Published in final edited form as:
Contemp Clin Trials. 2011 September ; 32(5): 685–693. doi:10.1016/j.cct.2011.04.007.

The relative efficiency of time-to-threshold and rate of change in longitudinal data

M. C. Donohue[*,a], **A. C. Gamst**[b], **R. G. Thomas**[b], **R. Xu**[c], **L. Beckett**[d], **R. C. Petersen**[e], **M. W. Weiner**[f], **P. Aisen**[b], and **the Alzheimer's Disease Neuroimaging Initiative**[g]

[a]9500 Gilman Dr. MC 0949 Division of Biostatistics and Bioinformatics University of California, San Diego La Jolla, CA 92093-0949

[b]Department of Neuroscience, University of California, San Diego, CA

[c]Department of Mathematics, University of California, San Diego, CA

[d]Department of Public Health Sciences, University of California, Davis, CA

[e]Department of Neurology, Mayo Clinic College of Medicine, Rochester, MN

[f]Department of Radiology, University of California, San Francisco, CA

[g]Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI) on November 30, 2009. 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. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Data/ADNIAuthorshipList.pdf).

Abstract

Randomized, placebo-controlled trials often use time-to-event as the primary endpoint, even when a continuous measure of disease severity is available. We compare the power to detect a treatment effect using either rate of change, as estimated by linear models of longitudinal continuous data, or time-to-event estimated by Cox proportional hazards models. We propose an analytic inflation factor for comparing the two types of analyses assuming that the time-to-event can be expressed as a time-to-threshold of the continuous measure. We conduct simulations based on a publicly available Alzheimer's disease data set in which the time-to-event is algorithmically defined based on a battery of assessments. A Cox proportional hazards model of the time-to-event endpoint is compared to a linear model of a single assessment from the battery. The simulations also explore the impact of baseline covariates in either analysis.

Keywords

longitudinal data; survival analysis; linear mixed models; marginal linear models; power

[*] *Preprint submitted to Contemporary Clinical Trials*.

1. Introduction

We explore the relative efficiency of linear models of repeated measures of a continuous outcome and the Cox proportional hazards model (PHM) [1] of time-to-threshold of a continuous outcome in randomized placebo-controlled studies. This comparison has practical implications for clinical trial design for Alzheimer's disease (AD) and human immunodeficiency virus (HIV), among other diseases. For instance, in the study of AD in pre-dementia elderly with mild cognitive impairment (MCI), clinical trials have historically used either the rate of progression to dementia over a period of time (typically three to 24 months) [2] or PHM of time-to-progression [3, 4, 5] as the primary analysis. There is no algorithmic definition of dementia, however, and in clinical trials a panel of experts is convened to review case reports to determine a consensus diagnosis at each visit (usually every six months). The time-to-progression endpoint has been preferred for its tangible clinical importance, as well as its acceptability to regulatory authorities. Though the dementia endpoint has face validity, it can be difficult to implement, subjective, variable from visit to visit, and analytically problematic due to non-proportional hazards [3] and interval censoring. We posit that a linear model of a continuous assessment of disease severity, for example the Alzheimer's Disease Assessment Scale, may be more efficient than a subjective dichotomization (“not demented” versus “demented”). To this end, we quantify the relative efficiency of a linear model analysis of rate of change of a repeated continuous measure and a PHM analysis of time-to-threshold. “Time-to-threshold” is also known as “time-to-event” in survival analysis literature, but we use the former to emphasize that we will model the event of interest as an observed continuous measure exceeding a predetermined threshold.

The issue is not new or unique to AD research. McKay et al. [6] analyzed continuous, categorical, and time-to-event cocaine use outcomes and found continuous outcomes to express the greatest effect sizes. A meta-analysis of the orthopedic surgery randomized trial literature found those trials with continuous outcomes had greater power on average than those with a dichotomous outcome, an outcome analytically equivalent to time-to-event [7], and a greater proportion of the continuous outcomes trials attained acceptable power (>80%) [8]. Similar observations were made in the fields of rheumatoid arthritis [9] and stroke [10]. Reliable continuous biomarker surrogates have accelerated the study of HIV [11], and are still actively being sought, for example, for prostate cancer [12, 13] and AD [14].

Lee and Whitmore [15] provide an extensive review of threshold regression or first-hitting-time models which are used to analyze the relationship between covariates and the time at which an observed or latent stochastic process first crosses a boundary. Though we will be exploiting aspects of this literature, we are not proposing a threshold regression method. Rather, we consider cases in which the threshold might be considered an arbitrary dichotomization of an observable continuous process. Such dichotomizations may facilitate interpretation, but it is our primary goal to elucidate whether this ease of interpretation comes at the cost of analytic efficiency.

