Computational Statistics and Data Analysis 113 (2017) 154–176

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Cross-sectional design with a short-term follow-up for prognostic imaging biomarkers☆

Joong-Ho Won [a], Xiao Wu [b], Sang Han Lee [c], Ying Lu [d,*]

[a] Department of Statistics, Seoul National University, Republic of Korea

[b] Department of Biostatistics, Harvard University, United States

[c] Center for Biomedical Imaging and Neuromodulation, The Nathan S. Kline Institute for Psychiatric Research, United States

[d] Departments of Biomedical Data Science, Radiology, and Health Research & Policy, Stanford University School of Medicine, United States

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form 30 December 2016
Accepted 31 December 2016
Available online 17 January 2017

Keywords:
Case-control study
Short-term follow-up
Cross-sectional design
Retrospective study
Prospective cohort study
Imaging biomarker
Imaging diagnostics
Prognosis
Osteoporosis
Osteoporotic fracture
Alzheimer's disease
Dementia
Hippocampus
MRI

ABSTRACT

Medical imaging techniques are being rapidly developed and used for diagnosis and prognostic predictions. To validate a prognostic predictive utility of a new imaging marker, a temporal association needs to be established to show an association between its baseline value with a subsequent chance of having the relevant clinical outcome. Validation of such techniques has several difficulties. First, different from techniques based on blood or tissue specimen, imaging techniques often have no historical samples to study and require new studies to collect data. For rare events, it can be costly. Second, the rapid technology evolution requires such validation studies to be short in order to keep the evaluation relevant. A new statistical design is proposed that extends traditional prospective cohort study by adding cases with known time of events and including a short-term follow-up to estimate the prospective odds ratio for the clinical endpoint of interest within a reasonably short duration of time. Under a Markov model, this new design can deliver a consistent estimate of the odds ratio and a formula for asymptotic variance. Numerical studies suggest that the new design induces a smaller variance than the corresponding prospective cohort study within three follow-ups. An application to Alzheimer's disease data demonstrates that the proposed design has a potential to be useful to rapidly establish a prognostic validity of a new imaging marker within a reasonable time, with a small sample size.

Published by Elsevier B.V.

1. Introduction

Precision medicine depends on successfully developed biomarkers to accurately predict a risk of adverse health outcomes, such as having a disease, osteoporotic fracture, heart failure, or death. A biomarker is defined as "a characteristic that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic response to therapeutic intervention (Biomarkers Definitions Working Group, 2001)". New biomarkers are

☆ For the Alzheimer's Disease Neuroimaging Initiative. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Corresponding author.
E-mail address: ylu1@stanford.edu (Y. Lu).

http://dx.doi.org/10.1016/j.csda.2016.12.017
0167-9473/Published by Elsevier B.V.

discovered in an accelerated pace due to recent advances in genetic, genomic, and protein-based methods of treating diseases as well as evolutions in *in vivo* imaging and tissue collection technologies (Altar, 2008). Biomarkers have multiple applications, including prediction of disease risk, detection of an existing disease, and estimation of patient prognosis (European Society of Radiology, 2010). To be useful as a prognostic tool, a biomarker has to be scientifically validated through statistical studies. As an illustration, suppose that we are interested in the clinical event of Alzheimer's disease (AD), the most common form of dementia, among those with mild cognitive impairment (MCI), a high risk group for this disease, and have a biomarker that takes a value of H or L, denoting "high" and "low" respectively. We want to see if the baseline measurement of H or L will predict the risk of AD in, say, the next year. Among various designs for scientific validation of the marker's predictive utility, prospective cohort study and (nested) case-control study are the most relevant.

Prospective cohort study (see, e.g., Porta et al., 2014) is a form of longitudinal study to evaluate a potential association of the biomarker and the event. In such a study, a group of participants from a population is selected, and each individual in the group is followed for many years to track the occurrence of the event of interest. In the example of AD above, we may measure the biomarker status for a group of elderly MCI subjects, and follow them annually to see who convert to AD. If the conversion rates turn out to be significantly different between those with the biomarker status of H and those with L, we conclude that the biomarker of question is indeed a valid prognostic biomarker for AD. A cohort study controls many confounding factors and is the most desirable approach in validating the desired association. However, it is expensive and not always feasible, especially if the event of interest is rare.

As an alternative, case-control study (see, e.g., Schlesselman, 1982) selects cases, e.g., those elderly who had a conversion from MCI to AD before the baseline of the study, and find those without a conversion with all the other known risk factors matched to those cases. We then look back their biomarker status in the past to see the association. Breslow (Breslow, 1976, 1996) showed that a case-control study can yield the same odds ratio as that would have been obtained from a prospective cohort study. The advantage of case-control study is that it can be performed at a lower cost for rare events than the corresponding prospective cohort study, and in a shorter time. The validity of a case-control study, however, depends upon its ability to evaluate the historical values of the potential biomarker. For the validation of serum- or tissue-based biomarkers, nested case-control study has been used conventionally. This design relies on historically conducted cohort studies, from which the cases and the controls are selected retrospectively. It then takes advantage of historically collected serum or tissue specimen in the cohort studies to determine the historical values of the potential biomarker.

Imaging biomarkers measure anatomical, functional or molecular parameters using radiological imaging technologies. Compared to tissue-based and serum-based biomarkers, characteristics of imaging biomarkers include non-invasiveness, spatiotemporal resolution, ability to be measured longitudinally, and adaptivity to technological advances. Due to these characteristics, imaging biomarkers are increasingly used in major medical areas such as cancer, cardiovascular diseases, musculoskeletal diseases, neurological diseases, and so on (European Society of Radiology, 2010). As for validation, longitudinal cohort study is particularly difficult for imaging biomarkers, because of the rapid pace of technological innovations in imaging sciences: often the technology becomes obsolete before the conclusion of a cohort study is made. Cost and feasibility issues for rare events still pertain. Nested case-control study design does not work for imaging biomarkers either, because many biomarkers depend on the technology that did not exist in the past, thus the historical values are impossible to evaluate. For example, in studying AD, the hippocampus as an imaging biomarker has drawn a significant attention by researchers because of the nature of neuronal loss related with memory loss in AD. As imaging technology develops, it has become feasible to observe hippocampus *in vivo* through non-invasive ways such as magnetic resonance imaging (MRI). Numerous studies have reported that hippocampus atrophy is associated with AD (Schuff et al., 2009, and references therein) and loss of hippocampal volume predicts progression to AD (Jack et al., 1999). Due to the insidious nature of AD, longitudinal study is desirable to confirm that temporal change of hippocampal volume has a prognostic potential as an MRI-based biomarker for AD; some large-scale brain MRI cohort studies were launched in more than 10 years ago, and are still on-going (Weiner et al., 2015). However, the most developed MRI scanner is not available to every participant in the study because of the rapidly developing technology. For instance, 1.5T MRI scanners were popular in a few years ago, but nowadays 3T scanners are demanded in most AD studies.

Another example is from the research of osteoporosis, where the incidence rate is low and standard diagnosis method is low-cost already. The broadband quantitative ultrasonometry (QUS) for studying bone mineral density and structure at low cost was developed before 1984 (Langton et al., 1984). However, its prognostic utility of measuring osteoporic fracture risk was not established until a longitudinal cohort study through the Study of Osteoporotic Fractures (SOF) confirmed it in 1997 (Bauer et al., 1997). Although the SOF cohort had previous visits, it was not possible to use a nested case-control study design and the past SOF data because the historical QUS measures were not available. In recent years, the costs of QUS and dual X-ray absorptiometry (DXA), a competing osteodensitometry method, have both gone down to less than $50 per scan, making future technologies much more difficult to validate their prognostic utility through prospective cohort studies.

Thus there is a strong demand for a design that is more efficient and rapid than prospective cohort designs so not to wait until all normal subjects to develop diseases, while estimating the odds ratio of a new imaging biomarker unbiasedly compared with cross-sectional designs. We propose a hybrid case-control study design with a short-term follow-up in order to estimate the odds ratio of a new (potential) biomarker consistently within a short time and with a feasible sample size. Our design takes advantage of a case-control design to enrich cases for non-fatal outcomes, and a feature of imaging biomarkers that they can be measured repeatedly for study participants via follow-ups. This design, however, can be applied to any other biomarkers as long as they share these characteristics, such as mobile biomarkers. We use statistical modeling to impute the

missing marker measurements prior to the outcome. We also assume clinical outcomes for which the time of occurrence is known (e.g., AD or fracture), while this assumption can be relaxed.

We organize our paper as follows. Section 2 describes the new design and the statistical model in detail. Section 3 considers the estimation procedure and the large-sample properties associated with the new design. In Section 4 we conduct numerical and simulation studies to compare the efficiency between our new design and prospective cohort study design. In Section 5 we apply the new design to a study of imaging biomarkers for AD. Section 6 provides conclusions and discussions.

Notation We borrow some notations from matrix calculus. A matrix is denoted by a capital letter. A vector is denoted by a lowercase letter. Every vector is a column vector otherwise mentioned. The transpose of matrix A is denoted by A^T. Notation $(A)_{ij}$ represents the component of matrix A on the ith row and jth column at the same time. Notation $(A)_{i\cdot}$ denotes the column vector made from the ith row of matrix A. Notation $(A)_{\cdot j}$ denotes the jth column of matrix A. The ith component of vector x is denoted by x_i. Symbol 0 is used to represent either number zero, zero vector, or zero matrix whose dimension is determined by the context and conformality. When necessary, notations $O_{m\times n}$ and O_m are used to refer to $m \times n$ and $m \times m$ zero matrices, respectively. Symbol **1** represents a vector of all ones, with a dimension conformal to the operation with other matrices. Operator vec() is used to vectorize a matrix: for an $m \times n$ matrix A, $\mathrm{vec}(A) = [(A)_{\cdot 1}^T, \ldots, (A)_{\cdot n}^T]^T \in \mathbb{R}^{mn}$. The inverse of the vec() operator is mat(): $\mathrm{mat}(\mathrm{vec}(A)) = A$. Notation $\mathrm{diag}(A_1, \ldots, A_n)$ refers to a (block) diagonal matrix that consists diagonal blocks $A_1, \ldots, A_n$; for a vector $x = (x_1, \ldots, x_n)^T$, $\mathrm{diag}(x) = \mathrm{diag}(x_1, \ldots, x_n)$. The Kronecker product is denoted by $\otimes$: if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, then

$$A \otimes B = \begin{bmatrix} (A)_{11}B & \cdots (A)_{1n}B \\ \vdots & \ddots \vdots \\ (A)_{m1}B & \cdots (A)_{mn}B \end{bmatrix} \in \mathbb{R}^{mp\times nq}.$$

Matrix differential is such that

$$dA = \begin{bmatrix} d(A)_{11} & \cdots d(A)_{1n} \\ \vdots & \ddots \vdots \\ d(A)_{m1} & \cdots d(A)_{mn} \end{bmatrix},$$

so $\mathrm{vec}(dA) = d\,\mathrm{vec}(A)$. The Kronecker delta is defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

2. Cross-sectional design with a short-term follow-up

The concept of the cross-sectional design with a short-term follow-up is illustrated in Fig. 1 in the context of the AD example of the previous section. The population at risk is elderly people who have MCI initially. The research question is whether the biomarker of interest, such as MRI-based hippocampal volume, is predictive of AD within a year from its measurement. Because the prevalence of AD is low, prospective designs are expected to be expensive and difficult to implement. Instead, a cross-sectional design can be conducted as follows. The study population, covariate-matched at the baseline, consists of two groups, one those who developed AD within t_0 time units before the study begins (case), and the other free of AD by the baseline (control). The time axis is normalized so that t_0 corresponds to the origin. For simplicity, it is assumed that the biomarker yields binary values of H and L. This may refer to relatively small (high risk of AD) and large (low risk) hippocampal volumes, respectively. Let the biomarker measurement at time t be $Y(t)$. This is a binary-valued stochastic process. The disease status is also a stochastic process and is denoted by $S(t)$, taking binary values of N (normal) and D (diseased). Thus the control group corresponds to those with $S(t_0) = \mathsf{N}$, and the case $S(t_0) = \mathsf{D}$. Initially, $S(0) \equiv \mathsf{N}$. Then the quantity of interest is the odds ratio

$$\frac{\Pr(S(t) = \mathsf{D} | S(0) = \mathsf{N}, Y(0) = \mathsf{L}) / \Pr(S(t) = \mathsf{D} | S(0) = \mathsf{N}, Y(0) = \mathsf{H})}{\Pr(S(t) = \mathsf{N} | S(0) = \mathsf{N}, Y(0) = \mathsf{L}) / \Pr(S(t) = \mathsf{N} | S(0) = \mathsf{N}, Y(0) = \mathsf{H})},$$

for $t = 1, \ldots, t_0$. Unfortunately, however, the historical values of the biomarker are not available, thus $Y(0)$ is unknown.

Nevertheless we can keep track of the progression of disease at population level as measured by the proportion of each biomarker level for $t > t_0$. Typically the proportion of H (small hippocampi) increases at a natural rate with aging but this rate is accelerated after the on-set of AD at time $\tau < t_0$. Therefore, if these rates do not vary abruptly with time, then a short-term follow-up study is sufficient to estimate them reliably. We can also extrapolate the progression to estimate $Y(0)$ using the estimated rates. Note that, under the same assumption of stable progression, the incidence rate for each level of the biomarker can also be estimated, hence the odds ratio. However, unless the incidence rates are large enough for all levels, they cannot be reliably estimated by a short-term follow-up with small to moderate sample sizes; if this is indeed the case, a prospective design will become feasible.

To be able to extrapolate the progression, we need a model for the temporal progression of the disease. A simple 2×2-state Markov chain model captures the key features of the assumed progression. This model consists of the measurement

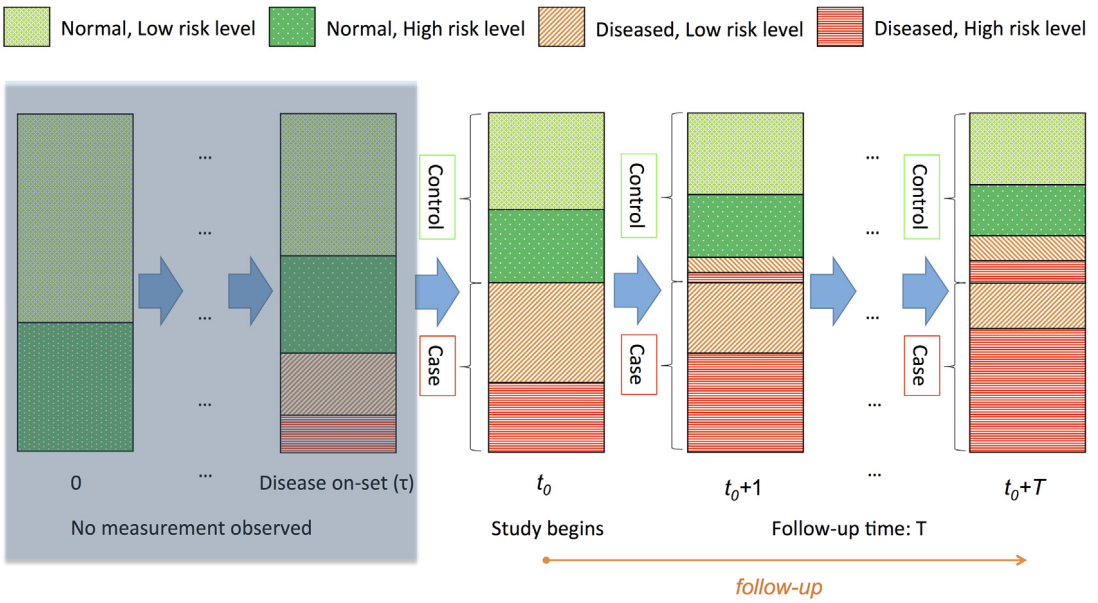

Fig. 1. Illustration of the cross-sectional design with a short-term follow-up. A measurement of the biomarker level is made in a cross-sectional fashion at time t_0 for each subject in the control group and the case group. Measurements are repeated for T time units during the short-term follow-up period in order to estimate the rate of change in the proportions of biomarker levels, which differs between the groups after the disease on-set τ if the biomarker is associated with the disease progression. Using this information the past proportions of the biomarker levels are imputed for each group (blurred region), in order to estimate the prospective odds ratio of the biomarker levels. Areas filled with green and dots represent the proportions of both biomarker levels in the undiseased population, whereas those with red stripes the diseased population. The darkened areas indicate the proportions of diseased subjects during the follow-up, which is difficult to observe as the incidence is low.

process $\{Y_i(t)\}$ and the disease status process $\{S_i(t)\}$, where i stands for a subject. We assume $(Y_i(t), S_i(t)) \overset{i.i.d.}{=} (Y(t), S(t))$, where $Y(t) \in \{\mathsf{H}, \mathsf{L}\}$, and $S(t) \in \{\mathsf{N}, \mathsf{D}\}$, as the illustration in the previous paragraphs. We further assume that the joint process $X_t = (Y(t), S(t))$ forms a discrete-time Markov chain. The key assumptions of this Markov chain model are that (1) the "diseased" state is irreversible, and (2) at the disease on-set, the biomarker level does not change abruptly (Chiang, 1980). The first assumption is a common characteristic of many non-fatal irreversible diseases, such as AD and osteoporotic fracture. The second assumption is due to that the probability of joint events at the same time is low and can be neglected in an approximation. A similar Markov chain model is considered to analyze temporal changes in MRI-based hippocampal volume measurements (Schuff et al., 2009).

