Bias in Markov Models of Disease

Daniel M. Faissol, Paul M. Griffin, and Julie L. Swann

Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205
Phone: 404-894-2300; Fax: 404-894-2301
{dfaissol,jswann,pgriffin}@isye.gatech.edu

27 June 2008

Abstract

We examine bias in Markov models of diseases, including both chronic and infectious diseases. We consider two common types of Markov disease models: ones where disease progression changes by severity of disease, and ones where progression of disease changes in time or by age. We find sufficient conditions for bias to exist in models with aggregated transition probabilities when compared to models with state/time dependent transition probabilities. We also find that when aggregating data to compute transition probabilities, bias increases with the degree of data aggregation. We illustrate by examining bias in Markov models of Hepatitis C, Alzheimer's disease, and lung cancer using medical data and find that the bias is significant depending on the method used to aggregate the data. A key implication is that by not incorporating state/time dependent transition probabilities, studies that use Markov models of diseases may be significantly overestimating or underestimating disease progression. This could potentially result in incorrect recommendations from cost-effectiveness studies and incorrect disease burden forecasts.

Keywords: Markov Model, Bias, Monte Carlo Simulation
1 Introduction

Markov models are commonly used to simulate diseases when evaluating various medical interventions. Modeling diseases allows for the consideration of long term consequences and other implications not practical via clinical trials. Examples include an analysis of cost-effectiveness of expanded HIV screening in the US (1; 2), an analysis of the optimal age of vaccination (3), the optimal timing of liver transplantation (4; 5), and an analysis of dynamic multi-drug therapies for HIV (6). Markov models are also used to forecast and estimate future morbidity, mortality, prevalence, and costs of various diseases (7). Markov models of diseases are often coupled with clinical trial data to determine the cost-effectiveness of medical interventions. Given the limited amount of resources available, Markov models of diseases can potentially be a relatively inexpensive, yet powerful tool to evaluate medical interventions.

However, due to limited disease data or model complexity, simplifying assumptions are made in many Markov disease models. In this paper, we analyze the effect of a common simplification, namely, that of modeling diseases that have nonlinear progression with Markov models that assume constant progression. The assumption of constant disease progression is used in many disease studies (8; 9; 10; 11; 12; 13; 14; 15; 16). Studies often assume constant disease progression when there is insufficient patient data to characterize non-linearities or changes over time. Constant progression is also often assumed in order to reduce the model’s complexity, particularly when deriving analytical results. In Kirkizlar, et al, (17), for example, the authors used a reduced the state space model of Hepatitis C (HCV) for analytical tractability. Consequently, the model did not allow for disease progression to change as a function of disease severity. The reduced state space was necessary to construct a dynamic policy of testing for the disease.

This study explores bias in Markov models of disease when constant disease progression is assumed between states, where bias is the over- or underestimation of overall disease progression. We consider diseases where progression (i.e., transition probabilities) depends on the severity of the disease (i.e., state) and diseases where progression varies with time (i.e., age, time spent in a state). In both cases, we use Markov models and compare the use of state/time dependent transition probabilities with the use of transition probabilities that assume constant progression. We make such a comparison in order to determine if, and under what conditions, Markov models that assume linear progression underestimate or overestimate disease progression. We then use medical data to assess the magnitude of the
bias for HCV, Alzheimer’s disease, and lung cancer.

Previous work on this topic is limited. Scherrer (18), motivated by the same problem, considered bias in the first two periods of the Markov chain in a Markov model of lung cancer and found conditions for bias to exist. Scherrer compared a model with time dependent transition probabilities to one where the probability of acquiring lung cancer did not depend on the number of years an individual has smoked. A study by Bruinvels and colleagues (19) is related in that it demonstrated the effect that using yearly probabilities instead of rates has on a Markov cohort simulation. Their study contained numerical analysis which illustrated the magnitude of the bias. Davis, et al. (20) examines the problem of estimating transition probabilities from aggregate data generated, and Yi, et al. (21) estimates transition probabilities for the specific case of HCV.

In non-health related literature, there has been a great deal of study on Markov chains in general. State space reduction techniques and the lumpability of Markov chains are particularly noteworthy. State space reduction techniques can be employed on a Markov chain that has prohibitively many states or dimensions to reduce it to one with fewer in order to make computations feasible and arrive at a steady state distribution (22; 23). Under special conditions, the states of a Markov chain may partitioned into “lumps” to form aggregated states that also satisfy the Markov property. Aggregation/disaggregation techniques have also been developed to estimate stationary distributions for large Markov chains. Marek, et. al. (24), for example, developed an iterative algorithm where the steady state distribution of a large Markov chain is estimated by computing the steady state distribution of an aggregated Markov chain. The algorithm finishes if, and when, the probability distribution of the disaggregated Markov chain converges to a vector under a given norm. Fortunately, Markov chains of diseases often have relatively small state spaces. The diseases considered in this study have the additional benefit that they are modeled with Markov chains that have special structure. We are able to exploit this fact to arrive at closed form expressions rather than using iterative algorithms or approximations.

Also noteworthy in regards to our study are papers that use the Kullback-Leibler (K-L) divergence rate in comparing two Markov chains (25; 26). The K-L divergence rate can be used to compare two Markov chains over the same state space. As the divergence rate increases, the Markov chains are considered more “different.” It is well known that the K-L divergence rate, however, is not a distance metric since it is not symmetric, nor does it satisfy the triangle inequality. More importantly, at present the K-L divergence rate for comparing Markov chains with different state spaces does not have a closed form expression. Since we
are interested specifically in Markov chains of disease progression, we are able to compare two different such Markov chains by examining the expected time to death in each under their respective steady state distributions. Using the expected time to death as a measure in this way gives us an intuitive method of comparing Markov chains with different state spaces. Since the resulting measure is in life years, the significance of the difference, or lack thereof, between the Markov disease models is easily understood and the health implications apparent.

In summary, our contribution is in explicitly characterizing the magnitude and direction of the bias for several ways of calculating transition probabilities and calculating the impact for several diseases.

2 Markov Models of Diseases

To analyze Markov disease progression models that have state dependent transition probabilities, we consider models of the type in Figure 1 and compare it to the case where the probability transitions between states are not identical. Figure 1 is a standard disease progression model used for liver disease that uses the METAVIR standard for liver disease, which is the motivating disease for this analysis (13; 27; 16). METAVIR is a scoring system specifically designed for HCV patients in which the scores F0-F4 represent different degrees of liver fibrosis. F0 represents no liver scarring and F4 represent cirrhosis or advanced liver scarring; the states in between represent intermediate levels of liver damage. Patients are determined to be in one of the five METAVIR states via a liver biopsy. Each state in Figure 1 corresponds to a score on the METAVIR system, which are ordered by increasing liver damage. The $x$ values in the figure are determined by dividing the time since infection by
the number of states progressed during that time, and averaging over many patients. For example, if a patient is determined to be in state F3 (as determined by a liver biopsy) with an infection length of 30 years, the rate of liver disease progression is thus $3/30 = 0.1$, which is then assumed to be the constant rate of progression between all states. Consequently, the transition probabilities are identical by construction. This method of arriving at a single aggregate transition probability is referred to as the indirect method and is used in the majority of studies analyzing progression rates of HCV (28; 21; 29). We will refer to the value computed using the indirect method as the indirect value.

When few patients in a data set have serial biopsies (i.e., biopsies at different points in time for the same patient), the modeler has little choice but to use the indirect method to estimate transition probabilities and implicitly assume constant disease progression. However, when patient data with serial biopsies is available, the modeler can potentially estimate the transition probabilities between METAVIR states so that the disease progression is not assumed to be constant. This study provides insight into the value of having richer disease data (i.e., serial biopsies)

Studies involving liver disease progression typically assume a constant progression of liver disease (12; 13; 16) as described above. Few studies allow for state dependent liver disease progression such as done in Bennett et al. (30). Several studies have shown, however, that liver disease progression is not constant and varies significantly between METAVIR states (31; 32; 33; 21). Matsumura (31) and Yi, et al., (21) computed estimates for the transition probabilities between the METAVIR states using patient data with serial biopsies in the former, and using a Markov maximum likelihood method on single biopsy patient data in the latter. Both studies also compute the $x$ value using the indirect method. No studies have analyzed, however, the consequences of using the single transition probability value that assumes constant progression versus using the transition probabilities that vary between states. Our study considers this issue.