Section 2 introduces an inflation factor for quantifying relative efficiency, in terms of the required sample size, for the true marginal linear model (MLM) [16] and the PHM in general terms when we can assume an underlying Wiener process with drift. Section 3 provides simulation studies to demonstrate the utility of the inflation factor, and other comparisons for which the inflation factor does not directly apply because the underlying process is not a Wiener process or is not linear. In the simulations we apply linear mixed models (LMM) [17] that are commonly used in practice. In Section 4 we present an example of an event, onset of dementia, defined by multiple continuous outcomes based on publicly available data from a large MCI cohort.

2. Inflation factor for PHM versus MLM assuming underlying Wiener process with drift

Assume that clinical disease progression for an individual i follows an underlying Wiener process with drift,

$$Y_i(t) = t\theta + \sigma W_{it}, \tag{1}$$

where $i = 1, \ldots, n$, $t > 0$, θ is a treatment-specific modifying effect on the rate of decline, and W_{it} is a standard Wiener process. The advantage of model (1), is that it allows a closed form expression for the distribution of the time-to-threshold, as seen below. Model (1), which was considered in [18], also allows for the variance of $Y_i(t)$ to increase with time as in a mixed-effect model with subject specific random slopes. We can observe disease progression in one of two ways: as continuous repeated measures with added error σW_t or as the time that the measurements cross a threshold c, $T_i = \min_t\{Y_i \geq c\}$. Each measure can be thought of as an imperfect observation of an underlying and unobservable latent variable representing, for example, disease state. We assume that the process is observed at times t_j, where $j = 1 < \ldots < m$ and there are two groups, A and B, with possibly different slope parameters, θ_A and θ_B. We will assume throughout that the two groups have equal sample size, n. This observed data is analyzed in one of two ways: a PHM of time-to-threshold or a MLM of the continuous repeated measures. The parameter of interest under PHM is the hazard ratio for the two groups, with inference obtained by the score test (equivalent to the log-rank test). The parameter of interest under the MLM is the estimated group difference in slopes: $\theta_B - \theta_A$. More specifically, we apply the PHM, which assumes an instantaneous probability of event via the hazard function

$$\lambda(t;Z_i)\,dt = P[t \leq T_i < t+dt | T_i \geq t; Z_i] = \lambda_0(t)\ \exp\{Z_i\theta_{\mathrm{HR}}\}\,dt, \tag{2}$$

where T_i is defined above, $Z_i = 1\{\text{subject } i \text{ in group } A\}$ is a group indicator variable, and θ_{HR} represents the log hazard ratio. Alternatively, we model the observations for an individual from group A with the MLM:

$$Y_{ij} = t_j\theta_A + \epsilon_{ij}; \quad i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{3}$$

with independently and identically distributed subject-specific vectors of residual errors, $(\varepsilon_{i1} \ldots \varepsilon_{im}) \sim N(0, \Sigma)$; and similarly for individuals from group B. Note that, by the properties of the assumed Wiener process, $\mathrm{VAR}(Y_{ij}) = \sigma^2 t_j$, $\mathrm{COV}(Y_{ij}, Y_{ik}) = \sigma^2 \mathrm{COV}(W_{tj}, W_{tk}) = \sigma^2 \min(t_j, t_k)$, and $\mathrm{COR}(Y_{ij}, Y_{ik}) = \min(t_j, t_k) / \sqrt{t_j t_k}$. Equivalently, the true variance-covariance matrix is

$$\Sigma = \sigma^2 \begin{pmatrix} t_1 & t_1 & t_1 & \cdots & t_1 \\ t_1 & t_2 & t_2 & \cdots & t_2 \\ t_1 & t_2 & t_3 & & \\ \vdots & \vdots & & \ddots & \\ t_1 & t_2 & & & t_m \end{pmatrix} \tag{4}$$

Assuming equal group sizes, the required total *number of events* for a two-tailed Cox proportional hazards score test with specified power $1 - \beta$, Type I error α, and log hazard ratio θ_{HR} can be estimated for the PHM design using Schoenfeld's [19] formula:

$$E_{\mathrm{PH}}=\frac{4\left(z_{\alpha/2}+z_{\beta}\right)^{2}}{\theta_{\mathrm{HR}}^{2}}, \tag{5}$$

where $p = \phi(z_p)$ and ϕ is the standard normal cumulative distribution function. Similarly, we can use the formula from [20] for the total sample size under the MLM

$$n_{\mathrm{LM}}=\frac{4\left(z_{\alpha/2}+z_{\beta}\right)^{2}\xi}{\left(\theta_{B}-\theta_{A}\right)^{2}} \tag{6}$$

where

$$\xi=(0\quad 1)\left[\begin{pmatrix}1 & \cdots & 1\\ t_1 & \cdots & t_m\end{pmatrix}\Sigma^{-1}\begin{pmatrix}1 & t_1\\ \vdots & \vdots\\ 1 & t_m\end{pmatrix}\right]^{-1}\begin{pmatrix}0\\ 1\end{pmatrix}. \tag{7}$$

