

Prediction of Dementia Using Screening and Diagnosis History in Longitudinal Two-phase Studies

Changyu Shen^{1,2}

¹ *Division of Biostatistics, School of Medicine, Indiana University, 1050 Wishard Boulevard RG R4101,
Indianapolis, IN 46202, U.S.A.*

² *Regenstrief Institute for Health Care, 1050 Wishard Boulevard, Indianapolis, IN 46202*

SUMMARY

Longitudinal studies using a two-phase sampling design offer a great potential in identifying cases of rare chronic disease whose diagnosis is complex and expensive. In addition to ethical reasons, the detected cases provide precious resource for many other studies that require sufficient number of cases to make reliable conclusions. In this paper, we proposed a formal statistical model to identify subjects at high risk of dementia based on the screening/diagnosis history in order to increase screening yield. We show that compared with cross-sectional method, up to 10% improvement in sensitivity can be achieved at several specificity levels of practical interest and the overall gain in prediction accuracy as measured by the AUC under the ROC curve is 3%-5%. Moreover, the inclusion of screening/diagnosis history is more helpful when the screening is less accurate. Our model provides a general framework

*Correspondence to: Division of Biostatistics, School of Medicine, Indiana University, 1050 Wishard Boulevard RG R4101, Indianapolis, IN 46202, U.S.A. E-mail: chashen@iupui.edu Phone: 317-274-1641 Fax: 317-274-2678

Contract/grant sponsor: National Institute of Health; contract/grant number: R01 AG15813, R01 AG09956 and P30 AG10133

that can be applied to many two-phase longitudinal studies of various diseases. It also serves as a preliminary investigation on an “adaptive” study design, in which probabilities of disease at each data collection wave are calculated based on the screening/diagnosis history and used to select subjects for diagnosis. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Dementia; Longitudinal; Prediction; Screening; Two-phase

1. INTRODUCTION

Epidemiological research often involves identification of cases of rare disorders. The diagnosis of the disorder of interest can be very expensive and complicated, which inevitably limits its use in a general population or a large sample from such a population. The difficulty can be overcome by applying the two-phase sampling, or screening, which was first proposed by Neyman as a sampling technique [1]. The procedure of a two-phase study design is composed of two steps: (i) a cheap, but less accurate test (screening) is first applied to a sample from the target population and the sample is divided into a number of strata based on the screening; (ii) a sub-sample is drawn from each stratum and a formal diagnosis is made on the sub-sample. In its simplest form, the original sample is divided into two strata: screening-positive and screening-negative. The idea is that the screening serves as a filter to roughly split the diseased from non-diseased so that (a) the screening-positive is much more concentrated with diseased subjects than the screening-negative and (b) the formal diagnosis can be more orientated toward screening-positive by applying a higher sampling rate to this stratum than the other stratum. The gain in efficiency will then

depend on the cost of the screening and its accuracy when other factors are fixed.

The major application of two-phase sampling has been focused on the estimation of prevalence of a rare disease [2, 3]. Optimal designs that minimize the variance of the estimator given a fixed budget have been proposed by several authors [4, 5, 6, 7]. It was shown that screening is most effective in the reduction of cost when the prevalence of disease is low and the screening is highly successful in separating cases from normal people [5]. Clayton also discussed incidence estimation in a multi-phase sampling framework [8].

A longitudinal study using two-phase design serves multiple purposes in addition to prevalence and incidence estimation, which includes case detection, identification of risk factors, life expectancy and so on. Case detection holds an important position for two reasons. First, ethical concern requires that all screening-positive subjects be diagnosed at the second phase [6]. This is particularly of our concern when the cost of the false negative is extremely high (e.g. early diagnosis is crucial for the victims' health benefit). Second, for a rare disease, a relative large diseased sample is a precious resource for other studies [2, 9]. For example, DNA-microarray and mass spectrometry techniques have been used to seek biomarkers of various diseases. Due to the high-dimension nature of data yielded from these techniques, a decent sample size is of critical importance for improving statistical power and making reliable conclusions.

Longitudinal studies offer a great potential in case detection since each participant in the study is evaluated multiple times over time. Hence, each subject who has not been diagnosed of having a disease possesses a profile of screening/diagnosis history that can potentially predict his/her current disease status. Such profile is particularly

informative for progressive diseases that manifest themselves gradually over time. For instance, Alzheimer's disease (AD) is an irreversible progressive neurological disorder, whose manifestation in general includes unusual cognitive decline in addition to overall low cognitive performance. The screening of AD usually includes a neuropsychological survey which generates a number of measurements used to evaluate the subject's cognitive performance. Compared with cross-sectional screening, the longitudinal design then is advantageous since it has cognitive decline information embedded in the screening history, which is a strong predictor of AD. From another point of view, a person whose last diagnosis as normal was made 10 years ago is likely to have a higher chance to have AD now than does a person whose last diagnosis as normal was made 2 years ago, given other relevant factors are fixed at the same level. Therefore, the screening/diagnosis history provides extra information on the likelihood of AD in addition to current screening. Then we are interested in the magnitude of possible gain in prediction accuracy by including such information. To our best knowledge, a formal statistical framework to address this question seems lacking in the literature.

