Introduction to Clinical Trials

Lecture 9: Data Analysis Approaches (continued)

Giorgos Bakoyannis, PhD
Associate Professor
Department of Biostatistics and Health Data Science
Indiana University Indianapolis

Notes by Prof. Constantin T. Yiannoutsos, City University of New York

Section 1

Time to event studies

Cross-over and non-compliance in time-to-event studies

Cross-over from or to active treatment at toxicity or disease progression may lead to statistical challenges in the analysis of overall survival because crossover leads to information loss and dilution of comparative clinical efficacy.

Cross-over (as well as other forms of non-compliance) has potentially significant implications for effect estimates on survival (e.g., hazard ratios).

The following follows the article by Jönsson and colleagues (Value in Health, 2014), where various methods to account for non-compliance and cross-over are reviewed.

The overall conclusion is that the results of the statistical analysis are potentially very different depending on the method used.

Case study: Sunitimib

We will describe two studies of the drug sunitinib (Sutent; Pfizer, Inc., New York, NY), an orally administered, multitargeted tyrosine kinase inhibitor.

We described here two trials of sunitinib:

1. The study by Motzer et al., (NEJM,) of sunitinib for the treatment of metastatic renal cell carcinoma (mRCC)
2. The study by Demetri et al., (Lancet,) of sunitinib for imatinib-resistant gastrointestinal stromal tumors (GIST)

The issue that complicates routine ITT analyses in both studies is the structured cross-over from the non-sunitinib arm (interferon-alpha and placebo respectively) to sunitinib anticipated by the trial design.

Analytical strategies in the sunitinib trias

There are four analytical strategies that were considered in this study:

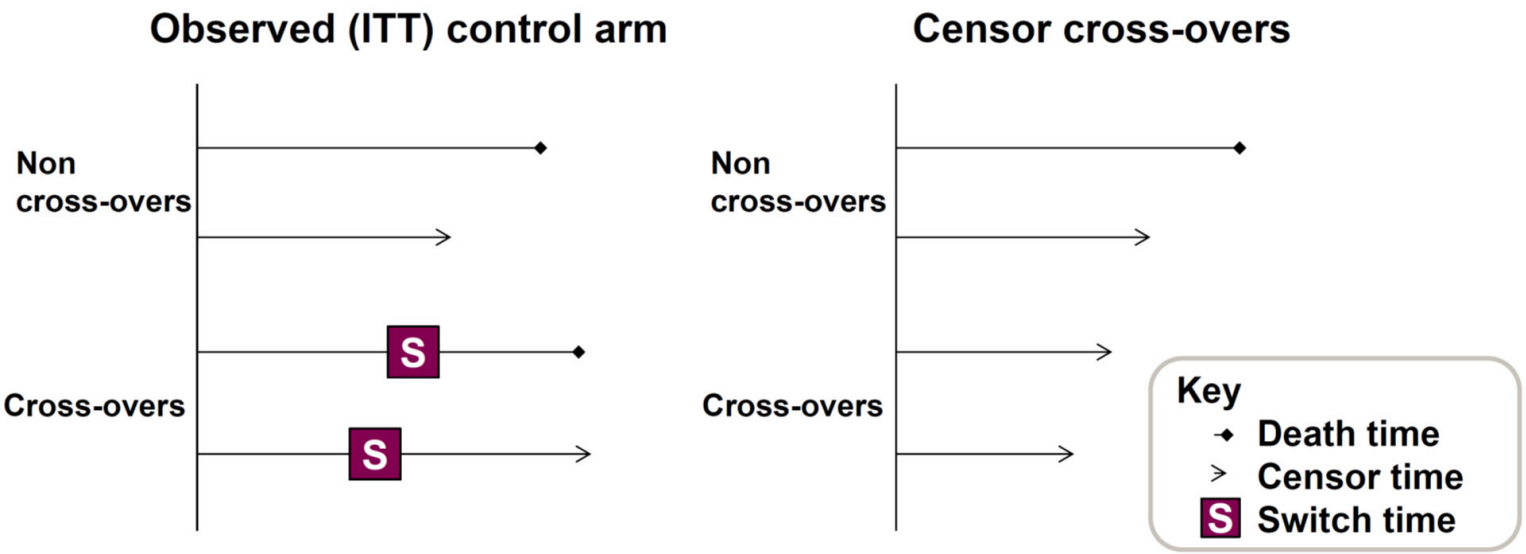
- The intent-to-treat (ITT) or as-randomized approach
- Censoring subjects at the time of cross-over (on-treatment) analysis
- Inverse probability of censoring weights (IPCW) modeling
- Rank-preserving structural failure time (RPSFT) model.

