STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1035–1049
Published online 13 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2241

A discrete-state discrete-time model using indirect observation

Deanna J. M. Isaman[1,*,†], William H. Herman[2] and Morton B. Brown[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[2]*Department of Internal Medicine and Epidemiology, University of Michigan, Ann Arbor, MI 48109, U.S.A.*

SUMMARY

This research was motivated by a desire to model the progression of a chronic disease through various disease stages when data are not available to directly estimate all the transition parameters in the model. This is a common occurrence when time and expense make it infeasible to follow a single cohort to estimate all the transition parameters.

One difficulty of developing a model of chronic disease progression from such data is that the available studies often do not include the transitions of interest. For example, in our model of diabetic nephropathy, many clinical studies did not differentiate between patients without nephropathy and those who had microalbuminuria (a pre-clinical stage of nephropathy). Another difficulty was a lack of data to directly estimate parameters of interest. We consider models which can accommodate such difficulties.

In this paper we consider the problem of estimating parameters of a discrete-time Markov process when longitudinal data describing the entire process are not available. First, we present a likelihood approach to estimate parameters of a discrete-time Markov model. Next, we use simulation to investigate the finite-sample behaviour of our approach. Finally, we present two examples: a model of diabetic nephropathy and a model of cardiovascular disease in diabetes. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: grouped Markov chain; chronic disease; study-specific transition matrix

1. INTRODUCTION

This research was motivated by a desire to model the progression of a chronic disease and of its complications through phases (stages) of severity when data are unavailable to directly estimate transition parameters between all the stages. In particular, we were interested in developing a model of diabetes progression which would include diabetic complications such as nephropathy and cardiovascular disease. Such models are useful in many applications and can incorporate many facets of clinical interest including rates of progression, relationships

*Correspondence to: Deanna J. M. Isaman, School of Nursing, 400 North Ingalls, Room 4245, University of Michigan, Ann Arbor, MI 48109, U.S.A.
†E-mail: djmisaman@umich.edu

Contract/grant sponsor: Michigan Diabetes Research and Training Center; contract/grant number: NIH:P60-DK20572

Received 11 August 2004
Accepted 4 March 2005

between diseases and their complications, costs of treatment, quality of life, and benefits of intervention. However, these models may include many parameters, and it is difficult or expensive to collect data to estimate all the parameters of the processes.

In our model of diabetes, longitudinal data were not available to estimate the transitions between stages of progression for all of the complications of diabetes. This is common when time and expense make it infeasible to follow a cohort through all stages of a disease. Often, the only available data for disease modelling are provided by small clinical studies which focus on a single transition in the disease process.

Current approaches to disease models involve either (a) simplifying the model to one whose parameters can be estimated by a single study [1], (b) making distributional assumptions about unmeasured transitions of the model [2, 3] or (c) using multiple studies: one for each transition in the model [4–6].

Compiling estimates from various studies is beneficial in that it allows a model to be continually updated as new data become available. However, the studies that provide estimates may not estimate all the transitions of interest. For example, in our model of diabetic nephropathy, many clinical studies did not differentiate between patients without nephropathy and those who had microalbuminuria (a pre-clinical stage of nephropathy). Although estimates of these transitions cannot be used to directly estimate model parameters, these estimates which involve transitions from normal to microalbuminuria and from microalbuminuria to nephropathy may improve estimates of the parameters of interest when combined with estimates from other studies.

Another difficulty encountered when using the existing techniques is the amount of data available. Some transitions have no data available while other transitions are well-studied. It is occasionally possible to indirectly estimate transitions for which we have no direct estimates. In this paper we discuss requirements for estimability when direct estimates of transitions are not available. Conversely, when transitions have several estimates, it can be difficult to select a single best estimate to include in the model. Our method is also appropriate for pooling estimates from multiple studies.

We assume that data are available in the form of cumulative counts of transitions from some set of paths in the model. Since transition estimates are rarely available as a function of time, we do not consider these types of transitions. Another characteristic of our data is that it comes from studies with varying lengths of follow-up. Thus, our approach considers the design features of each study. These characteristics of the data drive the development of our model.

In this paper, we consider the problem of estimating parameters of a discrete-time Markov process when longitudinal data that describe the entire process are not available. First, we present a likelihood-based approach to estimate parameters of a discrete-time Markov model. Next, we use simulation to investigate the finite-sample behaviour of our approach. Finally, we present two examples: a model of diabetic nephropathy and a model of cardiovascular disease in diabetes.

2. THE MODEL

2.1. Defining the model

We begin by making the following assumptions:

1. the disease process operates as a first-order Markov chain,

2. the data are independent realizations generated from either the process, one of its subprocesses, or a grouping of the process,
3. the data collected from published medical studies are unbiased estimates of the parameters that they measure,
4. the data are homogeneous with respect to risk factors among subjects in a given state, and
5. the data are informative.

We also define the following nomenclature: let *direct data* denote data that provide an estimate of a parameter in our model. Transitions in the theoretical model are denoted as *primary transitions* and the transition probabilities for these primary transitions are the parameters of interest. *Indirect data* are estimates of transitions that are not explicitly stated in the model (i.e. not primary transitions), such as those that omit stages. Paths for which we have indirect data estimates will be called *augmentary transitions*; i.e. these are paths whose cumulative probabilities of progression are functions of more than one parameter in the model.