Many diseases are modeled by Markov chains similar to that of Figure 1, particularly those where the disease states represent scores on an ordinal rating system such as the METAVIR standard. Examples include Alzheimer’s disease (AD), Glaucoma, and Amyotrophic Lateral Sclerosis (ALS). These diseases are progressive, sometimes degenerative, and may or may not be fatal. The severity of AD, for example, is commonly ranked using the Mini-Mental State Examination (MMSE), which is based on a scale of 0-30. The Alzheimer’s Disease Assessment Scale and the Blessed Information Memory Concentration are other common, ordered rating systems for Alzheimer’s disease. In practice, the states are
Figure 2: Markov models for a disease with time dependent transition probabilities. Typically grouped to represent states such as “healthy,” “mild,” “moderate,” and “severe.”

The majority of studies of AD assume a constant progression rate and use the same single transition value between states of different severity (9; 10; 11; 14; 15). Studies that analyze the effect of different medical interventions on AD typically measure the degree to which the intervention decreases the constant progression rate. Jonsson (34), for example, determined that treatment with donepezil 5 mg corresponds to multiplying the transition probability to a state with a lower MMSE score by $(1 - 0.4636)$ and by $(1 - 0.4807)$ for treatment with donepezil 10 mg. Some studies have shown, however, that the progression is far from constant and is typically slower in less severe states (35; 36; 37). Accordingly, some studies do use state dependent transition probabilities (38; 37). As in the case of HCV, studies have not considered the consequences of using the single transition probability value that assumes constant progression versus using the transition probabilities that vary between states.

Another example of a disease that can be modeled by a Markov chain similar to Figure 1 is the progression of ALS, which is often scored on a scale of 1-5 (Mild, Moderate, Severe, Terminal, Death) (39). Progression is often modeled as constant over time (8). Similar to the previous examples, studies have shown, however, that ALS has nonlinear progression (39).

For Markov disease progression models that have time dependent transition probabilities, we consider models of the type in Figures 2(a) and 2(b). In this case, we can interpret states 1-N in Figure 2(b) as ages or as time spent in a risk state. The former is appropriate to model diseases that progress faster with age such as HCV or AD; the latter is appropriate...
to model diseases such as lung cancer whose risk of occurrence increases with the number of years of smoking. The model in Figure 2(b) allows for the transitions to vary with time since the $a'_i$s can each have different values, whereas the model in Figure 2(a) does not capture the time dependency and is equivalent to constraining the $a'_i$s to be identical in Figure 2(b). Scherrer (18) began a framework for analyzing bias in models with time dependent transition probabilities and compared the two models in the case of two risk states plus a disease state for two time periods where the disease state is an absorbing state.

Scherrer (18) showed that the single transition probability of moving from the risk state to the disease state obtained by averaging the true transition probabilities of many risk states underestimates the likelihood of ending up in the disease state (when the probability of transitioning to the disease state increases with time spent in the risk state, and the entire population begins in the risk state). Scherrer also showed that the result also holds when the population is initially evenly split between the two risk states. We extend the Markov model in Scherrer and solve for a stationary distribution for $n$ risk states and arrive at conditions for bias that holds for the Markov model in steady state.

Markov models with time dependent transition probabilities of the type in Figure 2(b) are appropriate for diseases where the transitions vary over time or by age, including HCV, AD, lung cancer, cardiovascular disease and diabetes. In the case of HCV, many studies utilize transition probabilities that vary by age (13; 40; 41), whereas others only use a single transition probability that applies for all ages (16; 42) thereby not capturing the large differences in progression by age. Our analysis studies the difference in the expected time to progress through the METAVIR states of the HCV model using the single transition probability versus using age dependent transition probabilities.

This model can also be used to capture other measures such as incidence. For example, incidence of AD increases with age (43); in this case, each age of person’s life can represent a Markov state and can have a different transition rate to the disease state. Similarly, since the probability of acquiring lung cancer increases with the number of years of smoking (44), the Markov states can represent the number of years of smoking, each with different transition rates to lung cancer.
3 Monte Carlo Simulations of Markov Models of Diseases

A common method of analyzing Markov chains of disease progression in the medical literature is to employ the use of Monte Carlo simulations (2; 1; 16). In these models, each individual patient’s clinical course is followed from the time of entry into the model until death or the final state of interest is reached. The amount of time spent in each health state is determined along with associated costs and health outcome measures such as quality adjusted life years (QALY). Upon the patient’s death (or arrival in the final state), the next patient is introduced into the model. The process repeats over a large number of patients and the results are averaged over all patients. Figure 3(a) graphically represents the scenario where we require the transition probabilities for the disease progression to be identical in a Monte Carlo simulation, and we have generalized the number of states and introduced additional states before and after the disease progression states. In modeling HCV, for example, state “Healthy” represents uninfected, states “F0” – “F4” represent the METAVIR states, and the final state before death represents disease complications including hepatocellular carcinoma and liver decompensation. Figure 3(b) shows a disease progression with state dependent transition probabilities.
We are interested in comparing the model shown in Figure 3(a) with the model shown in Figure 3(b). When the final state is “Death” then the direct transitions to it (from states “Healthy - “N” ) represent death from causes other than the disease being modeled. It is worth noting that the final state need not be “Death.” When the final state represents a diseased state (perhaps a non-fatal disease), the direct transitions to it represent the probability of acquiring the disease for reasons other than those captured by the model’s states. In the case of time-dependent transition probabilities, we are interested in comparing the model in Figures 2(b) and 2(a) where we also assume that the “death” states feeds back into state 1 as it would in a Monte Carlo simulation. Similarly, “Death” need not be considered the final state.

The remainder of the paper is organized as follows. We first derive at stationary distributions for models where the progression changes with state or with time. Then we use the stationary distribution to arrive at conditions for bias. We then discuss factors that affect the degree of the bias. We apply the results to HCV, AD, and lung cancer using medical data. Finally, we discuss the implications of the results when Markov models of disease progression are used in studies of cost-benefit analysis, forecasting future prevalence, and estimating future disease burden.

4 Stationary Distributions

4.1 State dependent transition probabilities

The stationary distribution of a Markov chain describes the steady state behavior of the Markov chain, which allows us to draw conclusions about Markov models in the long run. The results derived from the stationary distribution will hold regardless of the initial health state or initial distribution of health states (in the case of modeling populations). The Markov models in Figure 3 are irreducible aperiodic Markov chains where all of the states are positive recurrent (i.e., they are ergodic). If we add a direct path from “death” to state 1 in the models in Figure 2 (as would be the case in Monte Carlo simulations), then they are also ergodic. By Theorem 4.3.3 in Ross (45), there exists a unique stationary distribution $\pi$ for each of those ergodic Markov chains. Moreover, by Theorem 4.3.1 of Ross, the expected return time to state $i$ (once the state is entered) is given by $\frac{1}{\pi_i}$. The death state is included for illustrative purposes and the time spent in it is always one time period. Keeping the death state in the model implies that $\frac{1}{\pi_{Death}} - 1$ represents the expected time from state
“Healthy” to state “Death”. It should be clear that solving the stationary distribution of the Markov chain in Figure 3(b) is equivalent to solving the stationary distribution of Figure 4, only differing in the labels given to the transition probabilities. Without loss of generality, we use the Markov chain in Figure 4 in our analysis for the remainder of this paper. That is, the results throughout this paper assume the comparison of Figure 4 to the corresponding single probability transition Markov chain (i.e., where $a_i = x \forall i$). We subsequently prove, however, that all of our results remain true in the more general case of comparing Figures 3(a) and 3(b), even when we add additional states before state 1 and after state $N$ (as long as they do not directly communicate with any of the states $1 - N$ as shown in the Figures).