In this paper, we construct a transitional model that allows one to calculate the probability of disease given the subject's screening/diagnosis history, adjusting for potentially non-ignorable drop-out [10]. We compare our model with cross-sectional analysis through real data analysis and simulations. Our model is motivated by a longitudinal two-phase study of dementia and developed in the context of the design of this study. Nevertheless, it presents a rather general framework that can fit into a broad range of two-phase studies of irreversible rare disorders. As McIntosh and Pepe pointed out [11], the probability of disease given the data is the optimal risk score in

the sense that the Receiver Operating Characteristic (ROC) curve is maximized at every point, which is equivalent to the Neyman-Pearson theory in hypothesis testing [12]. In other words, a selection rule based on the posterior probability of disease will maximize the sensitivity at any specificity level. Our work also serves as a preliminary investigation on an efficient design of two-phase longitudinal studies to maximize the number of cases detected. Essentially, such a design is carried out in a Bayesian context, in which the data model and the prior distribution of the parameters are pre-specified. Then we can calculate the probability of disease at each data collection wave (after the screening) given the data and select subjects with high probabilities for diagnosis.

The remaining of this article is organized as follows. In Section 2, we introduce the Indianapolis-Ibaden Dementia Project (IIDP) that motivates our work. We provide model details in Section 3. In Section 4, we apply the proposed model to the IIDP and conduct a simulation study to investigate the gain in prediction accuracy using our model as compared with prediction model based on cross-sectional analysis. We conclude this paper in Section 5.

2. THE INDIANAPOLIS-IBADEN DEMENTIA PROJECT

The Indianapolis-Ibaden Dementia Project is an on-going longitudinal study of dementia and Alzheimer's disease in the elderly starting 1992 [13]. Currently, this project is at its 5th data collection wave (including baseline). The study participants are 2212 African Americans living in Indianapolis (U.S.A.) and 2494 native Africans living in Ibaden (Nigeria). All participants were 65 or older at enrollment. A population-based two-phase survey was conducted at each data collection wave. There

was first an in-home screening using the Community Screening Interview for Dementia (CSID) [14], which yields a continuous score as an initial evaluation of the participants' cognitive performance. Subjects are then divided into 3 performance groups (good, intermediate and poor) based on their CSID scores. A full clinical assessment was performed for a random sub-sample of participants from each of the 3 groups with sampling rates 5%, 50% and 100%, respectively. In order to assure an old enough sub-sample was drawn from the "good" performance group, a stratified sampling was further conducted selecting 75% of those aged 75 or older in this group. Subjects diagnosed as dementia are excluded for future follow-up.

We choose the data collected at the Indianapolis site up to wave 4 to illustrate our model. We will fit our model to the data of the first 3 waves and use the estimated model to predict dementia at wave 4. For the sake of simplicity, we eliminate subjects who had intermittent missing values on CSID and subjects who have missing values on the covariates included in our model. This leads to 2115 subjects and the time between each consecutive follow-up is 2.6 years on average. In Table I, we summarize some demographic factors, CSID scores and the number of dementia cases detected at each time point for this cohort. It appears that females and subjects with higher education level stay longer in the study as compared with males and subjects with lower education level. Our previous work also suggests that the drop-out might be non-ignorable in the sense that the drop-out mechanism is associated with dementia status [15]. In next section, we introduce a transition model to describe the CSID score, diagnosis and drop-out process. The model is then used to predict the likelihood of dementia after the parameters are estimated. We present two types of likelihood

based approaches to predict dementia for subjects still in the study using their CSID/diagnosis profile. The first one is a full-likelihood approach that treats CSID score, diagnosis and drop-out process as random variables and the second one is a conditional-likelihood approach that treats CSID as fixed.

3. A TRANSITION MODEL

3.1. General framework

To facilitate the description of our statistical framework, we first introduce some notations. For simplicity of presentation, we drop subject index for now. Let S_i be a continuous screening score at wave i and $Y_i = 1$ if demented at wave i and 0 otherwise. Denote $D_i = 1$ if a participant leaves the study (due to any reason) before wave $i + 1$ and 0 otherwise. Let $W = \min(i : D_i = 1)$ be the last wave, at which a subject is in the study ($W \geq 1$). Then before a participant leaves the study, S_i is always observed and Y_i is observed occasionally due to the study design. We let $C_i = 1$ if Y_i is observed at wave i (a diagnosis is made) and 0 otherwise for $i = 1, 2, \dots, W$. We will assume that C_i is dependent only on S_i due to the two-phase design. Let $Z_i = (S_i, Y_i, C_i)$. Finally, we will use subscript (i) to indicate history up to wave i (e.g. $Z_{(i)} = (Z_k, k = 1, 2, \dots, i)$). Then we have the following properties:

Property(i) $Y_i = 1 \rightarrow Y_j = 1$ for $j > i$ (irreversibility of dementia);

Property(ii) $D_i = 1 \rightarrow D_j = 1$ for $j > i$ (definition of D_i);

Property(iii) $Y_i C_i = 1 \rightarrow D_i = 1$ (study design).

