Published in final edited form as:
Stat Med. 2013 January 30; 32(2): 220–229. doi:10.1002/sim.5480.

Power of the Mantel-Haenszel and Other Tests for Discrete or Grouped Time-to-event Data Under a Chained Binomial Model

John M. Lachin
The Biostatistics Center Departments of Epidemiology and Biostatistics, and Statistics The George Washington University 6110 Executive Boulevard, Suite 750 Rockville, Maryland USA 20852

Summary

Power for time-to-event analyses is usually assessed under continuous time models. Often, however, times are discrete or grouped, as when the event is only observed when a procedure is performed. Wallenstein and Wittes (*Biometrics*, 1993) describe the power of the Mantel-Haenszel test for discrete life-tables under their chained binomial model for specified vectors of event probabilities over intervals of time. Herein the expressions for these probabilities are derived under a piecewise exponential model allowing for staggered entry and losses to follow-up.

Radhakrishna (*Biometrics*, 1965) showed that the Mantel-Haenszel test is maximally efficient under the alternative of a constant odds ratio and derived the optimal weighted test under other alternatives. Lachin (*Biostatistical Methods: The Assessment of Relative Risks*, 2011) describes the power function of this family of weighted Mantel-Haenszel tests. Prentice and Gloeckler (*Biometrics*, 1978) describe a generalization of the proportional hazards model for grouped time data, and the corresponding maximally efficient score test. Their test is also shown to be a weighted Mantel-Haenszel test and its power function is likewise obtained.

There is trivial loss in power under the discrete chained binomial model relative to the continuous-time case provided that there is a modest number of periodic evaluations. Relative to the case of homogeneity of odds ratios, there can be substantial loss in power when there is substantial heterogeneity of odds ratios, especially when heterogeneity occurs early in a study when most subjects are at risk, but little loss in power when there is heterogeneity late in a study.

Keywords

Sample size; power; Mantel-Haenszel test; chained binomial model; Cox Proportional Hazards Model; Prentice-Gloeckler discrete Proportional Hazards Model

1. INTRODUCTION

Wittes and Wallenstein [1] describe the power function of the Mantel-Haenszel test [2] for stratified 2 × 2 tables; and Wallenstein and Wittes [3] describe the power function of the Mantel-Haenszel test for lifetables [4] where the events are observed in discrete or grouped time. They distinguish two cases. Under their multinomial model, all event times are observed exactly (continuously) but the event times are grouped, and during an interval subjects may be lost to follow-up, all of whom are known to be event free at the time lost. This would apply to a standard actuarial lifetable. However, under their chained binomial model it is assumed that a subject's status is only known when an evaluation is performed at the close of an interval, at which time it can be determined whether a subject experienced an event since the last evaluation. This type of data structure is common. Under this model a subject who has the event during the interval, but is not evaluated at the end of the interval, is not observed to have had the event.

An example of a chained binomial data structure is provided by the analysis of onset of nephropathy (kidney disease) observed on a renal function evaluation performed annually in the Diabetes Control and Complications Trial [5]. If a subject developed nephropathy between baseline and the 1 year evaluation, for example, it was only observable if the renal evaluation was performed at 1 year. Likewise, a subject who was lost to follow-up between baseline and 1 year was last known to be "event free" at baseline, not at the time of loss. Wallenstein and Wittes [3] describe such losses as obliterating any information about whether the subject experienced the event. In such cases, the analysis employs a Mantel-Haenszel, or similar, test exaiminig the numbers of subjects observed to have the event among those evaluated, as described by Koch, McCanless and Ward [6] and by Peto [7], among others. Lachin [8] describes the corresponding survival function estimate as a modified Kaplan-Meier estimate.

The Mantel-Haenszel test for lifetables is simply an application of the Mantel-Haenszel test for stratified 2×2 tables, and it is well known that the latter test is maximally efficient under the alternative hypothesis that there is a common odds ratio. However, some other alternative hypothesis might apply, such as a common reltaive risk (risk ratio) among strata, in which case the original unweighted Mantel-Haenszel test will not be optimal. Thus, Radhakrishna [9] derived a family of weighted Mantel-Haenszel tests for stratified 2 × 2 tables that employs a specific set of weights that are optimal, or provide a fully efficient test, under a specific alternative with a constant difference in risks on some scale.

With continuous event-time data (i.e. ungrouped and untied event times), the unweighted Mantel-Haenszel test is equivalent to the logrank test [10] and is optimal under a proportional hazards (PH) model, as demonstrated by Cox [11], among others. The Tarone-Ware [12] and G^{ρ} [13] families of weighted Mantel-Haenszel (or weighted logrank) tests then provide tests that are optimal over a range of alternatives that includes proportional hazards or proportional odds, among others. However, with grouped or discrete time data, these tests are no longer optimal for the corresponding continuous time alternative, so that the unweighted Mantel-Haenszel test is not fully efficient under the popular proportional hazards model. Prentice and Gloeckler [14], however, present a score test for discrete time data that yields a fully efficient weighted Mantel-Haenszel test under this model.

