2010

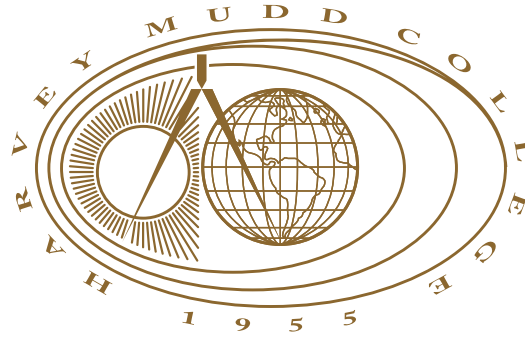# A Multistage Incidence Estimation Model for Diseases with Differential Mortality

Alyssa W. Dray
Harvey Mudd College

A Multistage Incidence Estimation Model for Diseases with Differential Mortality

Alyssa Dray

Talithia Williams, Advisor

Johanna Hardin, Reader
Susan Lewallen, Reader

May, 2010

Harvey Mudd College

COLLEGE

Department of Mathematics

Abstract

According to the World Health Organization, surgically removable cataract remains the leading cause of blindness worldwide. In sub-Saharan Africa, cataract surgical rate targets should ideally be set based on cataract incidence (the number of new cataracts developed each year). Unfortunately, the longitudinal studies necessary to measure incidence have not yet been feasible in these areas. Our research instead proposes a method for estimating incidence based on available cataract prevalence data.

We extend a method proposed by Podgor and Leske (1986) to estimate age-specific incidence from age-specific prevalence in single diseases with differential mortality. A two-stage disease extension is created in order to differentiate between unilateral cataract and bilateral cataract. The new model, along with a numerical simulation method to generate confidence intervals, is implemented in the statistical programming language R. The model is then applied to Rapid Assessment of Avoidable Blindness survey data from parts of Eritrea, The Gambia, Kenya (two regions), Mali, Rwanda and Tanzania. Our results suggest significant geographic variations in cataract incidence, a hypothesis to be further investigated as the RAAB survey expands and improves. We also show how the model can be further extended to model any n-stage progressive disease with differential mortality.

Contents

List of Figures

List of Tables

Acknowledgments

Chapter 1

Introduction

Surgically removable cataract remains the leading cause of blindness world-wide. Blindness due to cataract is much more common in developing countries due to the absence of ophthalmologists who can perform cataract surgery and the lack of infrastructure to utilize existing resources. Though many factors can influence cataract development, the vast majority of cases are age-related and develop in persons over 50 years old. Studies show that women are more susceptible to cataract and, in Africa, women tend to have less access to treatment. The World Health Organization's VISION 2020 project, which seeks to eliminate the main causes of avoidable blindness by 2020, includes an important focus on increasing the number of cataract surgeries in Africa (World Health Organization, 2009b). Ideally the number of surgeries performed each year would at least equal the number of incident cataracts (new cataracts developed) that year. Unfortunately, measuring cataract incidence directly would require longitudinal studies that survey the same group of people over a number of years to see when cataracts were developed, a procedure that has not been possible in Africa. For the past several years, because of the lack of data, sub-Saharan Africa has been assumed to be homogeneous in terms of cataract incidence, and cataract surgical rate (CSR) targets have been set equally across these regions.

1.1   The Need for Incidence Estimation

While cataract incidence is difficult to measure, new Rapid Assessment of Avoidable Blindness (RAAB) surveys provide data about age-specific cataract prevalence (the percentage of the population with cataracts in one or

both eyes). RAAB surveys are conducted on an age-stratified sample of people over 50 years of age from a district of one to two million people. Each person surveyed receives an eye exam and any eyes found to have visual acuity (VA) $< 6/18$ (indicating blindness or low vision in that eye) are re-examined to determine visual acuity and the cause of limited vision in that eye (Limburg and Meester, 2007). Previous cataract surgeries are also noted in this exam. Lewallen et al. (2010) computed age-specific cataract prevalence at several visual acuity levels based on data from seven RAAB surveys in sub-Saharan Africa. The authors also demonstrated that cataract prevalence can be easily computed from future RAAB survey data.

The RAAB survey methodology is well suited to the assessment of avoidable blindness. The survey is feasible to carry out in sub-Saharan Africa. There are some limitations on the data collected, notably that only a few visual acuity levels can be distinguished using RAAB methodology. However, since the goal is to assess blindness, the data is sufficient for our purposes. An important advantage of RAAB surveys is that the same survey methodology was used in all seven districts studied, allowing comparison between regions. The consistency of our data means that we expect that our methodology will be easily applicable to new RAAB data as it becomes available for additional districts.

The challenge addressed by our research is the estimation of cataract incidence from the available age-specific prevalence data. In simple chronic diseases that are not age-dependent and do not affect death rate, there is a simple dependence between incidence, prevalence, and disease duration, such that any of these variables can be calculated from the other two. However, cataract has been shown to affect death rates, and age dependence is an important factor. Podgor and Leske (1986) propose one incidence estimation strategy for a single disease with differential mortality. We extend their method in order to treat cataract as a bilateral disease. While our methodology is inspired by characteristics of cataract disease progression, especially differential mortality, the method itself is not cataract-specific and could likely apply to other bilateral diseases that affect mortality.

This report is organized as follows. Chapter 2 extends Podgor and Leske's model to bilateral diseases. Chapter 3 illustrates a parametric bootstrap method for computing confidence intervals for incidence estimates. In Chapter 4, cataract incidence in sub-Saharan Africa is discussed in detail and incidence is calculated for those African countries where RAAB survey data is available. Chapter 5 further extends the model to enable incidence estimation for an n-stage progressive disease. Chapter 6 discusses the impact of our model and gives suggestions for future work.

Chapter 2

Estimating Incidence of a Bilateral Disease

Our main research goal is to estimate cataract incidence using data on age-specific cataract prevalence available from RAAB surveys. This is a challenging task because "incidence" (the number of eyes developing cataract each year) is a dynamic measure of the impact of cataract on a population, whereas only static "prevalence" data (the percentage of the population having cataract at the time the survey was taken) is available. However, previous work by Podgor and Leske (1986) uses the age-dependence of prevalence data to estimate incidence of a single disease. Previous incidence estimation work is described in Section 2.1. In Section 2.2, we extend Podgor and Leske's work in order to apply it to a two-stage disease. Our two-stage extension is intended to be applicable to the case of cataract, where unilateral cataract (clouding of one eye) and bilateral cataract (clouding of both eyes) form the two disease stages. Here the new model is described in general terms that could also apply to other two-stage diseases; Chapter 4 discusses the application of this model to cataract.

2.1 Previous Work on Incidence Estimation

Incidence estimation for various diseases has been of interest to epidemiologists and others since at least the 1970s. However, the most common strategies rely on the availability of data that is prohibitively difficult or expensive to obtain in Africa, specifically longitudinal or serial prevalence data showing how prevalence evolves in the same population over time. Hallot et al. (2008) and Sakarovich et al. (2007) both published recent pa-

pers estimating HIV incidence in certain African populations. Unfortunately, a major part of each research effort was a longitudinal study of HIV prevalence evolution over time in a single population of interest. This type of study is too expensive to carry out in a systematic way across many areas, as would be necessary in order to investigate geographic variations in the incidence of a disease across sub-Saharan Africa. Brunet and Struchiner (1999) and Marschner (1997) develop non-parametric methods for incidence estimation. Both authors do sophisticated work: Brunet's model allows modeling of rapid changes in incidence and Marschner's paper includes sensitivity analysis to certain parameters. In a later paper, Brunet (2002) develops a compartment model that gives important relationships between incidence and prevalence. However, both authors again rely on serial prevalence data that is unavailable to us.

While many sophisticated incidence estimation techniques exist in the literature, most depend on the availability of a great deal of data obtained through expensive survey methods that are difficult to repeat on a large scale. In the case of cataract disease in Africa, the RAAB survey represents a breakthrough in data collection methodology, because it provides a relatively cheap and standardized way to collect age-specific prevalence data. However, RAAB surveys often take several years to complete, and it has not yet been feasible to collect any kind of longitudinal data, because it is prohibitively difficult to find the same people again after a number of years. These same challenges presumably apply to longitudinal surveys or the collection of serial prevalence data on other diseases in Africa.

Podgor and Leske (1986) do develop a method for incidence estimation based on age-specific prevalence from a single prevalence survey. The authors make certain assumptions that the district of interest represents a closed, steady-state system, then use age-dependence to estimate the time-dependence of cataract prevalence and thus estimate incidence. Their method is described in some detail in the next section, and the assumptions made are reasonable in many African districts. However, the method only envisioned a single disease of interest, whereas in the case of cataract we wish to separate unilateral and bilateral cataract. In Section 2.2, we extend Podgor's method to a model for bilateral diseases.

2.1.1   Incidence Estimation for a Single Disease

Podgor and Leske (1986) propose a method to estimate incidence in a single, irreversible disease with differential mortality. Podgor's method allows one to estimate the age-specific incidence of a single disease based on

known age-specific prevalence and mortality rates. The method consists of a closed, three-state model in which all people are described as healthy (*H*), infected (*I*), or dead (*D*), as shown in Figure 2.1.
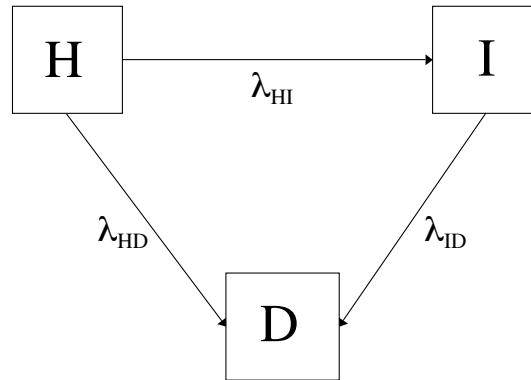
Figure 2.1: Podgor and Leske's Model of Disease Progression.

Because the disease is irreversible, there is no transition from state *I* back to state *H*. Podgor assumes the distribution of each transition time to be exponential with parameters λ_{HD}, λ_{HI}, and λ_{ID} giving the rates of healthy mortality, disease incidence, and diseased mortality respectively. By distinguishing healthy mortality λ_{HD} from diseased mortality λ_{ID}, the model accounts for mortality to be affected by the disease, an important feature in the case of cataract. Figure 2.1 shows a visual representation of the allowed transitions in Podgor's model.

Podgor and Leske assume that the system is closed (no people move into or out of the region) and that the prevalence at all ages is static in time. Based on these assumptions, they argue that to maintain the (assumed) steady-state, the population in a particular age range *a* must evolve, through the processes of mortality and incidence over a time interval *t*, to have the composition of people of age *a* + *t*. Given these assumptions, the age dimension of prevalence can be treated as a time dimension. The authors are then able to make conservation arguments that relate prevalence and transition probabilities. Specifically, because the disease is irreversible, the healthy people at time *a* + *t* must all have been healthy at time *a*. So the number of healthy people at age *a* + *t* must be the number of healthy people at time *a* times the probability, P_{HH}, that a healthy person at age *a* remains

alive and healthy during the time period. This can be expressed

$$N_1(1 - \pi_1) = N_0(1 - \pi_0)P_{HH}, \tag{2.1}$$

where N_0 and N_1 are the size of the total population at ages a and $a + t$ and π_0 an π_1 give the disease prevalence at ages a and $a + t$. Similarly, the number of infected people at age $a + t$ is the number of healthy people at age a times the probability, P_{HI}, that they become infected but survive, plus the number of infected people at age a times the probability, P_{II}, that they survive. This gives

$$N_1\pi_1 = N_0(1 - \pi_0)P_{HI} + N_0\pi_0 P_{II}. \tag{2.2}$$

Based on the exponential model, Podgor and Leske calculate the probabilities P_{HH}, P_{HI} and P_{II} in terms of λ_{HD}, λ_{HI}, and λ_{ID}. Though I elaborate more fully on the details of their derivation when I extend it in Section 2.2.3, their transition probabilities, for comparison, are given by

$$P_{HH} = e^{-(\lambda_{HD} + \lambda_{HI})}, \tag{2.3}$$

$$P_{HI} = \frac{\lambda_{HI}}{\lambda_{HD} + \lambda_{HI} - \lambda_{ID}} \left(e^{-\lambda_{ID}} - e^{-(\lambda_{HD} + \lambda_{HI})} \right), \text{ and} \tag{2.4}$$

$$P_{II} = e^{-\lambda_{ID}}. \tag{2.5}$$

Here we have taken the time interval between prevalence age groups to be $t = 1$ and therefore will get incidence rates expressed per time interval (for example if $t = 5$ years, then λ_{HI} will give incidence per 5 years.

Combining these equations eliminates N_1 and N_0 and leaves a single equation. Incidence (λ_{HI}) is related to known prevalence (π_0 and π_1) and mortality (λ_{HD} and λ_{ID}) by

$$\frac{(1 - \pi_0)\pi_1}{(1 - \pi_1)} e^{-(\lambda_{HD} + \lambda_{HI})} = \frac{\lambda_{HI}(1 - \pi_0)}{\lambda_{HD} + \lambda_{HI} - \lambda_{ID}} \left(e^{-\lambda_{ID}} - e^{-(\lambda_{HD} + \lambda_{HI})} \right) + \pi_0 e^{-\lambda_{ID}}. \tag{2.6}$$

This equation can be solved numerically using Newton's method. In this way, incidence in each age range can be estimated for a single disease with differential mortality if age-specific prevalence and mortality rates are known.

