Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Power of the Wilcoxon-Mann-Whitney test for non-inferiority in the presence of death-censored observations

Irene Schmidtmann[*,1], **Stavros Konstantinides** [2], and **Harald Binder** [3]

[1] Institute for Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Johannes Gutenberg University Mainz, D 55101 Mainz

[2] Center for Thrombosis and Hemostasis (CTH), University Medical Center Johannes Gutenberg University Mainz, D 55101 Mainz

[3] Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center – University Freiburg, Stefan-Meier-Str. 26, D 79104 Freiburg

Received zzz, revised zzz, accepted zzz

In clinical trials with patients in a critical state, death may preclude measurement of a quantitative endpoint of interest, and even early measurements, e.g. for intention-to-treat analysis, may not be available. For example, a non-negligible proportion of patients with acute pulmonary embolism will die before 30 day measurements on the efficacy of thrombolysis can be obtained. As excluding such patients may introduce bias, alternative analyses and corresponding means for sample size calculation are needed. We specifically, consider power analysis in a randomized clinical trial setting in which the goal is to demonstrate non-inferiority of a new treatment as compared to a reference treatment. Also, a non-parametric approach may be needed due to the distribution of the quantitative endpoint of interest. While some approaches have been developed in a composite endpoint setting, our focus is on the continuous endpoint affected by death-related censoring, for which no approach for non-inferiority is available. We propose a solution based on ranking the quantitative outcome and assigning "worst rank" scores to the patients without quantitative outcome because of death. Based on this, we derive power formulae for a non-inferiority test in the presence of death-censored observations, considering settings with and without ties. The approach is illustrated for an examplary clinical trial in pulmonary embolism. The results there show a substantial effect of death on power, also depending on differential effects in the two trial arms. Therefore, use of the proposed formulae is advisable whenever there is death to be expected before measurement of a quantitative primary outcome of interest.

Key words: Censoring by death; Global rank test; Non-inferiority; Combined endpoints; Pulmonary embolism;

Supporting Information for this article is available on the WWW under https://github.com/IreneSchmidtmann/GlobalRankTest

1 Introduction

In clinical trials for cardiovascular diseases, often the aim of an intervention is to improve functional capacity, measured by some quantitative surrogate endpoint such as the six minute walking distance, biomarker levels, or echocardiographic parameters of cardiovascular function.

Typically, unfavourable outcomes of the quantitative endpoint are related to a higher risk of cardiovascular death and consequently there is a non-negligible probability of an early fatal outcome. Thus, censoring by death may occur if a patient dies before the quantitative outcome can be determined. Censoring by death leads to missing values which are most unlikely to be missing at random. Therefore, ignoring this kind of missing values and excluding them from analysis does not only decrease power but may also lead to

*Corresponding author: e-mail: Irene.Schmidtmann@uni-mainz.de, Phone: +49-6131-173951, Fax: +49-6131-172968

biased estimates of treatment effect. It also contradicts the intention-to-treat principle when not all patients included in the trial are included in the analysis.

Worst-rank scores have been considered e. g. by Wittes et al (1989) and Lachin (1999) to tackle this problem. Without loss of generality we assume that higher values of the quantitative outcome are more favourable than lower values. We further assume that any quantitative outcome is better than death and that earlier death is worse than later death. This leads to an obvious ordering of outcomes which motivates the definition of worst-rank scores.

Tied worst-rank scores are obtained by allocating a rank score corresponding to a single value below the minimum observed value of the quantitative endpoint. Untied worst-rank scores can be obtained, if the time of death is taken into account. The lowest rank score is then allocated to the patient that has died first, subsequent deceased individuals receive ranks according to their time of death. After the deceased have been ranked, the subsequent ranks are allocated to the surviving patients according to their quantitative outcome.

Lachin (1999) has shown that such an approach is unbiased against a restricted one-sided alternative which states that the new treatment is better with respect to both, mortality and the quantitative endpoint or at least better with respect to one criterion and equal with respect to the other.

Another view of this approach is to consider it as a global ranking of multiple endpoints, e. g. some kind of event - such as death - and some quantitative measurement - such as a biomarker. This approach was originally suggested by O'Brien (1984) for multiple endpoints and applied in cardiology e. g. by Felker at al (2008) and Felker and Maisel (2010) combining several binary and quantitative endpoints.

We here consider the situation where a new treatment is compared to a standard treatment and the intention is to show that the new treatment is non-inferior to the standard treatment when comparing a quantitative measurement that may be censored by death. As above this corresponds to a one-sided hypothesis. The alternative hypothesis in this case states that the new treatment is non-inferior to the standard treatment with respect to both the quantitative endpoint and the mortality risk.

A Mann-Whitney test for equivalence is e. g. described by Wellek (2010); the non-inferiority test is the special case in which one equivalence boundary is set to infinity. Wellek provides mean and standard deviation under the (equivalence) null hypothesis.

Matsouaka and Bentensky (2015) have derived power and sample size formulae for the one-sided test of the null hypothesis of equality against a restricted alternative as described by Lachin (1999). They present specific formulae for common distributions of quantitative endpoint and time to event, amongst them a normally distributed quantitative endpoint and exponentially distributed time to event. We extend this approach to encompass a more general null hypothesis.

However, to the best of our knowledge, so far no power or sample size formula for the Wilcoxon-Mann-Whitney test for non-inferiority in the presence of death-censored observations is available.

In section 2, we present the application, a clinical trial in pulmonary embolism, that motivated our methodological work. After setting out the notation in section 3, the power formulae are derived in section 4. Section 5 illustrates the application of the derived formulae to the planned clinical trial. The validity of the derived formula is explored in a small simulation study. A discussion and concluding remarks are provided in section 6.

2 Thrombolysis application

The methodological considerations presented in this paper were motivated by planning a clinical trial studying treatment of patients with intermediate-high-risk pulmonary embolism. These are patients who are normotensive and appear hemodynamically stable at presentation, but have evidence of right ventricular dysfunction on echocardiography in addition to elevated cardiac troponin levels in the circulation.

In this trial, the standard would be to administer standard-dose systemic thrombolytic treatment. The new experimental treatment would be catheter-directed, ultrasound-assisted low dose thrombolysis. The

right to left ventricular (RV/LV) diameter ratio on echocardiography 24 hours after the intervention was considered a suitable primary endpoint. Unfortunately, fatal outcomes do occur within hours after treatment in patients with this condition. This precludes the observation of the RV/LV diameter ratio 24 hours after the intervention. Therefore, taking into account censoring by death of the quantitative endpoint was necessary.