The ITT and on-treatment analyses

The ITT analysis considers everyone's outcome within their randomly assigned treatment. Based on what we saw earlier, the ITT analysis will likely underestimate the treatment effect in the presence of crossover.

The on-treatment approach censors the time of a patient at their cross-over time. The two methods are shown graphically in Figure 1.

Figure: ITT versus on-treatment analyses



The IPCW model[1]

In the Inverse probability of censoring weights (IPCW), the model addresses the counterfactual question of what would be the treatment effect in the absence of cross-over. Similarly, it can answer the question with "cross-over" replaced by "non-compliance".

In this method, patients who cross over (or exhibit non-compliance) are censored, while patients remaining in their randomized arm (or continuing to exhibit compliance with their treatment) are weighted to compensate for missing data.

The weights are determined by the predicted probability of not being censored at a given time. Then, a survival analysis is carried out with weights made up of the inverse of the probability of remaining uncensored based on each patient's profile (measured covariates both at baseline and obtained over time during the study).

[1] Robins & Finkelstein, 2000.

Constructing the weights

The IPCW weights are constructed in two steps:

1. Step 1 Calculate the probability of cross-over at each time point based on baseline characteristics only; these are the: *numerator* weights
2. Step 2 Calculate the probability of cross-over at each time point based on both baseline and time-updated characteristics: *denominator* weights

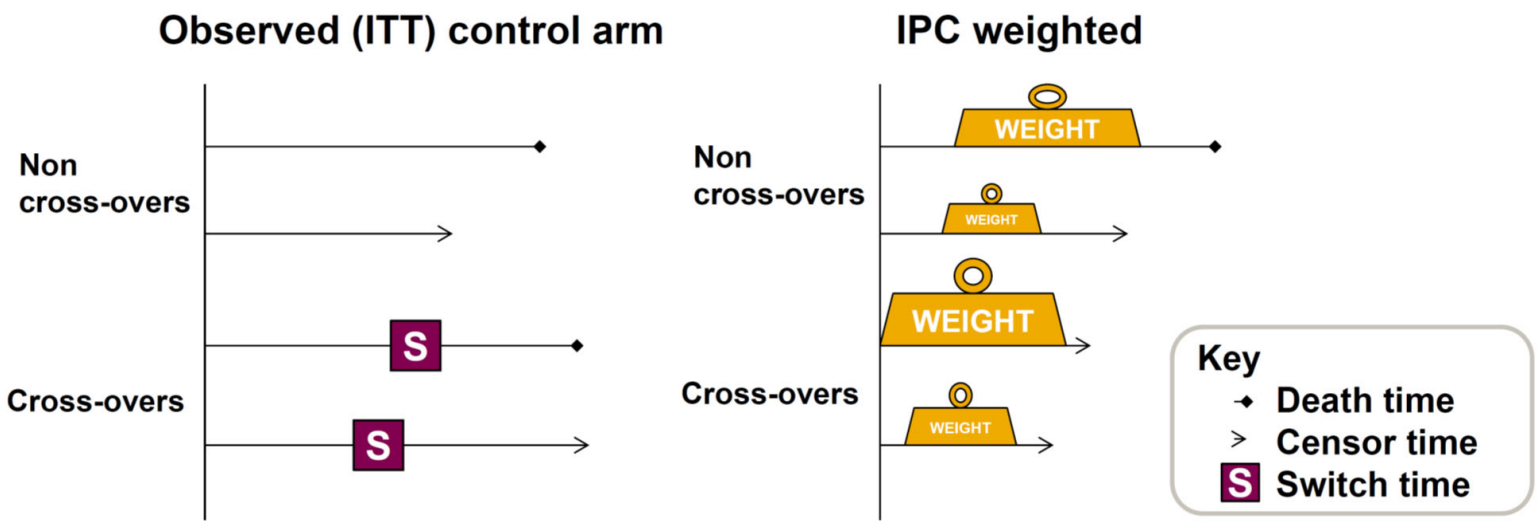
IPCW (stabilized) weights are calculated as the ratio of the numerator over the denominator weights.

A weighted analysis of the usual survival model (e.g., Cox, parametric models, etc.) is run. However, the interpretation concerns the (possibly counterfactual) *policy* of no cross over.

Schematic representation of the IPCW method

Overall, patients who belong to groups with high cross-over rate will get higher weights (they will "represent" more patients who crossed over) while patients in groups with low cross-over rate will get lower weights (they will represent a smaller number of patients who crossed over). This is shown schematically in Figure 2.