2.2 Extension to Bilateral Diseases

Our extension of Podgor and Leske's method follows the same basic strategy as the original method. Section 2.2.1 explains the need for a two-stage

model for cataract and identifies a new compartment model based on the possible transitions. Section 2.2.2 derives new equations relating various transition probabilities to prevalence, and Section 2.2.3 gives these transition probabilities as a function of known mortality rates and unknown incidence rates based on our exponential model. Finally, Section 2.2.4 describes a way of gaining additional insight into the disease progression by comparing incidence estimates (once obtained) to each other and to mortality rates.

2.2.1   Compartment Model of Disease Progression

In the case of cataract, Podgor and Leske's method is very appealing because it uses age-specific prevalence and mortality rates (obtainable from RAAB surveys and a WHO database, respectively, as described in Chapter 4), to calculate age-specific incidence. The method also accounts for differential mortality in the presence of the disease and makes reasonable assumptions about the stability of the population. However, the method accounts for only a single disease. In the case of cataract, the disease has two stages of interest: "unilateral cataract" (cataract in one eye only) and "bilateral cataract" (cataract in both eyes). These disease stages are likely to have the same mortality because increased mortality is primarily due to physical weakening from the disease rather than blindness itself (Lewallen, 2010). However, it is important to distinguish between the states because a person with bilateral cataract has twice as many operable eyes than a person with unilateral cataract. This chapter describes a four-compartment extension of Podgor and Leske's method that is effective in estimating both unilateral and bilateral cataract incidence.

   We define four states: healthy (H), unilateral cataract (U), bilateral cataract (B), and deceased (D), with allowed transitions as shown in Figure 2.2. Like Podgor, we considered mortality with a separate death state because of differential mortality (people with cataract are more likely to die). We assume that everyone who develops cataract first develops opacity in one eye, then at any later time may develop cataract in the second eye. Because we use a continuous model, these transitions might or might not both occur in the same five-year period.

   We assume that the transition time between any two states i and j is governed by an exponential distribution with parameter λ_{ij}, although certain aspects of the model do not depend on this assumption. The transition rates λ_{HD}, λ_{UD}, and λ_{BD} (for transition to state D from states H, U, and B respectively) are known or approximated mortality rates, which can
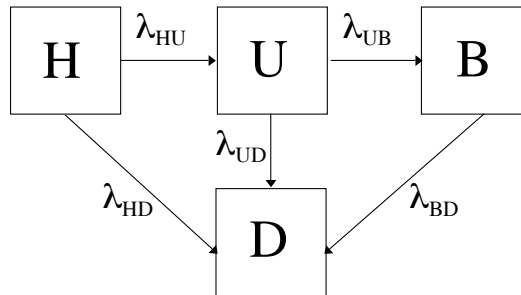
Figure 2.2: Bilateral Model of Disease Progression.

be different for each stage of the disease. Mortality rates can generally be obtained at least as readily as disease prevalence data. For example, in the case of cataract, healthy mortality rates were obtained from the World Health Organization and diseased mortality rates were estimated as discussed in Section 4.2. λ_{HU} and λ_{UB} are closely related to unilateral and bilateral incidence, respectively, and are the target quantities to calculate.

2.2.2   Disease Progression Equations

In our extended model, as in Podgor and Leske's model, we assume a closed group of people with cataract prevalence at each age fixed in time. In the case of cataract, these assumptions are likely appropriate and are forced by the scarcity of data. For example, if prevalence were non-static, to see whether it increased or decreased in time we would need to know prevalence at two time points, and that data is unavailable. It therefore seems reasonable to treat the age-dependence of cataract as a time variable, and derive model equations based on the conservation of people as they age.

The pool of people who are healthy at age a_1 would be those who were healthy at age a_0 and neither died nor developed cataract in the first eye

(second eye incidence is not possible directly from this state). That is,

$$N_1^H = N_0^H P_{HH}$$
$$N_1(1 - \pi_1^U - \pi_1^B) = N_0(1 - \pi_0^U - \pi_0^B)P_{HH}, \tag{2.7}$$

where N_0 and N_1 are the total population at a_0 and a_1, respectively and P_{HH} is the probability that a person who is healthy at time a_0 stays alive and healthy until time a_1.

The pool of people with unilateral cataract at time a_1 would be those who had unilateral cataract at time a_0 and survived without developing cataract in the second eye, plus those who were cataract free at time a_0 but developed cataract in one eye only. This give us

$$N_1^U = N_0^H P_{HU} + N_0^U P_{UU}$$
$$N_1\pi_1^U = N_0(1 - \pi_0^U - \pi_0^B)P_{HU} + N_0\pi_0^U P_{UU}, \tag{2.8}$$

where P_{HU} is the probability that an initially healthy person stays alive but develops cataract in one eye only, and P_{UU} is the person that a person with unilateral cataract at time a_0 survives without developing bilateral cataract.

Finally the pool of people with bilateral cataract (2nd eye affected) at time a_1 has three sources: healthy individuals at time a_0 who develop unilateral, then bilateral cataract; those with unilateral cataract at time a_0 who survive and develop bilateral cataract; and those with bilateral cataract at time a_0 who survive. This gives

$$N_1^B = N_0^H P_{HB} + N_0^U P_{HU} + N_0^B P_{BB}$$
$$N_1\pi_1^B = N_0(1 - \pi_0^U - \pi_0^B)P_{HB} + N_0\pi_0^U P_{HU} + N_0\pi_0^B P_{BB}, \tag{2.9}$$

where P_{HB} is the probability that a person healthy at a_0 develops unilateral, then bilateral, cataract by age a_1. P_{UB} is the probability that a person with unilateral cataract at a_0 survives but develops bilateral cataract by age a_1. P_{BB} is the probability that a person with bilateral cataract at a_0 survives until age a_1.

We combined the three equations above, eliminating the total number of people at each time, to yield two equations for the two unknowns λ_{HU} and λ_{UB} in terms of prevalence and probability of death only. In terms of probabilities P_{ij}, these equations are

$$\frac{(1 - \pi_0^U - \pi_0^B)}{(1 - \pi_1^U - \pi_1^B)}\pi_1^U P_{HH} = (1 - \pi_0^U - \pi_0^B)P_{HU} + \pi_0 P_{UU}, \text{ and} \tag{2.10}$$

$$\frac{(1 - \pi_0^U - \pi_0^B)}{(1 - \pi_1^U - \pi_1^B)}\pi_1^B P_{HH} = (1 - \pi_0^U - \pi_0^B)P_{HB} + \pi_0^U E + \pi_0^B. \tag{2.11}$$

Notice that these equations are derived directly from our disease progression model (which transitions are possible) and are independent of the distribution of time it takes to progress from any state to any other state. In the next section, we choose a specific model for these transition times in order to calculate the transition probabilities in terms of known mortality rates and the incidence rates we wish to calculate.

2.2.3 Transition Probabilities

To find the transition probabilities, we follow the logic described by Lagakos (Lagakos, 1976) and later used by Podgor (Podgor and Leske, 1986) to derive the transition probabilities in his model. We assume that the time at which people make any allowed transition can be described by an exponential model with appropriate characteristic rate, λ. For transitions to state D, the rate parameters $\lambda_{HD}, \lambda_{UD}, \lambda_{BD}$ are determined by known death rates with and without cataract. The other rate parameters, λ_{HU} and λ_{UB}, are the first and second eye cataract development rates we wish to calculate and are closely related to the unilateral and bilateral incidence rates.

The first set of probabilities, P_{HH}, P_{UU}, and P_{BB}, represent, respectively, the chance of remaining in the state H, U, or B for the full time period (zero state transitions). To calculate P_{HH}, for example, notice that because incidence and mortality are independent events, the probability density function (pdf) for leaving the H state by any means is

$$f_T(t) = (\lambda_{HD} + \lambda_{HU})e^{-(\lambda_{HD}+\lambda_{HU})t}.$$

So the probability of remaining in state H for at least one unit of time is

$$P_{HH} = \int_1^\infty (\lambda_{HD} + \lambda_{HU})e^{-(\lambda_{HD}+\lambda_{HU})t}dt$$
$$= e^{-(\lambda_{HD}+\lambda_{HU})}. \tag{2.12}$$

Similarly, transition out of state U has an exponential distribution with parameter $(\lambda_{UD} + \lambda_{UB})$ and transition out of B has rate parameter λ_{BD} since mortality is the only remaining transition option. This gives us

$$P_{UU} = e^{-(\lambda_{UD}+\lambda_{UB})} \tag{2.13}$$
$$P_{BB} = e^{-\lambda_{BD}}. \tag{2.14}$$

To evaluate the integrals, we make the explicit choice that $a_1 = a_0 + 1$. The result is that all rates will be expressed in units of people per time interval between age groups in the data set (usually $1 - 5$ years). Notice that we

were able to take $t = 0$ at the beginning of the time interval we're considering (age a_0), when the known distribution of T has $t = 0$ when the person first enters the state of interest. This simplifies the calculation greatly, and it is possible because of the memoryless property of the exponential distribution. P_{HH} is really the conditional probability that a person remains healthy at age $a_1 = a_0 + 1$ given that they were healthy at age a_0. However, due to memorylessness, this conditional probability is the same as the probability that a person is healthy at age 1 given that they are healthy at age 0.

Transition probabilities P_{HU} and P_{UB} represent the chance of progressing by exactly one state, from H to U and from U to B, respectively. These are analogous to one-stage advancement probability in Podgor and Leske's model (probability B in that model, or the probability of advancing to state I from state H). In the case of P_{HU}, we have the joint pdf

$$f_{T_D,T_U}(t_D|t_U) = \frac{\lambda_{HU}}{\lambda_{HD} + \lambda_{HU}}(\lambda_{HD} + \lambda_{HU})e^{-(\lambda_{HD}+\lambda_{HU})t_U}$$
$$\cdot (\lambda_{UD} + \lambda_{UB})e^{-(\lambda_{UD}+\lambda_{UB})(t_D-t_U)},$$

for the time T_U of transition from H to U and time T_D of leaving state U (often due to death but also due to a transition to state B). Here $\frac{\lambda_{HU}}{\lambda_{HD}+\lambda_{HU}}$ is the probability that a person who leaves the healthy state develops unilateral cataract (rather than dying). So

$$P_{HU} = P(T_D > 1, 0 < T_U < 1)$$
$$= \int_1^\infty \int_0^1 \lambda_{HU}e^{-(\lambda_{HD}+\lambda_{HU})t_U}(\lambda_{UD} + \lambda_{UB})e^{-(\lambda_{UD}+\lambda_{UB})(t_D-t_U)}dt_Udt_D$$
$$= \int_0^1 \lambda_{HU}e^{-(\lambda_{HD}+\lambda_{HU})t_U}e^{-(\lambda_{UD}+\lambda_{UB})(1-t_U)}dt_U$$
$$= \frac{\lambda_{HU}}{\lambda_{HD} + \lambda_{HU} - \lambda_{UD} - \lambda_{UB}}\left(e^{-(\lambda_{UD}+\lambda_{UB})} - e^{-(\lambda_{HD}+\lambda_{HU})}\right). \quad (2.15)$$

Similarly, probability P_{UB} is given by

$$P_{UB} = \int_0^1 e^{-(\lambda_{UD}+\lambda_{UB})t_B}\lambda_{UB}e^{-\lambda_{BD}(1-t_B)}dt_B$$
$$= \frac{\lambda_{UB}}{\lambda_{UD} + \lambda_{UB} - \lambda_{BD}}\left[e^{\lambda_{BD}} - e^{-(\lambda_{UD}+\lambda_{UB})}\right]. \quad (2.16)$$

The probability, P_{HB}, of having bilateral cataract at age $a + t$ given that a person was healthy at age a is more complicated than any transition probability in Podgor and Leske's model because it represents a two-stage

disease progression. Nevertheless, the extension is fairly straightforward when we assume exponential distributions for all the processes. The joint pdf for transition times T_U to unilateral cataract, T_B to bilateral cataract, and time of death T_D is

$$f_{T_U,T_B,T_D}(t_U, t_B, t_D) = \lambda_{HU}e^{-(\lambda_{HD}+\lambda_{HU})t_U}\lambda_{UB}e^{-(\lambda_{UD}+\lambda_{UB})(t_B-t_U)}\lambda_{BD}e^{-\lambda_{BD}(t_D-t_B)}.$$

So P_{HB}, the probability that $0 < T_U < T_B < 1 < T_D$, is given by

$$
\begin{aligned}
P_{HB} &= \int_0^1 \int_{t_U}^1 \int_1^\infty \lambda_{HU}e^{-(\lambda_{HD}+\lambda_{HU})t_U} \\
&\quad \cdot \lambda_{UB}e^{-(\lambda_{UD}+\lambda_{UB})(t_B-t_U)}\lambda_{BD}e^{-\lambda_{BD}(t_D-t_B)}dt_D dt_B dt_U \\
&= \int_0^1 \int_{t_U}^1 \lambda_{HU}e^{-(\lambda_{HD}+\lambda_{HU})t_U}\lambda_{UB}e^{-(\lambda_{UD}+\lambda_{UB})(t_B-t_U)}e^{-\lambda_{BD}(1-t_B)}dt_B dt_U \\
&= \frac{\lambda_{HU}\lambda_{UB}}{\lambda_{UD}+\lambda_{UB}-\lambda_{BD}} \\
&\quad \cdot \left[\frac{e^{-(\lambda_{HD}+\lambda_{HU})}-e^{-(\lambda_{UD}+\lambda_{UB})}}{\lambda_{HD}+\lambda_{HU}-\lambda_{UD}-\lambda_{UB}} - \frac{e^{-(\lambda_{HD}+\lambda_{HU})}-e^{-\lambda_{BD}}}{\lambda_{HD}+\lambda_{HU}-\lambda_{BD}} \right].
\end{aligned}
\tag{2.17}
$$