We begin with Figure 4 where $d = 0$ (i.e., there are no direct transitions to the final state). In this case the analysis is simple. Solving
\[
\begin{pmatrix}
\pi_0 & \pi_1 & \cdots & \pi_N
\end{pmatrix} =
\begin{pmatrix}
1 - a_1 & a_1 & 0 & 0 & \cdots & 0 \\
0 & 1 - a_2 & a_2 & 0 & \cdots & 0 \\
0 & 0 & 1 - a_3 & a_3 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 - a_N & a_N \\
1 & 0 & 0 & 0 & \cdots & 0
\end{pmatrix}
\]

for \(\pi\) gives us the stationary distribution. It can easily be shown that

\[
\pi_N = \frac{1}{1 + \sum_{1 \leq j \leq N} \frac{1}{a_j}}.
\]

The proof of the solution to Equation 1 is a special case (where \(d = 0\)) of Proposition 1. All proofs are provided in the Appendix. In the case that \(a_i = x \forall i\), then \(\pi_N = \frac{1}{1 + \frac{x}{2}}\).

Next we consider Figure 4 where \(d > 0\) (i.e., there are direct transitions to the final state). In this case the analysis is more complicated. Solving

\[
\begin{pmatrix}
\pi_0 & \pi_1 & \cdots & \pi_N
\end{pmatrix} =
\begin{pmatrix}
1 - a_1 - d & a_1 & 0 & 0 & \cdots & d \\
0 & 1 - a_2 - d & a_2 & 0 & \cdots & d \\
0 & 0 & 1 - a_3 - d & a_3 & \cdots & d \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 - a_N - d & a_N + d \\
1 & 0 & 0 & 0 & \cdots & 0
\end{pmatrix}
\]

for \(\pi\) gives the stationary distribution.

**Proposition 1.** The stationary distribution to the Markov chain in Figure 4 is

\[
\pi_N = \frac{1}{1 + \sum_{1 \leq j \leq N} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left(\frac{a_i}{a_i + d}\right)}.
\]
Consequently, the expected time from state 0 to N is equal to

\[
E_a = \sum_{1 \leq j \leq N} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left( \frac{a_i}{a_i + d} \right). \tag{5}
\]

If we consider the case where the \(a_i = x\) \(\forall i\) as it is in the METAVIR standard and in Figure 4, then the expected time from state 0 to N is equal to

\[
E_x = \sum_{1 \leq j \leq N} \frac{1}{x} \prod_{1 \leq i \leq j} \left( \frac{x}{x + d} \right). \tag{6}
\]

4.2 Time dependent transition probabilities

Next we solve for the stationary distribution to the Markov chain in Figure 2(b). In this case, we must solve

\[
\begin{bmatrix}
\pi_1 & \pi_2 & \cdots & \pi_N & \pi_C & \pi_D \\
0 & 1-a_1-d & 0 & 0 & \cdots & a_1 & d_1 \\
0 & 0 & 1-a_2-d & 0 & \cdots & a_2 & d_2 \\
0 & 0 & 0 & 1-a_3-d & \cdots & a_3 & d_3 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & a_N & d_N \\
0 & 0 & 0 & \cdots & 0 & 1-D & D \\
1 & 0 & 0 & 0 & \cdots & 0 & 0
\end{bmatrix}
\]

for \(\pi\).

**Proposition 2.** The stationary distribution to the Markov chain in Figure 2(b) (i.e., the solution to Equation 7) is

\[
\pi_D = \frac{1}{1 + \sum_{1 \leq j \leq N} (1 + \frac{a_j}{D}) \prod_{1 \leq i \leq j-1} (1 - a_i - d_i)}. \tag{8}
\]

Consequently, the expected time from state 1 to Death is equal to

\[
E_a = \sum_{1 \leq j \leq N} (1 + \frac{a_j}{D}) \prod_{1 \leq i \leq j-1} (1 - a_i - d_i). \tag{9}
\]
Figure 2(a) is a special case (N = 3) of Figure 3(b). Consequently, we can use the corresponding development to arrive at the expected time from state 1 to Death as

\[ E_x = \frac{x + d + D}{(x + d)(d + D)}. \]  

(10)

If we consider the case where the \( a_i = x \) \( \forall i \) where the \( d'_i \)s may vary, then the expected time from state 1 to Death is equal to

\[ E_x = \sum_{1 \leq j \leq N} \left( 1 + \frac{x}{D} \right) \prod_{1 \leq i \leq j-1} (1 - x - d_i). \]  

(11)

5 Analysis of Bias

In this section, we analyze bias in the models described above. In our analysis, we take the model with state/time dependent transition probabilities to be the status quo, and define bias as the change in the expected time it takes to reach the final state from the initial state when using constant transition probabilities (in place of the state/time dependent transition probabilities).

5.1 Conditions for the existence of model bias

5.1.1 State dependent transition probabilities

For the special case where there are no direct transitions to the final state (i.e., \( d = 0 \)) in Figure 4, it follows from Equation 2 that the bias, \( B_{a,x} \), resulting from the use of a single probability transition, \( x \), instead of the state dependent probability transitions \( a_i \), is equal to

\[ B_{a,x} = E_a - E_x = \frac{1}{1 + \sum_{1 \leq j \leq N} \frac{1}{a_j}} - \frac{1}{1 + \frac{N}{x}}. \]

The model with the single transition probabilities overestimates (underestimates) the disease progression, therefore, if and only if \( x > (<) \frac{1}{\sum_{1 \leq j \leq N} \frac{1}{a_j}} \) which is the harmonic mean of the \( a'_i \)s.

We turn our attention to the case when there are direct transitions to the final state (i.e., \( d > 0 \)), which represents death from natural causes (or an alternative means of obtaining
disease). We are interested in comparing the models in Figures 3(a) and 3(b) to analyze the consequences of using a single transition probability. With the framework introduced in the previous section, we can arrive at sufficient conditions for bias.

We start by subtracting Equation 6 from 5 to arrive at the expression for the difference in the expected time from state 0 to N when using a single transition probability versus allowing the transition probabilities to vary by state, which gives us the following Theorem.

**Theorem 1.** For a disease progression of the type in Figure 4, the bias resulting from the use of a single transition probability (i.e., \( a_i = x \ \forall \ i \)) versus allowing the transition probabilities to vary is

\[
B_{a,x} = \sum_{1 \leq j \leq N} \frac{1}{x} \prod_{1 \leq i \leq j} \left( \frac{x}{x + d} \right) - \sum_{1 \leq j \leq N} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left( \frac{a_i}{a_i + d} \right)
\]  

(12)

where we measure bias by the difference in expected time from state 0 to state N.

Theorem 1 indicates the magnitude and direction of the bias. When \( x \) is calibrated such that the expression in 12 is zero, then there is no bias. That is, the expected time from state 0 to state N is the same when using the state dependent probabilities or that particular value of \( x \). Using any other value for \( x \) results in bias in the direction indicated by Theorem 1. Next, we describe sufficient conditions on the value of \( x \) for the presence of bias.

**Theorem 2.** For a disease progression of the type in Figures 3(a) and 3(b), when \( x \) is less than or equal to the harmonic mean of the \( a'_i \)'s, and the \( a'_i \)'s do not all equal \( x \), then the model with aggregated transition probabilities (Figure 3(a)) strictly underestimates disease progression when compared to the model with state dependent transition probabilities (Figure 3(b)).

Theorem 2 states that the expected time from state 0 to N is larger when the harmonic mean of the state dependent transition probabilities is used as a constant transition probability than when the state dependent transition probabilities themselves are used. That is, model bias is strictly negative. The result holds for any value less or equal to than the harmonic mean as well. It follows from Theorem 2 that when a value smaller than the harmonic mean for \( x \) is used, as the value of \( x \) decreases the degree of underestimation (i.e., bias) increases. This is true since as \( x \) decreases, the time from state 0 to N increases, which is apparent by examining Figure 4. The importance of Theorem 2 is that it shows that when modeling a disease whose progression is not constant, using the harmonic mean of the state dependent transition probabilities (instead of the state dependent transition probabilities
themselves) will cause the modeler to underestimate the progression of the disease. When using a value smaller than the harmonic mean, the modeler will increasingly underestimate the progression of the disease.

**Corollary 1.** Consider a disease progression of the type in Figures 3(a) and 3(b) where the $a_i'$s are increasing. When $x$ is computed using the indirect method, the model with aggregated transition probabilities (Figure 3(a)) strictly underestimates disease progression when compared to the model with state dependent transition probabilities (Figure 3(b)).