While little difference in the primary endpoint and mortality was expected, the potential benefit of the low dose thrombolysis would be lower bleeding risk. Therefore, as primary endpoint analysis, we chose to test for non-inferiority and to apply a Wilcoxon-Mann-Whitney test for non-inferiority in the presence of death-censored observations.

3 Notation and hypotheses

Let $\tau(\tau > 0)$ denote the time from treatment to planned determination of the quantitative endpoint. This is intended to be identical for all patients. All patients are observed at least until time τ unless they die before this time. If patients die before time τ it is assumed that the time of death can be determined sufficiently precisely. Let further $X_{01}, \ldots, X_{0n_0}$ denote the values of the quantitative endpoint in the n_0 indiviudals in the reference group with standard treatment ($i = 0$) and $X_{11}, \ldots, X_{1n_1}$ the values of the quantitative endpoint in the n_1 indiviudals in the group with the new experimental treatment ($i = 1$). The individual event times in group i are given by $T_{i1}, \ldots, T_{in_i}$. They may not be observed if patients are alive at time τ. Let $D_{ik} = 1$ indicate that subject k in group i dies before the quantitative endpoint X_{ik} can be determined, i. e. $D_{ik} = 1$ if $T_{ik} < \tau$, $D_{ik} = 0$ otherwise. The survival probability in group i up to time τ is given by $q_i = S_i(\tau)$, where $S(t)$ is the survival function. Accordingly, the cumulative mortality in group i up to time τ is given by $p_i = 1 - q_i = \mathrm{P}(D_{ik} = 1)$.

Here, without loss of generality, we assume that high values of X_{ik} are favourable. Otherwise we could consider $-X_{ik}$. We further assume that death is worse than any quantitative outcome and – in the untied worst rank case – that early death is worse thant later death.

Following Matsouaka and Betensky (2015), we introduce a new variable, on which ranking can be based. This variable is constructed such that all patients who die before time τ have lower values than the patients who survive past τ and are ranked according to their survival times. Patients who survive past time τ are ranked according to their observed value of the quantitative endpoint. The new variable is defined as $\widetilde{X}_{ik} = D_{ik}(\eta + T_{ik}) + (1 - D_{ik})X_{ik}$ with $\eta = \min(X_{01}, \ldots, X_{0n_0}, X_{11}, \ldots, X_{1n_1}) - 1 - \tau$ in the untied case. In the tied case all patients who die before the quantitative endpoint can be determined have the same rank. Hence, a different definition of $\widetilde{X}_{ik}$ is needed and we set $\widetilde{X}_{ik} = D_{ik}\zeta + (1 - D_{ik})X_{ik}$ with $\zeta = \min(X_{01}, \ldots, X_{0n_0}, X_{11}, \ldots, X_{1n_1}) - 1$. In both cases, the minimum is taken over the observed X_{ik}. Note, that $\widetilde{X}_{ik} = X_{ik}$ holds for patients who survive past τ in both cases.

The non-inferiority hypotheses are

$$H_0 : \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) \leq \frac{1}{2} - \varepsilon$$
$$H_1 : \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) > \frac{1}{2} - \varepsilon$$

with non-inferiority margin $\varepsilon > 0$ for any pair (k, l) of observations from the two treatment groups. As $\varepsilon > 0$ is arbitrary in our approach, this is an extention of the suggestion by Matsouaka and Betensky in which $\varepsilon = 0$.

4 Power calculation

4.1 Untied case

The $\widetilde{X}_{ik}$ as defined in section 3 are used to determine ranks and to compute the Wilcoxon-Mann-Whitney test statistic $U = (n_0 n_1)^{-1}(\sum_{k=1}^{n_0}\sum_{l=1}^{n_1} I(\widetilde{X}_{0k} < \widetilde{X}_{1l}))$.

In order to evaluate the power of the Wilcoxon-Mann-Whitney test, we use the standardised version of the Wilcoxon-Mann-Whitney U statistic, i. e. the test statistic is given by $U^* = (U - \mu_0(U))/\sigma_0(U)$, where $\mu_0(U)$ is the expectation and $\sigma_0^2(U)$ the variance of U under the null hypothesis. Then under the null hypothesis U^* asymptotically follows a standard normal distribution. The null hypothesis is rejected at level α if $U^* > z_{1-\alpha}$ where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-percentile of the normal distribution.

Therefore we need expectation $\mu(U)$ and variance $\sigma^2(U)$ of U. They were derived by Matsouaka and Betensky (2015) and are as follows:

$$\begin{aligned}\mu(U) &= \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) \text{ for arbitrary } \widetilde{X}_{0k}, \widetilde{X}_{1l} \\ &= p_0 p_1 \pi_{t1} + p_0 q_1 + q_0 q_1 \pi_{x1} \\ &= \pi_{U1}\end{aligned}$$

with

$$\begin{aligned}\pi_{X1} &= \mathrm{P}(X_{0k} < X_{1l}) \\ \pi_{T1} &= \mathrm{P}(T_{0k} < T_{1l} | D_{0k} = D_{1l} = 1) \\ \sigma^2(U) &= (n_0 n_1)^{-1}\left\{\pi_{U1}(1-\pi_{U1}) + (n_0-1)(\pi_{U2} - \pi_{U1}^2) + (n_1-1)(\pi_{U3} - \pi_{U1}^2)\right\}\end{aligned}$$

with

$$\begin{aligned}\pi_{U2} &= \mathrm{P}(\widetilde{X}_{0k'} < \widetilde{X}_{1l}, \widetilde{X}_{0k'} < \widetilde{X}_{1l}) \\ &= p_0^2 q_1 + p_0^2 p_1 \pi_{t2} + 2 p_0 q_0 q_1 \pi_{x1} + q_0^2 q_1 \pi_{x2} \\ \pi_{X2} &= \mathrm{P}(X_{0k} < X_{1l}, X_{0k'} < X_{1l}) \\ \pi_{T2} &= \mathrm{P}(T_{0k} < T_{1l}, T_{0k'} < T_{1l} | D_{0k} = D_{0k'} = D_{1l} = 1) \\ \pi_{U3} &= \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}, \widetilde{X}_{0k} < \widetilde{X}_{1l'}) \\ &= p_0 q_1^2 + p_0 p_1^2 \pi_{t3} + 2 p_0 p_1 q_1 \pi_{t1} + q_0 q_1^2 \pi_{x3} \\ \pi_{X3} &= \mathrm{P}(X_{0k} < X_{1l}, X_{0k} < X_{1l'}) \\ \pi_{T3} &= \mathrm{P}(T_{0k} < T_{1l}, T_{0k} < T_{1l'} | D_{0k} = D_{1l} = D_{1l'} = 1) \\ &\quad \text{for arbitrary } \widetilde{X}_{0k}, \widetilde{X}_{0k'}, \widetilde{X}_{1l}, \widetilde{X}_{1l'}, k \neq k', l \neq l'\end{aligned}$$