In matrix form, the transition matrix of our Markov chain can be written as

$$P = \begin{bmatrix} \bar{\Lambda} P_N & \Lambda \\ & P_D \end{bmatrix}, \tag{1}$$

where

$$P_N = \begin{matrix} \\ \mathsf{LN} \\ \mathsf{HN} \end{matrix} \begin{matrix} \begin{matrix} \mathsf{LN} & \mathsf{HN} \end{matrix} \\ \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \end{matrix}, \qquad P_D = \begin{matrix} \\ \mathsf{LD} \\ \mathsf{HD} \end{matrix} \begin{matrix} \begin{matrix} \mathsf{LD} & \mathsf{HD} \end{matrix} \\ \begin{bmatrix} r & 1-r \\ 1-s & s \end{bmatrix} \end{matrix}, \tag{2}$$

and

$$\Lambda = \mathrm{diag}(1-\alpha, 1-\beta), \qquad \bar{\Lambda} = \mathrm{diag}(\alpha, \beta). \tag{3}$$

Matrix P_N is the conditional transition matrix of the biomarker *before* the disease on-set given that the subject remains normal after the transition, i.e., conditioned on $\tau > t+1$, since

$$\begin{aligned} \Pr(X_{t+1} = \mathsf{LN} | X_t = \mathsf{LN}, \tau > t+1) &= \frac{\Pr(X_{t+1} = \mathsf{LN}, X_t = \mathsf{LN}, \tau > t+1)}{\Pr(X_t = \mathsf{LN}, \tau > t+1)} \\ &= \frac{\Pr(X_{t+1} = \mathsf{LN}, X_t = \mathsf{LN})}{\Pr(X_{t+1} = \mathsf{LN}, X_t = \mathsf{LN}) + \Pr(X_{t+1} = \mathsf{HN}, X_t = \mathsf{LN})} \\ &= \frac{p\alpha}{p\alpha + (1-p)\alpha} = p, \end{aligned}$$

the proportion of unchanged L marker readings among the normal in one year, etc. Likewise, P_D is the conditional transition matrix of the biomarker *after* the disease on-set, or $\tau \le t$. Matrix Λ represents the transition probabilities from normal

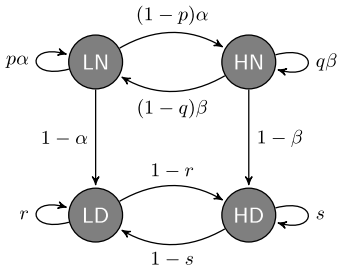

Fig. 2. State transition diagram of the 2×2-state Markov chain model of the biomarker level and the disease progression.

states to diseased. That is, α is the probability of transition from LN to LD, and β is that from HN to HD. This matrix is diagonal reflecting the second assumption in the previous paragraph. The chain starts from "normal" states, i.e.,

$$(\pi^{(0)})^T = \overset{\text{LN}}{\pi^{(0)}_{LN}} \quad \overset{\text{HN}}{\pi^{(0)}_{HN}} \quad \overset{\text{LD}}{0} \quad \overset{\text{HD}}{0} \;] = [(\pi^{(0)}_N)^T, 0^T] \tag{4}$$

in accordance with the population at risk. Given the distribution of the states at time $t-1$, the distribution at time t is obtained by the relation

$$(\pi^{(t)})^T = \left[(\pi^{(t)}_N)^T \quad (\pi^{(t)}_D)^T\right] = (\pi^{(t-1)})^T P.$$

The transition diagram of the Markov model is shown in Fig. 2.

Under this model, the basic quantity of interest is the 1-step odds ratio of the disease given that the subject is normal at age t given by

$$\begin{aligned}\frac{\Pr(\tau = t+1 | X_t = \text{HN}) / \Pr(\tau > t+1 | X_t = \text{HN})}{\Pr(\tau = t+1 | X_t = \text{LN}) / \Pr(\tau > t+1 | X_t = \text{LN})} &= \frac{\Pr(X_{t+1} = \text{HD} | X_t = \text{HN}) / \Pr(X_{t+1} = \text{HN or LN} | X_t = \text{HN})}{\Pr(X_{t+1} = \text{LD} | X_t = \text{LN}) / \Pr(X_{t+1} = \text{LN or HN} | X_t = \text{LN})} \\ &= \frac{(1-\beta)/\beta}{(1-\alpha)/\alpha},\end{aligned} \tag{5}$$

or its logarithm, which is a function of Λ (or $\bar{\Lambda}$). Even though the individual components of Λ are not estimable with the cross-sectional design, the (log) odds ratio can be reliably estimated. Quantities P_N, P_D, and $\pi^{(0)}_N$ are nuisance parameters affecting the accuracy of the estimation of the odds ratio; the former two can be well estimated by the short-term follow-up.

3. Maximum likelihood estimation and large-sample properties

In this section we consider the maximum likelihood estimation of the parameters of the 2×2-state model (1)–(4), and study the large-sample properties thereof. We first look at a reparameterization of the model to ease computation, and construct the log likelihood functions for individual parts of the design, that is, for the cross-sectional part and for the short-term follow-up part separately. Since the 2×2-state model (1)–(4) is easily extended to an $L \times 2$-state model with L ordinal risk levels, we describe the estimation procedure in this more general setting in the sequel.

3.1. Estimation

Reparameterization We use the following reparameterization

$$\bar{\Lambda}_{jj} = \frac{1}{1+e^{\lambda_j}}, \qquad \pi^{(0)}_{Nk} = \frac{e^{\gamma_{0k}}}{\sum_{l=1}^{L}(1+e^{\gamma_{0l}})}, \qquad (P_N)_{jk} = \frac{e^{\gamma_{jk}}}{\sum_{l=1}^{L}(1+e^{\gamma_{jl}})}, \qquad (P_D)_{jk} = \frac{e^{\bar{\gamma}_{jk}}}{\sum_{l=1}^{L}(1+e^{\bar{\gamma}_{jl}})},$$

and $\gamma_{jL} = 0$, $\bar{\gamma}_{jL} = 0$ for $j = 1, \ldots, L$ and $k = 1, \ldots, L$. We also set $\gamma_{0L} = 0$ for identifiability. With this parameterization each parameter is unconstrained in $(-\infty, \infty)$. We denote the parameters collectively by a vector

$$\theta = [\lambda^T, \gamma_0^T, \gamma^T, \bar{\gamma}^T]^T \in \mathbb{R}^{2L^2-1},$$

where

$$\begin{aligned}&\lambda = (\lambda_1, \ldots, \lambda_L)^T, \\ &\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0,L-1})^T, \\ &\gamma = (\gamma_1^T, \ldots, \gamma_L^T)^T, \qquad \gamma_j = (\gamma_{j1}, \ldots, \gamma_{j,L-1})^T, \quad j = 1, \ldots, L \\ &\bar{\gamma} = (\bar{\gamma}_1^T, \ldots, \bar{\gamma}_L^T)^T, \qquad \bar{\gamma}_j = (\bar{\gamma}_{j1}, \ldots, \bar{\gamma}_{j,L-1})^T, \quad j = 1, \ldots, L.\end{aligned}$$

Table 1
Data structure of the cross-sectional part of the design for the $L \times 2$-state Markovian model.

		Control (N)	Case (D)			
		$\tau > t_0$	$\tau = 1$	$\tau = 2$	$\cdots$	$\tau = t_0$
Risk level	L_1	n_1	m_{11}	m_{12}	$\cdots$	m_{1t_0}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	L_L	n_L	m_{L1}	m_{L2}	$\cdots$	m_{Lt_0}
	Sum	n	m			

The log odds ratio of any two levels L_j and L_k is then given by $\lambda_k - \lambda_j$. In particular, for the 2×2-state model ($L = 2$),

$$\alpha = 1/(1 + e^{\lambda_1}),$$
$$\beta = 1/(1 + e^{\lambda_2}).$$

so that the log odds ratio is $\lambda_2 - \lambda_1$.

Cross-sectional part Let the control group size be n and the case group size be m. At the beginning of the study, the number of observations for each group for each risk level can be summarized as in Table 1.
In this table, n, the total number of the control subjects, and m, the total number of the case subjects, are fixed by design; $\tau = 1, \ldots, t_0$ refer to the observed disease on-set times for the case subjects; we assume that these are known for each subject.

The probability of the observation in the control group (normal by time t_0) is multinomial with L cells, and the probability of the jth cell is

$$\Pr(X_{t_0} = \mathsf{L}_j\mathsf{N} | T > t_0) = \Pr(X_{t_0} = \mathsf{L}_j\mathsf{N} | X_{t_0} = \mathsf{L}_1\mathsf{N} \text{ or } \cdots \text{ or } \mathsf{L}_L\mathsf{N}) = \frac{\pi_{N,j}^{(t_0)}}{\pi_{N,1}^{(t_0)} + \cdots + \pi_{N,L}^{(t_0)}} = \frac{(\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} e_j}{(\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} \mathbf{1}},$$

where $\pi_N^{(t)}$ is a vector that consists of the first L entries of $\pi^{(t)} = P^T \pi^{(0)}$, i.e., $\pi^{(t)} = [(\pi_N^{(t)})^T, (\pi_D^{(t)})^T]^T$. Thus the log-likelihood of the control group is

$$l_{\text{ctrl}}^{(CS)}(\theta) = \sum_{j=1}^{L} n_j \log(\pi_N^{(0)} (\bar{\Lambda} P_N)^{t_0} e_j) - n \log(\pi_N^{(0)} (\bar{\Lambda} P_N)^{t_0} \mathbf{1}) + \text{const.} \tag{6}$$

The sampling distribution of the case group is also a multinomial with $L \times t_0$ cells. The probability of the (j, τ)th cell is given by

$$p_{\mathsf{L}_j\mathsf{D}}^{(\tau)} = \frac{\Pr(X_{t_0} = \mathsf{L}_j\mathsf{D}, T = \tau)}{\Pr(T \le t_0)}, \quad j = 1, \ldots, L, \ \tau = 1, \ldots, t_0.$$

The numerator is expressed as

$$\begin{aligned}\Pr(X_{t_0} = \mathsf{L}_j\mathsf{D}, T = \tau) &= \Pr(X_{\tau-1} = \mathsf{L}_1\mathsf{N}, X_\tau = \mathsf{L}_1\mathsf{D}, X_{t_0} = \mathsf{L}_j\mathsf{D}) + \cdots + \Pr(X_{\tau-1} = \mathsf{L}_L\mathsf{N}, X_\tau = \mathsf{L}_L\mathsf{D}, X_{t_0} = \mathsf{L}_j\mathsf{D}) \\ &= (\pi_N^{(\tau-1)})^T \Lambda P_D^{t_0-\tau} e_j = (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{\tau-1} \Lambda P_D^{t_0-\tau} e_j.\end{aligned}$$

The denominator is

$$\Pr(T \le t_0) = \Pr(X_{t_0} = \mathsf{L}_1\mathsf{D} \text{ or } \cdots \text{ or } \mathsf{L}_L\mathsf{D}) = \pi_{D,1}^{(t_0)} + \pi_{D,L}^{(t_0)} = (\pi_D^{(t_0)})^T \mathbf{1} = (\pi_N^{(0)})^T \left(\sum_{k=0}^{t_0-1} (\bar{\Lambda} P_N)^k \right) \Lambda \mathbf{1},$$

because

$$\begin{aligned}(\pi_D^{(t)})^T &= (\pi_{(0)})^T \Lambda P_D^{t-1} + (\pi_N^{(1)})^T \Lambda P_D^{t-2} + \cdots + (\pi_N^{(t-1)})^T \Lambda \\ &= (\pi_N^{(0)})^T \Lambda P_D^{t-1} + (\pi_N^{(0)})^T (\bar{\Lambda} P_N) \Lambda P_D^{t-2} + \cdots + (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t-1} \Lambda\end{aligned}$$

and $P_D^k \mathbf{1} = \mathbf{1}$. Thus

$$p_{\mathsf{L}_j\mathsf{D}}^{(\tau)} = \frac{(\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{\tau-1} \Lambda P_D^{t_0-\tau} e_j}{\sum_{t=1}^{t_0} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t-1} \Lambda \mathbf{1}},$$

and the log likelihood of the case group is

$$l^{(CS)}_{\text{case}}(\theta) = \sum_{t=1}^{t_0}\sum_{j=1}^{L} m_{jt} \log\left((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t}e_j\right) - m\log\left(\sum_{t=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^t\Lambda\mathbf{1}\right) + \text{const.} \tag{7}$$

The log likelihood of the cross-sectional part of the data is thus given by

$$\begin{aligned} l^{(CS)}(\theta) &= l^{(CS)}_{\text{ctrl}}(\theta) + l^{(CS)}_{\text{case}}(\theta) \\ &= \sum_{j=1}^{L} n_j \log(\pi_N^{(0)}(\bar{\Lambda}P_N)^{t_0}e_j) - n\log(\pi_N^{(0)}(\bar{\Lambda}P_N)^{t_0}\mathbf{1}) \\ &\quad + \sum_{t=1}^{t_0}\sum_{j=1}^{L} m_{jt}\log\left((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t}e_j\right) - m\log\left(\sum_{t=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^t\Lambda\mathbf{1}\right) + \text{const.} \end{aligned} \tag{8}$$

Short-term follow-up part Conditioning on the cross-sectional observations $(n_1, \ldots, n_L; m_{11}, \ldots, m_{Lt_0})$, the likelihood of the short-term follow-up part of the data is that of the Markov chain, governed by the transition matrix P, with two samples of initial observations $(n_1, \ldots, n_L; 0, \ldots, 0)$ and $(0, \ldots, 0; m_{1\cdot}, \ldots, m_{L\cdot})$ for states $(\mathsf{L}_1\mathsf{N}, \ldots, \mathsf{L}_L\mathsf{N}; \mathsf{L}_1\mathsf{D}, \ldots, \mathsf{L}_L\mathsf{D})$, where $m_{j\cdot} = \sum_{t=1}^{t_0} m_{jt}, j = 1, \ldots, L$.

