Supplementary material to 'Design and testing for clinical trials faced with misclassified causes-of-death'

BART VAN ROMPAYE*,[1], SHABBAR JAFFAR[2],
ELS GOETGHEBEUR[1]

[1] Department of Applied Mathematics and Computer Science,
Ghent University, Krijgslaan 281, S9, 9000 Ghent, Belgium

[2]Department of Epidemiology and Population Health,
London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.

Contents

[1]* To whom correspondence should be addressed: bart.vanrompaye@ugent.be, Tel : + 32 9 264 47 56, Fax : + 32 9 264 49 95

In these supplementary materials, notation is taken from the main text of the article.

1 A nonparametric estimator for $\xi(t)$

The weights of the contributions to (4.1) and (4.3) from the article are explicitly dependent of t, which can enter through a time-dependence both in the misclassification rates and in the proportion of the cause-specific baseline hazards, $\xi(t)$. The method itself thus makes no restriction whatsoever on $\xi(t)$, and the value of ξ may indeed differ for all observations. However, for the analysis all $\xi(t_i)$ values need to specified, and unless these are known to some (often unrealistic) level of detail, they need to be estimated. The main text proposes the use of a constant value for ξ, which is practical since it is easy to calculate, and which captures the most coarse feature of a possible time-dependence: the average value, averaged over the observation density. However, as Klein (2006) points out, for many settings the true $\xi(t)$ will vary strongly over time, for example when one cause is dominant in younger infants and the other in older infants (as is the case for ALRI in Gambia, Jaffar *and others*, 1997). To address this issue, we introduce in section 4.1 a simple parametric estimation procedure using a piecewise constant model for $\xi(t)$. While the calculations there remain simple, it already constitutes a strong improvement in flexibility and it protects against overfitting the data.

Nevertheless, to justify such simplifying assumptions on $\xi(t)$ (or to avoid oversimplifying assumptions), we now introduce a kernel-weighted version of the full log-likelihood $\log(L^*)$ (derived later, in section 2). This allows nonparametric smoothed estimation for $\xi(t)$ under H_0. The kernel-weighted log-likelihood l^{sm} is defined by introducing a Gaussian kernel in $\log(L^*)$ and is up to a constant equal to:

$$\begin{aligned} l_t^{sm}(\xi) \;=\; & \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{t-t_i}{h}\right)^2/2} \times \\ & \left[\log\left(\frac{1}{1+e^{-\xi(t_i)}}\right).dN_{Ci}(t_i) + \log\left(1-p_1(t_i)+e^{-\xi(t_i)}p_0(t_i)\right).dN_{1i}(t_i) \right. \\ & \left. + \log\left(p_1(t_i)+e^{-\xi(t_i)}(1-p_0(t_i))\right).dN_{0i}(t_i)\right] \end{aligned}$$

Here, we used counting process notation where $N_{Ci}(t)$ is equal to 0 until time t_i, at which it becomes 1 if individual i undergoes an event, otherwise it remains zero. Similarly, $N_{1i}(t)$ ($N_{0i}(t)$) is zero until individual i suffers a type 1 (0) failure, after which it becomes one.

Since we use a Gaussian kernel, each observation contributes at each timepoint t, but its contribution is downweighted by the factor in the first line as the distance between t and t_i grows. For an event time t_i $dN_{Ci}(t_i)$ and either $dN_{1i}(t_i)$ or $dN_{0i}(t_i)$ will be different from 0 (and equal to 1), with the second and third line deciding what its unweighted contribution will be.

To obtain a smoothed estimate of $\xi(t)$, the function $l_t^{sm}(\xi)$ needs to be optimized with respect to ξ at any timepoint t. This can be done by a Newton-

Raphson routine where the estimator (4.4) provides the starting value. The search for an optimal value for the bandwidth h lies outside the aims of this paper, but one can use practical expectations (e.g. to capture a yearly fluctuation one needs a bandwidth of only a few months) recognizing it is best to base the estimate at each point in time on at least 10 not severely down-weighted observations of each type. By varying the bandwidth one should be able to get an idea of what an appropriate value might be. We further propose that if a simple parametric shape is indicated, one uses this in the test statistic, rather than the smoothed values. This will make the final analysis less difficult to understand and thus hopefully more acceptable.

To illustrate the use of the estimator we performed a small simulation study, generating data on 20,000 individuals using a 4 year accrual period with maximum follow-up of 4.5 years. Events were generated using a constant hazard of 0.3. The true event types were determined from a bernoulli-experiment, where the probability of a type 1 event had the sinusoidal shape $0.5+\sin(2\pi t)/5$, implying $\exp(-\xi(t)) = (0.5 - \sin(2\pi t)/5) \,/\, (0.5 + \sin(2\pi t)/5)$. The observed event type was then obtained by misclassifying at a rate $p_1 = 0.6$ and $p_0 = 0.1$. Since the estimation procedure is derived under the null no treatment effects were included. Figure 1 shows in full line the true $\exp(-\xi(t))$ and in dotted line the smoothed estimator of $\exp(-\xi(t))$ using a bandwidth of $h = 0.5$ years. Overall, the estimator follows the true value closely and is at least able to identify the main trends. Larger bandwidths smear out effects causing the wave pattern to disappear, while smaller bandwidths cause the estimate to be unstable. Especially on the right hand side a large boundary effect is seen. Optimization of kernel shapes and parameters is a technical issue which falls beyond the scope of this discussion.

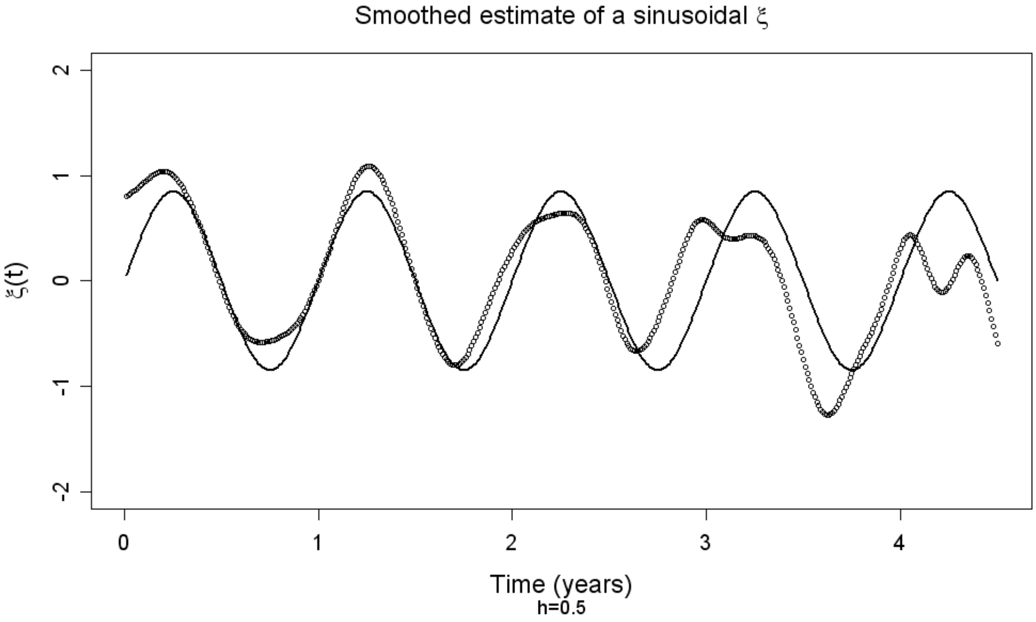

Figure 1: *Illustration of the fit from simulated data with a sinusoidal pattern.*

2 Derivation of equation (4.4) from the article

Later stages of the article use the simplifying assumption 3', that the relative cause-specific baseline hazard $\exp(-\xi)$ is constant over time. To obtain a consistent estimator for ξ we cannot use the partial likelihood L_p defined in the main text, since it conditions on the observed type of event, hence losing all information concerning the contrast between the two cause-specific hazards, and that is exactly governed by ξ. We therefore use an 'unconditional' or 'full' likelihood L^*, based on the observation of three types of events all occurring with their own probability. We define these probabilities under the null, assuming any additional censoring is non-informative. We introduce the notation $h(t) = h_1(t) + h_0(t)$.

The first type of event is the observation of a type 1 failure. The likelihood of such an event occurring at a time t_i is the sum of two densities. First, we look at the probability of a type 1 event being observed as type 1. This is the probability of event-free survival up to t_i times the hazard of an event happening at t_i times the probability of the event being of type 1 times the conditional probability no misclassification occurs. Second, there is the likelihood of a type 0 event being observed as type 1. This is the probability of event-free survival up to t_i times the hazard of an event happening at t_i times the probability of the event being of type 0 times the conditional probability misclassification occurs. More formally, we have:

$$f(T_i = t_i, C_i = 1, F_i = 1) = e^{-\int_0^{t_i} h(s)ds} h(t_i) \left(\frac{1}{1+e^{-\xi}}(1 - p_1(t_i)) + \frac{e^{-\xi}}{1+e^{-\xi}} p_0(t_i) \right)$$

For ease of interpretation, the order of the terms matches the explanation above.