If assumptions about the distribution of quantitative endpoint and event time are made, $\mu(U)$ and $\sigma(U)$ can be computed. When determining the power, we consider the boundary of the null hypothesis, i. e. $\mu_0(U) = \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) = 1/2 - \varepsilon$. The power of the test for a specific alternative, where $\mu(U) = \mu_1(U)$

and $\sigma(U) = \sigma_1(U)$, is given by

$$\begin{aligned}
1-\beta &= \mathrm{P}\left(\frac{U-\mu_0(U)}{\sigma_0(U)} > z_{1-\alpha} \mid H_1\right) \\
&= \mathrm{P}\left(\frac{U-\mu_1(U)}{\sigma_1(U)} > z_{1-\alpha}\frac{\sigma_0(U)}{\sigma_1(U)} + \frac{\mu_0(U)-\mu_1(U)}{\sigma_1(U)}\right) \\
&= \Phi\left(-z_{1-\alpha}\frac{\sigma_0(U)}{\sigma_1(U)} - \frac{\mu_0(U)-\mu_1(U)}{\sigma_1(U)}\right) \\
&= \Phi\left(z_{\alpha}\frac{\sigma_0(U)}{\sigma_1(U)} + \frac{\mu_1(U)-\mu_0(U)}{\sigma_1(U)}\right).
\end{aligned}$$

If the specific alternative considered is equivalence of treatments, i. e. $\widetilde{X}_{0k}$ and $\widetilde{X}_{1l}$ are from the same distribution, we have $\mu_1(U) = \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) = 1/2$ and $\sigma_1^2(U) = (n_0+n_1+1)/12n_0n_1$. The power is then given by

$$\begin{aligned}
1-\beta &= \Phi\left(z_{\alpha}\frac{\sigma_0(U)}{\sigma_1(U)} + \frac{\varepsilon}{\sigma_1(U)}\right) \\
&= \Phi\left(z_{\alpha}\sqrt{\frac{12\left\{\pi_{U1}(1-\pi_{U1}) + (n_0-1)(\pi_{U2}-\pi_{U1}^2) + (n_1-1)(\pi_{U3}-\pi_{U1}^2)\right\}}{n_0+n_1+1}}\right. \\
&\qquad \left. +\varepsilon\sqrt{\frac{12n_0n_1}{n_0+n_1+1}}\right).
\end{aligned}$$

4.2 Tied case

In the tied case we use the modified Wilcoxon-Mann-Whitney statistic $\widetilde{U}$ which allows for ties and is defined as $\widetilde{U} = (n_0n_1)^{-1}\sum_{k=1}^{n_0}\sum_{l=1}^{n_1}\left\{I(\widetilde{X}_{0k} < \widetilde{X}_{1l}) + I(\widetilde{X}_{0k} = \widetilde{X}_{1l})/2\right\}$. As before, for determining the power the standardised version of the test statistic is used: $\widetilde{U}^* = (\widetilde{U} - \mu_0(\widetilde{U}))/\sigma_0(\widetilde{U})$ where $\mu_0(\widetilde{U})$ is the expectation and $\sigma_0^2(\widetilde{U})$ the variance of $\widetilde{U}$ under the null hypothesis. Under the null hypothesis $\widetilde{U}^*$ also follows asymptotically a standard normal distribution. Again, the null hypothesis is rejected at level α if $\widetilde{U}^* > z_{1-\alpha}$.

Matsouaka and Betensky (2015) also derived expectation and variance in the tied case which are as follows:

$$\begin{aligned}
\mu(\widetilde{U}) &= \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) + \mathrm{P}(\widetilde{X}_{0k} = \widetilde{X}_{1l})/2 \text{ for arbitrary } \widetilde{X}_{0k}, \widetilde{X}_{1l} \\
&= \frac{p_0p_1}{2} + p_0q_1 + q_0q_1\pi_{x1} \\
&= \pi_{\widetilde{U}1} \\
\sigma^2(\widetilde{U}) &= (n_0n_1)^{-1}\left\{\pi_{\widetilde{U}1}(1-\pi_{\widetilde{U}1}) + (n_0-1)\left(\pi_{\widetilde{U}2} - \pi_{\widetilde{U}1}^2 - \frac{p_0^2p_1}{12}\right)\right. \\
&\qquad \left. +(n_1-1)\left(\pi_{\widetilde{U}3} - \pi_{\widetilde{U}1}^2 - \frac{p_0p_1^2}{12}\right) - \frac{p_0p_1}{4}\right\}
\end{aligned}$$

with

$$\pi_{\widetilde{U}2} = p_0^2 q_1 + \frac{p_0^2 p_1}{3} + 2p_0 q_0 q_1 \pi_{X1} + q_0^2 q_1 \pi_{X2}$$
$$\pi_{\widetilde{U}3} = p_0 q_1^2 + \frac{p_0 p_1^2}{3} + p_0 p_1 q_1 + q_0 q_1^2 \pi_{X3}, \text{ where } \pi_{X1}, \pi_{X2}, \pi_{X3} \text{ are as in the untied case.}$$

Again, $\mu(\widetilde{U})$ and $\sigma(\widetilde{U})$ under the null hypothesis or under the alternative can be computed if assumptions about the distributions of quantitative endpoint and event time are made. As previously, we consider the boundary of the null hypothesis, i. e. $\mu_0(\widetilde{U}) = \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) = \frac{1}{2} - \varepsilon$ when assessing the power of the test. The power is given by

$$\begin{aligned} 1 - \beta &= \mathrm{P}\left(\frac{\widetilde{U} - \mu_0(\widetilde{U})}{\sigma_0(\widetilde{U})} > z_{1-\alpha} \mid H_1 \right) \\ &= \Phi\left(z_\alpha \frac{\sigma_0(\widetilde{U})}{\sigma_1(\widetilde{U})} + \frac{\mu_1(\widetilde{U}) - \mu_0(\widetilde{U})}{\sigma_1(\widetilde{U})} \right) \end{aligned}$$