Let

(i, j) denote the path from node i to node j,
N denote the number of nodes in the theoretical model,
P denote the transition matrix for the theoretical model with elements $\{P\}_{ij} = \pi_{ij}$,
$\pi_{lij}(t_l)$ denote the cumulative transition probability for the lth study,
t_l denote the units of time observed in the lth study, and
x_{lij} denote the number of subjects in the lth study progressing from state i to j by time t_l.

Then, from our assumption of independent studies, the likelihood can be expressed as

$$\mathscr{L} \propto \prod_l \prod_i \prod_j \pi_{lij}(t_l)^{x_{lij}} \tag{1}$$

where $\sum_j \pi_{lij} = 1$.

Then we define P_l such that $\pi_{lij}(t) = \{P_l^t\}_{ij}$. Thus P_l generates the cumulative probability of transitions for the likelihood. Note that P_l depends on both the structure of P and the design of the clinical study. In general, P_l can be defined by rewriting P with appropriate absorbing states to represent the counting process of the clinical study. These absorbing states preclude counting subjects who progress along alternative paths (not included in the clinical study) and generate cumulative probabilities of progression without enumerating all possible paths beyond the study endpoint. Using this technique, which we call 'designed absorption', and exploiting the structural zeros specified in P_l greatly reduces the computational burden of evaluating the likelihood.

When the study design generates realizations from a grouped Markov chain (i.e. the study does not differentiate between several states), P_l must reflect the grouping. To construct P_l in this setting, we reindex the nodes of the grouped process via an 'onto' mapping $h(j) = i^*$, $i^* \in \{0, \ldots, m\}$, $m < N$ which takes the jth node of the ungrouped process and maps it to node i^* in the grouped process. Then, let $K_l(w)$, a $(m \times N)$ matrix, $m < N$, be defined as

$$\{K_l\}_{i^*,j}(w) = \begin{cases} w_j & \text{if } h(j) = i^* \\ 0 & \text{otherwise} \end{cases}$$

$$w_j = \frac{M_j}{\sum_{j \ni h(j)=i^*} M_j}$$

where M_j is the prevalence at state j. Also let $K_l(1)$ $(n \times m)$ be defined as $K_l(w)$ where $w_j = 1 \ \forall j$. Then $P_l^t = K_l(w)P^t K_l(1)$ in the likelihood. As an example, for one grouped node which, without loss of generality, combines nodes i–j, K^w would be

$$K_l(w) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ & \cdots & & & & \cdots & & \cdots & \\ 0 & 0 & 0 & \cdots & w_i & \cdots & w_j & \cdots & 0 \\ & \cdots & & & & \cdots & & \cdots & \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

and

$$P_l^1 = K_l(w)PK_l(1) = \begin{pmatrix} \pi_{11} & \cdots & \sum_{l=i}^{j} \pi_{1l} & \cdots & \pi_{1n} \\ \pi_{21} & \cdots & \sum_{l=i}^{j} \pi_{2l} & \cdots & \pi_{2n} \\ \cdots & & \cdots & & \cdots \\ \sum_{k=i}^{j} w_k \pi_{k1} & \cdots & \sum_{l=i}^{j} \sum_{k=i}^{j} w_k \pi_{kl} & \cdots & \sum_{k=i}^{j} w_k \pi_{kn} \\ \cdots & & \cdots & & \cdots \\ \pi_{n1} & \cdots & \sum_{l=i}^{j} \pi_{nl} & \cdots & \pi_{nn} \end{pmatrix}$$

2.2. Estimability

In general, MLEs for cumulative probabilities (e.g. $\pi_{lij}(t_l)$) will be estimable when the data are informative. However, we are interested in the estimability of the annual probabilities, π_{ij}. When data are not available for some primary transitions, model parameters may or may not be estimated from a combination of available primary and augmentary data. Thus to determine estimability of the annual probabilities we show the existence of an identifiable mapping between the cumulative and annual probabilities.

Let

s denote the number of primary paths in the model,
r denote the number of observations, $r \geqslant s$,
$e = (e_1 \ldots e_s)'$ denote the vector of primary edges in the graph indexed by a,
E_{ij} denote the set of edges in all paths connecting nodes i to j,
$\pi = (\pi_1 \ldots \pi_s)'$ denote a vector of primary transition probabilities corresponding to e,
$\tau = (\tau_1 \ldots \tau_s)'$ denote a vector of transition times for the edges corresponding to e,
$d = (d_1 \ldots d_r)'$ denote a $(r \times 1)$ vector $(r \geqslant s)$ of expected transition times for the data indexed by c,

B_a denote the set of indices of edges branching from the initial node of edge e_a,
W be an $(r \times s)$ matrix

$$w_{ca} = \begin{cases} 1 & \text{if } a \in E_{ij} \\ 0 & \text{otherwise} \end{cases}$$

$z = (z_1 \ldots z_s)'$ denote the vector of conditional probabilities indexed by a such that $z_a = \pi_a / (\sum_{i \in B_a} \pi_i)^2$, the conditional probability of traversing edge a from the initial node of edge e_a.

Lemma 1
If the z's defined above are estimable, then π is estimable.