Substituting the transition probabilities derived above into Equations 2.10 and 2.11, we obtain

$$
\begin{aligned}
\frac{(1-\pi_0^U-\pi_0^B)}{(1-\pi_1^U-\pi_1^B)}\pi_1^U e^{-(\lambda_{HD}+\lambda_{HU})} &= \frac{(1-\pi_0^U-\pi_0^B)\lambda_{HU}}{\lambda_{HD}+\lambda_{HU}-\lambda_{UD}-\lambda_{UB}} \\
&\quad \cdot \left(e^{-(\lambda_{UD}+\lambda_{UB})}-e^{-(\lambda_{HD}-\lambda_{HU})} \right) + \pi_0 e^{-(\lambda_{UD}+\lambda_{UB})},
\end{aligned}
\tag{2.18}
$$

and

$$
\begin{aligned}
\frac{(1-\pi_0^U-\pi_0^B)}{(1-\pi_1^U-\pi_1^B)}\pi_1^B e^{-(\lambda_{HD}+\lambda_{HU})} &= (1-\pi_0^U-\pi_0^B)\frac{\lambda_{HU}\lambda_{UB}}{\lambda_{UD}+\lambda_{UB}-\lambda_{BD}} \\
&\quad \cdot \left[\frac{e^{-(\lambda_{HD}+\lambda_{HU})}-e^{-(\lambda_{UD}+\lambda_{UB})}}{\lambda_{HD}+\lambda_{HU}-\lambda_{UD}-\lambda_{UB}} - \frac{e^{-(\lambda_{HD}+\lambda_{HU})}-e^{-\lambda_{BD}}}{\lambda_{HD}+\lambda_{HU}-\lambda_{BD}} \right] \\
&\quad + \pi_0^U\frac{\lambda_{UB}}{\lambda_{UD}+\lambda_{UB}-\lambda_{BD}}\left[e^{\lambda_{BD}}-e^{-(\lambda_{UD}+\lambda_{UB})} \right] + \pi_0^B e^{-\lambda_{BD}}.
\end{aligned}
\tag{2.19}
$$

Equations 2.18 and 2.19 are two equations in terms of the two unknown quantities of interest, λ_{HU} and λ_{UB}, and can be solved numerically. It is important to note that since Equations 2.18 and 2.19 are nonlinear, we have no

guarantee that they uniquely determine incidence. In practice for bilateral cataract, we solve these equations for two different starting values for every set of prevalence data points to test for consistency. The two solutions are consistent for cataract prevalence data or realistic simulated data. (In fact, in practice inconsistent solutions can be used to detect unrealistic randomly generated data used for confidence intervals, as described in Section 3.1.1). Future work should include a more formal check for uniqueness. For example, a grid of starting values spanning a square feasible region of incidence from 0 to 5 for both unilateral and bilateral incidence could be covered by 900 points in a 30 by 30 square with resolution smaller than our confidence intervals for incidence.

The λ_{HU} value we find in this way can be interpreted directly as five-year (if the time interval between the two prevalence endpoints is five years), first eye cataract incidence. To find the bilateral incidence from λ_{UB} (which indicates incidence of second eye cataract among those who already have cataract in one eye), we multiply by the prevalence of unilateral cataract averaged over the initial and final age periods. Notice that the calculated incidence is a rate and can be easily converted to any desired time interval. For example, an incidence rate per five years can be divided by five to give an annual incidence rate.

2.2.4   Comparing Unilateral and Bilateral Incidence

The memorylessness of the exponential model invites an additional comparison between unilateral incidence, bilateral incidence, and mortality that could be helpful to ophthalmologists. Once mortality and both incidences are known (or estimated), we can easily compute both the expected time to remain in a given disease stage and the percentage of people who die before progressing to the next stage.

For example, for people in the healthy stage, there are two possible transitions leaving that stage: death and development of unilateral cataract. Adding the two transition rates, the overall time to leave state H is exponentially distributed with parameter $\lambda_{HD} + \lambda_{HU}$. Therefore the expected time to remain in stage H is

$$E(T_{\text{remain healthy}}) = \int_0^\infty (\lambda_{HD} + \lambda_{HU})e^{-(\lambda_{HD}+\lambda_{HU})t}dt$$

$$E(T_{\text{remain healthy}}) = \frac{1}{\lambda_{HD} + \lambda_{HU}}. \tag{2.20}$$

Additionally, the percentage of people age a who develop unilateral

cataract before dying is

$$P(U|\text{transition at } a) = \frac{\lambda_{HU}}{\lambda_{HD} + \lambda_{HU}}. \tag{2.21}$$

Similarly, for people with unilateral cataract we have an expected time

$$E(T_{\text{remain unilateral}}) = \frac{1}{\lambda_{UD} + \lambda_{UB}}. \tag{2.22}$$

to remain in the unilateral cataract stage by either developing bilateral cataract or dying. For people who leave the unilateral stage at age a, the probability of developing unilateral cataract is

$$P(B|\text{transition at } a) = \frac{\lambda_{UB}}{\lambda_{UD} + \lambda_{UB}}. \tag{2.23}$$

Chapter 3

Confidence Intervals for Incidence

Confidence intervals for incidence were calculated by a parametric bootstrap method, described in Section 3.1. We made the assumption that the primary source of error in incidence was from random error in our prevalence data. Section 3.2 contains a discussion of excluded sources of error, efforts to minimize them, their likely impact on incidence estimates. However, because our motivation to estimate incidence stems from the limited amount of prevalence data, we first estimate the impact of prevalence data limitations on incidence results.

3.1 Numerical Simulation for Confidence Intervals

Prevalence data is directly used to calculate incidence data points for each age group, geographic district, and visual acuity level. We therefore first derive (Section 3.1.1) confidence intervals for these incidence data points based on simulated prevalence data. Confidence intervals for age-specific incidence are then combined (Section 3.1.2) to form confidence intervals for overall incidence in a particular geographic district at the specified visual acuity level.

3.1.1 Confidence Intervals for Age-Specific Incidence

Each incidence data point (for a particular district, VA level, and age group) is calculated based on four prevalence values (unilateral and bilateral prevalence at times a and $a + 1$), known healthy mortality rates, and a mortality

ratio estimate (1.5 for cataract). To generate confidence intervals, we simulate the four prevalence values, but leave the mortality values fixed (assuming they are not a significant source of error, comparatively).

For each age-specific incidence value, our simulation implemented the following steps:

1. The specified number of trial prevalence values are generated (we use 200 trials) based on distributions that model the expected random error in the original prevalence values. Trial unilateral and bilateral prevalence at time $t = 0$ (p_0^U and p_0^B) and at time $t = 1$ (p_1^U and p_1^B) are generated from two multinomial distributions. That is, each person is randomly chosen to be in either state H, U, or B with probabilities based on prevalence. The probability mass function (pmf) for this distribution is

$$f(n_t^U, n_t^B; n_t, \pi_t^U, \pi_t^B) = \Pr(N_t^U = n_t^U, N_t^B = n_t^B$$
$$\text{and } N_t^H = n_t - n_t^U - n_t^B) \qquad (3.1)$$
$$= \frac{n_t!}{n_t^U! n_t^B! (n_t - n_t^U - n_t^B)!} (\pi_t^U)^{n_t^U}$$
$$\cdot (\pi_t^B)^{n_t^B} \cdot (1 - \pi_t^U - \pi_t^B)^{(n_t - n_t^U - n_t^B)}, \qquad (3.2)$$

   where n_t is the total number of people at time t and π_t^U and π_t^B give data-generated unilateral and bilateral prevalence, respectively, at time t. Trial prevalence for state s at time t is then computed from the number of people in each state by the normalizing relation $p_t^s = n_t^s / n_t$.

2. Unilateral and bilateral incidence are calculated from each set of trial prevalence values. Incidence calculations are complicated by the fact that it is possible to randomly generate unrealistic prevalence value combinations, for example if prevalence decreases more with age than can be accounted for by mortality, even with zero incidence. However, in practice such unrealistic prevalence values occur only in $1 - 2\%$ of trials, and can be found by solving for incidence twice using two different sets of starting values for Newton's method. If the trial prevalence values are unrealistic, the two solutions for incidence will be inconsistent (and perhaps very large, since Newton's method will not converge). These trials are discarded so they do not influence mean and variance computations.

3. The mean and variance of each resulting vector of incidence values are computed.

4. 95% (age-group specific) confidence intervals for incidence are given by

$$(\bar{I}_a - 1.96 \cdot \sigma_{I_a}, \bar{I}_a + 1.96 \cdot \sigma_{I_a}), \tag{3.3}$$

where $\bar{I}_a$ is the mean and σ_{I_a} is the square root of the variance of the incidence trials for the age group a. This confidence interval is based on an approximation that simulated incidence values are normally distributed (see Section 3.1.3 for discussion).

3.1.2   Confidence Intervals for Overall Incidence

Earlier, we calculated overall incidence for each district and visual acuity level by taking a weighted average of incidence in each age group, with weights equalling the percentage of the surveyed population falling into each age group. That is, the overall incidence I is given by

$$I = \sum_{\text{age groups } a} p_a \cdot I_a, \tag{3.4}$$

where p_a and I_a are, respectively, the proportion of this district's population having age a and the age-specific incidence at age a.

Therefore, by properties of linear combinations of random variables, the mean and variance of overall incidence are given by

$$\bar{I} = \sum_{\text{age groups } a} p_a \cdot \bar{I}_a, \tag{3.5}$$

$$\sigma_I^2 = \sum_{\text{age groups } a} p_a^2 \cdot \sigma_{I_a}^2. \tag{3.6}$$

To obtain confidence intervals from sample mean and variance, we again make the approximation that I is normally distributed; this assumption is justified in the next section.

3.1.3   Normality

The method just described for computing confidence intervals for incidence depends on the assumption that simulated incidence trials are normally distributed. For age-specific incidence, the mean and variance were computed directly from trial data, however the confidence intervals themselves

assumed normality of the trial data. Similarly, since overall incidence is simply a linear combination of age-specific incidences, variance in overall incidence could be computed in a distribution-independent way based on the variance of age-specific incidences. However, to obtain confidence intervals from variance we assumed that simulated overall incidence was also normally distributed. In this section, we argue that in practice simulated age-dependent cataract incidence is normally distributed. We use this observation to justify our normality assumption about overall incidence as well.

We looked at histograms and quantile-quantile plots of a few sets of age-specific incidence trials that were least likely to be normally distributed because of their small sample size. For example, incidence trials for Mali at age 83-87 at VA < 6/18 are shown in Figure 3.1a and Figure 3.1b.

Both the histogram and quantile-quantile plot support the claim that the age-specific incidence trials are normally distributed. It therefore seems reasonable to assume that incidence trials for other age and visual acuity combinations with larger sample sizes would also be normally distributed. If age-specific incidence trials are normally distributed, then their linear combinations will also be normally distributed.

3.2 Discussion

Because the difficulty of obtaining large quantities of data in Africa was a primary motivation for our research, random error in prevalence (based on sample sizes relative to the proportion of people with cataract) was the main source of error considered when estimating confidence intervals for incidence. Other possible sources of error would depend on specific applications of our model, such as prevalence data collection methodology and sources of mortality rate data.

In the case of cataract prevalence data in sub-Saharan Africa, possible sources of error include systematic error between RAAB district surveys conducted by different teams, error in the mortality rates reported by the World Health Organization or systematic error introduced by our use of the same mortality rates for all countries, or error introduced by our assumptions about the ratio of diseased to healthy mortality. Two of these sources of error are especially important: systematic biases in data collection methodology between districts (discussed in Section 4.6), and error in mortality parameters that were also input into our model (investigated through sensitivity analysis in Section 4.4.1).

Histogram of IncU.trials[i,]
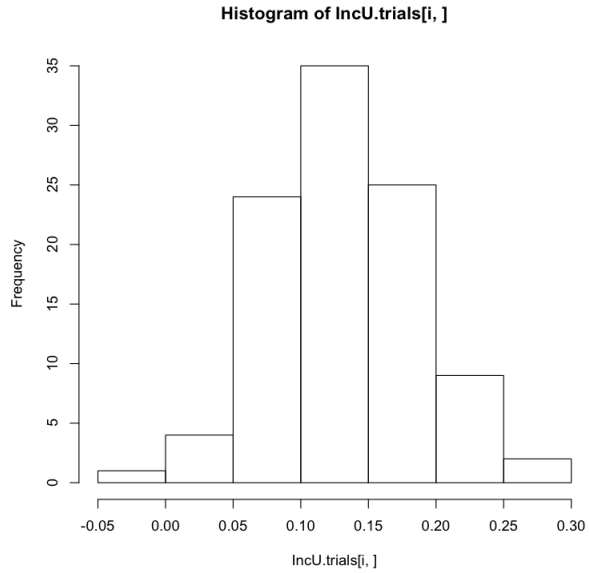
(a) Sample Histogram of Incidence Trials.