Wallenstein and Wittes [3] describe the power function of the unweighted Mantel-Haenszel test in general terms for a set of specified probabilities of the event and loss to follow-up in each interval over time, and specified numbers of subjects evaluated in each interval. They also presented examples demonstrating a loss in power due to the heterogeneity of odds ratios that can result from the interval observations. Herein the expressions for these probabilities in the chained binomial model are derived under a piecewise exponential model for the event of interest and for loss to follow-up. The power of the Mantel-logrank test under this model is then derived. We then generalize the expression of Wallenstein and Wittes to the family of weighted Mantel-Haenszel tests for lifetables that includes the Prentice-Gloeckler test.

Under a PH model with time-varying hazard rates it is well-known that there will be heterogeneity of survival odds and likewise heterogeneity of the odds ratio of events within intervals over time, thus reducing the power of the unweighted Mantel-Haenszel test. Power can also be reduced, relative to continuous-time observations, owing to the reduced numbers of subjects observed to experience the event when evaluated periodically. Examples explore the effects of each of these factors on the power of the unweighted Mantel-Haenszel test.

2. Chained Binomial Piecewise Exponential Model

Assume the staggered entry of subjects over an interval of $(0, R]$ units of calendar time and that the total study duration is $T > R$ units. For now consider a single group of N subjects and let $j = 1, ..., N$ designate the individual subjects in the order entered over the interval $(0, R]$. Assume that the entry or accrual times within the cohort, say $\{a_j\}$, are distributed with density $g(a)$ and distribution function $G(a)$, $a \in (0, R]$, where $G(0) = 0$ and $G(R) = 1$. Subjects entered by time a will have a possible administrative censoring or maximal follow-up time of $e = (T - a) \in [T - R, T)$, where $e = T - R$ for the last subject entered ($a = R$), and $e = (T - \varepsilon)$ for the first subject entered (for some ε close to zero). The pattern of administrative censoring, measured in study time since randomization, is then a reflection of the pattern of recruitment in calendar time. Under the simplest case of uniform entry, $g(a) = 1/R$, and $G(a) = a/R$.

Let $\tau_1, ..., \tau_I$ designate a sequence of fixed follow-up "visit" times after entry (i.e. study times) at which subjects are evaluated to determine whether the event has occured in the preceding interval $(\tau_{i-1} - \tau_i]$, where $\tau_0 = 0$ and it is assumed that the last time $\tau_I \leq T$. Assume that all subjects are evaluated at the exact specified visit times following entry. Let n_i denote the number entering the ith interval at time τ_{i-1}. Of these, let ℓ_i denote the number of subjects who were lost to follow-up (i.e. right censored) during the ith interval, all of whom were evaluated at the end of the prior interval and known to be event free at that time. These subjects are not evaluated at the end of the interval and thus $r_i = n_i - \ell_i$ is the number at risk (evaluated) at the end of the interval at time τ_i. Then let d_i designate the number among the r_i who are observed to have the outcome event when evaluated at time τ_i, where $E(d_i) = r_i\pi_i$, π_i being the probability of the event occuring during the ith interval. Note that d_i is not the actual number of subjects who may have the event during the interval, since some may not be observed at the end of the interval.

For a given subject, let t denote the event time and u the loss-to-followup time, the two being statistically independent. For the event times, assume a piecewise exponential model with an underlying continuous time hazard of the event λ_i that is assumed constant over the ith interval. A special case is the exponential assumption of a constant hazard λ over all intervals. Then the probability that an event occurs during the ith interval among those who remain at risk at the end of the prior interval is obtained as

$$\pi_i = \frac{\int_{\tau_{i-1}}^{\tau_i} \lambda_i e^{-\lambda_i t} dt}{S(\tau_{i-1})} = \frac{e^{-\lambda_i \tau_{i-1}} - e^{-\lambda_i \tau_i}}{S(\tau_{i-1})} \quad (1)$$

where the survival function at time τ_i-1 is provided by

$$S(\tau_{i-1}) = \exp\left[-\sum_{m=1}^{i-1} \lambda_m(\tau_m - \tau_{m-1})\right] = 1 - \sum_{m=1}^{i-1} [e^{-\lambda_m \tau_{m-1}} - e^{-\lambda_m \tau_m}]. \quad (2)$$

For the jth subject, the administrative censoring time is $T - a_j$. Under an entry distribution $g(.)$, the probability of administrative censoring in the ith interval is provided by

$$\nu_i = I_{(\tau_{i-1} \geq T-R)} \left[\frac{G(T - \tau_{i-1}) - G(T - \tau_i)}{G(T - \tau_{i-1})}\right]. \quad (3)$$

In the simple case of uniform entry over the interval $(0, R]$, then $g(a) = 1/R$ and $G(a) = a/R$ in which case

$$\nu_i=I_{(\tau_{i-1}\geq T-R)}\left[\frac{\tau_i-\tau_{i-1}}{T-\tau_{i-1}}\right]. \quad (4)$$