The second type of event is the observation of a type 0 failure. The likelihood of this type of event is assembled in a manner similar to the probability above:

$$f(T_i = t_i, C_i = 1, F_i = 0) = e^{-\int_0^{t_i} h(s)ds} h(t_i) \left(\frac{e^{-\xi}}{1+e^{-\xi}}(1 - p_0(t_i)) + \frac{1}{1+e^{-\xi}} p_1(t_i) \right)$$

The third type of event is the occurence of an administrative censoring at t_i:

$$P(D_i > t_i, F_i) = e^{-\int_0^{t_i} h(s)ds}$$

The full likelihood L^* is then:

$$\begin{aligned} L^* &= \prod_{i:C_i=1,F_i=1} e^{-\int_0^{t_i} h(s)ds} h(t_i) \left(\frac{1}{1+e^{-\xi}}(1 - p_1(t_i)) + \frac{e^{-\xi}}{1+e^{-\xi}} p_0(t_i) \right) \\ &\quad \cdot \prod_{i:C_i=1,F_i=0} e^{-\int_0^{t_i} h(s)ds} h(t_i) \left(\frac{e^{-\xi}}{1+e^{-\xi}}(1 - p_0(t_i)) + \frac{1}{1+e^{-\xi}} p_1(t_i) \right) \cdot \prod_{i:C_i=0} e^{-\int_0^{t_i} h(s)ds} \end{aligned}$$

L^* factorizes into parts containing the parameter of interest (ξ) and parts not containing ξ:

$$\begin{aligned} L^* &= \prod_{i:C_i=1} e^{-\int_0^{t_i} h(s)ds} \prod_{i:C_i=0} e^{-\int_0^{t_i} h(s)ds} \prod_{i:C_i=1} h(t_i) \\ &\cdot \prod_{i:C_i=1} \frac{1}{1+e^{-\xi}} \prod_{i:C_i=1,F_i=1} \big(1-p_1(t_i)+e^{-\xi}p_0(t_i)\big) \\ &\cdot \prod_{i:C_i=1,F_i=0} \big(e^{-\xi}(1-p_0(t_i))+p_1(t_i)\big) \end{aligned}$$

To maximize L^* w.r.t. ξ, we only need to maximize the log of the second and third line. To do this elegantly, we first assume constant misclassification rates and rewrite the last three factors:

$$L^* \propto \left(\frac{1}{1+e^{-\xi}}\right)^{\sum_i C_i} \big(1-p_1+e^{-\xi}p_0\big)^{\sum_i C_i F_i} \big(e^{-\xi}(1-p_0)+p_1\big)^{\sum_i C_i(1-F_i)}$$

The expressions in the exponents are the numbers of observed events of each type, O_0 and O_1:

$$L^* \propto \left(\frac{1}{1+e^{-\xi}}\right)^{O_0+O_1} \big(1-p_1+e^{-\xi}p_0\big)^{O_1} \big(e^{-\xi}(1-p_0)+p_1\big)^{O_0}$$

This becomes:

$$\log(L^*) \propto -(O_0+O_1)\log(1+e^{-\xi})+O_1\log(1-p_1+e^{-\xi}p_0)+O_0\log(e^{-\xi}(1-p_0)+p_1)$$

from which:

$$\frac{\partial\log(L^*)}{\partial\xi} = (O_0+O_1)\frac{e^{-\xi}}{1+e^{-\xi}} - O_1\frac{e^{-\xi}p_0}{1-p_1+e^{-\xi}p_0} - O_0\frac{e^{-\xi}(1-p_0)}{e^{-\xi}(1-p_0)+p_1}$$

Ignoring the solution at $\xi = \infty$, we get:

$$0 = (O_0+O_1)\frac{1}{1+\widehat{e^{-\xi}}} - O_1\frac{p_0}{1-p_1+\widehat{e^{-\xi}}p_0} - O_0\frac{1-p_0}{\widehat{e^{-\xi}}(1-p_0)+p_1}$$

which becomes:

$$O_1\frac{1-p_1+\widehat{e^{-\xi}}p_0-(1+\widehat{e^{-\xi}})p_0}{1-p_1+\widehat{e^{-\xi}}p_0} = O_0\frac{(1-p_0)(1+\widehat{e^{-\xi}})-\widehat{e^{-\xi}}(1-p_0)-p_1}{\widehat{e^{-\xi}}(1-p_0)+p_1}$$

This simplifies to:

$$\frac{O_1}{O_0} = \frac{1-p_1+\widehat{e^{-\xi}}p_0}{(1-p_0)\widehat{e^{-\xi}}+p_1}$$

The estimator for $e^{-\xi}$ now becomes:

$$\widehat{e^{-\xi}} = \frac{O_1p_1-O_0(1-p_1)}{O_0p_0-O_1(1-p_0)}$$

This expression can also be derived in an intuitive manner. We know that

$$e^{-\xi} = \frac{h_0(t)}{h_1(t)}$$

Under the null this can intuitively be estimated by looking at the proportion of the true number of people who failed for each type:

$$\widehat{e^{-\xi}} = \frac{\sum_i C_i(1-\delta_i)}{\sum_i C_i\delta_i} \tag{2.1}$$

Because of the misclassification we cannot use the true δ_i, but base our estimate on the observed F_i instead. A naive approach ignoring the misclassification uses:

$$\widehat{e^{-\xi}} = \frac{\sum_i C_i(1-F_i)}{\sum_i C_iF_i}$$

We can however incorporate the misclassification in the same way as in the score statistic by using weights based on the misclassification probabilities:

$$\widehat{e^{-\xi}} = \frac{\sum_i \left\{C_i(1-F_i).P\left(\delta_i=0|F_i=0\right) + C_iF_i.P\left(\delta_i=0|F_i=1\right)\right\}}{\sum_i \left\{C_iF_i.P\left(\delta_i=1|F_i=1\right) + C_i\left(1-F_i\right).P\left(\delta_i=1|F_i=0\right)\right\}}$$

This reduces to the naive estimator (2.1) if there is no misclassification (the probabilities become 0 and 1). By introducing the explicit expressions for the conditional probabilities this becomes:

$$\widehat{e^{-\xi}} = \frac{\sum_i \left\{C_i(1-F_i).\left(1-\frac{1}{1+\frac{(1-p_0(t_i))e^{-\xi}}{p_1(t_i)}}\right) + C_iF_i.\left(1-\frac{1}{1+\frac{p_0(t_i)e^{-\xi}}{1-p_1(t_i)}}\right)\right\}}{\sum_i \left\{C_iF_i.\frac{1}{1+\frac{p_0(t_i)e^{-\xi}}{1-p_1(t_i)}} + C_i\left(1-F_i\right).\frac{1}{1+\frac{(1-p_0(t_i))e^{-\xi}}{p_1(t_i)}}\right\}}$$

Since the probability that an event of a certain observed type is truly (not) of that type depends on ξ, $e^{-\xi}$ itself enters the expression. After replacing $e^{-\xi}$ throughout by its estimator, and by assuming constant p_0 and p_1 the expression resolves into a simple equation for $\widehat{e^{-\xi}}$:

$$\widehat{e^{-\xi}} = \frac{O_1p_1 - O_0(1-p_1)}{O_0p_0 - O_1(1-p_0)}$$

which is the same as from the previous derivation.

It should be understood that this estimator is asymptotically unbiased only under the null. It can also yield meaningless results, for example when $p_0 = p_1 = 0.5$. The cause-of-death diagnosis is then completely random, and the estimator is identically -1, in sharp contrast with its interpretation as hazard ratio. Note also that asymptotically one expects $O_0 = O_1$ here, making the expression ill-defined at 0/0. Of course one cannot expect to be able to estimate ξ here, since absolutely no information on cause-of-death is available. However, other cases may also yield negative estimates. One can show this occurs when the observed number of type 1 failures is higher or lower than both the number of type 1 failures one would observe if all failures were truly type 1 and the number one would observe if all failures were truly type 0. This makes sense since these two cases are the most extreme ones one would expect at a given p_0, p_1 and O_0+O_1 (asymptotically, thus ignoring the play of chance). Of course, in practice and at finite sample sizes this condition may be fulfilled if purely by chance an extreme number of misclassifications occur, but in practical settings with not too small numbers for each true failure type this will rarely occur.

3 Derivation of the asymptotic noncentrality parameter μ

We first note that the proof of the asymptotic normal distribution relies on the martingale central limit theorem (e.g. Andersen and Borgan (1985)). Here we focus only on deriving the asymptotic noncentrality parameter μ refered to in the main text (using assumption 3'). This derivation will be done in counting process notation, using four processes:

$$\begin{aligned} N_{j1}(t) &= \sum_{i=1}^{n} I\left(D_i \le t, F_i = 1, Z_i = j\right) \\ N_{j0}(t) &= \sum_{i=1}^{n} I\left(D_i \le t, F_i = 0, Z_i = j\right) \end{aligned}$$

with $j \in \{0,1\}$ denoting the treatment group. We can decompose each counting process into a compensator and a martingale:

$$\begin{aligned} dN_{10}(t) &= \bar{Y}_1(t)(p_1(t)e^{\phi z_i} + (1-p_0(t))e^{-\xi})h_1(t)dt + dM_{10}(t) \\ dN_{00}(t) &= \bar{Y}_0(t)(p_1(t) + (1-p_0(t))e^{-\xi})h_1(t)dt + dM_{00}(t) \\ dN_{11}(t) &= \bar{Y}_1(t)((1-p_1(t))e^{\phi z_i} + p_0(t)e^{-\xi})h_1(t)dt + dM_{11}(t) \\ dN_{01}(t) &= \bar{Y}_0(t)(1-p_1(t) + p_0(t)e^{-\xi})h_1(t)dt + dM_{01}(t) \end{aligned}$$

where $\bar{Y}_j(t)$ denotes the number at risk at time t in group j, and $\bar{Y}_0(t) + \bar{Y}_1(t) = \bar{Y}(t)$.

We start the derivation from the counting process expression for the test statistic U:

$$\begin{aligned} U^n \;=\; & \int_0^t \frac{\bar{Y}_0(s)}{\bar{Y}(s)} \frac{p_1(s)}{e^{-\xi}(1-p_0(s)) + p_1(s)} dN_{10}(s) - \int_0^t \frac{\bar{Y}_1(s)}{\bar{Y}(s)} \frac{p_1(s)}{e^{-\xi}(1-p_0(s)) + p_1(s)} dN_{00}(s) \\ & + \int_0^t \frac{\bar{Y}_0(s)}{\bar{Y}(s)} \frac{1-p_1(s)}{1-p_1(s)+e^{-\xi}p_0(s)} dN_{11}(s) - \int_0^t \frac{\bar{Y}_1(s)}{\bar{Y}(s)} \frac{1-p_1(s)}{1-p_1(s)+e^{-\xi}p_0(s)} dN_{01}(s) \end{aligned}$$

By inserting the decomposition, we can rewrite this:

$$
\begin{aligned}
U^n \;=\;& \int_0^t \frac{\bar{Y}_0(s)}{\bar{Y}(s)} \left[w_0 dM_{10}(s) + w_1 dM_{11}(s)\right] - \int_0^t \frac{\bar{Y}_1(s)}{\bar{Y}(s)} \left[w_0 dM_{00}(s) + w_1 dM_{01}(s)\right] \\
&+ \int_0^t \frac{\bar{Y}_0(s)\bar{Y}_1(s)}{\bar{Y}(s)} w_0 p_1(s) \left[h_1(s;1) - h_1(s)\right] ds \\
&+ \int_0^t \frac{\bar{Y}_0(s)\bar{Y}_1(s)}{\bar{Y}(s)} w_1 (1-p_1(s)) \left[h_1(s;1) - h_1(s)\right] ds
\end{aligned}
$$

In this we can introduce the contiguous alternative:

$$\lim_{n\to\infty} n^{1/2} \log \frac{h_1(t;1)}{h_1(t)} = \phi^* g(t)$$

from which:

$$h_1(t;1) - h_1(t) = \left(\frac{\phi^* g(t)}{n^{1/2}} + o(n^{-1/2}) \right) h_1(t)$$