Corollary 1 is a special case of Theorem 2. It is important to note that when using the indirect method as described in section 2, the modeler can average the computed indirect value of many patients by taking the arithmetic mean of the indirect values, or more correctly, by taking the harmonic mean of the indirect values. The choice of which mean to use is left to the modeler, however, the harmonic mean is more appropriate when averaging rates. When averaging the expected times between states, the arithmetic mean is more appropriate. Indeed, the harmonic mean of the progression rates (indirect values) is equivalent to the arithmetic mean of the expected times between states. The proof of Corollary 1 in the Appendix discusses the use of the arithmetic and harmonic means further. In Corollary 1, we assume $x$ is computed by taking the harmonic mean of the indirect values of many patients (or equivalently, the arithmetic mean of the expected times between states). Corollary 2 discusses the case where we use the arithmetic mean of the indirect values. The importance of Corollary 1 is that it suggests that studies that use the indirect method (the most common method of computing constant transition probabilities) due to lack of disease data are underestimating the disease progression for diseases such as HCV and AD that are believed to have increasing progression rates.

Note that the converses of Theorem 2 and Corollary 1 do not necessarily follow. We can, however, arrive at a result for model bias in the opposite direction (overestimation) using the development in the proof of Theorem 2.

**Theorem 3.** For a disease progression of the type in Figures 3(a) and 3(b), when $x$ is greater than or equal to the geometric (or arithmetic) mean of the $a_i'$s, and the $a_i'$s do not all equal $x$, then the model with aggregated transition probabilities (Figure 3(a)) strictly overestimates disease progression when compared to the model with state dependent transition probabilities (Figure 3(b)).

Theorem 3 states that the expected time from state 0 to N is smaller when the geometric mean of the state dependent transition probabilities is used as a constant transition proba-
probability than when the state dependent transition probabilities themselves are used. That is, model bias is strictly positive. The result holds for any value greater than or equal to the geometric mean, including the arithmetic mean. It follows from Theorem 3 that when a value larger than the geometric mean for $x$ is used (e.g., the arithmetic mean), as the value of $x$ increases the degree of overestimation (i.e., bias) increases. This is true since as $x$ increases, the time from state 0 to N decreases. The significance of Theorem 3 is that when modeling a disease whose progression is not constant, using the geometric mean of the state dependent transition probabilities (instead of the state dependent transition probabilities themselves) will cause the modeler to overestimate the progression of the disease. When using a value larger than the geometric mean, the modeler will increasingly overestimate the progression of the disease.

Corollary 2. Consider a disease progression of the type in Figures 3(a) and 3(b) where the $a_i's$ are decreasing. When $x$ equals the arithmetic mean of the computed indirect values for each patient, the model with aggregated transition probabilities (Figure 3(a)) strictly under-estimates disease progression when compared to the model with state dependent transition probabilities (Figure 3(b)).

Corollary 22 is a special case of Theorem 3. Contrary to Corollary 1, we assume $x$ is computed by taking the arithmetic mean of the computed indirect values of the patients. The importance of Corollary 2 is that it suggests that studies that use the indirect method in this way are overestimating the disease progression for diseases that have decreasing progression.

5.1.2 Time dependent transition probabilities

Next, we consider Markov models with time dependent transition probabilities. Subtracting Equation 11 from 9 gives us the expression for the difference in the expected time from state 1 to state to Death when using a single transition probability versus allowing the transition probabilities to vary over time. Solving for $x$ gives us a condition for bias, which we formulate as our next theorem.

**Theorem 4.** A Markov model without time varying transition probabilities of the type in Figure 2(a) overestimates (underestimates) the expected time to death when compared to a model with time dependent transition probabilities (Figure 2(b)) when

$$x > (\leq) \frac{E_a d - 1}{\gamma - E_a}$$

(13)
where $E_a$ is defined above and $\gamma = \frac{1}{d+D}$

Theorem 4 indicates that when $x$ equals the right hand side of inequality 13 the bias is zero. That is, the expected time from state 1 to death is the same when using the time dependent transition probabilities or that particular value of $x$. Using any other value for $x$ will result in bias in the direction indicated by Theorem 4.

The sufficiency theorems that are true for the state dependent transition probabilities model are not true for the time dependent transition probabilities model. They can be shown to be false by simple numerical counter examples, which we show in Section 6

5.2 Factors that affect the degree of model bias

Thus far we have compared models where the transition probabilities are allowed to vary between states with models that use a single aggregate transition probability between several states. In this section, we consider intermediate levels of data aggregation. That is, instead of considering only the use of a single transition probability, we combine transition probabilities in groups as shown in Figure 5, where $b_1$ is formed by taking some mean of $a_1$ and $a_2$ only. By comparing the three models in Figure 5 we will be able to analyze the effect of the degree of data aggregation on model bias. The degree of data aggregation is increasing as we switch from using the $a'_i$s to using the $b'_i$s to using $c$ as the transition probabilities for the Markov chain.

**Theorem 5.** When using any value greater than or equal to the geometric mean (e.g., the arithmetic mean) of the transition probabilities in Figure 5, the models with aggregated transition probabilities increasingly overestimate disease progression as the degree of aggregation increases.

It is worth recalling that the value computed using the indirect method (and averaging over many patients using the arithmetic mean) is greater than the arithmetic mean of the $a'_i$s for diseases that have decreasing progression.

**Theorem 6.** When using any value less than or equal to the harmonic mean of the transition probabilities in Figure 5, the models with aggregated transition probabilities increasingly underestimate disease progression as the degree of aggregation increases.

Recall that the value computed using the indirect method (and averaging over many patients using the harmonic mean) is less than the harmonic mean of the $a'_i$s for diseases that have decreasing progression.
6 Examples

In this section we use medical data to calculate the bias in models of three diseases introduced in Section 2. We start with models whose transition probabilities vary by state and use HCV and AD as examples. We also use HCV, AD and lung cancer to show bias in models with time dependent transitions.

6.1 Bias in diseases with state dependent transition probabilities

In this section we use HCV and AD as examples for bias in Markov models of diseases with state dependent transition probabilities when we aggregate the data and assume constant transition probabilities. In the case of HCV, the progression through the METAVIR states has been shown to vary by liver disease severity (31; 21). In this case, model states represent METAVIR states where the transitions between states vary as shown in Table 1, where $a_i$ is
the transition probability of going from state $i$ to state $i + 1$. We consider several different data sets for the same disease. The transition probability labeled “indirect” corresponds to the single transition probability obtained using the indirect method. The indirect value reported in each case was obtained by the same corresponding study that computed the $a_i's$ in order to ensure that we do not introduce extra bias by using parameters derived from different data sets.

In the case of AD, the progression has also been shown to vary significantly by severity (36; 37). The Markov model states in this case represent the grouped MMSE scores of an individual with AD, where the probability of transitioning between states varies according to Table 1.

There are very few studies that compute $a_i's$ for HCV, while several studies have computed them for AD. We chose studies that arrived at very different values for the $a_i's$ (due to studying different populations) so that we can test for bias under very different scenarios. For example, the $a_i's$ for Yi 1 pertain to 1138 chronic HCV patients in liver clinics who have disproportionately faster progression than those in Yi 2, which is made up of previously healthy women infected by exposure to contaminated anti-D immune globulin who were then screened for HCV. The data from Matsumura (31) was obtained from Japanese patients infected with HCV that have chronic liver disease. Similarly, while the Stern (36) and Suh (37) data sets are on different scales, they also report very different progressions based on the populations considered. Note that in these numerical results we assume the initial state is the first diseased state (rather than a healthy state) and set the (disease related) death rate from the final disease state equal to one. We perform the numerical analyses in this way in order to highlight the bias in the disease progression without confounding it with incidence and death rates. When we performed the analyses with realistic incidence and death rates we found that the bias directions and relative magnitudes agreed with the results contained herein.

Table 2 displays the results for comparing the use of a single transition probability versus using the $a_i's$ for HCV and AD. The expected time to death is displayed using the indirect value as well as the harmonic (harm), geometric (geo) and arithmetic (arith) means, respectively. We use the death rates from the Centers for Disease Control (CDC) (46). For each data set, we compute the percent change ($\% \Delta$) relative to using the $a_i's$. Note that the indirect method results in greater bias in all cases but one (the arithmetic mean using data from Matsumura (31) for HCV). In all cases, the harmonic mean results in less bias than any other single transition probability considered.
Table 1: Parameter values for transition probabilities

The arithmetic mean consistently results in the largest bias among the three means. We also compare the use of the $a_i'$s to the use of two probability transition values as in Figure 5 where we use the three different means. Using two probability transition values represents a lower degree of data aggregation than using only one. The direction of the bias remains constant, while the magnitude decreases as predicted by Theorems 5 and 6. In most cases, the bias decreases significantly between using a single transition probability versus two. It is also interesting that for the data sets where the $a_i'$s vary a great deal, the bias is much larger.