Lachin and Foulkes [15] also consider the truncated exponential entry distribution that allows either a concave recruitment pattern where enrollment lags initially but then catches up over time, or a convex pattern, where

$$g(a)=ce^{-ca}/(1-e^{-cR}), 0<a\leq R, c\neq 0, \ G(a)=\frac{1-e^{-ca}}{1-e^{-cR}}. \quad (5)$$

Also assume that the loss to follow-up time u has cumulative distribution function $H(u)$. Then the probability of being lost in the ith interval is provided by

$$\xi_i=\frac{H(\tau_i)-H(\tau_{i-1})}{1-H(\tau_{i-1})}. \quad (6)$$

Under a piecewise exponential model with hazard rate η_i over the ith interval, this yields

$$\xi_i=\frac{e^{-\eta_i\tau_{i-1}}-e^{-\eta_i\tau_i}}{1-H(\tau_{i-1})}. \quad (7)$$

where

$$H(\tau_{i-1})=1-\exp\left[-\sum_{m=1}^{i-1}\eta_m(\tau_m-\tau_{m-1})\right]=\sum_{m=1}^{i-1}[e^{-\eta_m\tau_{m-1}}-e^{-\eta_m\tau_m}]. \quad (8)$$

The probability of exiting the study during an interval due to loss-to-follow-up or administrative censoring, i.e. not being followed and evaluated at the end of the interval, is

$$\gamma_i=\nu_i+\xi_i-\nu_i\xi_i. \quad (9)$$

Thus, the probability that a subject enters the ith interval of follow-up is

$$\prod_{m=1}^{i-1}(1-\gamma_m)(1-\pi_m), \quad (10)$$

the probability that a subject is evaluated at the end of the interval is $1-\gamma_i$, and the probability that an event is observed at the end of the interval is $\pi_i(1-\gamma_i)$. Thus, $E(n_1)=N$ and

$$E(n_i)=N\prod_{m=1}^{i-1}(1-\gamma_m)(1-\pi_m), i>1=E(n_{i-1})(1-\gamma_{i-1})(1-\pi_{i-1}) \ E(r_i)=E(n_i)(1-\gamma_i) \ E(\ell_i)=E(n_i)-E(r_i)=E(n_i)\gamma_i \ E(d_i)=E(r_i)\pi_i. \quad (11)$$

3. Weighted Tests

Now consider a study with two groups ($k=1, 2$) and let n_{ik}, ℓ_{ik}, and r_{ik} refer to the numbers in the two groups to enter, exit from, and be evaluated at the end of the ith interval, of whom d_{ik} are observed to have had the event. Let n_i, ℓ_i, r_i and d_i denote the corresponding quantities in the combined cohort.

Using the conditional hypergeometric variance, the Mantel-Haenszel (1959) test can be expressed as

$$Z_{MH}=\left[\sum_{i=1}^{I}\left(d_{i1}-\frac{d_i r_{i1}}{r_i}\right)\right]\left[\sum_{i=1}^{I}\left(\frac{r_{i1}r_{i2}d_i(r_i-d_i)}{r_i^2(r_i-1)}\right)\right]^{-1/2} \quad (12)$$

that is distributed as standard normal under the compound null hypothesis H_0: $\pi_{i1} = \pi_{i2}$ for $i = 1, ..., I$.

The sample proportions are simply $p_{ik} = d_{ik}/r_{ik}$ where $E(p_{ik}) = \pi_{ik}$ in the kth group, and $p_i = d_i/r_i$ where $E(p_i) = \pi_{i(0)}$ in the combined cohort under the null hypothesis. Cochran [16] presents an alternate test in terms of the differences in proportions within the two groups ($p_{i1} - p_{i2}$) using the unconditional product-binomial variance. His test is equivalent to (12) except that it employs r_i^3 in lieu of $r_i^2(r_i - 1)$ in the denominator of the variance. This test is commonly referred to as the Cochran-Mantel-Haenszel (CMH) test. Radhakrishna [9] generalized this test to allow for different weights $\{w_i\}$ over the set of 2 × 2 tables. The weighted Mantel-Haenszel test using the unconditional variance is

$$Z_{WMH}=\frac{\sum_i w_i(p_{i1}-p_{i2})}{\left[\sum_i w_i^2\widehat{\sigma}_{i(0)}^2\right]^{1/2}} \quad (13)$$

where the estimate of the unconditional variance under the null hypothesis is

$$\widehat{\sigma}_{i(0)}^2=\widehat{V}[(p_{i1}-p_{i2})|H_0]=\left[p_i(1-p_i)\left(\frac{1}{r_{i1}}+\frac{1}{r_{i2}}\right)\right]. \quad (14)$$

For some function $f(\pi)$ with associated parameter $\theta_i = [f(\pi_{i1}) - f(\pi_{i2})]$, using a Taylor's expansion, it follows that $\theta_i \cong g'(\pi_{i(0)})[\pi_{i1} - \pi_{i2}]$. Then assuming that this quantity is constant over all tables, $\theta_i = \theta\ \forall_i$, the optimal weights $\{\omega_i\}$ that provide maximum efficiency are obtained [9]; cf. [8]. The test employs weights $\{w_i\}$ based on the estimated probabilities. The optimal test under the assumption of a common odds ratio is based on $f(\pi) = logit(\pi)$, θ_i being the log odds ratio, in which case $w_i = r_{i1}r_{i2}/r_i$ and $Z_{WMH} \cong Z_{MH}$.