If the specific alternative considered is equivalence of treatments, i. e. $\widetilde{X}_{0k}$ and $\widetilde{X}_{1l}$ are from the same distribution, we have $p_0 = p_1 = p, q_0 = q_1 = q, \pi_{X1} = 1/2, \pi_{X2} = \pi_{X3} = 1/3$ and hence $\pi_{\widetilde{U}1} = 1/2, \pi_{\widetilde{U}2} = \pi_{\widetilde{U}3} = 1/3$. Consequently, $\mu_1(\widetilde{U}) = \mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) = 1/2$ and
$\sigma_1^2(\widetilde{U}) = \left\{(n_0 + n_1 + 1) - p^2\left[3 + (n_0 + n_1 - 2)p\right]\right\} / (12 n_0 n_1)$.
The power is then given by

$$\begin{aligned} 1 - \beta &= \Phi\left(z_\alpha \frac{\sigma_0(\widetilde{U})}{\sigma_1(\widetilde{U})} + \frac{\varepsilon}{\sigma_1(\widetilde{U})} \right) \\ &= \Phi\left(z_\alpha \sqrt{\frac{\left\{ \pi_{\widetilde{U}1}(1-\pi_{\widetilde{U}1}) + (n_0-1)(\pi_{\widetilde{U}2} - \pi_{\widetilde{U}1}^2 - \frac{p_0^2 p_1}{12}) + (n_1-1)(\pi_{\widetilde{U}3} - \pi_{\widetilde{U}1}^2 - \frac{p_0 p_1^2}{12}) - \frac{p_0 p_1}{4} \right\}}{\{(n_0+n_1+1) - p^2[3+(n_0+n_1-2)p]\}/12}} \right. \\ &\qquad \left. + \varepsilon \sqrt{\frac{12}{(n_0 + n_1 + 1) - p^2\left[3 + (n_0 + n_1 - 2)p\right]}} \right) \end{aligned}$$

5 Application

For the clinical application we assume that the RV/LV reduction is approximately normally distributed with common variance for both treatment groups: $X_i \sim \mathcal{N}(\mu_i, \sigma^2), (i = 0, 1)$. Under the standard treatment RV/LV reduction is expected to be $\mu_0 = 0.3$ on average with standard deviation $\sigma = 0.1$. A difference of $\varepsilon^* = 0.05$ is deemed acceptable. As the quantitative endpoint is assumed to be normally distributed, the acceptable difference can be expressed as a multiple c of the standard deviation, here $c = 0.5$. This would be used as non-inferiority margin in a parametric test if no censoring by death occurred. The corresponding parametric hypotheses are $H_0 : \mu_1 \leq \mu_0 - \varepsilon^* = \mu_0 - c\sigma$ and $H_1 : \mu_1 > \mu_0 - \varepsilon^* = \mu_0 - c\sigma$ with $\mu_0 = 0.3$, $\sigma = 0.1$ and $c = 0.5$. We chose $\alpha = 0.025$ for this one-sided test. We explore the power as a function of sample size and the probabilities of death in both treatment groups for the specific alternative of equivalence of treatments. The ratio of $n_0 : n_1$ is chosen as $1 : 2$ in our application.

In a first step we determine the non-parametric non-inferiority margin $\tilde{\varepsilon}$ for the case of no censoring by death such that $H_0 : P(X_0 < X_1) \leq \frac{1}{2} - \tilde{\varepsilon}$ and $H_1 : P(X_0 < X_1) > \frac{1}{2} - \tilde{\varepsilon}$.

As $\Phi(x)$ is monotonically increasing under H_0 we have $P(X_0 < X_1) = \Phi((\mu_1 - \mu_0)/\sqrt{2}\sigma) \leq \Phi(-\varepsilon^*/\sqrt{2}\sigma) = \Phi(-c/\sqrt{2})$. Therefore, if no censoring by death occured, $\tilde{\varepsilon}$ would be chosen such that $1/2 - \tilde{\varepsilon} = \Phi(-c/\sqrt{2})$ and hence with $c = 0.5$ we chose $\tilde{\varepsilon} = 1/2 - \Phi(-1/\sqrt{8}) \approx 0.1382$.

All calculations were performed using `R 3.4.4`, packages `knitr, xtable, latex2exp,` and `flexsurv`.

5.1 Untied case

For the untied case we assume that the time to death follows an exponential distribution, i. e. $T_i \sim exp(\lambda_i), (i = 0, 1)$, i. e. $q_i = exp(-\lambda_i \tau)$. Without loss of generality we can assume that $\tau = 1$. To determine the overall non-inferiority margin ε, again consider the boundary of the null hypothesis, i. e. $\mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) = 1/2 - \varepsilon$.

$$\begin{aligned}
\mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) &= p_0 p_1 \mathrm{P}(T_{0k} < T_{1l}) + p_0 q_1 + q_0 q_1 P(X_{0k} < X_{1l}) \\
&= p_0 p_1 \frac{\lambda_0}{\lambda_0 + \lambda_1} + p_0(1 - p_1) + (1 - p_0)(1 - p_1)(\frac{1}{2} - \tilde{\varepsilon}) \\
&= \frac{1}{2} - \left((1 - p_0)(1 - p_1)\tilde{\varepsilon} + \gamma p_0 p_1 + \frac{1}{2}(p_1 - p_0) \right) \text{ with } \gamma = \frac{1}{2} - \frac{\lambda_0}{\lambda_0 + \lambda_1}
\end{aligned}$$

and hence, as $\tilde{\varepsilon} = 1/2 - \Phi(-c/\sqrt{2})$,

$$\begin{aligned}
\varepsilon &= (1 - p_0)(1 - p_1)\tilde{\varepsilon} + \gamma p_0 p_1 + (p_1 - p_0)/2 \\
&= (1 - p_0)(1 - p_1)(1/2 - \Phi(-c/\sqrt{2})) + \gamma p_0 p_1 + (p_1 - p_0)/2
\end{aligned}$$

For $p_0 = p_1 = p$ this simplifies to

$$\varepsilon = (1 - p)^2 \tilde{\varepsilon} = (1 - p)^2 (1/2 - \Phi(-c/\sqrt{2})).$$

```
## pdf
##   2
```

