (1) age by period (age-period), (2) birth year by period (birth year-period), and (3) birth year by age (birth year-age). Using these graphs, it is possible to qualitatively assess which effects are operative. The age-period graph, for instance, allows the influence of period or cohort effects to be considered within each age group for the outcome of interest. Cohort and period effects are unlikely to exist if age-specific rates do not vary over time and it may be futile to proceed with the multiphase method. If age-specific estimates change in a parallel manner during a specific period (Figure 9), period effects may be present. Similarly, cohort effects may be present if age-specific estimates change for certain age

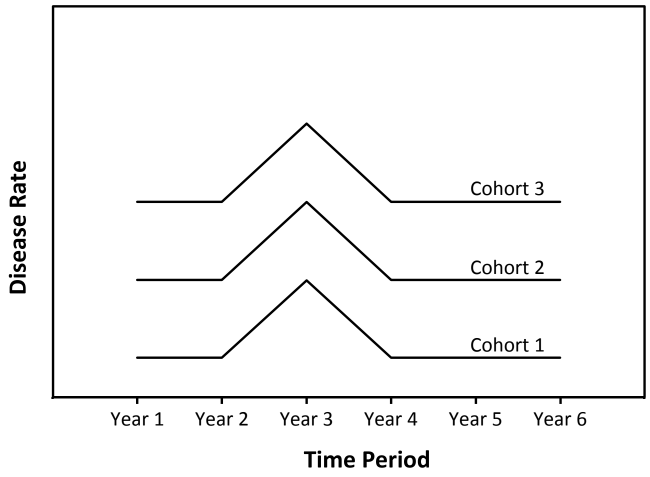


Figure 9. Hypothetical rate of disease in three birth cohorts over time with period and cohort effects operative.
(Reproduced from Keyes and Li, 2011)

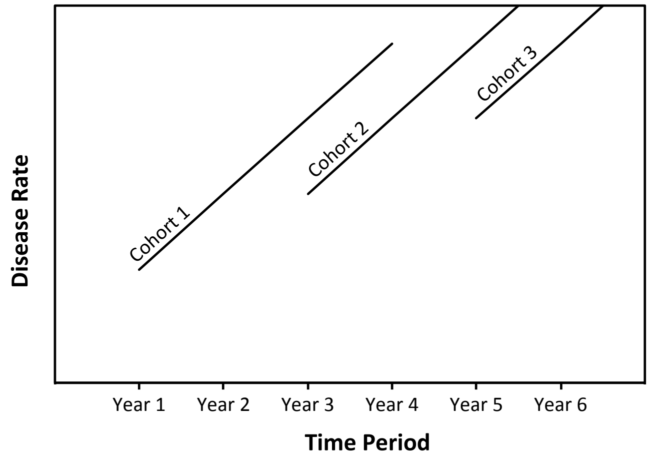


Figure 10. Hypothetical rate of disease in three birth cohorts over time with age and cohort effects operative.
(Reproduced from Keyes and Li, 2011)

groups differently than other groups (Figure 10). Analogous interpretations can be made for cohort-period and cohort-age graphs.

Phase II: Median Polish

The median polish is applied to an *m* x *n* contingency table to remove the additive effects of age and period. This method works by alternately subtracting row (age) and column (period) medians from the table. The detailed steps of the algorithm are:

1. Calculate the median of each row[iii] and record the value to the side of that row.
2. Subtract the row median from each cell in the row.
3. Compute the median of each column and record the value beneath that column.
4. Subtract the column median from each cell in the column
5. Repeat steps 1 and 2 until no change occurs in the row or column medians.

To illustrate the simplicity of this technique, Figure 11 shows one iteration of the median polish for a hypothetical example showing case counts by age and period.

[iii] Instead of rows, the algorithm can start with columns but most pre-coded algorithms begin with rows. The result will be slightly different but similar.

Figure 11. Hypothetical example of one iteration of the median polish with age groups as rows and period as columns demonstrating step 1–4 of the algorithm

Step 1. Calculate median of each row (shaded in blue) and record value to side of the row

		Death Year		
Age group	Period 1	Period 2	Period 3	Row Median
Age 1	3	4	5	4
Age 2	5	4	6	5
Age 3	5	6	5	5

Step 2. Subtract row median from each cell in the row

		Death Year	
Age group	Period 1	Period 2	Period 3
Age 1	-1	0	1
Age 2	0	-1	1
Age 3	0	1	0

Step 3. Compute median of each column (shaded in blue) and record value beneath the column

		Death Year	
Age group	Period 1	Period 2	Period 3
Age 1	-1	0	1
Age 2	0	-1	1
Age 3	0	1	0
Column Median	0	0	1

Step 4. Subtract column median from each cell in the column

		Death Year	
Age group	Period 1	Period 2	Period 3
Age 1	-1	0	1
Age 2	0	-1	1
Age 3	0	1	0

Typically, rows and column medians will not converge perfectly to zero. In most cases, the median polish will stabilize within a few iterations and there will be no appreciable change in the row and column medians. Consequently, algorithms frequently specify a threshold of change (e.g., 0.0001) that is essentially zero or some other stopping rule (e.g., a maximum number of iterations). The values in remaining in the cells after the median polish has completed are the non-additive residuals. They are the components of the original contingency table, which cause it to deviate from a perfectly additive model (as specified in Equation 3). Large residual values indicate less agreement between an additive model and the observed data, indicating potentially important joint effects. In the case of analysis with the multiphase method, these non-additive residuals represent the cohort effects combined with random error. While the multiphase method utilizes the median polish at this stage to separate the additive and non-additive elements of the model, alternative techniques such as quantile regression are possible and will be discussed later.

The rates may be log transformed to evaluate the interaction on the log-additive scale.[4] Applying a log transformation to models of rates is routine in epidemiology and, in some ways, facilitates the analysis. If employed, the log transformation should be applied prior to the median polish. Note that the results of the polish do not depend on the scale of the interaction, but log transformations have some useful properties such as reducing positive skewness by compressing the upper end of the distribution, extending the lower end, and normalizing the residuals around zero. It is worth remarking, however, that log transformation of data that contain zero values may complicate the median polish step. The value of log(0) is negative infinity and, thus, undefined for the purposes of computation. In his explanation of the median polish, Tukey suggests two possible solutions for addressing this obstacle: set zero cells to be lower than any other value in the table (the “easy way”) or add a small constant (e.g., 0.1, 0.25. 0.5, etc.) to all values before log transformation (“the careful way”).[6] Tukey prefers the “careful way” but states that either way usually works well.

Other data transformations (e.g., taking the square root or reciprocal of the original value) are possible but this step cannot be prescribed. Each dataset must be individually assessed for the appropriate transformation. From this point forward, the exposition will assume that the data from the original contingency table have been transformed using the natural logarithm (from this point forward referred to as “log transformed”) as this is a likely procedure for most applications in public health and produces easily interpretable results familiar to epidemiologists. To emphasize, other transformations may be valid and the median polish residuals used for regression in the next step despite the focus on log transformation. In fact, a linear regression may still be the appropriate model if the residuals have a normal distribution (as assumed in the alternate transformations suggested above.)

The residuals are then plotted against the birth cohort categories to qualitatively assess the presence and size of cohort effects. When using log-rates, the residuals will evenly distribute around zero in the absence of cohort effects. Residuals that deviate from zero indicate that age and period are not perfectly additive, suggestive of a cohort effect. Moreover, residuals from the median polish can be subtracted from the cells of the original data to produce a contingency table that reflects only the additive effects of period and age. This data can then be qualitatively compared with the original data to see how removing the cohort effect changes rate estimates. Note that data need to be on the same scale before subtraction is performed. For instance, log data should be exponentiated prior to being subtracted from the original rates.

Phase III: Regression

The final step of the multiphase method is to separate the median polish residuals into the systematic and non-systematic components – the cohort effect and error terms, respectively – by regressing them from the median polish on the cohort categories. Linear regression is a suitable model choice for partitioning residuals from log-transformed rates because they are assumed to be normally distributed; another model choice may be more appropriate for residuals derived from other types of outcome data. Hypothesis testing and contrasts may also be employed in this step to compare the relative magnitude of the effects and test for statistically significant differences among cohorts. The equation for the linear regression model for the residuals of the log-transformed rates is

$$r_k = \mu + \gamma_{k-1} + \epsilon_{ijk} \tag{5}$$

where r_k are the values of the residuals from the median polish for the k = 1, 2,…, m + n - 1 cohort categories; the intercept μ is the referent birth cohort category; γ_{k-1} is the vector of indicator variables for the k - 1 = m + n -2 cohort categories (note that the vector of indicator variables is one less than the cohort categories because the referent category is modeled as the intercept), and ϵ_{ijk} is the vector of random error across i age groups, j periods, and k cohort categories from the median polish and regression models.

This regression produces k parameter estimates – one for each cohort category including the referent. These estimates reflect the log-rate for each cohort relative to a referent for a log-additive model from which age and period effects have been removed.[iv] Thus, exponentiating these log-transformed estimates produces the rate attributable to each cohort category in absence of age and period effects. At this point, rate ratios for each cohort category relative to a referent can be calculated and the size of the cohort effects evaluated. Other hypothesis tests can be performed at this stage to compare different combinations of cohorts. Because these assessments are content and context-specific, it is not possible to provide more direction for conducting them and the details are left to each researcher to determine. All testing should be planned *a priori* rather than *ad hoc*.

[iv] Note that the overall F-test for the regression is unlikely to be significant. The F-test is used to assess the goodness of fit of the regression model. In the multiphase method, estimation rather than fit is the goal.

Finally, the residuals (ϵ_{ijk}) from the regression model should be examined to verify parametric assumptions. To this end, data should be plotted to confirm that observed values are relatively linear and that residuals are roughly normally distributed with equal variance. If parametric assumptions are violated, a couple corrective options may be possible. First, a different model may be used for the third step in lieu of a linear regression model such as parametric or semi-parametric robust regression methods. Another option is to return to the original contingency table and try a different data transformation – like the square root or reciprocal – with the hope that it will produce normally distributed residuals when the median polish is applied.

Application of Multiphase Method to Viral Hepatitis Mortality

Population-level rates were calculated for 4-year intervals for age groups and periods (Table 1). Oregon viral hepatitis mortality data for 1995–2010 was used as the numerator and Oregon population estimates as the denominator.

Table 1. Age groups and periods used for analysis

Age groups	Periods
25–28	1995–1998
29–32	1999–2002
33–36	2003–2006
37–40	2007–2010
41–44	
45–48	
49–52	
53–56	
57–60	
61–64	
65–68	
69–72	
73–76	
77–80	
81–84	
85+	

Mortality data were abstracted from 1995–2010 death certificates obtained electronically from the Oregon Health Authority Center for Health Statistics (Portland, Oregon) via the Multnomah County Health Department. Deaths were restricted to Oregon residents including those who died out-of-state. Individuals under 25 years old were excluded due to the small number of viral hepatitis related deaths in younger ages (less than 10 deaths over the total study period), as were individuals who died in Oregon but resided elsewhere. National Center for Health Statistics (NCHS) considers rates based on fewer than 20 deaths statistically unreliable.[68] While the age groups 25–28, 29–32, 33–36, 37–40, 77–80, 81–84, and 85+ had fewer than 20 deaths for at least one time period, no groups were excluded on this basis. Compiled mid-year estimates of Oregon's population size were obtained for the rates' denominator through VistaPHw software, a web application used by Oregon Health Division and other governmental agencies for community health assessment.[69] Data obtained from VistaPHw included population estimates for 1995–2010 by age from the Portland State University Population Research Center (PSU).

Deaths in the United States require a death certificate and states use the U.S. Standard Certificate of Death (1989 revision for 1995–2005, 2003 revision for 2006–2010) to collect uniform data on the decedent as well as the circumstances and cause of death. The attending physician, medical examiner, or coroner completes the medical portion of the death certificate and often provides identifying information such as name, residence, race, and sex; the funeral director or other person in charge of interment completes the remainder of the document which is mostly demographic, usually with the assistance of a family member of the deceased.[70] Both revisions of the U.S. Standard death certificate have two sections for obtaining information on the cause of death.[71,72] The first part collects the immediate and underlying causes of death, the second part significant conditions that contributed to, but did not result in the underlying cause of death from Part I. The causes listed in these sections are coded according to the *International Classification of Diseases* (ICD). Together, they constitute the multiple-cause of death variables, which encompass up to 20 codes for each year of mortality data. Viral hepatitis deaths were identified from the multiple-cause of death variables using the ICD revision 9 (ICD-9) code 070 and revision 10 (ICD-10) codes B15-B19 and B94.2 (Table 2). Viral hepatitis deaths were categorized into 4-year intervals based on decedents' age and year of death, which yielded 16 age groups and 4 periods (generating 19 birth cohorts). Four-year intervals were employed rather than more typical five year intervals because contributing cause of death variables contributing cause of death variables were not available for data prior to 1994 and a minimum of four time periods were required for applying the multiphase method. Further, deaths identified with ICD-9 codes needed to be grouped separately from those using ICD-10 to allow for the detection of potential period effects resulting from differences in the revisions. Consequently, 4-year intervals were necessary.

Table 2. ICD codes used to identify viral hepatitis deaths

ICD version	ICD code	Description
ICD-9	070	Viral hepatitis
ICD-10	B15-B19	Viral hepatitis
	B94.2	Sequelae of viral hepatitis

To calculate mortality rates, the number of viral hepatitis deaths for each age and period group was divided by the corresponding age group for the Oregon population. First, population estimates were formatted for the same 4-year intervals by age group and period. Because data from VistaPHw were categorized into 5-year age groups, each estimate was divided by 5 then reallocated into 4-year age groups matching those of the intended numerators (see Table 1 for a list of the age groups). Subsequently, population estimates for each age group were averaged into the 4-year period intervals to create denominators for each age and period category. Once computed, rates were arranged into a two-way contingency table and analyzed with the multiphase method.