Table 2: Results for HCV and AD with state dependent transition probabilities for multiple degrees of data aggregation
6.2 Bias in diseases with time/age dependent transition probabilities

In this section we use HCV, AD and lung cancer as examples for bias in Markov models of diseases with transition probabilities that vary by time/age. In the case of HCV, the progression through the METAVIR states have been shown to vary by age (13; 40; 41). The Markov model states in this case represent ages of an individual (while infected with HCV), where the transition to the disease state (cirrhosis) vary by age as shown in Table 3. In the case of AD, the risk of acquiring AD has been shown to increase with age (43). The Markov model states in this case represent the ages of an individual without AD, where the probability of transitioning to the disease state (i.e., acquiring AD) varies by age according to Table 4. In the case of lung cancer, we use the model in Figure 2(b) where each state represents the number of years an individual has smoked cigarettes. In this case we use the function

\[ p(n) = 1.845 \times 10^{-10} n^{4.5} \]  

(14)
as the probability of transitioning to lung cancer after smoking for \( n \) years reported in Scherrer (18), based on Doll and Peto (47) for individuals who smoke 20 cigarettes per day beginning at age 16.

<table>
<thead>
<tr>
<th>Annual Fibrosis Progression by age ( a_i )</th>
<th>0.0125</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ&lt;30</td>
<td>0.0225</td>
</tr>
<tr>
<td>31-40</td>
<td>0.0135</td>
</tr>
<tr>
<td>41-50</td>
<td>0.03125</td>
</tr>
<tr>
<td>51-60</td>
<td>0.05525</td>
</tr>
<tr>
<td>61-70</td>
<td>0.07525</td>
</tr>
<tr>
<td>70+</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Liver Fibrosis estimates for HCV (13)

Table 5 displays the results for comparing the use of a single transition probability versus using the \( a_i \) for HCV and AD. For HCV, we assume the individual acquires the disease at age 20 and compute the expected time until death from age 20 onwards using the death rates from the CDC (46). For AD, we consider the incidence starting at age 60 (since the incidence before age 60 is very small) and compute the expected time to death from age 60 onwards using the same death rates. We compare the use of the harmonic, geometric and arithmetic means. Note that the results serve as counterexamples to show that Theorems...
2 and 3 do not hold for models with time dependent transition probabilities as they do for models with state dependent transition probabilities. The harmonic mean results in smaller bias for both cases, however this is not necessarily always the case.

Table 5: Results for HCV and AD with time dependent transition probabilities

Table 6 displays the results for comparing the use of a single transition probability versus using the $a_i'$s for lung cancer. We assume the individual begins smoking 20 cigarettes per day starting at age 15 and use same death rates as previously. We compare the use of the harmonic, geometric and arithmetic means. Note that the harmonic mean does not always result in smaller bias. We also show the results for different levels of data aggregation. We compare the use of the $a_i'$s to the use of two, three and four probability transition values in the way described in Figure 5 where we use the three different means. Note that the direction of the bias does not always remain constant. However, the bias does decrease significantly when decreasing the degree of data aggregation. The results in Tables 5 and 6
are not sensitive to the death rates since they are driven by the $a'_i$s.

<table>
<thead>
<tr>
<th>$a'_i$s</th>
<th>1 transition probability</th>
<th>2 transition probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected time to death from age 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59.98</td>
<td>45.64</td>
<td>61.47</td>
</tr>
<tr>
<td>%$\Delta$</td>
<td>-23.92%</td>
<td>2.47%</td>
</tr>
<tr>
<td>%$\Delta$</td>
<td>-3.86%</td>
<td>2.46%</td>
</tr>
</tbody>
</table>

Table 6: Results for Lung Cancer with time dependent transition probabilities for multiple degrees of data aggregation (smoking 20 cigarettes/day since 15 years old)

6.3 Transient Analysis

Since the analyses in the previous sections are based on stationary distributions, the results hold on average and in the long term. In this section we consider the bias in the transient periods of the Markov chain. That is, we assume an individual is initially in disease state 0, and compute the probability of the individual having reached state 4 or death at each time period. Figures 6 and 7 illustrate this probability using the HCV data from the first column of Table 1 and the lung cancer data described in Section 6.2, respectively. The results indicate that the bias in disease progression during the transient periods is in the same direction as it is in steady state, further supporting our analytical results. The same is true for all other disease data used in this paper.

7 Discussion

When constant transition probabilities must be used due to limited data, or for analytical tractability, caution must be exercised when choosing the value for the transition probability in order to reduce the affect of bias. We derived sufficient conditions to test for the presence
of bias in two different Markov models of diseases. The disease models can include state dependent transition probabilities or time/age dependent probabilities. For the case of state dependent transition probabilities, we also show analytically that the bias increases as the degree of data aggregation increases, and that the increased bias is in the same direction as determined by Theorems 2 and 3.

Bias can be significant in many cases when models use constant transition probabilities. Using the indirect method, which is a common practice in each of the diseases discussed, resulted in larger bias than any of the three means considered in all cases except for one. Additionally, the bias increases dramatically when the degree of data aggregation increases.

The case where insufficient disease data causes the modeler to use the indirect method of obtaining a constant transition probability is of particular interest since it is such a common occurrence. We showed that the transition probability derived using the indirect method is less than harmonic mean of the state dependent transition probabilities when they are increasing. This result is important because it means that the use of the indirect method for diseases with increasing progression rates leads to an underestimation of disease progression.
We also showed that when the arithmetic mean is used to average the indirect values of many patients, the disease progression is overestimated for diseases that have decreasing progression rates.

It is noteworthy that many studies that produce estimates for the state dependent transition probabilities of diseases with increasing progression rates don’t always arrive at strictly increasing transition probabilities. Results often include progression rates that increases until the final states where progression then slows down, such as those used in section 6. In light of the progressive nature of the diseases studied, the authors typically attribute the observed slower progression in the final states due to ceiling effects of the scoring system (e.g. MMSE), and not to the nature of the disease itself, such as in Park, et al, (48) and Aguero-Torres et al, (49). Nevertheless, we used the transition rates as reported in the literature and found the bias to be significant when using the indirect method.

The implication of our results is that many of the studies that assume constant transition probabilities are arriving at conclusions based on biased models, bias potentially as large as the examples in Section 6. Analyses that use Markov models of disease progres-
sion that ignore the state/time dependency of the probability transitions to determine the cost-effective of medical interventions such as pharmaceutical drugs, therapies or screening programs may be significantly biased towards or away from cost-effectiveness depending on the type of aggregation used. A cost-effectiveness study of HCV, for example, that assumes a constant disease progression and uses the harmonic mean (or any smaller value) for the disease progression could potentially determine a cost-effective medical intervention to not be cost-effective because the disease progression was underestimated.

Similarly, when estimating future prevalence, if a disease model overestimates (underestimates) the progression to death, ceteris paribus, then future prevalence is underestimated (overestimated). For many diseases, the health states become costlier and associated with lower quality of life as the disease progresses, as is the case with HCV, AD and lung cancer. In these cases, it is clear that if we calibrate the models such that they have the same expected time from the first state to the final state, when the disease progression increases (decreases) in time/state, the Markov models that use a single probability transition overestimate (underestimate) the time spent in the later states. This can be important since the later states can be significantly costlier than the earlier states in the disease progression causing the disease burden forecast to be biased.

Finally, additional bias introduced into Markov models of disease progression is due to the fact that transition probability estimates are typically inherently underestimated. This is true because when a patient is determined to be in a particular health state at some point in time (by a liver biopsy, for example), there is no way of knowing how long the patient was in that state. Consequently, the estimate for the rate of progression is commonly made assuming the patient entered that diagnosed state in the time period of the diagnosis. For this reason, estimates of the disease progression are typically lower bounds. As a result, the bias introduced by the use of a single transition probability (instead of state/time dependent probabilities) can be either increased or decreased by this measurement effect. In the case that the use of a single transition probability causes an underestimation of disease progression, the bias due to measurement will cause the disease progression to be underestimated even more so than the results in Section 6 suggest.