In order to relate the two sample sizes, we need to represent the effect under one model as an effect under the other. The time-to-threshold assumption, i.e. $T_i = \min_t\{Y_i \geq c\}$, provides the framework for discerning this connection under model (1). Since T_i is a so-called "first passage time" of a Wiener process with drift, it is known to have an inverse Gaussian (or Wald) distribution with location parameter $\mu = c/\theta$ and scale parameter $\gamma = c^2 / \sigma^2$ [21]. The resulting event time distribution function is

$$F(t;\mu,\gamma)=P[T\leq t]=\Phi\left(\sqrt{\gamma/t}\,(t/\mu-1)\right)+\exp(2\gamma/\mu)\,\Phi\left(-\sqrt{\gamma/t}\,(t/\mu+1)\right). \tag{8}$$

Let the random variable T_A represent the time-to-threshold for an individual randomized to group *A* and T_B represent that of an individual randomized to group *B*. Then for any time *t*, under the PHM, the log hazard ratio is

$$\begin{aligned}\theta_{\mathrm{HR}} &=\log\left(\frac{\log\left(P\left[T_A>t\right]\right)}{\log\left(P\left[T_B>t\right]\right)}\right)\\ &=\log\left(\frac{\log\left(1-F_A(t)\right)}{\log\left(1-F_B(t)\right)}\right).\end{aligned} \tag{9}$$

If we let *r* denoted the overall event rate, so that $n_{\mathrm{PH}} = E_{\mathrm{PH}}/r$, and substitute the above expression for θ_{HR} into (5), we have the inflation factor

$$\psi=\frac{n_{\mathrm{PH}}}{n_{\mathrm{LM}}}=\frac{E_{\mathrm{PH}}}{rn_{\mathrm{LM}}}=\frac{\left(\theta_{B}-\theta_{A}\right)^{2}}{\xi r}\log\left(\frac{\log\left(1-F_{A}(t)\right)}{\log\left(1-F_{B}(t)\right)}\right)^{2} \tag{10}$$

Note that the mean event rate *r*, assuming no censoring up to the maximum follow-up time τ, can be expressed in terms of *F* as $r = (F_A(\tau) + F_B(\tau))/2$. The modal design in AD clinical trials has every subject followed for a predetermined maximum duration, but dropout, which is common, must also be considered. Also note that the relation (9) should be constant in *t* to satisfy the proportional hazards assumption. The fact that it is not constant implies we should expect the assumption to be violated for time-to-threshold of an underlying linear process (3). Inference based on the PHM score (log-rank) test is still valid under the null

hypothesis, but non-proportionality can introduce bias that might affect power [22]. On left panel of Figure 1 we plot the hazard ratio $\exp(\theta_{HR}(t))$ for various values of c with $\sigma = 0.5$, $\theta_A = 0.2$, and $\theta_B = 0.1$. The right panel depicts the resulting total sample sizes. We see that even though the lower thresholds result in greater event rates, which should improve power and reduce sample size, it can also shrink the hazard ratio closer to one and the net effect can be a decrease in power. This is not always the case, as we see the sample size curves for $c = 2.0$ and $c = 3.0$ intersect.

The plots in Figure 2 demonstrate that Ψ is not always greater than 1, which would indicate that MLM generally dominates PHM in efficiency. However, the only cases we found in which the inflation factor favored PHM were in impractical scenarios in which the required sample size approached zero due large effect size (or small variance).

3. Simulations

3.1. A Wiener process

We generated data based on a Wiener process with drift as in (1). Group A (placebo) had slope parameter $\theta_A = 0.2$ and Group B (active) had slope parameter $\theta_B = 0.1$, yielding more rapid progression in Group A. We also assumed $\sigma = 0.5$, $t = 1; \ldots ; 10$, and varied the threshold from 1/2 to 3, yielding expected mean event rates in the placebo group from 85.5% (low threshold for event) to 20.9% (high threshold).

Using (6) and the known data generating parameters, we calculated the required sample size to be $n = 87.2$ under a MLM ($\alpha = 5\%$, power = 80%) with a correlation structure as in (3). Alternatively, given $\sigma = 0.5$ and applying a threshold of $c = 1$, the log hazard ratio θ_{HR} ranged from $\theta_{HR} = 0.371$ at $t = 1$ to $\theta_{HR} = 0.483$ at $t = 10$, which translates to a hazard ratio in the range $\exp(\theta_{HR}) = 1.45$ to 1.62. If we assume no loss to follow-up, resulting in an overall event rate of $(F_A(10) + F_B(10))/2 = 80.4\%$; we arrive at a required sample size under PHM in the range of $n = 168$ to 284, or an inflation of factor in the range of $\Psi = 1.92$ to 3.26.