Statistical analysis. Descriptive statistics were calculated to describe demographic information such as age, sex, race/ethnicity, veteran status, and place of birth as well as the type of hepatitis for decedents whose deaths were related to viral hepatitis; these individuals were compared to those whose cause of death was not associated with viral hepatitis. The same characteristics were also assessed over the four time periods using χ^2 tests to test for differences in categorical variables (e.g., sex, race, veteran status), one-way ANOVA to test for differences in the mean of decedents' age, and Poisson regression to test for a linear increase in mortality rates.

For the first step of the multiphase method, viral hepatitis mortality rates were graphed by age and period (age-period), birth year and period (birth year-period), and birth year and age (birth year-age). The second and third steps of the multiphase method were conducted as described previously using log-transformed rates prior to the median polish step. Results from the median polish were qualitatively evaluated using a scatterplot of the residuals by the cohort categories. To better assess the distribution of the residuals around zero, a loess curve was fit to the data using locally weighted regression (smoothing parameter, 1.0).[73] At this step, viral hepatitis rates were calculated with the cohort effect removed and these rates were graphed by age-period, birth year-period, and birth year-age. Another graph comparing the difference in specific cohorts with and without the cohort effects was also constructed. A linear regression model was run using residuals from the median polish as the dependent variable and design variables for the cohort categories as predictors. Each resulting parameter estimate was exponentiated and compared to the 1922–1925 referent cohort to obtain risk ratio estimates and 95% confidence intervals using linear contrasts. The 1922–1925 cohort was chosen as the referent group because it was the first earliest cohort with data for all of the four time points. Additional hypothesis testing using contrasts was performed to compare the average risk ratio for the 1950-1965 cohorts with proceeding and subsequent cohorts.

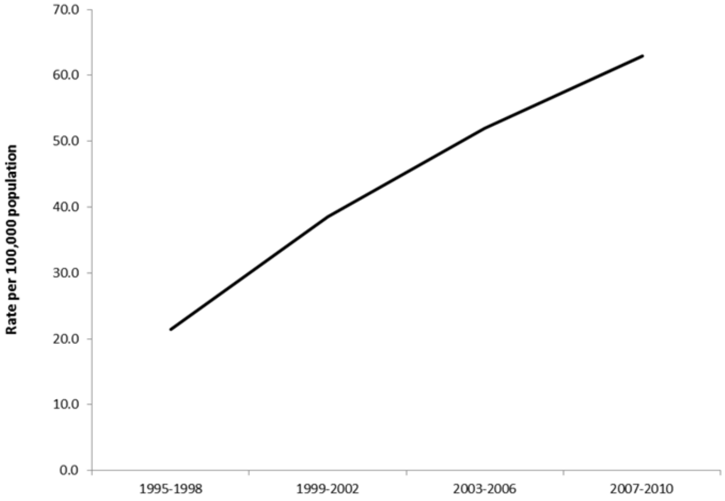
Data management was conducted with SAS (SAS Institute Inc., Cary, NC). The median polish and regression analyses were generated using R (R Development Core Team, Vienna, Austria). The following R packages were used for importing SAS datasets, preparing data for graphing, running regression analyses, and plotting results: *sas7bdat*,[74] *reshape*,[75] *gmodels*, [76] *gplots*,[77] *MASS*,[78] *robustbase*,[79] and *hett*.[80]

Results

Descriptive Analysis of Viral Hepatitis Mortality Rates

Between 1995 and 2010, a total of 473,615 deaths were registered for Oregon residents aged 25 years and older. Of these, 4,162 (0.9%) deaths had viral hepatitis listed as one of the multiple-causes of death. The majority (82.1%) of these deaths were related to hepatitis C. Less common were deaths due to hepatitis B (3.5%); other non-B, non-C hepatitis (4.9%); and unspecified hepatitis (9.5%). Over the study period, the viral hepatitis mortality rate for all ages increased 194% (Poisson test for trend, p<0.0001; Figure 12). Compared with deaths due to other causes, viral hepatitis deaths occurred overall in individuals that were younger (mean 55.7 vs 75.5 years; t-test with Satterthwaite adjustment, p<0.001), male (70.4 vs 49.0%, p<0.001), non-white (14.7 vs 4.7%; χ^2 test, p<0.001), and foreign born (6.9 vs 5.7%; χ^2 test, p=0.001). No difference was found between rates of mortality rates for hepatitis and other causes for veterans (29.0 vs 29.1%; χ^2 test, p=0.83).

Figure 12. Viral hepatitis mortality rates in Oregon residents 25 years and older, 1995-2010



The characteristics of decedents whose deaths were related to viral hepatitis by period are shown in Table 3. In summary, there were no significant changes in sex (χ^2 test, p=0.34), race/ethnicity (χ^2 test, p=0.28), or country of birth (χ^2 test, p=0.57) over the four periods. Decedents' age at death, veteran status, and type of hepatitis, however, did vary significantly (p <0.001 from one-way ANOVA, p=0.002 from χ^2 test and p<0.001 from χ^2 test, respectively) over this time period.

Table 3. Characteristics of Oregon residents with viral hepatitis-related deaths, 1995–2010

	1995–1998		1999–2002		2003–2006		2007–2010	
Characteristic	(n=446)		(n=867)		(n=1240)		(n=1611)	
Mean age (SD)	55.3	14.5	54.3	12.0	55.0	10.1	57.1	9.2
Sex								
Male	301	67.5	623	71.9	863	69.6	1144	71.0
Female	145	32.5	244	28.1	377	30.4	467	29.0
Unknown/Missing	0	0.0	0	0.0	0	0.0	0	0.0
Race/Ethnicity								
Non-Hispanic White	378	84.8	729	84.1	1050	84.7	1396	86.7
Not White	68	15.3	138	15.9	189	15.2	215	13.4
Unknown/missing	0	0.0	0	0.0	1	0.1	0	0.0
Ever in armed services								
Yes	130	29.2	269	31.0	364	29.4	431	26.8
No	314	70.4	593	68.4	863	69.6	1157	71.8
Unknown/missing	2	0.5	5	0.6	13	1.0	23	1.4
Birth place								
United States	388	87.0	786	90.7	1122	90.5	1486	92.2
Not United States	35	7.9	55	6.3	76	6.1	112	7.0
Unknown/Missing	23	5.2	26	3.0	42	3.4	13	0.8
Hepatitis type								
Hepatitis B	79	17.7	16	1.9	20	1.6	31	1.9
Hepatitis C	347	77.8	638	73.6	924	74.5	1509	93.7
Other hepatitis	5	1.1	71	8.2	61	4.9	68	4.2
Unspecified hepatitis	15	3.4	142	16.4	235	19.0	3	0.2

Cohort Analysis of Viral Hepatitis Mortality with Multiphase Method

Graphical Representation

Table 4 displays the contingency table for age-specific viral hepatitis mortality rates. This table displays the unadjusted data used for initial graphical assessment. Based on the data from the contingency table, the overall hepatitis mortality rate increased linearly between 1995 and 2010 from 21.5 deaths per 100,000 population to 63.0 deaths per 100,000 population. Rates were highest in ages 49–64, peaking at 176.0 deaths per 100,000 population in 2007–2010 among ages 57–60.

Table 4. Viral hepatitis mortality by age and year of death*

	Death Year			
Age group	1995–1998	1999–2002	2003–2006	2007–2010
25–28	1.83	2.72	0.51	0.48
29–32	2.26	2.13	2.02	1.93
33–36	8.75	7.57	6.41	5.85
37–40	16.71	21.82	15.29	8.32
41–44	29.67	35.87	35.82	26.29
45–48	33.52	76.42	69.41	58.31
49–52	28.68	79.47	127.97	119.03
53–56	23.84	70.49	152.21	157.53
57–60	21.45	58.68	89.13	176.04
61–64	28.91	45.46	54.66	111.99
65–68	30.37	41.55	50.29	70.95
69–72	28.39	31.79	50.07	44.46
73–76	35.68	40.26	42.60	51.40
77–80	33.01	26.74	35.55	48.76
81–84	32.15	45.04	31.83	32.85
85+	15.97	31.52	28.01	27.86
Total	21.46	38.51	51.92	62.99

*Rate per 100,000 population.

Viral hepatitis mortality rates are graphed by age across each time period in Figure 13. The most volatility in rates occurs in the 49–68 age groups. Rates for these ages start around 26.7 deaths per 100,000 population then rise. Among these groups, the greatest increase was in ages 57–60 which experienced a 7.2% increase between the first and last time periods. Less dramatic changes occurred in the other age groups. The rates for the 25–36 age groups were relatively constant and experienced less than 10 deaths per 100,000 population over the entire timeframe. Deaths in the remaining age groups increased after 1998 (with the exception of the 77–80 age group) then continued to modestly grew (ages 45–48, 69–72, 73–76, 77–80), decline slightly (age groups 37–40 and 45–48), or stabilize (age groups 41–44, 81–84, and 85+). Many of the age-specific trends appear non-parallel since mortality rates increased faster in some age groups than others, suggestive of cohort effects. Because of less variability in rates in the 1995–1998 period, which increases in subsequent years, period effects may be present. However, if existent, these effects do not appear to be consistent for the entire population.

Figure 13. Viral hepatitis mortality rates by age and period

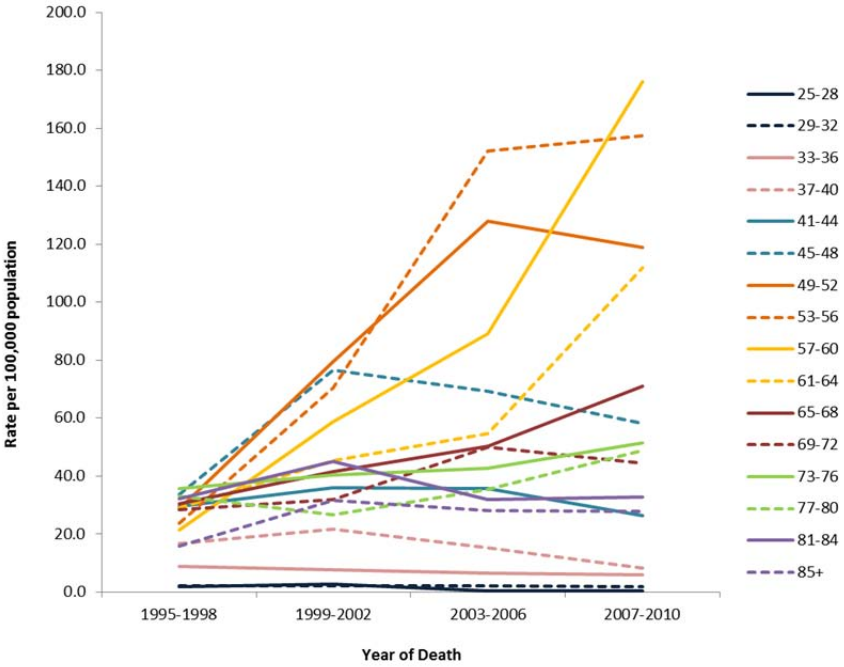


Figure 14 demonstrates that patterns observed by age and period do not directly translate to birth cohorts. In the birth-year-period graph, the most growth is seen in individuals born between 1946 and 1961 but rates are also rising in the 1942–1945 and 1962–1969 cohorts. Rates in the remaining cohorts are increasing slightly or relatively flat. Compared to the previous graph, there appears to be a clearer distinction between mortality trends experienced by individuals born in 1950–1961 compared to other years.

Figure 14. Viral hepatitis mortality rates by birth year and period

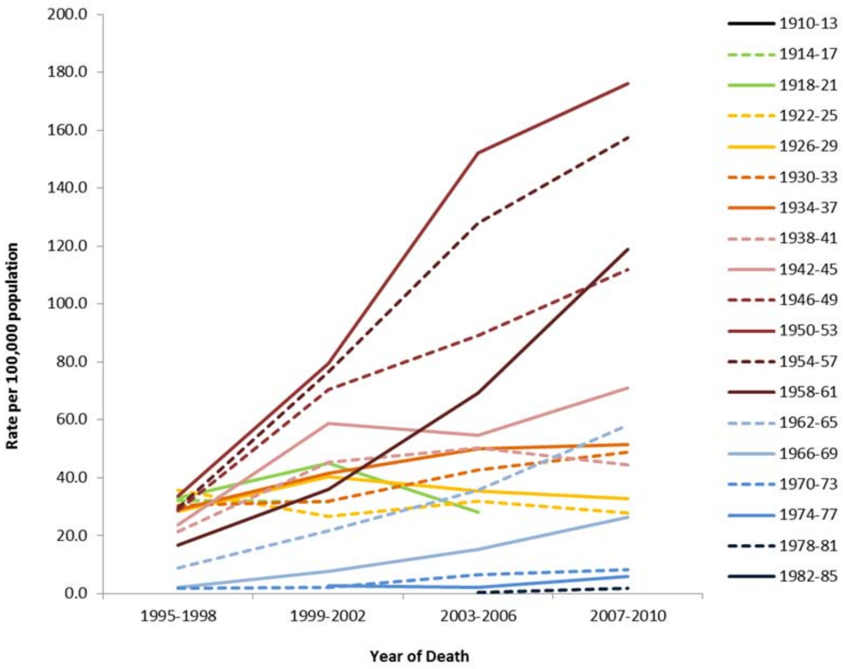
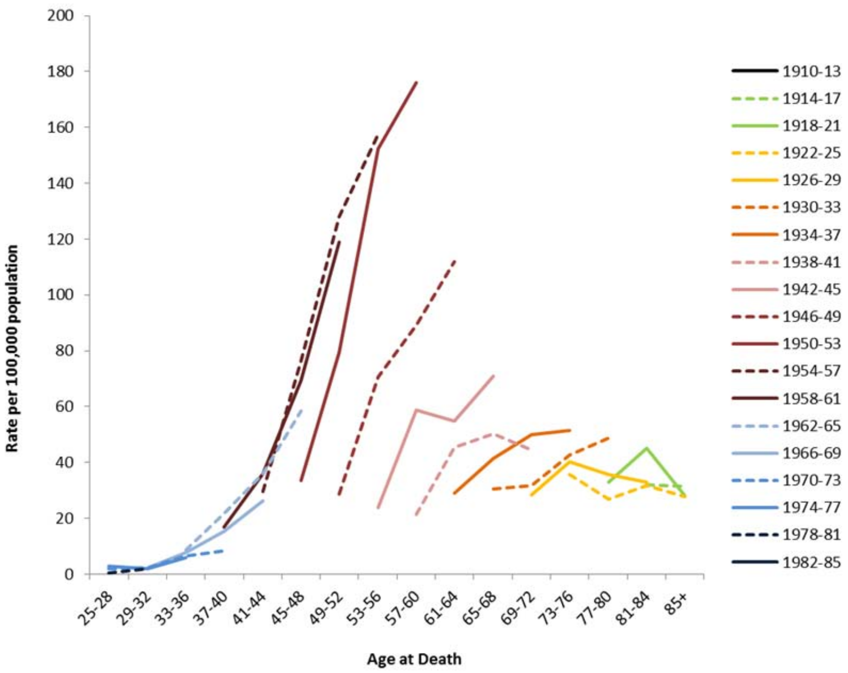


Figure 15 displays trends based on age and birth year. Viral hepatitis mortality reaches its apex when the 1950–1953 cohort is 57–60 years old. Age-specific rates in the 1950–1953, 1954–1957, 1958–1961, and 1962–1965 cohorts track similarly. Although the 1946–1949 cohort has a similar shape to the aforementioned groups, the magnitude of rates is much less especially for ages 53–56 and 57–60. The shape of trends for age-specific rates in the other cohorts manifest substantially differently.