The likelihood of the control group for the follow-up part is given by

$$L^{(FU)}_{ctrl}(\theta) \propto \prod_{l\in\mathcal{S}}\prod_{m\in\mathcal{S}} p_{lm}(\theta)^{\nu_{lm}}, \quad \mathcal{S} = \{\mathsf{L}_1\mathsf{N}, \ldots, \mathsf{L}_L\mathsf{N}, \mathsf{L}_1\mathsf{D}, \mathsf{L}_L\mathsf{D}\},$$

where $p_{lm}(\theta) = (P)_{lm}$, and ν_{lm} is the total number of transitions from state l to state m for $t = t_0 + 1, \ldots, t_0 + T$ in the control group. Similarly, the likelihood of the case group is

$$L^{(FU)}_{case}(\theta) \propto \prod_{l\in\mathcal{S}}\prod_{m\in\mathcal{S}} p_{lm}(\theta)^{\mu_{lm}},$$

where μ_{lm} is the total number of transitions from state l to state m for $t = t_0 + 1, \ldots, t_0 + T$ in the case group. Note that no transition from $\mathsf{L}_j\mathsf{D}$ to $\mathsf{L}_k\mathsf{N}$ $(j \neq k)$ is allowed in the model, hence $\mu_{lm} = 0$ for $l = \mathsf{L}_j\mathsf{D}, m = \mathsf{L}_k\mathsf{N}, j \neq k$. The log likelihood of the short-term follow-up part of the data is then given by

$$l^{(FU)}(\theta) = \sum_{l\in\mathcal{S}}\sum_{m\in\mathcal{S}} \nu_{lm}\log p_{lm}(\theta) + \sum_{l\in\mathcal{S}}\sum_{m\in\mathcal{S}} \mu_{lm}\log p_{lm}(\theta) + \text{const.} \tag{9}$$

Full likelihood The full log likelihood can therefore be written as

$$\begin{aligned} l(\theta) &= l^{(CS)}(\theta) + l^{(FU)}(\theta) \\ &= \sum_{j=1}^{L} n_j\log(\pi_N^{(0)}(\bar{\Lambda}P_N)^{t_0}e_j) - n\log(\pi_N^{(0)}(\bar{\Lambda}P_N)^{t_0}\mathbf{1}) + \sum_{t=1}^{t_0}\sum_{j=1}^{L} m_{jt}\log\left((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t}e_j\right) \\ &\quad - m\log\left(\sum_{t=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^t\Lambda\mathbf{1}\right) + \sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{S}}\nu_{jk}\log p_{jk}(\theta) + \sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{S}}\mu_{jk}\log p_{jk}(\theta) + \text{const.}, \end{aligned} \tag{10}$$

which is to be maximized.

3.2. Large-sample properties

Given the point estimate $\hat{\theta}$ of θ (hence that of the log odds ratio $\lambda_k - \lambda_j$), we are interested in its large-sample properties. The limiting condition of interest is when $n, m \to \infty$, while the ratio between the case and control group sizes remains fixed. First note that the full log likelihood (10) can be written as

$$l(\theta) = l_{ctrl}(\theta) + l_{case}(\theta),$$

where

$$l_{ctrl}(\theta) = l^{(CS)}_{ctrl}(\theta) + l^{(FU)}_{ctrl}(\theta), \tag{11}$$

$$l_{case}(\theta) = l^{(CS)}_{case}(\theta) + l^{(FU)}_{case}(\theta). \tag{12}$$

Samples for $l_{ctrl}(\cdot)$ and $l_{case}(\cdot)$ are independent. Since each of $l_{ctrl}(\theta)$ and $l_{case}(\theta)$ is a log likelihood of a multinomial distribution whose cell probabilities are a smooth function of the parameter θ, we have

$$\sqrt{n}(\hat{\theta}_{ctrl} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I_{ctrl}(\theta^0)^{-1}),$$

$$\sqrt{m}(\hat{\theta}_{case} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I_{case}(\theta^0)^{-1}),$$

as $n, m \to \infty$, where $\hat{\theta}_{ctrl}, \hat{\theta}_{case}$ are the respective maximizers of (11), (12), provided that the Fisher information matrices

$$I_{ctrl}(\theta) = \frac{1}{n} E_\theta \left[\left(\frac{\partial l_{ctrl}}{\partial \theta} \right) \left(\frac{\partial l_{ctrl}}{\partial \theta} \right)^T \right]$$
$$I_{case}(\theta) = \frac{1}{m} E_\theta \left[\left(\frac{\partial l_{case}}{\partial \theta} \right) \left(\frac{\partial l_{case}}{\partial \theta} \right)^T \right]$$

are full rank at $\theta = \theta^0$, the true but unknown value of the parameter, and all the elements of $\pi_N^{(0)}$, P_N, P_D, and $\mathrm{diag}(\Lambda)$ are positive (Rao, 1973; Agresti and Kateri, 2012, Ch. 16). These conditions are satisfied if θ^0 is bounded. Then the MLE $\hat{\theta}$, or the maximizer of (10) is asymptotically normally distributed as

$$\sqrt{n+m}(\hat{\theta} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I(\theta^0)^{-1}), \tag{13}$$

where

$$I(\theta) = \rho_1 I_{ctrl}(\theta) + \rho_2 I_{case}(\theta),$$

as $n, m \to \infty$, $n/(n+m) \to \rho_1$, $m/(n+m) \to \rho_2$, and $\rho_1 + \rho_2 = 1$ (Lehmann and Casella, 1998, Theorem 6.7.1). Thus it suffices to evaluate $I_{ctrl}(\theta)$ and $I_{case}(\theta)$.

The evaluation of $I_{ctrl}(\theta)$ and $I_{case}(\theta)$ are, however, somewhat involved, thus we summarize the results in the following lemmas.

Lemma 1. *Define matrix-valued functions*

$$G_k(a, B, C) = \sum_{l=0}^{k-1} \mathrm{diag}(a^T B^l) C B^{k-1-l} \in \mathbb{R}^{L \times L}, \qquad F_k(a, B, C) = \begin{bmatrix} \sum_{l=0}^{k-1} a^T B^l C e_1 B^{k-1-l} \\ \vdots \\ \sum_{l=0}^{k-1} a^T B^l C e_p B^{k-1-l} \end{bmatrix} \in \mathbb{R}^{L^2 \times L},$$

for $a \in \mathbb{R}^L$, $B, C \in \mathbb{R}^{L \times L}$. *Then*

$$I_{ctrl}(\theta) = \begin{bmatrix} K_\lambda D_1 K_\lambda^T & K_\lambda D_1 K_{\gamma_0}^T & K_\lambda D_1 K_\gamma^T & 0 \\ K_{\gamma_0} D_1 K_\lambda^T & K_{\gamma_0} D_1 K_{\gamma_0}^T & K_{\gamma_0} D_1 K_\gamma^T & 0 \\ K_\gamma D_1 K_\lambda^T & K_\gamma D_1 K_{\gamma_0}^T & K_\gamma D_1 K_\gamma^T & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} W_\lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & W_\gamma & 0 \\ 0 & 0 & 0 & W_{\tilde{\gamma}} \end{bmatrix} =: V_{11} + V_{12}, \tag{14}$$

where $D_1 = \mathrm{diag}(\pi^{(ctrl)})^{-1} - \mathbf{1}\mathbf{1}^T$, *with*

$$\pi^{(ctrl)} = (\pi_1^{(ctrl)}, \ldots, \pi_L^{(ctrl)})^T \quad \text{with } \pi_j^{(ctrl)} = \left((\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} \mathbf{1} \right)^{-1} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} e_j,\ j = 1, \ldots, L,$$

and

$$K_\lambda = \left((\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} \mathbf{1} \right)^{-1} \bar{\Lambda}(\bar{\Lambda} - I) G_{t_0}(\pi_N^{(0)}, \bar{\Lambda} P_N, P_N), \qquad K_{\gamma_0} = \left((\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} \mathbf{1} \right)^{-1} \bar{\Lambda}(\bar{\Lambda} - I) \Sigma_0^T (\bar{\Lambda} P_N)^{t_0},$$
$$K_\gamma = \left((\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t_0} \mathbf{1} \right)^{-1} \bar{\Lambda}(\bar{\Lambda} - I)\, \mathrm{diag}(\Sigma_1^T, \ldots, \Sigma_L^T) F_{t_0}(\pi_N^{(0)}, \bar{\Lambda} P_N, \bar{\Lambda}),$$

with

$$\Sigma_0 = \begin{bmatrix} \tilde{\Sigma}_0 \\ -\mathbf{1}^T \tilde{\Sigma}_0 \end{bmatrix} \in \mathbb{R}^{L \times (L-1)}, \quad \tilde{\Sigma}_0 = \mathrm{diag}(I_{L-1}, 0)(\mathrm{diag}(\pi_N^{(0)}) - \pi_N^{(0)} (\pi_N^{(0)})^T)\, \mathrm{diag}(I_{L-1}, 0),$$
$$\Sigma_j = \begin{bmatrix} \tilde{\Sigma}_j \\ -\mathbf{1}^T \tilde{\Sigma}_j \end{bmatrix} \in \mathbb{R}^{L \times (L-1)}, \quad \tilde{\Sigma}_j = \mathrm{diag}(I_{L-1}, 0)(\mathrm{diag}(P_N^T e_j) - (P_N^T e_j)(P_N^T e_j)^T)\, \mathrm{diag}(I_{L-1}, 0),\ j = 1, \ldots, L.$$

Finally,

$$
\begin{aligned}
W_\lambda &= \operatorname{diag}(\phi_1 \bar{\Lambda}_{11}^2 (1-\bar{\Lambda}_{11})^2 (e_1^T P_N \operatorname{diag}(P_N^T \bar{\Lambda} e_1)^{-1} P_N^T e_1 + 1/(1-\bar{\Lambda}_{11})), \ldots, \\
&\quad \phi_L \bar{\Lambda}_{LL}^2 (1-\bar{\Lambda}_{LL})^2 (e_L^T P_N \operatorname{diag}(P_N^T \bar{\Lambda} e_L)^{-1} P_N^T e_L + 1/(1-\bar{\Lambda}_{LL}))), \\
W_\gamma &= \operatorname{diag}(\phi_1 (1-\bar{\Lambda}_{11})^2 \Sigma_1^T \operatorname{diag}(P_N^T \bar{\Lambda} e_1)^{-1} \Sigma_1, \ldots, \phi_L (1-\bar{\Lambda}_{LL})^2 \Sigma_L^T \operatorname{diag}(P_N^T \bar{\Lambda} e_L)^{-1} \Sigma_L), \\
W_{\bar{\gamma}} &= \operatorname{diag}\left(\phi_{L+1} \bar{\Sigma}_1^T \operatorname{diag}(I_{L-1}, 0)(\operatorname{diag}(P_D^T e_1)^{-1} + (1/(P_D^T)_{1L}) \mathbf{1}\mathbf{1}^T) \operatorname{diag}(I_{L-1}, 0) \bar{\Sigma}_1, \ldots, \right. \\
&\quad \left. \phi_{2L} \bar{\Sigma}_L^T \operatorname{diag}(I_{L-1}, 0)(\operatorname{diag}(P_D^T e_L)^{-1} + (1/(P_D^T)_{LL}) \mathbf{1}\mathbf{1}^T) \operatorname{diag}(I_{L-1}, 0) \bar{\Sigma}_L\right),
\end{aligned}
$$

with $\phi_j = \sum_{k'=1}^{L} \sum_{t=1}^{T} \pi_{k'}^{(ctrl)}(\theta)(P^{t-1})_{k'j}$, $j = 1, \ldots, 2L$, *and*

$$
\bar{\Sigma}_j = \begin{bmatrix} \breve{\Sigma}_j \\ -\mathbf{1}^T \breve{\Sigma}_j \end{bmatrix} \in \mathbb{R}^{L \times (L-1)}, \quad \breve{\Sigma}_j = \operatorname{diag}(I_{L-1}, 0)(\operatorname{diag}(P_D^T e_j) - (P_D^T e_j)(P_D^T e_j)^T) \operatorname{diag}(I_{L-1}, 0),\ j = 1, \ldots, L.
$$

The proof can be found in the Appendix.

Lemma 2. *Let* G_k, F_k, $\Sigma_j, j = 0, 1, \ldots, L$, $\bar{\Sigma}_j, j = 1, \ldots, L$, *be those defined in Lemma 1. Then*

$$
I_{case}(\theta) = \begin{bmatrix} H_\lambda D_2 H_\lambda^T & H_\lambda D_2 H_{\gamma_0}^T & H_\lambda D_2 H_\gamma^T & H_\lambda D_2 H_{\bar{\gamma}}^T \\ H_{\gamma_0} D_2 H_\lambda^T & H_{\gamma_0} D_2 H_{\gamma_0}^T & H_{\gamma_0} D_2 H_\gamma^T & H_{\gamma_0} D_2 H_{\bar{\gamma}}^T \\ H_\gamma D_2 H_\lambda^T & H_\gamma D_2 H_{\gamma_0}^T & H_\gamma D_2 H_\gamma^T & H_\gamma D_2 H_{\bar{\gamma}}^T \\ H_{\bar{\gamma}} D_2 H_\lambda^T & H_{\bar{\gamma}} D_2 H_{\gamma_0}^T & H_{\bar{\gamma}} D_2 H_\gamma^T & H_{\bar{\gamma}} D_2 H_{\bar{\gamma}}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & U_{\bar{\gamma}} \end{bmatrix} =: V_{21} + V_{22}, \tag{15}
$$

where

$$
\begin{aligned}
D_2 &= \operatorname{diag}(\pi^{(case)})^{-1} - 2 \operatorname{diag}\left(\operatorname{diag}(\pi_{\cdot 1}^{(case)})^{-1} \tilde{\pi}_{\cdot 1}^{(case)} (\tilde{\pi}_{\cdot 1}^{(case)})^T, \ldots, \operatorname{diag}(\pi_{\cdot t_0}^{(case)})^{-1} \tilde{\pi}_{\cdot t_0}^{(case)} (\tilde{\pi}_{\cdot t_0}^{(case)})^T\right) \\
&\quad + \operatorname{diag}\left((\mathbf{1}^T \operatorname{diag}(\pi_{\cdot 1}^{(case)})^{-1} \mathbf{1}) \pi_{\cdot 1}^{(case)} (\pi_{\cdot 1}^{(case)})^T, \ldots, (\mathbf{1}^T \operatorname{diag}(\pi_{\cdot t_0}^{(case)})^{-1} \mathbf{1}) \pi_{\cdot t_0}^{(case)} (\pi_{\cdot t_0}^{(case)})^T\right)
\end{aligned}
$$

for $\pi^{(case)} = ((\pi_{\cdot 1}^{(case)})^T, \ldots, (\pi_{\cdot t_0}^{(case)})^T)^T$, *with*

$$
\begin{aligned}
\pi_{\cdot t}^{(case)} &= (\pi_{1t}^{(case)}, \ldots, \pi_{Lt}^{(case)})^T, \qquad \tilde{\pi}_{\cdot t}^{(case)} = ((\pi_{1t}^{(case)})^{1/2}, \ldots, (\pi_{Lt}^{(case)})^{1/2})^T, \\
\pi_{jt}^{(case)} &= \frac{(\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t-1} \Lambda P_D^{t_0 - t} e_j}{\sum_{t'=1}^{t_0} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t'-1} \Lambda \mathbf{1}}, \quad j = 1, \ldots, L,\ t = 1, \ldots, t_0,
\end{aligned}
$$

and

$$
\begin{aligned}
H_\lambda &= \frac{1}{\sum_{t'=1}^{t_0} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t'-1} \Lambda \mathbf{1}} \left[H_\lambda^{(1)}, \ldots, H_\lambda^{(t_0)}\right], \qquad H_{\gamma_0} = \frac{1}{\sum_{t'=1}^{t_0} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t'-1} \Lambda \mathbf{1}} \left[H_{\gamma_0}^{(1)}, \ldots, H_{\gamma_0}^{(t_0)}\right], \\
H_\gamma &= \frac{1}{\sum_{t'=1}^{t_0} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t'-1} \Lambda \mathbf{1}} \left[H_\gamma^{(1)}, \ldots, H_\gamma^{(t_0)}\right], \qquad H_{\bar{\gamma}} = \frac{1}{\sum_{t'=1}^{t_0} (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t'-1} \Lambda \mathbf{1}} \left[H_{\bar{\gamma}}^{(1)}, \ldots, H_{\bar{\gamma}}^{(t_0)}\right],
\end{aligned}
$$

with

$$
\begin{aligned}
H_\lambda^{(t)} &= \bar{\Lambda}(\bar{\Lambda} - I) \left(G_{t-1}(\pi_N^{(0)}, \bar{\Lambda} P_N, P_N) \Lambda - \operatorname{diag}((\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t-1})\right) P_D^{t_0 - t}, \\
H_{\gamma_0}^{(t)} &= \Sigma_0^T (\bar{\Lambda} P_N)^{t-1} \Lambda P_D^{t_0 - t}, \\
H_\gamma^{(t)} &= \operatorname{diag}(\Sigma_1^T, \ldots, \Sigma_L^T) F_{t-1}(\pi_N^{(0)}, \bar{\Lambda} P_N, \bar{\Lambda}) \Lambda P_D^{t_0 - t}, \\
H_{\bar{\gamma}}^{(t)} &= \operatorname{diag}(\bar{\Sigma}_1^T, \ldots, \bar{\Sigma}_L^T) F_{t_0 - t}(\Lambda (P_N^T \bar{\Lambda})^{t-1} \pi_N^{(0)}, P_D, I), \quad t = 1, \ldots, t_0.
\end{aligned}
$$