For the test statistic we then find:

$$
\begin{aligned}
\frac{U^n}{n^{1/2}} \;=\;& \frac{1}{n^{1/2}} \int_0^t \frac{\bar{Y}_0(s)}{\bar{Y}(s)} \left[w_0 dM_{10}(s) + w_1 dM_{11}(s)\right] - \frac{1}{n^{1/2}} \int_0^t \frac{\bar{Y}_1(s)}{\bar{Y}(s)} \left[w_0 dM_{00}(s) + w_1 dM_{01}(s)\right] \\
&+ \int_0^t \frac{\frac{\bar{Y}_0(s)}{n}\frac{\bar{Y}_1(s)}{n}}{\frac{\bar{Y}(s)}{n}} \left[w_0 p_1(s) + w_1(1-p_1(s))\right] \left(\phi^* g(s) + n^{1/2}.o(n^{-1/2})\right) h_1(s) ds
\end{aligned}
$$

with predictable variation $\langle \frac{U^n}{n^{1/2}} \rangle$:

$$
\begin{aligned}
&\int_0^t \frac{\frac{\bar{Y}_0(s)}{n}\frac{\bar{Y}_1(s)}{n}}{\frac{\bar{Y}(s)}{n}} \left[w_0^2(p_1 h_1(s) + (1-p_0)h_0(s)) + w_1^2((1-p_1)h_1(s) + p_0 h_0(s))\right] ds \\
&\quad + \int_0^t \left(\frac{\bar{Y}_0(s)/n}{\bar{Y}(s)/n} \right)^2 \frac{\bar{Y}_1(s)}{n} \left(\frac{\phi^* g(s)}{n^{1/2}} + o(n^{-1/2}) \right) \left(w_0^2 p_1 h_1(s) + w_1^2 (1-p_1(s)) h_1(s) \right) ds
\end{aligned}
$$

The first line of the test statistic converges to 0 when n goes to infinity, as does the term with $o(n^{-1/2})$ and the second line of the predictable variation. Further assuming that $g(t)$ is constant and equal to 1, the standardized test statistic $\frac{T^n/n^{1/2}}{\sqrt{\langle T^n/n^{1/2} \rangle}}$ can thus be shown to be an asymptotically standard normal random variable plus an additional term

$$\frac{\phi^*(w_0 p_1 + w_1(1-p_1))}{\left[w_0^2(p_1 + (1-p_0)e^{-\xi}) + w_1^2(1-p_1+p_0 e^{-\xi})\right]^{1/2}} \left[\int_0^\tau \frac{s_0(t)s_1(t)}{s(t)} h_1(t) dt \right]^{1/2}$$

4 Simplified test statistic for more realistic settings

Although the adapted logrank test in itself relies only on a limited amount of assumptions, some were made to simplify the expression for the test statistic.

Although these simplifying assumptions are sometimes biologically implausible, one must keep in mind that applications at the design stage require simple assumptions. Imposing a more elaborate structure requires more and more detailed knowledge which is not always available in advance. Nevertheless, for the adapted test to be practically feasable we should accommodate more complex situations, balancing the simplicity of modelling, calculating and interpreting with the desire to fully mimic the reality. Some simple extensions may accommodate very strong and well known deviations from the simplest case, while supporting the need for a parsimonious description of the reality.

4.1 Varying the baseline hazard ratio

Assumption 3', of a constant hazard ratio ξ between the two failure types, seems unreasonable in some settings (e.g. Klein, 2006 criticizes this point relating to the paper by Goetghebeur and Ryan, 1990, although there as well extensions are rather straightforward). Very often one failure type is common at a young age or at the onset of a disease, while other failure types dominate later ages or progressed stages of the disease. For example, in the Gambian setting malaria mortality is seen to increase from neonates over post-neonates to children between 1 and 4 years of age (Jaffar *and others*, 1997). ALRI mortality on the other hand peaks in post-neonates and diminishes thereafter. Since a complete specification of ξ as a function of time is infeasible, we model these variations through a piecewise constant model for ξ. This already constitutes a major improvement in flexibility, without unnecessary complicating the model. We assume two regimens with a change point t_c, which is either known from biological considerations or estimated from the data, e.g. by maximizing a partial likelihood.

The proportionality function becomes:

$$\log\left(\frac{h_1(t)}{h_0(t)}\right) = \xi(t) = \xi_1 + I(t > t_c)\xi_2$$

where I is the usual indicator function. The standardized score statistic then becomes:

$$\frac{\sum_{\substack{i:C_i=1\\ t_i\le t_c}} w_i(t_i,F_i)(Z_i-\bar{Z}_i) + \sum_{\substack{i:C_i=1\\ t_i>t_c}} w_i(t_i,F_i)(Z_i-\bar{Z}_i)}{\sqrt{\sum_{\substack{i:C_i=1,\\ t_i\le t_c}} w_i^2(t_i,F_i)\left[\left(\sum_{j\in\mathcal{R}_i} Z_j^2/n_i\right)-\bar{Z}_i^2\right] + \sum_{\substack{i:C_i=1,\\ t_i>t_c}} w_i^2(t_i,F_i)\left[\left(\sum_{j\in\mathcal{R}_i} Z_j^2/n_i\right)-\bar{Z}_i^2\right]}} \tag{4.1}$$

where w_i is now defined differently from before:

$$w_i(t_i, F_i) = \begin{cases} \frac{1}{e^{-\xi_1}\frac{1-p_0(t_i)}{p_1(t_i)}+1} & F_i = 0, t_i \leq t_c \\ \frac{1}{e^{-\xi_1}\frac{p_0(t_i)}{1-p_1(t_i)}+1} & F_i = 1, t_i \leq t_c \\ \frac{1}{e^{-\xi_1-\xi_2}\frac{1-p_0(t_i)}{p_1(t_i)}+1} & F_i = 0, t_i > t_c \\ \frac{1}{e^{-\xi_1-\xi_2}\frac{p_0(t_i)}{1-p_1(t_i)}+1} & F_i = 1, t_i > t_c \end{cases} \tag{4.2}$$

Note that the explicit separation into two sums in equation (4.1) is not necessary under this definition.

Expression (4.1) can again be expressed in terms of quantities used in weighted logrank tests. The first term in both the numerator and denominator refers to two cause-specific weighted logrank tests that use everyone in the study in their risk set but which only use events before t_c. The second term consists of two cause-specific weighted logrank tests that only use persons at risk at time t_c, and only use events after t_c. So the test statistic is now a weighted logrank test consisting of four contributions instead of two: one for each cause before t_c, and one for each cause after t_c. Again, if constant misclassification rates are assumed, one can use quantities from standard logrank tests.

In this setting, both ξ_1 and ξ_2 may have to be estimated. For ξ_1 this is done in the same way as before, but based only on failures before t_c. Assuming constant misclassification probabilities, this becomes:

$$\widehat{e^{-\xi_1}} = \frac{O_1 p_1 - O_0(1-p_1)}{O_0 p_0 - O_1(1-p_0)}$$

where O_0 and O_1 now denote the number of events observed as type 0 and type 1 respectively, before time t_c. Estimation of ξ_2 is done through:

$$\widehat{\xi_2} = -\widehat{\xi_1} - \log\left(\frac{O_1' p_1 - O_0'(1-p_1)}{O_0' p_0 - O_1'(1-p_0)}\right)$$

where O_0' and O_1' now have to be interpreted as the number of failures per type after t_c.

4.2 Varying the misclassification probabilities

Although the derivation of the score statistic in no way restricts the time-evolution of the misclassification probabilities, it is only under the assumption of constant probabilities that it reduces to the elegant statistic (4.5) from the article. While we prefer this to promote the acceptance of the new test, in many settings this constancy is violated. This may result from a learning process or a sudden change in diagnostic method, but it may also result from an evolution in the cause-of-death structure (Maude and Ross, 1997).

The rise in mortality during and after the rainy season seen in Jaffar *and others* (1997) is most pronounced in malaria. Since a well known overlap exists between symptoms of malaria and ALRI (e.g. Redd and Bloland, 1992; O'Dempsey *and others*, 1993), a higher proportion of malaria in the deaths from other causes than ALRI would make the misclassification rates go up, resulting in a seasonal change in misclassification probabilities.

To implement certain effects of time-varying misclassification probabilities without complicating the test statistic too much, we can allow for piecewise constant misclassification probabilities with one known changepoint t_c. The misclassification probabilities are named p_0 and p_1 before time t_c, and p'_0 and p'_1 after t_c.

Since we made no assumption concerning the time-dependence of the misclassification probabilities in the derivation of the score statistic, equations (4.1) and (4.3) from the main text still hold. It is therefore easy to see that the test statistic takes exactly the same form (4.1) as in the previous extension (piecewise constant hazard ratios):

$$\frac{T^n}{\sqrt{V^n}} = \frac{\sum\limits_{\substack{i:C_i=1\\t_i\le t_c}} w_i(t_i,F_i)(Z_i-\bar{Z}_i) + \sum\limits_{\substack{i:C_i=1\\t_i>t_c}} w_i(t_i,F_i)(Z_i-\bar{Z}_i)}{\sqrt{\sum\limits_{\substack{i:C_i=1\\t_i\le t_c}} w_i^2(t_i,F_i)\left[\left(\sum\limits_{j\in\mathcal{R}_i} Z_j^2/n_i\right)-\bar{Z}_i^2\right] + \sum\limits_{\substack{i:C_i=1\\t_i>t_c}} w_i^2(t_i,F_i)\left[\left(\sum\limits_{j\in\mathcal{R}_i} Z_j^2/n_i\right)-\bar{Z}_i^2\right]}}$$

Again the weighted logrank test consists of four terms, one for each cause of death before and after the change point t_c. However, the weights w_i are now defined differently from (4.2):

$$w_i(t_i,F_i) = \begin{cases} \frac{1}{e^{-\xi}\frac{1-p_0}{p_1}+1} & F_i=0, t_i\le t_c \\ \frac{1}{e^{-\xi}\frac{p_0}{1-p_1}+1} & F_i=1, t_i\le t_c \\ \frac{1}{e^{-\xi}\frac{1-p'_0}{p'_1}+1} & F_i=0, t_i> t_c \\ \frac{1}{e^{-\xi}\frac{p'_0}{1-p'_1}+1} & F_i=1, t_i> t_c \end{cases} \tag{4.3}$$

A new estimator for ξ can be derived from the full likelihood, which leads to a fourth degree polynomial equation for which the solution is rather complicated. It is however possible to use equation (4.4) from the main text, using only the data up to time t_c. This may then be extended by using the equation on the data after time t_c (of course with the apropriate values p'_0 and p'_1) to get a second estimate $\widehat{e^{-\xi}}'$. The two estimates can then be averaged, yielding an (under the null) asymptotically unbiased estimate of ξ. Finally, we note that estimators can also be derived for situations with rapidly fluctuating misclassification probabilities.