To study the accuracy of these sample size estimates under reasonable departures from the presumed MLM model, we simulated 1000 trials with total sample sizes of $n = 90$, $n = 170$, and $n = 290$; to examine whether the linear model attains simulated power of 80% with $n = 90$. Rather than using the known Wiener process correlation structure, we used the common mixed effects model with random intercept and slope:

$$Y_{ij} = \theta + b_{0i} + t_j(\theta_A + b_{1i}) + \epsilon_{ij}. \tag{11}$$

In contrast to the marginal model, ε_{ij} were assumed to be independently distributed $N(0, \sigma^2)$ and the within subject correlation was modeled by the random intercept and slope, b_{0i} and b_{1i}. Also in contrast to the marginal model covariance structure of the MLM (4), the LMM assumes $\text{VAR}(Y_t) = \text{VAR}(b_0)^2 + t^2\text{VAR}(b_1)^2 + \text{VAR}(\varepsilon)$ and $\text{COV}(Y_s, Y_t) = \text{VAR}(b_0) + st\text{VAR}(b_1) + (s + t)\text{COV}(b_0, b_1)$. Though the correlation structures are not identical, this model is particularly appropriate since it models a variance that grows with time, as in the Wiener process.

Table 1 demonstrates that the calculated power as described earlier (lower half of the table) is very consistent with the simulated results (upper half of the table). We also simulated data assuming no treatment ($\theta_A = \theta_B = 0.2$) to verify the specified Type I error (α). In particular the power of 77.9% with LMM at $n = 90$ was close to the power of 83.5% with PHM at $n = 170$ and a threshold of $c = 1$. This supports the conservative estimate of an inflation factor of $\Psi = 1.92$. We also see that at all sample sizes, power for PHM reached a maximum at a

threshold of $c = 2$. This is also reected in Figure 1 where we see the sample size curves for $c = 2$ and $c = 3$ intersect.

3.2. An autoregressive process

We repeated the previous simulation study, but generated data based on an autoregressive model, $Y_i(0) \sim N(0, 1)$, $Y_i(t) = \theta + Y_i(t-1) + \varepsilon_i(t)$, where $t = 1, \ldots, 10$ and $\text{VAR}(\varepsilon) = 0.25$. The event threshold was again set at 1 and we let θ, the slope parameter for each group, be either $\theta_B = 0.2$ or $\theta_A = 0.1$. Because we did not assume an underlying Wiener process, the survival times were no longer distributed according to the inverse Gaussian distribution. To approximate the required sample size under both models, we generated a large sample ($n = 2000$) and fit a mixed effects model with random intercept and slope to inform the LMM calculation and used the observed event rates to inform the PHM sample size calculation. Using pilot estimates from the LMM fit and assuming $\text{VAR}(Y_t) \approx \text{VAR}(b_0) + t^2\text{VAR}(b_1) + \text{VAR}(\varepsilon)$ and $\text{COV}(Y_s, Y_t) \approx \text{VAR}(b_0) + st\text{VAR}(b_1) + (s + t)\text{COV}(b_0, b_1)$, we arrive at a marginal model variance-covariance matrix that we can apply to (6). We found a sample size of $n_{\text{LM}} = 189$ was necessary to attain power 80% with $\alpha = 5\%$ and a two tailed test. We also found an event rate of $\hat{r} = 71.7\%$ and $\widehat{\theta}_{\text{HR}}(10) = 0.259$. Applying these estimates to (5), we find $n_{\text{PH}} = 652$. Therefore we simulated 1000 trials with the sample sizes of 170 and 650, and again modeled the data with either PHM (log-rank test) or LMM with random slope and intercept. Table 2 summarizes the results. As expected, the simulations attained about 80% power with $n = 170$ under LMM and $n_{\text{PH}} = 650$ under PHM, verifying the inflation factor of about $\Psi = 3.82$.

3.3. A non-linear trajectory

Next we simulated continuous longitudinal data according to a non-linear trajectory with random intercepts and slopes that flation out once a threshold is met. More specifically, for an individual i in group A at time t_j :

$$Y_{ij} = \begin{cases} b_{0i} + (\theta_A + b_{1i})\, t_j + \varepsilon_{ij} & \text{if} \quad b_{0i} + (\theta_A + b_{1i})\, t_j < c \\ c + \varepsilon_{ij} & \text{if} \quad b_{0i} + (\theta_A + b_{1i})\, t_j \geq c \end{cases}$$

where $t = 1, \ldots, 10$, $b_{0i} \sim N\left(0, \sigma_0^2\right)$, $b_{1i} \sim N\left(0, \sigma_1^2\right)$, and $\varepsilon \sim N(0, \sigma^2)$. We simulate two groups with $\theta_A = 0.2$ and $\theta_B = 0.1$ and let $\text{VAR}(b_0) = \text{VAR}(b_1) = 0.1$ and $\text{VAR}(\varepsilon) = 0.25$. We varied c from 1/2 to 3.