8 Appendix

For the proofs in this appendix, it is helpful to first introduce some mathematical tools that we will use in our analysis of model bias. We start by introducing elementary symmetric
polynomials, which are special cases of symmetric polynomials of degree $m$ in $n$ variables (where $m \leq n$).

**Definition 1.** The elementary symmetric polynomial of degree $m$ in $n$ variables is defined as

$$e_{n,m}(x_1, x_2, \cdots, x_n) = \sum_{1 \leq j_1 \leq j_2 \cdots \leq j_m \leq n} x_{j_1}x_{j_2} \cdots x_{j_m}.$$ 

**Definition 2.** An alternative definition for $e_{n,m}(x_1, x_2, \cdots, x_n)$ is that of the coefficient of $x^m$ in the expansion of $\prod_{1 \leq i \leq n} (x_i + y)$.

Next, we use the elementary symmetric polynomials in the following definition.

**Definition 3.** We define

$$S_{n,m} := \frac{e_{n,m}(n_m)}{n}.$$ 

We use $S_{n,m}$ extensively in the proofs below to simplify notation. Additionally, we use MacLaurin's inequality.

**MacLaurin's inequality** states that

$$S_{n,1} \geq (S_{n,1})^{1/2} \geq (S_{n,2})^{1/3} \geq \cdots \geq (S_{n,n})^{1/n}$$

with equality if and only if all of the $x_i$ values are equal.

Lastly, we introduce the following Lemma, which we use to reformulate the expression for the stationary distribution of Markov models with state dependent transition probabilities into a form that is easier to manipulate.

**Lemma 1.**

$$\sum_{1 \leq j \leq n} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left( \frac{a_i}{a_i + d} \right) = \frac{\left( \prod_{1 \leq i \leq n} (a_i + d) - \prod_{1 \leq i \leq n} a_i \right) \frac{1}{d}}{\prod_{1 \leq i \leq n} (a_i + d)} \quad (15)$$

which we prove below.

**Proof of Proposition 1**

Multiplying the right hand side of Equation 3 gives
\[
\begin{align*}
\pi_1 &= \pi_1(1 - a_1 - d) + \pi_N \\
\pi_2 &= \pi_1a_1 + \pi_2(1 - a_2 - d) \\
\pi_3 &= \pi_2a_2 + \pi_3(1 - a_3 - d) \\
&\vdots \quad \vdots \quad \vdots \\
\pi_{N-1} &= \pi_{N-2}a_{N-2} + \pi_{N-1}(1 - a_{N-1} - d) \\
\pi_N &= \pi_{N-1}a_{N-1} + d(\pi_1 + \pi_2 + \cdots + \pi_{N-1}) \quad (16)
\end{align*}
\]

which can be written as
\[
\begin{align*}
\pi_1a_1 &= \pi_2(a_2 + d) \\
\pi_2a_2 &= \pi_3(a_3 + d) \\
\pi_3a_3 &= \pi_4(a_4 + d) \\
&\vdots \quad \vdots \quad \vdots \\
\pi_{N-1}a_{N-1} &= \pi_N(1 + d) - d \\
\pi_N &= \pi_1(a_1 + d). \quad (17)
\end{align*}
\]

For a stationary distribution, the relationship
\[
\pi_1 + \pi_2 + \pi_3 + \cdots + \pi_N = 1 \quad (18)
\]

must hold. Rewriting Equation 18 in terms of \( \pi_N \) using the identities in Equations 17 we get
\[
\begin{align*}
\frac{\pi_N}{a_1 + d} + \frac{\pi_Na_1}{(a_1 + d)(a_2 + d)} + \frac{\pi_Na_1a_2}{(a_1 + d)(a_2 + d)(a_3 + d)} + \cdots + \frac{\pi_Na_1a_2\cdots a_n}{(a_1 + d)(a_2 + d)(a_3 + d)\cdots (a_{N-1})} + \pi_N = 1 \quad (19)
\end{align*}
\]

which can be rewritten as Equation 4. \( \Box \)

**Proof of Proposition 2**

28
Multiplying the right hand side of Equation 7 gives

\[ p_1 = p_D \]
\[ p_2 = p_1(1 - a_1 - d) = p_D(1 - a_1 - d) \]
\[ p_3 = p_2(1 - a_2 - d) = p_D(1 - a_1 - d)(1 - a_2 - d) \]
\[ \vdots \]
\[ p_N = p_{N-1}(1 - a_{N-1} - d) = p_D(1 - a_1 - d)(1 - a_2 - d) \cdots (1 - a_{N-1} - d) \]
\[ p_C = p_1a_1 + p_2a_2 + \cdots + p_Na_N + p_C(1 - D) \]
\[ p_D = p_1d_1 + p_2d_2 + \cdots + p_Nd_N + p_C(D). \]

Since Equation 18 must hold for a stationary distribution, rewriting Equation 18 in terms of \( p_D \) using the above equations gives us the desired expression for \( p_D \). \( \square \)

**Proof of Lemma 1**

Define the following string of identities

\[
\prod_{1 \leq i \leq N} (a_i + d) = d \prod_{2 \leq i \leq N} (a_i + d) + a_1 \prod_{2 \leq i \leq N} (a_i + d) \\
\prod_{2 \leq i \leq N} (a_i + d) = d \prod_{3 \leq i \leq N} (a_i + d) + a_2 \prod_{3 \leq i \leq N} (a_i + d) \\
\prod_{3 \leq i \leq N} (a_i + d) = d \prod_{4 \leq i \leq N} (a_i + d) + a_3 \prod_{4 \leq i \leq N} (a_i + d) \\
\vdots \\
\prod_{N-2 \leq i \leq N} (a_i + d) = d \prod_{N-1 \leq i \leq N} (a_i + d) + a_{n-2} \prod_{N-1 \leq i \leq N} (a_i + d) \\
\prod_{N-1 \leq i \leq N} (a_i + d) = d \prod_{N \leq i \leq N} (a_i + d) + a_{n-1} \prod_{N \leq i \leq N} (a_i + d). \]

Substituting from the bottom up recursively we have

\[
\prod_{1 \leq i \leq N} (a_i + d) = d \prod_{2 \leq i \leq N} (a_i + d) + da_1 \prod_{3 \leq i \leq N} (a_i + d) + da_1a_2 \prod_{4 \leq i \leq N} (a_i + d) + \cdots + da_1a_2 \cdots a_{N-2}(a_N + d) + da_1a_2 \cdots a_{N-1}. \quad (20)
\]
Now, we can rewrite the left hand side of Lemma 1 with a common denominator to be

\[
LHS = \sum_{1 \leq j \leq N} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left( \frac{a_i}{a_i + d} \right) \prod_{2 \leq i \leq N} (a_i + d) + a_1 \prod_{3 \leq i \leq N} (a_i + d) + a_1 a_2 \prod_{4 \leq i \leq N} (a_i + d) + \cdots + a_1 a_2 \cdots a_{N-1} \prod_{1 \leq i \leq N} (a_i + d).
\]

Define the numerator of LHS to be

\[
\gamma = \prod_{2 \leq i \leq N} (a_i + d) + a_1 \prod_{3 \leq i \leq N} (a_i + d) + a_1 a_2 \prod_{4 \leq i \leq N} (a_i + d) + \cdots + a_1 a_2 \cdots a_{N-1}.
\]

From Equation 20 we have

\[
\prod_{1 \leq i \leq N} (a_i + d) = \gamma \times d + \prod_{1 \leq i \leq N} (a_i).
\]

\[\square\]

**Proof of Theorem 2**

We start by analyzing the Markov disease model in Figure 4 and compare it to the case where \(a_i = x \forall i\). Then we extend the result to be true in the more general case of Figures 3(a) and 3(b). When \(x\) is less than or equal to the harmonic mean of the \(a’\)s, we write this mathematically as

\[
x \leq \frac{N}{\sum_{i=1}^{N} \frac{1}{a_i}}.
\]