Normal Q-Q Plot
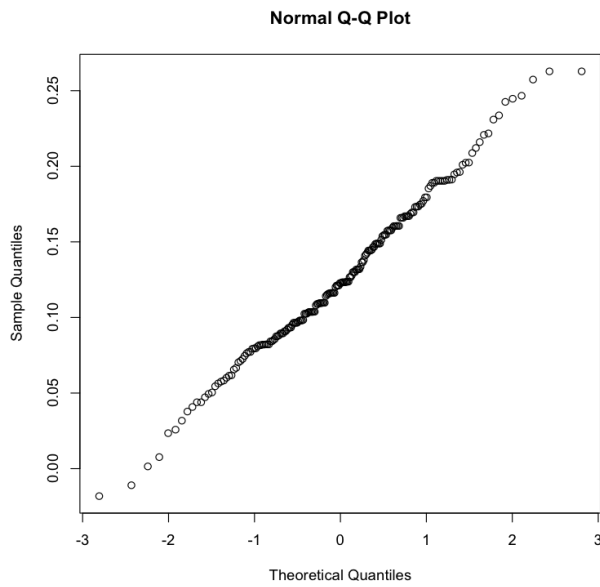
(b) Normal Quantile-Quantile Plot of Incidence Trials.

Figure 3.1: Example Normality Tests For Trial Incidence Data (From Mali, Ages 83-87 at VA $< 6/18$).

Chapter 4

Application to Cataract Incidence in Africa

Having developed an incidence estimation model for bilateral diseases, we are ready to address the challenge of cataract incidence estimation in sub-Saharan Africa. In Section 4.1, we discuss the survey methodology used to obtain data on cataract prevalence in Africa. Section 4.2 describes the details of how our incidence estimation model is implemented to process this data. Section 4.3 presents incidence estimation results for Africa. In Sections 4.4 and 4.5, we further analyze our results in order to address, respectively, questions about geographic variation in cataract incidence and differences between unilateral and bilateral incidence. Finally, Section 4.6 discusses the assumptions of our model and their validity in the case of cataract incidence estimation in sub-Saharan Africa.

4.1   Survey Methodology

Data on cataract prevalence in Africa was derived from Rapid Assessment of Avoidable Blindness (RAAB) surveys in seven African districts. The RAAB survey methodology was designed to assess avoidable blindness due to cataract and other diseases. Survey teams consist of ophthalmologists, ophthalmic assistants, and support staff including a data entry clerk. The teams attend training which includes a standardization workshop measuring the agreement of their surveys with surveys conducted by more experienced teams (Limburg and Meester, 2007).

Survey participants are selected via a cluster sampling method, where population units of 50 people over age 50 (experience has shown that teams

can generally survey 50 people in one day if they live within a small geographic area) are randomly chosen based on census data from among all population units in a 1-2 million person district of interest. Only people over age 50 are surveyed since 80% of cataract cases occur in people over 50, and this greatly reduces the sample size necessary compared to a survey of the entire population. Surveys generally reach $1500 - 3000$ people using this cluster sampling method (Limburg and Meester, 2007: p. 6-12).

Each survey participant receives a standard eye exam. People with visual acuity (VA) less than 6/18 on the Snellen scale in meters (20/60 vision in feet) are categorized as visually impaired and examined further using a pinhole to categorize VA as either $< 6/18$, $< 6/18$ but $\geq 6/60$, $< 6/60$ but $\geq 3/60$, $< 3/60$ but $\geq 1/60$, light perception, or no light perception. People with visual acuity less than 3/60 in the better eye with best optical correction are considered blind. To determine cause of blindness or visual impairment, the lens of the eye is also examined and classified as normal, obvious lens opacity, aphakic (likely because the lens was surgically removed in a cataract surgery), or pseudophakic. Participants found to have cataract or previous surgery are also asked about the surgery or why surgery was not performed (Limburg and Meester, 2007). Figure 4.1 shows the form used to collect RAAB survey data.

In deriving cataract prevalence from this survey data, Lewallen et al. (2010) considered all persons with lens opacity, aphakia or pseudophakia caused by cataract to have 'cataract' at a given visual acuity level. Thus both those with cataract and those who had previously had cataract removal surgery could be counted as having cataract.

4.2 Model Implementation

We applied our model to known prevalence and mortality data in order to estimate both unilateral and bilateral incidence in sub-Saharan Africa. Age-specific prevalence data for Africa was obtained from previous analysis of seven RAAB surveys by Lewallen et al. (2010). The survey districts used were: Kilimanjaro, Tanzania (Habiyakire et al., 2010); Kericho, Kenya (Kimani et al., 2008); Nakuru, Kenya (Mathenge et al., 2007a); Western Region, Rwanda (Mathenge et al., 2007b); The Gambia (Oye et al., 2009b); Koulikor, Mali (Oye et al., 2009a); and Eritrea (Mueller et al., 2009).

An important question raised by the work of Lewallen et al. (2010) is to what extent geographic variation in cataract prevalence translates into variation in incidence. The prevalence data suggests significant geographic

Annex 1.	**RAPID ASSESSMENT FOR AVOIDABLE BLINDNESS SURVEY RECORD**

A. General Information   Year – Month: ☐☐☐☐ - ☐☐
Survey area _____ ☐☐   **Cluster no.** ☐☐☐   **Individual no.**
Name _____   **Sex:**   Male O (1)   **Age** (years)   ☐☐
                                              Female O (2)

Optional 1 ☐☐   **Examination status:**
Optional 2   Examined: O (1) (go to B)   Refused: O (3) (go to E)
                      Not available: O (2) (go to E)   Not able to communicate: O (4) (go to E)

B. Vision – Presenting Vision

Glasses	without glasses: O (1)
	with available distance glasses: O (2)

	Right eye	**Left eye**
Can see 6/18	O (1)	O (1)
Cannot see 6/18, but can see 6/60	O (2)	O (2)
Cannot see 6/60, but can see 3/60	O (3)	O (3)
Cannot see 3/60 but can see 1/60	O (4)	O (4)
Light perception (PL+)	O (5)	O (5)
No light perception (PL-)	O (6)	O (6)

Best Vision – with best correction or pinhole

	Right eye	**Left eye**
Can see 6/18	O (1)	O (1)
Cannot see 6/18, but can see 6/60	O (2)	O (2)
Cannot see 6/60, but can see 3/60	O (3)	O (3)
Cannot see 3/60 but can see 1/60	O (4)	O (4)
Light perception (PL+)	O (5)	O (5)
No light perception (PL-)	O (6)	O (6)

C. Lens examination

	Right eye	**Left eye**	
Normal lens/minimal lens opacity	O (1)	O (1)	(go to D)
Visually impairing lens opacity	O (2)	O (2)	(go to D & F)
Lens absent (aphakia)	O (3)	O (3)	(go to D & G)
Pseudophakia without PCO	O (4)	O (4)	(go to D & G)
Pseudophakia with PCO	O (5)	O (5)	(go to D & G)
No view of Lens	O (6)	O (6)	(go to D)

D. Main cause presenting VA<6/18
    (mark only one cause for each eye)

	Right eye	**Left eye**	Mark one principal disorder
Refractive error	O (1)	O (1)	O (1)
Cataract, untreated	O (2)	O (2)	O (2)
Aphakia, uncorrected	O (3)	O (3)	O (3)
Surgical complications	O (4)	O (4)	O (4)
Trachoma	O (5)	O (5)	O (5)
Phthisis	O (6)	O (6)	O (6)
Other corneal scar	O (7)	O (7)	O (7)
Globe abnormality	O (8)	O (8)	O (8)
Glaucoma	O (9)	O (9)	O (9)
Diabetic retinopathy	O (10)	O (10)	O (10)
ARMD	O (11)	O (11)	O (11)
Onchocerciasis	O (12)	O (12)	O (12)
Post. segment / CNS disorder	O (13)	O (13)	O (13)
Not examined – can see 6/18	O (14)	O (14)	O (14)

E. History, if not examined
    (From relative or

Believed	**Right eye**	**Left eye**
Not blind	O (1)	O (1)
Blind due to cataract	O (2)	O (2)
Blind due to other causes	O (3)	O (3)
Operated for cataract	O (4)	O (4)

F. Why cataract operation was not done
(mark 1 or 2 responses, if VA<6/18, not improving with pinhole, with visually impairing lens opacity in one or both eyes)

Unaware that treatment is possible	O (1)
Believes it to be destiny / God's Will	O (2)
Told to wait for cataract to mature	O (3)
Surgical services not available or very far	O (4)
Don't know how to get surgery	O (5)
Cannot afford operation	O (6)
No one to accompany	O (7)
No time available / other priorities	O (8)
Old age and need not felt	O (9)
One eye adequate vision / need not felt	O (10)
Fear of operation	O (11)
Fear of loosing eyesight	O (12)
Other disease contra-indicating operation	O (13)

G. Details about cataract operation

	Right eye	**Left eye**
Age at operation (years)	☐☐	☐☐
Place of operation		
Government hospital	O (1)	O (1)
Voluntary / charitable hospital	O (2)	O (2)
Private hospital	O (3)	O (3)
Eye camp / improvised setting	O (4)	O (4)
Traditional setting	O (5)	O (5)
Type of surgery		
Non IOL	O (1)	O (1)
IOL implant	O (2)	O (2)
Couching	O (3)	O (3)
Cost of surgery		
Totally free	O (1)	O (1)
Partially paid	O (2)	O (2)
Totally paid	O (3)	O (3)
Cause of VA<6/18 after cataract surgery		
Ocular co-morbidity (Selection)	O (1)	O (1)
Operative complications (Surgery)	O (2)	O (2)
Refractive error (Spectacles)	O (3)	O (3)
Late complications (Sequelae)	O (4)	O (4)
Not applicable, can see 6/18	O (5)	O (5)
Are you satisfied with results of cataract surgery?		
Very satisfied	O (1)	O (1)
Partially satisfied	O (2)	O (2)
Indifferent	O (3)	O (3)
Partially dissatisfied	O (4)	O (4)
Very dissatisfied	O (5)	O (5)

Figure 4.1: RAAB survey instrument (Limburg and Meester, 2007: p. 65).

variation in prevalence, so as a preliminary measure of this variation, Lewallen et. al. grouped the seven surveyed districts into two groups based on 95% confidence intervals of prevalence. Group 2 (Eritrea, Mali and The Gambia), when pooled, showed prevalence 2.6, 2.6, and 2.5 times higher than Group 1 (Kenya, Tanzania, and Rwanda) at the three visual acuity levels of $< 6/18$, $6/60$, and $3/60$ respectively (Lewallen et al., 2010). In addition to computing incidence for each individual district, we also computed incidence for the pooled prevalence data of each of these groups.

The age-specific probability of death over the interval for each country was taken from life tables published by the World Health Organization (World Health Organization, 2009a). We used the average of the for the African countries from which we had RAAB survey data. The 5-year healthy death rate, λ_{HD}, was calculated from the 5-year probability of death, nq_X, as follows: $\lambda_{HD} = -\ln(1 - nq_X)$, where nq_X is a mortality table column described in depth in World Health Organization (2001). We assumed that the death rate in a person with cataract in one or both eyes would be 1.5 times that of someone without cataract. That is, we let $\lambda_{UD} = \lambda_{BD} = 1.5\lambda_{HD}$.

First and second eye cataract incidence were calculated for each age interval and using each of three visual acuity levels ($<6/18$, $<6/60$, $<3/60$) as cutoffs for blindness due to cataract. Incidence results are given in Section 4.3 and a discussion of time to develop unilateral and bilateral cataract is given in Section 4.5.

Confidence intervals for incidence were computed using the parametric bootstrap method described in Chapter 3. We generated 200 trial incidence values for each true (age-specific) incidence value estimated.

4.3   Incidence Results

Cataract incidence was found to increase with age at all visual acuity levels. To calculate the overall incidence in the survey population (50+ years old), we multiplied the age-specific incidence by the proportion of population in each age group. The 1-year incidence of blindness due to cataract in persons over 50 in each survey district is shown in Tables 4.1 and 4.2. Incidence values are given per 100 people, with overall incidence indicating the number of eyes per 100 people per year that develop cataract. Figure 4.2 shows the age dependence of incidence among four districts (Rwanda, Tanzania, and Kenya) that are both geographically close and show an overlap of 95% confidence intervals for prevalence. Figure 4.3 shows age dependence of

pooled prevalence for the same four districts.
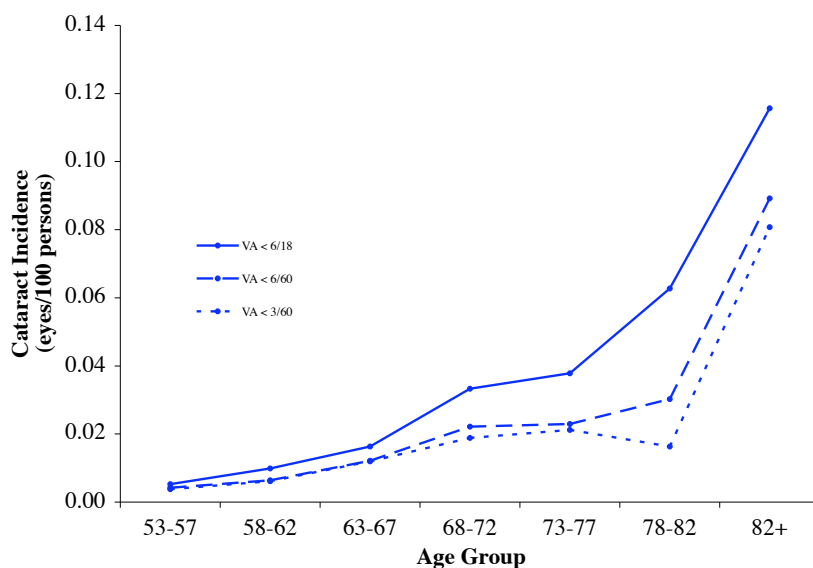
Figure 4.2: Age Dependence of Pooled Cataract Incidence.