Figure: Schematic representation of the IPCW procedure



The RPSFT model

The Rank-preserving structural failure time (RPSFT) model allows a direct comparison of the two (or more) randomization groups by adjusting the overall survival of patients who cross over to reflect the survival they would have had if they never received the experimental treatment.

In the RPSFT model, the observed failure time T_i for each subject i is associated with the counterfactual time U_i that would have been observed had the subject not crossed over and received the active drug. T_i and U_i are related through the treatment history $Q_i(u)$, $u \in (0, T_i)$ as follows:

$$U_i = \int_0^{T_i} \exp\{\psi Q_i(u)\}\, du$$

If switching occurs at discrete time points, this reduces to the following sum:

$$U_i = \begin{cases} T_i & \text{non} - \text{cross} - \text{overs} \\ T_i^{\text{c}} + \exp(\psi) T_i^{e} & \text{cross} - \text{overs} \end{cases}$$

T_i^{e} and T_i^{c} are the times spent in the experimental and control arms respectively and note that $T_i = T_i^c + T_i^e$.

The RPSFT model as an accelerated failure-time model

In the RPSFT model, the coefficient ψ *accelerates* the consumption of the survival time by a factor e^{ψ}. When $\psi < 0$ the untreated survival time U_i is less than T_i, the observed survival, and the treatment is beneficial; otherwise the treatment is detrimental.

The RPSFT model is related to the so-called "accelerated failure time models" in survival analysis where risk factors act multiplicatively on the time scale (i.e., by accelerating or decelerating the time until the occurrence of the failure).

The RPSFT model: An example

To understand how this works, consider a study in smoking cessation[2] resulting in $\psi = -0.1$ and an observed lifetime for a subject $T_i = 2.2$ years, given smoking history 0.2 years on, one year off, one year on. The survival if always smoked is

$$U_i = (2.2 - 2)e^{-0.1} + (2 - 1)e^{0} + (1)e^{-0.1} = 1 + 1.2e^{-0.1} \approx 2.09$$

This means that, had the subject smoked for the entire period, their survival would have been $U_i = 2.09$ years.

However, the one-year smoking cessation added about 0.11 years of life to this subject's survival, since the corresponding observed survival time was $T_i = 2.2$ years.

[2] Mark & Robins, 1993.

Core idea of the RPSFT model

The main idea of the RPSFT model is that each patient has an inherent failure time U_i, which, because of randomization, is independent of the treatment assignment R_i.

In other words, people with longer or shorter survival do not end up preferentially in one or the other treatment arm; randomization guarantees this (at least in expectation).

In addition, the decision to cross over from the control to the experimental arm is assumed to be independent of the true failure time if unexposed (in our case if not crossing over) U_i.

Note that this does not mean that the decision of crossing over is independent of T_i, the observed failure time.

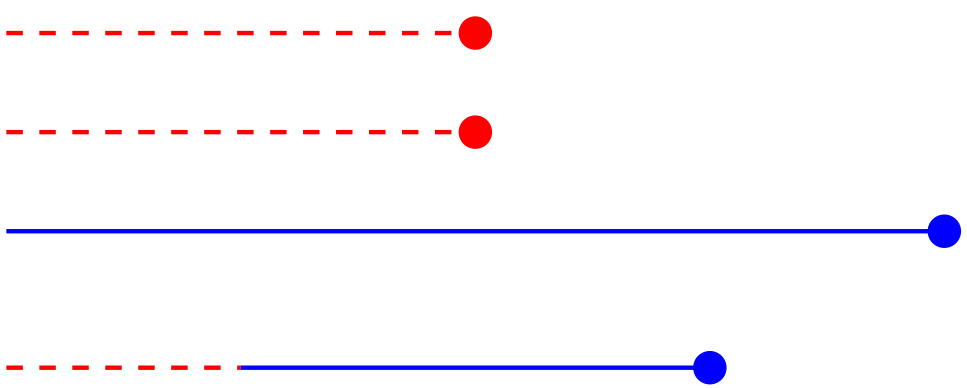
Interpretation of ψ

From the previous discussion, the coefficient ψ is related to the exposed and unexposed time by the following equation

$$e^{\psi} = \frac{T^c}{T^e} \Rightarrow e^{\psi} T^e = T^c$$

That is, $\psi < 0$ results in a longer survival if treated compared to untreated (treatment beneficial).

For example, if $e^{\psi} = 0.5$ (i.e., $\psi \approx -0.69$), this means that one year unexposed/untreated equals two years exposed/treated.