To the longitudinal data we applied two misspecified linear models: (1) the random intercept and slope LMM as used in the previous examples, and (2) a random intercept and slope model with quadratic fixed effect for time allowing for a non-linear trajectory (LMM2). The parameter of interest from LMM is the group difference in slopes. The parameter of interest from LMM2 is the estimated group difference at $t = 10$. Finally, we used the known value of c as the threshold to define the events to be modeled via PHM. With c in the range 1/2 to 3, we found overall event rates in the range 87.5% (low threshold) to 8.9% (high threshold). We let $n = 100$ and 200.

The results are summarized in Table 3. We found there was a clear advantage to PHM when there was a low threshold and high event rate, but this reversed as the threshold increased and event rate decreased. We also see that the quadratic time model, LMM2, was consistently better than the standard LMM, especially when the threshold was low.

4. Example: Mild Cognitive Impairment and time-to-progression

The Alzheimer's Disease Neuroimaging Initiative (ADNI), which began in 2004, is a collaborative project funded by National Institute on Aging and National Institute of Bioimaging and Bioengineering, the pharmaceutical and imaging industry, and several foundations (see *www.adni-info:org*). The study design and baseline characteristics are described in [23]. Briefly, the objective of ADNI is to study the rate of change of cognition, function, brain structure, and biomarkers in 200 elderly controls, 400 subjects with MCI, and 200 with Alzheimer's disease. For this analysis, publicly available data were downloaded from the ADNI web site *www.loni.ucla:edu/ADNI* on November 30, 2009. The data set contains repeated continuous measures of key assessments and progression events at 6-month intervals over 2 to 3 years, and is ideal for a more complex, clinically realistic simulation of our comparison of interest. Namely, we will simulate clinical trials to determine which experimental design can more efficiently detect a hypothesized intervention to slow cognitive and functional decline in a population with MCI.

In clinical practice and trials, the dementia endpoint is not algorithmically defined. It is a subjective transition based on the review of a battery of cognitive and functional assessments. Studies typically employ the consensus opinion of an expert panel. We took advantage of the rich ADNI data to develop a multivariate mixed-effects model for disease progression using multiple cognitive and functional measures, and to develop an algorithmic definition of progression for this process using the observed clinical diagnosis data. Our model of disease progression, richer than that hypothesized in Section 2, incorporated multiple measures: Alzheimer's Disease Assessment Scale, Cognitive Sub-scale, (ADAS-$_{COG}$; [24, 3]), Clinical Dementia Rating Sum of Boxes (CDR-SB, log transformed), and Functional Activities Questionnaire (FAQ) [25]. These measures were selected because they provide assessment of primary aspects of AD progression: cognitive performance, global clinical status and functional abilities. Each measure is converted to a z score to provide a common scale, and a multivariate mixed-effects model is fitted [26] to estimate mean rates of change, random variation in slopes and intercepts, and effects on slopes and intercepts of the presence of an apolipoprotein E4 (ApoE4) allele and baseline hippocampal volume. Specifically, the mixed effects model is of the form:

$$Y_{ik}\left(t_{ijk}\right)=X_{ik}t_{ijk}\beta+Z_{ik}b_{0i}+Z_{ik}t_{ijk}b_{1i}+\varepsilon_{ijk} \tag{12}$$

for individual i, at time t, outcome k = 1, 2, 3 (ADAS-$_{COG}$, CDR-SB, or FAQ), and covariate vectors

$$X_{ik}=\left(1\{k=1\},1\{k=2\},1\{k=3\},\text{Hippocampus}_i,\text{ApoE4}_i\right),Z_{ik}=\left(1\{k=1\},1\{k=2\},1\{k=3\}\right),$$

where Hippocampus$_i$ and ApoE4$_i$ are the baseline the standardized hippocampal volume and ApoE4 status for individual i. The vector β represents the fixed effects for time for each of the 3 outcomes and the shared time-by-hippocampus and time-by-ApoE4 effects. The 3 random intercepts and 3 random slopes for the 3 outcomes are represented by the vector b_i = (b_{0i}, b_{1i}), which is assumed to distributed $N(0, \Sigma)$. The residuals, ε_{ijk}, are assumed independent $N(0, \sigma^2)$.

Because progression to dementia is subjective and not algorithmically defined, we derived a diagnostic algorithm for progression diagnosis based on baseline and follow-up ADAS-$_{COG}$, CDR-SB, and FAQ z-scores, using a repeated binary outcome Generalized Estimating

Equation (GEE) logistic regression model [27] we regressed the observed progression outcomes on the z-scores:

$$\text{logit}\left(W_{ij}\right) = \left(\text{ADAS}_{i0}, \text{CDR}_{i0}, \text{FAQ}_{i0}, \text{ADAS}_{ij}, \text{CDR}_{ij}, \text{FAQ}_{ij}\right)\beta. \tag{13}$$

Here W_{ij} = 1 if progression to AD is observed for individual i at time t_j and 0 otherwise. The right hand side of model (13) provides a continuous linear predictor of progression, to which a progression threshold can be applied. The progression threshold was tuned to produce about a 40% progression rate over two years in simulated placebo group data, comparable to the actual progression rate observed in ADNI. Our modeled progression rule was in agreement with actual clinical decisions for 315/391 = 80.6% of MCI subjects with follow-up data. The sensitivity and specificity of the algorithm for detecting clinical progression decisions was 115/134 = 85.8% and 200/257 = 77.8%.