We show in Table II several patterns of (Y_i, C_i, D_i) that can occur to demonstrate the

data structure.

Because we are primarily concerned with subjects who are still in the study, we restrict any distribution in the following description to that conditional on $W \geq i$ unless otherwise noted. Specifically, for $i > 1$, we assume:

$$\begin{aligned} [Z_i, D_i | Z_{(i-1)}] &= [S_i | Z_{(i-1)}][Y_i | S_i, Z_{(i-1)}][C_i | S_i, Y_i, Z_{(i-1)}][D_i | S_i, Y_i, C_i, Z_{(i-1)}] \\ &= [S_i | Z_{i-1}][Y_i | S_i, Z_{i-1}][C_i | S_i][D_i | S_i, Y_i, C_i] \end{aligned} \quad (1)$$

$$\propto [S_i | Y_{i-1}, S_{i-1}][Y_i | S_i, S_{i-1}, Y_{i-1}][D_i | S_i, Y_i, C_i] \quad (2)$$

$$= P_i(Y_{i-1})Q_i(Y_{i-1}, Y_i)R_i(Y_i, C_i, D_i) \quad (3)$$

Equation (1) is obvious for the distribution of C_i . We now explain the rationale for each other term in equation (1). For S_i , it is clear that S_i does not depend on C_k conditional on S_k ($k < i - 1$). We also conjecture that it is reasonable to assume that the current screen score does not depend on score measured before last data collection wave conditional on last score and last dementia status. It is possible that S_i might depend on the history of Y_i (e.g. the longer the duration of the disease, the worse the performance). However, it should be relative minor when we condition on last screen performance. For Y_i , we only need to consider the scenario of $Y_{i-1} = 0$ due to property (i). Note that $Y_{i-1} = 0$ also implies that $Y_k = 0, k < i - 1$. Since the onset of dementia follows a great decline in cognitive performance, it is rather rare to observe a person who is normal at wave $i - 1$ but has been experiencing dramatic cognitive decline or low cognitive performance for a relatively long time (recall on average, two consecutive interviews are 2.6 years apart). Hence, after we include the screening at wave $i - 1$ and i , the screening history before wave $i - 1$ does not provide

much extra information with respect to the onset of dementia at wave i . Finally, $[D_i]$ is likely to depend on dementia and cognitive performance. It was shown that subjects whose cognitive performance are low tend to drop out the study [16]. It is possible that the drop-out mechanism depends on the history of cognitive performance and dementia status (e.g. cognitive decline over the past years) conditional on current dementia status and screening score, we assume they are relative minor. Since we are not trying to construct a selection model to address the drop-out issue in this paper, we believe this assumption suffices our prediction purpose. Since the distribution of C_i is determined by a three-level categorical variable defined by S_i , it can be assumed that $[C_i|S_i]$ is known. Then equation (2) follows, with \propto indicating that the density on the left side differs from the density on the right by a known constant. We omit the dependency of the terms in equation(3) as functions of S_i for notation simplicity. The dependency of R_i on C_i is due to the design that subjects are excluded for further follow-up when diagnosed as dementia. Finally, to initiate the transition model, we assume:

$$\begin{aligned}
[Z_1, D_1] &= [S_1][Y_1|S_1][C_1|S_1, Y_1][D_1|S_1, Y_1, C_1] \\
&\propto [S_1][Y_1|S_1][D_1|S_1, Y_1, C_1] \\
&= P_1 Q_1(Y_1) R_1(Y_1, C_1, D_1)
\end{aligned} \tag{4}$$

Then the density of the complete data up to wave W (the last wave, at which a participant is in the study) can be written as:

$$[Z_{(W)}] \propto P_1 Q_1(Y_1) R_1(Y_1, C_1, I(W = 1)) \prod_{i=2}^W P_i(Y_{i-1}) Q_i(Y_{i-1}, Y_i) R_i(Y_i, C_i, I(W = i)),$$

where $I(expr)$ is the indicator function that takes value 1 when $expr$ is true and 0 otherwise. The density of the observed data can be calculated by summing out the unobserved Y_i 's:

$$[Z_{(W)}^{obs}] \propto P_1 \sum_{Y_{miss}} \{Q_1(Y_1)R_1(Y_1, C_1, I(W = 1)) \prod_{i=2}^W P_i(Y_{i-1})Q_i(Y_{i-1}, Y_i)R_i(Y_i, C_i, I(W = i))\}. \quad (5)$$