In figure 1 the solid black line is identical in all panels; it depicts the situation when no censoring due to death occurs. Figure 1A describes the situation when mortality is the same in both treatment groups. For given sample size power decreases with an increase in probability of censoring by death. Figure 1B also exhibits this effect, however, here the risk of censoring by death is 20% higher in the new treatment group under null hypothesis and the power is less affected by the risk of censoring by death. If λ_0 and λ_1 are such that $\lambda_0/(\lambda_0 + \lambda_1) = 1/2 - \tilde{\varepsilon}$ then

$$\begin{aligned}
\mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) &= p_0 p_1 (1/2 - \tilde{\varepsilon}) + p_0(1 - p_1) + (1 - p_0)(1 - p_1)(1/2 - \tilde{\varepsilon}) \\
&= (1/2 - \tilde{\varepsilon})(1 + p_0 - p_1) - 2\tilde{\varepsilon} p_0 p_1 \\
&\approx 1/2 - \tilde{\varepsilon} \text{ if } p_0, p_1 \ll 1
\end{aligned}$$

This effect can be seen in figure 1C where the power for the smaller probabilities for death is very similar to the situation where there is no censoring by death. In figure 1D the assumption is that under the null hypothesis the risk for death ist 2.5 fold in the new treatment group compared to the reference group. In this case the computed power increases with increasing probability of censoring due to death. Values displayed in figure 1 are also provided in table 4 in the appendix.

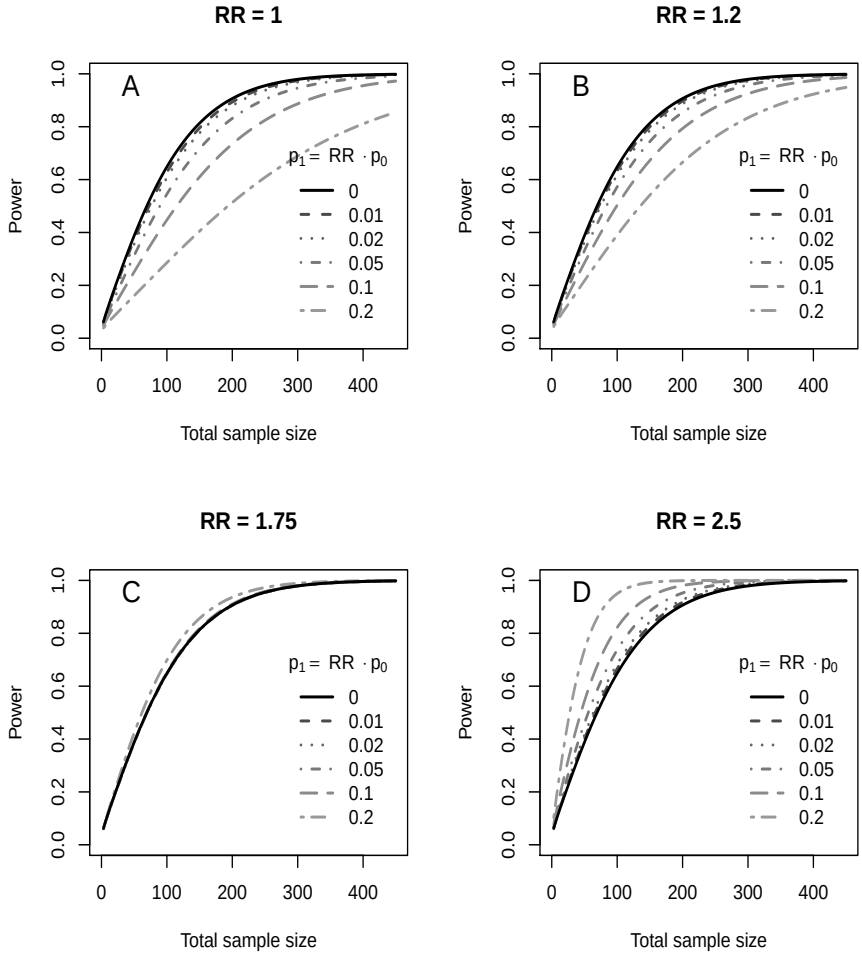

Figure 1 Power and sample size for untied case with sample size allocation for reference to new treatment 1:2. In each panel a different relative risk (RR) of death in the new treatment group (risk p_1) compared to the reference group (risk p_0) is assumed under the null hypothesis. The power is computed for the alternative of equivalence of both treatments. Different line types correspond to different risks in the reference group.

5.2 Tied case

For the tied case it suffices to make an assumption about the probability of death in each group under the null hypothesis. To determine the overall non-inferiority margin we again consider the boundary of the null hypothesis $\mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) = 1/2 - \varepsilon$.

$$\begin{aligned}\mathrm{P}(\widetilde{X}_{0k} < \widetilde{X}_{1l}) &= p_0p_1/2 + p_0q_1 + q_0q_1P(X_{0k} < X_{1l}) \\ &= p_0 - p_0p_1/2 + (1-p_0)(1-p_1)/2 - (1-p_0)(1-p_1)\tilde{\varepsilon} \\ &= 1/2 - (p_1-p_0)/2 - (1-p_0)(1-p_1)\tilde{\varepsilon}\end{aligned}$$

and hence, again considering that $\tilde{\varepsilon} = 1/2 - \Phi(-c/\sqrt{2})$,

$$\begin{aligned}\varepsilon &= (p_1 - p_0)/2 + (1 - p_0)(1 - p_1)\tilde{\varepsilon} \\ &= (p_1 - p_0)/2 + (1 - p_0)(1 - p_1)(1/2 - \Phi(-c/\sqrt{2}))\end{aligned}$$

For $p_0 = p_1 = p$ this also simplifies to

$$\varepsilon = (1 - p)^2\tilde{\varepsilon} = (1 - p)^2(1/2 - \Phi(-c/\sqrt{2})).$$

Table 1 Effective non-inferiority margin and total samples size as function of relative risk (RR) of death in the new treatment group (risk p_1) compared to the reference group (risk p_0) for given $\alpha = 0.025$ and power of 80% with sample size allocation for reference to new treatment 1:2.