We simulated data based on the multivariate linear mixed model (12) to produce simultaneous cognitive and functional measures. All three simulated measures were then entered into our derived progression algorithm, the predictive model (13). We also added a treatment effect to model (12) resulting in a 25% or 50% reduction in the rate of decline. We then apply LMMs to the simulated continuous outcomes to derive an estimated treatment effect for ADAS-$_{\text{COG}}$ and CDR-SB. Likewise, we applied the PHM to the simulated progression events to estimate the treatment effect on the time-to-progression. Note that the PHM utilized information from two assessments that are not available to the two univariate LMMs for longitudinal ADAS and CDR.

We also explored the efficiency of a pre-specified sample enrichment strategy in which the inclusion criteria requires subjects to exhibit amyloid beta (Aβ) dysregulation at baseline. Such a strategy would be particularly appropriate for testing anti-amyloid interventions. Simulations were repeated using estimates from the ADNI MCI subgroup, which we denote MCI-Aβ, defined by a cerebral spinal fluid (CSF) $A\beta_{1-42}$ cutpoint of 192 pg/Ml, independently derived by [28]. We also used baseline FreeSurfer hippocampal volumes provided by University of California, San Francisco, and serial ADAS-$_{\text{COG}}$, CDR-SB, and FAQ assessed every six months for two years. The available sample size with complete data necessary for estimating the model parameters was n = 393 for MCI and n = 144 for MCI-Aβ.

Dropout was simulated by assuming exponentially distributed dropout times resulting in about 30% attrition over 2 years. This is a conservative estimate of dropout consistent with the 230/769 = 29.9% dropout rate observed in the 3-year donepezil and vitamin E trial [3]; and the 656/1457 = 45% dropout rate observed in the 4-year Rofecoxib trial [4].

We simulated data from 1000 clinical trials over a range of sample sizes, analyzed using LMM and PHM with and without presence of an ApoE4 allele and/or baseline hippocampal volumes, and estimated statistical power by the proportion of trials that rejected the null hypothesis of no treatment effect (p¡0.05). The model fitting and simulation were done in the **R** statistical computing environment [29].

4.1. Results

Figure 3 summarizes the results in terms of power per total sample size n from simulated trials in MCI populations (bottom 2 panels) and MCI-Aβ populations with amyloid dysregulation (top two panels). Results from a simulated 25% treatment effect are displayed on the left and results from a 40% treatment effect are displayed on the right. The LMM results ("○" and "Δ") are clearly separated from the PHM results ("+"), demonstrating

consistently greater power across all sample sizes simulated. Including baseline hippocampal volumes or ApoE4 status (not shown) provides a small, but consistent, improvement in power that is more delineated in the MCI population.

5. Discussion

We found a quantifiable degradation of power with PHM compared to the alternative linear models in our scenarios, except when the underlying data was nonlinear and event rate was high. The inflation factor (10) demonstrates that this degradation is a function of the event rate, r, and the log hazard ratio, which can be expressed as a function of the threshold, slope, and variance parameters from an assumed underlying Wiener process with drift. The simulations also showed that the MLM power calculations, assuming known variance-covariance matrix, provided good estimates for the LMM. The autoregressive simulations demonstrated that power under the PHM was not monotone in the threshold or event rate. The MCI example showed that the degradation of power with PHM can have meaningful impact on the efficiency and costs of clinical trials in a realistic setting, even when clinical diagnosis is based on more outcome data than a single quantitative outcome measurement. These costs and comparisons should be considered, along with face validity, when evaluating the choice of endpoint in clinical trials.

In addition to the loss of power to detect a treatment effect, the MLM and LMM are generally more appropriate, robust, and efficient in many settings, particularly in studies of AD in MCI populations. The standard PHM analysis is not appropriate for the interval censored data that arise in these clinical trials settings; and the linear models obviate any bias that might be introduced by violations of the proportional hazards assumption. PHM also does not account for multi-state transitions, which are common. There are, of course, other analysis techniques that can handle the above issues, and their efficiency relative to the LMM is a question for future study. Another issue left to future study is a direct assessment of the effects of missing data on the inflation factor ψ, though the MCI simulation did attempt to replicate missingness observed in ADNI. Heuristically, the LMM and MLM make more efficient use of partially complete data, which should only amplify their relative efficiency over PHM given missing data. For the same reason, biases induced by informative missingness may be exacerbated by PHM relative to LMM. More specifically, the PHM uses no information regarding changes in performance that are below the threshold of the event of interest, whereas the LMM is informed by such changes. The fact that mixed-models use all available data helps make it robust in the face of data missing at random [30]. Using all of the data also makes the mixed-model less susceptible, relative to the Cox model, to bias induced by missing data mechanisms of all types. These open issues notwithstanding, the LMM and MLM are common, easily accessible, and robust alternatives to PHM; and the proposed inflation factor provides a means for making analytic efficiency comparisons with the PHM.