U_i for subject i

Untreated, life = 1 year

Always treated, life = 2 years

Untreated 6 months
then switch, life = 1.5 years

Construction of the RPSFT model

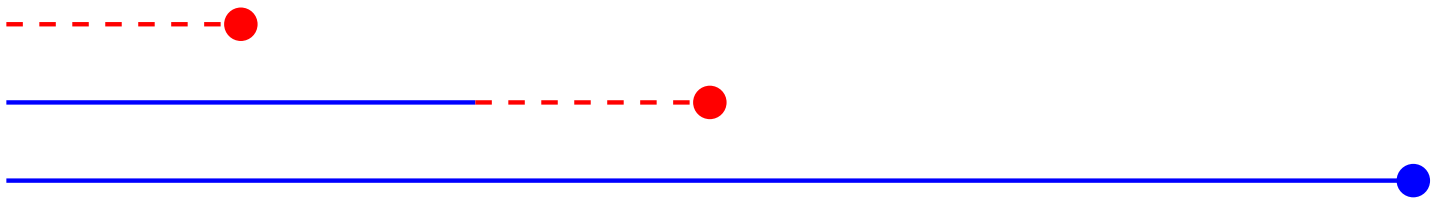
The RPSFT model is constructed in the following steps:

1. Define a model relating the observed event time T_i to the unobserved event time U_i *that would have been observed if crossover had not occurred*.
2. Compute U_i for a range of possible values of ψ (which includes all relevant confounders) and find the one for which a statistical test of the equality of U_i across the two groups has the highest (least significant) p value. This ensures that the U_i are independent of R_i.

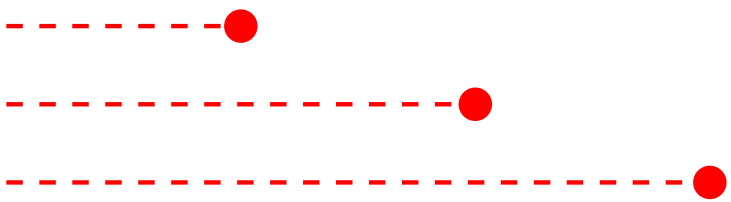
Hypothetical data example[4]: Switches only from active arm.

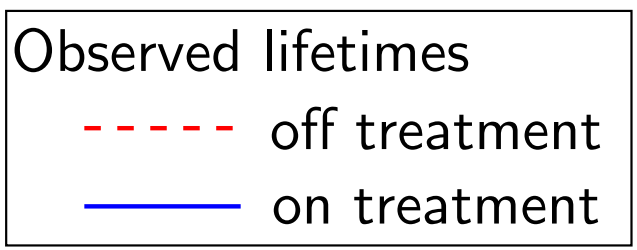
Treated and untreated subjects have equal U_i.[3]

Treated arm:



Control arm:



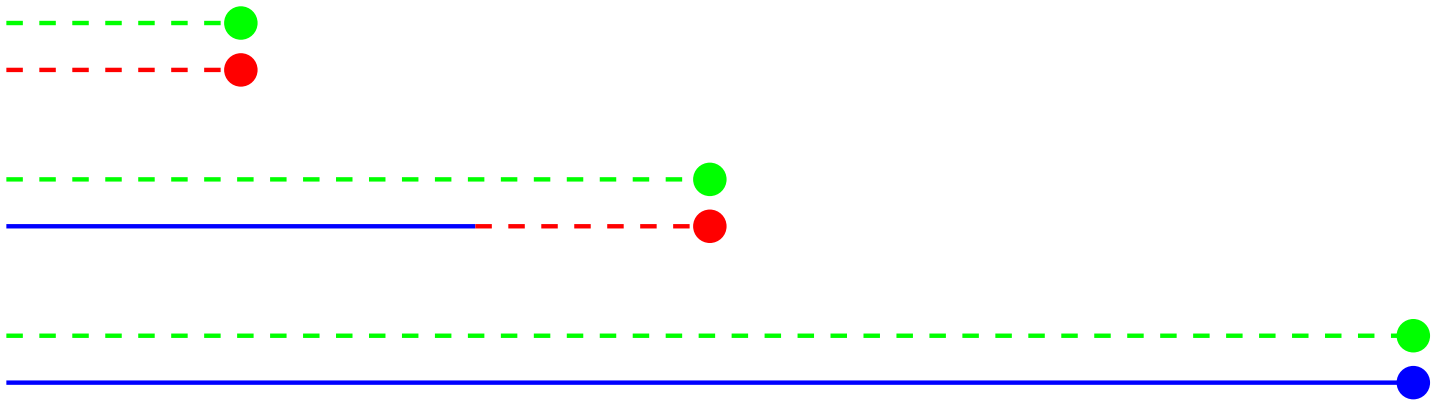


[3] Note that this is due to the randomization.