Proof
Partition the z's into sets, $Z_q = \{z_j | j \in B_q\}$. Then, for each partition

$$\sum_{z \in Z_q} z = \sum_{a \in B_q} \left(\frac{\pi_a}{(\sum_{i \in B_q} \pi_i)^2} \right)$$

$$= \frac{1}{(\sum_{i \in B_q} \pi_i)^2}$$

or $(\sum_{i \in B_a} \pi_i)^2 = 1 / \sum_{z \in Z_q} z$. Substituting this result into the definition of z_a, we get

$$\pi_a = \frac{z_a}{(\sum_{z \in Z_q} z)}$$

for all π_a such that $z_a \in Z_q$. Since the choice of partition was arbitrary, all π's are estimable. □

Theorem 2
If $\text{rank}(W) = s$, then the transitions in the model are estimable.

Proof
Under our Markovian assumption, the expected time to transition for a single edge is

$$E[\tau_a] = \sum_{k=0}^{\infty} k \left(1 - \sum_{i \in B_a} \pi_i \right)^{k-1} \pi_a$$

$$= -\pi_a \sum_{k=0}^{\infty} \frac{\partial}{\partial \pi_a} \left(1 - \sum_{i \in B_a} \pi_i \right)^k$$

$$= -\pi_a \frac{\partial}{\partial \pi_a} \left(\frac{1}{\sum_{i \in B_a} \pi_i} \right)$$

$$= \frac{\pi_a}{(\sum_{i \in B_a} \pi_i)^2}$$

Then, the expected time to transition for d_c, the cth observation which traverses some path from i to j, is

$$\begin{aligned} d_c &= \sum_{a \in E_{ij}} E[\tau_a] \\ &= \sum_{a \in E_{ij}} \frac{\pi_a}{(\sum_{i \in B_a} \pi_i)^2} \\ &= \sum_{a \in E_{ij}} w_{ca} z_a \end{aligned}$$

Thus, $d = Wz$, and if $\text{rank}(W) = s$ the z's are estimable. Finally, by Lemma 1, estimable z's imply estimable π's. □

This suggests an algorithm for determining estimability: create a matrix, W, of paths traversed by the data; if W has rank s, the model is estimable from the data.

2.3. *Extensions for covariates*

A common occurrence in disease modelling is to have at least one clinical study which gives transition estimates based upon some stratification of covariates. Manton [2] presents a similar problem for a single study modelled in continuous time. We adapt his approach for use with multiple studies and discrete time by rewriting the likelihood from equation (1) as

$$\mathscr{L} \propto \prod_l \prod_i \prod_j \prod_s \pi_{lijs}(t_l)^{x_{lijs}}$$

where s indexes the strata for each transition, and (as before) l, i, j index the studies, starting points, and ending points, respectively. This notation implies a block-diagonal structure to the transition matrix, P. The transition matrix can be written as sub-matrices, P_s, for the level of each covariate $s = 1, \ldots, z$.

In many situations, we will have unstratified data to incorporate into the likelihood in addition to stratified transition data. In these situations, we can view the strata as a grouped node and use prevalence data to proportionally allocate grouped data into strata. However, unstratified data alone will give no information about the strata-specific probabilities. Estimation of strata-specific transition probabilities requires either stratified data or additional assumptions.

3. SIMULATIONS

To investigate the finite-sample properties of our estimators, we performed a number of simulations. We ran 1000 replications of a 3-node model (illustrated in Figure 1). In this figure we represent the primary transitions as solid arrows and the augmentary transitions as dashed arrows. The simulations vary the number of studies per transition and number of subjects per study. Each observation was generated by simulating the number of progressions among the subjects over a 5-year study. Table I displays the results of our first set of simulations. The table reports the number of studies for the paths $(0,1), (1,2)$, and $(0,2)$, respectively, as ratios

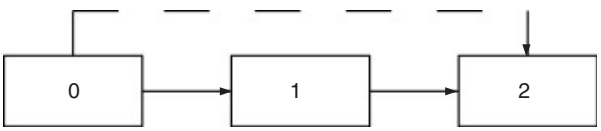

Figure 1. Model and data design for simulations.

Table I. Simulation results with 1000 replications of a 3-node model.

π_{01}	π_{12}	Features	Studies	Subjects	$\hat{\pi}_{01}$ (SE)	$\hat{\pi}_{12}$ (SE)
Varying number of studies						
0.1	0.2	No augmentary	1:1:0	500	0.1005 (0.0067)	0.2002 (0.0104)
0.1	0.2		1:1:1	500	0.1000 (0.0058)	0.2001 (0.0100)
0.1	0.2	Unbalanced	1:1:3	500	0.0999 (0.0052)	0.2007 (0.0994)
0.1	0.2	Unbalanced	1:1:9	500	0.0999 (0.0043)	0.2007 (0.0089)
0.1	0.2	No augmentary	3:3:0	500	0.1003 (0.0042)	0.2003 (0.0065)
0.1	0.2		3:3:3	500	0.1000 (0.0035)	0.2001 (0.0057)
0.1	0.2	Unbalanced	3:3:9	500	0.1001 (0.0030)	0.2001 (0.0053)
0.1	0.2	No primary 0 to 1	0:1:1	500	0.1004 (0.0136)	0.2004 (0.0104)
0.1	0.2	No primary 0 to 1	0:1:3	500	0.0996 (0.0082)	0.2011 (0.0100)
0.1	0.2	No primary 0 to 1	0:1:9	500	0.1003 (0.0062)	0.2004 (0.0104)
Varying number of subjects						
0.1	0.2		3:3:3	100	0.1001 (0.0076)	0.2007 (0.0124)
0.1	0.2		3:3:3	1000	0.1002 (0.0025)	0.2001 (0.0041)
Varying size of π_{01}						
0.2	0.2	No augmentary	3:3:0	500	0.2002 (0.0059)	0.2003 (0.0060)
0.2	0.2		3:3:3	500	0.2001 (0.0054)	0.2002 (0.0061)
0.3	0.2		3:3:3	500	0.3001 (0.0076)	0.2003 (0.0052)
0.5	0.2		3:3:3	500	0.5002 (0.0139)	0.2002 (0.0048)