Acknowledgments

We are grateful to the reviewers and editor for their insightful suggestions. We are also grateful to the ADNI community, including the collaborators at all of the cores, sites, and laboratories; and to the ADNI volunteers and their families. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee

organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

1. Regression models and life-tables.; Journal of the Royal Statistical Society Series B (Methodological). 1972. p. 187-220.URL http://www.jstor.org/stable/2985181
2. Safety and effiacy of galantamine in subjects with mild cognitive impairment. Neurology. 2008; 70(22):2024. ID: 231612481. [PubMed: 18322263]
3. Vitamin e and donepezil for the treatment of mild cognitive impairment. The New England journal of medicine. 2005; 352(23):2379–88. ID: 110248948. [PubMed: 15829527]
4. A randomized, double-blind, study of rofecoxib in patients with mild cognitive impairment. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2005; 30(6):1204–5. ID: 110328423. [PubMed: 15742005]
5. Effect of rivastigmine on delay to diagnosis of alzheimer's disease from mild cognitive impairment: the inddex study. Lancet Neurology. 2007; 6(6):501–512. ID: 442451988. [PubMed: 17509485]
6. Continuous, categorical, and time to event cocaine use outcome variables: degree of intercorrelation and sensitivity to treatment group differences. Drug and alcohol dependence. 2001; 62(1):19. ID: 97362401. [PubMed: 11173164]
7. Fitting cox's regression model to survival data using glim.; Journal of the Royal Statistical SocietySeries C (Applied Statistics). 1980. p. 268-275.URL http://www.jstor.org/stable/2346901
8. Effect of continuous versus dichotomous outcome variables on study power when sample sizes of orthopaedic randomized trials are small. Archives of orthopaedic and trauma surgery. 2002; 122(2): 96. ID: 94143036. [PubMed: 11880910]
9. Comparison of rheumatoid arthritis clinical trial outcome measures: A simulation study. Arthritis and rheumatism. 2003; 48(11):3031. ID: 97804303. [PubMed: 14613263]
10. Use of ordinal outcomes in vascular prevention trials: Comparison with binary outcomes in published trials. Stroke. 2008; 39(10):2817–2823. ID: 424315419. [PubMed: 18669897]
11. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids.; Journal of the American Statistical Association. 1995. p. 27-37.URL http://www.jstor.org/stable/2291126
12. Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. Journal of Applied Statistics. 2003; 30(2):235–247. ID: 366785396.
13. Prostate-specific antigen (psa) alone is not an appropriate surrogate marker of long-term therapeutic benefit in prostate cancer trials. European journal of cancer. 2006; 42(10):1344–1350. [PubMed: 16730974]
14. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. Brain. 2009; 132(5):1355–1365. ID: 362475022. [PubMed: 19339253]
15. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary.; Statistical Science. 2006. p. 501-513.URL http://www.jstor.org/stable/27645791
16. Marginalized multilevel models and likelihood inference. Statistical Science. 2000; 15(1):1–26.
17. Random-effects models for longitudinal data. Biometrics. 1982; 38(4):963–74. ID: 113328916. [PubMed: 7168798]
18. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association. 2001; 96(455) ID: 482249484.
19. Sample-size formula for the proportional-hazards regression model.; Biometrics. 1983. p. 499-503.URL http://www.jstor.org/stable/2531021
20. Analysis of Longitudinal Data. 2 ed.. Oxford University Press; USA: 2002. ISBN 0198524846. URL http://www.worldcat.org/isbn/0198524846
21. The inverse Gaussian distribution: theory, methodology, and applications. M. Dekker; New York: 1989. ISBN 0824779975 9780824779979. ID: 18163863

22. Martingale-based residuals for survival models.; Biometrika. 1990. p. 147-160.doi:10.1093/biomet/77.1.147. URL http://dx.doi.org/10.1093/biomet/77.1.147

23. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. Neurology. 2010; 74(3):201–9. ID: 568064771. [PubMed: 20042704]

24. A new rating scale for alzheimer's disease. The American Journal of Psychiatry. 1984; 141(11): 1356–64. ID: 114863527. [PubMed: 6496779]

25. Measurement of functional activities in older adults in the community. Journal of gerontology. 1982; 37(3):323–9. ID: 115094771. [PubMed: 7069156]

26. Multivariate longitudinal models for complex change processes. Statistics in medicine. 2004; 23(2):231–9. ID: 111663956. [PubMed: 14716725]

27. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42(1):121–30. ID: 115936579. [PubMed: 3719049]

