

The two *K-M* curves cross at approximately 70 weeks, with an apparent advantage to the *MM* group prior to this time and thereafter a disadvantage. In this situation the complementary log plots would also cross, suggesting that the relative hazards do not remain proportional. However, it should be stressed here that the curves become less reliable with time so that the crossing at 70 weeks could merely be a consequence of small numbers of events observed in that region.

### WEIGHTED MANTEL-HAENSZEL

A number of tests have been proposed which compare two survival curves for the non-*PH* situation. These tests can be generated from one basic formula and are known as weighted Mantel-Haenszel tests. For a two-group comparison of *A* and *B*, these are defined as

$$\chi_{MHw}^2 = \frac{\sum w_t (O_{At} - E_{At})^2}{\sum w_t^2 V_t} \quad (10.1)$$

Here,  $E_{At}$  and  $V_t$  are calculated as described by equations (3.1) and (3.11) respectively. The weights are the  $w_t$  and can vary as time,  $t$ , changes. These tests all give a statistic which is distributed as  $\chi^2$  with  $df = 1$  for the two group comparisons discussed here.

If we set  $w_t = 1$ , for all values of  $t$ , then this assigns equal weight to each death at whatever time,  $t$ , it occurs. In this situation, equation (10.1) becomes that of the (unweighted) Mantel-Haenszel test of equation (3.12). This is the most powerful test when the hazards are (at least approximately) proportional.

### GEHAN

If we set  $w_t \neq 1$  we are implicitly stating that differences between certain parts of the survival curves being compared are of greater interest than others. For example, if it is anticipated that a new treatment may help avoid the particular risk of early deaths when compared with the standard treatment but thereafter holds no particular advantage, then extra weight may be assigned to any early deaths. One way in which this can be achieved is by setting  $w_t = R_t$ , where  $R_t$  is the total number of patients at risk at time  $t$ , as we have described for equation (2.11). In most examples, this number declines over time,  $t$ , as the total number of deaths and censored observations increases with time. This test, therefore, places greater emphasis or weight on 'earlier' parts of the curves. This version of the test is known as the Gehan, generalised Wilcoxon or Breslow test, and is

$$\chi_{Gehan}^2 = \frac{\sum R_t (O_{At} - E_{At})^2}{\sum R_t^2 V_t} \quad (10.2)$$

To calculate  $\chi_{Gehan}^2$  it is necessary to calculate the components of both the numerator and denominator at each death time, that is  $O_{At}$ ,  $E_{At}$  and  $V_t$ , as was done in Table 3.3, and  $R_t$ .

*Example - weighted Mantel-Haenszel and Gehan tests - time to remission from incontinence*

Nolan, Debelle, Oberklaid and Coffey (1991) quote both the Mantel-Haenszel version of the Logrank test with  $p$ -value = 0.12, corresponding to  $\chi^2_{MHw} = 0.82$ ,  $df = 1$ , and the Gehan test, with  $p$ -value = 0.012, corresponding to  $\chi^2_{Gehan} = 6.30$ ,  $df = 1$ . They also comment that: "The proportionality assumption was not satisfied for the full follow-up period".

The difference in the  $p$ -values given by the two tests leads to quite different interpretations of the data and demonstrates how important it is to plot the survival curves for the two groups so that an appropriate interpretation of the significance tests can be made.

Care must be taken in applying the Gehan test as it can be very misleading when there are many censored observations or when the censoring is unequal in the two groups. These situations can arise in the early life of a prospective study, for example in a clinical trial which is still in the recruitment phase. In this case there may be many patients recently entered and therefore insufficient observation time for the critical events to occur. Unequal censoring can occur if, for whatever reason, information is systematically received on events occurring earlier in one group than in the other. For example, in a randomised trial to compare patients using a regular and predetermined appointments system against a system in which patients present only when they notice symptoms of recurrence, one could argue that those patients with fixed appointments will be examined more rigorously and early signs of recurrence thus detected and reported. In contrast, patients assigned to the other group will only report recurrence once the symptoms are clearly obvious to the patient.

#### TARONE-WARE

A similar but less extreme means of weighting the earlier part of the survival curve, than through the Gehan test, is given by using the Tarone-Ware version of the test. This is

$$\chi^2_{TW} = \frac{\sum R_i^{1/2} (O_{At} - E_{At})^2}{\sum R_i V_i} \quad (10.3)$$

where the  $w_i$  of equation (10.1) are set equal to  $R_i^{1/2}$ . This still gives more weight to the early deaths observed in the study but in a less extreme way than that of equation (10.2).

#### PETO-PRENTICE

Another definition for the weights has been suggested by both Peto and Prentice and their test is


$$\chi^2_{PP} = \frac{\sum w_{PPi} (O_{At} - E_{At})^2}{\sum w_{PPi} V_i} \quad (10.4)$$

# Survival ANALYSIS

A PRACTICAL APPROACH

SECOND EDITION

DAVID MACHIN  
YIN BUN CHEUNG  
MAHESH K.B. PARMAR

 WILEY