Tables 4.1 and 4.2 give overall incidence (with confidence intervals) by district at each visual acuity level and Figure 4.4 shows a representative subset of the data in visual form, specifically, incidence by eyes at visual acuity $< 6/18$. The significance of the VA $< 6/18$ cutoff is that the World Health Organization defines low vision due to cataract at this threshold, and blindness due to cataract at VA $< 3/60$ (Limburg and Meester, 2007). Both thresholds represent significant lifestyle impact from cataract, hence we present both thresholds
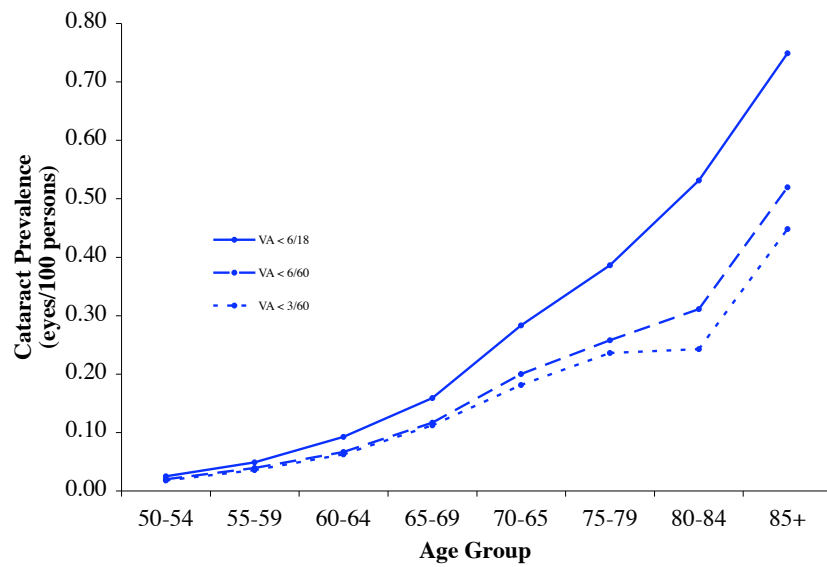
Figure 4.3: Age Dependence of Pooled Cataract Prevalence.

Study Site	< 6/18 VA		
	First Eye	Second Eye	Overall
Nakuru, Kenya	1.7	0.6	2.3
	(1.3,2.2)	(0.4,0.9)	(1.8,2.8)
Kilimanjaro, Tanzania	1.8	0.7	2.5
	(1.4, 2.3)	(0.5, 1.1)	(2.1, 3.1)
Western Province, Rwanda	1.5	0.6	2.2
	(1.1, 2.1)	(0.4, 1.1)	(1.6, 2.9)
Kericho, Kenya	2.2	0.8	3.0
	(1.7, 2.8)	(0.5, 1.3)	(2.4, 3.8)
Koulikor, Mali	4.6	1.8	6.4
	(3.9, 5.5)	(1.3, 2.4)	(5.5, 7.5)
Gambia	2.7	1.1	3.8
	(1.7, 4.2)	(0.6, 2.2)	(2.6, 5.5)
Eritrea	3.9	1.6	5.5
	(3.3, 4.7)	(1.2, 2.1)	(4.8, 6.4)

Study Site	< 6/60 VA		
	First Eye	Second Eye	Overall Cataract
Nakuru, Kenya	1.3	0.4	1.7
	(0.9, 1.7)	(0.2,0.7)	(1.3,2.2)
Kilimanjaro, Tanzania	1.1	0.4	1.5
	(0.8,1.5)	(0.2,0.7)	(0.8,1.8)
Western Province, Rwanda	0.9	0.3	1.2
	(0.6, 1.4)	(0.2, 0.7)	(0.8, 1.8)
Kericho, Kenya	1.7	0.6	2.2
	(1.2, 2.3)	(0.3, 0.9)	(1.7, 2.9)
Koulikor, Mali	3.0	1.1	4.1
	(2.4, 3.8)	(0.8, 1.6)	(3.4, 5)
Gambia	1.6	0.7	2.4
	(0.9, 2.9)	(0.3, 1.7)	(1.5, 3.8)
Eritrea	2.9	1.1	4.0
	(2.4, 3.6)	(0.8, 1.6)	(3.4, 4.8)

Table 4.1: Incidence of Visual Impairment Due to Cataract.

| Study Site | < 3/60 VA | | |
	First Eye	Second Eye	Overall Cataract
Nakuru, Kenya	1.1	0.4	1.5
	(0.8, 1.5)	(0.2, 0.6)	(1.1, 1.9)
Kilimanjaro, Tanzania	1.0	0.3	1.3
	(0.7,1.4)	(0.2,0.6)	(1,1.8)
Western Province, Rwanda	0.8	0.3	1.1
	(0.5, 1.3)	(0.1, 0.6)	(0.8, 1.6)
Kericho, Kenya	1.4	0.5	1.9
	(1, 2)	(0.3, 0.8)	(1.4, 2.5)
Koulikor, Mali	2.5	0.9	3.4
	(2, 3.2)	(0.6, 1.4)	(2.8, 4.2)
Gambia	1.3	0.8	2.1
	(0.7, 2.5)	(0.3, 1.8)	(1.2, 3.5)
Eritrea	2.6	1.0	3.6
	(2.1, 3.2)	(0.7, 1.4)	(3, 4.3)

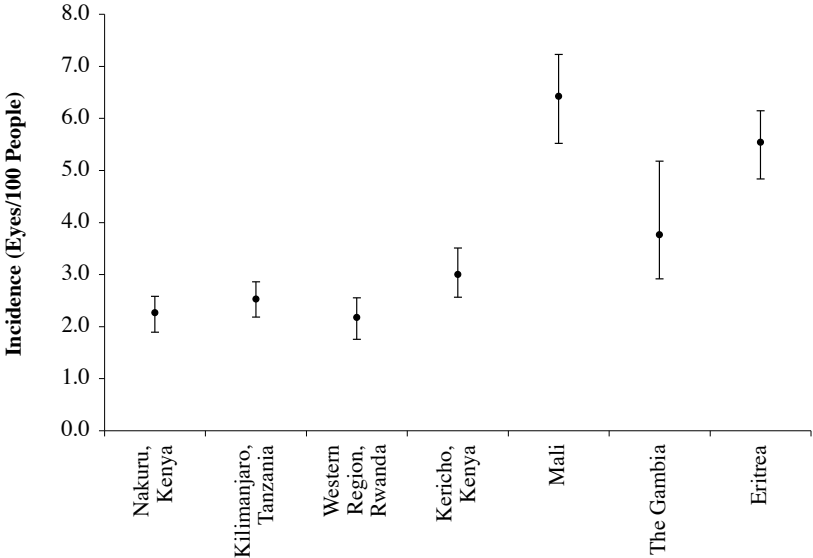Table 4.2: Incidence of Blindness Due to Cataract.

Figure 4.4: Incidence of Low Vision (Visual Acuity < 6/18) Due to Cataract.

4.4   Geographic Variation in Cataract Incidence

Our model provides new data regarding differences in cataract incidence among districts in sub-Saharan Africa. The difference in incidence between Mali and Kenya, for example, is large compared to the 95% confidence interval on either data point, indicating that random error in prevalence is unlikely to account for this difference. This type of geographic variation has great practical significance and should be explored further as more RAAB survey data becomes available.

We also compute incidence estimates for pooled cataract prevalence for each of the two rough groups suggested by Lewallen et al. based on confidence intervals for prevalence. Group 1, low-prevalence countries, consists of both Kenyan districts, Tanzania, and Rwanda. Group 2, high-prevalence countries, consists of Mali, The Gambia, and Eritrea Lewallen et al. (2010). This grouping is clearly not a perfect representation of the geographic variation in incidence, however it allows us to obtain a quasi-quantitative measure of geographic variation for comparison to other effects, such as variation based on age. Table 4.3 gives pooled incidence results for these two groups.

Study Site	< 6/18 VA		
	First Eye	Second Eye	Overall Cataract
Group 1	1.8	0.7	2.5
Kenya, Rwanda, Tanzania	(1.5,2.0)	(0.5,0.9)	(2.2,2.8)
Group 2	4.0	1.6	5.6
Eritrea, The Gambia, Mali	(3.5,4.5)	(1.3,2)	(5.1,6.2)
Study Site	< 6/60 VA		
	First Eye	Second Eye	Overall Cataract
Group 1	1.2	0.4	1.6
Kenya, Rwanda, Tanzania	(1,1.4)	(0.3,0.6)	(1.4,1.9)
Group 2	2.7	1.1	3.8
Eritrea, The Gambia, Mali	(2.4,3.2)	(0.8,1.3)	(3.3,4.3)
Study Site	< 3/60 VA		
	First Eye	Second Eye	Overall Cataract
Group 1	1.1	0.4	1.4
Kenya, Rwanda, Tanzania	(0.9,1.3)	(0.3,0.5)	(1.2,1.7)
Group 2	2.3	0.9	3.2
Eritrea, The Gambia, Mali	(2,2.7)	(0.7,1.2)	(2.8,3.7)

Table 4.3: Pooled Cataract Incidence.

4.4.1 Sensitivity to Mortality Parameter Estimate

When analyzing geographic variation in cataract incidence, it is important to consider whether any part of our model introduces geographic biases. Bias could originate either in data collection procedures (discussed in more detail in Section 4.6) or in the model itself. In addition to prevalence data, we use age-dependent mortality rates as data in our model. Since we use the same average mortality rate for all countries studied, it is unlikely that mortality introduces a geographic bias. However, we also introduce a mortality ratio parameter giving the ratio of diseased to healthy mortality, which could significantly impact incidence estimates. Although we also use the same mortality ratio for all countries, it is important to verify that different mortality ratios do not affect the relative scale of geographic differences and confidence intervals, which could render our results less significant. Here we give an initial, quantitative estimate of the impact of this parameter on incidence and discuss whether this parameter could account for the geographic variation in incidence noted in the previous section.

We varied the mortality ratio parameter over its range of reasonable values, from 1 (indicating cataract disease has no impact on mortality) to 2 (indicating cataract doubles mortality), in increments of 0.25. Figure 4.5 shows the impact of mortality ratio on incidence for the highest-incidence district (Mali) and lowest-incidence district (Nakuru, Kenya) surveyed. In both cases, the data suggests a linear relationship between incidence and the mortality rate parameter. For both countries, incidence estimates for a mortality ratio of 1.25 or 1.75 fell within a 95% confidence interval for our original incidence estimate with mortality ratio 1.5. This suggests that error in incidence due to our mortality ratio estimate is comparable to or less than our confidence intervals for incidence based on RAAB survey sample sizes.

While mortality ratio estimates are not the most significant source of error in our model, it is important to consider whether they could introduce a geographic bias between countries. We computed incidence by eyes with each mortality estimate for all seven districts at visual acuity $< 6/18$. Figure 4.6 shows the effect of decreasing the mortality ratio below 1.5 and Figure 4.7 shows the effect of increasing this ratio. While incidence increases with the mortality ratio parameter, the differences between incidence levels in different districts remain unaffected by the mortality ratio. In all cases, Eastern African districts (both Kenya districts, Tanzania, and Rwanda) had overlapping 95% confidence intervals. The difference between Eastern African districts and Mali and Eritrea remained large compared to the confidence intervals shown, and the confidence interval for
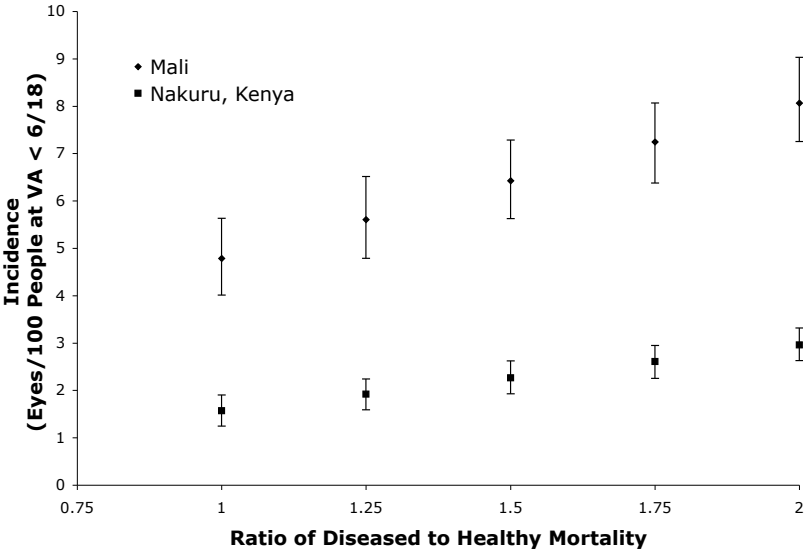
Figure 4.5: Incidence Dependence on Mortality Ratio.

the Gambia remained large and continued to overlap those of most other districts.

