Accurate Sample Size Calculations in Trials with Non-Proportional Hazards

James Bell – Elderbrook Solutions GmbH

PSI Conference - 4th June 2018

Sample Size Under NPH — Introduction

- Non-proportional hazards (NPH) are common in time-to-event (TTE) trials
  - E.g. heterogeneous populations, ‘cures’, immuno-oncology

- Cox and Log-Rank Test very common analysis even under NPH
  - HR meaningful if viewed as a weighted average over time
  - Both methods powerful under NPH (and often required by FDA/EMA…)

- RMST and landmark analyses are being increasingly investigated as alternatives
  - Other methods are available… (e.g. weighted log-rank test, average HR)

- However there are issues with sample size calculations:
  - Those for Cox/Log-Rank Test assume PH
  - Those for RMST/Landmark methods struggle with censoring
    - Simulations typically recommended….

- Here, accurate analytical methods for NPH planning are presented

Sample Size Under NPH

Log Rank Test

- Log Rank Test is the Score Test for a basic Cox Model
  - Sample size planning for one works for the other

- Power is typically calculated using the Schoenfeld Formula*:

$$Events = \frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{P_1 P_2 \log(\theta)^2}$$

- However, under NPH, we do not know θ (the HR).
  - Formula also derived under PH: Can it still be used?

- Hard to derive θ directly due to 'dynamic' event-driven weighting scheme

- Instead we use an indirect LRT-based method…

*Schoenfeld D, *Biometrics* 1983, **39(2)**:499-503

Sample Size Under NPH — Pike and Peto

- Log-Rank Formula:

$$Z_{LRT} = \frac{O_1 - E_1}{\sqrt{V}}$$

- Two literature-reported methods* use LRT-derived quantities to estimate θ:

$$\ln(\hat{\theta}_{peto}) = \frac{O_1 - E_1}{V}, \qquad Var(\ln(\hat{\theta}_{peto})) = 1/V$$

$$\hat{\theta}_{pike} = \frac{O_1 E_2}{O_2 E_1}, \qquad Var(\ln(\hat{\theta}_{pike})) = \frac{1}{E_1} + \frac{1}{E_2}$$

- Pike method reported** as more accurate, but conservative → Chosen method
  - Reasonable for $1/3 < \theta < 3$

- Expectations are calculable for all components

*Peto R, Peto J, *J R Stat Soc A Ser A-G*. 1972, **135**: 185-198

Berry G, Kitchin R, Mock P, *Stat Med*. 1991, **10: 749-755.

Sample Size Under NPH — Expectations

$$\boldsymbol{O_j = \sum d_{ij}} \qquad\qquad \boldsymbol{E_j = \sum \frac{n_{ij} d_i}{n_i}}$$

- To calculate expectations, we consider the distribution functions w.r.t. time:
  - Assuming independence of events, dropout (dr) and administrative censoring (c):

$$d_j(t) = N_j \left(1 - F_{dr,j}(t)\right)\left(1 - F_{c,j}(t)\right) f_j(t) \qquad = N_j C_j^-(t) f_j(t)$$

$$n_j(t) = N_j \left(1 - F_{dr,j}(t)\right)\left(1 - F_{c,j}(t)\right)\left(1 - F_j(t)\right) = N_j C_j^-(t) S_j(t)$$

- Therefore:

$$\boldsymbol{E}\left(O_j\right) = \int_0^T d_j(t)\, dt$$

$$\boldsymbol{E}\left(E_j\right) = \int_0^T \frac{n_j(t)\left(d_1(t) + d_2(t)\right)}{n_1(t) + n_2(t)} dt$$

$$\theta_{pike} = \frac{O_1 E_2}{O_2 E_1}, \qquad O = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{P_1 P_2 \log(\theta)^2}$$