Prentice and Gloeckler [14] present a generalization of the Cox Proportional Hazards model to the case of interval or grouped survival time. For two groups with constant hazard ratio θ over time, the hazard functions are $\lambda_1(t) = \theta\lambda_2(t)$, and cumulative hazard functions $\Lambda_1(t) = \theta\Lambda_2(t)$, with survival functions $S_1(t) = \exp(-\theta\Lambda_2(t)) = S_2(t)^\theta$. Thus, the model coefficient $\beta = \ln(\theta)$ or the log hazard ratio. Under a piecewise exponential model for each interval, from (1), the probabilities of the event in the two groups (π_{i1}, π_{i2}) are related as

$$\pi_{i1}=\frac{\int_{\tau_{i-1}}^{\tau_i}\theta\lambda_{i2}e^{-\theta\lambda_{i2}u}du}{S_2(\tau_{i-1})^\theta}=\frac{S_2(\tau_{i-1})^\theta-S_2(\tau_i)^\theta}{S_2(\tau_{i-1})^\theta}\neq\theta\pi_{i2}. \quad (15)$$

Thus, under the discrete PH model the event probabilities are not proportional and do not provide constant odds ratios over time, even under a simple exponential model.

The Prentice-Gloeckler model assumes a "background" probability of the event within each successive interval that is modified as a function of a linear function of covariates. With a single binary covariate representing treatment group, the model parameters consist of I background probabilities and the covariate coefficient (β). The efficient score test for H_0: $\beta = 0$ then is a C_α test that is also a weighted Mantel-Haenszel test as in (13) using weights

$$w_i=\left[\frac{r_{i1}r_{i2}}{d_i}\right]\ln\left[\frac{r_i}{r_i-d_i}\right]=\left[\frac{r_{i1}r_{i2}}{r_ip_i}\right]\ln\left[\frac{1}{1-p_i}\right]. \quad (16)$$

Note, however, that as $p_i \rightarrow 0$, then $\ln [1 - p_i]^{-1} \rightarrow p_i$ and the weights of this test are indistinguishable from those of the weighted Mantel-Haenszel test for a common odds ratio.

4. Power and Sample Size

Let ρ_k denote the group sample fractions on entry with initial sample sizes $N_k = \rho_k N$. Within each group, the accrual distribution $G_k(a)$, with the corresponding distribution of administrative censoring over time, and the distribution of losses to follow-up with the hazard rates of loss-to-follow-up $\{\eta_{ik}\}$ over time, yield the probabilities of a subject exiting each interval $\{\gamma_{ik}\}$ in the two groups over all I intervals as in Section 2. In a randomized study, the accrual distributions will be the same for the two groups but not necessarily the loss hazard rates and the associated exit probabilities. Under the alternative hypothesis, the groups will differ with respect to the hazard rates of the event $\{\lambda_{ik}\}$ that generate the probabilities of the event $\{\pi_{ik}\}$ over all I intervals. Then, the probability that a subject in the kth group is at risk at τ_i is:

$$\alpha_{ik}=\rho_k(1-\gamma_{ik})\prod_{m=1}^{i-1}(1-\gamma_{mk})(1-\pi_{mk}). \quad (17)$$

The expected number at risk at the end of the ith interval in the kth group, and the expected number of events, are provided by

$$E(r_{ik})=N\alpha_{ik} \;\; E(d_{ik})=\pi_{ik}E(r_{ik})=N\alpha_{ik}\pi_{ik}. \quad (18)$$

The corresponding totals in each interval are

$$E(r_i)=E(r_{i1}+r_{i2})=N\alpha_i;\alpha_i=\alpha_{i1}+\alpha_{i2} \;\; E(d_i)=E(d_{i1}+d_{i2})=N(\alpha_{i1}\pi_{i1}+\alpha_{i2}\pi_{i2}). \quad (19)$$

Under the null hypothesis, $\lambda_{i1} = \lambda_{i2} = \lambda_{i1(0)}$, or equivalently, $\pi_{i1} = \pi_{i2} = \pi_{i(0)}$, $i = 1, ..., I$. From the specification under the alternative, the assumed common hazard in each interval is obtained as $\lambda_{i(0)} = \alpha_{i1}\lambda_{i1} + \alpha_{i2}\lambda_{i2}$. For a computation with constant hazards and hazard ratios over time then this common hazard can be obtained as $\lambda_{i(0)} = \rho_1\lambda_i1 + \rho_2\lambda_{i2}$. Then the assumed common event probability $\pi_{i(0)}$ is obtained from (1) using the $\lambda_{i1(0)}$. These probabilities are employed in (17) to obtain the expected fractions at risk in each group, $\{\alpha_{i1(0)}\}$ and $\{\alpha_{i2(0)}\}$, and overall $(\{\alpha_{i(0)}\})$. Substituting into (18) for each group yields the expected frequencies within each group, $E(r_{ik(0)})$ and $E(d_{ik(0)})$, and in total, $E(r_{i(0)})$, and $E(d_{i(0)})$. This approach differs slightly from that used by Wallenstein and Wittes [3] who obtained the null event probability, π_i^0 in their notation, as the weighted average of the pre-specified probabilities π_{i1} and π_{i2} (equation 8 therein).