Figure 15. Viral hepatitis mortality rates by birth year and age



Median Polish

The median polish method was applied to the log transformed rates for overall viral hepatitis-related mortality (Table 5). The resulting residuals are plotted in Figure 16. From this scatterplot, the pattern of the residuals across cohorts appears to systematically differ from zero for cohorts between 1942 and 1965. The slight but sustained positive increase of residuals for 1950–1965 (Figure 16, blue shading) indicate that the non-additive component (i.e., cohort effect plus error) of these rates appears greater in these years than anticipated by a purely additive model.

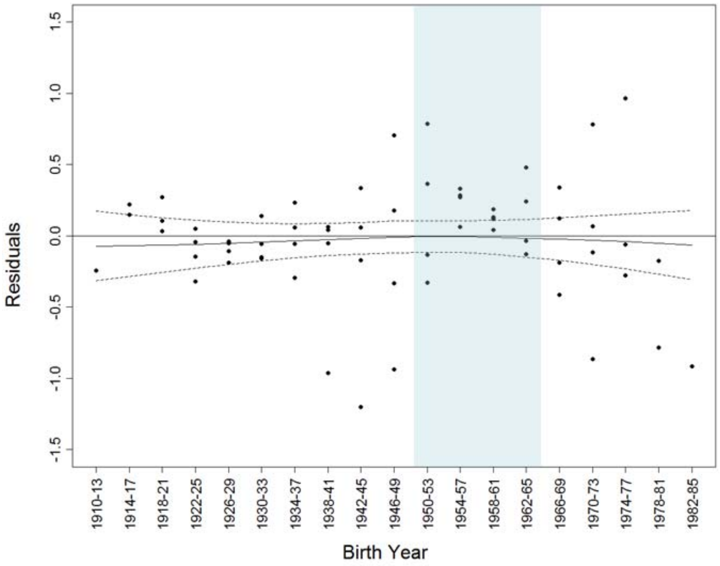
Table 5. Log transformed viral hepatitis mortality by age group and period*

	Period			
Age group	1995–1998	1999–2002	2003–2006	2007–2010
25–28	0.78	0.97	-0.78	-0.91
29–32	0.34	0.07	-0.06	-0.17
33–36	0.48	0.12	-0.11	-0.28
37–40	0.19	0.24	-0.19	-0.87
41–44	0.06	0.04	-0.03	-0.41
45–48	-0.33	0.29	0.12	-0.13
49–52	-0.94	-0.13	0.27	0.13
53–56	-1.20	-0.33	0.37	0.33
57–60	-0.96	-0.17	0.18	0.79
61–64	-0.29	-0.05	0.06	0.71
65–68	-0.16	-0.06	0.06	0.34
69–72	-0.05	-0.15	0.23	0.04
73–76	0.05	-0.04	-0.06	0.06
77–80	0.11	-0.32	-0.11	0.14
81–84	0.15	0.27	-0.15	-0.19
85+	-0.24	0.22	0.03	-0.04

*Rate per 100,000 population.

Additional graphs were created based on rates with the cohort effect removed. These graphs were made by subtracting the residuals from log-transformed rates and exponentiating the difference. For instance, the viral hepatitis mortality rate for ages 57–60 in the period 2007–2010 was 176.0 deaths per 100,000 population whereas the log-transformed rate from the median polish for the same cell was 0.79 (compare Tables 4 and 5). Thus, it is possible to estimate the rate for this cell with the cohort effect removed with the following calculation: exp(log(170.0)-(0.79))=79.88.[v] Performing this calculation for each cell of the original contingency table yields a table with perfectly additive estimates of the mortality rates (that is, in the hypothetical scenario where only age and period effects are operative). Note that these tables are only valid for assessing observed trends when no cohort effect exists. If the absence of cohort effects has been established, however, it may be simpler to skip this additional qualitative step and refer directly to the age-period graph created in the first step of the multiphase method (e.g., Figure 13).

Figure 16. Residual values from median polish by birth year for viral hepatitis mortality rates



*Note non-referent lines represent a loess curve (solid line) and 95% confidence interval (dotted lines; smoothing parameter, 1.0)

[v] Note that the number of significant digits for each element of this calculation especially impacts the precision of the solution because of the logarithm.

Figures 17, 18, and 19 show age-period, birth year-period, and birth year-age plots. These graphs are analogous to those in the previous section (Figures 13-15) except that the non-additive component consisting of the cohort effect and error has been removed. Comparing Figure 17 to Figure 13, it appears that a slight period effect may be present in the data as several of the age-specific mortality rates – for example, in the 45–68 age groups – begin to increase after 1998. However, not all age groups experience this change – for example, ages 33–40 change little if at all. If an effect exists, it appears not to be differential and does not impact the entire population. Also, while the 49–68 age groups still have the highest mortality rates, the rates are now parallel (Figure 17 compared to Figure 13) which is consistent with a purely additive model in which cohort effects are absent. Mortality rates are attenuated in Figure 18 compared to Figure 14 and differences in birth year correspond with increases seen among age groups; no clear trend in period effects is discernible. Finally, birth cohort-age graph (Figure 19) demonstrates a peak in viral hepatitis mortality between ages 53–56 for the available birth years.

Figure 17. Viral hepatitis mortality rates by age and period with cohort effect removed

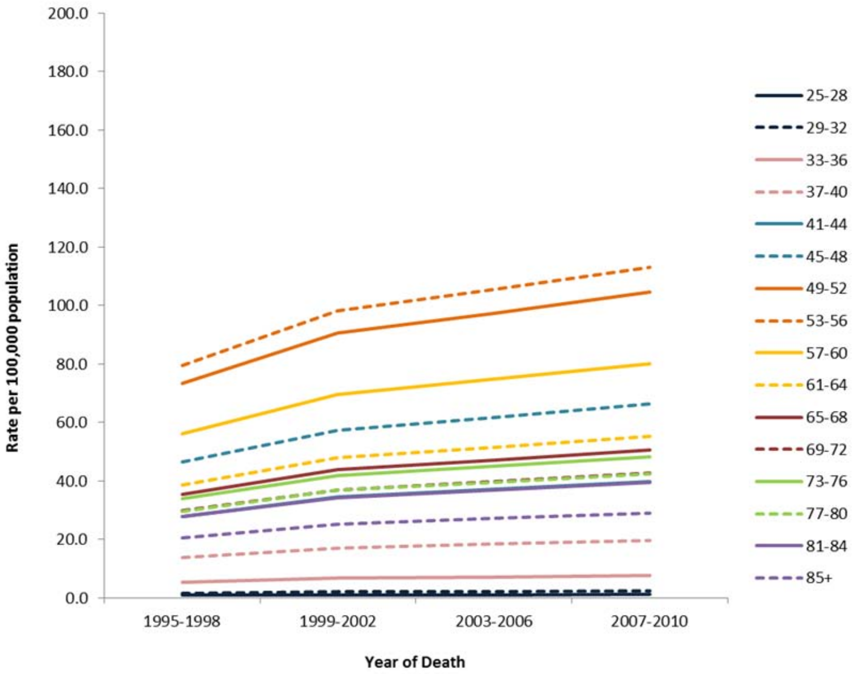


Figure 18. Viral hepatitis mortality rates by birth year and period with cohort effect removed

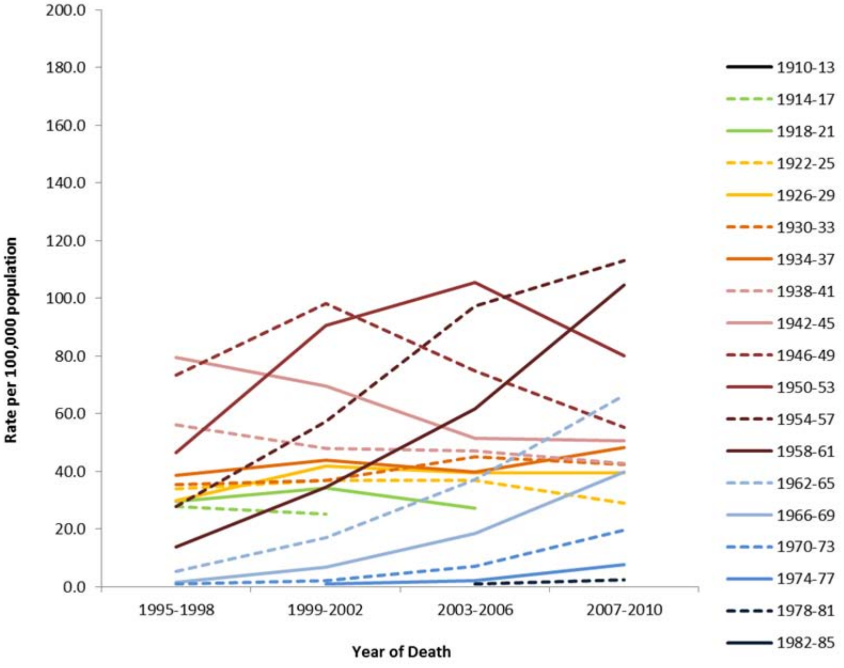
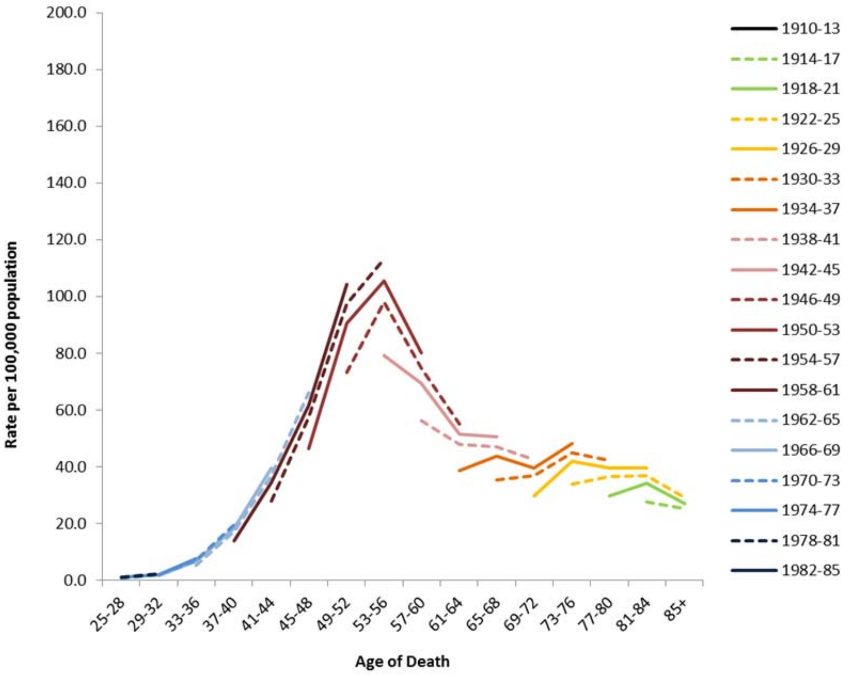
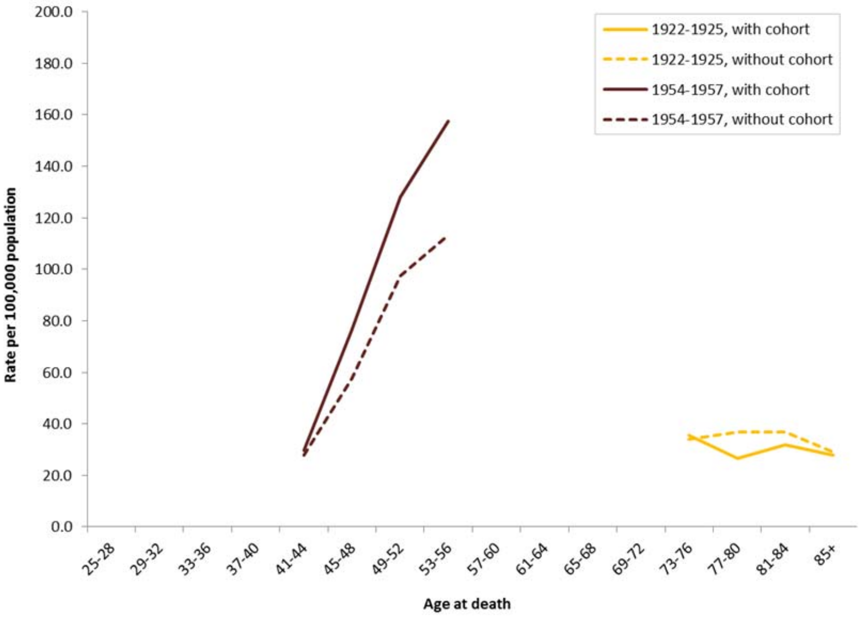


Figure 19. Viral hepatitis mortality rates by birth year and age with cohort effect removed



To directly compare rates with and without cohort effects, birth year-specific mortality rates from the original contingency table and from the same table with the multiplicative components removed were graphed together. Figure 20 shows viral hepatitis mortality rates for the 1922–1925 and 1954–1957 cohorts with and without the cohort effect. Removing the cohort effect for those born in 1922–1925 has little impact on the mortality rate except for a small increase for older individuals. Trends in mortality rates for individuals in these birth years are well described by the age and period effects alone. Comparatively, there is more evidence of a cohort effect in viral hepatitis mortality for individuals born between 1954 and 1957. When the cohort effect is subtracted, the rate drops for all but the youngest age group. The graph suggests that a purely additive age-period model would underestimate rates for these birth years.