- → Everything needed to predict HR and hence power

Sample Size Under NPH   RMST

- Royston & Parmar provided formulae for RMST sample size planning*:
  - $\mu_j = \mathbf{E}(RMST_j) = \int_0^R S_j(t)d\boldsymbol{t}$
  - $\mathbf{V}(RMST_j) = 2\int_0^R t\, S_j(t)dt - \left\{\int_0^R S_j(t)dt\right\}^2$
  - $\boldsymbol{SE}(\hat{\mu}_j) = \sqrt{\frac{\varphi^2 \mathrm{V}(\mathrm{RMST}_j)}{N_j}}$
- However, φ^2 is censoring/recruitment dependent, (1 if no censoring, increasing with censoring).
  - No direct estimation method provided (Suggested to back-estimate from existing trial data).
- Note that $\mathbf{V}(RMST_j)$is an intrinsic property of event distribution
  - Independent censoring does not affect KM plot, only the number at risk
- We therefore need to replace N_j by an **effective sample size**

*Royston, Parmar *BMC Medical Research Methodology* 2013, **13**:152

Sample Size Under NPH RMST

- On day 1, effective sample size is N_j but decreases over time due to censoring

- The overall effective sample size can be viewed as a **weighted average** of the changing **effective sample size over time**, using the **point variance function** as the weighting.

- We therefore derive:
  - Variance contribution at time x (by differentiation): $\mathrm{dV(x)} = 2S(x)\left(x - \int_0^x S(t)\,dt\right)$
  - Effective sample size at time x: $N_{eff}(x) = \frac{N\int_0^x C^{-}(t)f(t)dt}{F(x)}$

- Then:

$$N_{eff} = \frac{2N}{V(RMST)}\int_0^R \frac{S(x)\left(x-\int_0^x S(t)dt\right)\int_0^x C^{-}(t)f(t)dt}{F(x)}dx \quad \text{and} \quad \boldsymbol{SE}(\mu_j) = \sqrt{\frac{\mathrm{V(RMST}_j)}{N_{eff,j}}}$$

- Following Royston et al., sample size may then be calculated using standard approaches

Sample Size Under NPH **Landmark**

- Landmark analysis is typically performed using a normal approximation and **Greenwood's formula*** to calculate variance:

$$\boldsymbol{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{i:\, t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- We can again calculate expectations based upon distribution functions, similar to O and E
  - $(n_i - d_i) \xrightarrow[\delta t \to 0]{} n(t)$ since the 'point' number of events tends to 0

$$\boldsymbol{V}(S(t)) = \frac{S(t)^2}{N} \int_0^T \frac{f(t)}{S(t)^2 C^{-}(t)} dt$$

  - Note: This also corresponds more directly to Tsiatis' formula**

- Standard normal-approximation based methods may then be applied

*Greenwood M *Reports on Public Health and Medical Subjects*. 1926, **33**: 1–26.
** Tsiatis A *Annals of Statistic* 1981 **9**, 93-108

Sample Size Under NPH GESTATE

- These are complex integrals where any distributions could be specified. **Problems!**
  - Most integrals **not analytically-solvable**
  - Don’t want to limit distribution choice; particular issue for NPH
    - **Combinatorics** become prohibitive with even a handful of distribution types

- Solutions:
  - **Numerical integration**; most relevant integrals evaluable
  - **Generic formula coding**; no distribution-specific code
  - **Object-oriented programming**; distributions are from interchangeable **Curve** objects
  - **Self-writing code**; integration functions written at run-time

- To implement this, an R package has been written: GEneralised Survival Trial Assessment Tool Environment (**GESTATE**).
- Core code can handle any ‘well-behaved’ distribution, or combination of distributions
  - Adding new distributions is straightforward

Sample Size Under NPH Simulation

- **Curve architecture** also allows for a **generalised simulation approach**:
  - Wide variety of event, censoring and recruitment distributions supported
  - **Shared inputs/syntax** with analytic approach – simple to validate
  - Note: still relies on independent censoring