From Equation 6, for the disease progression in Figure 4, we know that the expected time from state 0 to state N is

\[
E_a = \sum_{1 \leq j \leq N} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left( \frac{a_i}{a_i + d} \right).
\]

By Lemma 1, we can rewrite this as

\[
E_a = \frac{\left( \prod_{1 \leq i \leq N} (a_i + d) - \prod_{1 \leq i \leq N} a_i \right) \frac{1}{d}}{\prod_{1 \leq i \leq N} (a_i + d)}.
\]
Using Definition 2 on the numerator and the denominator we can rewrite the above expression as

\[
E_a = \frac{(d^N + e_{N,1}d^{N-1} + e_{N,2}d^{N-2} + \cdots + e_{N,N-1}d + e_{N,N} - \prod_{1 \leq i \leq N} a_i)\frac{1}{d}}{d^N + e_{N,1}d^{N-1} + e_{N,2}d^{N-2} + \cdots + e_{N,N-1}d + e_{N,N}}. \tag{21}
\]

Since \(\prod_{1 \leq i \leq N} a_i = e_{N,N}\), we can rewrite Equation 21 as

\[
E_a = \frac{d^{N-1} + e_{N,1}d^{N-2} + e_{N,2}d^{N-3} + \cdots + e_{N,N-1}d}{d^N + e_{N,1}d^{N-1} + e_{N,2}d^{N-2} + \cdots + e_{N,N-1}d + e_{N,N}}.
\]

Finally, dividing both numerator and denominator by \(e_{N,N}\) we get

\[
E_a = \frac{\bar{e}_{N,N}d^{N-1} + \bar{e}_{N,N-1}d^{N-2} + \bar{e}_{N,N-2}d^{N-3} + \cdots + \bar{e}_{N,1}}{\bar{e}_{N,N}d^N + \bar{e}_{N,N-1}d^{N-1} + \bar{e}_{N,N-2}d^{N-2} + \cdots + \bar{e}_{N,1}d + 1} \tag{22}
\]

where \(\bar{e}_{i,j}\) are the elementary symmetric polynomials of \(\frac{1}{a_1}, \frac{1}{a_2}, \cdots, \frac{1}{a_N}\).

Note that since the denominator can be written by multiplying the numerator by \(d\) and adding 1, Equation 22 can be written as

\[
E_a = \frac{A}{Ad + 1}
\]

where

\[
A = \bar{e}_{N,N}d^{N-1} + \bar{e}_{N,N-1}d^{N-2} + \bar{e}_{N,N-2}d^{N-3} + \cdots + \bar{e}_{N,1}.
\]

We can arrive at a similar expression for the expected time from state 0 to state N in the model with aggregated transition probabilities where we have \(x's\) instead of \(a's\). Using the same procedure as above we get that the expected time from state 0 to state N (using \(x's\)) is

\[
E_x = \frac{\frac{1}{x^N}d^{N-1} + \binom{N}{1}\frac{1}{x^{N-1}}d^{N-2} + \binom{N}{2}\frac{1}{x^{N-2}}d^{N-3} + \cdots + \frac{N}{x}}{\frac{1}{x^N}d^N + \binom{N}{1}\frac{1}{x^{N-1}}d^{N-1} + \binom{N}{2}\frac{1}{x^{N-2}}d^{N-2} + \cdots + \frac{N}{x} + 1}. \tag{23}
\]

Similarly, since the denominator can be written by multiplying the numerator by \(d\) and adding 1, Equation 23 can be written as

\[
E_x = \frac{B}{Bd + 1}
\]
where
\[ B = \frac{1}{x^N} d^{N-1} + \binom{N}{1} \frac{1}{x^{N-1}} d^{N-2} + \binom{N}{2} \frac{1}{x^{N-2}} d^{N-3} + \cdots + \frac{N}{x} \cdot \]

We would like to compare the expected time from state 0 to state N using \( a' \)'s with the expected time using \( x' \)'s. If we can show that
\[ \frac{B}{Bd+1} > \frac{A}{Ad+1} \tag{24} \]
then it follows that the model in which \( x' \)'s are used underestimates the disease progression since the expected time from state 0 to state N is longer. Since the inequality in 24 is always true whenever \( B > A \), we need only to show that
\[
\begin{align*}
&\frac{\sum a_i d^N}{\sum a_i} > \frac{\sum b_i d^{N-1}}{\sum b_i} + \frac{\sum c_i d^{N-2}}{\sum c_i} + \cdots + \frac{\sum d_i d}{\sum d_i} \\
&\frac{1}{x^N} d^N + \binom{N}{1} \frac{1}{x^{N-1}} d^{N-1} + \binom{N}{2} \frac{1}{x^{N-2}} d^{N-2} + \cdots + \frac{N}{x} \cdot \sum d_i d < \\
&\frac{1}{x^N} d^N + \binom{N}{1} \frac{1}{x^{N-1}} d^{N-1} + \binom{N}{2} \frac{1}{x^{N-2}} d^{N-2} + \cdots + \frac{N}{x} \cdot d_i d. \tag{25}
\end{align*}
\]

The expression labeled \( d \) is less than or equal to the expression labeled \( d' \) since
\[ \tau_{N,1} = \sum_{1 \leq i \leq N} \frac{1}{a_i} \]
by hypothesis, and \( x \leq \sum_{1 \leq i \leq N} \frac{1}{a_i} \) implies \( \frac{N}{x} \geq \sum_{1 \leq i \leq N} \frac{1}{a_i} \).

Next we must show that the terms labeled \( a, b, c, \cdots \) are strictly less than the terms labeled \( a', b', c', \cdots \) (whenever the \( a_i \) are not equal), which is true when
\[
\begin{align*}
\frac{\binom{N}{k}}{x^{N-k}} &> \tau_{N,N-k} \quad \forall \; k \\
\frac{\binom{N}{m}}{x^m} &> \tau_{N,m} = \binom{N}{m} S_{N,m} \quad \forall \; m \\
\frac{1}{x^m} &> S_{N,m} \quad \forall \; m. \tag{26}
\end{align*}
\]

But since \( \frac{1}{x} \geq \frac{1}{N} \sum_{1 \leq i \leq N} \frac{1}{a_i} = S_{N,1} \) and by Maclaurin’s inequality \( S_{N,1} > (S_{N,m})^{\frac{1}{m}} \) for all \( m \).
whenever the $a_i$ are not all equal, then inequality 26 is always true. □

**Proof of Corollary 1**

Given Theorem 2, we need only to show that when the $a'_i$'s are increasing, the value computed using the indirect method is less than the harmonic mean of the $a'_i$'s. The indirect method is computed without taking into account death from other causes (i.e., when $d = 0$). It is a well known property of the harmonic mean that it equals the arithmetic mean of the reciprocals. Since $a_i$ represents the transition probability from state $i$ to $i + 1$, when $d = 0$ the expected time from state $i$ to $i + 1$ is $\frac{1}{a_i} := t_i$. Consequently, the harmonic mean of the $a'_i$'s is equal to the arithmetic mean of the $t'_i$'s. When the $a'_i$'s are increasing (i.e., $t'_i$'s are decreasing), we can see that the value obtained using the indirect method,

$$x_{\text{indirect}} = \frac{1}{N} \left( \frac{1}{t_1 + \frac{t_1 + t_2}{2} + \cdots + \frac{t_1 + t_2 + \cdots + t_N}{N}} \right)$$

is less than the value using the arithmetic mean of the $t'_i$'s,

$$x_{\text{arithmetic}} = \frac{1}{N} \left( t_1 + t_2 + \cdots + t_N \right) = \frac{N}{(1/a_1 + 1/a_2 + \cdots + 1/a_N)}$$

which is equal to the harmonic mean of the $a'_i$'s. □

**Proof of Theorem 3**

As in the proof of the previous theorem, we start by analyzing the Markov disease model in Figure 4 and compare it to the case where $a_i = x \forall i$. We start with Equations 22 and 23 from the proof of Theorem 2. In this case, however, we want to show that $E_x < E_a$ or equivalently that

$$\frac{B}{Bd + 1} < \frac{A}{Ad + 1}$$

which is true when $B < A$. In this case, we must show that the expression on the left hand side of Equation 25 is greater than the expression on right hand side of Equation 25. By hypothesis, $a$ is greater than or equal to the expression labeled $a'$ since