(e.g. 1:1:0). In all cases, the distribution of the estimates, $\hat{\pi}_{01}$ and $\hat{\pi}_{12}$ were approximately normal.

The model (1:1:0) can be fitted by the current techniques for disease modelling. The addition of a single augmentary study (model 1:1:1) decreases the standard error of $\hat{\pi}_{01}$ by a relatively small amount. However, the use of 9 augmentary studies decreases the standard error by approximately 65 per cent. Comparing the model fit by existing techniques (model 1:1:0) to models which allow pooling of data (e.g. model 3:3:0 or 3:3:3), the use of multiple studies substantially decreases the standard error. In general, the standard error decreases proportional to the amount of primary data, and decreases more slowly with augmentary data.

Our simulations also demonstrate our technique in scenarios where the existing techniques cannot be used. When there are no primary data available for a transition, our method can estimate transition probabilities from augmentary data, at the cost of increasing the standard error. Comparing models 0:1:1 and 1:1:0 we see that replacing a primary study with an augmentary study doubles the standard error. However, the parameter is estimable. Note that it requires about nine augmentary studies to reduce the variance to a level comparable to using a single primary study.

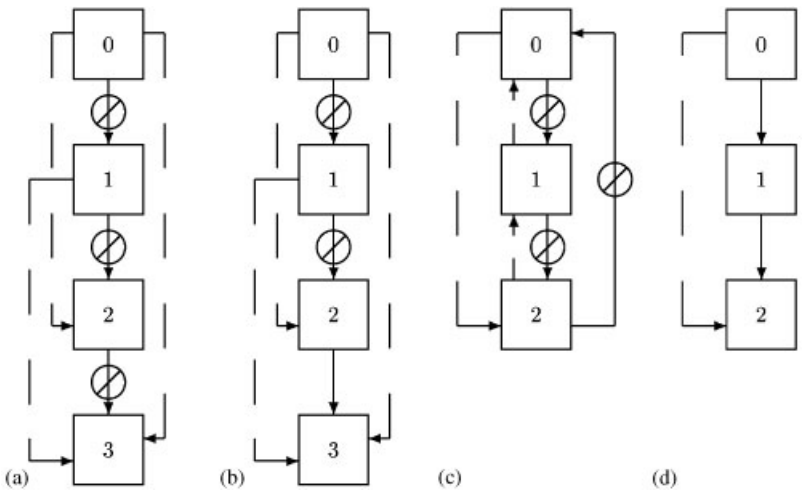

Figure 2. Model and data design for simulations.

Table II. Simulation results with 1000 replications of models displayed in Figure 2 with data from 5-year study periods.

Transition From–To	Value	Estimate	Simulated SE	Asymptotic SE	Studies/ Path	Subjects	Model (Figure 2)
0–1	0.1	0.100	0.013	0.012	3:3:3	500	A
1–2	0.2	0.204	0.035	0.036			
2–3	0.1	0.100	0.013	0.013			
0–1	0.1	0.101	0.012	0.011	3:3:3:1	500	B
1–2	0.2	0.201	0.023	0.020			
2–3	0.1	0.100	0.007	0.007			
0–1	0.1	0.100	0.019	0.020	1:1:1	200	C
1–2	0.2	0.208	0.049	0.060			
2–0	0.1	0.101	0.019	0.015			
0–1	0.1	0.101	0.013	0.021	1:1:1	50	D
1–2	0.2	0.202	0.031	0.030			
0–1	0.1	0.101	0.021	0.022	2:2:2	20	D
1–2	0.2	0.202	0.036	0.038			
0–1	0.1	0.102	0.030	0.024	1:1:1	20	D
1–2	0.2	0.205	0.059	0.063			

To investigate our method's performance in more difficult scenarios and small sample sizes, we performed additional simulations. Figure 2 illustrates several simulated models and Table II provides details about each simulation. For each model we present a variance estimate based on likelihood theory and a comparable simulated variance based on 1000 replications. The first three simulations have limited data for the primary edges. Models A and B contrast models with and without a single study measuring a primary edge; model C illustrates the

estimability of a model where no primary data are available, but augmentary data are collected as subjects progress two stages in the loop. Finally, we present three simulations with model D where we decrease the number of observations and subjects to investigate performance with very small sample sizes.