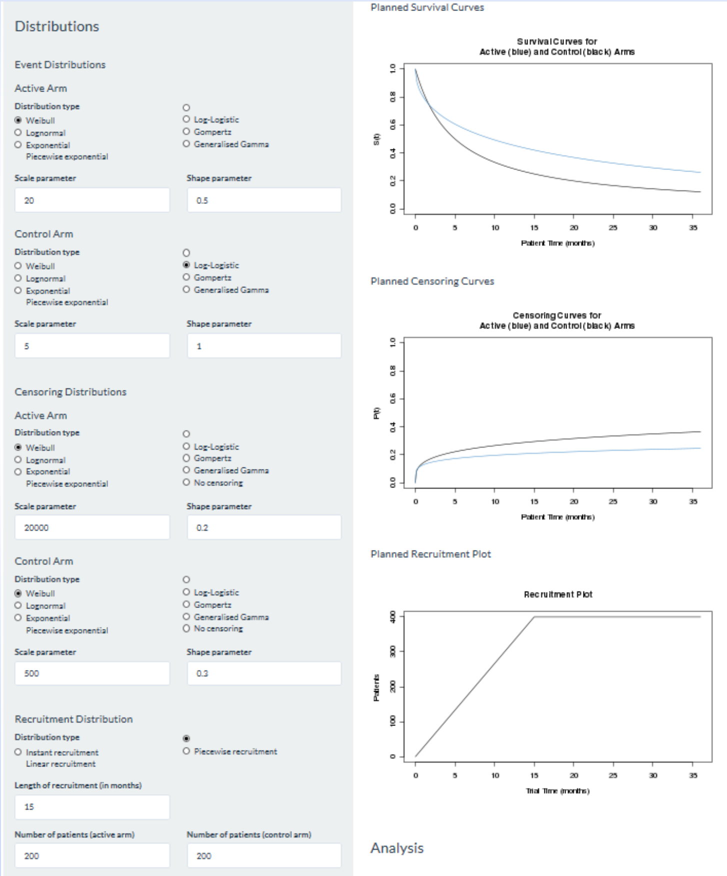
- Designed to be straightforward to use
  - Automatic analysis and summary functions covering each analysis method
  - Parallel processing options included for speed

Sample Size Under NPH

R Shiny UI

- **Interactive R Shiny UI** written
    - Real-time plots of S(t), censoring CDF and recruitment input distributions
    - Analytic and simulation approaches run through same interface
    - Exportable outputs

- Inputs for an example are displayed:
    - **Weibull** active event curve,
    - **Log-logistic** control event curve
    - Differential **Weibull** censoring between arms.
- Analysis performed after 36 months
    - Restriction time: 30 months
    - Landmark analysis: 30 months
- 20,000 simulations performed



Sample Size Under NPH Example

Simulation Summary:

log(HR)	HR	Log(HR) SE	Cox Z-value	Cox P-Value	Log-Rank Z-Value	Log-Rank P-Value	Observed Events (Active)	Observed Events (Control)	Observed Events (Total)	Failed (Log-Rank)	Log-Rank Power	Simulations	Mean Assessment Time	RMST (Active)	RMST (Active) SE	RMST (Control)	RMST (Control) SE	RMST Difference	RMST Difference SE	Failed (RMST)	RMST Power	Landmark Survival (Active)	Landmark Survival (Active) SE	Landmark Survival (Control)	Landmark Survival (Control) SE	Landmark Survival Difference	Landmark Survival Delta SE	Failed (Landmark)	Landmark Power
-0.3691	0.6914	0.12793	-2.8787	0.002	-2.8993	0.0019	117.0942	134.6266	251.721	0	0.8282	20000	36	13.8469	0.9493	9.7385	0.827	4.1084	1.259	0	0.9016	0.2934	0.0382	0.1433	0.0315	0.1501	0.0496	0	0.8526