Figure 20. Viral hepatitis mortality rates for individuals born in 1954-1957 and 1922-1925 with and without the cohort effect removed



Regression

There were no differences in viral hepatitis mortality risk for individual cohort categories (Table 6 and Figure 21). Although the point estimates for the risk ratio exceeded one for some birth years, none were statistically significant. However, as hypothesized, the average mortality risk for the generations born between 1950 and 1965 was significantly higher than preceding birth cohorts for 1910–1949 (p=0.03) and subsequent birth cohorts for 1966–1985 (p=0.003). The relative risk of viral hepatitis mortality in the baby boomer generation was 1.3 (95% confidence interval (CI): 1.0–1.6) times greater than older cohorts and1.5 (95% CI: 1.2–2.1) times greater than younger cohorts.

Table 6. Estimated risk ratio and 95% confidence interval for the effect of cohort on viral hepatitis mortality

Birth Year	Risk ratio	95% Confidence Interval	
1910–13	0.98	0.30	3.20
1914–17	1.51	0.55	4.20
1918–21	1.44	0.55	3.77
1922–25	1.00	Reference	
1926–29	1.14	0.45	2.90
1930–33	1.19	0.47	3.02
1934–37	1.24	0.49	3.15
1938–41	1.00	0.39	2.54
1942–45	0.98	0.39	2.50
1946–49	1.14	0.45	2.90
1950–53	1.50	0.59	3.80
1954–57	1.59	0.63	4.05
1958–61	1.42	0.56	3.59
1962–65	1.45	0.57	3.67
1966–69	1.21	0.48	3.08
1970–73	1.22	0.48	3.09
1974–77	1.55	0.59	4.06
1978–81	0.78	0.28	2.16
1982–85	0.50	0.16	1.64

The value of R^2 for the linear regression model was 0.26, indicating that cohort effects explain approximately 26% of the variance in the age-period interaction while the remaining variance is unexplained. Note that this is 26% more variation than explained by a model with age and period alone. The residuals from the linear regression are plotted by cohort in Figures 21 and 22. These residual plots do not appear to have major deviations from the parametric assumptions. While some of the residuals demonstrate more variation, it falls within ± 1 standard deviation (Figure 23).

Figure 21. Estimated risk ratio and 95% confidence intervals for the effect of cohort on viral hepatitis mortality

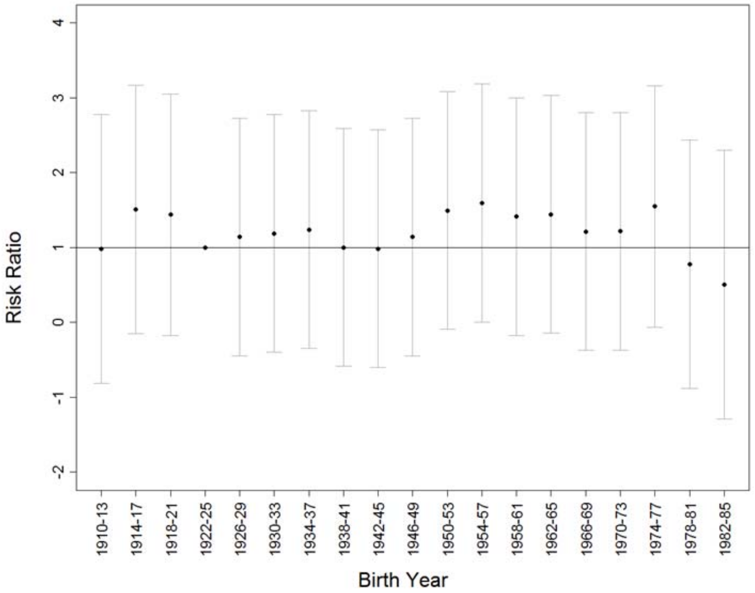


Figure 22. Residual error from linear regression model by birth year

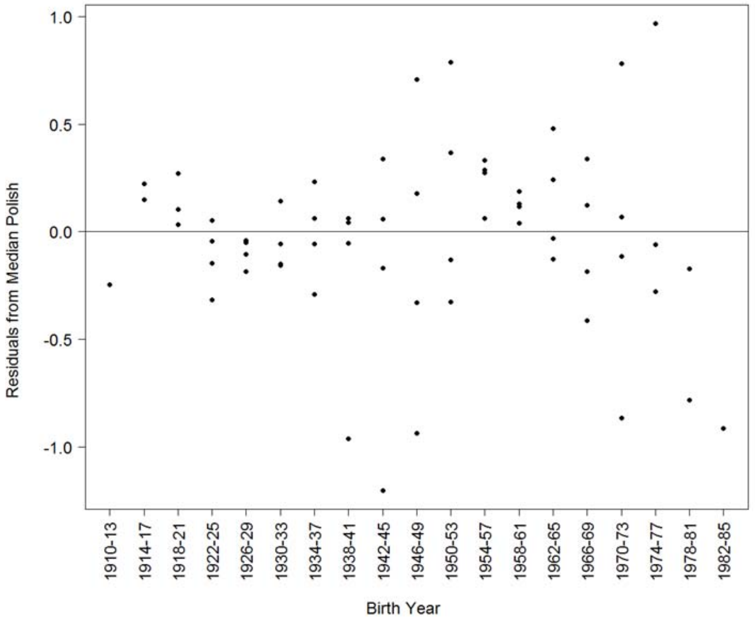
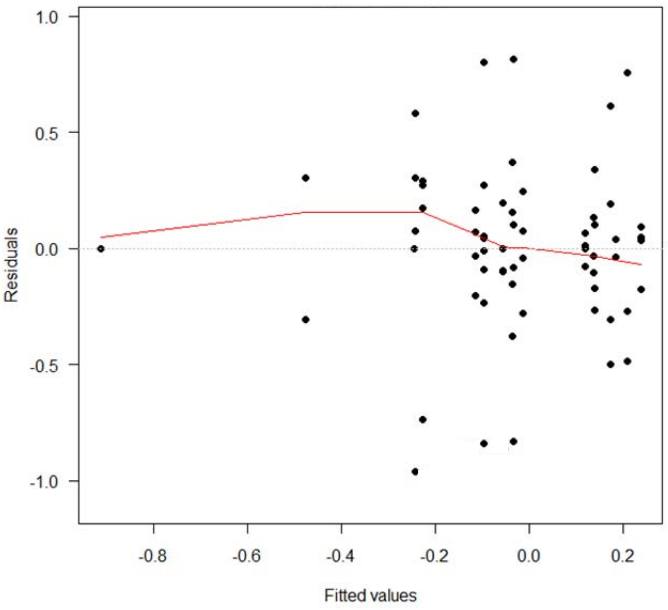


Figure 23. Residual error by fitted values



DISCUSSION

Application of Multiphase Method to Viral Hepatitis Mortality

Cohort analysis using the multiphase method indicates that differences exist in the risk of Oregon viral hepatitis mortality by birth year. Specifically, baby boomers born between 1950 and 1965 had increased risk of dying from viral hepatitis – especially hepatitis C which represents the majority of viral hepatitis deaths[vi] – than other generations. These findings correspond with other studies that propose that individuals born between 1940 and 1965 have the greatest risk of infection from hepatitis C.[1-3] Armstrong et al. (2000) postulate that baby boomers were likely infected in their 20s and 30s during 1970–1990, a period of high hepatitis C incidence.[1] Given the risk factors for both hepatitis B and C transmission, it is likely that the majority of these individuals were exposed to the virus when they experimented with injection drug use or received blood transfusions prior to routine screening of blood products. Although hepatitis B represents a small proportion of viral hepatitis deaths in Oregon, two additional risk factors should be considered for this these infections: high-risk sexual behaviors (e.g., multiple sex partners) and immigration trends from countries where infection is endemic.

Although not quantitatively assessed in the analyses, age effects appear to have a strong influence on viral hepatitis deaths. Risk was highest in the 49–64 year old age groups and lowest in both the youngest (25–40) and oldest (≥80) age groups. This pattern was true regardless of the inclusion of cohort effects. Because age effects are likely, the birth cohort following the 1950–1965 cohort may be at increased risk of hepatitis mortality. For instance, the

[vi] In addition to hepatis C cases representing the majority of cases in each time period, it is likely that the "unspecified" viral hepatitis deaths listed in Table 3 are primarily attributable to hepatitis C.

rates for the 1966–1969 cohort are only 9.5 deaths per 100,000 less than the 1962–1965 cohort for the 41–44 age group. Analysis of adjacent groups is limited by the number of years of period data that are available. Based on the pattern of increased deaths in the 49–60 age groups, it is possible that the 1966–1969 cohort will catch up with the baby boomers as they age. Also, the findings of this study suggest there might be some heterogeneity among risk of viral hepatitis deaths within the baby boomer generation. Signs of a cohort effect were weak in the 1946–1949 cohort compared to other baby boomer cohorts even though rates in this group appeared elevated overall.

If the results of this analysis are reliable, state and local health departments in Oregon will likely to continue to see high rates of viral hepatitis mortality at least through 2029 (this calculation is based on cohort 1950–1965 being at greatest risk and the number of deaths peaking in the 49–64 year old age, thus 1965+64=2029). Epidemiologists should monitor mortality rates in subsequent cohorts to see whether this increase is likely to be sustained. Individuals born between 1966 and 1969 are of particular interest as they progress into the high risk years in middle age. Consequently, it might be worth duplicating this analysis in a few years when more years of data are available. As for program planning, public health officials should consider recommendations that encourage providers to screen individuals born in 1942–1969 – the cohorts at highest risk and those adjacent to them – for viral hepatitis, particularly hepatitis C. This suggestion is supported by other analyses and may prove cost effective.[3,81] Officials should anticipate a continuing health and economic burden as individuals infected with chronic viral hepatitis are diagnosed and need additional health services such as antiviral therapy and liver transplantation.

A subtle additive effect appeared in qualitative stages of the analyses and differentially affected age groups. It is unclear whether to consider this influence a period effect because it does not operate on the population level and, thus deviates from the standard definition of this concept. Theories about this observed variation must also explain why it impacts some age groups and birth cohorts more than others. For instance, it is unlikely to have resulted from the shift between ICD-9 and -10 coding unless coding changes differently impacted hepatitis ascertainment among age groups.

A number of limitations may influence these results. Chiefly, some of the viral hepatitis rates were based on less than 20 deaths, making the estimates potentially unreliable. Unfortunately, there was no way to increase the sample size for these age groups while keeping the analysis targeted on Oregon. To exclude these age groups would have meant fewer cohorts containing all time points and potentially compromised the analysis. Further, it is possible that the analysis contains survival bias for the oldest cohorts. Based on results of the present analysis, it appears that most viral hepatitis deaths occur between ages 49 and 64. Assuming this risk has not changed over time, it is likely that the majority of individuals from earlier cohorts – those born in birth years before 1938 – died during these high risk age groups. Deaths in these individuals would have occurred prior to the study years and would not be included in this analysis. If this supposition is true, then risk in the older cohorts may be underestimated.

Where possible, sources of bias have been anticipated and minimized. Misclassification of variables from death certificates including cause of death may have occurred. While I anticipate that misclassification due to age or year of death is likely to be negligible and random, it is possible that ascertainment of viral hepatitis deaths may be incomplete due to misclassification and/or underreporting on the death certificates. I anticipated the underreporting of this condition and sought to reduce the effect as much as possible by identifying viral hepatitis-related deaths based on multiple rather than single cause of death variables. However, research that assessed viral hepatitis by multiple cause of death from mortality records – including a study for Multnomah County, largest county in Oregon with a large proportion of reported cases – concludes that viral hepatitis mortality was underestimated nevertheless.[82,83] Misclassification of deaths associated with viral hepatitis would result in an underestimate of the true magnitude of death rates. In interpreting the results of the multiphase method, I also assume that these biases have been relatively constant over time. This assumption, however, may be optimistic. Although screening for hepatitis C became available in 1989 and I chose to begin my analysis in 1995 after it had been available for several years, methods for detecting both hepatitis B and C have improved over the study period.[84,85]

Changes in cause of death coding occurred in 1999, further suggesting that our findings must be viewed with caution. A revision in mortality coding occurred when death certificates in the United States switched from ICD-9 to ICD-10 coding. In planning this project, I deliberately separated time periods between 1998 and 1999. Consequently, any impact related to differences in coding should show up as a period effect, which was not evident overall in the data. In addition, I carefully weighed the decision to implement a comparability ratio to bridge the ICD-9 and ICD-10 data. Comparability studies conducted by the NCHS have shown that the classification of viral hepatitis was impacted by coding revisions and viral hepatitis was more likely to be selected as the underlying cause of death in ICD-9 than in ICD-10.[86] According to the NCHS study, the decrease in viral hepatitis classification following the revision was mostly attributable to viral hepatitis being considered a consequence of HIV in ICD-10 but not ICD-9 (i.e., HIV tended to be listed as the cause of death on death certificates rather than viral hepatitis). The study provided a comparability ratio between ICD-9 and ICD-10 codes to aid studies in bridging differences in ICD revisions for assessing trends across years. I opted not to apply the comparability ratio in my analysis. First, the comparability study was conducted only with the underlying cause of death variable whereas I planned to use multiple causes of death. As mentioned previously, I anticipated that multiple cause of death variables would better capture deaths for which viral hepatitis was the contributory but not primary cause of death. This decision was bolstered by preliminary analysis of the Oregon mortality data which showed 33.8-52.6% of deaths related to viral hepatitis were classified by multiple cause of death (Table 7). The observed increase in the number of hepatitis-related deaths by using multiple cause of death is reasonable based on the epidemiology of the disease as individuals infected with chronic hepatitis are likely to die of sequelae (e.g., cirrhosis, HCC) as opposed to the infection itself. Secondly, Oregon has a relatively low number of persons living with HIV/AIDS relative to the rest of the United States.[87] As a result, I determined that the comparability correction was not be appropriate for regional data being assessed.