The sensitivity analysis presented here is certainly preliminary. Future work could further investigate the role of mortality rate estimates in our model, especially our assumption that mortality rates with unilateral and bilateral cataract are equal. It is likely that allowing the ratio of unilateral to bilateral mortality to vary as well would affect unilateral and bilateral incidence, the ratio of unilateral to bilateral incidence, and incidence by eyes. A more complex version of our model could also include a mortality ratio parameter that varied with age. However, while more sophisticated mortality estimates could improve the accuracy of incidence calculations, they are unlikely to account for the strong geographic variations in incidence we observe.
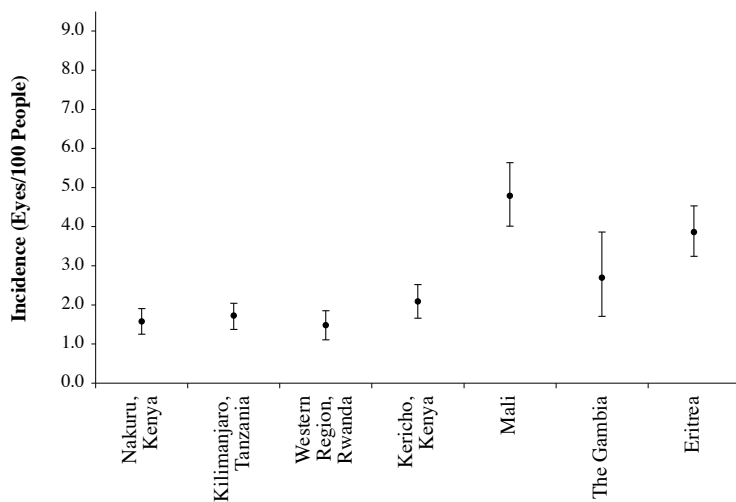
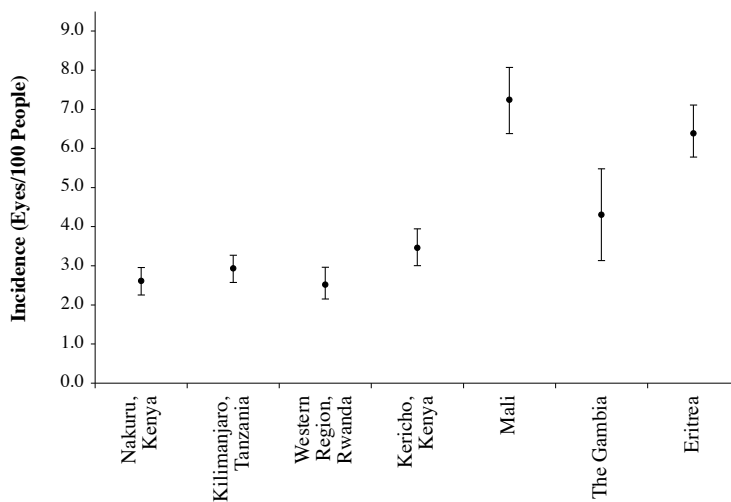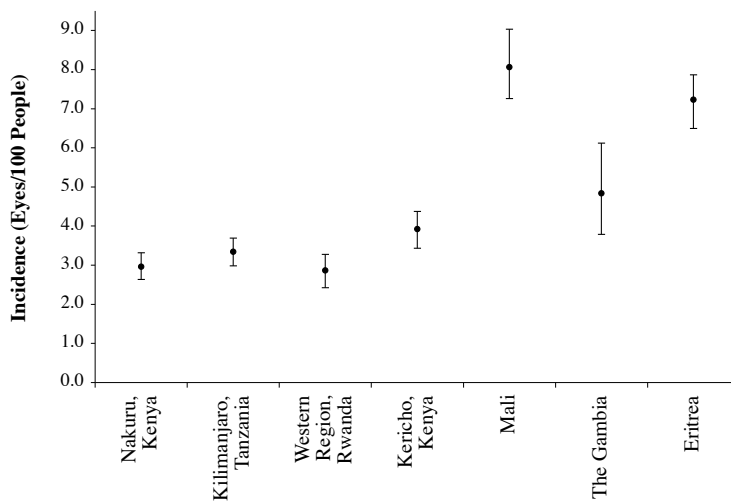(a) Incidence with Mortality Ratio 1.
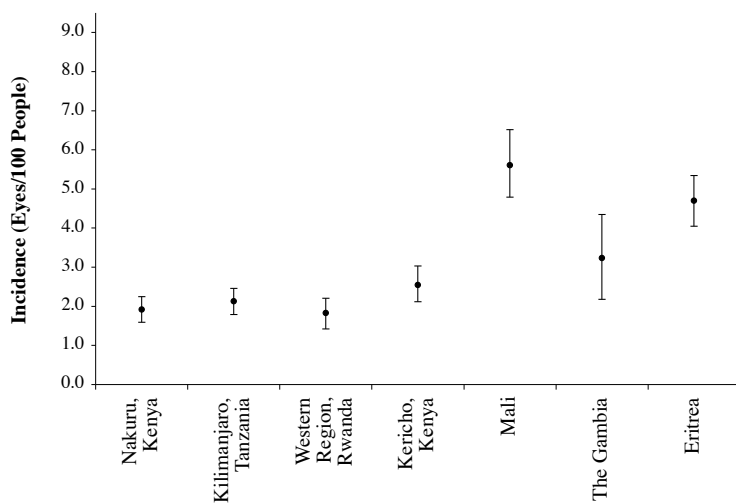
(b) Incidence with Mortality Ratio 1.25.

Figure 4.6: Effect of Reduced Mortality Ratio.

(a) Incidence with Mortality Ratio 1.75.


(b) Incidence with Mortality Ratio 2.

Figure 4.7: Effect of Increased Mortality Ratio.

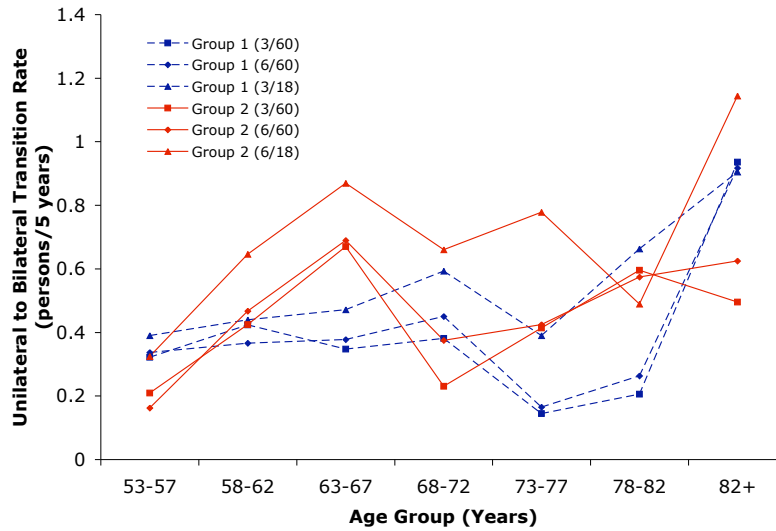## 4.5   Comparison of Unilateral and Bilateral Incidence

One feature of our model is that it allows separate calculation of unilateral and bilateral incidence. These two types of incidence behave qualitatively differently in a way that suggests future research in this area could give more information about the development of cataract disease itself. Here we explore some of these differences. Figure 4.8 shows pooled unilateral incidence and bilateral transition rates for Group 1 and Group 2. Unilateral incidence (λ_{HU} in our model) is compared with the rate of transition to bilateral cataract among those who already have unilateral cataract (λ_{UB}, or "bilateral transition rate") in order to separate the processes of developing first and second-eye cataract. The bilateral transition rate was used instead of bilateral incidence since bilateral incidence in the full population depends on the prevalence of unilateral cataract as well as the rate people with unilateral cataract develop bilateral cataract.

Interestingly, a comparison of the two graphs shows qualitatively different behavior of unilateral and bilateral transition rates. Unilateral incidence increases with age and there is a clear distinction between incidence in Group 1 and in Group 2; the difference between groups is usually larger than the difference between different visually acuity levels. The latter difference is known to be significant to policy decisions since in many areas cataract surgical rate (CSR) targets are set differently for the visual acuity levels shown in an attempt to manage scarce resources. In contrast, in the case of bilateral transition rates, noise seems to dominate over both age-based and country group distinctions. Further work in both data collection and statistical analysis of the data presented would help illuminate this issue. One extremely interesting hypothesis is that perhaps the progression from unilateral to bilateral cataract is more characteristic of the cataract disease itself rather than of specific environmental or genetic influences in any given region.

The transition from unilateral to bilateral cataract may be especially interesting to ophthalmologists. Due to limited resources, ophthalmologists may need to choose between operating on a person with unilateral cataract or waiting until they develop bilateral cataract to do a more efficient double surgery. In order to evaluate such a strategy, two types of data from our computed transition rates may be helpful. Our model allows us to calculate the probability of a person in a given state (healthy or unilateral) progressing to the next state, that is, the probability that a person ever develops the next stage of cataract before passing away. Figure 4.9 compares probability of progression to unilateral and bilateral cataract states. A sec-
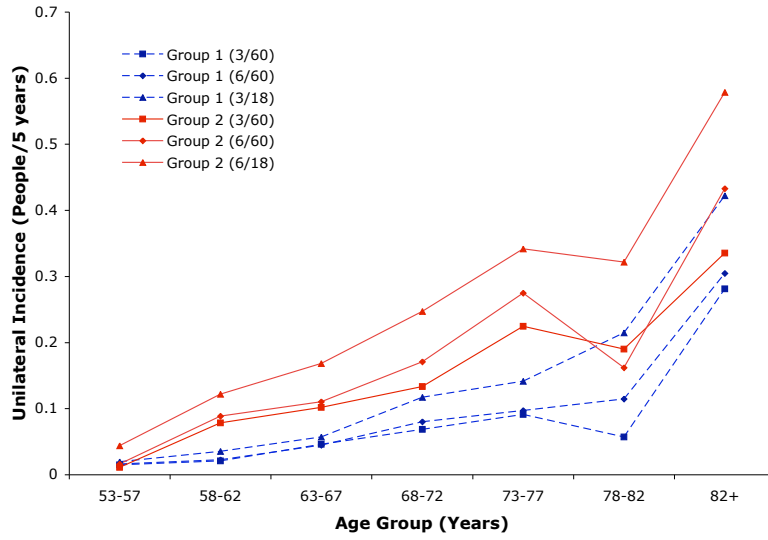
(a) Unilateral Incidence by Group.

(b) Bilateral Transition Rate by Group.

Figure 4.8: Comparison of Unilateral and Bilateral Transition Rates.

ond, related calculation gives the expected time to remain in a given state (again, healthy or unilateral) before the transition (either developing the next stage of cataract disease or death) takes place. Figure 4.10 compares expected transition times to leave the healthy and unilateral states, which correspond respectively to transitions to the unilateral (or deceased) and bilateral (or deceased) states.

Our initial results suggest that the strategy of waiting until a person develops bilateral cataract before operating is problematic for two reasons. For younger persons, the waiting time to develop cataract in the second eye may be significant, causing a significant period of lower quality of life for that person due to partial blindness. For older persons, the percentage who develop bilateral cataract at all drops well below 50%, so this policy would bar the majority from access to any type of cataract surgery. Our data complements the safety argument based on our colleague's experience that earlier cataract surgeries are safer and have higher success rates (Lewallen, 2010). Again, future work especially additional statistical analyses of this data could prove fruitful. Our intention here is to show the rich variety of information available in the transition rate data estimated by our model, and suggest avenues that our work opens for future study.
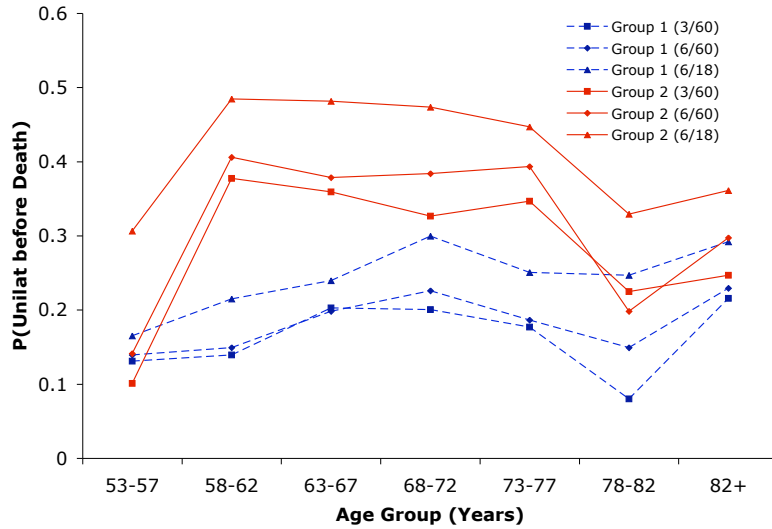
(a) Probability of Developing Unilateral Cataract.
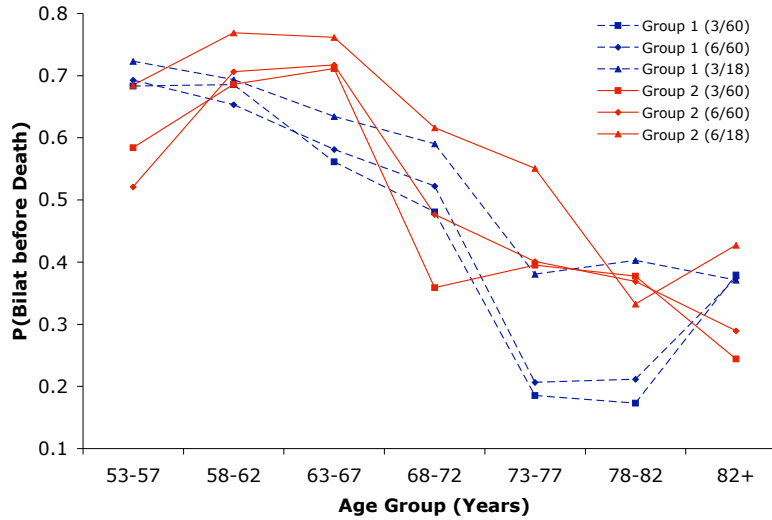
(b) Probability of Developing Bilateral Cataract.