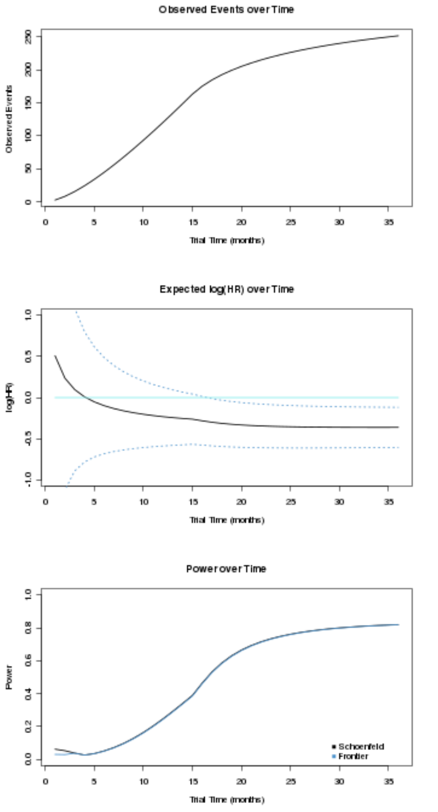
Analytic Properties:

Values

Below is a table with estimated values for several key trial parameters. Note that the estimated required SS column provides an estimate at each assessment time of the sample size required to reach the pre-specified power if all parameters other than patient numbers are kept the same (based on Schoenfeld). This sample size should be used only as a tool to guide future runs of GESTATE and should not be reported and/or used in CTPs.