Finally,

$$U_{\tilde{\gamma}} = \operatorname{diag}\left(\psi_1 \bar{\Sigma}_1^T \operatorname{diag}(I_{L-1}, 0)(\operatorname{diag}(P_D^T e_1)^{-1} + (1/(P_D^T)_{1L})\mathbf{1}\mathbf{1}^T)\operatorname{diag}(I_{L-1}, 0)\bar{\Sigma}_1, \ldots,\right.$$
$$\left.\psi_L \bar{\Sigma}_L^T \operatorname{diag}(I_{L-1}, 0)(\operatorname{diag}(P_D^T e_L)^{-1} + (1/(P_D^T)_{LL})\mathbf{1}\mathbf{1}^T)\operatorname{diag}(I_{L-1}, 0)\bar{\Sigma}_L\right),$$

with $\psi_j = \sum_{k'=1}^{L}\sum_{t=1}^{T}\sum_{\tau=1}^{t_0}\pi_{k'\tau}^{(case)}(\theta)(P_D^{t-1})_{k'j}$, $j = 1, \ldots, L$.

The proof can be found in the Appendix.

We are ready to establish an asymptotic normality of the estimate $\hat{\theta}$:

Theorem 1. *Assume that θ^0, the value of θ with which the data of the study is generated, is bounded, and $I_{ctrl}(\theta)$ and $I_{case}(\theta)$, given in (14) and (15), are full rank at $\theta = \theta^0$. Then $\hat{\theta}$, which maximizes the log likelihood (8), satisfies*

$$\sqrt{n+m}(\hat{\theta} - \theta^0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}),$$

as both n and m tend to infinity, while $n/(n+m) \to \rho_1 > 0$ and $m/(n+m) \to \rho_2 > 0$ with $\rho_1 + \rho_2 = 1$, and $\mathcal{I} = \rho_1 I_{ctrl}(\theta^0) + \rho_2 I_{case}(\theta^0)$.

Proof. The conclusion follows immediately from Lemmas 1 and 2, and the argument leading to (13). □

Remark 1. The first terms V_{11} and V_{21} of (14) and (15) are due to the cross-sectional part of the design, whereas the second terms V_{12} and V_{22} are due to the short-term follow-up part. The 3×3 block principal submatrices of these terms are associated with the conditional transition probabilities *before* disease on-set, whereas the last diagonal block with those *after* disease on-set. Thus in the control group, measurements before and after disease on-set are uncorrelated with each other, which is expected; in the case group, follow-up does not provide any information on the pre-disease period, and this is due to the Markovian assumption.

Remark 2. The prospective cohort counterpart of this design contains only the control group, starting with $t = 0$. Thus for this design we have $V = V_{12}$ with sample size of $n + m$.

Remark 3. The assumption that θ^0 is bounded also ensures that each term of the log likelihood is non-degenerate, because positivity of the transition probabilities implies that each $n_j, m_{jt} > 0$ almost surely.

4. Numerical studies

4.1. Large-sample asymptotic efficiency

In order to see how the cross-sectional study design with a short-term follow-up has a merit over the prospective cohort study design, we compare their large-sample asymptotic variances of the estimated log odds ratio by using the 2×2-state model (1)–(4) and the formula in Theorem 1 for a variety of scenarios. In particular, we consider null cases in which the incidence rates $1 - \alpha$ and $1 - \beta$ are the same regardless of the biomarker level, i.e. the biomarker is non-prognostic, and non-null cases in which it is prognostic with the odds ratio (5) being approximately 2.0 to the first digit below the decimal point. We use four different combinations of α and β for each set of cases. We set $t_0 = 5$ and compute the large-sample asymptotic standard deviation of the maximum likelihood estimate of the log odds ratio for $T = 2, 3, 4, 5$. We also vary ρ_2, the ratio of the case sample size to the total sample size, from 0.1 to 0.9 with an increment of 0.1. The ratio $\rho_2 = 0$ corresponds to the prospective design started at $t = t_0$, and $\rho_2 = 1$, the case-only scenario, is non-identifiable because of the assumed irreversibility of the disease. Both extremes are ruled out. The transition matrices before and after the disease on-set are set as

$$P_N = \begin{matrix} \\ \text{LN} \\ \text{HN} \end{matrix}\begin{matrix} \begin{matrix} \text{LN} & \text{HN} \end{matrix} \\ \begin{bmatrix} 0.85 & 0.15 \\ 0.05 & 0.95 \end{bmatrix} \end{matrix}, \qquad P_D = \begin{matrix} \\ \text{LD} \\ \text{HD} \end{matrix}\begin{matrix} \begin{matrix} \text{LD} & \text{HD} \end{matrix} \\ \begin{bmatrix} 0.90 & 0.10 \\ 0.05 & 0.95 \end{bmatrix} \end{matrix}$$

to incorporate an accelerated disease progression, e.g., the loss of bone mineral, after the disease on-set. These matrices are fixed throughout the scenarios. The initial distribution is set to $\pi_N^{(0)} = (0.5, 0.5)^T$.

In the null cases we use the combinations of incidence rates $(1 - \alpha, 1 - \beta) = (0.001, 0.001)$, $(0.010, 0.010)$, $(0.020, 0.020)$, and $(0.050, 0.050)$. The computed large-sample standard deviations are plotted in Fig. 3. No normalization by the sample size is applied. With small to moderate incidence rates (0.1%–1.0%) the cross-sectional design with two to three follow-ups suggests a definite merit over the prospective cohort design for reasonably balanced case-control sample sizes, say $\rho_2 = 0.5$. For the medium rate of 2.0% the new design becomes comparable to the prospective one, whereas for higher rates it loses its competitive edge.

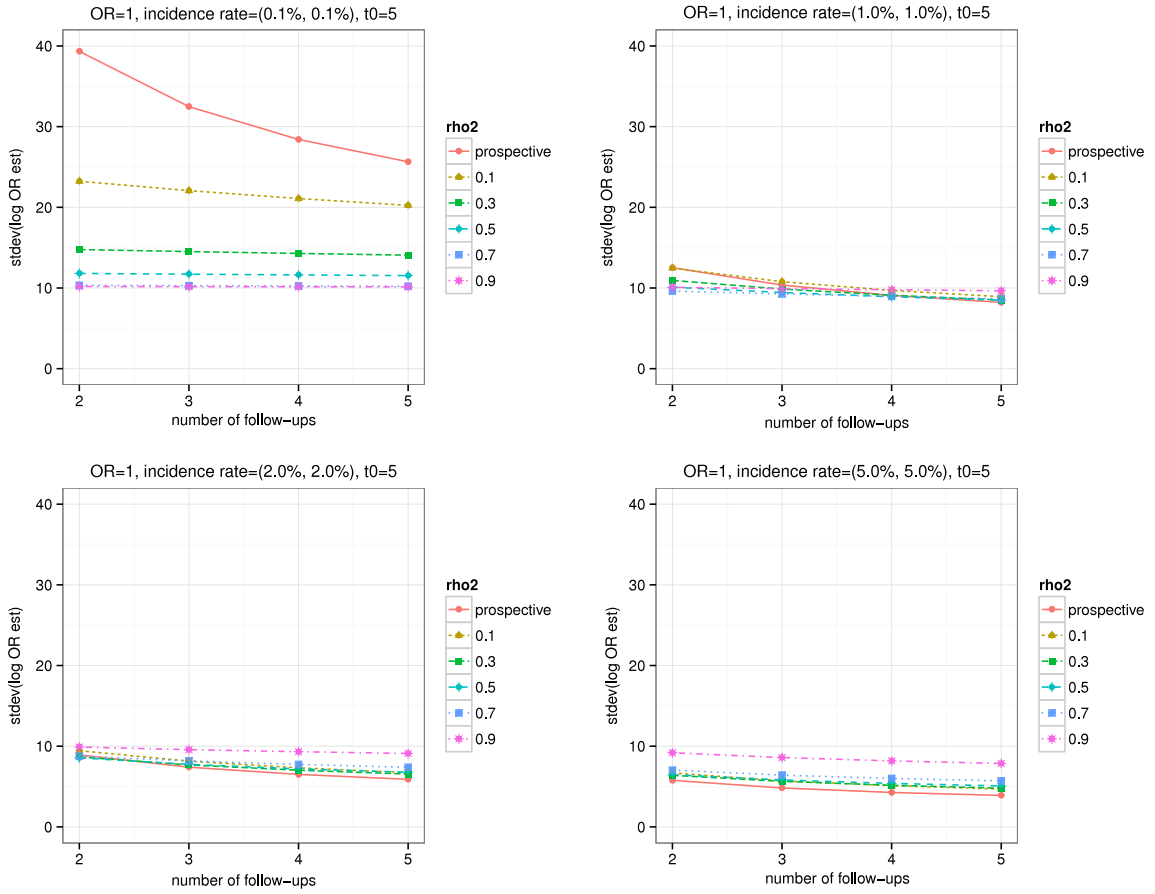

Fig. 3. Comparison of the large-sample asymptotic standard deviation of the estimated log odds ratio for the cross-sectional design with a short-term follow-up with various case/control sample size ratios and for the prospective cohort design. Null cases (odds ratio = 1). Note that the standard deviations are *not* normalized by the sample size.

In the non-null cases we use the incidence rate combinations $(1-\alpha, 1-\beta) = (0.005, 0.010)$, $(0.010, 0.020)$, $(0.020, 0.040)$, and $(0.050, 0.100)$. Note the odds ratio is roughly 2.0 for each combination. The large-sample asymptotic standard deviations are computed and plotted in Fig. 4. Again, no normalization by the sample size is applied. The relative asymptotic efficiency of the new design is retained in even higher incidence rates of 2.0% and 4.0%, up to four follow-ups. Thus the new design exhibits a relatively higher gain in sensitivity than specificity.

4.2. Finite-sample performance

To see if the large-sample asymptotic efficiency of the new design is maintained in finite-sample settings, we conducted a simulation study comparing the standard errors of the cross-sectional design with a short-term follow-up and the prospective cohort study. Specifically, we generated samples from the Markov chain model (1)–(4), with the parameters set as in Section 4.1. To simulate the case-control design part, we conducted a rejection sampling, i.e., run as many chains until collecting m diseased subjects, and resample n normal (undiseased) subjects uniformly from the rejected samples. For the generated dataset, the model was fit by maximizing (10). This procedure was repeated 1000 times to obtain a sample standard deviation of the estimated log odds ratio. The log likelihood function (10) is highly nonlinear and nonconcave with $2L^2 - 1$ variables (seven for $L = 2$), thus is difficult to maximize globally using Newton's method. To find a solution close to the global maximizer, we used the GenSA package (Xiang et al., 2013), which shows the best performance in a comparative study by Mullen (2014). We obtained reasonably accurate estimates using a random initialization and the maximum number of iterations of 40; anything else is set default. The 1000 repetition of the maximum likelihood estimation took less than 0.2 s on average, per dataset per combination of incidence rates. The number of follow-ups T and the ratio of the case sample size to the total sample size ρ_2 were fixed at 3 and 0.5, respectively, and the total sample size $n + m$ tested were varied between 500, 1000, 2000, 3000, 4000, and 5000. Note that the same combinations of α and β are used as the large-sample computation.

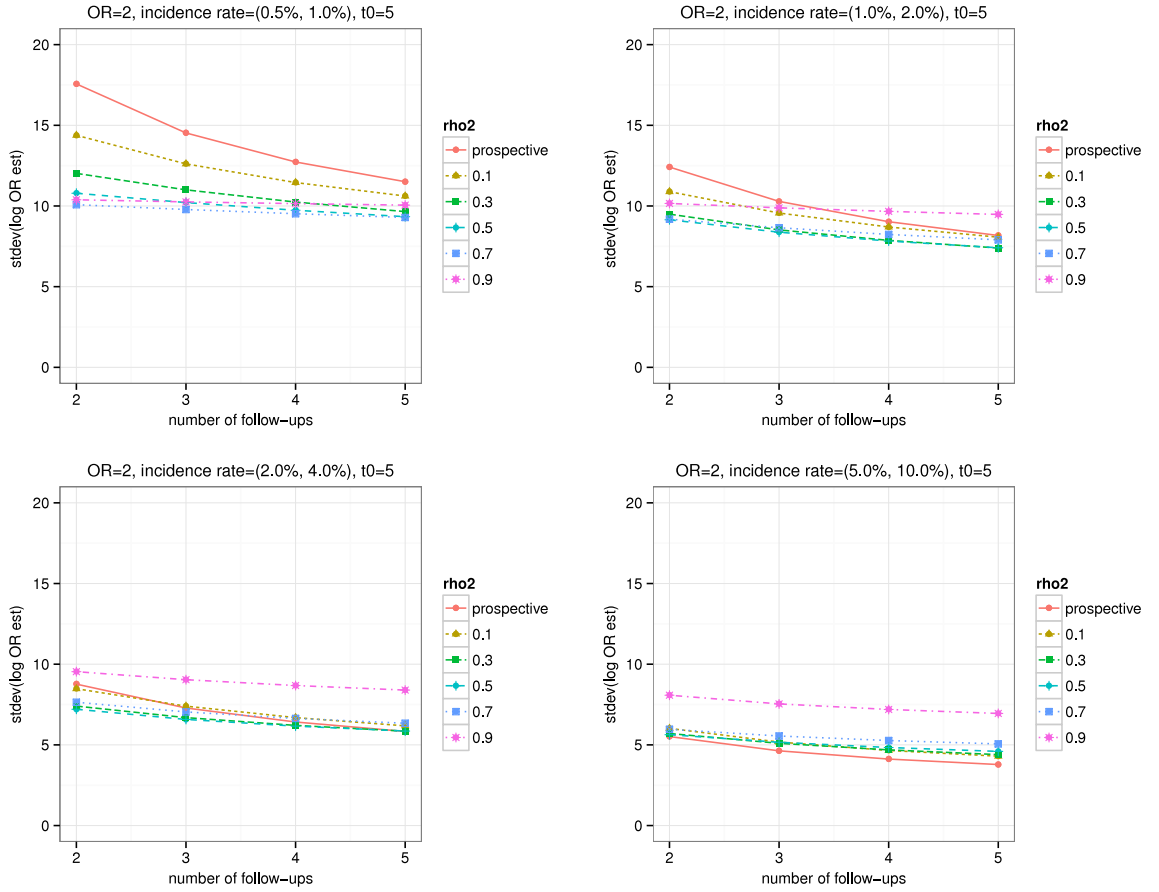

Fig. 4. Comparison of the large-sample asymptotic standard deviation of the estimated log odds ratio for the cross-sectional design with a short-term follow-up with various case/control sample size ratios and for the prospective cohort design. Non-null cases (odds ratio $\approx$ 2). Note that the standard deviations are *not* normalized by the sample size.