Similar to the results of the first simulations, the results in Table II show agreement between point estimates and their theoretical values. Also, the simulated variances are close to the variance estimates derived from likelihood theory. In particular, the last lines of Table II display results from a model with 3 nodes, 2 parameters, 3 observations and 20 subjects per observation. Even in this sparse simulation, the difference between standard errors from the simulation and asymptotic results was not more than 0.006.

4. APPLICATIONS

This research was motivated by a model of progression and complication in diabetes mellitus. Because of the time, expense, and difficulty involved in conducting a longitudinal study of diabetes progression, data for such disease-progression models typically come from many small studies, each of which estimates parameters of one or two steps in the process [4, 6]. Since these smaller studies were not designed to investigate our theoretical model, they do not necessarily provide direct estimates of our disease model. For example, we desired to use a study of diabetes progression that did not differentiate between healthy patients and patients with microalbuminuria (a pre-clinical stage of proteinuria).

The foundations for the theoretical model were chosen in collaboration with clinical investigators, and the data were extracted from the medical literature by a clinical researcher after an extensive literature review. This researcher provided us with the best available literature providing primary and augmentary estimates for the model and then providing us with the 'best' estimate for each of the primary transitions as per the standard approach. In the following sections, we use the available literature to compute estimates of disease progression, and we compare those results to estimates using the standard approach.

4.1. A discrete-time model of diabetic nephropathy

Our theoretical model for progression of diabetic nephropathy included 6 stages: normal (no nephropathy), microalbuminuria, proteinuria, end-stage renal disease (ESRD) with dialysis, ESRD with transplant, and death due to ESRD (denoted as states 0–5, respectively). Figure 3 illustrates the states and transitions. The corresponding transition matrix for probability of transitions in one year is

$$P = \begin{pmatrix} 1-\pi_{01}-\pi_{02} & \pi_{01} & \pi_{02} & 0 & 0 & 0 \\ 0 & 1-\pi_{12} & \pi_{12} & 0 & 0 & 0 \\ 0 & 0 & 1-\pi_{23} & \pi_{23} & 0 & 0 \\ 0 & 0 & 0 & 1-\pi_{34}-\pi_{35} & \pi_{34} & \pi_{35} \\ 0 & 0 & 0 & 0 & 1-\pi_{45} & \pi_{45} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \text{Normal} \\ \text{Microalb.} \\ \text{Proteinur.} \\ \text{Dialysis} \\ \text{Transpl.} \\ \text{Death} \end{matrix}$$

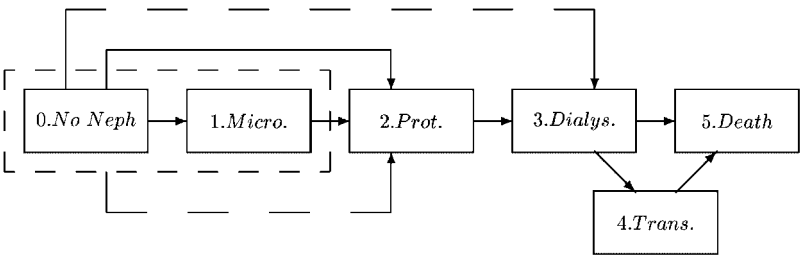

Figure 3. A discrete-time model of nephropathy.

Table III. Medical data for a discrete-time model of diabetic nephropathy.

Transition	N	Incident count	Years	Annualized	Reference
0–1	79	15	6	0.034	Ravid [7]
0–1	90	31	9	0.046	Forsblom [8]
{0, 1}–2	398	69	4	0.046	Klein [9]
0–2	176	36	6	0.037	Gall [10]
1–2	49	26	6	0.118	Tanaka [11]
1–2	45	19	5	0.104	Ravid [12]
2–3	202	16	5	0.016	Humphrey [13]
3–4	1000	40	1	0.04	USRDS [14]
3–5	231	144	3	0.278	Koch [15]
3–5	11 929	5356	5	0.112	Byrne [16]
4–5	23	4	5	0.037	Meigham [17]

Table IV. Comparative solutions for diabetic nephropathy model with and without pooling data.

Transition	State names	Single estimate	SE	Pooled estimate	SE
0–1	Normal to microalbuminuria	0.05	0.0088	0.046	0.0067
0–2	Normal to proteinuria	0.03	0.0061	0.026	0.0060
1–2	Microalbuminuria to proteinuria	0.10	0.0228	0.091	0.0105
2–3	Proteinuria to dialysis (ESRD)	0.01	0.0041	0.017	0.0039
3–4	Dialysis to transplant	0.04	0.0062	0.039	0.0061
3–5	Dialysis to death	0.11	0.0015	0.119	0.0030
4–5	Transplant to death	0.04	0.0184	0.045	0.0215

One clinical assumption was to include a transition from no nephropathy to proteinuria which bypasses microalbuminuria. Our interpretation of this parameter is patients' progression from normal through microalbuminuria to proteinuria within a single year. Setting $\pi_{02} > 0$ allows us to model these rapid progressions.

Table III displays the data extracted from the medical literature for use in our model of diabetic nephropathy. Grouped nodes are indicated by braces. The results are shown in Table IV for both the traditional 'single estimate' method and our 'pooled' method. Note that

Table V. Estimated correlation matrix for a model of diabetic nephropathy.