4.3 More than one diagnostic method

The use of different diagnostic methods for different deaths is closely related to the previous topic and will occur in reality. In the Gambian setting most deaths will occur at home, but at least a small number will occur in hospital allowing for more exact determination of the underlying cause. For these observations, the misclassification probabilities will be much smaller and they should receive a different weight factor. This is essentially the same problem as before and the solution is therefore identical. If we only consider two diagnostic methods, the adapted logrank test becomes a weighted sum consisting of four terms, one for each cause-of-death at each diagnostic method. For ease of explanation, we ignore possible differences in cause-specific hazards related to the different settings using the different diagnostic methods.

We introduce a binary diagnosis indicator K_i. As before, the general adapted test statistic based on equations (4.1) and (4.3) from the main text still holds, and we can now assume constant misclassification probabilities depending on K_i. When $K_i = 0$ the misclassification probabilities are p_0 and p_1, when $K_i = 1$ the misclassification probabilities are p_0' and p_1'. The standardized test statistic now becomes

$$\frac{\sum\limits_{\substack{i:C_i=1\\K_i=0}} w_i(K_i,F_i)(Z_i-\bar{Z}_i)+\sum\limits_{\substack{i:C_i=1\\K_i=1}} w_i(K_i,F_i)(Z_i-\bar{Z}_i)}{\sqrt{\sum\limits_{\substack{i:C_i=1\\K_i=0}} w_i^2(K_i,F_i)\left[\left(\sum\limits_{j\in\mathcal{R}_i} Z_j^2/n_i\right)-\bar{Z}_i^2\right]+\sum\limits_{\substack{i:C_i=1\\K_i=1}} w_i^2(K_i,F_i)\left[\left(\sum\limits_{j\in\mathcal{R}_i} Z_j^2/n_i\right)-\bar{Z}_i^2\right]}}$$

with similar weights as before:

$$w_i(K_i,F_i)=\begin{cases}\frac{1}{e^{-\xi\frac{1-p_0}{p_1}}+1} & F_i=0,K_i=0\\ \frac{1}{e^{-\xi\frac{p_0}{1-p_1}}+1} & F_i=1,K_i=0\\ \frac{1}{e^{-\xi\frac{1-p_0'}{p_1'}}+1} & F_i=0,K_i=1\\ \frac{1}{e^{-\xi\frac{p_0'}{1-p_1'}}+1} & F_i=1,K_i=1\end{cases}$$

4.4 Introducing missing failure types

Problems with diagnosis may not only lead to misclassified causes-of-death, but also to failure in assigning a cause-of-death. Goetghebeur and Ryan (1990) used techniques similar to ours to correct logrank tests for missing causes-of-death. Since the same model assumptions are used by and large, only a minor extension of the partial likelihood is needed to integrate both problems. However, the model for the missingness depends strongly on the mechanism which induces it. We assume the probabilities of missingness to be equal among both treatment groups but different for the two causes of death. In reality, they can depend on the treatment if this induces or masks symptoms leading to a significantly more complex expression. In any case, the equality between treatment groups can be used as an approximation.

Notation is slightly changed to support the extra complexity. The probability of a missing cause-of-death is $p_i^u(t)$, with i the true cause-of-death (0 or 1, as before). If a cause-of-death is assigned, the probability of misclassification becomes $p_i^m(t)$, depending on the true failure type i. We name the misclassification indicator M (0 if the true failure type is observed, 1 else), and the 'unknown' indicator U (0 if a failure type is assigned, 1 else). After including the intensity processes for the missing data in the partial likelihood, the form of the score statistic (4.1) from the article is preserved:

$$T^n = \sum_{\substack{i:F_i=0\\U_i=0}} w_i^0(t_i)\left[z_i - \bar{Z}_i\right] + \sum_{\substack{i:F_i=1\\U_i=0}} w_i^1(t_i)\left[z_i - \bar{Z}_i\right] + \sum_{i:U_i=1} w_i^u(t_i)\left[z_i - \bar{Z}_i\right]$$

The weights w_i depend on the event time t_i and the observation type for individual i (now type 0, 1 or unspecified) indicated in the superscript:

$$\begin{cases} w_i^0(t_i) = \frac{1}{e^{-\xi}\frac{(1-p_0^u(t_i))(1-p_0^m(t_i))}{(1-p_1^u(t_i))p_1^m(t_i)}+1} & F_i = 0, U_i = 0 \\ w_i^1(t_i) = \frac{1}{e^{-\xi}\frac{(1-p_0^u(t_i))p_0^m(t_i)}{(1-p_1^u(t_i))(1-p_1^m(t_i))}+1} & F_i = 1, U_i = 0 \\ w_i^u(t_i) = \frac{1}{e^{-\xi}\frac{p_0^u(t_i)}{p_1^u(t_i)}+1} & U_i = 1 \end{cases}$$

The contributions of observations with a specified failure type are different from before, where leaving a cause-of-death unspecified was not possible. Also, an extra term deals with the information contributed by observations without a specified failure type. The weights reflect the probability that an observation of a given type at a given time is truly a type 1 failure.

The variance of the test statistic becomes

$$\begin{aligned} V^n \;\; = \;\; & \sum_{\substack{i:F_i=0\\U_i=0}} w_i^0(t_i)^2\left[\left(\sum_{j\in\mathcal{R}_i}\frac{z_j^2}{n_i}\right) - \bar{Z}_j^2\right] + \sum_{\substack{i:F_i=1\\U_i=0}} w_i^1(t_i)^2\left[\left(\sum_{j\in\mathcal{R}_i}\frac{z_j^2}{n_i}\right) - \bar{Z}_j^2\right] \\ & + \sum_{i:U_i=1} w_i^u(t_i)^2\left[\left(\sum_{j\in\mathcal{R}_i}\frac{z_j^2}{n_i}\right) - \bar{Z}_j^2\right] \end{aligned}$$

Assuming constant p_i^0, p_i^1 and p_i^u (i=true failure type), we introduce the notation

$$\begin{cases} A = \frac{(1-p_1^u)(1-p_1^m)}{(1-p_0^u)p_0^m} \\ \\ B = \frac{(1-p_1^u)p_1^m}{(1-p_0^u)(1-p_0^m)} \\ \\ C = \frac{p_1^u}{p_0^u} \end{cases} \tag{4.4}$$

Under the null, ξ can be estimated through the equation

$$0 = O_1 \frac{1-A}{A+e^{-\xi}} + O_0 \frac{1-B}{B+e^{-\xi}} + O_u \frac{1-C}{C+e^{-\xi}}$$

which reduces to a standard quadratic equation, where the negative square root of the discriminant should be used.

The test statistic now becomes an easy-to-implement weighted logrank test consisting of three terms:

$$Z^n = \frac{w^0 T_0^n + w^1 T_1^n + w^u T_u^n}{((w^0)^2 V_0^n + (w^1)^2 V_1^n + (w^u)^2 V_u^n)^{1/2}} \tag{4.5}$$

Here, T_0^n, T_1^n and T_u^n are the numerators of the standard logrank tests comparing the number of failures of type 0, type 1 and unspecified type respectively between treatment groups, while V_0^n, V_1^n and V_u^n are the squares of the denominators of the respective tests.

5 Sensitivity to misspecification of p_0 and p_1

5.1 Summary

As shown in the main paper, the corrected cause-specific analysis outperforms both the naive cause-specific and the all-cause analysis. However, these conclusions are based on asymptotic results assuming the misclassification probabilities are exactly known. Except for some very specific examples this will rarely be the case. This section aims to give a sense of the degree of uncertainty which can be allowed on the knowledge of p_0 and p_1, so that the adapted test remains a viable alternative for the all-cause analysis. We will refer to the problem of assigning the wrong cause-of-death as *misclassification*, while we will refer to the problem of using the wrong probability of misclassification as *misspecification*.

From a practical point of view two situations can arise. In the first one the goal is to analyse a given data set and we need to know what the impact of misspecification is on the power of the analysis. The second situation concerns the design stage, where one computes the needed sample size through formula (5.1) (main text) using misspecified misclassification probabilities. One can then wonder in what power the miscalculated sample size combined with an analysis based on misspecified probabilities results.

The impact of misspecification of p_0 and p_1 is assessed by means of simulations, using 1000 simulation steps for over 900 realistic settings, leading to a standard error on the estimated power of less than 1.6%. Sample sizes are calculated using formula (5.1) from the main text. Throughout we will assume the failures under study are detrimental, and the anticipated treatment effect is to lower their hazard. Formula (5.1) is then slightly conservative.

For the first situation misspecification of p_1 indeed leads to a loss of power when using the adapted method. However, even with moderate to large misspecification (up to 20%) this loss is smaller than or of the same magnitude as the excess power which results from using the conservative sample size formula (5.1). Misspecification of p_0 has no influence on the power, because of the estimate used for ξ. When introducing this estimate into the test statistic, the explicit dependence on p_0 disappears. This makes sense, since knowledge of p_1, of the number of events for each cause and of ξ completely determines p_0.

For the second situation simulation followed the same approach as before, only now the sample sizes were based on the misspecified p_0 and p_1, as they would be when considering use of the adapted statistic in the design of a study. Formula (5.1) then yields sample sizes which may be either too high or too low, leading to deviations in resulting power in both directions. Although the size of these deviations can be larger than 10%, combined with a conservative sample size formula they are expected to be between -10% and +5% if the misspecification of p_1 is kept below 20% and the misspecification of p_0 is kept below 6%.

Throughout all this, the type I error rate is controlled, a theoretical result which is confirmed by simulation.

It appears that the setting (defined by ξ, p_0 and p_0) has more impact than the size of the misspecification itself. In a typical setting as the Gambian one (p_0 small, p_1 large and cause-specific hazard for event of interest low) misspecification of p_0 has a larger effect, but the misspecification itself is expected to be smaller, resulting in a smaller impact. We found that overestimation of p_1 is probably to be prefered to underestimation. This leads to conservativeness however, meaning that the sample will be larger than strictly necessary, and thus to inefficiency.

5.2 Design of the simulation study

The adapted test uses various quantities, of which some are assumed to be known in advance and some can be estimated. These quantities are:

- p_1: the probability of misclassifying an event of interest as a competing risk
- p_0: the probability of misclassifying a competing risk as an event of interest
- ξ: log of the cause-specific baseline hazard ratio (log of the ratio of the hazard for the event of interest and the competing risk, in the control group)

In practical settings one often relies on previous research to estimate the misclassification probabilities p_1 and p_0. In rare cases, this may also be the case for ξ. In this sensitivity analysis, we assume ξ needs to be estimated from the data at hand. Therefore, the misclassification probabilities are the only parameters which may be misspecified (apart from model misspecification) and are the only parameters studied here.