Now, suppose we are at wave t and already obtained S_t . Then we can calculate the probability of dementia for a subject who has not been diagnosed as demented and is still in the study based on the framework just described. Suppose that the last diagnosis on the subject was made at wave $q, 0 \leq q \leq t - 1$ with $q = 0$ indicating no diagnosis was made on the subject. This also implies $C_{q+1} = \dots = C_{t-1} = 0$. Then the probability can be calculated as follows after some algebra (see appendix for details):

$$\Pr[Y_t = 1|q, S_{(t)}, W \geq t] = \begin{cases} Q_t(0, 1), & \text{if } q = t - 1 \\ \frac{(\sum_{j=0}^{t-2-q} M_j) + M_{t-1-q} Q_t(0,1)}{\sum_{j=0}^{t-1-q} M_j}, & \text{if } q < t - 1 \end{cases} \quad (6)$$

where $M_j = Q_{q+1}R_{q+1} \left(\prod_{k=q+2}^{t-1} P_k Q_k R_k \right) P_t$ evaluated at $Y_q = \dots = Y_{q+j} = 0, Y_{q+j+1} = \dots = Y_{t-1} = 1, C_{q+1} = \dots = C_{t-1} = 0$, and $D_q = \dots = D_{t-1} = 0$.

It can be seen from equation (6) that the probability of dementia does not depend on the screening scores before the wave of last diagnosis. In other words, our model assumes that conditional on the subject is normal at wave q , all screening scores at wave $i < q$ do not provide extra information regarding the current likelihood of dementia. Since P_1 is not involved in the prediction and can be estimated separately from other distributions (equation (5)), we will not model this term in next subsection, in which we provide further details on how to model and estimate parameters for other terms

in equation (5).

3.2. Model and estimation

We provide model details for P_i , Q_i and R_i and estimation of parameters in this subsection. We will include some covariates in each of the three terms in addition to what has been discussed in previous subsection. We use X_i to denote a column covariate vector that includes: sex (*female*, 1:female, 0:male), years of education (*grade*, continuous), age at wave i (*age*, continuous, time-dependent), time interval in years between wave $i - 1$ and i (*tint*, continuous, time-dependent, X_1 for Q_1 and X_i for R_i do not include *tint*). Since the estimation is based on the Maximum Likelihood (ML) principle, we will write P_i , Q_i and R_i as functions of corresponding parameters.

(a) P_i , $i > 1$

We assume a normal distribution for S_i . Specifically,

$$P_i(\alpha) = [S_i | S_{i-1}, Y_{i-1}, X_i] \sim N(G_i, \sigma^2) \quad (7)$$

$$\alpha = (\alpha_C, \alpha_S, \alpha_Y, \alpha_X, \sigma^2)$$

$$G_i = \alpha_C + \alpha_S S_{i-1} + \alpha_Y Y_{i-1} + \alpha_X X_i.$$

(b) Q_i , $i \geq 1$

For $i > 1$, we use a logistic structure:

$$Q_i(\beta) = [Y_i | S_i, S_{i-1}, Y_{i-1}, X_i] = \begin{cases} 1, & \text{if } Y_{i-1} = Y_i = 1 \\ 0, & \text{if } Y_{i-1} - Y_i = 1 \\ K_i^{Y_i} (1 - K_i)^{1 - Y_i}, & \text{otherwise} \end{cases} \quad (8)$$

$$\beta = (\beta_C, \beta_S, \beta_{DLN}, \beta_X)$$

$$K_i = \text{logit}^{-1}(\beta_C + \beta_S S_i + \beta_{DLN}(S_{i-1} - S_i) + \beta_X X_i).$$

For $i = 1$, we assume:

$$\begin{aligned} Q_1(\lambda) &= [Y_i|S_1, X_1] = K_1^{Y_1}(1 - K_1)^{1-Y_1} \\ \lambda &= (\lambda_C, \lambda_S, \lambda_X) \\ K_1 &= \text{logit}^{-1}(\lambda_C + \lambda_S S_1 + \lambda_X X_1). \end{aligned} \tag{9}$$

(c) $R_i, i \geq 1$

We assume a logistic structure for R_i :

$$\begin{aligned} R_i(\gamma) &= [D_i|S_i, Y_i, C_i, X_i] = \begin{cases} 1, & \text{if } Y_i C_i = D_i = 1 \\ 0, & \text{if } Y_i C_i - D_i = 1 \\ H_i^{D_i}(1 - H_i)^{1-D_i}, & \text{otherwise} \end{cases} \\ \gamma &= (\gamma_C, \gamma_S, \gamma_Y, \gamma_X) \\ H_i &= \text{logit}^{-1}(\gamma_C + \gamma_S S_i + \gamma_Y Y_i + \gamma_X X_i). \end{aligned} \tag{10}$$

Then we can write the likelihood function of $(\alpha, \beta, \lambda, \gamma)$ for a subject as: (see (5))

$$L(\alpha, \beta, \lambda, \gamma) \propto \sum_{Y_{miss}} \{Q_1(\lambda)R_1(\gamma) \prod_{i=2}^W P_i(\alpha)Q_i(\beta)R_i(\gamma)\}. \tag{11}$$