The power of a Mantel-logrank test under a chained binomial model can then be evaluated using the resulting non-centrality parameter of the test that is simply the expectation of the test statistic under this model. This is provided by

$$\psi=\left[\sum_{i=1}^{I}\left(E(d_{i1})-\frac{E(d_i)E(r_{i1})}{E(r_i)}\right)\right]\left[\sum_{i=1}^{I}\left(\frac{E(r_{i1})E(r_{i2})E(d_i)E(r_i-d_i)}{E(r_i)^2E(r_i-1)}\right)\right]^{-1/2}. \quad (20)$$

The power and sample size for a one-sided test at level α with a given non-centrality parameter then satisfy the relationship

$$|\psi|=Z_{1-\alpha}+Z_{1-\beta} \quad (21)$$

where one would use $Z_{1-\alpha/2}$ for a two-sided test. Then the power of the test for given N and values of the other parameters is provided by $\Phi(Z_{1-\beta})$ where

$$Z_{1-\beta}=|\psi|-Z_{1-\alpha}. \quad (22)$$

Owing to the denominator for the variance, a closed form solution for the required N is not possible, although N can be readily obtained from a recursive procedure such as the secant method [17]. Alternately, $E(r_i)^3$ may be substituted for $E(r_i)^2E(r_i-1)$ to yield an expression of the form $\psi=\sqrt{N}\varphi$ where φ is then a function of the corresponding parameters $\{\alpha_{i1}, \alpha_{i2}, \alpha_i, \pi_{i1}, \pi_{i2}, \pi_i\}$ from (18) and (19). Then the required N is provided by $N=(Z_{1-\alpha}+Z_{1-\beta})^2/\varphi^2$.

A generalization of the Wallenstein-Wittes equation may also be employed. Lachin [8] presents equations to determine the power (or sample size) for a weighted Mantel-Haenszel test. The test in (13) is a function of the difference in proportions $p_{i1}-p_{i2}$ for the ith interval and weights that are also a function of the sample sizes and possibly the proportions. From Slutsky's theorem [8], the distribution of this test converges to that of

$$\frac{\sum_i \omega_{i(0)}(p_{i1}-p_{i2})}{\left[\sum_i \omega_{i(0)}^2 \sigma_{i(0)}^2\right]^{1/2}}=\frac{T}{\sqrt{V(T|H_0)}} \quad (23)$$

where $E(w_i)=N\omega_i$, and $\omega_{i(0)}$ is the corresponding quantity evaluated under H_0.

Let

$$\sigma_i^2=V[(p_{i1}-p_{i2})|H_1]=\frac{1}{N}\left[\frac{\pi_{i1}(1-\pi_{i1})}{\alpha_{i1}}+\frac{\pi_{i2}(1-\pi_{i2})}{\alpha_{i2}}\right]=\frac{\varphi_i^2}{N} \; \sigma_{i(0)}^2=V[(p_{i1}-p_{i2})|H_0]=\frac{1}{N}\left[\pi_{i(0)}(1-\pi_{i(0)})\left(\frac{1}{\alpha_{i1(0)}}+\frac{1}{\alpha_{i2(0)}}\right)\right]=\frac{\varphi_{i(0)}^2}{N}. \quad (24)$$

Thus, under H_1,

$$\mu_1=E(T|H_1)=\sum_i \omega_i(\pi_{i1}-\pi_{i2}) \quad (25)$$

The relationship between sample size and power is expressed as

$$\sqrt{N}|\mu_1|=Z_{1-\alpha}\left[\sum_i \omega_{i(0)}^2 \varphi_{i(0)}^2\right]^{1/2}+Z_{1-\alpha}\left[\sum_i \omega_i^2 \varphi_i^2\right]^{1/2}. \quad (26)$$

Since $\omega_i^2\varphi_i^2 \cong \omega_{i(0)}^2\varphi_{i(0)}^2$, then approximately

$$\sqrt{N}|\mu_1|=[Z_{1-\alpha}+Z_{1-\beta}]\left[\sum_i \omega_{i(0)}^2 \varphi_{i(0)}^2\right]^{1/2}. \quad (27)$$

The Mantel-Haenszel test is known to be maximally efficient against the alternative that there is a constant odds ratio over all 2 × 2 tables. The test uses weights with expectation

$$E(w_i)=E\left[\frac{r_{i1}r_{i2}}{r_i}\right]=N\frac{\alpha_{i1}\alpha_{i2}}{\alpha_i} \quad (28)$$

so that

$$\omega_i=\frac{\alpha_{i1}\alpha_{i2}}{\alpha_i},\ \omega_{i(0)}=\frac{\alpha_{i1(0)}\alpha_{i2(0)}}{\alpha_{i(0)}}. \quad (29)$$