In the sensitivity analysis, we take as input all combinations with parameters taken from the following possibilities:

Table 1: Input values for the sensitivity analysis.

ξ	p_0 (%)	p_1 (%)			
$\log \frac{0.0059}{0.0275}$	10	20	40	60	80
	15	20	40	60	80
	20	20	40	60	80
0	10	20	40	60	80
	15	20	40	60	80
	20	20	40	60	80
$\log \frac{0.0275}{0.0059}$	10	20	40	60	80
	15	20	40	60	80
	20	20	40	60	80

- p_1: is taken to be high: (20, 40, 60, 80) (%)
- p_0: is taken to be low: (10, 15, 20) (%)
- ξ: three cases are studied: rare event of interest, equal occurrence, rare competing risk: $\left(\log \frac{0.0059}{0.0275}, 0, \log \frac{0.0275}{0.0059}\right)$

The input values for ξ are based on the expected hazards for the event of interest and the competing risks in the setting described by Jaffar *and others* (2003). It is unlikely that the cause-of-interest is assigned by mistake, and more often the cause-of-interest is not recognized if present.

We generate data from the various combinations of these numbers, using 0.0059 and 0.0275 as the cause-specific hazards for the event of interest and the competing risks, or vice versa. The 'equal occurrence' case uses hazards of 0.01 for both event types. We assume constant cause-specific hazards, with staggered accrual over 4 years followed by an extra follow-up period of half a year (roughly approximating the trial described in Cutts *and others*, 2005). The treatment effect is taken from the article by Jaffar, and entails a reduction of the cause-specific hazard with 31.5%.

We assume the analysis of the generated data may use wrong misclassification probabilities. The difference between the value used for analysis and the true simulation value is chosen from two vectors:

- Δp_1: large misspecification is possible: (-18, -9, 0, 9, 18) (%)
- Δp_0: the misspecification is somewhat smaller: (-6, -3, 0, 3, 6) (%)

At all combinations of input parameters, all possible combinations of misspecifications are used. For each unique combination of input and analysis quantities (in total 3x4x3x5x5=900 combinations) 1000 simulations are done. The sample size for each simulation is determined from the sample size formula (5.1), and is therefore unique to each combination of input parameters.

From these simulations, we estimate the power to detect the treatment effect at a 5% significance level as the rejection rate of the null hypothesis.

When designing methods aiming to gain power, one should be careful not to inflate the type I error rate. We illustrate that for our method the type I error rate is controlled, even under misspecification of the model parameters. When this implies that ξ needs to be estimated, this follows theoretically by noting that the ξ-estimator is a consistent estimator under the null. By applying a Slutsky-like approach to the test statistic one can show that the statistic's asymptotic distribution under the null is the same regardless of whether the used ξ was known or estimated. Thus, asymptotically the type I error rate is not inflated when estimating ξ. The type I error rate under misspecification of p_0 and p_1 is discussed in the following sections.

5.3 Results for misspecification in the analysis

Effects on type I error rate at finite samples are shown by simulation under the null (ϕ=0). We used 10,000 simulations at each of the 900 settings discussed above. The minimum observed rejection rate was 4.54%, the maximum was 5.62% which illustrates that the rejection rate is not increased. Figure 2 compares the observed distribution of rejection rates with the expected distribution using 10,000 simulation steps if the 5% α-level is respected in each simulation setting. The figure shows no shift in location, indicating the type I error rate is generally not increased, and shows no increased dispersion, indicating all simulations have the same rejection rate, i.e. 5%.

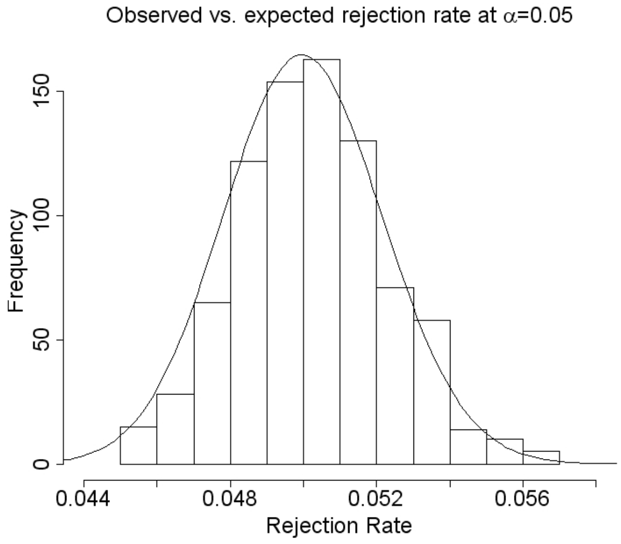

Figure 2: *Histogram of the empirical rejection rates and comparison with expectation under a controlled type I error rate, for various settings faced with misspecified p_0 and p_1 at the analysis stage.*

We can measure powerloss due to misspecification of p_0 and p_1 in two ways. The first one estimates the powerloss, comparing the analyses with and without misspecification. The second one takes the conservativeness of the used sample size formula into account, meaning that the power actually obtained with the calculated sample size is higher than the power that was asked for. The size of this conservativeness depends on the setting, but is easily in the range of some percentage points, and can therefore procure a buffer against the powerloss due to misspecification. The second method compares the obtained powers to the desired level of 80%.

The first method provides insight in the impact of misspecification, telling us what the true resulting loss of power is. This may be important when considering whether or not to invest in a pilot study to estimate p_0 and p_1. The second method on the other hand has more practical implications, telling us whether or not a study planned using the proposed sample size formula will attain the desired power. It can also be used to determine what level of misspecification is acceptable, in the sense that it will not adversely affect the power once combined with the proposed sample size formula.

5.3.1 Comparison to analysis without misspecification

We first compare the power for the analyses with and without a given misspecification. Figure 3 shows the estimated cumulative density of the powerloss induced by the misspecification, in black when ξ is estimated from the data, in red when it is known.

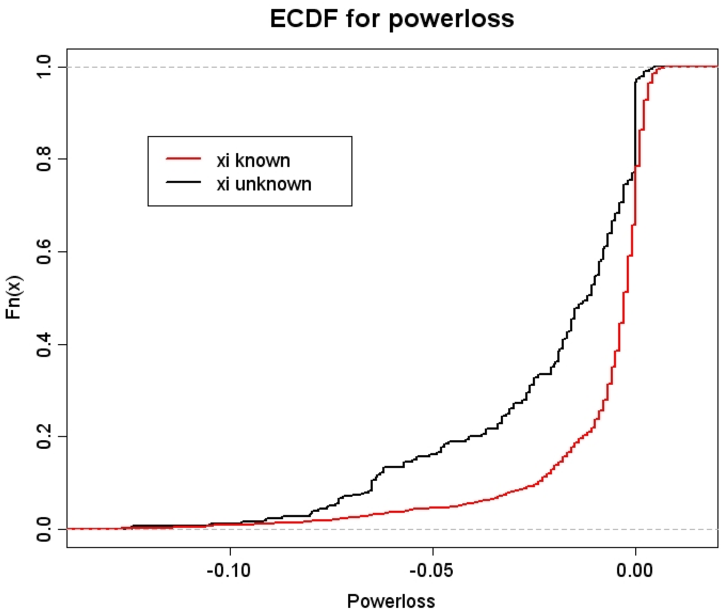

Figure 3: *Cumulative distribution of the difference in power between an analysis based on the correct p_0 and p_1 and an analysis when these are misspecified by Δp_0 and Δp_1. The black curve is for an analysis estimating ξ, the red for an analysis with exact knowledge of ξ.*

When ξ is estimated (black) the powerloss exceeds 5% in just under 20%

of the cases, but exceptions may yield a powerloss of up to 13%. A number of simulations yield a powerloss of zero, these are mainly simulations with Δp_0 and Δp_1 equal to zero. Exact knowledge of ξ seems to lower the impact of misspecification (the impact being at least 5% in less than 10% of the cases - red curve), although a small number of extreme powerlosses are still present. Similar conclusions are drawn from the boxplots in figure 4. The appearance of less datapoints in the righthand boxplot is due to overlap coming from independence of p_0 when estimating ξ.

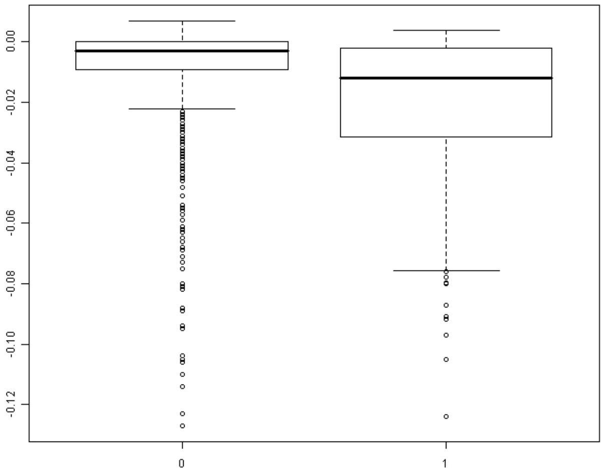

Figure 4: *Powerloss with ξ known in advance (0) or estimated (1).*

We conclude that knowing ξ reduces the impact of misspecification on power. However, we assume that in most cases ξ is unknown and focus on the case where ξ is estimated.

Our true interest lies in the connection between misspecification and powerloss. We therefore build two regressions models linking the two: one model containing all five parameters (ξ, true p_0 and p_1 and Δp_0 and Δp_1) and one model pruned by backwards elimination based on p-values. For all parameters quadratic effects will be considered, since we expect that overestimating the misclassification by 10% has a similar effect as underestimating by 10%.