Table 7. Viral hepatitis deaths identified using single and multiple cause of death variables, Oregon Residents, 1995–2010

	1995-1998 (n=447)		1999-2002 (n=870)		2003-2006 (n=1244)		2007-2010 (n=1611)		Total (n=4172)	
Death variable	n	%	n	%	n	%	n	%	n	%
Single cause	212	47.4	528	60.7	824	66.2	905	56.2	2469	59.2
Multiple cause	235	52.6	342	39.3	420	33.8	706	43.8	1703	40.8

Population estimates from PSU were used as the denominator of the mortality rates. Although other population estimates exist (e.g., from the U.S. Census Bureau), these projections were preferred as a source of denominator data. Because these estimates are calculated using both regional and national data,[88] these data an attractive source of Oregon and county-level population estimate as attested by their wide use in local and state governments. However, these estimates tend to be slightly lower than the estimates from the U. S. Census Bureau. If these estimates undercount the population, mortality rates will be overestimated.

Multiphase Method

The multiphase method is a straightforward and robust way to conduct a cohort analysis with trend data provided that the premise that cohort effects are an interaction of age and time period conceptually makes sense. Given this assumption, this approach provides a clever resolution to the identifiably problem which plagues other APC techniques. To recapitulate the process, the median polish is used to partition additive (i.e., background rate + age effects + period effects) and non-additive (i.e., cohort effects + random error) components of the model for the observed rates. Afterward, regression is used to separate the systematic (i.e., cohort effect) and non-systematic (i.e., random error) components. This approach is especially attractive because the median polish step is nonparametric, makes minimal assumptions about the data being analyzed, and may be applied to diverse types of data. Furthermore, the multiphase method allows for cohort effects to be detected as well as quantified. When rates are log-transformed prior to the median polish, estimates of cohort effects are output as rate ratios, which are relatively easy to interpret and familiar to epidemiologists.

Modifications to the multiphase method are possible. For instance, the median polish is not the only technique to separate additive and non-additive elements. A mean polish or quantile regression could also be employed. However, as mentioned previously, its flexibility and robustness for multiple types of data are an advantage of the median polish. Analyses that substitute another procedure should consider what assumptions are made, what will be gained, and what (if anything) will be lost. Similarly, techniques such robust regression can be used in the final step of the multiphase method instead of a linear model. In fact, such approaches may be more appropriate for the final step of the multiphase method, depending on the distribution of the residuals from the previous step.

While a useful tool for public health practitioners attempting to delve into temporal trends beyond age and period, the multiphase method is not without limitations. For instance, as developed by Keyes and Li, this technique does not provide a way to control for predictors or easily assess subpopulations. With the current approach, stratification is the only means to conduct an analysis of a subpopulation. The contingency table must be set up for each stratum individually before the multiphase method can be applied. To conduct an analysis of viral hepatitis mortality in males, for example, rates for males by age group must be calculated and a corresponding contingency table constructed prior to the median polish or regression steps. Further, there are limitations in comparing analyses between subpopulations. While rates between analyses can be informally compared, it is not possible to directly compare the risk estimates from different models without additional statistical techniques. Depending on the number of events for the health outcome of interest in the dataset, the reliability of rates also may be decreased.

Another shortcoming of the multiphase method is that the precision of rates is lost when data are evaluated with the median polish. That is, a rate of 2.5 per 1,000 population based on 5 cases in a population of 2,000 individuals at risk is equivalent to one based on 500 cases in a population of 200,000 individuals at risk. It does not take into account the differences in the standard error for each measurement of 2.5 per 1,000 population – 0.0011 and 0.00011, respectively. It is possible that future refinements of the multiphase method could address this drawback. By altering the median polish and regression steps to include precision estimates, the analysis could be weighted and, thereby, allow rates based on smaller numbers of events to have less influence in the analysis.

The multiphase method also requires a long period of data with an ample number of age groups to obtain a good estimate of the cohort effects. The minimum number of time periods and age groups to adequately power a cohort analysis using the multiphase method has not yet been determined, but it appears that the current analysis came close to the minimum. The addition of additional years of data would probably increase the power of the analysis and permit comparison of additional cohorts.

APC analyses that use contingency tables to compute birth cohorts require that the length of age and period intervals be equal. The multiphase method is no different in this regard. When age and period intervals are uneven, birth cohorts will be indistinct. Table 8 demonstrates a contingency table with unequal age group and period interval widths. In this table, the age groups are arranged into 5-year intervals and periods into 4-year intervals. The result is data that cannot be separated into cohorts. Take for example the “cohort” starting with the age group 25–29 year olds in 1995–1998 and extending downwards to the right. The cells diagonally read 1970–1973, 1969–1972, 1968–1971, and 1967–1970. Clearly, no distinct cohort is formed. The same will be true in the reverse scenario, where period is wider than age group.

Table 8. Contingency table with unequal age and period intervals

	Period							
Age group	1995–1998		1999–2002		2003–2006		2007–2010	
25–29	1970	1973	1974	1977	1978	1981	1982	1985
30–34	1965	1968	1969	1972	1973	1976	1977	1980
35–39	1960	1963	1964	1967	1968	1971	1972	1975
40–44	1955	1958	1959	1962	1963	1966	1967	1970
45–49	1950	1953	1954	1957	1958	1961	1962	1965
50–54	1945	1948	1949	1952	1953	1956	1957	1960
55–59	1940	1943	1944	1947	1948	1951	1952	1955
60–64	1935	1938	1939	1942	1943	1946	1947	1950
65–69	1930	1933	1934	1937	1938	1941	1942	1945
70–74	1925	1928	1929	1932	1933	1936	1937	1940
75–79	1920	1923	1924	1927	1928	1931	1932	1935
80–84	1915	1918	1919	1922	1923	1926	1927	1930
85+	1910	1913	1914	1917	1918	1921	1922	1925

As described previously, the multiphase method has two additional limitations inherent to APC analyses that use contingency tables – the problems of overlapping cohorts and missing data. I will briefly re-describe each and comment on their significance. First, overlapping cohorts result from the labeling convention for birth cohorts. During this process, some individuals are misclassified into erroneous birth cohorts. Mutually exclusive cohort risks cannot be estimated because each cohort has individuals assigned to it who should be properly designated to preceding or subsequent cohorts. Second, missing data will always impact the youngest and oldest cohort categories. Fewer data points are available for these groups because they are formed by the diagonal cells of the contingency table. When the diagonal cells do not span every age groups and periods on the contingency table, the corresponding cohort category will have fewer than the maximum number of data points. The sparse data points result in less reliable estimates for the affected cohort categories.

SUMMARY AND CONCLUSIONS

Despite the limitations outlined, the multiphase method provides information useful for understanding disease trends in situations where differences in generations exist as was hypothesized for viral hepatitis. This technique not only allows for detection of cohort effects but also for their quantification. Furthermore, such analysis may provide results and insight distinct from the tools typically used for trend assessment. In the motivating example presented here, cohort analysis suggests that individuals born around 1950–1965 may be at higher risk of death related to viral hepatitis, particularly hepatitis C, than those born at other times. Assessing these results in context with apparent age effects indicates that current levels of morbidity and mortality due to viral hepatitis will continue for at least a decade. If these findings are accurate, public health departments have additional information to help guide planning and prevention efforts.

Refinements of the multiphase method should be explored. First, modifications to the each step in the two stage modeling process – i.e., the median polish and regression phases – are possible and may provide desirable characteristics. For instance, robust regression may be used instead of linear regression and better address issues of heteroskedasticity in heavy-tailed distributions. Second, methods to incorporate the precision of the underlying rates should be evaluated. Third, future research should assess whether the revisions to the modeling process – for example, replacing the median polish with quantile regression – would faciliate analysis of subpopulations or allow for controlling covariates. Any proposed modifications, however, should be evaluated and their contributions to improving the method weighed with possible limitations.

In conclusion, this thesis was written with the intention of elucidating the multiphase method in the hope of adding an analytic tool to the field of applied public health. There are other areas in public health beside viral hepatitis that may benefit from age-period-cohort analysis and for which the multiphase method may prove both accessible and illuminating. In maternal and child health, for instance, understanding the relative contributions of maternal age, time period, and mother's birth cohort may help identify etiologic factors for the increasing incidence of low birth weight infants. Public health practitioners should consider the multiphase method alongside other techniques when devising an analysis to assess trends.

REFERENCES

1. Armstrong G, Alter M, McQuillan G, Margolis H. The past incidence of hepatitis C virus infection: implications for the future burden of chronic liver disease in the United States. *Hepatology.* Mar 2000;31(3):777-782.

2. Armstrong G, Wasley A, Simard E, McQuillan G, Kuhnert W, Alter M. The prevalence of hepatitis C virus infection in the United States, 1999 through 2002. *Ann Intern Med.* May 16 2006;144(10):705-714.

3. Ly KN, Xing J, Klevens RM, Jiles RB, Ward JW, Holmberg SD. The increasing burden of mortality from viral hepatitis in the United States between 1999 and 2007. *Ann Intern Med.* Feb 21 2012;156(4):271-278.

4. Keyes KM, Li G. A multiphase method for estimating cohort effects in age-period contingency table data. *Ann Epidemiol.* Oct 2010;20(10):779-785.

5. Keyes KM, Li G. Age–Period–Cohort Modeling in *Injury Research*. In: Li G, Baker SP, eds: Springer US; 2012:409-426.

6. Tukey J. *Exploratory data anlaysis* Reading, MS: Addison-Wesley Publishing Company; 1977.

7. Selvin S. *Statistical analysis of epidemiologic data.* Vol 17. New York: Oxford University Press; 1996.

8. Holford TR. Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annu Rev Public Health.* 1991;12:425-457.

9. Andvord K, Wijsmuller G, Blomberg B. What can we learn by following the development of tuberculosis from one generation to another? *The International Journal of Tuberculosis and Lung Disease.* 2002;6(7):562-568.

10. Case R. Cohort analysis of mortality rates as an historical or narrative technique. *Br J Prev Soc Med.* Oct 1956;10(4):159-171.

11. Frost W. The age selection of mortality from tuberculosis in successive decades. 1939. *Am J Epidemiol.* Jan 1 1995;141(1):4-9; discussion 3.

12. Kermack WO, McKendrick AG, McKinlay PL. Death-rates in Great Britain and Sweden. Some general regularities and their significance. *Int J Epidemiol.* Aug 2001;30(4):678-683.

13. Macmahon B, Terry WD. Application of cohort analysis to the study of time trends in neoplastic disease. *J Chronic Dis.* Jan 1958;7(1):24-35.

14. Kupper LL, Janis JM, Karmous A, Greenberg BG. Statistical age-period-cohort analysis: a review and critique. *J Chronic Dis.* 1985;38(10):811-830.

15. Robertson C, Gandini S, Boyle P. Age-period-cohort models: a comparative study of available methodologies. *J Clin Epidemiol.* Jun 1999;52(6):569-583.

16. Keyes KM, Utz RL, Robinson W, Li G. What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971-2006. *Soc Sci Med.* Apr 2010;70(7):1100-1108.

17. Centers for Disease Control and Prevention. Disease burden from viral hepatitis A, B, and C in the United States. http://www.cdc.gov/hepatitis/PDFs/disease_burden.pdf. Accessed March 5, 2012.

18. Centers for Disease Control and Prevention. Viral Hepatitis Surveillance - United States, 2009. 2011; http://www.cdc.gov/hepatitis/Statistics/2009Surveillance/PDFs/2009HepSurveillanceRpt.pdf. Accessed May 16, 2012.

19. Butler RJ, Heron J. The prevalence of infrequent bedwetting and nocturnal enuresis in childhood. A large British cohort. *Scand J Urol Nephrol.* 2008;42(3):257-264.

20. Sakata R, Grant EJ, Ozasa K. Long-term follow-up of atomic bomb survivors. *Maturitas.* Jun 2012;72(2):99-103.

21. Nathanson N, Kew OM. From emergence to eradication: the epidemiology of poliomyelitis deconstructed. *Am J Epidemiol.* Dec 1 2010;172(11):1213-1229.

22. Susser M. Period effects, generation effects and age effects in peptic ulcer mortality. *J Chronic Dis.* 1982;35(1):29-40.

23. Ryder NB. The cohort as a concept in the study of social change. *Am Sociol Rev.* Dec 1965;30(6):843-861.

24. Mason KOM, W.M., Winsborough HH, Poole WK. Some methodological issues in cohort analysis of archival data. *American Sociological Review.* 1973;38(2):242-258.

25. Holford TR. Analysing the temporal effects of age, period and cohort. *Stat Methods Med Res.* 1992;1(3):317-337.

26. Clayton D, Schifflers E. Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Stat Med.* Jun 1987;6(4):449-467.

27. Greenberg BG, Wright JJ, Sheps CG. A technique for analyzing some factors affecting the incidence of syphilis. *Journal of the American Statistical Association.* 1950;45(251):373-399.