Figs. 5 and 6 illustrate that the relative efficiency of the new design is retained throughout the various sample sizes considered. The gap between the new design and the prospective cohort design decreases as the incidence rates increase, as expected. The large drop of standard deviation for small incidence rates in the prospective design when the sample size increases from 500 to 1000 appears to be due to that, at a small incidence rate, transition from the normal to the diseased state is rarely observed especially for these small sample sizes. More importantly, the standard deviation of the new design is smaller than the prospective one even for large incidence rates, unlike the asymptotic comparison in Section 4.1. Thus the large-sample variance formula due to Theorem 1 appears to be conservative.

5. Alzheimer's disease data example

In this section we illustrate an application of our new design to a dataset arising from a large-scale longitudinal study of AD, in order to assess the prognostic potential of an MRI-based biomarker. Note that, as our design is brand new, we are not aware of any existing study to which our design can be applied directly. Therefore we chose the Alzheimer's Disease Neuroimaging Initiative (ADNI), a large-scale MRI and fluorodeoxyglucose positron emission tomography (FDG-PET) study started in 2004, and sampled a subset of subjects to virtually conduct a cross-sectional study with a short-term follow-up. In neurodegeneration research, MCI is generally considered as the prodromal phase of AD, thus significant efforts have been made to find biomarkers predictive of conversion from MCI to AD. The hippocampus is one of the main targets as an imaging biomarker of AD, and MRI-based hippocampal volume measurements have shown great potential for this purpose (Schuff et al., 2009; Lee et al., 2016).

5.1. ADNI dataset

Data used in the preparation of this section were obtained from the ADNI database (http://adni.loni.usc.edu), currently in its third phase. The primary goal of the ADNI has been to test whether serial MRI, positron emission tomography (PET), other

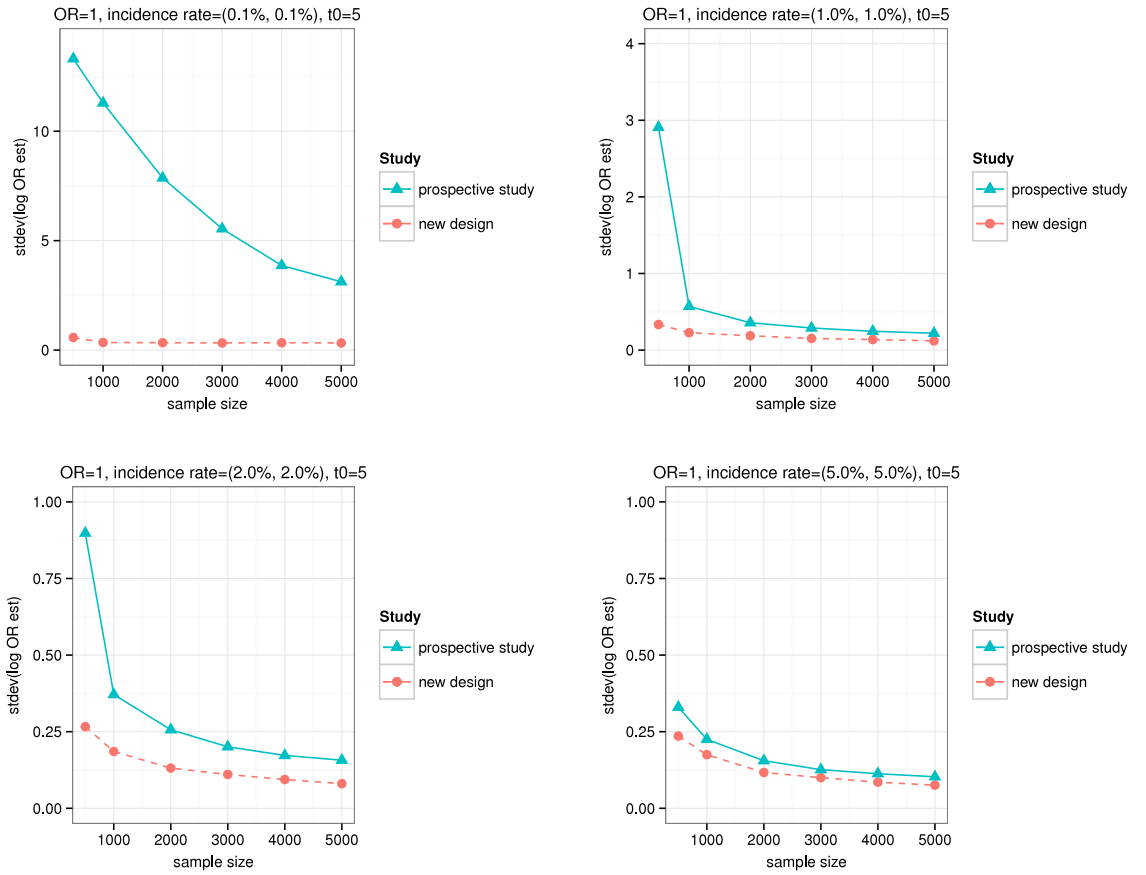

Fig. 5. Comparison of the finite-sample standard deviation of the estimated log odds ratio for the cross-sectional design with a short-term follow-up with the same case/control sample sizes ($\rho_2 = 0.5$) and three follow-ups ($T = 3$), and for the prospective cohort design. Null cases (odds ratio $= 1$).

biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific biomarkers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of the ADNI is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. For up-to-date information on the ADNI, see http://www.adni-info.org.

For our data example we used the "ADNI MERGE" dataset, which combines data from three ADNI funding cycles (ADNI 1, ADNI GO, and ADNI 2). This dataset includes at least three (up to seven) annual longitudinal measurements of the volume of the hippocampus derived from T1-weighted structural MRI scans (ADNI 1 used 1.5T scanners, ADNI GO and 2 used 3T), from 265 individuals initially diagnosed as either early or late MCI, as well as the baseline measurements of the hippocampal volume from 365 cognitively normal (CN) individuals. Among the MCI subjects, 37 converted to AD over the course of the follow-up, with at least two subsequent annual measurements available. We regarded three years before a subject's last measurement as the age at the beginning of the virtual cross-sectional study with a short-term follow-up for the subject. We chose 26 individuals with the age at this virtual, cross-sectional baseline greater than 69.0 and less than 81.0 (the age was measured in a continuous scale) from the 37 converters as the case; 22 were males and 4 were females. The mean and standard deviation of the age was 75.9 and 3.49, respectively. We then chose 26 individuals from the non-converters by age and gender matching, regarding the earliest age from which three consecutive annual measurements were available as the virtual baseline. The gender distribution was identical to the case; the mean and standard deviation of the age was 74.4 and 3.78, respectively.

5.2. Hippocampal volume measurement

The volume of hippocampus was measured based on a semi-automated brain mapping method. As initial guidance, trained operators manually identified 22 landmarks for hippocampal segmentation (five equally spaced image slices perpendicular to the long axis of the ipsilateral hippocampus from head to tail), then a high-dimensional fluid transformation

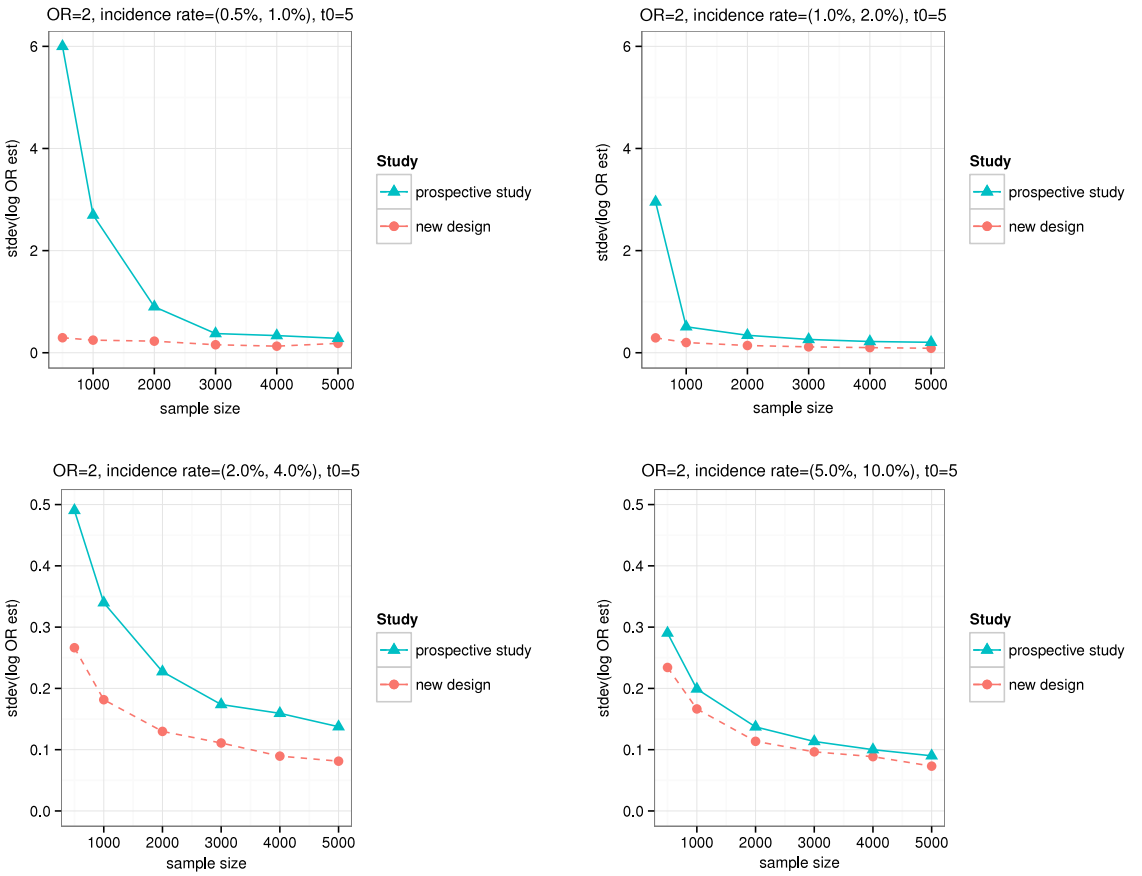

Fig. 6. Comparison of the finite-sample standard deviation of the estimated log odds ratio for the cross-sectional design with a short-term follow-up with the same case/control sample sizes ($\rho_2 = 0.5$) and three follow-ups ($T = 3$), and for the prospective cohort design. Non-null cases (odds ratio ≈ 2).

algorithm (Christensen et al., 1997) transformed a marked hippocampal MRI template from a reference brain to match the target images of each subject.

Because hippocampal volume (HV) tends to increase with the brain size, often measured by the intracranial volume (ICV), proper normalization of HV by ICV is an important task in evaluating this imaging biomarker. One of the most widely used methods in the literature is to regress the HV by ICV using the reference population and use the prediction residual to adjust for the volume in the data. Also in the ADNI data, it is known that the association between age and hippocampal volume is significant while that between gender and the volume is not (Voevodskaya et al., 2014). Thus we used a linear regression with ICV and age at scan used to adjust for HV, using the CN group as the reference population. To fit our model, we needed to discretize the adjusted HV. We chose to dichotomize it for simplicity: as the residuals from the CN group was distributed roughly from -4000 to $+2000$, we chose -1000 in the prediction residual as the cutoff to construct a binary marker. We may call this marker BHV (binarized hippocampal volume).

5.3. Results

For the case and control groups selected by the criteria described in the previous subsections, we applied our method described in Section 3 to estimate the odds ratio of the imaging biomarker BHV. In order to verify the Markovian assumption, we conducted a chi-square test on the two-year transition counts involved with the follow-up ($p > 0.456$). Thus the Markov chain approach in our method appears reasonable in this data. We set $t_0 = 10$, rendering the population of interest the MCI group at age between 60 and 70; $m = n = 26$, $T = 3$ and $L = 2$ by construction. We maximized the log likelihood (10) in the same fashion as Section 4.2 to obtain the MLE of the log odds ratio (OR) of the BHV, and computed the asymptotic standard deviation of the estimated log OR by using the formula in Theorem 1, with $\rho_1 = \rho_2 = 0.5$. The estimated log OR was 2.73, with estimated standard deviation of 0.684. In order to see the merit of the new design, we sampled an addition subset of 26 age and gender matched control subjects, and counted state transitions in the group of total 52 controls. No sufficient transitions from normal (MCI) to diseased (AD) for calculating the log OR were observed; the asymptotic

formula from Theorem 1 gives a standard deviation of 1.69 if the parameters were the same as those obtained from the new design, and predicts that more than 20 visits are required to achieve the same level of standard deviation as the new design.

The log OR of 2.73 ± 0.684 is optimistic because of the selection bias of the virtual study as well as the interval censoring. Nonetheless it is illustrative to demonstrate that temporal trend in hippocampal volume has a potential as a prognostic biomarker for AD, which has been confirmed in the literature (Rosen et al., 1984; Jack et al., 1999; Schuff et al., 2009). The cutoff for dichotomization is somewhat arbitrary, although the estimated log OR is not much sensitive to this choice: when the cutoff was lowered to -1500, it became 2.34 ± 0.824. If the proposed design is implemented, it may provide a more reasonable figure, with significantly reduced time and efforts compared to large-scale imaging biomarker studies such as the ADNI.

6. Conclusions

We have proposed and studied the properties of a new statistical study design that extends the traditional prospective cohort study design by adding cases with known time of events and including a short-term follow-up, in order to estimate the prospective odds ratio of a clinical endpoint of interests within a reasonably short duration of time. This kind of design has been called for because of the lack of historical samples and the demand for a short duration of studies originating from the rapid pace of evolution in prognostic imaging biomarkers. A temporal association need to be established to show an association between its baseline value and the subsequent chance of the relevant clinical outcome in order to validate the prognostic utility of markers. We have employed a Markov chain model to establish the necessary temporal association, and showed that under this model, our new design yields a consistent estimate of odds ratio, and can induce a smaller variance than from the corresponding prospective cohort design within a short duration for follow-ups when the incidence rate of the outcome is low (say, below 5%–10%): an asymptotically normal variance formula has been derived, and it has been demonstrated that in most reasonable scenarios a follow-up with at most three visits is sufficient to obtain a reasonable and more efficient estimate than the prospective design. Thus our design has a potential to be useful to rapidly establish prognostic validity of a new imaging biomarkers within a reasonable time and with a smaller sample sizes for rare clinical endpoints; we have demonstrated this through an application to the ADNI data. In this sense, it is a cost-effective study design. While our design is devised having non-invasive imaging biomarkers primarily in mind, it can be generalized to other biomarkers sharing the key features of no historical samples and rapid pace of development, e.g., mobile biomarkers.