$\hat{\pi}_{01}$	1						
$\hat{\pi}_{02}$	−0.014	1					
$\hat{\pi}_{12}$	−0.046	−0.444	1				
$\hat{\pi}_{23}$	0.001	0.0003	−0.002	1			
$\hat{\pi}_{34}$	0	0	0	0	1		
$\hat{\pi}_{35}$	0	0	0	0	0.312	1	
$\hat{\pi}_{45}$	0	0	0	0	0.064	−0.761	1

the standard errors for the pooled estimates are generally on the same order as the single estimates. This is not surprising given the limited amount of additional or augmentary data. However, in cases like the estimate between microalbuminuria and proteinuria, the standard error is reduced 50 per cent as the sample size doubles by adding a second study and augmentary data. In contrast, the standard error for the transition between dialysis to death is larger for the pooled estimate than for the single estimate. This can be explained by examining the annualized estimates displayed in Table III. The two studies providing estimates for the progression between dialysis and death [15, 16] suggest very different estimates. Moreover, if Koch's estimate had been selected by our clinical expert as the best single estimate, the standard error for reference would be 0.02. Because Koch's sample size is substantially lower than Byrne's the pooled point estimate reflects Byrne's estimate; however, the variance is based on the mean square error which is increased due to the large difference between the estimates in the two studies.

Table V displays the correlation matrix for our estimates of transition probabilities. The block zero pattern in the lower left corner is due to the model design and availability of data. As Figure 3 illustrates, there are no primary paths which bypass state 3 (ESRD with dialysis) and there are no augmentary data which include dialysis as an intermediary state. Thus, our estimates of transitions between states 3, 4, and 5 are independent of transitions between states 0, 1, 2, and 3. When additional augmentary data are included which has stage 3 as an intermediary state (e.g. transition from proteinuria to death), this block diagonal pattern is lost.

4.2. A discrete-time model of cardiovascular disease in diabetes

We also applied our method to a model of cardiovascular disease in diabetes. The states were ordered from 0 to 4, respectively, as normal (no CVD), angina, MI, history of MI, and death due to CVD. The transition matrix was defined as

$$P = \begin{pmatrix} 1-\pi_{01}-\pi_{02} & \pi_{01} & \pi_{02} & 0 & 0 \\ 0 & 1-\pi_{12}-\pi_{14} & \pi_{12} & 0 & \pi_{14} \\ 0 & 0 & 0 & \pi_{23} & 1-\pi_{23} \\ 0 & 0 & \pi_{32} & 1-\pi_{32}-\pi_{34} & \pi_{34} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \text{Normal} \\ \text{Angina} \\ \text{MI} \\ \text{History} \\ \text{Death} \end{matrix}$$

This model defines a myocardial infarction (MI) as an event such that patients pass through MI and either die or enter a state called 'History of MI'. Because of our assumption of discrete

Table VI. Medical data for discrete-time CVD model.

Transition	N	Incident count	Years	Annualized	Reference
0–2	890	180	7	0.032	Haffner [18]
0–3	1138	101	10	0.009	UKPDS 33 [19]
0–4	1138	98	10	0.009	UKPDS 33 [19]
1–2	569	61	2	0.055	Malmberg [20]
1–4	569	53	2	0.048	Malmberg [20]
2–4	475	85	1	0.179	Miettinen [21]
3–2	73	31	5	0.179	Ulvenstam [22]
3–2	169	76	7	0.082	Haffner [18]
3–2	78	33	5	0.104	Ravid [12]
3–4	468	137	5	0.067	Lowel [23]

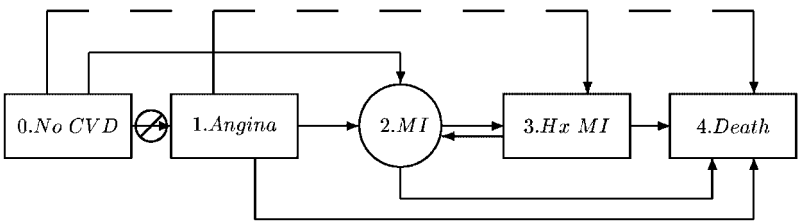

Figure 4. A discrete-time model of cardiovascular disease.

Table VII. Comparative solutions for CVD model with and without pooling data.

State transition	State names	Single estimate	SE	Pooled estimate	SE
0–1	Normal to angina			0.007	0.0009
0–2	Normal to MI	0.03	0.0021	0.026	0.0014
1–2	Angina to MI	0.06	0.0067	0.058	0.0071
1–4	Angina to death	0.05	0.0063	0.049	0.0066
2–4	MI to death	0.18	0.0177	0.271	0.0171
3–2	History of MI to MI	0.11	0.0180	0.097	0.0077
3–4	History of MI to death	0.06	0.0061	0.020	0.0046

time, this parameterization means that patients can return to the event MI (reinfarction) but multiple infarctions in one year are not allowed. Future work to incorporate a continuous-time model will avoid these simplifying assumptions.

The data extracted from the medical literature are presented in Table VI. Note that there are no data estimating the transition from normal to angina. This unobserved transition is estimable from augmentary data measuring transition from normal to history of MI and from normal to death. Also, in this model we use multiple observations from a single study. For example, the data from the UKPDS-33 was used to estimate progression from no CVD to incident MI (Figure 4).