RR	p_0	$\varepsilon^{(\text{untied})}$	$n_{\text{total}}^{(\text{untied})}$	$\varepsilon^{(\text{tied})}$	$n_{\text{total}}^{(\text{tied})}$
1	0.00	0.138	147	0.138	147
	0.01	0.135	153	0.135	153
	0.02	0.133	162	0.133	162
	0.05	0.125	186	0.125	186
	0.10	0.112	237	0.112	237
	0.20	0.088	390	0.088	390
1.2	0.00	0.138	147	0.138	147
	0.01	0.136	153	0.136	153
	0.02	0.134	156	0.134	156
	0.05	0.128	174	0.128	174
	0.10	0.119	204	0.119	204
	0.20	0.104	276	0.104	276
1.75	0.00	0.138	147	0.138	147
	0.01	0.138	147	0.138	147
	0.02	0.138	147	0.138	147
	0.05	0.139	147	0.139	147
	0.10	0.140	144	0.140	144
	0.20	0.148	129	0.147	129
2.5	0.00	0.138	147	0.138	147
	0.01	0.141	141	0.141	141
	0.02	0.144	135	0.144	135
	0.05	0.152	120	0.152	120
	0.10	0.169	96	0.168	96
	0.20	0.209	60	0.205	60

In table 1 the effective non-inferiority margin ε for the Wilcoxon-Mann-Whitney test for non-inferiority and the total number of cases needed to obtain a power of 80% at signifcance level $\alpha = 0.025$ are shown as function of the probability of death in the reference group p_0 and the relative risk RR of death in the new treatment group compared to the reference group under H_0. In our example there is little difference in power between the tied case and the untied case for given sample size. Specifically, effective non-inferiority margins and sample sizes are identical in the tied and untied case if the probability of death is identical in both treatment groups. The values are displayed in table 4 in the appendix.

So far we based the choice of the non-inferiority margin on assumptions concerning the quantitative endpoint, starting from the assumption of zero probability of censoring by death in both groups under the null hypothesis. This was extended to investigating several non-zero constellations of probability of censoring by death in both treatment groups. An alternative would be to start from the assumption of identical

distributions of the quantitative endpoint and to derive the non-inferiority margin from an acceptable difference in risks of death. In this case $c = 0$ and the formula for the non-inferiority margin ε in the tied case simplifies to $\varepsilon = (p_1 - p_0)/2$. In the untied case the formula for the non-inferiority margin is given by $\varepsilon = \gamma p_0 p_1 + (p_1 - p_0)/2$.

5.3 Simulation

To assess whether the accuracy of the power formula holds if the time to death does not follow an exponential distribution we performed a small simulation study. In this we assumed that time to event follows a log-logistic distribution with hazard $h(t; \alpha, \beta) = \beta t^{\beta-1}/(\alpha^\beta + t^\beta)$, with scale parameter $\alpha > 0$ and shape parameter $\beta > 0$. Scale α was chosen such that corresponding cumulative incidence functions matched the pre-specified probabilities of death at time $\tau = 1$. While the shape β was assumed to be identical in both groups, it was varied between simulation scenarios assuming values 0.8, 1.0, and 1.2. The probabilities of censoring by death and the relative risks between the reference group and new treatment group were chosen as in 5.2, the total sample size was chosen as determined in table 1. 1000 datasets were generated and evaluated in each scenario. The results of the simualation are shown in table 2.

Table 2 Observed power as function of relative risk (RR) of death under the null hypothesis in the new treatment group (risk p_1) compared to the reference group (risk p_0) and total sample size n. Significance level $\alpha = 0.025$, tied and untied case, shape parameter β of the log-logistic time to event distribution. Sample size allocation for reference to new treatment 1:2.

RR	p_0	n	untied: $\beta = 0.8$	$\beta = 1.0$	$\beta = 1.2$	tied: $\beta = 0.8$	$\beta = 1.0$	$\beta = 1.2$
1	0.00	147	0.80	0.82	0.79	0.82	0.81	0.81
	0.01	153	0.78	0.79	0.80	0.80	0.83	0.78
	0.02	162	0.82	0.80	0.83	0.79	0.80	0.79
	0.05	186	0.80	0.81	0.81	0.80	0.82	0.79
	0.10	237	0.80	0.81	0.80	0.81	0.80	0.82
	0.20	390	0.80	0.80	0.78	0.79	0.81	0.80
1.2	0.00	147	0.82	0.80	0.83	0.80	0.80	0.79
	0.01	153	0.83	0.81	0.80	0.82	0.83	0.81
	0.02	156	0.81	0.81	0.80	0.80	0.78	0.82
	0.05	174	0.80	0.80	0.80	0.81	0.82	0.79
	0.10	204	0.82	0.78	0.80	0.81	0.78	0.79
	0.20	276	0.78	0.79	0.80	0.80	0.81	0.80
1.75	0.00	147	0.80	0.81	0.81	0.80	0.80	0.79
	0.01	147	0.80	0.79	0.79	0.80	0.80	0.79
	0.02	147	0.82	0.81	0.80	0.81	0.80	0.81
	0.05	147	0.81	0.80	0.80	0.80	0.79	0.81
	0.10	144	0.79	0.79	0.78	0.80	0.81	0.79
	0.20	129	0.80	0.78	0.81	0.78	0.81	0.79
2.5	0.00	147	0.81	0.80	0.81	0.82	0.82	0.80
	0.01	141	0.80	0.80	0.80	0.80	0.81	0.79
	0.02	135	0.80	0.81	0.82	0.80	0.78	0.84
	0.05	120	0.80	0.83	0.80	0.81	0.80	0.82
	0.10	96	0.80	0.81	0.82	0.81	0.80	0.81
	0.20	60	0.80	0.80	0.82	0.81	0.81	0.80

In the simulation, the power was found to be close to the nominal power of 80% in all scenarios studied. In the untied case we found the power to be 80.4 % on average, range 77.9 % - 83.5 %. In the tied case we observed average power 80.4 %, range 77.6 % - 83.6 %.

6 Discussion

In section 4 we derived power formulae for the Wilcoxon-Mann-Whitney test for non-inferiority for the situation were observations may be censored due to death, relying on the work of Matsouaka and Betensky (2015) for the ordinary Wilcoxon-Mann-Whitney test in the presence of death censored observations. Both, the untied situation in which time of death is taken into account for ranking and the tied situation in which all deaths are ranked identically were considered. The formulae use approximations which seem to give valid results.

We applied the derived formulae to an example from cardiology and found in this situation that there is little to no difference in power for given sample size between the tied and untied case. In the simulation we also found little difference between the tied and untied case.