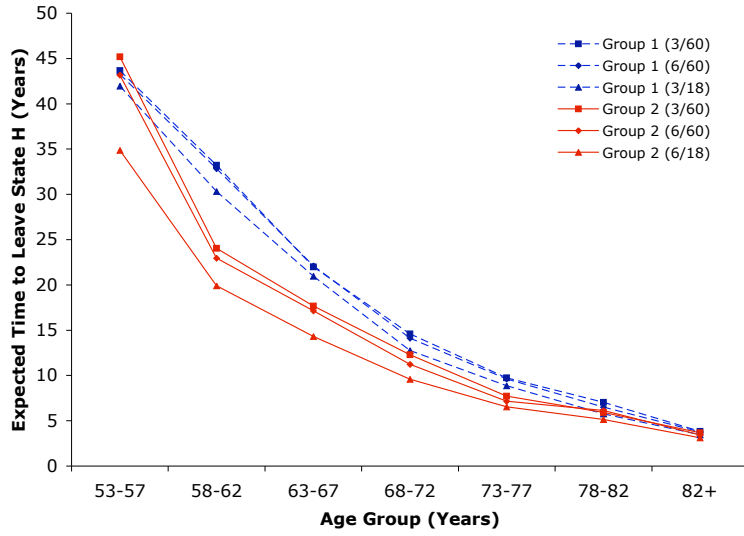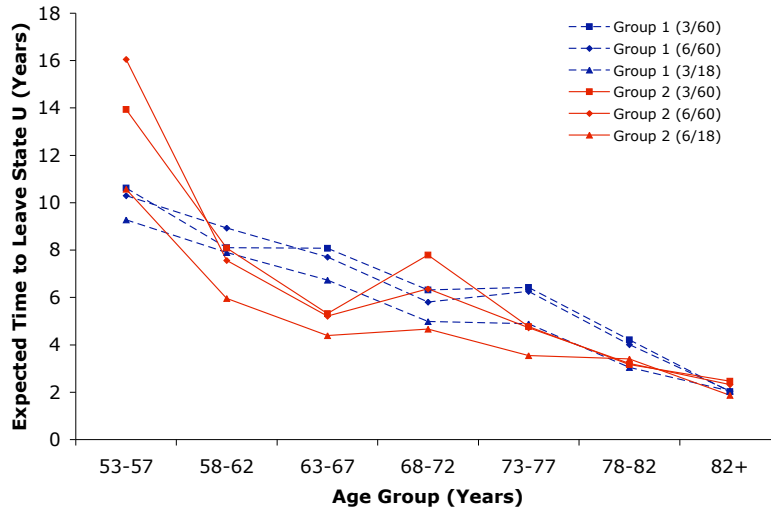Figure 4.9: Probability of Developing Unilateral versus Bilateral Cataract.

(a) Expected Time to Remain in Healthy State.

(b) Expected Time to Remain in Unilateral State.

Figure 4.10: Comparison of Expected Transition Times.

4.6   Discussion

The RAAB survey provides an important breakthrough in survey methodology that provides a great deal of new data on cataract prevalence in Africa. It is standardized, feasible and much less expensive than the longitudinal studies historically used to measure incidence in developed countries. Data from seven surveyed districts already suggests significant geographic variations in the impact of cataract that is important to take into account when allocating resources to meet the ambitious but necessary goal of eliminating blindness due to cataract in Africa. The incidence estimation method developed in our research fills a key role in the analysis of these data, because it allows the incidence estimates needed by policymakers to be generated from RAAB and mortality data alone. Further work collecting additional data and improving data collection and analysis methods can build on our research and continue to assist efforts to send ophthalmological teams where they are needed most.

 Because of the importance of our work, it is important to thoroughly discuss the assumptions and limitations of our model. Due to the limited data available, our model can only estimate, rather than calculate, incidence. Our model makes several key assumptions in order to estimate incidence:

1. People in each district form a closed system (immigration and emigration are neglected).

2. Prevalence and mortality rates are time independent (though they are age dependent).

3. Mortality and disease development follow an exponential distribution.

4. Mortality is the same for patients with unilateral and bilateral cataract.

5. The ratio of diseased to healthy mortality is the same for all age groups and geographic regions.

6. Because visual acuity levels at the time of past surgeries are unknown, all people with previous cataract surgery are counted as having "cataract" at whatever visual acuity level is relevant at the date of the surgery.

Many of the above assumptions are entirely appropriate in the case of cataract or reflect unavoidable limitations due to the lack of available data. For example, since we only know prevalence at one point in time, we must assume that the district forms a closed, steady-state system in order to use age as a time-like variable. Fortunately, the steady-state assumption is reasonable as cataract incidence depends on genetic or environmental factors rather than transmission, and these factors are unlikely to vary enough to affect cataract within the 35-year age span of the study. Similarly, the assumption that mortality and disease development follow exponential distributions is common in the literature, and we have no evidence that cataract does not follow this familiar pattern.

However, the assumption that each district forms a closed system bears further investigation. Our model assumes that the elderly population of each district remains fixed, with no significant immigration or emigration among people 50 years old or older over the 35-year span preceding the study. In many parts of Africa, the elderly do indeed have limited mobility and this assumption is valid. However, our model should not be applied to districts where turmoil has caused a great deal of recent immigration or emigration. It is not clear how to estimate the extent of this effect since survey participants do not give information about their immigration history.

Ophthalmologists and survey teams in Africa recognize the danger of having separate survey teams collect data in different regions, and have spent a great deal of time and energy working to standardize the process. Standardization workshops are included as part of the training of survey teams, as well as explicit warnings about the importance of certain survey procedures (for example, trying to re-contact people who are not home) for good data collection. In addition, the cataract categories (both clouded lens and aphakia from a previous surgery) diagnosed were coarse enough that it is likely that different ophthalmologists would agree on almost all classifications (Lewallen, 2010). The survey data we used was mainly based on medical diagnoses rather than verbal questions that could be more culturally sensitive in some areas than in others. It is impossible to quantify the possible error from differences in survey teams and comparisons between countries should be regarded as preliminary. However, a critique of RAAB survey methodology is beyond the scope of our research. We hope that as survey and standardization practices continue to improve, the importance of this source of error will diminish. Therefore, we have focused our analysis on sources of error introduced by the model itself.

It is also important to consider error arising from mortality rates. Healthy mortality rates are drawn from a World Health Organization database. How-

ever, diseased mortality rates were based on a parameter estimate of 1.5 for the ratio of diseased to healthy mortality. Our sensitivity analysis in Section 4.4.1 begins to address this issue, and suggests that mortality ratio estimates are unlikely to be responsible for the geographic variation in incidence we observed. However, the mortality ratio used certainly does affect incidence estimates across the board in a way that has practical implications for resource allocation. Future work should include further investigation into this issue.

In spite of the limitations, our results already provide valuable information on cataract incidence that can aid policy decisions and guide future research. A specific geographic grouping of countries into regions of similar incidence is not yet possible, especially in the case of The Gambia where the sample size was not large enough to clearly distinguish the district from any other district. However, the difference in incidence between Eastern African countries (Kenya, Rwanda and Tanzania) and other African countries (Eritrea and Mali) is of great practical significance and unlikely to be accounted for by variations due to RAAB sample size, inaccurate estimates of the mortality ratio parameter, or lack of standardization between survey teams. This suggests that a more systematic analysis of geographic patterns as more RAAB data becomes available may be extremely fruitful.

Our separation of unilateral and bilateral incidence also sheds light on the development of cataract disease itself. Our comparison of unilateral and bilateral incidence suggests that while unilateral incidence does vary between districts, the transition rate from unilateral to bilateral cataract varies much less. This could indicate that while the incidence of cataract disease depends on genetic or environmental factors, disease progression does not depend as much on these factors. As additional RAAB data becomes available, it will be interesting to see whether this pattern persists. In answering this question, it will also be important to continue sensitivity analysis to the mortality ratio parameter. In particular, it will be necessary to test whether the surprising lack of consistent patterns in bilateral incidence persists if the mortality rates for people with unilateral and bilateral cataract are allowed to vary with respect to each other, rather than being an artifact of our assumption that these mortality rates are equal. A more sophisticated model for mortality ratios, for example allowing mortality ratio to vary with age, would also be an important goal for future work.

Chapter 5

Estimating Incidence of a n-stage Progressive Disease

Some features of our incidence estimation model for bilateral diseases suggest that the model could easily be extended to an irreversible, n-stage progressive disease. In the present chapter, we exploit these similarities to generalize our two-stage disease model to a more robust model for any n-stage progressive disease.

In Section 2.2.2, we extended Podgor and Leske's two equations for disease progression to three equations that modeled the addition of a disease state. This process can be extended arbitrarily to an n-stage disease, and we do so explicitly here. Furthermore, we already noted in Section 2.2.3 that transition probabilities may be grouped by the number of stages traversed in a time step. For example, the expressions for P_{HH}, P_{UU}, and P_{BB} were extremely similar and only differed by a substitution of the relevant transfer and mortality rates. The two one-stage transition probabilities, P_{HU} and P_{UB} were also similar. Viewed in this way, we have already derived the zero-, one- and two-stage transition probabilities needed for any n-stage disease progression. We now formalize this generalization and extend it to probabilities for arbitrary transitions through m successive disease stages.

Our generalized model applies to irreversible, n-stage progressive diseases, where a person at stage j of the disease can only move forward to stage $j + 1$ or die; a person can never move backward and must pass at least briefly through all disease stages even if not all stages are observed. That is, if a person is observed to be in disease stage 1 at the beginning of a time interval and in stage 3 at the end of the time interval, the person is assumed to have passed through stage 2 for at least a short portion of that

interval. This could be represented by the flow chart shown in Figure 5.1 with healthy state 0, diseased states 1 through *n*, and an allowed death state *D*.
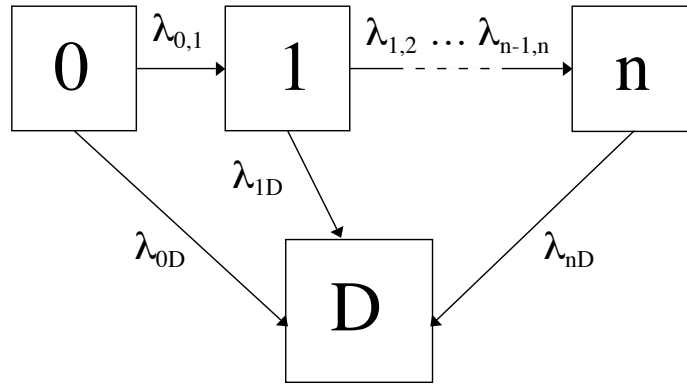

Figure 5.1: Generalized Model of Disease Progression.

Here, the allowed transitions leaving each state j are to state $j + 1$ and D, both with exponential distributions characterized by parameters $\lambda_{j,j+1}$ and $\lambda_{j,D}$, respectively. Since state n is the final disease state, there is no allowed transition from state n to $n + 1$; we write $\lambda_{n,n+1} = 0$ to simplify the formulas that follow. The model allows mortality rates $\lambda_{0,D}, \ldots, \lambda_{n,D}$ for different disease stages to be set separately, which is very desirable in many diseases such as cancer where progression to a new stage implies a new mortality rate. Given this information and the prevalence values $\pi_1, \ldots, \pi_n$ for each diseased state, our goal is to compute the n unknown transition rates $\lambda_{0,1}, \ldots, \lambda_{n-1,n}$. These can later be converted to incidence rates $I_{j,j+1}$, if desired, by taking

$$I_t^{j,j+1} = \pi_t^j \cdot \lambda_{j,j+1}. \tag{5.1}$$

That is, incidence is the conditional transfer rate $\lambda_{j,j+1}$ given that a person is in state j, times the probability π_t^j that the person is in state j at time t.

As in the two-stage disease case, we can find $n + 1$ equations representing the conservation of people as they progress through different disease stages. As before, $\pi_t^1 \ldots \pi_t^n$ represent the prevalence of all diseased stages at time $t = 0, 1$ and N_t is used to represent the number of people at time t ($t = 0$ denotes the beginning of the timestep whereas $t = 1$ denotes the end). Equation 5.2 represents the possible ways people could arrive in state j at time $t = 1$. Since people are allowed to move forward through arbitrarily many disease stages in a time interval, the $N_1 \pi_1^j$ people who end up in state j could have started the time interval in any state i satisfying $0 \leq i \leq j$. $N_0 \pi_0^i$ gives the number of people in state i at the beginning of the time period, and $P_{i,j}$ gives the probability (which we subsequently compute) that these people will end up in state j at the end of the time period. Summing over all allowed i, we obtain

$$N_1 \pi_1^j = \sum_{i=0}^{j} N_0 \pi_0^i P_{i,j}, \qquad 1 \leq j \leq n. \tag{5.2}$$

Note that for consistency we have introduced π_t^0, denoting the "healthy prevalence" and given by

$$\pi_t^0 = 1 - \sum_{i=1}^{n} \pi_t^i. \tag{5.3}$$

This healthy prevalence condition represents the requirement that all people not in any disease stage are considered healthy for purposes of our model. In other words, all of the people observed alive at each time point must be in one of the $n + 1$ living stages of our model, so the prevalences of all $n + 1$ stages counting the healthy stage must sum to 1.