We first report the first model, based on the data where ξ is estimated ($p_0, p_1, \Delta p_0, \Delta p_1$ are all expressed as probabilities, between 0 and 1):

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.336e-03  8.184e-03   0.530 0.596316
xi          -1.278e-03  3.423e-04  -3.733 0.000201 ***
true p1      2.410e-02  1.093e-02   2.206 0.027664 *
true p0      7.044e-02  1.100e-01   0.640 0.522168
delta p1    -4.404e-02  3.380e-03 -13.029  < 2e-16 ***
delta p0    -3.604e-05  1.014e-02  -0.004 0.997165
```

```
xi^2         -6.346e-04  3.852e-04  -1.648 0.099795 .
tp1^2        -6.524e-02  1.076e-02  -6.066 1.94e-09 ***
tp0^2        -2.138e-01  3.650e-01  -0.586 0.558148
dp1^2        -1.429e+00  3.174e-02 -45.023  < 2e-16 ***
dp0^2         6.475e-03  2.857e-01   0.023 0.981920
```

The second (reduced) model looks like this:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.008744   0.002448   3.572 0.000373 ***
xi          -0.001278   0.000342  -3.736 0.000199 ***
true p1      0.024098   0.010916   2.208 0.027525 *
delta p1    -0.044039   0.003377 -13.041  < 2e-16 ***
tp1^2       -0.065240   0.010745  -6.071 1.87e-09 ***
dp1^2       -1.429031   0.031712 -45.063  < 2e-16 ***
```

The model is much reduced. There is no significant effect of the misspecification of p_0 whatsoever (further investigation showed that p_0 did have a quadratic effect when ξ is exactly known). Misspecification of p_1 has a strong effect, with a quadratic trend as expected. There is also a quadratic effect of the true p_1 itself. The effects of Δp_1 are the same whether or not Δp_0 is in the model. It may be noted that although it would be interesting to see if any interactions occur, this would make the model more complicated and it would cloud the general features we are mainly looking for.

For a more quantitive view of the effect of misspecification we use the complete model to estimate the powerloss, for example at $\Delta p_1 = 20\%$ and $\Delta p_0 = 0\%$ in the Gambian setting ($\xi = -1.539$, $p_0 = 0.1$, $p_1 = 0.6$). The powerloss due to misspecification then becomes 6.5%. Although this is a significant reduction of the power, we have to keep in mind that the used misspecification is quite large. If we reduce Δp_1 to 10% for example, the expected powerloss is only 1.8%. Furthermore, the power which is gained by using this method (in comparison to all-cause analyses) is in general larger for the problems we focus on. This means that even under considerable misspecification the method is still favourable compared to others.

Finally, we illustrate the loss of power as estimated by the first model, by plotting it against Δp_1 for ξ equal to -1.539, 0 and 1.539 (figure 5). We assume $\Delta p_0 = 0\%$, $p_1 = 60\%$ and $p_0 = 10\%$. The figure shows the quadratic dependence, and an almost complete overlap between ξ equal to 0 and equal to -1.539. Extrapolating, when Δp_1 is -30%, the analysis loses about 12% power, when it is +30% the analysis loses about 14% power.

5.3.2 Comparison to prespecified power of 80%

We now wonder if the conservativeness of the proposed sample size formula is enough to counter the powerloss due to possible misspecification. The cumulative density of the power minus 0.8 is plotted for both an analysis with estimation of ξ and one with exact knowledge of ξ (figure 6).

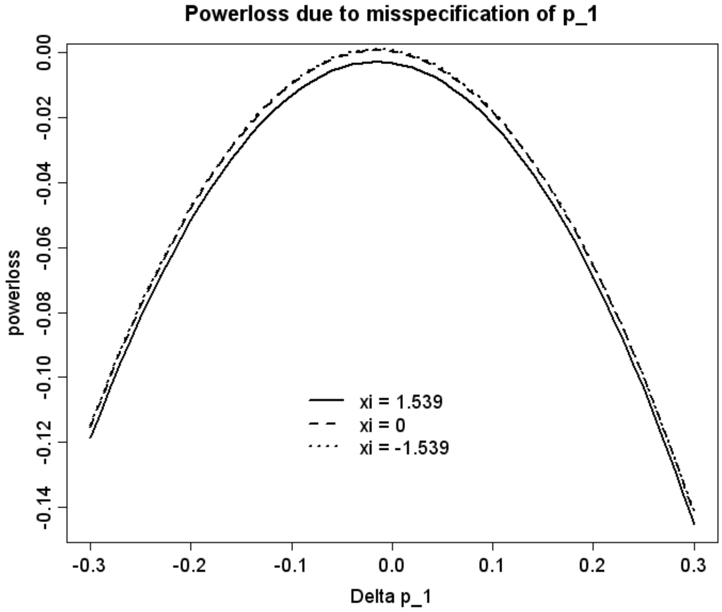

Figure 5: *Loss of power due to misspecification of p_1 at various values for ξ.*

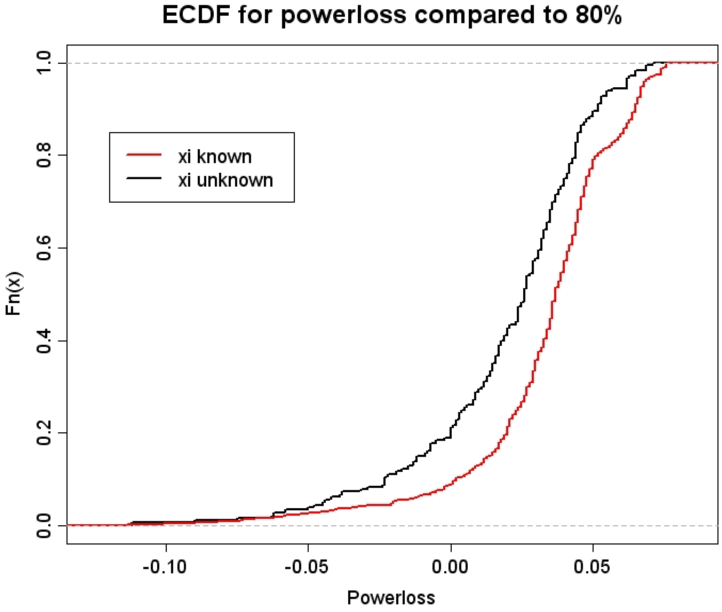

Figure 6: *Cumulative distribution of the difference between the prespecified power of 80% and the power of an analysis based on $p_0 + \Delta p_0$ and $p_1 + \Delta p_1$ (misspecified). The black curve is for an analysis estimating ξ, the red for an analysis with exact knowledge of ξ.*

Compared to figure 3, the distributions are shifted some 7% to the right, illustrating the conservativeness of the sample size formula. For ξ estimated, the power drops below 80% in a little less than 20% of the cases, but there are still some exceptions where the power is as low as 70%.

When ξ is known in advance, the power is lower than 80% in only 10% of the cases but exceptions again go as low as 70%. The difference between knowing and estimating ξ is still distinct (confirming the smaller impact of misspecification when ξ is known in advance), but smaller than before.

As before, we build two regression models:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.733e-02  9.765e-03   3.823 0.000141 ***
xi           8.040e-03  4.084e-04  19.685  < 2e-16 ***
true p1      6.796e-02  1.304e-02   5.213 2.31e-07 ***
true p0      1.510e-01  1.313e-01   1.150 0.250321
delta p1    -4.404e-02  4.033e-03 -10.920  < 2e-16 ***
delta p0    -3.604e-05  1.210e-02  -0.003 0.997624
xi^2         1.193e-04  4.596e-04   0.260 0.795258
tp1^2       -1.310e-01  1.283e-02 -10.208  < 2e-16 ***
tp0^2       -5.172e-01  4.356e-01  -1.187 0.235374
dp1^2       -1.429e+00  3.787e-02 -37.733  < 2e-16 ***
dp0^2        6.475e-03  3.408e-01   0.019 0.984847
```

Pruning the model based on the p-values leads to a more concise model:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0476859  0.0029175  16.345  < 2e-16 ***
xi           0.0080401  0.0004077  19.723  < 2e-16 ***
true p1      0.0679645  0.0130116   5.223 2.19e-07 ***
delta p1    -0.0440385  0.0040253 -10.941  < 2e-16 ***
tp1^2       -0.1309995  0.0128083 -10.228  < 2e-16 ***
dp1^2       -1.4290310  0.0377996 -37.805  < 2e-16 ***
```

Again, no significant effect of misspecification of p_0 appears in the model (when ξ is known, this effect is borderline significant). Interestingly, the intercept shows that globally the conservativeness of the sample size formula leads to a power which is about 4% higher than expected. When the event of interest occurs less frequently than the competing risk (as in the Gambian setting) this conservativeness can be 1% lower, when it occurs more often this can be 1% higher. The strength of the conservativeness as a buffer against powerloss due to misspecification therefore depends on the precise setting.

We now use the first model to illustrate the impact of the misspecification. In the Gambian example, a misspecification of 20% of p_1 leads to a power which is 3.7% lower than the anticipated 80%. We now plot the power against Δp_1 for ξ equal to -1.539, 0 and 1.539, assuming $\Delta p_0 = 0\%$, $p_1 = 60\%$ and $p_0 = 10\%$ (figure 7). As a reference, the powers for a naive cause-specific and an all-cause analysis using the same sample size are shown (respectively 71.9% and 67.2%). The higher power for the naive analysis illustrates its higher asymptotic efficiency, relative to the all-cause analysis (section 5.2 of the main text). A misspecification of over 20% is needed to make the powers for the two cause-specific analyses comparable, and well over 30% to make the all-cause analysis a viable alternative.

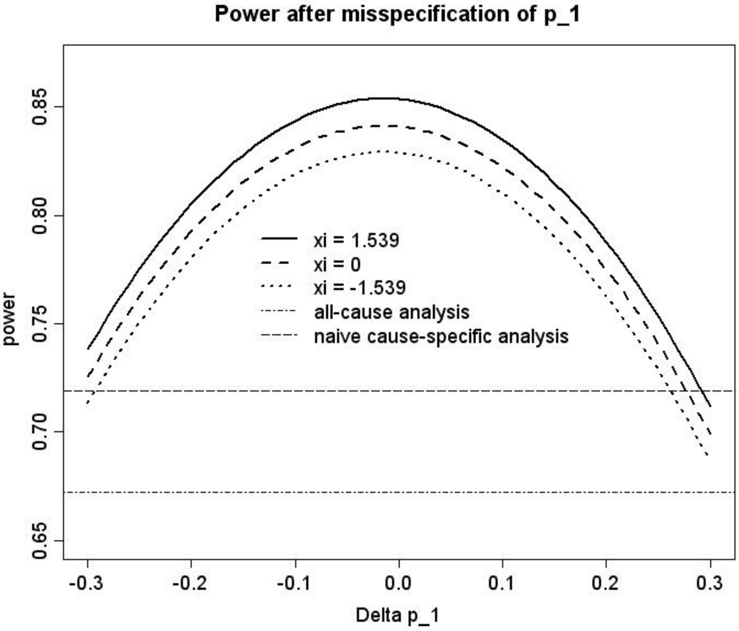

Figure 7: *Power after misspecification of* p_1 *at various values for* ξ.

We generally conclude that our adapted cause-specific test is a strong alternative for a standard all-cause analysis, even under severe misspecification of the misclassification probabilities. To protect the power of the study however, the use of the conservative sample size formula is recommended.