The time-homogeneous discrete-time Markov chain model that we have employed can be an oversimplification. For example, subject-specific effects are difficult to model with the Markovian assumption. While we are aware of these limitations of our model, it is a first-order approximation that can provide a reasonable initial estimation of odds ratios within manageable time and size of the study for new imaging biomarkers. Naturally, an elaboration of the temporal association model would be an intriguing future work. For instance, the assumption of homogeneous transitions can be lifted to incorporate the effect of aging, and the discrete-time model can be extended to continuous-time models. Nevertheless the discrete-time aspect of our model is common in research studies as most often, the visits are scheduled in time intervals. Although several changes of risk level may occur between follow-ups, we can minimize such possibilities by assuming slow progression of the disease, which is true in AD and osteoporosis, and by an appropriate choice of time window in the implementation of the design. Including subject-specific random effects via non-Markov models, e.g., semi-Markov models, will be another interesting outlet for an extension of the Markov model. In any case, complete temporal modeling without any assumption will not be possible due to identifiability problem, unless a prospective design is used.

When we discretize continuous measurements, the cut-off values may change the number of subjects in each risk level. While this may affect our design efficiency, such as how rare the high-risk patients are in the population, the main driving factor is the event (disease) rate, which does not depend on the levels of risk. Discretization equally affects prospective designs, and we seek for a relative merit over these designs rather than absolute superiority; for the continuous marker considered in our AD example, we dichotomized the marker levels and found the impact of the binary cutoff value to the final odds ratio estimate was minimal.

With further improvement of the model, e.g., for continuous levels and with age- or subject-dependent transition rates, we are optimistic about the adaption of our design in future projects.

Acknowledgments

A preliminary result for this research is presented in the Statistical Evaluation of Medical Diagnostic Tests session of the 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics), 12–14 December 2015, London, United Kingdom. Joong-Ho Won's research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP, Nos. 2013R1A1A1057949 and 2014R1A4A1007895).

Appendix. Proofs

A.1. Proof of Lemma 1

Note

$$
\begin{aligned}
nI_{ctrl}(\theta) &= E_\theta\left[\left(\frac{\partial l_{ctrl}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}}{\partial\theta}\right)^T\right] \\
&= E_\theta\left[\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial\theta}\right)^T\right] + E_\theta\left[\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial\theta}\right)^T\right] \\
&\quad + E_\theta\left[\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial\theta}\right)^T\right] + E_\theta\left[\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial\theta}\right)^T\right]. \qquad (16)
\end{aligned}
$$

The first term in (16) is the information matrix of n observations of the multinomial of L cells with cell probability $\pi^{(ctrl)} = (\pi_1^{(ctrl)}, \ldots, \pi_L^{(ctrl)})^T$, where

$$\pi_j^{(ctrl)} = \pi_j^{(ctrl)}(\theta) = \frac{f(\theta)e_j}{f(\theta)\mathbf{1}}, \quad j = 1, \ldots, L,$$

for

$$f(\theta) = (\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t_0}.$$

Therefore

$$E_\theta\left[\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial\theta}\right)^T\right] = n\left(\frac{\partial\pi^{(ctrl)}}{\partial\theta}\right)\mathrm{diag}(\pi^{(ctrl)})^{-1}\left(\frac{\partial\pi^{(ctrl)}}{\partial\theta}\right)^T.$$

In order to evaluate $\partial\pi_j^{(ctrl)}/\partial\theta$, we first evaluate the matrix differential of $f(\theta)$ to see

$$df = (d\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t_0} + \sum_{l=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^l(d\bar{\Lambda})P_N(\bar{\Lambda}P_N)^{t_0-1-l} + \sum_{l=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^l\bar{\Lambda}(dP_N)(\bar{\Lambda}P_N)^{t_0-1-l}$$

after some algebra. In particular, we use the fact $d\bar{\Lambda} = -d\Lambda$ and

$$d(A^n) = (dA)A^{n-1} + A(dA)A^{n-2} + \cdots + A^{n-1}(dA) = \sum_{j=0}^{n-1}A^j(dA)A^{n-j-1}$$

for a square matrix A (we interpret $A^0 = I$). Now note that

$$d\bar{\Lambda} = -\mathrm{diag}\left(\bar{\Lambda}_{11}(1-\bar{\Lambda}_{11})d\lambda_1, \ldots, \bar{\Lambda}_{pp}(1-\bar{\Lambda}_{pp})d\lambda_p\right) = \sum_{j=1}^{L}\bar{\Lambda}_{jj}(\bar{\Lambda}_{jj}-1)e_je_j^Td\lambda_j, \qquad (17)$$

$$d\pi_N^{(0)} = \Sigma_0 d\gamma_0, \qquad (18)$$

$$(dP_N^T)_j = \Sigma_j d\gamma_j, \quad j = 1, \ldots, L, \qquad (19)$$

$$(dP_D^T)_j = \bar{\Sigma}_j d\bar{\gamma}_j, \qquad (20)$$

with $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0,L-1})^T$, $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{j,L-1})^T$. Then in the evaluation of df,

$$(d\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t_0} = d\gamma_0^T\Sigma_0^T(\bar{\Lambda}P_N)^{t_0} \qquad (21)$$

and

$$
\begin{aligned}
\sum_{l=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^l d\bar{\Lambda}P_N(\bar{\Lambda}P_N)^{t_0-1-l} &= \sum_{l=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^l\left(\sum_{j=1}^{L}\bar{\Lambda}_{jj}(\bar{\Lambda}_{jj}-1)e_je_j^Td\lambda_j\right)P_N(\bar{\Lambda}P_N)^{t_0-1-l} \\
&= \sum_{j=1}^{L}d\lambda_j\bar{\Lambda}_{jj}(\bar{\Lambda}_{jj}-1)\sum_{l=0}^{t_0-1}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^le_je_j^TP_N(\bar{\Lambda}P_N)^{t_0-1-l} \\
&= d\lambda^T\bar{\Lambda}(\bar{\Lambda}-I)G_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, P_N)x, \qquad (22)
\end{aligned}
$$

where

$$G_k(a,B,C)=\begin{bmatrix}\sum_{l=0}^{k-1}a^TB^l(e_1e_1^TC)B^{k-1-l}\\ \vdots\\ \sum_{l=0}^{k-1}a^TB^l(e_Le_L^TC)B^{k-1-l}\end{bmatrix}=\sum_{l=0}^{k-1}\mathrm{diag}(a^TB^l)CB^{k-1-l},$$

for $a\in\mathbb{R}^L,\ B,C\in\mathbb{R}^{L\times L}$, and for any $x\in\mathbb{R}^L$,

$$\begin{aligned}\sum_{l=0}^{t_0-1}(\pi_N^{(0)})^T(\bar\Lambda P_N)^l\bar\Lambda(dP_N)(\bar\Lambda P_N)^{t_0-1-l}x&=\sum_{l=0}^{t_0-1}a_l^T(dP_N)b_l\quad(a_l^T=\bar\Lambda(P_N^T\bar\Lambda)^l\pi_N^{(0)},\ b_l=(\bar\Lambda P_N)^{t_0-1-l}x)\\&=\sum_{l=0}^{t_0-1}\mathrm{vec}(b_l^T(dP_N^T)a_l)\\&=\sum_{l=0}^{t_0-1}(a_l^T\otimes b_l^T)\,\mathrm{vec}(dP_N^T)\\&=\sum_{l=0}^{t_0-1}(a_l^T\otimes b_l^T)\sum_{j=1}^{L}\mathrm{vec}(\Sigma_j(d\Gamma)e_je_j^T)\quad(\Gamma=\mathrm{mat}(\gamma))\qquad(23)\\&=\sum_{l=0}^{t_0-1}\sum_{j=1}^{L}\mathrm{vec}(b_l^T\Sigma_j(d\Gamma)e_je_j^Ta_l)\\&=\sum_{l=0}^{t_0-1}\sum_{j=1}^{L}\left((a_l^Te_je_j^T)\otimes(b_l^T\Sigma_j)\right)\mathrm{vec}(d\Gamma)\\&=\sum_{l=0}^{t_0-1}d\gamma^T\sum_{j=1}^{L}\left((e_je_j^Ta_l)\otimes(\Sigma_j^Tb_l)\right)\\&=d\gamma^T\left(\sum_{j=1}^{L}(e_je_j^T\otimes\Sigma_j^T)\sum_{l=0}^{t_0-1}(a_l\otimes b_l)\right)\qquad(24)\\&=d\gamma^T\,\mathrm{diag}(\Sigma_1^T,\dots,\Sigma_L^T)\sum_{l=0}^{t_0-1}\left((\bar\Lambda(P_N^T\bar\Lambda)^l\pi_N^{(0)}\cdot1)\otimes((\bar\Lambda P_N)^{t_0-1-l}x)\right)\\&=d\gamma^T\,\mathrm{diag}(\Sigma_1^T,\dots,\Sigma_L^T)\sum_{l=0}^{t_0-1}\left((\bar\Lambda(P_N^T\bar\Lambda)^l\pi_N^{(0)})\otimes(\bar\Lambda P_N)^{t_0-1-l}\right)x\qquad(25)\\&=d\gamma^T\,\mathrm{diag}(\Sigma_1^T,\dots,\Sigma_L^T)F_{t_0}(\pi_N^{(0)},\bar\Lambda P_N,\bar\Lambda)x,\qquad(26)\end{aligned}$$

where

$$F_k(a,B,C)=\sum_{l=0}^{k-1}\left((a^TB^lC)^T\otimes B^{k-1-l}\right)=\begin{bmatrix}\sum_{l=0}^{k-1}a^TB^lCe_1B^{k-1-l}\\ \vdots\\ \sum_{l=0}^{k-1}a^TB^lCe_LB^{k-1-l}\end{bmatrix}\in\mathbb{R}^{L^2\times L},$$

for $a\in\mathbb{R}^L,\ B,C\in\mathbb{R}^{L\times L}$. In the derivation of (26), identity

$$\mathrm{vec}(AXB)=(B^T\otimes A)\,\mathrm{vec}(X)$$

is used multiple times. Eq. (23) is due to

$$\mathrm{vec}(dP_N^T)=\begin{bmatrix}\Sigma_1d\gamma_1\\ \vdots\\ \Sigma_Ld\gamma_L\end{bmatrix}=\mathrm{diag}(\Sigma_1,\dots,\Sigma_L)d\gamma=\sum_{j=1}^{L}((e_je_j^T)\otimes\Sigma_j)\,\mathrm{vec}(d\Gamma)=\sum_{j=1}^{L}\mathrm{vec}(\Sigma_j(d\Gamma)e_je_j^T).$$

In (24) and (25), identity

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

is used. Collecting (21), (22), and (26), we see

$$df = [d\lambda^T, d\gamma_0^T, d\gamma^T] \begin{bmatrix} \bar{\Lambda}(\bar{\Lambda} - I) & & \\ & \Sigma_0^T & \\ & & \operatorname{diag}(\Sigma_1^T, \ldots, \Sigma_L^T) \end{bmatrix} \begin{bmatrix} G_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, P_N) \\ (\bar{\Lambda}P_N)^{t_0} \\ F_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, \bar{\Lambda}) \end{bmatrix}, \tag{27}$$

or

$$\frac{\partial f(\theta)}{\partial \theta} = \begin{bmatrix} \bar{\Lambda}(\bar{\Lambda} - I)G_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, P_N) \\ \Sigma_0^T(\bar{\Lambda}P_N)^{t_0} \\ \operatorname{diag}(\Sigma_1^T, \ldots, \Sigma_L^T)F_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, \bar{\Lambda}) \\ 0 \end{bmatrix}.$$

It follows that

$$\begin{aligned} \frac{\partial \pi_j^{(ctrl)}}{\partial \theta} &= \frac{(f(\theta)\mathbf{1})(\partial f(\theta)/\partial\theta)e_j - (f(\theta)e_j)(\partial f(\theta)/\partial\theta)\mathbf{1}}{(f(\theta)\mathbf{1})^2} \\ &= \frac{1}{(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t_0}\mathbf{1}} \begin{bmatrix} \bar{\Lambda}(\bar{\Lambda} - I)G_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, P_N)e_j \\ \Sigma_0^T(\bar{\Lambda}P_N)^{t_0}e_j \\ \operatorname{diag}(\Sigma_1^T, \ldots, \Sigma_L^T)F_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, \bar{\Lambda})e_j \\ 0 \end{bmatrix} \\ &\quad - \frac{(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t_0}e_j}{((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t_0}\mathbf{1})^2} \begin{bmatrix} \bar{\Lambda}(\bar{\Lambda} - I)G_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, P_N)\mathbf{1} \\ \Sigma_0^T(\bar{\Lambda}P_N)^{t_0}\mathbf{1} \\ \operatorname{diag}(\Sigma_1^T, \ldots, \Sigma_L^T)F_{t_0}(\pi_N^{(0)}, \bar{\Lambda}P_N, \bar{\Lambda})\mathbf{1} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} K_\lambda \\ K_{\gamma_0} \\ K_\gamma \\ 0 \end{bmatrix} e_j - \pi_j^{(ctrl)} \begin{bmatrix} K_\lambda \\ K_{\gamma_0} \\ K_\gamma \\ 0 \end{bmatrix} \mathbf{1} = Ke_j - \pi_j^{(ctrl)}K\mathbf{1} \end{aligned}$$

for $j = 1, \ldots, L$. Thus

$$\frac{\partial \pi^{(ctrl)}}{\partial \theta} = K - K\mathbf{1}(\pi^{(ctrl)})^T.$$

Let

$$J_1 = \frac{\partial \pi^{(ctrl)}}{\partial \theta} \operatorname{diag}(\pi^{(ctrl)})^{-1/2} = K \operatorname{diag}(\pi^{(ctrl)})^{-1/2} - K\mathbf{1}(\tilde{\pi}^{(ctrl)})^T,$$

where $\tilde{\pi}^{(ctrl)} = ((\pi_1^{(ctrl)})^{1/2}, \ldots, (\pi_L^{(ctrl)})^{1/2})^T$. Then,

$$\begin{aligned} \left(\frac{\partial \pi^{(ctrl)}}{\partial \theta}\right) \operatorname{diag}(\pi^{(ctrl)})^{-1} \left(\frac{\partial \pi^{(ctrl)}}{\partial \theta}\right)^T &= J_1 J_1^T \\ &= K(\operatorname{diag}(\pi^{(ctrl)})^{-1} - \operatorname{diag}(\pi^{(ctrl)})^{-1/2}\tilde{\pi}^{(ctrl)}\mathbf{1}^T \\ &\quad - \mathbf{1}(\tilde{\pi}^{(ctrl)})^T \operatorname{diag}(\pi^{(ctrl)})^{-1/2} + \mathbf{1}(\tilde{\pi}^{(ctrl)})^T\tilde{\pi}^{(ctrl)}\mathbf{1}^T)K^T \\ &= K(\operatorname{diag}(\pi^{(ctrl)})^{-1} - \mathbf{1}\mathbf{1}^T)K^T \\ &= \begin{bmatrix} K_\lambda \\ K_{\gamma_0} \\ K_\gamma \\ 0 \end{bmatrix} D_1 \begin{bmatrix} K_\lambda^T & K_{\gamma_0}^T & K_\gamma^T & 0 \end{bmatrix} = V_{11}. \end{aligned}$$