As expected, power decreases with increasing probability of death – while the relative risk in the new treatment group is not too high compared to the reference group. This effect is most pronounced if the probability of death is identical in both groups under the non-inferiority null hypothesis. It decreases with increasing relative risk in the the new treatment group compared to the reference group . We also observed that sample size and power are hardly affected by censoring due to death if $\lambda_0/(\lambda_0+\lambda_1) \approx 1/2 - P(X_0 < X_1)$ where λ_0, λ_1 are the hazards – given exponential distribution of time to death – and $P(X_0 < X_1)$ the probability that the quantiative outcome tends to smaller values in the reference distribution. If the relative risk of death in the new treatment group is much higher relative to the reference group the effects are reversed when applying the formulae, i. e. power increases with increasing probability of death due to censoring. This effect could be corroborated in the simulation.

The choice of non-inferiority margin presents an additional challenge when global ranks are used to tackle censoring by death. Wellek (2010) gives some general suggestions for the choice of the equivalence or non-inferiority margin in the non-parametric situation, while Munzel and Hauschke (2003) and Zhang et al (2015) have investigated the special case of ordered categorical data. However, neither of these approaches can be transferred directly to the problem presented here.

Furthermore, Munzel (2009) has suggested to choose the non-inferiority margin such that establishing non-inferiority of a new treatment to a reference treatment would also allow to conclude that the new treatment is better than placebo. This requires either a three-armed trial including placebo in addition to the new treatment and the reference treatment or reliable data from a previous study in which the reference treatment was shown to be better than placebo. The former is – as Munzel has pointed out clearly – unethical in a situation where death is a realistic threat. The latter would require both information on the quantitative endpoint and probability of death for the no treatment option – which is not avaible in our context.

Therefore, in our example, we chose to derive the non-inferiority margin based on clinical considerations about equivalence in the outcome of the quantitative endpoint. As we could assume normality for the quantitative outcome, we used this to derive the a joint non-inferiority margin. If normality cannot be assumed other criteria would have to be used.

Using a joint non-inferiority margin for both, the probability of death and the quantitative outcome, may not always be appropriate. It is possible to use separate margins when deriving the power formula or to focus on an acceptable non-inferiority margin for the probability of death. However, the separate margins will effectively be combined into one overall margin.

In our application we delibarately considered only relatively small probabilities of censoring due to death in the reference group and small to moderate relative risks in the new treatment group compared to the reference group. If there is a substantial proportion of early deaths, mortality is likely to be a more relevant primary endpoint than any functional parameter. Also, allowing much higher risk of death as non-inferior does not seem to be a reasonable assumption for clinical trials, either.

As usual when using any combination of endpoints, there may be conflicting results, i. e. there may be an advantage of the new treatment with respect to the quantitative outcome and a disadvantage with respect to death or vice versa. This possibility needs careful consideration.

Another possibility would be to assign different weights to time to death and the quantitative outcome such as suggested by Matsouaka et al (2016). This would allow to prevent to a certain extent to take too high risks of death when the new treatment is beneficial only with respect to the quantitative endpoint.

However, our focus is on finding an appropriate way of including cases with missing quantitative outcome due to death in the analyses. In the end, clinical judgement is necessary and in a clinical trial the two outcomes have to be carefully assessed when the non-inferiority null hypothesis is rejected.

The presented approach does not allow for inclusion of covariates. If it is important to include covariates rank regression models could be applied. For such an approach the power formulae presented here will at best give a rough estimate of required power or sample size. A separate power analysis will be needed in this case.

Also, so far risk of death and underlying values of the quantitative variable have been assumed to be independent. This could be relaxed by considering joint models that take into account shared or correlated frailty.

Appendix

A.1. Untied case

In table 3 an abridged list of values used to obtain figure 1 is shown. Total sample size is given in steps of 30; only sample sizes for which at least one observed power is less than or equal to 0.95 are displayed.

A.2. Tied case

In table 4 an abridged list of values for power in the tied case is shown. Again, total sample size is given in steps of 30; only sample sizes for which at least one observed power is less than or equal to 0.95 are displayed.

References

Lachin, J. L. (1999) Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials* **20**, 408–422.

Wittes, J., Lakatos, E., and Probstfield, J. (1989) Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in Medicine* **8**, 415–425.

O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

Felker, G. M., Anstrom, K. J., Rogers, J. G. A. (2008) global ranking approach to end points in trials of mechanical circulatory support devices. *Journal of Cardiac Failure* **14**, 368–372.

Felker, G. M., Maisel, A. S. (2010) A Global Rank End Point for Clinical Trials in Acute Heart Failure. *Circulation: Heart Failure* **3**, 643–646.

Matsouaka, R. A., Betensky, R. A. (2015) Power and sample size calculations for the Wilcoxon—Mann—Whitney test in the presence of death-censored observations. *Statistics in Medicine* **34**, 406–431.

Matsouaka, R. A., Singhal, A. B.Betensky, R. A. (2016) An optimal Wilcoxon—Mann—Whitney test of mortality and a continuous outcome. *Statistical Methods in Medical Research* **epub ahead of print**, doi: 10.1177/0962280216680524.

Munzel, U. (2009)Nonparametric non-inferiority analyses in the three–arm design with active control and placebos. *Statistics in Medicine* **28**, 3643–3656.

Munzel, U., Hauschke, D. (2003) Nonparametric test for proving noninferiority in clinical trials with ordered categorical data. *Pharmaceutical Statistics* **2**, 31–37.

Wellek, S. (2010) *Testing Statistical Hypotheses of Equivalence and Noninferiority, Second Edition*. Chapman and Hall / CRC, Boca Raton.

Zhang, F., Miyaoka, E., Huang, F., Tanaka, Y. (2015) Test Statistics and Confidence Intervals to Establish Noninferiority between Treatments with Ordinal Categorical Data. *Journal of biopharmaceutical statistics* **5**, 921–938.

Table 3 Computed power as function of relative risk (RR) of death in the new treatment group (risk p_1) compared to the reference group (risk p_0), for given $\alpha = 0.025$, total sample size n_{total}, Sample size allocation for reference to new treatment 1:2. Untied case.