To find transition probabilities $P_{j,j+m}$ from stage j to $j + m$ in one time step, we first consider the pdf for transition times $T_{j+1}, \ldots T_{j+m}, T_D$. The differences between consecutive transition times are independent, exponentially distributed random events, with distribution

$$f_{T_{k+1} - T_k}(t_{k+1} - t_k) = \frac{\lambda_{k,k+1}}{\lambda_{k,k+1} + \lambda_{k,D}} (\lambda_{k,k+1} + \lambda_{k,D}) e^{-(\lambda_{k,k+1} + \lambda_{k,D})(t_{k+1} - t_k)},$$
$$\tag{5.4}$$

because the time spent in state k before leaving (either through death or by progressing to state $k + 1$) is exponentially distributed with parameter $(\lambda_{k,k+1} + \lambda_{k,D})$. Also, the conditional probability that the person leaves

state k because of disease progression (rather than death) is given by $\frac{\lambda_{k,k+1}}{\lambda_{k,k+1}+\lambda_{k,D}}$.
Simplifying, we obtain

$$f_{T_{k+1}-T_k}(t_{k+1}-t_k) = \lambda_{k,k+1}e^{-(\lambda_{k,k+1}+\lambda_{k,D})(t_{k+1}-t_k)}, \tag{5.5}$$

for $j \le k < j+m$. The final time difference, $T_D - T_{j+m}$, differs slightly from
the others because in order to represent a person who stays in state $j+m$,
we consider both death and disease progression as exits from this state and
therefore do not include the factor of $\frac{\lambda_{j+m,j+m+1}}{\lambda_{j+m,j+m+1}+\lambda_{j+m,D}}$. The pdf for this term
is

$$f_{T_D-T_{j+m}(t_D-t_{j+m})} = \left(\lambda_{j+m,j+m+1} + \lambda_{j+m,D}\right) e^{-(\lambda_{j+m,j+m+1}+\lambda_{j+m,D})(t_D-t_{j+m})}. \tag{5.6}$$

Multiplying these independent terms, the overall joint pdf is

$$f_{T_{j+1},T_{j+2},\dots,T_{j+m},T_D}(t_{j+1},t_{j+2},\dots,t_{j+m},t_D) =$$
$$\left[\prod_{k=j}^{j+m-1} \lambda_{k,k+1}e^{-(\lambda_{k,k+1}+\lambda_{k,D})(t_{k+1}-t_k)}\right]$$
$$\cdot \left(\lambda_{j+m,j+m+1} + \lambda_{j+m,D}\right) e^{-(\lambda_{j+m,j+m+1}+\lambda_{j+m,D})(t_D-t_{j+m})}. \tag{5.7}$$

The transition probability from the jth stage to the $j+m$th stage is then
the probability that $T_{j+1}\dots T_{j+m}$ occur sequentially between times 0 and 1,
but that T_D of leaving the $j+m$th state occurs after time 1. That is,

$$P_{j,j+m} = P(0 \le T_{i+1} \le T_{i+2} \le \dots \le T_{i+m} \le 1 \le T_D). \tag{5.8}$$

This probability corresponds an integral of the pdf:

$$P_{j,j+m} = \int_0^1 dt_{i+1} \int_{t_{i+1}}^1 dt_{i+2} \cdots \int_{t_{i+m-1}}^1 dt_{i+m} \int_1^\infty dt_D$$
$$f_{T_{j+1},T_{j+2},\dots,T_{j+m},T_D}(t_{j+1},t_{j+2},\dots,t_{j+m},t_D) \tag{5.9}$$
$$P_{j,j+m} = \int_0^1 dt_{i+1} \int_{t_{i+1}}^1 dt_{i+2} \cdots \int_{t_{i+m-1}}^1 dt_{i+m} \int_1^\infty dt_D$$
$$\left[\prod_{k=j}^{j+m-1} \lambda_{k,k+1}e^{-(\lambda_{k,k+1}+\lambda_{k,D})(t_{k+1}-t_k)}\right]$$
$$\cdot \left(\lambda_{j+m,j+m+1} + \lambda_{j+m,D}\right) e^{-(\lambda_{j+m,j+m+1}+\lambda_{j+m,D})(t_D-t_{j+m})}. \tag{5.10}$$

Notice that, as in the case of our two-stage model, we have a set of $n+1$ equations in $n+1$ unknowns (n desired λ values plus the ratio N_1/N_0, which we eliminate). The dependence of these equations on $\lambda_{0,1}, \ldots, \lambda_{n-1,n}$ is via probabilities $P_{j,j+m}$. All of these integrals can be evaluated in closed form because they consist of constants times exponential functions. The bounds on each integral do depend on the subsequently integrated variable, which adds some complexity at each step. However, neither the integration steps nor the evaluation of bounds changes the integrand from its current form of an exponential function of a linear combination of variables, which is easily integrable in closed form. The $n+1$ equations found in this way are by no means linear, and before solving them one would need to check numerically that they have a unique solution within the feasible range of interest. Confidence intervals could still be computed by the simulation method described in Chapter 3, again generating all prevalence values at initial and final times from two multinomial distributions, and using the n-stage incidence computation method just discussed to simulate all trials.

For n above 3 or 4, the time to numerically solve the system of equations generated by our model may be prohibitive. Some approximations might be helpful to simplify computations. For example, in some diseases it may be very unlikely that a person would progress from the healthy stage to stage n of the disease in a single time interval (likely $1-5$ years). In this case, some of the most computationally intensive transition probabilities could be eliminated by setting a maximum number of states through which a person could progress in a single time step. For example, if the cutoff were a maximum of 2 states, only transition probabilities of the form $P_{j,j}$, $P_{j,j+1}$ and $P_{j,j+2}$ would be used, with the remaining probabilities set to zero. This would ensure that each model equation had a maximum of three terms, simplifying the computation.

Chapter 6

Conclusions

Inspired by the need in sub-Saharan Africa for cataract incidence estimation, we have developed an incidence estimation method based on age-specific prevalence for any n-stage disease with differential mortality. Our method includes a numerical simulation method to compute confidence intervals for incidence estimates based on random error in prevalence. We explicitly solved for incidence and implemented our solution numerically in the case of bilateral diseases, and applied the bilateral solution to RAAB survey data to generate incidence estimates for unilateral and bilateral cataract that can be compared across Africa. We hope that the incidence data we provide provides a useful first look at variations in cataract incidence across sub-Saharan Africa and the different resources that will be needed to achieve the goal of eliminating blindness and visual impairment due to cataract in Africa. More importantly, we regard this work as a proof of concept that incidence estimates can be generated at all based on data from one-time, feasible RAAB surveys.

Most assumptions of the bilateral disease model apply to our generalized model of n-stage progressive diseases as well. In both cases, the disease is assumed to be irreversible and progress at least briefly through all disease stages in order. In the case of cataract, it is clear that no person with bilateral cataract will again develop unilateral cataract. However, many diseases do not show this straightforward progression. Importantly, our generalized method also assumes that the group surveyed is representative of a closed system, with no immigration or emigration. The model should not be applied to regions in political or economic turmoil or to any area where people are mobile on during the time period of interest, usually at least the previous 30 years. This assumption becomes more problematic

when we consider applying our model beyond our original population of elderly persons in developing countries.

However, several of the above assumptions invite future research. In the case of cataract in Africa, previously operated eyes were counted as having developed cataract on the date of the surgery. However, more data is available through RAAB surveys related to these eyes, for example, time of surgery in each eye. This data was collected in a censored way and was difficult to include in our model. Perhaps other statistical techniques could be used to include this data. Survival analysis techniques may be especially suited to the inclusion of censored data without biasing the remaining analysis (See, for example, Kleinbaum and Klein, 2005; Tableman and Kim, 2004). This approach could also be useful beyond cataract as many data collection methods could result in additional, censored data beyond age-specific prevalence.

Future investigation could also reevaluate assumptions we made regarding diseased mortality. So far, we have data only for healthy mortality and made an assumption that mortality rates with unilateral and bilateral cataract are equal and 1.5 times the healthy rate. This was an important first step in considering differential mortality, but potentially misses certain subtleties. Perhaps unilateral and bilateral cataract cause different mortality rates. Alternately, it seems likely that the ratio of diseased to healthy mortality changes with age. Our incidence estimation code could be used analyze in greater detail the sensitivity of our results to differential mortality assumptions. New models for mortality may need to be developed, especially in the case of multiple-stage diseases where the differential mortality between stages is of interest, or where the mortality ratio between disease stages is likely to be influenced by a third variable, such as age.

In spite of the assumptions required, the model developed here is quite general, describing a disease that progresses through distinct phases with age dependence and differential mortality. Very little in the model is specific to cataract. Our methodology could likely be applied to other diseases in the developing world or anywhere incidence data is difficult to obtain directly. In particular, most of the other incidence estimation work in the literature is related to modeling of HIV/AIDS in Africa, so perhaps this work can contribute to the discourse on AIDS modeling.

Disease modeling in developing countries is not easy. We have tremendous respect for the ophthalmologists and RAAB survey teams on the ground in Africa. Without their daily perseverance, no amount of theoretical work in this area would be meaningful. We recognize the major frustrations involved in working, with extremely limited resources and funding, to meet

the challenge of eliminating blindness due to cataract in Africa. Yet these same teams, impelled by a sense of human dignity, often seek to go beyond the removal of cataracts at a certain advanced threshold of blindness. They see the need to provide higher quality surgeries, perform surgery earlier in the development of cataract and thus reduce the risk of surgery, and provide access to surgery in more remote regions outside the districts of currently practicing ophthalmologists. In this context, it is heartening to us that ophthalmic teams have begun to receive theoretical support in the form of RAAB methodology development and analysis. We are excited that, from halfway around the world in Claremont, California, we have been able to contribute to their efforts. We are glad to have found a research area that both is intellectually rewarding and has social impact, and have high hopes for future collaborative efforts of this kind.

Bibliography

Brunet, R. C. 2002. Convenient links between time varying incidence rates and current status information for epidemiological models with heterogeneity. *Journal of Biological Systems* 10(2):85–96.

Brunet, R.C., and C.J. Struchiner. 1999. A non-parametric method for the reconstruction of age- and time-dependent incidence from the prevalence data of irreversible diseases with differential mortality. *Theoretical Population Biology* 56:76–90.

Habiyakire, C., G. Kabona, P. Courtright, and S. Lewallen. 2010. Rapid assessment of avoidable blindness and cataract surgical services in Kilimanjaro region, Tanzania. In press.

Hallot, T. B., B. Zaba, J. Todd, B. Lopman, M. Wambura, S. Biraro, S. Gregson, and J. T. Boerma. 2008. Estimating incidence from prevalence in generalised HIV epidemics: Methods and validation. *PLoS Medicine* 5(4):611–622.

Keiding, N. 1991. Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society, Series A* 154(3):371–412.

Kimani, K., W. Mathenge, M. Shiela, et al. 2008. Cataract surgical services, outcome and barriers in Kericho, Bureti and Bomet Districts, Kenya. *East Africa Journal of Ophthalmology* 13:36–41.

Kleinbaum, David G., and Mitchel Klein. 2005. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health, New York: Springer.

Lagakos, S.W. 1976. A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometrics* 32(3):551–559.

Lewallen, Susan. 2010. Personal Communication.

Lewallen, Susan, Talithia Williams, Alyssa Dray, Brian Stock, Wanjiku Mathenge, Joseph Oye, John Nkurikiye, Kahaki Kimani, Andreas Muller, and Paul Courtright. 2010. Estimating incidence of vision-reducing cataract in Africa: A new model with implications for program targets. In press.

Limburg, H., and W. Meester. 2007. *Rapid Assessment of Avoidable Blindness*. International Centre for Eye Health, London School of Hygiene and Tropical Medicine, London, 4th ed.

Marschner, I. C. 1997. A method for assessing age–time disease incidence using serial prevalence data. *Biometrics* 53(4):1284–1398.

Mathenge, W., H. Kuper, H. Limburg, S. Polack, O. Onyango, G. Nyaga, and A. Foster. 2007a. Rapid assessment of avoidable blindness in Nakuru District, Kenya. *Ophthalmology* 114(3):599–605.

Mathenge, W., J. Nkurikiye, H. Limburg, and H. Kuper. 2007b. Rapid assessment of avoidable blindness in Western Rwanda: Blindness in a postconflict setting. *PLoS Medicine* 4(7):e217.

Mueller, Andreas, et al. 2009. RAAB survey data, Eritrea. Unpublished data obtained from author.

Oye, Joseph, et al. 2009a. RAAB survey data, Koulikor, Mali. Unpublished data obtained from author.

———. 2009b. RAAB survey data, The Gambia. Unpublished data obtained from author.

Podgor, Marvin J., and M. Cristina Leske. 1986. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in Medicine* 5:573–578.

Podgor, M.J., M.C. Leske, and F. Ederer. 1983. Incidence estimates for lens changes, macular changes, open-angle glaucoma and diabetic retinopathy. *American Journal of Epidemiology* 118(2):206–212.

Sakarovich, Charlotte, Ahmadou Alioum, Didier K. Ekouevi, Philippe Msellati, Valeriane Leroy, and Francois Dabis. 2007. Estimating incidence of HIV infection in childbearing African women using serial prevalence data from antenatal clinics. *Statistics in Medicine* 26(2):320–335.

Tableman, Mara, and Jong Sung Kim. 2004. *Survival Analysis Using S*. Texts in Statistical Science, New York: Chapman and Hall.

World Health Organization. 2001. *National Burden of Disease Studies: A Practical Guide*. Geneva, 2nd ed.

———. 2009a. URL `http://apps.who.int/whosis/database/life_tables/life_tables.cfm`. Accessed in November, 2009.

———. 2009b. Vision 2020 mission, goals, aims, and objectives. URL `http://www.v2020.org/page.asp?section=000100010009`. Accessed in October, 2009.