28. Daniels D, Grytdal S, Wasley A. Surveillance for acute viral hepatitis - United States, 2007. *MMWR Surveill Summ.* May 22 2009;58(3):1-27.

29. Centers for Disease Control and Prevention. Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. *MMWR Recomm Rep.* Oct 16 1998;47(RR-19):1-39.

30. Briggs ME, Baker C, Hall R, et al. Prevalence and risk factors for hepatitis C virus infection at an urban Veterans Administration medical center. *Hepatology.* Dec 2001;34(6):1200-1205.

31. Tomkins SE, Elford J, Nichols T, et al. Occupational transmission of hepatitis C in healthcare workers and factors associated with seroconversion: UK surveillance data. *J Viral Hepat.* Mar 2012;19(3):199-204.

32. Wicker S, Cinatl J, Berger A, Doerr HW, Gottschalk R, Rabenau HF. Determination of risk of infection with blood-borne pathogens following a needlestick injury in hospital workers. *Ann Occup Hyg.* Oct 2008;52(7):615-622.

33. Savey A, Simon F, Izopet J, Lepoutre A, Fabry J, Desenclos JC. A large nosocomial outbreak of hepatitis C virus infections at a hemodialysis center. *Infect Control Hosp Epidemiol.* Sep 2005;26(9):752-760.

34. Thompson ND, Perz JF, Moorman AC, Holmberg SD. Nonhospital health care-associated hepatitis B and C virus transmission: United States, 1998-2008. *Ann Intern Med.* Jan 6 2009;150(1):33-39.

35. Macedo de Oliveira A, White KL, Leschinsky DP, et al. An outbreak of hepatitis C virus infections among outpatients at a hematology/oncology clinic. *Ann Intern Med.* Jun 7 2005;142(11):898-902.

36. Sexual transmission of hepatitis C virus among HIV-infected men who have sex with men--New York City, 2005-2010. *MMWR Morb Mortal Wkly Rep.* Jul 22 2011;60(28):945-950.

37. van de Laar TJ, Paxton WA, Zorgdrager F, Cornelissen M, de Vries HJ. Sexual transmission of hepatitis C virus in human immunodeficiency virus-negative men who have sex with men: a series of case reports. *Sex Transm Dis.* Feb 2011;38(2):102-104.

38. Tohme RA, Holmberg SD. Is sexual contact a major mode of hepatitis C virus transmission? *Hepatology.* Oct 2010;52(4):1497-1505.

39. Weinbaum CM, Williams I, Mast EE, et al. Recommendations for identification and public health management of persons with chronic hepatitis B virus infection. *MMWR Recomm Rep.* Sep 19 2008;57(RR-8):1-20.

40. Kim WR. Epidemiology of hepatitis B in the United States. *Hepatology.* May 2009;49(5 Suppl):S28-34.

41. Davis GL, Roberts WL. The healthcare burden imposed by liver disease in aging Baby Boomers. *Curr Gastroenterol Rep.* Feb 2010;12(1):1-6.

42. Thomas DL, Di Bisceglie AM, Alter HJ, Terrault NA. Understanding the natural history of chronic HBV and HCV infections. *J Fam Pract.* Apr 2010;59(4 Suppl):S17-22.

43. Colvin H, Michell A, eds. *Hepatitis and Liver Cancer: A National Strategy for Prevention and Control of Hepatitis B and C*. Washington, DC: National Academies Press; 2010.

44. El-Serag H, Mason A. Rising incidence of hepatocellular carcinoma in the United States. *N Engl J Med.* Mar 11 1999;340(10):745-750.

45. Perz JF, Armstrong GL, Farrington LA, Hutin YJ, Bell BP. The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *J Hepatol.* Oct 2006;45(4):529-538.

46. Miniño A, Arias E, Kochanek KD, Murphy SL, Smith BL. Deaths: FInal Data for 2000. *National Vital Statistics Reports.* 2002;50(15):1-120.

47. Arias E, Anderson R, Kung H, Murphy S, Kochanek K. Deaths: Final Data for 2001. *National Vital Statistics Reports.* 2003;52(3):1-116.

48. Kochanek KD, Murphy SL, Anderson RN, Scott C. Deaths: Final Data for 2002. *National Vital Statistics Reports.* 2004;53(5):1-116.

49. Hoyert DL, Heron MP, Murphy SL, Kung H-C. Deaths: Final Data for 2003. *National Vital Statistics Reports.* 2006;54(13):1-120.

50. Miniño A, Heron MP, Murphy SL, Kochanek KD. Deaths: Final Data for 2004. *National Vital Statistics Reports.* 2007;55(19):1-120.

51. Kung H-C, Hoyert DL, Xu J, Murphy SL. Deaths: Final Data for 2005. *National Vital Statistics Reports.* 2008;56(10):1-121.

52. Heron MP, Hoyert DL, Murphy SL, Xu J, Kochanek KD, Tejada-Vera B. Deaths: Final Data for 2006. *National Vital Statistics Reports.* 2009;57(14):1-135.

53. Xu J, Kochanek KD, Murphy SC, Tejada-Vera B. Deaths: Final Data for 2007. *National Vital Statistics Reports.* 2010;58(19):1-135.

54. Miniño AM, Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2008. *National Vital Statistics Reports.* 2011;59(10):1-157.

55. Murphy SL, Xu J, Kochanek KD. Deaths: Preliminary Data for 2010. *National Vital Statistics Reports.* 2012;60(4):1-69.

56. El-Serag HB. Hepatocellular carcinoma. *N Engl J Med.* Sep 22 2011;365(12):1118-1127.

57. Centers for Disease Control and Prevention's Divison of Viral Hepatitis. Surveillance Data for Acute Viral Hepatitis - United States, 2008. http://www.cdc.gov/hepatitis/Statistics/2008Surveillance/index.htm. Accessed April 5, 2012.

58. Kanwal F, Hoang T, Kramer JR, et al. Increasing prevalence of HCC and cirrhosis in patients with chronic hepatitis C virus infection. *Gastroenterology.* Apr 2011;140(4):1182-1188 e1181.

59. Hepatocellular carcinoma - United States, 2001-2006. *MMWR Morb Mortal Wkly Rep.* May 7 2010;59(17):517-520.

60. Kershenobich D, Razavi HA, Cooper CL, et al. Applying a system approach to forecast the total hepatitis C virus-infected population size: model validation using US data. *Liver Int.* Jul 2011;31 Suppl 2:4-17.

61. Freeman AJ, Dore GJ, Law MG, et al. Estimating progression to cirrhosis in chronic hepatitis C virus infection. *Hepatology.* Oct 2001;34(4 Pt 1):809-816.

62. Davis G, Alter M, El-Serag H, Poynard T, Jennings L. Aging of hepatitis C virus (HCV)-infected persons in the United States: a multiple cohort model of HCV prevalence and disease progression. *Gastroenterology.* Feb 2010;138(2):513-521, 521 e511-516.

63. Wasley A, Kruszon-Moran D, Kuhnert W, et al. The prevalence of hepatitis B virus infection in the United States in the era of vaccination. *J Infect Dis.* Jul 15 2010;202(2):192-201.

64. Mitchell T, Armstrong GL, Hu DJ, Wasley A, Painter JA. The increasing burden of imported chronic hepatitis B--United States, 1974-2008. *PLoS One.* 2011;6(12):e27717.

65. Ioannou G. Hepatitis B virus in the United States: infection, exposure, and immunity rates in a nationally representative survey. *Ann Intern Med.* 2011 154(5):319-328.

66. Surveillance for chronic hepatitis B virus infection - New York City, June 2008-November 2009. *MMWR Morb Mortal Wkly Rep.* Jan 13 2012;61(1):6-9.

67. Characteristics of Persons with Chronic Hepatitis B --- San Francisco, California, 2006. *MMWR Morb Mortal Wkly Rep.* 2007;56(18):446-448.

68. National Center for Health Statisitics. Vital Statitics of the United States: Mortality, 1999. Technical Appendix. In: Branch MS, ed. Hyattsville, MD1999.

69. Oregon Health Authority. VistaPHw in Oregon: Health Assessment. 2012; http://public.health.oregon.gov/BirthDeathCertificates/VitalStatistics/VistaPHw/Pages/VistaPHw.aspx. Accessed March 26, 2012.

70. National Center for Health Statistics. Medical Examiners' and Coroners' Handbook on Death Registration and Fetal Death Reporting, 2003 Revision. In: Services DoHaH, ed. Hyattsville, Maryland: Centers for Disease Control and Prevention; April 2003.

71. Freedman MA, Gay GA, Brockert JE, Potrzebowski PW, Rothwell CJ. The 1989 revisions of the US Standard Certificates of Live Birth and Death and the US Standard Report of Fetal Death. *Am J Public Health.* Feb 1988;78(2):168-172.

72. National Center for Health Statistics. U.S. Standard Death Certificate. 2003; http://www.cdc.gov/nchs/data/dvs/death11-03final-acc.pdf. Accessed March 30, 2012.

73. Cleveland W, Gross E, Shyu W. Local regression models. In: Chambers J, Hastie T, eds. *Statistical Models in S*. Pacific Grove, California: Chapman and Hall/CRC 1992.

74. Shotwell M. sas7bdat: SAS Database Reader (experimental). 2011. http://CRAN.R-project.org/package=sas7bdat.

75. Wickham H. Reshaping data with the reshape package. *Journal of Statistical Software.* 2007;21(12). http://www.jstatsoft.org/v21/i12/paper.

76. Warnes G. gmodels: Various R programming tools for model fitting. 2011. http://CRAN.R-project.org/package=gmodels.

77. Warnes G. gplots: Various R programming tools for plotting data. 2011. http://CRAN.R-project.org/package=gplots.

78. Venables W, Ripley B. *Modern Applied Statistics with S. Fourth edition*. New York: Springer; 2002.

79. Rousseeuw P, Croux C, Todorov V, et al. robustbase: Basic Robust Statistics. 2011. http://CRAN.R-project.org/package=robustbase.

80. Taylor J. hett: Heteroscedastic t-regression. 2012. http://CRAN.R-project.org/package=hett.

81. Rein DB, Smith BD, Wittenborn JS, et al. The cost-effectiveness of birth-cohort screening for hepatitis C antibody in U.S. primary care settings. *Ann Intern Med.* Feb 21 2012;156(4):263-270.

82. Thomas AR, Zaman A, Bell BP. Deaths from chronic liver disease and viral hepatitis, Multnomah County, Oregon, 2000. *J Clin Gastroenterol.* Oct 2007;41(9):859-862.

83. Wu C, Chang HG, McNutt LA, Smith PF. Estimating the mortality rate of hepatitis C using multiple data sources. *Epidemiol Infect.* Feb 2005;133(1):121-125.

84. Gretch DR. Diagnostic tests for hepatitis C. *Hepatology.* Sep 1997;26(3 Suppl 1):43S-47S.

85. Zoulim F. New nucleic acid diagnostic tests in viral hepatitis. *Semin Liver Dis.* Nov 2006;26(4):309-317.

86. Anderson R, Minino A, Hoyert D, Rosenberg H. Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates. *Natl Vital Stat Rep.* May 18 2001;49(2):1-32.

87. Centers for Disease Control and Prevention. *HIV Surveillance Report, 2010.* 2012;22. http://www.cdc.gov/hiv/surveillance/resources/reports/2010report/pdf/2010_HIV_Surveillance_Report_vol_22.pdf.

88. Proehl R. Annual Population Estimates. http://pdx.edu/prc/annual-population-estimates. Accessed March 29, 2012.

89. Koenker R. quantreg: Quantile Regression. 2011. http://CRAN.R-project.org/package=quantreg}.

90. R Development Core Team. R Data Import/Export. 2012; http://cran.r-project.org/doc/manuals/R-data.pdf. Accessed June 10, 2012.

APPENDIX I: DOCUMENTATION FOR R CODE

The sample R code included in Appendix II is intended to provide guidance for individuals interested in applying the multiphase method. The code was written for the specific purpose of assessing Oregon viral hepatitis mortality data for this thesis project and will need to be modified for other analyses (e.g., steps explicitly for which cohort categories are explicitly defined or manipulated will need to be changed). The following section will outline the steps of the code and point out some specific considerations that need to be made when adapting it.

Installation of R Packages

Prior to running any part of the code, the necessary packages must be installed in R. The included code requires the following R packages: *sas7bdat*,[74] *reshape*,[75] *gmodels*, [76] *gplots*,[77] and *MASS*.[78] Additional packages are available for quantile (*quantreg*[89]), robust (*robustbase*[79]), and heterogenous t-distribution (*hett*[80]) regressions; however, these analyses are not included with the provided code.

Set Up

Once these R packages have been installed, the code comprising the "Set Up" section from the included code may be executed. This code removes any prior variables or graphics from the R system and loads the requisite packages.

Import Data

The next step is to import data from an age-period contingency table into R. The table should be formatted with age groups as rows and time periods as columns as shown in Table 4. Note that age group and period intervals need to be of equal width to form distinct birth cohorts as mentioned previously. Also, rates are the recommended format of data for the provided code. Although other data (e.g., cases or proportions) may also be used in the multiphase method and will be compatible with the median polish step of this code, the regression step, as written, may not be appropriate for non-rate data. Thus, additional modifications to this code – not detailed in this thesis – may be necessary for applying the multiphase method to other types of data.

Management of the viral hepatitis mortality data used in this thesis was done in SAS. Consequently, the included code is written to import data from a permanent SAS dataset into R. While R can read data saved as other file types (e.g., .csv, .txt), only code for importing SAS datasets is provided. Individuals who wish to import other file types should refer to the "R Data Import/Export" manual[90] on the Comprehensive R Archive Network or documentation for specific R packages designed to import specific file types.

Tables for Graphical Representation

To facilitate the creation of graphs for the first phase of the multiphase method (graphical representation), birth cohorts from the imported age-period contingency table are identified and the table reformatted with birth cohorts as rows and periods as columns. A table for birth cohort by age is also produced from the original contingency table. Both the birth cohort-period and birth cohort-age tables are output as a .csv file which can be easily incorporated into Microsoft Excel or equivalent spreadsheet software. Although reformatting the contingency table is not strictly necessary to produce birth cohort-period or birth cohort-age graphs in Excel for the graphical representation, having pre-formatted data to plot greatly simplifies the process.