Table VIII. Estimated correlation matrix for CVD model.

$\hat{\pi}_{01}$	1						
$\hat{\pi}_{02}$	−0.030	1					
$\hat{\pi}_{12}$	0.015	−0.045	1				
$\hat{\pi}_{14}$	−0.002	0.007	−0.056	1			
$\hat{\pi}_{24}$	−0.001	−0.021	0.001	−0.0001	1		
$\hat{\pi}_{32}$	−0.002	0.045	−0.002	0.0003	−0.235	1	
$\hat{\pi}_{34}$	−0.009	−0.018	−0.001	−0.0001	−0.383	0.518	1

The results of our method are reported and compared to single estimates in Table VII. The pooled estimates seem reasonable compared to the initial data. In the case of progression from history of MI to recurrent MI, the multiple studies available provide a reduction in the standard error of the estimate. The primary benefit of our technique, compared to the standard method of choosing a single best estimate, is our ability to estimate the transition probability from normal to angina. The correlation matrix for estimators is displayed in Table VIII.

5. DISCUSSION

The method described in this paper provides two advantages over the existing techniques: the use of multiple studies per transition and the use of augmentary data. By pooling data from multiple studies, we can reduce the variance of our estimates in the usual fashion. Using augmentary data provides several benefits. First, although there are methods for pooling estimates for a single transition in situations when all the studies are drawn from the same study design and the same definition of states, some well-designed studies do not define their population and measurements based on our model of interest. Using augmentary data, many more studies are available for use in our model. We saw from our simulations that augmentary data provide less information than primary data, but in some applications more augmentary data are available than primary data.

The second benefit of using augmentary data is the ability to estimate transition probabilities when no primary data are available. Some transitions are not well-measured. This can occur due to rapid progression through a state or due to the discovery of new states in the clinical model. Using augmentary data, the transition probabilities are estimable.

Although we have focused our work on Markov models where all parameters are estimable, our work can be viewed from a variety of other contexts. It is an extension of meta-analysis applied to multi-state models. Much research has focused on meta-analysis for observation of a single outcome [24]; however, no known work has considered meta-analysis in the context of non-iid observations (i.e. observations arising from differing study designs and different initial states). Our work can also be viewed within the framework of a missing information problem. We assumed that the parameters were estimable from the available data through direct evaluation of the likelihood. However, it would be possible to relax this assumption and use a technique such as the EM algorithm [25] to provide estimates of unknown transition probabilities from augmentary data.

Our method can be further generalized to accommodate a larger family of distributions (e.g. inhomogeneous processes) and dependent observations; however such generalizations are

computationally burdensome. Moreover, we rarely have data available to fit such sophisticated models. In a situation where more data are available, existing techniques would probably be adequate.

It is well-known that methods which glean data from published results can suffer from publication bias [26, 27] since significant findings are over-represented in the literature. However, in our application, many of the data are available from sources other than clinical research trials. For example, it is often possible to find data in a published registry or from large demographic studies such as the Framingham study or the UKPDS. Moreover, in the cases where we extract data from a clinical research trial, the data are generally extracted from the control arm of the trial. Thus, the effect of publication bias would tend to yield conservative estimates in our model.

There are several directions for future research in this area. First is the extension of our methods to continuous-time models. In our application of CVD, it was necessary to assume not more than one MI in a fixed time period (one year) due to the discrete-time constraint. A continuous-time model would allow us to posit multiple events in unit of time. Another direction for future research is finding a parsimonious expression for covariates. In the current expression, use of covariates is greatly limited by computational feasibility. Further research to find a parsimonious expression for covariates that is computationally simple (similar to our use of designed absorption to simplify the likelihood) would be useful. Finally, a random effects model to accommodate the extra variability between studies would be a contribution to this research. Again, this extension is limited by availability of data and computational feasibility.

In conclusion, we have introduced a theoretical framework for discrete-time models of disease progression and extensions of our work for covariates. We also presented designed absorption as a method for reducing the computational burden of evaluating the likelihood. Using this framework we presented simulations which suggested that our method was well-behaved in most settings. Finally, we presented two applications of our approach for discrete-time models which incorporated unobserved transitions and grouped nodes. Thus, in the context of chronic disease progression, our technique has been shown to be flexible and well-behaved, giving us a new approach to modelling discrete-time discrete-state processes.

ACKNOWLEDGEMENTS

We thank Dr Michael Brandle for his time and effort invested in developing the clinical model which motivated this research and extracting the estimates from the literature.

This study was supported by the Biostatistics Core of the Michigan Diabetes Research and Training Center (NIH:P60-DK20572).

REFERENCES

1. Barendregt JJ, Baan CA, Bonneux L. An indirect estimate of the incidence of non-insulin-dependent diabetes mellitus. *Epidemiology* 2000; **11**:274–279.
2. Manton KG, Lowrimore G, Yashin A. Methods for combining ancillary data in stochastic compartment models of cancer mortality: generalization of heterogeneity models. *Mathematical Population Studies* 1993; **4**:133–147.
3. Manton KG, Stallard E. A stochastic compartment model representation of chronic disease dependence: techniques for evaluating parameters of partially unobserved age inhomogeneous stochastic processes. *Theoretical Population Biology* 1980; **18**:57–75.
4. Brown JB, Russel A, Chan W, Pedula K, Aickin M. The global diabetes model: user friendly version 3.0. *Diabetes Research and Clinical Practice* 2000; **50**:S15–S46.