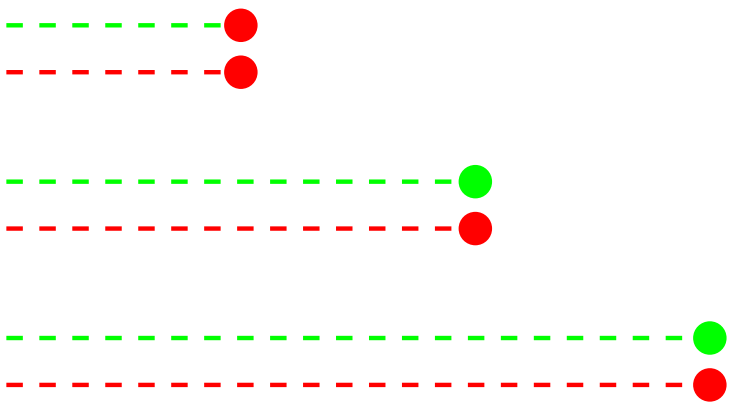
[4] Closely following the presentation by Ian White, HTMR network workshop on Methods for adjusting for treatment switches in late-stage cancer trials. London, 20th February 2012

Hypothetical data example: $e^{\psi} = 1$

Treated arm:



Control arm:



Observed lifetimes

off treatment

on treatment

Fitted untreated lifetimes

Comments

The estimate of ψ resulting in $e^{\psi} = 1$ (i.e., $\hat{\psi} = 0$) above, is not a good estimate for ψ since the estimated untreated survival times are not equal between the two groups!

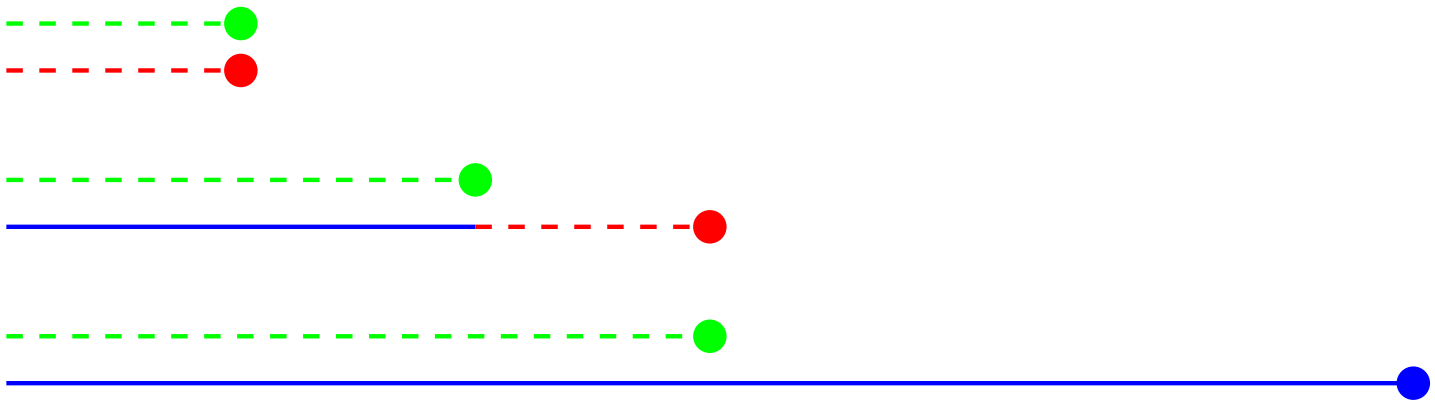
Note that, because of randomization, the estimated untreated survival times must be equal in the two groups. Since the estimate $\hat{\psi} = 0$ resulted in unequal survival times between the two groups, this estimate does not reflect the data.

Note also that $\hat{\psi} = 0$ is equivalent to no treatment effect since then $e^{\psi} = \frac{T^c}{T^e} = 1$. In other words, the survival under exposure would be equal to the survival under non-exposure.