For a data set of size n , we can write the likelihood function as:

$$L_n(\alpha, \beta, \lambda, \gamma) \propto \prod_{j=1}^n L^{(j)}(\alpha, \beta, \lambda, \gamma), \tag{12}$$

where $L^{(j)}$ is the likelihood function for subject j . The parameters can be estimated by ML using various numerical algorithms. Since no algorithm can guarantee a global maximum, a common practice is to initiate the algorithm with various starting points. In this paper, we will first fit models (7)- (10) separately and randomly select 200 starting points at the neighborhood of the estimates (α_Y and γ_Y are not estimable by solely fitting (7) or (10), starting values for α_Y are selected at the neighborhood of

the mean difference (over time and individuals) of screening scores between subjects diagnosed of dementia and others and starting values for γ_Y are selected at a neighborhood of 0). We use the Nelder-Mead algorithm [17] to optimize the likelihood function. Since we are trying to predict the likelihood of dementia at wave t , the likelihood function (11) in practice will include data up to wave $t - 1$ instead of the time at which the subjects leave the study. Then we can calculate the probability of dementia at wave t based on the parameter estimates and (6).

One way to approximate the likelihood function and prediction is to treat S_i as fixed and omit the term P_i in (11) and (6). We will call the approach based on (11) the Full Likelihood (**FL**) approach and the one treating S_i as fixed the Conditional Likelihood (**CL**) approach. In next section, we will examine the performance of FL, CL and a cross-sectional approach through the analysis of Indianapolis data from IIDP and simulation.

4. DATA ANALYSIS AND SIMULATION

We applied the proposed model described in last section to the data collected at Indianapolis site from the Indianapolis-Ibaden Dementia Project. Parameters are estimated using data up to wave 3. The estimates are then plugged into (6) to predict the likelihood of dementia at wave 4 for the 201 subjects who were selected for diagnosis (Table I). In addition to the FL and CL approaches, we also include a cross-sectional (**CS**) analysis. Specifically, we fit model (9) to the baseline data and then use the parameter estimates to compute the probability of dementia at wave 4 based on screening scores and other covariates observed at that wave. The three parameters

that measures how the CSID predicts dementia are β_S , β_{DLN} and λ_S . The estimates of the three parameters from the FL are -0.107, 0.108 and -0.150, all of which have p -value less than 0.001. Then, after some simple calculation based on the estimate of σ^2 , these estimates translate into the following interpretation:

1. Conditional on the CSID decline from last wave, normal at last wave and other covariates, the odds ratio of being demented now is 2.05 for every one standard deviation decrease in current CSID score ($OR_S = 2.05$)
2. Conditional on the current CSID, normal at last wave and other covariates, the odds ratio of being demented now is 2.07 for every one standard deviation increase in CSID decline from last wave ($OR_{SDLN} = 2.07$)
3. Conditional on the covariates, the odds ratio of being demented at baseline is 2.74 for every one standard deviation decrease in baseline CSID ($OR_{SBASE} = 2.74$)

We show the ROC curves based on computed probabilities for the three approaches (FL, CL and CS) in Figure 1. Although the FL and CL are slightly better than the CS, the difference is negligible. It seems that incorporating the history of CSID/diagnosis does not improve the prediction accuracy substantially. It might be that the CSID itself is already fairly good in selecting the demented subjects so that including the history of screening/diagnosis does not help much. Then the question is whether or not the incorporation of this information is helpful when the screening is not as accurate. Another possibility is that we did not include useful covariates in the model or the model structure is not close to the real mechanism so that the overall prediction accuracy of the three approaches is undermined to various extent.

To address these issues, we conduct a simulation study. We generate data sets

based on models (7)- (10) and four sets of parameters. The distribution of the covariates are generated according to the distributions found in the Indianapolis data from IIDP. Specifically, the four sets of parameters are the same except that (i) $(OR_S, OR_{SDLN}, OR_{SBASE})$ are set to various values to reflect difference in screening accuracy, and (ii) for each set of values in (i), $(\alpha_C, \beta_C, \lambda_C, \alpha_Y)$ are adjusted so that the number of subjects and cases at each wave for each set in (i) are close. For each set of the parameters, we generate 100 data sets that is composed of 2000 subjects and 4 waves of screening/diagnosis. Then data from the first three waves are used to estimate parameters in FL and CL, and data from the baseline is used to estimate the parameters in CS. Prediction is again based on computed probabilities. The results are summarized in Table III. For the first three simulations, the values of $(OR_S, OR_{SDLN}, OR_{SBASE})$ are set in a “monotone” way to represent a decreasing trend in screening accuracy in the sense that both absolute screening score and its decline has less prediction capability as we proceed from simulation 1 to 3. It can be seen that the difference in AUC between FL and CS goes from 2.9% in Simulation 1 to 5% in Simulation 3. Thus, the improvement in overall prediction accuracy is more apparent when the screen is not very efficient. When looking at the sensitivity at the high specificity end, there is quite some difference between FL and CS, ranging from 5% to close to 10%. In general, the gain in sensitivity is more when the specificity is fixed at higher level. The improvement in sensitivity at fixed specificity level is practically meaningful because often the purpose of implementing a screening is to maximize the sensitivity at a tolerable specificity level. For the fourth simulation, both OR_S and OR_{SBASE} are the same as simulation 1 whereas OR_{SDLN} is set to 1. The purpose of

this simulation is to investigate whether or not the improvement of FL/CL over CS seen in previous simulations is simply due to the fact that cognitive decline is included in the FL and CL. If this is true, we should not see much difference between FL/CL and CS in simulation 4. However, as can be seen from Table III, FL/CL still predict better than CS, though at an attenuated level compared with simulation 1. Therefore, the gain in prediction accuracy seen in simulation 1-3 is at least partially attributed to the inclusion of screening history.