Median Polish

Prior to the median polish, rates are log transformed. If rate data are not being used, the necessity of this transformation should be carefully weighted and this step removed if appropriate. Next, the median polish is conducted on the log transformed data from the age-period contingency table with a convergence tolerance set at 0.0001. The convergence tolerance may be adjusted and other options added to this step. It may be helpful to review what options are available for the median polish procedure by entering “help(medpolish)” in the R command line.

After the median polish has been performed, the resulting residuals are extracted and used for a few different purposes. First, they are subtracted from the log transformed contingency table created prior to the median polish. The resulting values are exponentiated to generate age-period, birth cohort-period, and birth cohort-age tables for assessing age and period effects with cohort effects removed (as in Figures 17–19). These tables are exported as .csv files for incorporation in a spreadsheet or graphing program. Second, an age-period table of the log transformed median polish residuals (like Table 5) is also exported as a .csv file. Third, the median polish residuals are concatenated into a single column vector. A corresponding vector is created, which identifies which residual value each birth cohort corresponds with. Finally, the residuals are plotted against each birth cohort categories and a loess curve fit to the data using locally weighted regression with a smoothing parameter of 1.0.[73]

Linear Regression

Before the linear regression is conducted, two preliminary steps are performed: 1) the vector of labeling the birth cohorts corresponding to the median polish residuals is assigned as a categorical variable and 2) a referent cohort is selected from the birth cohorts. Then, the linear regression is run with the median polish residuals as the dependent variable and the birth cohort categories as the predictor. Using linear contrasts, risk ratios and corresponding 95% confidence intervals for each cohort category compared to a referent are constructed from the predicted values and confidence limits output from the regression. This analysis is repeated for birth cohorts grouped into the 1950–1965 baby boomer group compared to preceding and subsequent birth cohorts. Note that the estimates and related hypothesis testing done for risk of viral hepatitis mortality in baby boomers compared to other groups may not be appropriate for analyses of other health outcomes. Depending on the project and proposed hypotheses, similar methods of estimation and hypothesis testing may be appropriate and the provided code can be modified.

The risk ratio and 95% confidence interval for individual birth cohorts compared to a referent are graphed as are the standard residual plots for assessing whether parametric assumptions are met. Finally, risk ratio estimates and confidence intervals are exported as .csv files.