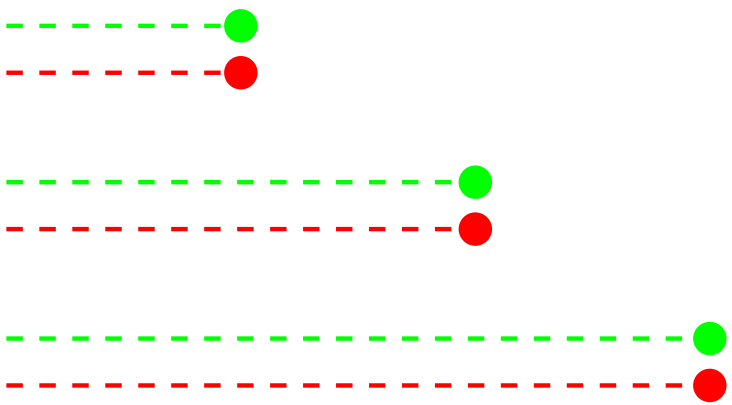
In the above situation, the assumption of no treatment effect is incongruous with what we see in practice, i.e., generally longer observed survival times among the treated patients compared to the untreated patients.

Hypothetical data example: $e^{\psi} = 0.5$

Treated arm:



Control arm:



Thus, $e^{\psi} = 0.5$ (i.e., $\psi \approx -0.69$) is a good estimate for ψ since the estimated untreated survival times are balanced between the two groups.

Recensoring

Censoring introduces unexpected complications into the RSFMT model.

This is because, if there is a beneficial treatment effect (which extends survival), then failure times in the treated group will be more likely to be censored. Thus, censoring is informative and excluding censored observations will lead to attenuation bias of the estimate of the treatment effect[5].

Mathematically, censoring implies that, instead of the failure time T_i we observe $X_i = \min(T_i, C_i)$ where C_i is the censoring time, which runs from randomization to the common closure of the study.

Unfortunately, replacing U_i with X_i in the calculations will not work unless the null hypothesis is true (i.e., if $\psi = 0$) because if non-compliance is non-random, X_i and R_i are not independent from each other.

[5] The exact opposite happens when a treatment or exposure is detrimental.

Dealing with censoring in the RPSFT model

The core argument for developing a procedure to deal with censoring is to understand that, just as with U_i, C_i (the maximum follow-up time) is (at least in theory) known at randomization and is thus independent of R_i the randomly assigned treatment.

Thus, any function of U_i and C_i will be independent of R_i as well. We define such a function as $X_i(\psi) = \min(U_i(\psi), C_i(\psi))$ as follows:

$$C_i(\psi) = \begin{cases} C_i, & \text{if} \psi \geq 0 \\ C_i \exp(\psi) & \text{if} \psi < 0 \end{cases}$$

where we have made the dependence on ψ explicit. This is called *recensoring* the data. The new censoring indicator is

$$\Delta_i(\psi) = I\left\{C_i(\psi) < U_i(\psi)\right\}$$

We say then that the ith individual is ψ-censored.

Statistical implementation of the RPSFT model

Using the fundamental equation

$$U_i = \int_0^{T_i} \exp\left[\psi Q_i(u)\right] du$$

we can estimate $U_i(\psi)$ for a specific value of ψ.

Then, considering these $U_i(\psi)$ as the true failure times, we carry out a statistical test (e.g., a log-rank test) of the treatment arms.

The value ψ_0 we seek is the solution satisfying the equation[6]

$$\Pr(U_i(\psi_0) \geq t | R_i = 1) = \Pr(U_i(\psi_0) \geq t | R_i = 0), \qquad t \geq 0$$

Thus, the desired value of ψ_0 is that which results in the most non-significant log-rank (or other statsitical) test.

[6] Note that this is the definition of independence between U_i and R_i.

Practical implementation of the RPSFT model

In practice, we carry a grid search to find this value of ψ_0.