Simulation 1 is based on the exact same parameter estimates from Indianapolis data from IIDP. The AUC shows a higher level than that shown in Figure 1, which implies that we might have ignored covariates that predict dementia or the model structure deviates from the truth. Such deviations seem to have more impact on FL and CL than on CS. The CL approach has very similar AUC as the FL, with around 3% decrease in sensitivity level across simulations and various specificity levels. Hence, it provides a reasonable approach to balance model complexity and accuracy.

5. CONCLUSIONS

In this paper, we formally proposed a statistical model to predict dementia using the screening/diagnosis history in the longitudinal two-phase study setting. It was shown that although the overall gain in prediction accuracy by including such information as compared with a cross-sectional model is small, the sensitivity is improved up to 10% in a range of specificity values that are often desirable for screening purpose. The improvement is more obvious for screening test that are less accurate. We also demonstrated that the performance of the conditional likelihood falls in between the full likelihood and cross-sectional methods. In spite of its motivation by dementia

study, our method presents a general framework that can be used to deal with longitudinal two-phase studies of other diseases. Our study laid out a basis for future research on optimal design of longitudinal two-phase studies in terms of case detection. The underlying principle would be to select subjects for the second phase evaluation according to the probabilities of disease computed based on a formal model of screening/diagnosis history, which essentially is a Bayesian approach. The theoretical significance and practical feasibility of such an “adaptive” design would require further investigation.

ACKNOWLEDGEMENTS

This research was supported by NIH grants: R01 AG15813, R01 AG09956 and P30 AG10133. We would like to thank Dr. Sujuan Gao for helpful discussions and critical comments on the manuscript.

APPENDIX: Derivation of equation (6)

(i) $q = t - 1$

$$\begin{aligned} \Pr[Y_t = 1 | t - 1, S_{(t)}, W \geq t] &= \Pr[Y_t = 1 | Y_{t-1} = 0, S_t, S_{t-1}, W \geq t] \quad (\text{equations (1) - (2)}) \\ &= Q_t(0, 1) \quad (\text{equations (3)}) \end{aligned}$$

(ii) $0 \leq q < t - 1$

Let V be a vector. We denote $V_{a,b} = (V_a, V_{a+1}, \dots, V_b)$ if $a \leq b$ and a null vector if $a > b$. Then

$$\begin{aligned} P = \Pr[Y_t = 1 | q, S_{(t)}, W \geq t] &= \frac{\Pr[Y_t = 1, S_{q+2}, \dots, S_t, W \geq t | q, S_{(q+1)}, W \geq q + 1]}{\Pr[S_{q+2}, \dots, S_t, W \geq t | q, S_{(q+1)}, W \geq q + 1]} \\ &= \frac{\Pr[Y_t = 1, S_{q+2,t}, W \geq t | q, S_{1,q+1}, W \geq q + 1]}{\Pr[S_{q+2,t}, W \geq t | q, S_{1,q+1}, W \geq q + 1]}. \end{aligned}$$

According to equations (1)-(3), we have

$$\begin{aligned}
 & [Y_{q+1,t-1}, S_{q+2,t}, W \geq t|q, S_{1,q+1}, W \geq q+1] \\
 = & [Y_{q+1}|Y_q = 0, S_{q+1}, S_q, D_q = 0][D_{q+1}|Y_{q+1}, S_{q+1}, C_{q+1} = 0, D_q = 0] \\
 & \left(\prod_{k=q+2}^{t-1} [S_k|Y_{k-1}, S_{k-1}, D_{k-1} = 0][Y_k|Y_{k-1}, S_k, S_{k-1}, D_{k-1} = 0][D_k|Y_k, S_k, C_k = 0, D_{k-1} = 0] \right) \\
 & [S_t|Y_{t-1}, S_{t-1}, D_{t-1} = 0] \\
 = & Q_{q+1}R_{q+1} \left(\prod_{k=q+2}^{t-1} Q_kR_kP_k \right) P_t \text{ evaluated at } Y_q = 0, Y_{q+1,t-1}, C_{q+1,t-1} = 0, D_{q,t-1} = 0
 \end{aligned}$$