5.4 Misspecification at the design stage

The sensitivity analysis so far does not mimic a realistic course for a study. Indeed, sample size calculations were based on the true parameters, while in real life the statistician only has access to the possibly misspecified estimates of those parameters. It would therefore make sense to use these to determine the sample size, thus giving a more complete and realistic view of the impact that misspecification has on a study, from design to analysis.

The simulation study of this effect uses the same set-up as before, only now the sample size is determined from the misspecified probabilities. The resulting variation in sample sizes (at a given ξ, p_0, p_1-combination), was large enough to lead to differences of up to 41% for the power of the all-cause analysis. Variations decreased with increasing ξ and p_1, while the effect of p_0 was less pronounced. Similar effects on the power of the naive cause-specific analysis were observed for ξ, effects of p_0 and p_1 were less pronounced.

To assess the control of the type I error rate simulations were performed under the null. The conclusions are similar to those for misspecification at the analysis stage: no increase in type I error rate is detected. As a matter of completeness, figure 8 shows the comparison between the distribution of the empirical and the expected rejection rate.

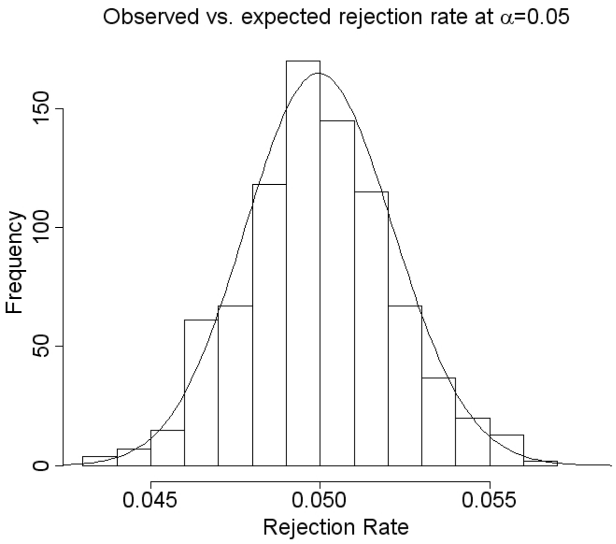

Figure 8: *Histogram of the empirical rejection rates and comparison with expectation under a controlled type I error rate, for various settings faced with misspecified p_0 and p_1 at the design stage.*

For the impact on power, we first look at the estimated cumulative density function for the difference in power between the analysis using a correct and a misspecified p_0 and p_1 (figure 9). This difference is now much more widely distributed than in figure 3: around 30% of the simulations show a difference which is smaller than -10%, while around 30% of the simulations show a positive difference, even up to 10%.

Knowledge of ξ reduces the impact of the misspecification roughly in the same way as before. This can also be seen on the boxplot (figure 10). From now on we focus on the setting where ξ is estimated.

As before we build regression models for the difference in power. All parameters (ξ, p_0, p_1 and the amounts of misspecification) were significantly present, but only for ξ, p_1 and Δp_1 the dependence showed a quadratic trend. For the misclassification of p_0 the quadratic trend is borderline significant.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.031670   0.026034  -1.216 0.224125
dss1         0.004212   0.001089   3.868 0.000118 ***
dss2         0.157491   0.034758   4.531 6.67e-06 ***
dss3         0.190668   0.349996   0.545 0.586048
dss4         0.195851   0.010753  18.214  < 2e-16 ***
dss5         0.371735   0.032258  11.524  < 2e-16 ***
```

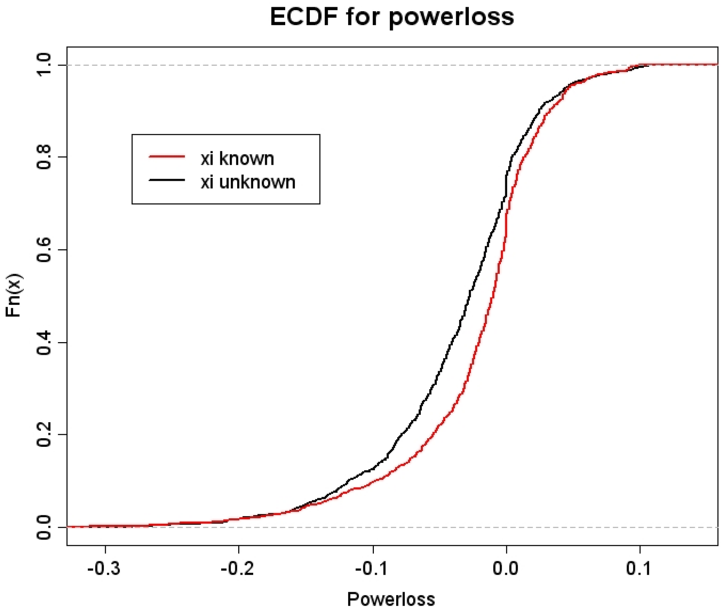

Figure 9: *Cumulative distribution of the difference in power between an analysis based on and designed through the correct p_0 and p_1 and an analysis when these are misspecified by Δp_0 and Δp_1. The black curve is for an analysis estimating ξ, the red for an analysis with exact knowledge of ξ.*

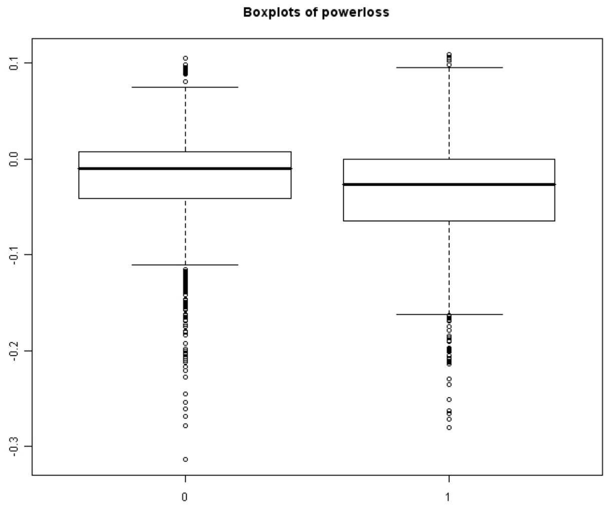

Figure 10: *Comparison of power differences with ξ known in advance (0) or estimated (1).*

```
dss12         -0.003456   0.001225  -2.820 0.004902 **
dss22         -0.218998   0.034215  -6.401 2.50e-10 ***
dss32         -0.360656   1.161288  -0.311 0.756204
```

```
dss42       -1.822911   0.100974 -18.053  < 2e-16 ***
dss52       -1.384933   0.908766  -1.524 0.127872
```

Or pruned:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.026649   0.009477  -2.812 0.005032 **
dss1         0.004212   0.001089   3.867 0.000118 ***
dss2         0.157491   0.034766   4.530 6.70e-06 ***
dss3         0.082471   0.033531   2.460 0.014102 *
dss4         0.195851   0.010755  18.210  < 2e-16 ***
dss5         0.371735   0.032266  11.521  < 2e-16 ***
dss12       -0.003456   0.001226  -2.820 0.004912 **
dss22       -0.218998   0.034223  -6.399 2.52e-10 ***
dss42       -1.822911   0.100998 -18.049  < 2e-16 ***
```

We can again illustrate the loss of power by plotting it against Δp_1 for ξ equal to -1.539, 0 and 1.539 (figure 11), assuming $\Delta p_0 = 0\%$, $p_1 = 60\%$ and $p_0 = 10\%$. Although the figure is quite similar to figure 5 the powerloss is clearly much greater, with losses of up to 25%. Also, the asymmetry is now reversed, showing a larger powerloss at underestimation of p_1. This is caused by the larger sample sizes that are used at the right hand side of the figure (see also figure 13).

Figure 12 compares the final power to the prespecified level of 80%, showing that only in around 40% of the cases the power falls below 80%.

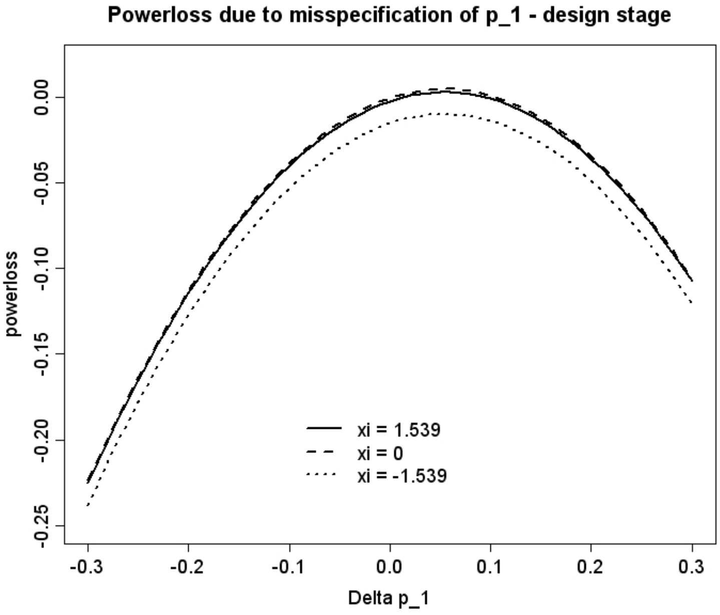

Figure 11: *Loss of power due to misspecification of p_1 at the design stage at various values for ξ.*

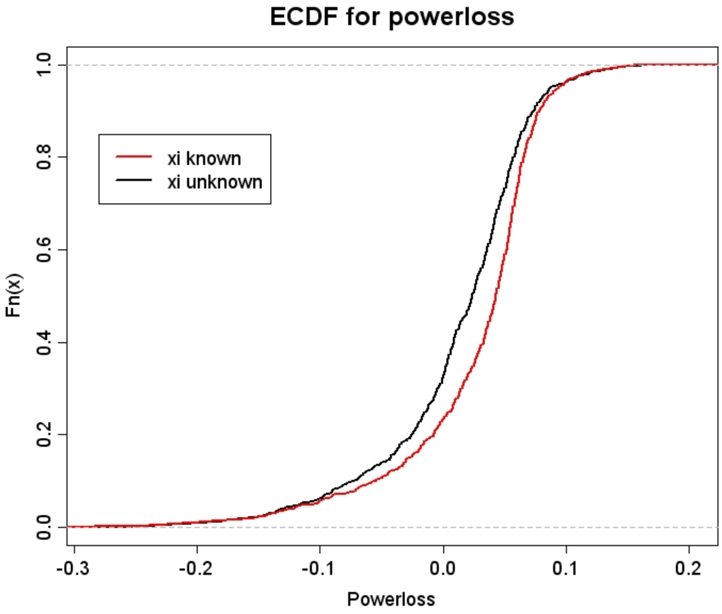

Figure 12: *Cumulative distribution of the difference between the prespecified power of 80% and the power of an analysis based on and designed through* $p_0 + \Delta p_0$ *and* $p_1 + \Delta p_1$ *(misspecified). The black curve is for an analysis estimating* ξ*, the red for an analysis with exact knowledge of* ξ*.*