The second term in (16) is the information matrix of n observations of the Markov chain with transition matrix $P = P(\theta)$ for T steps, conditioning on the initial state frequencies of $n_1, \ldots, n_L$. Thus

$$E_\theta \left[\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial \theta}\right) \left(\frac{\partial l_{ctrl}^{(FU)}}{\partial \theta}\right)^T \right] = \left(\frac{\partial \operatorname{vec}(P^T)}{\partial \theta}\right) nI_{ctrl}(P) \left(\frac{\partial \operatorname{vec}(P^T)}{\partial \theta}\right)^T,$$

where $I_{ctrl}(P)$ is a $4L^2 \times 4L^2$ matrix so that

$$\begin{aligned} n(I_{ctrl}(P))_{jk,gh} &= E_\theta\left[E_\theta\left[\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial p_{jk}}\right)\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial p_{gh}}\right)\,\middle|\, n_1,\ldots,n_L\right]\right] \\ &= E_\theta\left[E_\theta\left[\left(\frac{\nu_{jk}-\nu_j p_{jk}}{p_{jk}} - \frac{\nu_{j,2L}-\nu_j p_{j,2L}}{p_{j,2L}}\right)\left(\frac{\nu_{gh}-\nu_g p_{gh}}{p_{gh}} - \frac{\nu_{g,2L}-\nu_j p_{g,2L}}{p_{g,2L}}\right)\,\middle|\, n_1,\ldots,n_L\right]\right] \\ &= \begin{cases} n\phi_j(1/p_{jk} + 1/p_{j,\min(L+j,2L)}), & \text{if } (j,k)=(g,h), \\ n\phi_j/p_{j,\min(L+j,2L)}, & \text{if } j=g,\ k\neq h, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

with $\nu_j = \sum_{k'=1}^{2L} \nu_{jk'}$ and

$$\phi_j = \sum_{k'=1}^{L}\sum_{t=1}^{T} \pi_{k'}^{(ctrl)}(P^{t-1})_{k'j},$$

using the fact

$$\begin{aligned} E(\nu_{jk} - \nu_j p_{jk} | n_1,\ldots,n_L) &= 0 \\ \operatorname{cov}(\nu_{jk} - \nu_j p_{jk}, \nu_{gh} - \nu_g p_{gh} | n_1,\ldots,n_L) &= \begin{cases} \displaystyle\sum_{t=1}^{T}\sum_{k'=1}^{L}(n_{k'})(P_{kj}^{t-1})p_{jk}(1-p_{jk}), & (j,k)=(g,h), \\ \displaystyle -\delta_{ig}\sum_{t=1}^{T}\sum_{k=1}^{L}(n_{k'})(P_{kj}^{t-1})p_{jk}p_{gh}, & k\neq h \end{cases} \end{aligned}$$

for $j, g = 1,\ldots,2L$, and $k, h = 1,\ldots,2L-1$ (Anderson and Goodman, 1957, Section 2). In matrix form,

$$\begin{aligned} I_{ctrl}(P) &= \operatorname{diag}(I_{11}(P_N,\bar{\Lambda}), I_{22}(P_D)), \\ I_{11}(P_N,\bar{\Lambda}) &= \operatorname{diag}(\phi_1\Xi_1,\ldots,\phi_L\Xi_L), \\ I_{22}(P_D) &= \operatorname{diag}(\phi_{L+1}\Xi_{L+1},\ldots,\phi_{2L}\Xi_{2L}), \end{aligned} \tag{28}$$

where

$$\Xi_j = \begin{cases} \operatorname{diag}(\operatorname{diag}(P_N^T\bar{\Lambda}e_j)^{-1} + (1/\Lambda_{jj})\mathbf{1}\mathbf{1}^T, O_L) \in \mathbb{R}^{2L\times 2L}, & j=1,\ldots,L, \\ \operatorname{diag}\left(O_L, \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix}(\operatorname{diag}(P_D^Te_j)^{-1} + (1/(P_D^T)_{jL})\mathbf{1}\mathbf{1}^T)\begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix}\right) \in \mathbb{R}^{2L\times 2L}, & j=L+1,\ldots,2L. \end{cases}$$

Now, from (17)–(20),

$$d\operatorname{vec}(P^T) = \begin{bmatrix} d\lambda^T & d\gamma_0^T & d\gamma & d\bar{\gamma}^T \end{bmatrix}^T \begin{bmatrix} J_1 & 0 \\ 0 & 0 \\ J_2 & 0 \\ 0 & J_3 \end{bmatrix},$$

where

$$\begin{aligned} J_1 &= \operatorname{diag}(\bar{\Lambda}_{11}(\bar{\Lambda}_{11}-1)e_1^T[P_N,-I],\ldots,\bar{\Lambda}_{11}(\bar{\Lambda}_{11}-1)e_L^T[P_N,-I]) \in \mathbb{R}^{L\times 2L^2} \\ J_2 &= \operatorname{diag}([\Lambda_{11}\Sigma_1^T, O_{(L-1)\times L}],\ldots,[\Lambda_{LL}\Sigma_L^T, O_{(L-1)\times L}]) \in \mathbb{R}^{L(L-1)\times 2L^2} \\ J_3 &= \operatorname{diag}([O_{(L-1)\times L}, \bar{\Sigma}_1^T],\ldots,[O_{(L-1)\times L}, \bar{\Sigma}_L^T]) \in \mathbb{R}^{L(L-1)\times 2L^2}. \end{aligned}$$

Therefore

$$E_\theta\left[\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial\theta}\right)\left(\frac{\partial l_{ctrl}^{(FU)}}{\partial\theta}\right)^T\right] = n\begin{bmatrix} J_1 I_{11} J_1^T & 0 & J_1 I_{11} J_2^T & 0 \\ 0 & 0 & 0 & 0 \\ J_2 I_{11} J_1^T & 0 & J_2 I_{11} J_2^T & 0 \\ 0 & 0 & 0 & J_3 I_{22} J_3^T \end{bmatrix}, \tag{29}$$

in which

$$
\begin{aligned}
J_1 I_{11} J_1^T &= \mathrm{diag}(\phi_1 \bar{\Lambda}_{11}^2 (1-\bar{\Lambda}_{11})^2 e_1^T P_N (\mathrm{diag}(P_N^T \bar{\Lambda} e_1)^{-1} + (1/\Lambda_{11})\mathbf{1}\mathbf{1}^T) P_N^T e_1, \ldots, \\
&\quad \phi_L \bar{\Lambda}_{LL}^2 (1-\bar{\Lambda}_{LL})^2 e_L^T P_N (\mathrm{diag}(P_N^T \bar{\Lambda} e_L)^{-1} + (1/\Lambda_{LL})\mathbf{1}\mathbf{1}^T) P_N^T e_L) \\
&= \mathrm{diag}(\phi_1 \bar{\Lambda}_{11}^2 (1-\bar{\Lambda}_{11})^2 (e_1^T P_N \,\mathrm{diag}(P_N^T \bar{\Lambda} e_1)^{-1} P_N^T e_1 + 1/(1-\bar{\Lambda}_{11})), \ldots, \\
&\quad \phi_L \bar{\Lambda}_{LL}^2 (1-\bar{\Lambda}_{LL})^2 (e_L^T P_N \,\mathrm{diag}(P_N^T \bar{\Lambda} e_L)^{-1} P_N^T e_L + 1/(1-\bar{\Lambda}_{LL}))), \\
J_2 I_{11} J_2^T &= \mathrm{diag}(\phi_1 \Lambda_{11}^2 \Sigma_1^T (\mathrm{diag}(P_N^T \bar{\Lambda} e_1)^{-1} + (1/\Lambda_{11})\mathbf{1}\mathbf{1}^T) \Sigma_1, \ldots, \\
&\quad \phi_L \Lambda_{LL}^2 \Sigma_L^T (\mathrm{diag}(P_N^T \bar{\Lambda} e_L)^{-1} + (1/\Lambda_{LL})\mathbf{1}\mathbf{1}^T) \Sigma_L) \\
&= \mathrm{diag}(\phi_1 (1-\bar{\Lambda}_{11})^2 \Sigma_1^T \,\mathrm{diag}(P_N^T \bar{\Lambda} e_1)^{-1} \Sigma_1, \ldots, \phi_L (1-\bar{\Lambda}_{LL})^2 \Sigma_L^T \,\mathrm{diag}(P_N^T \bar{\Lambda} e_L)^{-1} \Sigma_L), \\
J_3 I_{22} J_3^T &= \mathrm{diag}\left(\phi_{L+1} \bar{\Sigma}_1^T \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} (\mathrm{diag}(P_D^T e_1)^{-1} + (1/(P_D^T)_{1L})\mathbf{1}\mathbf{1}^T) \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} \bar{\Sigma}_1, \ldots, \right. \\
&\quad \left. \phi_{2L} \bar{\Sigma}_L^T \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} (\mathrm{diag}(P_D^T e_L)^{-1} + (1/(P_D^T)_{LL})\mathbf{1}\mathbf{1}^T) \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} \bar{\Sigma}_L \right) \\
J_1 I_{11} J_2^T &= \mathrm{diag}\left(-\phi_1 \bar{\Lambda}_{11} (1-\Lambda_{11})^2 e_1^T P_N (\mathrm{diag}(P_N^T \bar{\Lambda} e_1)^{-1} + (1/\Lambda_{11})\mathbf{1}\mathbf{1}^T) \Sigma_1, \ldots, \right. \\
&\quad \left. -\phi_L \bar{\Lambda}_{LL} (1-\Lambda_{LL})^2 e_L^T P_N (\mathrm{diag}(P_N^T \bar{\Lambda} e_L)^{-1} + (1/\Lambda_{LL})\mathbf{1}\mathbf{1}^T) \Sigma_L \right), \\
&= \mathrm{diag}\left(-\phi_1 \bar{\Lambda}_{11} (1-\Lambda_{11})^2 ((1/\bar{\Lambda}_{11})\mathbf{1}^T + (1/\Lambda_{11}) e_1^T \mathbf{1}\mathbf{1}^T) \Sigma_1, \ldots, \right. \\
&\quad \left. -\phi_L \bar{\Lambda}_{LL} (1-\Lambda_{LL})^2 ((1/\bar{\Lambda}_{LL})\mathbf{1}^T + (1/\Lambda_{LL}) e_L^T \mathbf{1}\mathbf{1}^T) \Sigma_L \right) \\
&= 0,
\end{aligned}
$$

due to $\mathbf{1}^T \Sigma_j = 0$ for $j = 1, \ldots, L$. Thus the right hand side of (29) is equal to nV_{12}.

The third and fourth terms in (16) vanish because

$$
\begin{aligned}
E_\theta \left[\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial p_{jk}} \right) \left(\frac{\partial l_{ctrl}^{(FU)}}{\partial p_{gh}} \right) \right] &= E_\theta \left[\left(\frac{\partial l_{ctrl}^{(CS)}}{\partial p_{jk}} \right) E_\theta \left[\left(\frac{\nu_{gh} - \nu_g p_{gh}}{p_{gh}} - \frac{\nu_{g,2L} - \nu_j p_{g,2L}}{p_{g,2L}} \right) \,\middle|\, n_1, \ldots, n_L \right] \right] \\
&= 0
\end{aligned} \tag{30}
$$

for $j, g = 1, \ldots, 2L$, and $k, h = 1, \ldots, 2L-1$.

A.2. Proof of Lemma 2

Note

$$
\begin{aligned}
m I_{case}(\theta) &= E_\theta \left[\left(\frac{\partial l_{case}}{\partial \theta} \right) \left(\frac{\partial l_{case}}{\partial \theta} \right)^T \right] \\
&= E_\theta \left[\left(\frac{\partial l_{case}^{(CS)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(CS)}}{\partial \theta} \right)^T \right] + E_\theta \left[\left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right)^T \right] \\
&\quad + E_\theta \left[\left(\frac{\partial l_{case}^{(CS)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right)^T \right] + E_\theta \left[\left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(CS)}}{\partial \theta} \right)^T \right].
\end{aligned} \tag{31}
$$

The first term in (31) is the information matrix of m observations of the multinomial of Lt_0 cells with cell probability $\pi^{(case)} = ((\pi_{\cdot 1}^{(case)})^T, \ldots, (\pi_{\cdot t_0}^{(case)})^T)^T = (\pi_{11}^{(case)}, \ldots, \pi_{L1}^{(case)}, \ldots, \pi_{1t_0}^{(case)}, \ldots, \pi_{Lt_0}^{(case)})^T$, where

$$
\pi_{jt}^{(case)} = \frac{g_t(\theta) e_j}{g_0(\theta)\mathbf{1}}, \quad j = 1, \ldots, L, \ t = 1, \ldots, t_0,
$$

for

$$
g_t(\theta) = (\pi_N^{(0)})^T (\bar{\Lambda} P_N)^{t-1} \Lambda P_D^{t_0 - t}, \quad \text{for } t = 1, \ldots, t_0, \qquad g_0(\theta) = \sum_{t=1}^{t_0} g_t(\alpha, \beta, \pi_0).
$$

Therefore

$$
E_\theta \left[\left(\frac{\partial l_{case}^{(CS)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(CS)}}{\partial \theta} \right)^T \right] = m \left(\frac{\partial \pi^{(case)}}{\partial \theta} \right) \mathrm{diag}(\pi^{(case)})^{-1} \left(\frac{\partial \pi^{(case)}}{\partial \theta} \right)^T.
$$

Similar to the case of $\partial\pi^{(ctrl)}/\partial\theta$, we need to evaluate the matrix differential of $g_t(\theta)$ in order to evaluate $\partial\pi_{jt}^{(case)}/\partial\theta$:

$$
\begin{aligned}
dg_t &= (d\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t} + \sum_{l=0}^{t-2}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^l(d\bar{\Lambda})P_N(\bar{\Lambda}P_N)^{t-2-l}\Lambda P_D^{t_0-t} - (\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}(d\bar{\Lambda})P_D^{t_0-t} \\
&\quad + \sum_{l=0}^{t-2}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^l\bar{\Lambda}(dP_N)(\bar{\Lambda}P_N)^{t-2-l}\Lambda P_D^{t_0-t} + \sum_{l=0}^{t_0-t-2}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^l(dP_D)P_D^{t_0-t-1-l} \\
&= d\gamma_0^T\Sigma_0^T(\bar{\Lambda}P_N)^{t-1}P_D^{t_0-t} \\
&\quad + \sum_{j=1}^{L}d\lambda_j\bar{\Lambda}_{jj}(\bar{\Lambda}_{jj}-1)\left(\sum_{l=0}^{t-2}(\pi_N^{(0)})^T(\bar{P}_N)^le_je_j^T(\bar{\Lambda}P_N)^{t-2-l}\Lambda P_D^{t_0-t} - (\pi_N^{(0)})^T(\bar{P}_N)^{t-1}e_je_j^TP_D^{t_0-t}\right) \\
&\quad + d\gamma^T\sum_{j=1}^{L}(e_je_j^T\otimes\Sigma_j^T)\sum_{l=0}^{t-2}\left((\bar{\Lambda}(P_N^T\bar{\Lambda})^l\pi_N^{(0)})\otimes(\bar{\Lambda}P_N)^{t-2-l}\right)\Lambda P_D^{t_0-t} \\
&\quad + d\bar{\gamma}^T\sum_{j=1}^{L}(e_je_j^T\otimes\bar{\Sigma}_j^T)\sum_{l=0}^{t_0-t-1}\left(((P_D^T)^l\Lambda(P_N^T\bar{\Lambda})^{t-1}\pi_N^{(0)})\otimes P_D^{t_0-t-1-l}\right) \\
&= [d\lambda^T, d\gamma_0^T, d\gamma^T, d\bar{\gamma}^T]\cdot\begin{bmatrix}\bar{\Lambda}(\bar{\Lambda}-I) & & & \\ & \Sigma_0^T & & \\ & & \mathrm{diag}(\Sigma_1^T,\ldots,\Sigma_L^T) & \\ & & & \mathrm{diag}(\bar{\Sigma}_1^T,\ldots,\bar{\Sigma}_L^T)\end{bmatrix} \\
&\quad\cdot\begin{bmatrix}\left(G_{t-1}(\pi_N^{(0)},\bar{\Lambda}P_N,P_N)\Lambda - \mathrm{diag}((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1})\right)P_D^{t_0-t} \\ (\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t} \\ F_{t-1}(\pi_N^{(0)},\bar{\Lambda}P_N,\bar{\Lambda})\Lambda P_D^{t_0-t} \\ F_{t_0-t}(\Lambda(P_N^T\bar{\Lambda})^{t-1}\pi_N^{(0)},P_D,I)\end{bmatrix},
\end{aligned}
$$

using (17)–(20). Thus

$$
\frac{\partial g_t(\theta)}{\partial\theta} = \begin{bmatrix}\bar{\Lambda}(\bar{\Lambda}-I)\left(G_{t-1}(\pi_N^{(0)},\bar{\Lambda}P_N,P_N)\Lambda - \mathrm{diag}((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1})\right)P_D^{t_0-t} \\ \Sigma_0^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t} \\ \mathrm{diag}(\Sigma_1^T,\ldots,\Sigma_L^T)F_{t-1}(\pi_N^{(0)},\bar{\Lambda}P_N,\bar{\Lambda})\Lambda P_D^{t_0-t} \\ \mathrm{diag}(\bar{\Sigma}_1^T,\ldots,\bar{\Sigma}_L^T)F_{t_0-t}(\Lambda(P_N^T\bar{\Lambda})^{t-1}\pi_N^{(0)},P_D,I)\end{bmatrix}.
$$