Assessment Time	Patients Recruited	Events (Control)	Events (Active)	Events (Total)	Hazard Ratio	Log (HR)	Log (HR) SE	Power (Schoenfeld)	Power (Frontier)	Estimated Required SS	RMST (Control)	RMST (Active)	RMST Delta	RMST SE	RMST Power	RMST Failure	Landmark (Control)	Landmark (Active)	Landmark Delta	Greenwood Delta SE	Landmark Power
1	27	1.065	1.681	2.747	1.6642	0.5093	1.225	0.062	0.0303	23596	0.9116	0.8627	-0.0489	0.1282	0.0572	1	NA	NA	NA	NA	NA
2	53	3.75	4.545	8.295	1.2676	0.2371	0.696	0.0528	0.0292	36053	1.6824	1.6244	-0.058	0.2159	0.0454	1	NA	NA	NA	NA	NA
3	80	7.573	8.081	15.654	1.1024	0.0975	0.5055	0.0386	0.0376	113023	2.35	2.3272	-0.0229	0.2876	0.03	1	NA	NA	NA	NA	NA
4	107	12.244	12.114	24.357	1.0093	0.0092	0.4051	0.0264	0.0261	8127360	2.9389	2.9856	0.0467	0.3485	0.0339	1	NA	NA	NA	NA	NA
5	133	17.57	16.546	34.116	0.9489	-0.0524	0.3423	0.0354	0.0343	179197	3.4657	3.6082	0.1424	0.4015	0.0542	1	NA	NA	NA	NA	NA
6	160	23.417	21.315	44.732	0.9065	-0.0982	0.299	0.0514	0.0499	39013	3.9423	4.2002	0.2579	0.4484	0.0831	1	NA	NA	NA	NA	NA
7	187	29.689	26.373	56.063	0.8751	-0.1334	0.2671	0.0721	0.0701	16845	4.3773	4.7658	0.3885	0.4905	0.1214	1	NA	NA	NA	NA	NA
8	213	36.313	31.687	68	0.8509	-0.1615	0.2425	0.0978	0.0954	9482	4.7776	5.308	0.5304	0.5286	0.1694	1	NA	NA	NA	NA	NA
9	240	43.234	37.228	80.461	0.8317	-0.1843	0.2229	0.1285	0.1257	6153	5.1481	5.8291	0.681	0.5634	0.2263	1	NA	NA	NA	NA	NA
10	267	50.407	42.974	93.381	0.8162	-0.2031	0.2069	0.1639	0.1608	4363	5.4931	6.3312	0.8381	0.5954	0.2904	1	NA	NA	NA	NA	NA
11	293	57.798	48.907	106.706	0.8033	-0.219	0.1935	0.2036	0.2002	3285	5.8158	6.8158	1	0.625	0.3594	1	NA	NA	NA	NA	NA
12	320	65.38	55.012	120.392	0.7926	-0.2325	0.1821	0.2468	0.2432	2585	6.1189	7.2843	1.1654	0.6526	0.4309	1	NA	NA	NA	NA	NA
13	347	73.128	61.275	134.403	0.7835	-0.244	0.1723	0.2927	0.289	2101	6.4047	7.7379	1.3332	0.6783	0.5022	1	NA	NA	NA	NA	NA
14	373	81.023	67.685	148.708	0.7757	-0.254	0.1638	0.3405	0.3367	1753	6.675	8.1777	1.5027	0.7025	0.5711	1	NA	NA	NA	NA	NA
15	400	89.05	74.232	163.281	0.769	-0.2627	0.1562	0.3892	0.3855	1492	6.9315	8.6045	1.673	0.7253	0.6356	1	NA	NA	NA	NA	NA
16	400	96.129	79.225	175.354	0.7539	-0.2825	0.1507	0.4644	0.4606	1202	7.1754	9.0192	1.8437	0.7534	0.687	1	NA	NA	NA	NA	NA
17	400	101.695	83.155	184.85	0.7406	-0.3003	0.1468	0.5324	0.5286	1009	7.408	9.4224	2.0144	0.7861	0.7265	1	NA	NA	NA	NA	NA
18	400	106.218	86.527	192.745	0.7304	-0.3141	0.1436	0.5873	0.5837	884	7.6303	9.8148	2.1846	0.8211	0.7582	1	NA	NA	NA	NA	NA
19	400	109.983	89.507	199.49	0.7226	-0.3249	0.1411	0.631	0.6276	799	7.8431	10.1971	2.354	0.8572	0.7841	1	NA	NA	NA	NA	NA
20	400	113.169	92.188	205.357	0.7166	-0.3333	0.139	0.6657	0.6626	737	8.0472	10.5696	2.5225	0.8941	0.8055	1	NA	NA	NA	NA	NA
21	400	115.907	94.63	210.537	0.7119	-0.3399	0.1372	0.6935	0.6907	692	8.2433	10.933	2.6897	0.9313	0.8233	1	NA	NA	NA	NA	NA
22	400	118.289	96.871	215.16	0.7081	-0.3451	0.1356	0.7161	0.7135	657	8.432	11.2876	2.8556	0.9687	0.8384	1	NA	NA	NA	NA	NA
23	400	120.38	98.945	219.324	0.7052	-0.3493	0.1343	0.7344	0.7321	629	8.6138	11.6338	3.02	1.0062	0.8511	1	NA	NA	NA	NA	NA
24	400	122.232	100.872	223.104	0.7029	-0.3526	0.133	0.7495	0.7474	607	8.7893	11.9721	3.1828	1.0437	0.8621	1	NA	NA	NA	NA	NA
25	400	123.884	102.674	226.558	0.7011	-0.3552	0.1319	0.762	0.7601	589	8.9588	12.3027	3.3439	1.081	0.8715	1	NA	NA	NA	NA	NA
26	400	125.367	104.361	229.728	0.6996	-0.3573	0.131	0.7726	0.7708	574	9.1227	12.626	3.5033	1.1182	0.8796	1	NA	NA	NA	NA	NA
27	400	126.707	105.951	232.658	0.6985	-0.3589	0.13	0.7814	0.7797	562	9.2815	12.9424	3.6609	1.1552	0.8867	1	NA	NA	NA	NA	NA
28	400	127.923	107.452	235.374	0.6976	-0.3601	0.1292	0.7889	0.7873	551	9.4353	13.2519	3.8166	1.192	0.8929	1	NA	NA	NA	NA	NA
29	400	129.031	108.869	237.9	0.6969	-0.3611	0.1285	0.7953	0.7938	542	9.5846	13.555	3.9704	1.2286	0.8983	1	NA	NA	NA	NA	NA
30	400	130.046	110.214	240.26	0.6965	-0.3618	0.1278	0.8006	0.7993	535	9.7296	13.8519	4.1223	1.2649	0.903	1	NA	NA	NA	NA	NA
31	400	130.979	111.498	242.477	0.6961	-0.3622	0.1271	0.8051	0.8039	529	9.7296	13.8519	4.1223	1.2623	0.9042	0.3237	0.1429	0.2938	0.151	0.0632	0.6654
32	400	131.838	112.709	244.548	0.6958	-0.3627	0.1265	0.8094	0.8083	523	9.7296	13.8519	4.1223	1.2602	0.9051	0.0848	0.1429	0.2938	0.151	0.0578	0.7426
33	400	132.634	113.87	246.504	0.6957	-0.3628	0.126	0.8128	0.8118	519	9.7296	13.8519	4.1223	1.2585	0.9058	0.0236	0.1429	0.2938	0.151	0.0547	0.7888
34	400	133.372	114.979	248.351	0.6957	-0.3629	0.1255	0.8158	0.8148	515	9.7296	13.8519	4.1223	1.2571	0.9065	0.0066	0.1429	0.2938	0.151	0.0525	0.82
35	400	134.058	116.041	250.099	0.6957	-0.3628	0.125	0.8183	0.8174	511	9.7296	13.8519	4.1223	1.256	0.907	0.0018	0.1429	0.2938	0.151	0.0509	0.8425
36	400	134.698	117.055	251.753	0.6958	-0.3627	0.1245	0.8206	0.8198	508	9.7296	13.8519	4.1223	1.2551	0.9073	5e-04	0.1429	0.2938	0.151	0.0497	0.8591