We now build a linear regression model for the difference between the power and 80%.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.085751   0.025689   3.338 0.000879 ***
dss1        -0.014269   0.001075 -13.279  < 2e-16 ***
dss2         0.041822   0.034297   1.219 0.223020
dss3        -0.232693   0.345359  -0.674 0.500633
dss4         0.176266   0.009549  18.459  < 2e-16 ***
dss5         0.371735   0.031831  11.678  < 2e-16 ***
dss12       -0.003163   0.001209  -2.616 0.009045 **
dss22       -0.131568   0.033762  -3.897 0.000105 ***
dss32        0.678457   1.145905   0.592 0.553954
dss42       -1.476558   0.080706 -18.296  < 2e-16 ***
dss52       -1.384933   0.896728  -1.544 0.122840
```

Or pruned:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.064750   0.007926   8.169 1.06e-15 ***
dss1        -0.014269   0.001075 -13.275  < 2e-16 ***
dss2         0.041822   0.034307   1.219 0.223151
dss4         0.176266   0.009552  18.453  < 2e-16 ***
dss5         0.371735   0.031840  11.675  < 2e-16 ***
```

```
dss12        -0.003163   0.001209  -2.615 0.009064 **
dss22        -0.131568   0.033771  -3.896 0.000105 ***
dss42        -1.476558   0.080729 -18.290  < 2e-16 ***
```

Again we find all dependencies to be quadratic, except for p_0. A remarkable distinction with misspecification in the analysis alone is that here misspecification of p_0 has an even stronger influence than misspecification of p_1, where before it had none.

What we are most interested in here is the size of the powerloss, and more specifically whether or not the conservativeness of the sample size formula is enough to counter this. First we look at the effect of misspecifying p_1 as before (figure 13). As a reference, the powers for a naive cause-specific and an all-cause analysis using the same sample size are shown (respectively 71.9% and 67.2%). For good interpretation, one should realize these powers depend on the sample sizes which in turn depend on Δp_1, as shown in the bottom part of the plot.

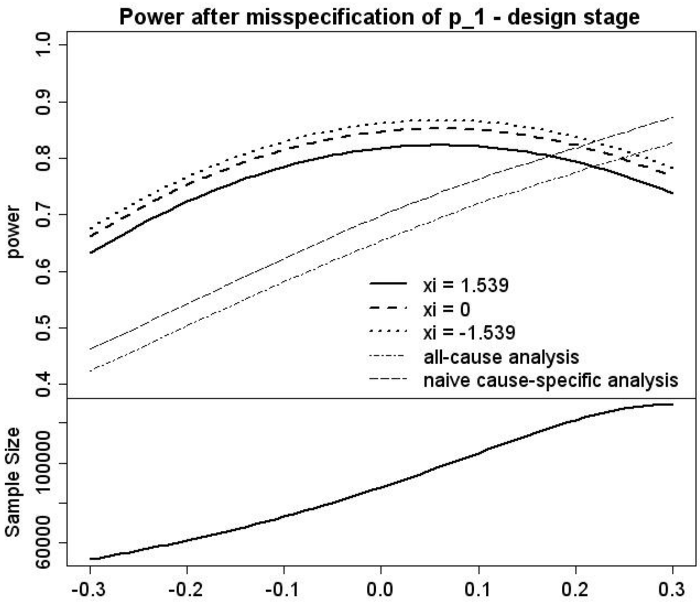

Figure 13: *Power in function of Δp_1, the misspecification of p_1 at the design stage, at various values for ξ, with indication of sample sizes as calculated through formula (5.1) from the main text.*

We now see that the naive cause-specific analysis obtains a similar (or even higher) power when we overestimate p_1 by more than 15%. The all-cause analysis obtains a similar (or even higher) power when we overestimate it by more than 20%. At this point, the sample size has risen to more than 120,000, the level derived from the *ARE*'s in section 5.4 of the main text.

Looking further into the Gambian setting, we get a power of 83.77% if p_1 is underestimated by 20%. When we additionally underestimate p_0

by 5%, we even gain power and get 85.28%. Of course, this doesn't come without a price: the sample size calculation yields a larger number making the study more costly (e.g. when the Δp_1 is -20%, the sample size used is 61,285, when Δp_1 is +20% it is 121,311). Also, the power doesn't always go up. To illustrate this, we plot the power expected from the full model as a function of misspecification size in figure 14. By extrapolating it appears that misspecifying p_0 by 10% and p_1 by 30% always has a negative impact on the power, although the severity depends on the direction of the misspecification.

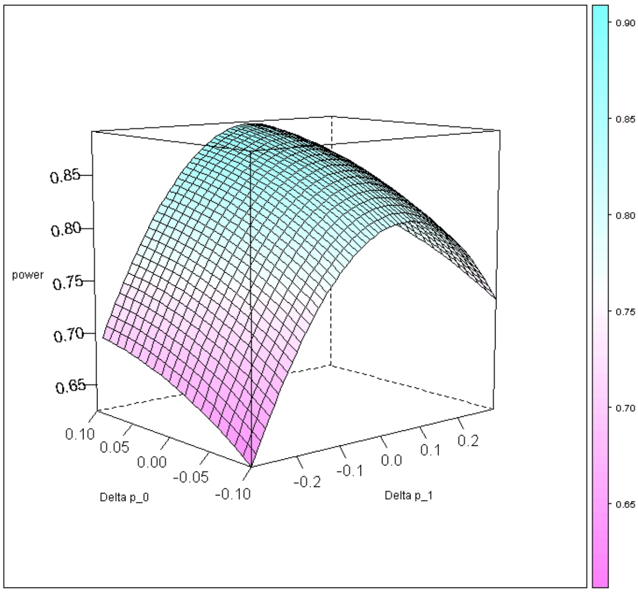

Figure 14: *Power as a function of* p_0 *and* p_1 *misspecification for the Gambian example.*

5.5 Conclusions

The setting (defined by ξ, p_0 and p_0) has more impact than the misspecification itself it appears. In a typical setting as the Gambian one (p_0 small, p_1 large and cause-specific hazard for event of interest low) misspecification of p_0 has a larger effect, but the misspecification itself is of course expected to be smaller, resulting in a smaller impact. From the models we found that overestimation of p_1 is probably to be prefered to underestimation (positive coefficient for the main effect of the misspecification). This leads to conservativeness however, meaning that the sample will be larger than strictly necessary.

5.6 Unadressed issues

This sensitivity analysis focussed on the impact on power of the misspecification of misclassification probabilities. Although this was assessed at various levels of event versus competing risk occurrence, the sensitivity to relative occurrence ξ itself was not part of the analysis. This is irrelevant if one estimates ξ, but it may be important if ξ is prespecified by the analyst.

A further issue may be model misspecification. We only adressed the case where the proportionality holds, where both event types are independent, where there are no additional parameters influencing the survival rates. In reality most of these assumptions will at best be approximately true, and it can be useful to see how the power of the test reacts to such deviations. However, extensions of the theory exist which allow some assumptions to be relaxed to some extent. Although this may resolve the issue it opens up new ways in which the analyst may misspecify the model, and thus new possible issues for a more thorough sensitivity analysis.

A final issue is the use of the conservativeness of the sample size formula in adressing sensitivity issues. We stated before that using sample size formula (5.1) leads to a conservative view on sample size, leading to a higher power than anticipated. This increase in power counteracts the powerloss due to any reasonable misspecification of the misclassification probabilities. However, it is clear that the tendency towards conservativeness of the sample size formula depends on input parameters (such as ξ, p_0 and p_1). In this sense, the general conclusion that combined use of the adapted statistic and the sample size formula leads to the expected power, even under mild misspecification of p_0 and p_1, could itself be subject to a larger sensitivity analysis, mainly using more variation in ξ.

References

ANDERSEN, P.K. AND BORGAN, 0. (1985). Counting Process models for Life History Data: A Review. *Scandinavian Journal of Statistics* **12,** 97–158.

CUTTS, F.T., ZAMAN, S.M.A., ENWERE, G., JAFFAR, S., LEVINE, O.S., OKOKO, J.B., OLUWALANA, C., VAUGHAN, A., OBARO, S.K., LEACH, A., MCADAM, K.P., BINEY, E., SAAKA, M., ONWUCHEKWA, U., YALLOP, F., PIERCE, N.F., GREENWOOD, B.M., ADEGBOLA, R.A., for the Gambian Pneumococcal Vaccine Trial Group (2005). Efficacy of nine-valent pneumococcal conjugate vaccine against pneumonia and invasive pneumococcal disease in The Gambia: randomised, double-blind, placebo-controlled trial. *The Lancet* **365,** 1139–1146.

GOETGHEBEUR, E. AND RYAN, L. (1990). A Modified Log Rank Test for Competing Risks with Missing Failure Type. *Biometrika* **77,** 207–211.

JAFFAR, S., LEACH, A., GREENWOOD, A.M., JEPSON, A., MULLER, O., OTA, M.O.C., BOJANG, K., OBARO, S. AND GREENWOOD, B.M. (1997). Changes in the pattern of infant and childhood mortality in Upper River Division, The Gambia, from 1989 to 1993. *Tropical Medicine and International Health* **2,** 28–37.

JAFFAR, S., LEACH, A., SMITH, P.G., CUTTS, F. AND GREENWOOD, B. (2003). Effects of misclassification of causes of death on the power of a trial to assess the efficacy of a pneumococcal conjugate vaccine in The Gambia. *International Journal of Epidemiology* **32,** 430-436.

KLEIN, J.P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine* **25,** 1015–1034.

MAUDE, G.H AND ROSS, D.A. (1997). The Effect of Different Sensitivity, Specificity and Cause-Specific Mortality Fractions on the Estimation of Differences in Cause-Specific Mortality Rates in Children from Studies Using Verbal Autopsies. *International Journal of Epidemiology* **26,** 1097–1106.

O'DEMPSEY, T.J.D., MCARDLE, T.F., LAURENCE, B.E., LAMONT, A.C., TODD, J.E. AND GREENWOOD, B.M. (1993). Overlap in the clinical features of pneumonia and malaria in African children. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **87,** 662-665.

REDD, S.C. AND BLOLAND, P.B. (1992). Usefulness of Clinical Case-definitions in Guiding Therapy for African Children with Malaria or Pneumonia. *The Lancet* **340,** 1140–1143.