5. Eastman RC, Javitt JC, Herman WH, Dasbach EJ, Zbrozek AS, Dong F, Manninen D, Garfield SA, Copley-Merriman C, Maier W, Eastman J, Kotsanos J, Cowie CC, Harris M. Model of complications of NIDDM I. model construction and assumptions. *Diabetes Care* 1997; **20**:725–734.
6. Palmer AJ, Brandt A, Gozzoli V, Weiss C, Stock H, Wenzel H. Outline of a diabetes disease management model: principles and applications. *Diabetes Research and Clinical Practice* 2000; **50**:S47–S57.
7. Ravid M, Brosh D, Levi Z, Bar-Dayan Y, Ravid D, Rachmani R. Use of enalapril to attenuate decline in renal function in normotensive, normoalbuminuric patients with type 2 diabetes mellitus. *Annals of Internal Medicine* 1998; **128**:982–988.
8. Forsblom CM, Sane T, Groop P, Saloranta C, Ekstrand A, Groop L, Totterman KJ. Predictors of progression from normoalbuminuria to microalbuminuria in NIDDM. *Diabetes Care* 1998; **21**:1932–1938.
9. Klein R, Klein BE, Moss SE. Incidence of gross proteinuria in older-onset diabetes. *Diabetes* 1993; **42**:381–389.
10. Gall MA, Hougaard P, Borch-Johnson K, Parving HH. Risk factors for development of incipient and overt diabetic nephropathy in patients with non-insulin dependent diabetes mellitus. *British Medical Journal* 1997; **314**:783–788.
11. Tanaka Y, Onuma T, Atsum Y, Thojima T, Matsuoda K, Kawamori R. Role of glycemic control and blood pressure in the development and progression of nephropathy in elderly Japanese NIDDM patients. *Diabetes Care* 1998; **21**:116–120.
12. Ravid M, Savin H, Jutrin I, Bental T, Katz B, Lishner M. Long-term stabilizing effect of angiotension-converting enzyme inhibition on plasma creatinine and on proeinuria in normotensive type II diabetic patients. *Annals of Internal Medicine* 1993; **118**:557–581.
13. Humphrey LL, Ballard DJ, Frohnert PP, Chu CP, O'Fallon WM, Palumbo PJ. Chronic renal failure in non-insulin-dependent diabetes mellitus. *Annals of Internal Medicine* 1989; **111**:788–796.
14. USRDS United States Renal Data System. Incidence of reported ESRD: 2000 annual report, 2000.
15. Koch M, Kutkuhn B, Grabensee B, Ritz E. Apolipoprotein a, fibrinogen, age, and history of stroke are predictors of death in dialysed diabetic patients. *Nephrology Dialysis Transplantation* 1997; **12**:2603–2611.
16. Byrne C, Vernon P, Cohen JJ. Effect of age and diagnosis on survival of older patients beginning chronic dialysis. *Journal of the American Medical Association* 1994; **271**:34–45.
17. Meigham AV, Fonck C, Coosemans W, Vandeleene B, Vanrenterghem Y, Squifflet J, Pirson Y. Outcome of cadaver kidney transplantation in 23 patients with type 2 diabetes mellitus. *Nephrology Dialysis Transplantation* 2001; **16**:1686–1691.
18. Haffner SM, Lehto S, Ronnemaa T, Pyorala K, Laasko M. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. *New England Journal of Medicine* 1998; **339**:229–234.
19. UK Prospective Diabetes Study UKPDS Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes UKPDS 33. *Lancet* 1998; **352**:837–853.
20. Malmberg K, Yusuf S, Gerstein HC, Brown J, Zhao F, Hunt D, Piegas L, Calvin J, Keltai M, Budaj A. Impact of diabetes on long-term prognosis in patients with unstable angina and non-q-wave myocardioal infarction. *Circulation* 2000; **102**:1014–1019.
21. Miettinen H, Lehto S, Salomaa V, Mahonen M, Niemela M, Haffner SM, Pyorala K, Tuomilehto J. Impact of diabetes on mortality after the first myocardial infarction. *Diabetes Care* 1998; **21**:69–75.
22. Ulvenstam G, Aberg A, Bergstrand R, Johanssom S, Pennert K, Vedin A, Wilhelmsen L, Wilhelmsson C. Long term prognosis after myocardial infarction in men with diabetes. *Diabetes* 1985; **34**:787–792.
23. Lowel H, Koenig W, Engel S, Hormann A, Keil U. Impact of diabetes on survival after myocardial infarction. *Diabetologia* 2000; **43**:218–226.
24. Hedges LV, Olkin L. *Statistical Methods for Meta-analysis*. Academic Press: New York, 1985.
25. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 1977; **39**:1–38.
26. Felson DT. Bias in meta-analytic research. *Journal of Clinical Epidemiology* 1992; **45**:885–892.
27. Fleiss JL. The statistical basis of meta-analysis. *Statistical Methods in Medical Research* 1993; **2**:121–145.