The power of this test is then obtained by solving for $Z_{1-\beta}$ in (27) to yield

$$Z_{1-\beta}=\sqrt{N}\frac{\sum_i\left[\frac{\alpha_{i1(0)}\alpha_{i2(0)}}{\alpha_{i(0)}}\right](\pi_{i1}-\pi_{i2})}{\left[\sum_i\left[\frac{\alpha_{i1(0)}\alpha_{i2(0)}}{\alpha_{i(0)}}\right]^2\varphi_{i(0)}^2\right]^{1/2}}-Z_{1-\alpha}. \quad (30)$$

This is equivalent to the Wallenstein-Wittes equation except for the manner in which the null probability $\pi_{i(0)}$ is computed.

The power function of the Prentice-Gloeckler test can likewise be obtained from these expressions. The test uses weights in (16) so that

$$\omega_i=\left[\frac{\alpha_{i1}\alpha_{i2}}{\alpha_i\pi_i}\right]\ln\left[\frac{1}{1-\pi_{i(0)}}\right],\ \omega_{i(0)}=\left[\frac{\alpha_{i1(0)}\alpha_{i2(0)}}{\alpha_{i(0)}\pi_{i(0)}}\right]\ln\left[\frac{1}{1-\pi_{i(0)}}\right]. \quad (31)$$

The expectation under the PH alternative is

$$E(T|H_1)=\mu_1=\sum_i\omega_i(\pi_{i1}-\pi_{i2})=\sum_i\left(\ln\left[\frac{1}{1-\pi_i}\right]\right)\left[\frac{\alpha_{i1}\alpha_{i2}}{\alpha_i\pi_i}\right](\pi_{i1}-\pi_{i2}) \quad (32)$$

and the null variance is

$$V(T|H_0)=\sum_i\omega_{i(0)}^2\varphi_{i(0)}^2/N=\sum_i\left(\ln\left[\frac{1}{1-\pi_{i(0)}}\right]\right)^2\left[\frac{\alpha_{i(0)1}\alpha_{i(0)2}}{\alpha_{i(0)}\pi_{i(0)}}\right]^2\frac{\varphi_{i(0)}^2}{N}. \quad (33)$$

Then the basic equation relating sample size and power is

$$|\mu_1|=Z_{1-\alpha}\sqrt{V(T|H_0)}+Z_{1-\beta}\sqrt{V(T|H_1)}\cong(Z_{1-\alpha}+Z_{1-\beta})\sqrt{V(T|H_0)} \quad (34)$$

which yields power as a function of

$$Z_{1-\beta}=\sqrt{N}\frac{\sum_i\ln\left[\frac{1}{1-\pi_{i(0)}}\right]\left[\frac{\alpha_{i(0)1}\alpha_{i(0)2}}{\alpha_{i(0)}\pi_{i(0)}}\right](\pi_{i1}-\pi_{i2})}{\left[\sum_i\left(\ln\left[\frac{1}{1-\pi_{i(0)}}\right]\right)^2\left[\frac{\alpha_{i(0)1}\alpha_{i(0)2}}{\alpha_{i(0)}\pi_{i(0)}}\right]^2\varphi_{i(0)}^2\right]^{1/2}}-Z_{1-\alpha}. \quad (35)$$

As $\pi_{i(0)}\rightarrow 0$, such as when the number of intervals increases, then ln $[1-\pi_{i(0)}]^{-1}\rightarrow\pi_{i(0)}$ and the above equation simplifies to the expression in (30) that provides the power of the weighted Mantel-Haenszel test for a common odds ratio. Also, if $\pi_{i(0)}=\pi_{(0)}$ is constant over all intervals, such as under an exponential model, then the power of the Prentice-Gloeckler test is identical to that of Mantel-Haenszel test.

6. Examples

Consider a study with uniform enrollment of $N = 100$ patients over $R = 3$ years and a total duration of $T = 5$ years with constant hazards $\lambda = 0.30$ and $\eta = 0.05$ for all visits and with 2 visits per year. Using the expressions in Lachin and Foulkes [15], 59.4 subjects with the event would be expected with continuous time observations (e.g. the day of the event).

However, for semi-annual outcome assessments, Table 1 provides the expected numbers of subjects to enter each interval (n_j), to exit during the interval (ℓ_j), to be evaluated at the end of the interval (r_j) and to have the event present at the end of the interval (d_j). Summing the last column, the total expected number of subjects observed with the event is 56.3. For any other sample size, the values above provide the expected percentages.