Then we can define

$$\begin{aligned}
 M_j &= \Pr[Y_{q+1,q+j} = 0, Y_{q+j+1,t-1} = 1, S_{q+2,t}, W \geq t|q, S_{1,q+1}, W \geq q+1] \\
 &= Q_{q+1}R_{q+1} \left(\prod_{k=q+2}^{t-1} Q_kR_kP_k \right) P_t \\
 &\text{evaluated at } Y_{q,q+j} = 0, Y_{q+j+1,t-1} = 1, C_{q+1,t-1} = 0, D_{q,t-1} = 0 \\
 j &= 0, 1, \dots, t-1-q
 \end{aligned}$$

Due to the irreversibility property of dementia, we have

$$\begin{aligned}
 M_j &= \Pr[Y_{q+1,q+j} = 0, Y_{q+j+1,t-1} = 1, Y_t = 1, S_{q+2,t}, W \geq t|q, S_{1,q+1}, W \geq q+1] \\
 j &= 0, 1, \dots, t-2-q
 \end{aligned}$$

Then we have

$$\begin{aligned}
 P &= \frac{\sum_{j=0}^{t-1-q} \Pr[Y_{q+1,q+j} = 0, Y_{q+j+1,t-1} = 1, Y_t = 1, S_{q+2,t}, W \geq t|q, S_{1,q+1}, W \geq q+1]}{\sum_{j=0}^{t-1-q} \Pr[Y_{q+1,q+j} = 0, Y_{q+j+1,t-1} = 1, S_{q+2,t}, W \geq t|q, S_{1,q+1}, W \geq q+1]} \\
 &= \frac{\sum_{j=0}^{t-2-q} M_j + M_{t-1-q} \Pr[Y_t = 1|Y_{t-1} = 0, S_{1,t}, W \geq t]}{\sum_{j=0}^{t-1-q} M_j} \\
 &= \frac{\sum_{j=0}^{t-2-q} M_j + M_{t-1-q} Q_t(0, 1)}{\sum_{j=0}^{t-1-q} M_j}
 \end{aligned}$$

REFERENCES

1. J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33:101–116, 1938.
2. L. A. Beckett, P. A. Scherr, and D. A. Evans. Population prevalence estimates from complex samples. *J Clin Epidemiol*, 45(4):393–402, 1992. 0895-4356 Journal Article.

3. G. Dunn, A. Pickles, M. Tansella, and J. L. Vazquez-Barquero. Two-phase epidemiological surveys in psychiatric research. *Br J Psychiatry*, 174:95–100, 1999. 0007-1250 Editorial.
4. W. G. Cochran. *Sampling Techniques*. Wiley, New York, 3rd edition, 1977.
5. W. E. Deming. An essay on screening, or two-phase sampling. *International Statistical Review*, 45:28–37, 1978.
6. P. E. Shrout and S. C. Newman. Design of two-phase prevalence surveys of rare disorders. *Biometrics*, 45(2):549–55, 1989. 0006-341x Journal Article.
7. R. McNamee. Two-phase sampling for simultaneous prevalence estimation and case detection. *Biometrics*, 60(3):783–92, 2004. 0006-341x Journal Article.
8. D. G. Clayton, D. Spiegelhalter, G. Dunn, and A. Pickles. Analysis of longitudinal binary data from multi-phase sampling. *Journal of the Royal Statistical Society Series B*, 60:71–87, 1998.
9. C. Dowrick, P. Casey, O. Dalgard, C. Hosman, V. Lehtinen, J. L. Vazquez-Barquero, and G. Wilkinson. Outcomes of depression international network (odin). background, methods and field trials. odin group. *Br J Psychiatry*, 172:359–63, 1998. 0007-1250 Journal Article Multicenter Study.
10. R.J.A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition, 2002.
11. M. W. McIntosh and M. S. Pepe. Combining several screening tests: optimality of the risk score. *Biometrics*, 58(3):657–64, 2002. 0006-341x Journal Article.
12. J. Newman and E. S. Pearson. On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337, 1933.
13. H. C. Hendrie, B. O. Osuntokun, K. S. Hall, A. O. Ogunniyi, S. L. Hui, F. W. Unverzagt, O. Gureje, C. A. Rodenberg, O. Baiyewu, and B. S. Musick. Prevalence of alzheimer’s disease and dementia in two communities: Nigerian africans and african americans. *Am J Psychiatry*, 152(10):1485–92, 1995. 0002-953x Journal Article.
14. K. S. Hall, S. Gao, C. L. Emsley, A. O. Ogunniyi, O. Morgan, and H. C. Hendrie. Community screening interview for dementia (csi ’d’); performance in five disparate study sites. *Int J Geriatr Psychiatry*, 15(6):521–31, 2000. 0885-6230 Journal Article.
15. C. Shen and S. Gao. A mixed-effects model for cognitive decline with non-monotone non-response from a two-phase longitudinal study of dementia. *Statistics in Medicine (in press)*, 2005.
16. C. Shen and L. Weissfeld. Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. *Biostatistics*, 6(2):333–47, 2005. 1465-4644

Journal Article.