Plots



Sample Size Under NPH Example

	Events	HR	Log-Rank Power	RMST 30 Δ (months)	RMST 30 Δ SE (months)	RMST Power	Δ S(30)	Δ S(30) SE	Δ S(30) Power
Simulation	251.72	0.691	82.8%	4.108	1.259	90.2%	0.150	0.0496	85.3%
Analytic	251.75	0.696	82.0%	4.122	1.255	90.7%	0.151	0.0497	85.9%

- **Properties are a good match between simulation and analytic approaches**
- HR of 0.69 accurately predicted for this time of assessment (36 months)
  - For comparison, HR at 12 months predicted to be 0.79
- RMST and landmark standard errors accurately calculated, despite high censoring:
  - Overall, 37% censoring
  - Without any dropout, 308.6 events predicted (18.5% events censored due to dropout)
  - For comparison, binomial-derived SE would be 0.0406
- Powers are close, although in 2-3σ range for Monte Carlo error

Sample Size Under NPH Discussion

- Methods work well for most cases, but have a few limitations:

  - HR calculation becomes less accurate as planned HR moves further from 1
    - Limitation of the Pike approximation

  - Power calculation becomes less accurate for all three methods the more extreme the non-proportional hazards
    - The normal assumptions start breaking down:
      - Variance becomes correlated with point estimate
      - Properties may be predicted well, but power less so

- For RMST it is important to calculate the probability of analysis 'failure' (due to undefined analysis curve at point of restriction)
  - Also implemented analytically, but not shown here

Summary

- Accurate numerical-integration methods for prediction of many time-to-event trial properties under Non-Proportional Hazards have been presented

- Prediction of the Cox hazard ratio at a given assessment time is demonstrated

- Direct, analytic power calculations under censoring are presented for RMST and Landmark analyses

- The GESTATE R package has been written to implement these methods in a uniquely flexible fashion, allowing for simple input of complex assumption combinations
    - It also separately includes simulation functionality

- The GESTATE package should be publically available later in 2018.

Acknowledgements

- **Boehringer Ingelheim** for the ongoing collaboration on GESTATE
- **Jasmin Ruehl** for creating the R Shiny GUI, testing, and implementation of additional event distributions
- **Oliver Sailer** for detailed testing and feedback