It follows that

$$
\begin{aligned}
\frac{\partial\pi_{jt}^{(case)}}{\partial\theta} &= \frac{(g_0(\theta)\mathbf{1})(\partial g_t(\theta)/\partial\theta)P_D^{t_0-t}e_j - (g_t(\theta)P_D^{t_0-t}e_j)(\partial g_0(\theta)/\partial\theta)\mathbf{1}}{(g_0(\theta)\mathbf{1})^2} \\
&= \frac{1}{\sum_{t'=1}^{t_0}(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t'-1}\Lambda\mathbf{1}}\begin{bmatrix}\bar{\Lambda}(\bar{\Lambda}-I)\left(G_{t-1}(\pi_N^{(0)},\bar{\Lambda}P_N,P_N)\Lambda - \mathrm{diag}((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1})\right)P_D^{t_0-t}e_j \\ \Sigma_0^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t}e_j \\ \mathrm{diag}(\Sigma_1^T,\ldots,\Sigma_L^T)F_{t-1}(\pi_N^{(0)},\bar{\Lambda}P_N,\bar{\Lambda})\Lambda P_D^{t_0-t}e_j \\ \mathrm{diag}(\bar{\Sigma}_1^T,\ldots,\bar{\Sigma}_L^T)F_{t_0-t}(\Lambda(P_N^T\bar{\Lambda})^{t-1}\pi_N^{(0)},P_D,I)e_j\end{bmatrix} \\
&\quad - \frac{(\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t-1}\Lambda P_D^{t_0-t}e_j}{\left(\sum_{t'=1}^{t_0}((\pi_N^{(0)})^T\bar{\Lambda}P_N)^{t'-1}\Lambda\mathbf{1}\right)^2}\sum_{t'=1}^{t_0}\begin{bmatrix}\bar{\Lambda}(\bar{\Lambda}-I)\left(G_{t'-1}(\pi_N^{(0)},\bar{\Lambda}P_N,P_N)\Lambda - \mathrm{diag}((\pi_N^{(0)})^T(\bar{\Lambda}P_N)^{t'-1})\right)\mathbf{1} \\ \Sigma_0^T(\bar{\Lambda}P_N)^{t-1}\Lambda\mathbf{1} \\ \mathrm{diag}(\Sigma_1^T,\ldots,\Sigma_L^T)F_{t'-1}(\pi_N^{(0)},\bar{\Lambda}P_N,\bar{\Lambda})\Lambda\mathbf{1} \\ \mathrm{diag}(\bar{\Sigma}_1^T,\ldots,\bar{\Sigma}_L^T)F_{t_0-t'}(\Lambda(P_N^T\bar{\Lambda})^{t'-1}\pi_N^{(0)},P_D,I)\mathbf{1}\end{bmatrix} \\
&= \begin{bmatrix}H_\lambda^{(t)} \\ H_{\gamma_0}^{(t)} \\ H_\gamma^{(t)} \\ H_{\bar{\gamma}}^{(t)}\end{bmatrix}e_j - \pi_{jt}^{(case)}\begin{bmatrix}H_\lambda^{(t)} \\ H_{\gamma_0}^{(t)} \\ H_\gamma^{(t)} \\ H_{\bar{\gamma}}^{(t)}\end{bmatrix}\mathbf{1} = H_te_j - \pi_{jt}^{(case)}H_t\mathbf{1}
\end{aligned}
$$

for $j = 1, \ldots, L$, $t = 1, \ldots, t_0$. Thus

$$\begin{aligned}\frac{\partial \pi^{(case)}}{\partial \theta} &= \begin{bmatrix} H_1 & H_2 & \cdots & H_{t_0} \end{bmatrix} - \begin{bmatrix} H_1 \mathbf{1}(\pi_{\cdot 1}^{(ctrl)})^T & H_2 \mathbf{1}(\pi_{\cdot 2}^{(ctrl)})^T & \cdots & H_{t_0} \mathbf{1}(\pi_{\cdot t_0}^{(ctrl)})^T \end{bmatrix} \\ &= H - H \operatorname{diag}(\mathbf{1}(\pi_{\cdot 1}^{(case)})^T, \ldots, \mathbf{1}(\pi_{\cdot t_0}^{(case)})^T).\end{aligned}$$

Let

$$\begin{aligned}J_2 &= \frac{\partial \pi^{(case)}}{\partial \theta} \operatorname{diag}(\pi^{(case)})^{-1/2} \\ &= H \operatorname{diag}(\pi^{(case)})^{-1/2} - H \operatorname{diag}\left(\operatorname{diag}(\tilde{\pi}_{\cdot 1}^{(case)})^{-1} \mathbf{1}(\pi_{\cdot 1}^{(case)})^T, \ldots, \operatorname{diag}(\tilde{\pi}_{\cdot t_0}^{(case)})^{-1} \mathbf{1}(\pi_{\cdot t_0}^{(case)})^T \right),\end{aligned}$$

where $\tilde{\pi}_{\cdot t}^{(case)} = ((\pi_{1t}^{(case)})^{1/2}, \ldots, (\pi_{Lt}^{(case)})^{1/2})^T$ for $t = 1, \ldots, t_0$. Then,

$$\begin{aligned}&\left(\frac{\partial \pi^{(case)}}{\partial \theta} \right) \operatorname{diag}(\pi^{(case)})^{-1} \left(\frac{\partial \pi^{(case)}}{\partial \theta} \right)^T = J_2 J_2^T \\ &= H \Big(\operatorname{diag}(\pi^{(case)})^{-1} - \operatorname{diag}(\pi^{(case)})^{-1/2} \operatorname{diag}\left(\operatorname{diag}(\tilde{\pi}_{\cdot 1}^{(case)})^{-1} \mathbf{1}(\pi_{\cdot 1}^{(case)})^T, \ldots, \operatorname{diag}(\tilde{\pi}_{\cdot t_0}^{(case)})^{-1} \mathbf{1}(\pi_{\cdot t_0}^{(case)})^T \right) \\ &\quad - \operatorname{diag}\left(\pi_{\cdot 1}^{(case)} \mathbf{1}^T \operatorname{diag}(\tilde{\pi}_{\cdot 1}^{(case)})^{-1}, \ldots, \pi_{\cdot t_0}^{(case)} \mathbf{1}^T \operatorname{diag}(\tilde{\pi}_{\cdot t_0}^{(case)})^{-1} \right) \operatorname{diag}(\pi^{(case)})^{-1/2} \\ &\quad + \operatorname{diag}\left((\mathbf{1}^T \operatorname{diag}(\pi_{\cdot 1}^{(case)})^{-1} \mathbf{1}) \pi_{\cdot 1}^{(case)} (\pi_{\cdot 1}^{(case)})^T, \ldots, (\mathbf{1}^T \operatorname{diag}(\pi_{\cdot t_0}^{(case)})^{-1} \mathbf{1}) \pi_{\cdot t_0}^{(case)} (\pi_{\cdot t_0}^{(case)})^T \right) \Big) H^T \\ &= \begin{bmatrix} H_\lambda \\ H_{\gamma_0} \\ H_\gamma \\ H_{\bar{\gamma}} \end{bmatrix} D_2 \begin{bmatrix} H_\lambda^T & H_{\gamma_0}^T & H_\gamma^T & H_{\bar{\gamma}}^T \end{bmatrix} = V_{21}.\end{aligned}$$

The second term in (31) is the information matrix of m observations of the Markov chain with transition matrix $P_D = P_D(\theta)$ for T steps, conditioning on the initial state frequencies of $m_1, \ldots, m_L$, where $m_j = \sum_{t=1}^{t_0} m_{jt}$. Following (27), we have

$$E_\theta \left[\left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right)^T \right] = \left(\frac{\partial \operatorname{vec}(P_D^T)}{\partial \theta} \right) m I_{case}(P_D) \left(\frac{\partial \operatorname{vec}(P_D^T)}{\partial \theta} \right)^T,$$

where

$$I_{case}(P_D) = \operatorname{diag}(\psi_1 \Psi_1, \ldots, \psi_L \Psi_L) \in \mathbb{R}^{L^2 \times L^2}$$

with

$$\psi_j = \sum_{k'=1}^{L} \sum_{t=1}^{T} \sum_{\tau=1}^{t_0} \pi_{k'\tau}^{(case)} (P_D^{t-1})_{k'j},$$

and

$$\Psi_j = \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} (\operatorname{diag}(P_D^T e_j)^{-1} + (1/(P_D^T)_{jL}) \mathbf{1}\mathbf{1}^T) \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} \in \mathbb{R}^{L \times L}, \quad j = 1, \ldots, L.$$

for $j, g = 1, \ldots, L$, and $k, h = 1, \ldots, L-1$. Now, from (17)–(20),

$$d \operatorname{vec}(P_D^T) = \begin{bmatrix} d\lambda^T & d\gamma_0^T & d\gamma & d\bar{\gamma} \end{bmatrix}^T \begin{bmatrix} 0 \\ 0 \\ 0 \\ J_4 \end{bmatrix},$$

where

$$J_4 = \operatorname{diag}(\bar{\Sigma}_1^T, \ldots, \bar{\Sigma}_L^T) \in \mathbb{R}^{L(L-1) \times L^2}.$$

Therefore

$$E_\theta \left[\left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right) \left(\frac{\partial l_{case}^{(FU)}}{\partial \theta} \right)^T \right] = m \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & J_4 I_{case} J_4^T \end{bmatrix}, \tag{32}$$

in which

$$J_4 I_{case} J_4^T = \operatorname{diag}\left(\psi_1 \bar{\Sigma}_1^T \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} (\operatorname{diag}(P_D^T e_1)^{-1} + (1/(P_D^T)_{1L})\mathbf{1}\mathbf{1}^T) \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} \bar{\Sigma}_1, \ldots, \right.$$
$$\left. \psi_L \bar{\Sigma}_L^T \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} (\operatorname{diag}(P_D^T e_L)^{-1} + (1/(P_D^T)_{LL})\mathbf{1}\mathbf{1}^T) \begin{bmatrix} I_{L-1} & \\ & 0 \end{bmatrix} \bar{\Sigma}_L \right).$$

Thus the right hand side of (32) is equal to mV_{22}.

The third and fourth terms in (31) vanish just as (30).

References

Agresti, A., Kateri, M., 2012. Categorical Data Analysis, third ed. John Wiley & Sons.

Altar, C.A., 2008. The biomarkers consortium: On the critical path of drug discovery. Clin. Pharmacol. Ther. 83 (2), 361–364.

Anderson, T.W., Goodman, L.A., 1957. Statistical inference about Markov chains. Ann. Math. Statist. 89–110.

Bauer, D.C., Glüer, C.C., Cauley, J.A., Vogt, T.M., Ensrud, K.E., Genant, H.K., Black, D.M., 1997. Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women: a prospective study. Arch. Intern. Med. 157 (6), 629–634.

Biomarkers Definitions Working Group, , 2001. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin. Pharmacol. Ther. 69 (3), 89–95.

Breslow, N., 1976. Regression analysis of the log odds ratio: A method for retrospective studies. Biometrics 409–416.

Breslow, N.E., 1996. Statistics in epidemiology: the case-control study. J. Amer. Statist. Assoc. 91 (433), 14–28.

Chiang, C.L., 1980. Introduction to Stochastic Process and their Applications. Robert E. Krieger Publishing Co., New York.

Christensen, G.E., Joshi, S.C., Miller, M.I., 1997. Volumetric transformation of brain anatomy. IEEE Trans. Med. Imaging 16 (6), 864–877.

European Society of Radiology, , 2010. White paper on imaging biomarkers. Insights Imaging 1 (2), 42–45.

Jack, C., Petersen, R.C., Xu, Y.C., OBrien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Waring, S.C., Tangalos, E.G., Kokmen, E., 1999. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 52 (7), 1397–1403.

Langton, C., Palmer, S., Porter, R., 1984. The measurement of broadband ultrasonic attenuation in cancellous bone. Eng. Med. 13 (2), 89–91.

Lee, S.H., Bachman, A.H., Yu, D., Lim, J., Ardekani, B.A., Initiative, A.D.N., et al., 2016. Predicting progression from mild cognitive impairment to Alzheimer's disease using longitudinal callosal atrophy. Alzheimer's Dement.: Diagn., Assess. Dis. Monit. 2, 68–74.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, second ed. Springer.

Mullen, K.M., 2014. Continuous global optimization in R. J. Stat. Softw. 60 (6), 1–45.

Porta, M.S., Greenland, S., Hernán, M., dos Santos Silva, I., Last, J.M., 2014. A Dictionary of Epidemiology, sixth ed. Oxford University Press.

Rao, C.R., 1973. Linear Statistical Inference and its Applications. John Wiley & Sons.

Rosen, W.G., Mohs, R.C., Davis, K.L., 1984. A new rating scale for Alzheimer's disease. Am. J. Psychiatry.

Schlesselman, J.J., 1982. Case Control Studies: Design, Conduct, Analysis. Oxford University Press.

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L., Trojanowski, J., Thompson, P., Jack, C., Weiner, M., Initiative, D.N., et al., 2009. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 132 (4), 1067–1077.

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., Wahlund, L.-O., Larsson, E.-M., Westman, E., Initiative, A.D.N., et al., 2014. The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy ageing and Alzheimer's disease. Front. Aging Neurosci. 6, 264.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Cedarbaum, J., Donohue, M.C., Green, R.C., Harvey, D., Jack, C.R., et al., 2015. Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. Alzheimer's Dement. 11 (7), 865–884.

Xiang, Y., Gubian, S., Suomela, B., Hoeng, J., 2013. Generalized simulated annealing for global optimization: the GenSA package. R J. 5 (1), 13–28.