17. J. A. Nelder and R. Mead. A simplex method for function minimisation. *Computer Journal*, 7:303–313, 1965.

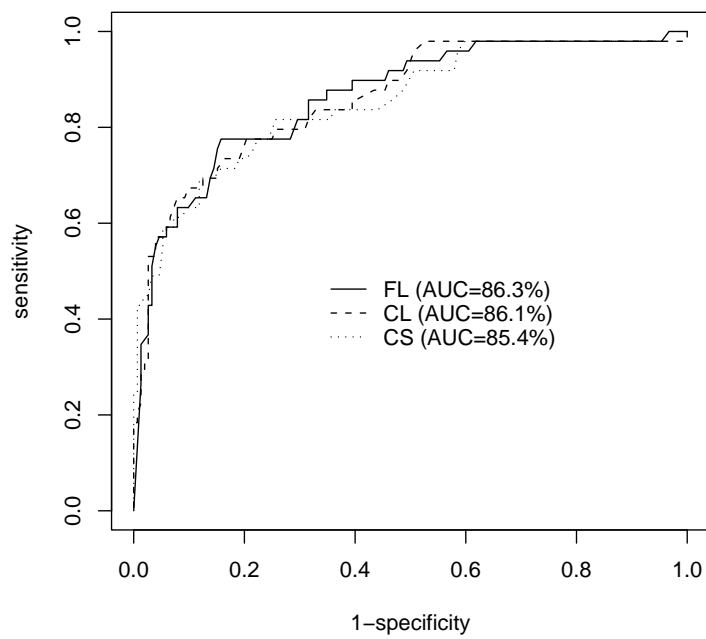


Figure 1. ROC curves based on probabilities of dementia computed by Full Likelihood (FL), Conditional Likelihood (CL) and Cross-Sectional (CS) approaches. Curves are based on 201 subjects (49 cases) selected for diagnosis at wave 4 from the Indianapolis cohort of the Indianapolis-Ibaden Dementia Project

Table I. Summary of demographic factors, CSID score and number of dementia cases detected. Mean (S.E.) is shown for continuous variables and percentage is shown for binary variable

	wave 1 (baseline)	wave 2	wave 3	wave 4
subjects in the study	2115	1648	1181	686
age	74.0 (7.0)	75.2 (6.7)	77.4 (6.4)	80.0 (5.6)
female	64.7%	66.3%	68.8%	72.3%
years of education	9.6 (3.1)	9.8 (3.0)	10.0 (3.0)	10.4 (2.7)
CSID	65.6 (9.2)	67.5 (7.2)	63.7 (9.1)	63.9 (10.2)
subjects selected for clinical evaluation	327	218	228	201
dementia cases identified	59	22	60	49

Table II. Demonstration of some patterns of (Y_i, C_i, D_i) for 3 waves. Numbers in bold font are not observed

id	Y_1	C_1	D_1	Y_2	C_2	D_2	Y_3	C_3	D_3
1	0	0	0	0	1	0	0	0	0
2	0	0	1	1		1	1		1
3	1	1	1	1		1	1		1
4	1	0	0	1	1	1	1		1
5	0	0	0	0	0	0	1	0	0
6	1	0	0	1	0	0	1	1	1

Table III. Summary of simulation results for Full Likelihood (FL), Conditional Likelihood (CL) and Cross-Sectional (CS) approaches (100 runs, $n=2000$, 4 waves, SEN: sensitivity, SPC: specificity)

mean (S.E.)	FL	CL	CS
Simulation 1: $OR_S = 2.05$, $OR_{SDLN} = 2.07$, $OR_{SBASE} = 2.74$			
AUC	0.925 (0.019)	0.913 (0.022)	0.896 (0.024)
SEN (SPC=0.9)	0.771	0.744	0.680
SEN (SPC=0.85)	0.840	0.812	0.775
SEN (SPC=0.8)	0.885	0.857	0.830
Simulation 2: $OR_S = 1.29$, $OR_{SDLN} = 1.30$, $OR_{SBASE} = 1.72$			
AUC	0.824 (0.025)	0.806 (0.026)	0.786 (0.029)
SEN (SPC=0.9)	0.512	0.489	0.440
SEN (SPC=0.85)	0.604	0.568	0.539
SEN (SPC=0.8)	0.671	0.645	0.612
Simulation 3: $OR_S = 1.11$, $OR_{SDLN} = 1.12$, $OR_{SBASE} = 1.48$			
AUC	0.785 (0.024)	0.768 (0.024)	0.735 (0.031)
SEN (SPC=0.9)	0.452	0.425	0.356
SEN (SPC=0.85)	0.539	0.509	0.445
SEN (SPC=0.8)	0.609	0.578	0.522
Simulation 4: $OR_S = 1.11$, $OR_{SDLN} = 1$, $OR_{SBASE} = 1.48$			
AUC	0.949 (0.011)	0.942 (0.012)	0.928 (0.017)
SEN (SPC=0.9)	0.836	0.817	0.775
SEN (SPC=0.85)	0.895	0.875	0.841
SEN (SPC=0.8)	0.930	0.915	0.888