$$\overline{e}_{N,N} = \left( \prod_{1 \leq i \leq N} \frac{1}{a_i} \right)^{\frac{1}{N}}.$$ 

which is the reciprocal of the geometric mean of the $a'_i$'s. It remains to show that the terms labeled $b$, $c$, $\cdots$, $d$ are strictly greater than the terms labeled $b'$, $c'$, $\cdots$, $d'$ in inequality 25.
(whenever the $a_i$ are not equal), which is true when

$$\binom{N}{k} \frac{x^N}{x^{N-k}} < \tau_{N,N-k} \quad \forall \; k$$

$$\binom{N}{m} \frac{x^m}{x^m} < \tau_{N,m} = \binom{N}{m} S_{N,m} \quad \forall \; m$$

$$\frac{1}{x^m} < S_{n,m} \quad \forall \; m.$$

But since $\frac{1}{x} \leq \left( \prod_{1 \leq i \leq N} \frac{1}{a_i} \right)^\frac{1}{N} = (S_{N,N})^{\frac{1}{N}}$ and by Maclaurin’s inequality $(S_{N,N})^{\frac{1}{N}} < (S_{N,m})^{\frac{1}{m}}$ for all $m$ whenever the $a_i$ are not all equal, then inequality 31 is always true. Since the geometric mean is always less than or equal to the arithmetic mean, the result holds for the arithmetic mean as well. □

**Proof of Corollary 2**

Given Theorem 3, we need only to show that when the $a'_i$s are decreasing, the arithmetic mean of the indirect values of the patients is greater than the arithmetic mean of the $a'_i$s. When the $a'_i$s are decreasing, the $t'_i$s are increasing, and we can see that the arithmetic mean of the indirect values,

$$x_{\text{indirect}} = \frac{1}{N} \left( \frac{1}{t_1} + \frac{2}{t_1 + t_2} + \cdots + \frac{N}{t_1 + t_2 + \cdots + t_N} \right)$$

is greater than

$$x_{\text{arithmetic}} = \frac{1}{N} \left( \frac{1}{t_1} + \frac{1}{t_2} + \cdots + \frac{1}{t_N} \right) = \frac{1}{N} (a_1 + a_2 + \cdots + a_N),$$

which is equal to the arithmetic mean of the $a'_i$s. □

**Proof that Theorems 2 and 3 are true in the general case of the model in Figure 3(b).**

We have that

$$\sum_{1 \leq j \leq N} \frac{1}{a_j} \prod_{1 \leq i \leq j} \left( \frac{a_i}{a_i + \frac{x}{a_i}} \right) > \sum_{1 \leq j \leq N} \frac{1}{x} \prod_{1 \leq i \leq j} \left( \frac{x}{x + \frac{1}{a_i}} \right)$$
which expands to

\[
\frac{1}{a_1+d} + \frac{a_1}{(a_1+d)(a_2+d)} + \frac{a_1a_2}{a_1a_2 \cdots a_n} \\
+ \cdots + \frac{x^2}{(a_1+d)(a_2+d)(a_3+d) \cdots (a_{N-1})}
\]

\[
> \frac{1}{i+d} + \frac{i}{(i+d)(a_1+d)} + \frac{ia_1}{i(a_1+d)(a_2+d)} \\
+ \cdots + \frac{ix^2}{(i+d)(a_1+d)(a_2+d) \cdots (a_{N-1})(z+d)(D+d)}
\]

We want to show that the inequality remains true when we add the transition probabilities \(i, z,\) and \(D\) to the stationary distribution according to the Markov model in Figure 3(b). By updating Equation 19 (where we derived the stationary distribution) to include \(i, z,\) and \(D,\) and dividing by \(\pi_N,\) it should be clear that we want to show

\[
\frac{1}{i+d} + \frac{i}{(i+d)(a_1+d)} + \frac{ia_1}{i(a_1+d)(a_2+d)} \\
+ \cdots + \frac{ix^2}{(i+d)(a_1+d)(a_2+d) \cdots (a_{N-1})(z+d)(D+d)}
\]

which is clearly true whenever inequality 34 is true. The result is also true for the reverse inequality. \(\Box\)

**Proof of Theorem 5**

To show that bias increases as the degree of aggregation increases, we must prove that

\[ E_a > E_b > E_x \]

when the \(a'_i\)'s and \(b'_i\)'s are not all equal to \(x,\) and \(b_1 = \sqrt{a_1a_2}\) and \(b_2 = \sqrt{a_3a_4}.\) \(E_a > E_x\) and \(E_b > E_x\) are true by Theorem 3 from the previous section. It remains to show that \(E_a - E_b\)
is strictly positive, which expands to

\[
E_a - E_b = \frac{a_1a_2a_3a_4}{(a_1 + d)(a_2 + d)(a_3 + d)(a_4 + d)(\sqrt{a_1a_2} + d)^2(\sqrt{a_3a_4} + d)^2}
\]

\[
-2\sqrt{a_1a_2}d^2 - 2d^2\sqrt{a_3a_4} + a_2da_4 + a_1a_3a_4 + a_1da_3 + a_1da_4 + d^2a_3 + a_2a_3a_4
\]

\[
+ a_2d^2 + a_2a_3d + a_1a_2a_4 + a_1a_2a_3 + a_1d^2 + d^2a_1 - 2a_3a_4\sqrt{a_1a_2} - 4\sqrt{a_1a_2}\sqrt{a_3a_4}d
\]

\[
-2a_1a_2\sqrt{a_3a_4}/((a_1 + d)(a_2 + d)(a_3 + d)(a_4 + d)(\sqrt{a_1a_2} + d)^2(\sqrt{a_3a_4} + d)^2).
\]

Since the fraction in line 36 is always positive, it suffices to show that lines 37 - 38 are strictly positive. These lines can be rewritten as a polynomial in \(d\) as

\[
d^2 \left(-2\sqrt{a_1a_2} + a_1 + a_2 + a_3 + a_4 - 2\sqrt{a_3a_4}\right)
\]

\[
+d \left(a_1a_3 + a_1a_4 + a_2a_4 + a_2a_3 - 4\sqrt{a_1a_2}\sqrt{a_3a_4}\right)
\]

\[
+ \left(-2a_1a_2\sqrt{a_3a_4} - 2a_3a_4\sqrt{a_1a_2} - a_1a_2a_4 + a_1a_3a_4 + a_2a_3a_4 + a_1a_2a_3\right).
\]

Next we show that each coefficient of the powers of \(d\) is strictly positive. Rewriting the above expression, we can see that this is true if and only if

\[
\frac{a_1 + a_2}{2} + \frac{a_3 + a_4}{2} > \sqrt{a_1a_2} + \sqrt{a_3a_4}
\]

\[
\frac{a_1 + a_2}{2} + \frac{a_3 + a_4}{2} > \sqrt{a_1a_2}\sqrt{a_3a_4}
\]

\[
\frac{a_1a_2}{2} + \frac{a_3a_4}{2} + \frac{a_1 + a_2}{2} > a_1a_2\sqrt{a_3a_4} + a_3a_4\sqrt{a_1a_2}.
\]

The above inequalities are true by repeatedly applying the inequality of the arithmetic and geometric mean, which states

\[
\frac{a_1 + a_2}{2} > \sqrt{a_1a_2}
\]

whenever \(a_1 \neq a_2\) (which is a special case of the Maclaurin inequality where \(n=2\)). It should be clear that the result holds for any mean greater than or equal to geometric mean. In the same way that Theorems 2 and 3 were shown to still be valid in the more general setting where we include additional states before and after the disease progression, this theorem is also true in the more general setting. The proof is omitted since it follows the same steps.
Proof of Theorem 6

We omit the proof since it very closely follows that of Theorem 5, where instead of using the inequality of the arithmetic and geometric means, we use the identity

\[ a_1^2 + a_2^2 > 2a_1a_2 \]

whenever \( a_1 \neq a_2 \) and \( a_1, a_2 > 0 \) which can be seen to be true since \((a_1 + a_2)^2 > 0\). It should be clear that the result holds for any mean less than or equal to the harmonic mean. In the same way that Theorems 2 and 3 were shown to still be valid in the more general setting where we include additional states before and after the disease progression, this theorem is also true in the more general setting. □

References