Note that no subjects are evaluated beyond 4.5 years of follow-up because all subjects entered during the first 6 months of recruitment are followed for [4.5, 5) years, none evaluated at 5 years of follow-up. Thus all of these are administratively censored during the last interval. Likewise, the 1/6 of subjects recruited in the second half of the third year are followed for [2, 2.5) years and last evaluated at 2 years. This results in a reduction in the numbers of observed events and the pursuant power. This can be addressed by shortening the duration of the evaluation intervals, or extending the study. For example, if the study were extended by 3 months to 5.25 years, then half the subjects whose last visit occurs within an interval would be available for evaluation at the next interval. In this case the total expected number of observed subjects with the event is 60. Alternately, keeping the study at 5 years and increasing the frequency of visits to 4 per year, with 20 intervals, the total expected number of subjects with the event is 57.8.

Also, the probability of new cases among those at risk $\{\pi i\}$ under an exponential model is constant over time. With constant hazard λ, the probability of an event in an interval of length I is simply $1 - \exp(-\lambda I)$. Thus for the above example with half-yearly intervals, $\pi_j = 1 - \exp(-0.3/2) = 0.139$. Therefore, under an exponential model, there is a constant relative risk and constant odds ratio over time within each of the 2 × 2 tables over time.

Now consider the potential loss in power compared to the continuous-time test power under an exponential model. Lachin and Foulkes [15] describe the power of the test of the difference between two hazard rates under an exponential model for observations observed continuously over time. In their Table 3.B, using this test, they show that $N = 406$, equally divided between two groups, provides 90% power to detect a relative hazard of 2/3 with a control hazard $\lambda_2 = 0.3$ per year and losses to follow-up at hazard rate $\eta = 0.05$ per year in each group in a study with $R = 3$ years of uniform recruitment and total duration $T = 5$ years using a test at the $\alpha = 0.05$ level (one-sided). Under the alternative, the expected numbers of events in the control and treated groups are 120.5 and 93.1, respectively, and the expected numbers lost to follow-up (not administratively censored) are 20 and 23, respectively. Asymptotically, the continuous time Mantel-Haenszel (logrank) test should have equivalent power under this model.

Table 2 presents the expected numbers of events within each group and the resulting levels of power of the Mantel-Haenszel test and the Prentice-Gloeckler test under the chained binomial model with varying numbers of visits (intervals) per year. With only one evaluation per year, even though the number of subjects observed to have the event is reduced by 11%, there is a small reduction in power. As expected, as the number of intervals (assessment visits) increases, the numbers of events observed increases and likewise power. The power obtained from the non-centrality parameter (22) is trivially less than that from the weighted test expression (30). With frequent intervals, the power estimates exceed the 0.90 estimated from the continuous-time expressions of Lachin and Foulkes for the exponential model-based test.

Additional computations were performed for the above Lachin-Foulkes design with $N = 406$ under a piecewise exponential model with 2 intervals per year, but where the control group hazards $\{\lambda_{j2}\}$ and the hazard ratios either increased or decreased over time. As would be expected from similar computations performed for continuous time observations using the

Lakatos [18] model, power is greater when the hazard function decreases rather than increases over time, and when the hazard ratio is further from 1 earler rather than later in the study. The power was virtually identical using either the expressions for the Mantel-Haenszel test or the Prentice-Gloeckler test.

7. Discussion

Other than the work of Wallenstein and Wittes [3] there has been little assessment of the impact of grouping of event times on the resulting power of a weighted Mantel-Haenszel test. Virtually all existing methods for assessment of sample size or power are based on the observation of event times continuously, some under a simple exponential model such as Lachin and Foulkes [15], or a piecewise exponential model such as Lakatos [18]. While the Lakatos procedure, as herein, is based on specified hazard rates over discrete intervals of time, it actually assesses the power of the continuous time logrank test, not that of a discrete time test, under the piecwise exponential model assumptions.

The Mantel-Haenszel test for a stratified analysis of 2 × 2 tables is known to be fully efficient under the alternative of a constant odds ratio over all 2×2 tables and thus Wallenstein and Wittes [3] describe the power of the test for grouped-time survival analysis under this alternative. Their expression allows for variation in the odds ratios over intervals of time, and they showed that heterogeneity of odds ratios over time leads to a loss of power relative to the case of a constant odds ratio over time, as is well known for the original Mantel-Haenszel test for stratified 2 × 2 tables. However, they did not explore the properties of the test under the popular proportional hazards model, or the simplest special case of an exponential model. Under an exponential model with evenly spaced evaluations, the conditional probability of the event is constant within each interval, and thus also the odds ratio, so that there is little loss in power relative to the continuous time case, and that loss is attributable to the reduction in observed events owing to the periodic outcome assessments (Table 2). Under the PH model with time-varying hazard rates, it is readily shown that a constant odds ratio can not apply, and thus there is some loss in power relative to the continuous time case. However, depending on the pattern of recruitment, follow-up and the extent of losses to follow-up, the impact of this heterogeneity can be slight, especially when the event probabilities are small in which case the relative risk (risk ratio) is approximately equal to the odds ratio.