28. Cerebrospinal fluid biomarker signature in alzheimers disease neuroimaging initiative subjects. Annals of Neurology. 2009; 65(4):403–413. ID: 327390388. [PubMed: 19296504]

29. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2009. ISBN 3-900051-07-0; URL http://www.R-project.org

30. Inference and missing data. Biometrika. 1976; 63:581–592.

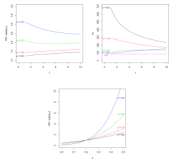

Figure 1.
The hazard ratio $\exp(\theta_{HR})$, as defined in (9) depends on the threshold (c), the Wiener process parameters (σ^2, θ_A and θ_B) and time t. The slopes, as depicted on the left, though not drastic, can dramatically impact the resulting sample size calculation, as depicted on the right. We have found in our Wiener process simulation that choosing a large t results in sample size estimates that are generally consistent with our simulations. In the top panels we set $\sigma = 0.5$, $\theta_A = 0.2$, $\theta_B = 0.1$ and vary t and c. In the bottom panel we $\sigma = 0.5$, $t = 10$, $\theta_B = 0.1$, and vary $\delta = \theta_A - \theta_B$.

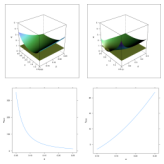

Figure 2.
In the plot on the upper left we set $\sigma = 0.5$, $\theta_A = \theta_B + \delta$, $\theta_B = 0.1$, and compute the inflation factor, Ψ, as a function of δ and c for a PHM with $t = 1, \ldots, 10$ and θ_{HR} evaluated at $t = 10$ according to (9). The plane $\Psi = 1$ is plotted for comparison. The inflation factor surface does in fact dip below 1 for c near 5 and δ between 0.4 and 0.5 (not pictured). However the efficiency gain with PHM in this case is not practical since the required sample size is below $n = 4$ with effect sizes this large (bottom left). In the upper right hand plot we see Ψ as a function of c and σ with $t = \theta_A = 0.1$ and $\theta_B = 0.2$. Again Ψ does achieve values less than 1, but with impractically small values of σ which result in $n = 4$ (bottom right).

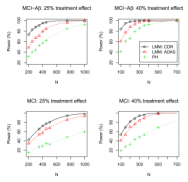

Figure 3.
Simulated power for studies in MCI and MCI with amyloid dysregulation (MCI-Abeta) versus total sample size, *n*. Lines represent LOESS smooths.

Table 1

Power (α) in percent out of 1000 simulated trials for the given total sample size, analysis method, and event threshold. The event rates associated with the thresholds ranged from a mean of 85.5% (c = 1/2) to a mean of 20.9% (c = 3).

	LMM	**PHM**			
		***c* = 0.5**	***c* = 1**	***c* = 2**	***c* = 3**
Simulated					
n = 90	77.9 (5.2)	42.2 (4.5)	56.1 (4.6)	64.4 (5.3)	53.6 (6.6)
n = 170	96.9 (4.7)	64.3 (5.5)	83.5 (4)	90.2 (4.5)	80.1 (5.4)
n = 290	100 (6.9)	87.7 (5.4)	96.5 (4.1)	98.3 (4.4)	95.5 (4.2)
Calculated					
n = 90	81	41	53	63	58
n = 170	97	66	80	88	84
n = 290	99	87	95	98	97

Contemp Clin Trials. Author manuscript; available in PMC 2012 September 1.

Table 2

Power (α) in percent based on 1000 simulated trials for the given total sample size, analysis method, and event threshold. The event rates associated with the thresholds ranged from a mean of 81.8% (c = 1/2) to a mean of 34.25% (c = 3).

	LMM	**PHM**			
		***c* = 0.5**	***c* = 1**	***c* = 2**	***c* = 3**
n = 170	77.5 (5.9)	17.8 (4.7)	26 (5.2)	42.5 (4.5)	44.8 (4.5)
n = 650	100 (7.3)	57.4 (5.3)	76.9 (4.3)	93.7 (4.6)	94.5 (4.6)

Table 3

Power (α) in percent based on 1000 simulated trials for the given total sample size, analysis method, and event threshold as described in Section 3.3. The event rates associated with the thresholds ranged from a mean of 87.5% (c = 1/2) to a mean of 8.9% (c = 3).

	***c* = 0.5**	***c* = 1**	***c* = 2**	***c* = 3**
n = 100				
PHM	88.4 (5.5)	97.3 (6.3)	95 (5.4)	64.9 (2.3)
LMM	7 (5.4)	38.1 (5.4)	98.6 (6.2)	99.7 (6)
LMM2	53.5 (3.3)	80.7 (2.7)	99 (4.7)	99.6 (5.4)
n = 200				
PHM	99.7 (5.3)	99.9 (5.8)	100 (5)	93.2 (4.3)
LMM	9.5 (5.2)	68.3 (6.4)	100 (5.4)	100 (5)
LMM2	88.8 (4.7)	98 (4.1)	100 (3.6)	100 (4.9)