APPENDIX II: R CODE

```
###################################

#                       SET UP                         #

###################################



rm(list = ls())                                                    # Clear all variables

graphics.off()                                                      # Close graphics windows

library(sas7bdat)                                                   # load sas7bdat package to read sas
code into R

library(reshape)                                                    # load reshape package

library(gmodels)                                                    # load gmodels package

library(gplots)                                                     # load gplots package

library(MASS)                                                       # load MASS package
```

```
########################################

#                    IMPORT DATA                    #

########################################


## read age-period contingency table data from SAS into R ##

cc <- read.sas7bdat(

"C:/Users/elmanm/dropbox/thesis/Data/Cohort_4yrGrp/Total/aprates_4YrGrp.sas7bdat")



###########################################################

#                    TABLES FOR GRAPHICAL RESPRESENTATION                    #

###########################################################


# reformat age-period table to birth cohort-period table for export/easy graphing in Excel #

# NOTE 1: birth cohort-period table is not used in analysis beyond graphical representation

# phase
# NOTE 2: This section will need to be modified for analyses with different numbers of age
# groups and periods #
a4  <-cc[,"heprate4"]                              # extract 4th column of contingency
                                                   # table
c.a4<-rev(a4)                                      # reverse order of extracted 4th column
b4<-append(NA,append(NA,append(NA,c.a4)))          # append 3 rows of missing values to
                                                   # extracted 4th column (diagonals
                                                   # without datapoints)
a3  <-cc[,"heprate3"]                               # extract 3rd column of contingency
                                                   # table
c.a3<-rev(a3)                                      # reverse order of extracted 3rd column
b3<-append(append(NA,append(NA,c.a3)),NA)          # append 3 rows of missing values to
                                                   # extracted 3rd column (diagonals
                                                   # without all datapoints)
a2  <-cc[,"heprate2"]                              # extract 2nd column of contingency
                                                   # table
c.a2<-rev(a2)                                      # reverse order of extracted 2nd
                                                   # column
```

```
b2<-append(append(append(NA,c.a2),NA),NA)        # append 3 rows of missing values to
                                                 # extracted 2nd column (diagonals
                                                 # without all datapoints)
a1  <-cc[,"heprate1"]                            # extract 1st column of contingency
                                                 # table
c.a2<-rev(a1)                                    # reverse order of extracted 1st column
b1<-append(append(append(c.a2,NA),NA),NA)        # append 3 rows of missing values to
                                                 # extracted 1st column (diagonals
                                                 # without all datapoints)
c <- cbind(b1,b2,b3,b4)                          # recombine reformatted columns to
                                                 # form birth cohort-period table


# reformat age-period table to birth cohort-age table for export/easy graphing in Excel #
# NOTE 1: birth cohort-period age is not used in analysis beyond graphical representation
# phase #
# NOTE 2: This section will need to be modified for analyses with different numbers of age
# groups and periods #
q1 <-t(cc[1,])                                # extract and transpose 1st column of
                                              # contingency table
p1 <-append(cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA),q1)               # append 15 rows of missing values to
                                              # extracted column
q2 <-t(cc[2,])                                # extract and transpose 2nd column of
                                              # contingency table
p2 <-append(append(cbind
(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA),q2),NA)              # append 15 rows of missing values to
                                              # extracted column
q3 <-t(cc[3,])                                # extract and transpose 3rd column of
                                              # contingency table
```

```
p3 <-append(append(

cbind(NA,NA,NA,NA,NA,NA,

NA,NA,NA,NA,NA,NA,NA),q3),

cbind(NA,NA))                                    # append 15 rows of missing values to

                                                 # extracted column

q4 <-t(cc[4,])                                   # extract and transpose 4th column of

                                                 # contingency table

p4 <-append(append(

cbind(NA,NA,NA,NA,NA,NA,

NA,NA,NA,NA,NA,NA),q4),

cbind(NA,NA,NA))                                 # append 15 rows of missing values to

                                                 # extracted column

q5 <-t(cc[5,])                                   # extract and transpose 5th column of

                                                 # contingency table

p5 <-append(append(

cbind(NA,NA,NA,NA,NA,NA,

NA,NA,NA,NA,NA),q5),

cbind(NA,NA,NA,NA))                              # append 15 rows of missing values to
```

```
                                        # extracted column
q6 <-t(cc[6,])                          # extract and transpose 6th column of
                                        # contingency table
p6 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA),q6),
cbind(NA,NA,NA,NA,NA))                  # append 15 rows of missing values to
                                        # extracted column
q7 <-t(cc[7,])                          # extract and transpose 7th column of
                                        # contingency table
p7 <-append(append(
cbind(NA,NA,NA,NA,NA,
NA,NA,NA,NA),q7),
cbind(NA,NA,NA,NA,NA,NA))               # append 15 rows of missing values to
                                        # extracted column
q8 <-t(cc[8,])                          # extract and transpose 8th column of
                                        # contingency table
```

```
p8 <-append(append(
cbind(NA,NA,NA,NA,NA,
NA,NA,NA),q8),
cbind(NA,NA,NA,NA,NA,NA,NA))               # append 15 rows of missing values to
                                           # extracted column
q9 <-t(cc[9,])                             # extract and transpose 9th column of
                                           # contingency table
p9 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,NA),q9),
cbind(NA,NA,NA,NA,NA,NA,NA,NA))            # append 15 rows of missing values to
                                           # extracted column
q10 <-t(cc[10,])                           # extract and transpose 10th column of
                                           #.contingency table
p10 <-append(append(
cbind(NA,NA,NA,NA,NA,NA),q10),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA))                                      # append 15 rows of missing values to
                                           # extracted column
```

```
q11 <-t(cc[11,])                       # extract and transpose 11th column of
                                       # contingency table
p11 <-append(append(
cbind(NA,NA,NA,NA,NA),q11),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA))                               # append 15 rows of missing values to
                                       # extracted column
q12 <-t(cc[12,])                       # extract and transpose 12th column of
                                       # contingency table
p12 <-append(append(
cbind(NA,NA,NA,NA),q12),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA))                            # append 15 rows of missing values to
                                       # extracted column
q13 <-t(cc[13,])                       # extract and transpose 13th column of
                                       # contingency table
```

```
p13 <-append(append(
cbind(NA,NA,NA),q13),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA))                          # append 15 rows of missing values to
                                        # extracted column
q14 <-t(cc[14,])                        # extract and transpose 14th column of
                                        # contingency table
p14 <-append(append(
cbind(NA,NA),q14),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA))                       # append 15 rows of missing values to
                                        # extracted column
q15 <-t(cc[15,])                        # extract and transpose 15th column of
                                        # contingency table
```

```
p15 <-append(append(
cbind(NA),q15),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA))                        # append 15 rows of missing values to
                                            # extracted column
q16 <-t(cc[16,])                            # extract and transpose 16th column of
                                            # contingency table
p16 <-append(q16,
cbind(NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,NA))                  # append 15 rows of missing values to
                                            # extracted column
p <- cbind(p1,p2,p3,p4,p5,p6,p7,p8,
p9,p10,p11,p12,p13,p14,p15,p16)             # recombine reformatted columns to
                                            # form birth cohort-age table


# output birth cohort-period tables as .csv files #
write.csv(c,file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_CP.csv')
write.csv(p,file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_CA.csv')
```

```
########################################

#                    MEDIAN POLISH                    #

########################################


# run median polish on age-period contingency table and extract residuals for regression step #

ct<-log(cc)                                             # log transform age-period contingency

                                                        # table


med.ct <- medpolish(ct,eps=0.0001)                      # run median polish on contingency

                                                        # table with convergence tolerance set

                                                        # at 0.0001

med.ct                                                  # examine dataset

r <-med.ct$residuals                                    # extract median polish residuals


# make age-period table with cohort effects removed #

ft1 <- exp(ct-r)
```

```
# make birth cohort-period table with median polish residuals removed for qualitative
# assessment of graphs without cohort effects    #
# NOTE: This section will need to be modified for analyses with different numbers of age
# groups and periods #
a4 <-ft1[,"heprate4"]                          # extract 4th column of residual table
f.a4<-rev(a4)                                  # reverse order of extracted 4th column
f4<-append(NA,append(NA,append(NA,f.a4)))      # append 3 rows of missing values to
                                               # extracted 4th column (diagonals
                                               # without all datapoints)
a3 <-ft1[,"heprate3"]                          # extract 3rd column of residual table
f.a3<-rev(a3)                                  # reverse order of extracted 3rd column
f3<-append(append(NA,append(NA,f.a3)),NA)      # append 3 rows of missing values to
                                               # extracted 3rd column (diagonals
                                               # without all datapoints)
a2 <-ft1[,"heprate2"]                          # extract 2nd column of residual table
f.a2<-rev(a2)                                  # reverse order of extracted 2nd
                                               # column
```

```
f2<-append(append(append(NA,f.a2),NA),NA)          # append 3 rows of missing values to
                                                   # extracted 2nd column (diagonals
                                                   # without all datapoints)
a1 <- ft1[,"heprate1"]                             # extract 1st column of residual table
f.a2<-rev(a1)                                      # reverse order of extracted 1st column
f1<-append(append(append(f.a2,NA),NA),NA)          # append 3 rows of missing values to
                                                   # extracted 1st column (diagonals
                                                   # without all datapoints)
ft2 <- cbind(f1,f2,f3,f4)                          # recombine reformatted columns to
                                                   # form birth cohort-period table
                                                   # without cohort effects


# make birth cohort-age table with median polish residuals removed for qualitative assessment
# of graphs without cohort effects #
# NOTE: This section will need to be modified for analyses with different numbers of age
# groups and periods #
v1  <-t(ft1[1,])                                   # extract and transpose 1st column of
                                                   # contingency table
```

```
w1<-append(cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA),v1)              # append 15 rows of missing values to
                                             # extracted column
v2 <-t(ft1[2,])                              # extract and transpose 2nd column of
                                             # contingency table
w2 <-append(append(cbind
(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA),v2),NA)             # append 15 rows of missing values to
                                             # extracted column
v3 <-t(ft1[3,])                              # extract and transpose 3rd column of
                                             # contingency table
w3 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA),v3),
cbind(NA,NA))                                # append 15 rows of missing values to
                                             # extracted column
v4 <-t(ft1[4,])                              # extract and transpose 4th column of
                                             # contingency table
```

```
w4 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA),v4),
cbind(NA,NA,NA))                      # append 15 rows of missing values to
                                      # extracted column
v5 <-t(ft1[5,])                       # extract and transpose 5th column of
                                      # contingency table
w5 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA),v5),
cbind(NA,NA,NA,NA))                   # append 15 rows of missing values to
                                      # extracted column
v6 <-t(ft1[6,])                       # extract and transpose 6th column of
                                      # contingency table
```

```
w6 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA),v6),
cbind(NA,NA,NA,NA,NA))                        # append 15 rows of missing values to
                                              # extracted column
v7 <-t(ft1[7,])                               # extract and transpose 7th column of
                                              # contingency table
w7 <-append(append(
cbind(NA,NA,NA,NA,NA,
NA,NA,NA,NA),v7),
cbind(NA,NA,NA,NA,NA,NA))                     # append 15 rows of missing values to
                                              # extracted column
v8 <-t(ft1[8,])                               # extract and transpose 8th column of
                                              # contingency table
w8 <-append(append(
cbind(NA,NA,NA,NA,NA,
NA,NA,NA),v8),
cbind(NA,NA,NA,NA,NA,NA,NA))                   # append 15 rows of missing values to
                                               # extracted column
v9 <-t(ft1[9,])                                # extract and transpose 9th column of
                                               # contingency table
w9 <-append(append(
cbind(NA,NA,NA,NA,NA,NA,NA),v9),
cbind(NA,NA,NA,NA,NA,NA,NA,NA))                # append 15 rows of missing values to
                                               #. extracted column
v10 <-t(ft1[10,])                              # extract and transpose 10th column of
                                               #.contingency table
w10 <-append(append(
cbind(NA,NA,NA,NA,NA,NA),v10),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA))                                          # append 15 rows of missing values to
                                               # extracted column
v11 <-t(ft1[11,])                              # extract and transpose 11th column of
                                               # contingency table
```

```
w11 <-append(append(
cbind(NA,NA,NA,NA,NA),v11),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA))                              # append 15 rows of missing values to
                                      # extracted column
v12 <-t(ft1[12,])                     # extract and transpose 12th column of
                                      # contingency table
w12 <-append(append(
cbind(NA,NA,NA,NA),v12),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA))                           # append 15 rows of missing values to
                                      # extracted column
v13 <-t(ft1[13,])                     # extract and transpose 13th column of
                                      # contingency table
```

```
w13 <-append(append(
cbind(NA,NA,NA),v13),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA))                      # append 15 rows of missing values to
                                    # extracted column
v14 <-t(ft1[14,])                   # extract and transpose 14th column of
                                    # contingency table
w14 <-append(append(
cbind(NA,NA),v14),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA))                   # append 15 rows of missing values to
                                    # extracted column
v15 <-t(ft1[15,])                   # extract and transpose 15th column of
                                    # contingency table
```

```
w15 <-append(append(
cbind(NA),v15),
cbind(NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA))                          # append 15 rows of missing values to
                                              # extracted column
v16 <-t(ft1[16,])                             # extract and transpose 16th column of
                                              # contingency table
w16 <-append(v16,
cbind(NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,NA))                    # append 15 rows of missing values to
                                              # extracted column
w <- cbind(w1,w2,w3,w4,w5,w6,w7,w8,
w9,w10,w11,w12,w13,w14,w15,w16)               # recombine reformatted columns to
                                              # form birth cohort-age table
```

```
d <-melt(c)                                   # concatenates columns of residuals

                                              # (i.e., forms one long vector of the

                                              # values from the median polish with

                                              # numeric indicators corresponding to

                                              # to each birth cohort



# output age-period, birth cohort-period, and birth cohort-age median polish residual tables as

# csv files #

write.csv(r,

file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_AP_resid.csv')

write.csv(ft1,

file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_AP_noCEffect.csv')

write.csv(ft1,

file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_CP_noCEffect.csv')

write.csv(w,

file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_CP_noCEffect.csv')



# plot residuals from median polish #
```

```
par(mar=c(9,6,3,2)+.01)                                   # set margins w/extra room on bottom
                                                          # for axis label
plot(d$X1, d$value,xlab="", ylab="",pch=19,
col="black",ylim=c(-1.5,1.5),xlim=c(1,19)
,xaxt="n",cex.axis=1.5)                                   # make scatterplot of median polish
                                                          # residuals by birth cohort
abline(h=0, col="black")                                  # add reference line at 0
axis(1,at=1:19,lab=c("1910-13","1914-17","1918-21",
"1922-25","1926-29","1930-33","1934-37","1938-41",
"1942-45","1946-49","1950-53","1954-57","1958-61",
"1962-65","1966-69","1970-73","1974-77","1978-81",
"1982-85"),las=2,cex.axis=1.5)                            # label birth cohorts on x-axis
                                                          # NOTE: This step will need to be
                                                          # modified in other analyses
par(mar=c(5,4.5,3.5,2)+.01)                               # adjust margins again for labels
title("Residual values from median polish by
birth cohort", xlab="Birth Year",ylab="Residuals",
cex.lab=2)                                                # add main title and bottom and left
```

```
                                                  # axis labels
fit<-loess(d$value~d$X1,span=1.0,
data.frame(x=X1,y=value),degree=1)                # fit loess curve predicted values and
                                                  # standard error (to calculate
                                                  # confidence intervals) with smoothing
                                                  # parameter of 1.0
pred<-predict(fit,data.frame(x=X1),se=T)          # extract predicted values for loess
                                                  # curve
p1 <-pred$fit[1:19]                               # restrict predicted values of loess curve
                                                  # to first 19 values (remaining values
                                                  # are repeats of the first 19, their
                                                  # inclusion complicates plotting)
                                                  # NOTE: This step will need to be
                                                  # modified in other analyses
```

```
se1<-pred$se[1:19]                 # restrict standard error of loess curve
                                   # to first 19 values (remaining values
                                   # are repeats of the first 19, their
                                   # inclusion complicates plotting)
                                   # NOTE: This step will need to be
                                   # modified in other analyses
lines(p1)                          # add predicted values for loess curve
                                   # to plot
lines(p1+(1.96*se1), lty=2)        # add upper 95% confid. interval for
                                   # loess curves to plot
lines(p1-(1.96*se1), lty=2)        # add lower 95% confid. interval for
                                   # loess curves to plot
```

```
########################################

#                    LINEAR REGRESSION               #

########################################



# run linear regression on median polish residuals #

x.f <- factor(d$X1)                              # create factor (i.e., categorical)

                                                 # variable for cohort categories

x.f <- relevel(x.f,ref="4")                      # set referent category

                                                 # NOTE: This step will need to be

                                                 # modified in other analyses

reg <- reg <- lm(d$value~(x.f))                  # run linear regression


# calculate exponentiated risk ratios and associated 95% confidence intervals for each cohort

# category vs referent category using linear contrasts for estimation #

# NOTE: This section will need to be modified for analyses with different numbers of age

# groups and periods #

t1 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f1'=1), conf.int=0.95))

t2 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f2'=1), conf.int=0.95))
```

```
t3 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f3'=1), conf.int=0.95))

t5 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f5'=1), conf.int=0.95))

t6 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f6'=1), conf.int=0.95))

t7 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f7'=1), conf.int=0.95))

t8 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f8'=1), conf.int=0.95))

t9 <- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f9'=1), conf.int=0.95))

t10<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f10'=1),conf.int=0.95))

t11<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f11'=1),conf.int=0.95))

t12<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f12'=1),conf.int=0.95))

t13<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f13'=1),conf.int=0.95))

t14<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f14'=1),conf.int=0.95))

t15<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f15'=1),conf.int=0.95))

t16<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f16'=1),conf.int=0.95))

t17<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f17'=1),conf.int=0.95))

t18<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f18'=1),conf.int=0.95))

t19<- exp(estimable(reg,cm=cbind('(Intercept)'=-1,'x.f19'=1),conf.int=0.95))
```

```
# calculate exponentiated average risk ratios and associated 95% confidence intervals for
# preceding cohorts and baby boomers in 1950-1965 cohorts using linear contrasts for
# estimation #
# NOTE: This section may not be appropriate for other analyses applying this method #
tt1 <- exp(estimable(reg,cm=cbind('(Intercept)'=-.1,
'x.f1'= -.1,
'x.f2'= -.1,
'x.f3'= -.1,
'x.f5'= -.1,
'x.f6'= -.1,
'x.f7'= -.1,
'x.f8'= -.1,
'x.f9'= -.1,
'x.f10'=-.1,
'x.f11'=.25,
'x.f12'=.25,
'x.f13'=.25,
'x.f14'=.25
), conf.int=0.95))

# calculate exponentiated average risk ratios and associated 95% confidence intervals for baby
# boomers in 1950-1965  cohorts and subsequent cohorts using linear contrasts for estimation #
# NOTE: This section may not be appropriate for other analyses applying this method #
tt2 <- exp(estimable(reg,cm=cbind('x.f11'=.25,
'x.f12'=.25,
'x.f13'=.25,
'x.f14'=.25,
'x.f15'=-.2,
'x.f16'=-.2,
'x.f17'=-.2,
'x.f18'=-.2,
'x.f19'=-.2
), conf.int=0.95))
```

```
# extract and combine estimated risk ratios and 95% confidence intervals for individual birth
# cohorts #
# NOTE: This section will need to be modified for analyses with different numbers of age
# groups and periods #
t <- cbind(c(1:19),
rbind(t1$Estimate,t2$Estimate,t3$Estimate,1,t5$Estimate,t6$Estimate,
t7$Estimate,t8$Estimate,t9$Estimate,t10$Estimate,t11$Estimate,
t12$Estimate,t13$Estimate,t14$Estimate,t15$Estimate,t16$Estimate,
t17$Estimate,t18$Estimate,t19$Estimate),
rbind(t1$Lower.CI,t2$Lower.CI,t3$Lower.CI,0,t5$Lower.CI,t6$Lower.CI,
t7$Lower.CI,t8$Lower.CI,t9$Lower.CI,t10$Lower.CI,t11$Lower.CI,
t12$Lower.CI,t13$Lower.CI,t14$Lower.CI,t15$Lower.CI,t16$Lower.CI,
t17$Lower.CI,t18$Lower.CI,t19$Lower.CI),
rbind(t1$Upper.CI,t2$Upper.CI,t3$Upper.CI,0,t5$Upper.CI,t6$Upper.CI,
t7$Upper.CI,t8$Upper.CI,t9$Upper.CI,t10$Upper.CI,t11$Upper.CI,
t12$Upper.CI,t13$Upper.CI,t14$Upper.CI,t15$Upper.CI
,t16$Upper.CI,t17$Upper.CI,t18$Upper.CI,
t19$Upper.CI))
```

```
colnames(t)<-c("Cohort","RR","LCL","UCL")            # add column names to combined data

t <- as.data.frame(t)                                # format t as a data frame



# extract and combine estimated risk ratios and 95% Confidence Intervals for preceding cohorts

# compared to baby boomer cohorts #

# NOTE: This section may not be appropriate for other analyses applying this method #

o_tt1 <- cbind(

rbind(tt1$Estimate),

rbind(tt1$Lower.CI),

rbind(tt1$Upper.CI))



# extract and combine estimated risk ratios and 95% Confidence Intervals for baby boomer

# cohorts compared to subsequent cohorts #

# NOTE: This section may not be appropriate for other analyses applying this method #

o_tt2 <- cbind(

rbind(tt2$Estimate),

rbind(tt2$Lower.CI),

rbind(tt2$Upper.CI))
```

```
# output risk ratios and 95% confidence intervals #
write.csv(t,file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_RR.csv')
write.csv(o_tt1,
file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_RRoldVBBm.csv')
write.csv(o_tt2,
file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_RRyngVBBm.csv')


# output risk ratio and 95% confidence intervals #
write.csv(t,file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_RR.csv')
write.csv(o_tt1,
file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_RRtt1.csv')
write.csv(o_tt2,
file='C:/Users/elmanm/Dropbox/thesis/Data/Cohort_4yrGrp/Total/APC_4Yr_RRtt2.csv')
```

```
# plots for linear regression #
# Plot estimated risk ratios with bars for 95% confidence intervals for individual cohort
# categories #
par(mar=c(9,6,3,2)+.01)                                # set margins w/extra room on bottom
                                                       # for axis label
plotCI(x = t$RR, uiw = se, lty = 1,
xaxt ="n", xlim = c(1,19), ylim = c(-2,4), gap = 0,
ylab="", xlab="", barcol="dark grey",pch=19,
cex.axis=1.5)                                          # plot risk ratios and confidence
                                                       # intervals
                                                       # NOTE: This step will need to be
                                                       # modified in other analyses
abline(h=1, col="black")                               # add reference line at 1
```

```
axis(1,at=1:19,lab=c("1910-13","1914-17","1918-21",
"1922-25","1926-29","1930-33","1934-37",
"1938-41","1942-45","1946-49","1950-53",
"1954-57","1958-61","1962-65",
"1966-69","1970-73","1974-77","1978-81",
"1982-85"),las=2,cex.axis=1.5)                       # label birth cohorts on x-axis
                                                     # NOTE: This step will need to be
                                                     # modified in other analyses
par(mar=c(5,4.5,4,2)+.01)                             # adjust margins again for labels
title("Estimated risk ratio and 95% confidence
intervals for the effect of birth cohort on
viral hepatitis mortality", xlab="Birth Year",
ylab="Risk Ratio",cex.lab=2)                         # add main title and bottom and left
                                                     # axis labels
# standard residual plots #
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))   # place 4 plots on same page
plot(reg, pch=19,  las = 1, sub="")                  # plot residuals
par(opar)                                            # reset plotting preferences
```

```
# plot for residual error from linear regression model by birth cohort

r1  <-data.frame(x=c(reg$residuals[1:16],
NA,NA,NA,NA,
reg$residuals[17:32],NA,NA,NA,NA,
reg$residuals[33:48],NA,NA,NA,NA,
reg$residuals[49:64]))                            # configure output from regression for
                                                  # plotting
                                                  # NOTE: This step will need to be
                                                  # modified in other analyses

par(mar=c(6,4,3,2)+.01)                           # adjust margins for labels
plot(d$X1, r1$x, xlab="", ylab="", pch=19,
ylim=c(-1.5,1.5),xlim=c(1,19),xaxt="n",cex.axis=.7)   # plot residuals from regression vs
                                                  # cohort categories
                                                  # NOTE: This step will need to be
                                                  # modified in other analyses
abline(h=0, col="black")                          # add reference line at 0
```

```
axis(1,at=1:19,lab=c("1910-13","1914-17",
"1918-21","1922-25","1926-29","1930-33",
"1934-37","1938-41","1942-45","1946-49",
"1950-53","1954-57","1958-61","1962-65",
"1966-69","1970-73","1974-77","1978-81",
"1982-85"), las=2,cex.axis=.7)                      # label birth cohorts on x-axis
                                                    # NOTE: This step will need to be
                                                    # modified in other analyses
par(mar=c(5,4,3.5,2)+.01)                           # adjust margins again for labels
title("Residual error from linear regression model
by birth cohort",xlab="Birth Cohort",
ylab="Residual")                                    # add main title and bottom and left
                                                    # axis labels
```