RR	n_{total}	$p_0 = 0$	$p_0 = 0.01$	$p_0 = 0.02$	$p_0 = 0.05$	$p_0 = 0.1$	$p_0 = 0.2$
1	30	0.258	0.247	0.236	0.207	0.167	0.111
	60	0.448	0.430	0.413	0.364	0.292	0.186
	90	0.604	0.584	0.563	0.503	0.410	0.261
	120	0.725	0.705	0.684	0.621	0.517	0.334
	150	0.813	0.795	0.776	0.717	0.610	0.405
	180	0.876	0.861	0.845	0.791	0.689	0.472
	210	0.919	0.907	0.894	0.849	0.755	0.534
	240	0.948	0.939	0.929	0.892	0.809	0.591
	270	0.967	0.960	0.952	0.923	0.852	0.642
	300	0.979	0.974	0.969	0.946	0.887	0.689
	330	0.987	0.984	0.980	0.963	0.914	0.731
	360	0.992	0.990	0.987	0.974	0.935	0.768
	390	0.995	0.994	0.992	0.982	0.951	0.801
	420	0.997	0.996	0.995	0.988	0.963	0.830
	450	0.998	0.998	0.997	0.992	0.973	0.855
1.2	30	0.258	0.249	0.241	0.219	0.189	0.146
	60	0.448	0.434	0.421	0.385	0.332	0.253
	90	0.604	0.589	0.574	0.530	0.463	0.357
	120	0.725	0.710	0.695	0.649	0.577	0.453
	150	0.813	0.800	0.786	0.744	0.673	0.541
	180	0.876	0.865	0.853	0.816	0.750	0.619
	210	0.919	0.910	0.901	0.870	0.812	0.687
	240	0.948	0.941	0.934	0.909	0.860	0.745
	270	0.967	0.962	0.957	0.938	0.897	0.793
	300	0.979	0.976	0.972	0.957	0.925	0.834
	330	0.987	0.985	0.982	0.971	0.946	0.867
	360	0.992	0.990	0.988	0.981	0.961	0.894
	390	0.995	0.994	0.993	0.987	0.972	0.917
	420	0.997	0.996	0.995	0.992	0.980	0.934
	450	0.998	0.998	0.997	0.994	0.986	0.949
1.75	30	0.258	0.257	0.256	0.255	0.257	0.279
	60	0.448	0.446	0.445	0.444	0.450	0.488
	90	0.604	0.603	0.602	0.601	0.608	0.653
	120	0.725	0.724	0.723	0.722	0.730	0.774
	150	0.813	0.812	0.812	0.812	0.818	0.856
	180	0.876	0.875	0.875	0.875	0.881	0.911
	210	0.919	0.918	0.918	0.918	0.923	0.946
	240	0.948	0.947	0.947	0.947	0.951	0.968
2.5	30	0.258	0.267	0.276	0.308	0.369	0.533
	60	0.448	0.463	0.478	0.528	0.617	0.807
	90	0.604	0.622	0.640	0.693	0.783	0.929
	120	0.725	0.742	0.759	0.809	0.882	0.976
	150	0.813	0.828	0.843	0.884	0.939	0.992
	180	0.876	0.888	0.900	0.932	0.969	0.998
	210	0.919	0.929	0.938	0.960	0.985	0.999
	240	0.948	0.955	0.962	0.978	0.993	1.000

Table 4 Computed power as function of relative risk (RR) of death in the new treatment group (risk p_1) compared to the reference group (risk p_0), for given $\alpha = 0.025$, total sample size n_{total}, Sample size allocation for reference to new treatment 1:2. Tied case.

RR	n_{total}	$p_0 = 0$	$p_0 = 0.01$	$p_0 = 0.02$	$p_0 = 0.05$	$p_0 = 0.1$	$p_0 = 0.2$
1	30	0.258	0.247	0.236	0.207	0.168	0.113
	60	0.448	0.430	0.413	0.364	0.293	0.188
	90	0.604	0.584	0.563	0.503	0.411	0.264
	120	0.725	0.705	0.684	0.621	0.517	0.338
	150	0.813	0.795	0.776	0.717	0.610	0.408
	180	0.876	0.861	0.845	0.791	0.690	0.475
	210	0.919	0.907	0.894	0.849	0.755	0.537
	240	0.948	0.939	0.929	0.892	0.809	0.594
	270	0.967	0.960	0.952	0.923	0.852	0.645
	300	0.979	0.974	0.969	0.946	0.887	0.692
	330	0.987	0.984	0.980	0.963	0.914	0.734
	360	0.992	0.990	0.987	0.974	0.935	0.770
	390	0.995	0.994	0.992	0.982	0.951	0.803
	420	0.997	0.996	0.995	0.988	0.963	0.832
	450	0.998	0.998	0.997	0.992	0.973	0.856
1.2	30	0.258	0.249	0.241	0.219	0.189	0.148
	60	0.448	0.434	0.421	0.385	0.332	0.256
	90	0.604	0.589	0.574	0.530	0.463	0.359
	120	0.725	0.710	0.695	0.649	0.577	0.456
	150	0.813	0.800	0.786	0.744	0.673	0.544
	180	0.876	0.865	0.853	0.816	0.751	0.621
	210	0.919	0.910	0.901	0.870	0.812	0.689
	240	0.948	0.941	0.934	0.910	0.860	0.746
	270	0.967	0.962	0.957	0.938	0.897	0.794
	300	0.979	0.976	0.972	0.957	0.925	0.835
	330	0.987	0.985	0.982	0.971	0.946	0.868
	360	0.992	0.990	0.988	0.981	0.961	0.895
	390	0.995	0.994	0.993	0.987	0.972	0.917
	420	0.997	0.996	0.995	0.992	0.980	0.935
	450	0.998	0.998	0.997	0.994	0.986	0.949
1.75	30	0.258	0.257	0.256	0.255	0.258	0.283
	60	0.448	0.446	0.445	0.444	0.450	0.489
	90	0.604	0.603	0.602	0.601	0.608	0.652
	120	0.725	0.724	0.723	0.722	0.730	0.772
	150	0.813	0.812	0.812	0.812	0.818	0.855
	180	0.876	0.875	0.875	0.875	0.880	0.909
	210	0.919	0.918	0.918	0.918	0.923	0.945
	240	0.948	0.947	0.947	0.947	0.951	0.967
2.5	30	0.258	0.267	0.277	0.308	0.369	0.530
	60	0.448	0.463	0.478	0.528	0.617	0.800
	90	0.604	0.622	0.640	0.693	0.782	0.924
	120	0.725	0.742	0.759	0.809	0.882	0.973
	150	0.813	0.828	0.843	0.884	0.938	0.991
	180	0.876	0.888	0.900	0.932	0.969	0.997
	210	0.919	0.929	0.938	0.960	0.985	0.999
	240	0.948	0.955	0.962	0.978	0.993	1.000