Prentice and Gloeckler [14] generalized the PH model to grouped or interval time data and under a PH alternative with grouped time data their score test will always be more powerful than the Mantel-Haenszel test considered by Wallenstein and Wittes, provided that the hazard rates vary over time. If not, and the exponential model applies, then the Prentice and Gloeckler test and the Mantel-Haenszel test are both fully efficient. Further, there are other patterns over time that depart from the proportional hazards model, and also from a constant odds ratio model, for which the Mantel-Haenszel and Prentice-Gloeckler tests will yield similar results, and will thus have similar power.

Acknowledgments

This work was partially supported by funding from the National Institute of Diabetes, Digestive and Kidney Diseases.

References

1. Wittes J, Wallenstein S. The power of the Mantel-Haenszel test. J. Amer. Statist. Assoc. 1987; 82:1104–1109.

2. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 1959; 22:719–748. [PubMed: 13655060]

3. Wallenstein S, Wittes J. The power of the Mantel-Haenszel test for grouped failure time data. Biometrics. 1993; 49:1077–1087. [PubMed: 8117902]

4. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother. Rep. 1966; 50:163–170. [PubMed: 5910392]

5. Diabetes Control and Complications Trial Research Group (DCCT). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N. Engl. J. Med. 1993; 329:977–986. [PubMed: 8366922]

6. Koch GG, McCanless I, Ward JF. Interpretation of statistical methodology associated with maintenance trials. Am. J. Med. 1984; 77(supplement 5B):43–50. [PubMed: 6095661]

7. Peto, J. The calculation and interpretation of survival curves. In: Buyse, ME.; Staquett, MJ.; Sylvester, RJ., editors. Cancer Clinical Trials, Methods and Practice. Oxford: Oxford University Press; 1984.

8. Lachin, JM. Biostatistical Methods: The Assessment of Relative Risks. Second Edition. New York: Wiley; 2011.

9. Radhakrishna S. Combination of results from several 2×2 contingency tables. Biometrics. 1965; 21:86–98.

10. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). J. Roy. Statist. Soc. A. 1972; 135:185–206.

11. Cox DR. Regression models and life-tables (with discussion). J. Roy. Statist. Soc. B. 1972; 34:187–220.

12. Tarone RE, Ware J. On distribution free tests for equality of survival distributions. Biometrika. 1977; 64:156–160.

13. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982; 69:553–566.

14. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. Biometrics. 1978; 34:57–67. [PubMed: 630037]

15. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification. Biometrics. 1986; 42:507–519. [PubMed: 3567285]

16. Cochran WG. Some methods for strengthening the χ^2 tests. Biometrics. 1954; 10:417–451.

17. Thisted, RA. Elements of Statistical Computing. New York: Chapman and Hall; 1988.

18. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics. 1988; 44:229–241. [PubMed: 3358991]

Table 1

Expected numbers to enter an interval (n_i), exit during an interval (ℓ_i), to be observed and at risk at the end of the interval (r_i), and the events to be observed (d_i) in study with uniform enrollment over $R = 3$ years, a total duration of $T = 5$ years with 2 visits per year and constant hazards $\lambda = 0.30$ and $\eta = 0.05$ for all visits, and an initial sample size of $N = 100$. For all intervals the probability of the event is $\pi = 0.139$.

Interval	**E(n_i)**	**E(ℓ_i)**	**E(r_i)**	**E(d_i)**
(0,0.5]	100.0	2.5	97.5	13.6
(0.5,1.0]	83.9	2.1	81.9	11.4
(1.0,1.5]	70.5	1.7	68.7	9.6
(1.5,2.0]	59.2	1.5	57.7	8.0
(2.0,2.5]	49.7	9.3	40.4	5.6
(2.5,3.0]	34.7	7.6	27.1	3.8
(3.0,3.5]	23.3	6.3	17.1	2.4
(3.5,4.0]	14.7	5.1	9.6	1.3
(4.0,4.5]	8.2	4.2	4.0	0.6
(4.5,5.0]	3.5	3.5	0.0000	0.0000

Table 2

The expected numbers of events within each group and the resulting levels of power of the Mantel-Haenszel (MH) test computed using the non-central distribution (eqn. 22), and asymptotically (30), and of the Prentice Gloeckler (PG) test (35), under the chained binomial model with varying numbers of visits (intervals) per year for the model described by Lachin and Foulkes (1986) with $N = 406$, a control hazard rate of 0.3/year, a 2/3 hazard ratio, losses with hazard 0.05/year, $R = 3$, and $T = 5$. Using the test of the difference in exponential hazard rates for continuous observations, this sample size provides 90% power with a one-sided 0.05-level test.

Intervals per year	Controls $E(d_C)$	Treated $E(d_E)$	MH Power Non-central	MH Power Asymptotic	PG Power
1	107.7	81.9	0.869	0.868	0.868
2	114.2	87.6	0.888	0.887	0.887
3	116.3	89.4	0.893	0.892	0.892
4	117.4	90.3	0.896	0.895	0.895
8	119.0	91.7	0.900	0.899	0.899
12	119.5	92.2	0.901	0.900	0.900
24	120.0	92.6	0.902	0.901	0.901
52	120.3	92.9	0.903	0.902	0.902