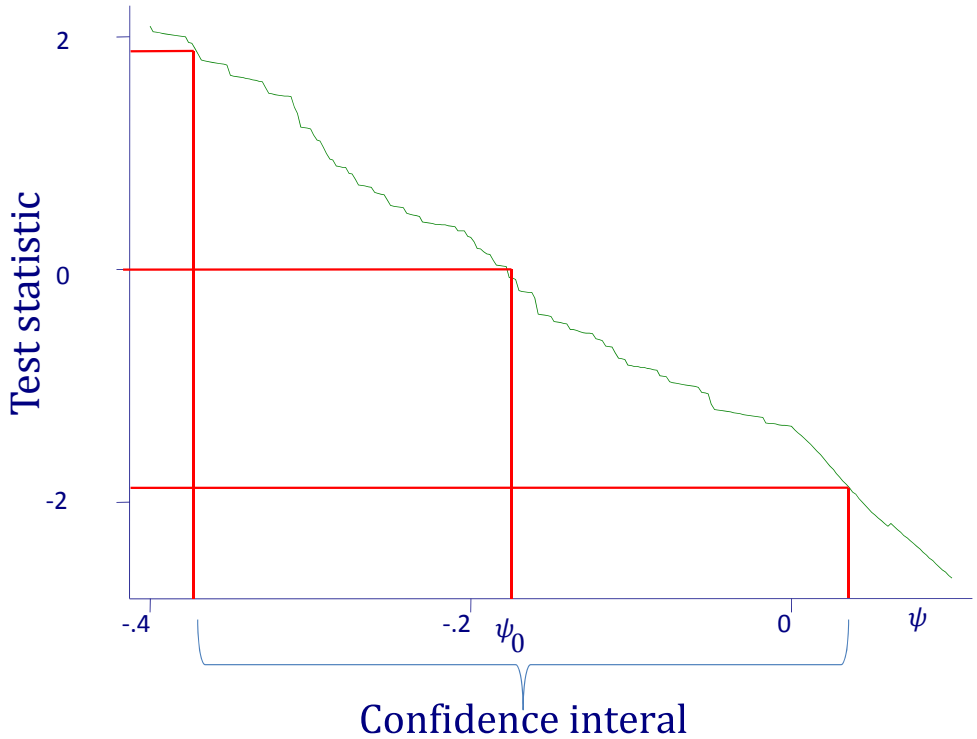


Figure: Hypothetical grid search for ψ_0

The confidence interval is the region where the test statistic does not reject the null hypothesis.

Remarks on the RPSFT model

The RPSFT model is "rank preserving" because a constant factor is used for adjusting the time to event for each patient. Thus, if two patients i and j are on the same treatment (either control or experimental), and patient i fails (dies) before patient j, before adjustment, patient i will also *always* fail before patient j after adjustment.

In other words, the ranking in failure times is preserved.

A key assumption of the RPSFT model is that the experimental treatment results in a constant change (reduction) in the time to failure or death, which is assumed equal for all patients before and after progression. This may not be a reasonable assumption in some cases, which may restrict the use of the method.

Advantages and disadvantages of the RPSFT model

A major advantage of the RPSFMT model is that it is "randomization respecting" method. In other words, it compares the two treatment groups as they were randomized.

Another major advantage is that, unlike the IPCW model, the RPSFT model does not assume no unmeasured confounding. In other words, it does not assume that we have accounted for the effect of all factors that are associated with both the outcome and the non-compliance.

A disadvantage of the method is the need to re-censor the data.

Another major disadvantage is the assumption that the treatment effect is constant regardless when, in the disease progression, the treatment is applied. This assumption in particular is unlikely to be 100% correct.

Sunitinib for metastatic renal-cell carcinoma (mRCC)

In this international phase III trial[7], 750 patients with mRCC were randomized to receive either sunitinib (n = 375) or interferon-alfa (IFN-α; n = 375).

Crossover was allowed only after an interim analysis had concluded a significant gain in the primary endpoint PFS.

Twenty five patients (7%) in the IFN-α group crossed over to sunitinib after an average of 70.8 weeks. There were 390 total deaths (190 in the sunitinib and 200 in the IFN-α arm).

All events were included in the ITT analysis. Censoring at the time of crossover (on-treatment analysis) led to the exclusion of five deaths in the IFN-α arm which occurred after crossover to sunitinib.

[7] Motzer et al., NEJM, 2007.

Analysis by RPSFT

In the RPSFT model, the estimated value for the acceleration parameter calculated using a grid search method was $\hat{\psi} = -0.244$, corresponding to a decrease in overall survival time of $\exp(\hat{\psi}) = 0.22$ with IFN-α than with sunitinib.

The results of all analyses are presented in the following Figure:

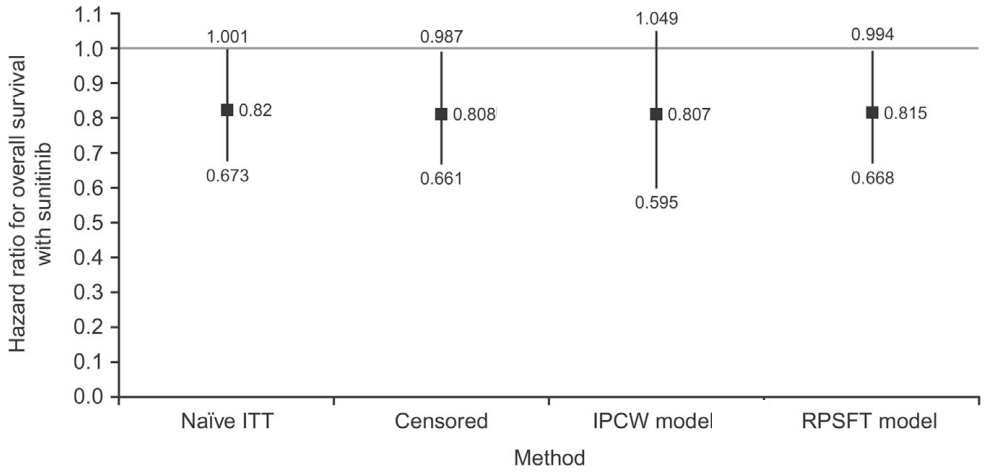


Figure: ITT and alternative analyses of the mRCC study

While the results are similar, it is notable that the IPCW model has wider confidence intervals as the censoring model and the impact of the RPSFT model was minimal, most likely because of the limited cross-over in the

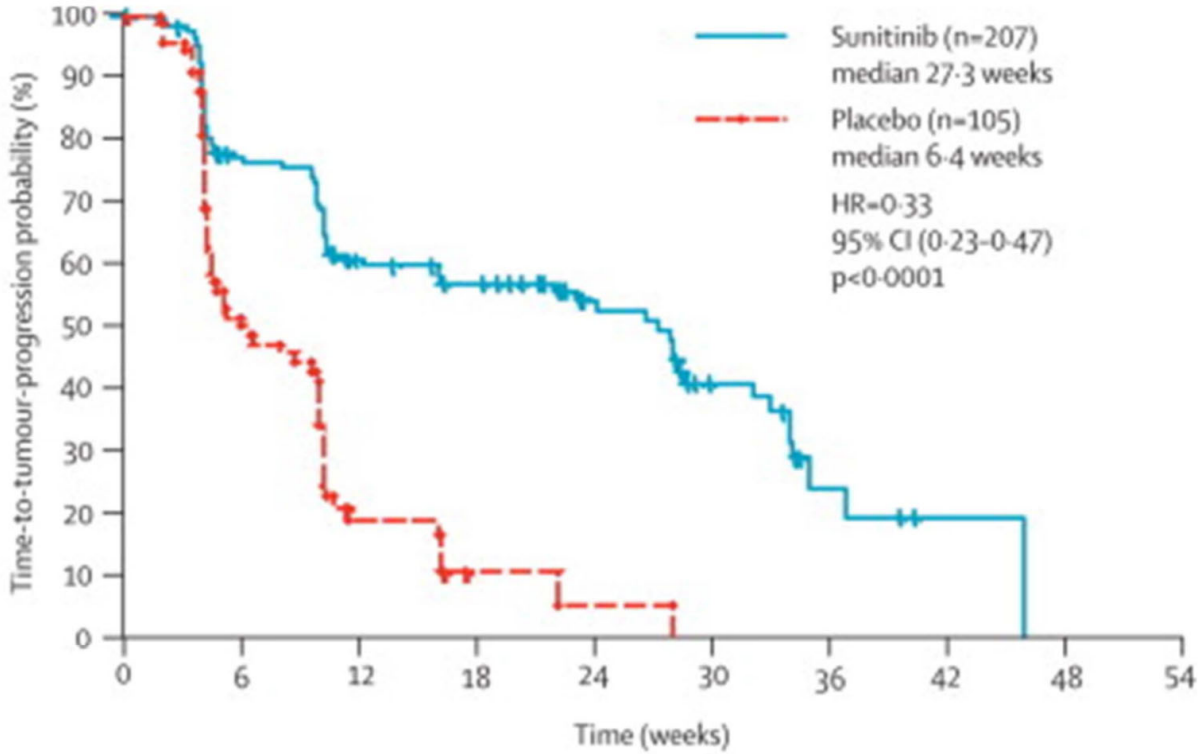
Sunitinib for gastrointestinal stromal tumors (GIST)

In this international phase III, multi-center, randomised, double-blind, placebo-controlled study of sunitinib for the management of gastrointestinal stromal tumors (GIST), 312 patients with with advanced and documented imatinib resistance were randomized to receive either sunitinib ($n = 207$) or placebo ($n = 105$).

Crossover was allowed after an interim analysis concluded a significant gain in the time to progression (Figure 5), all patients in the placebo arm were allowed to switch.

Sunitinib trial: Time to tumor progression

During the interim analysis, the following results were observed[8]



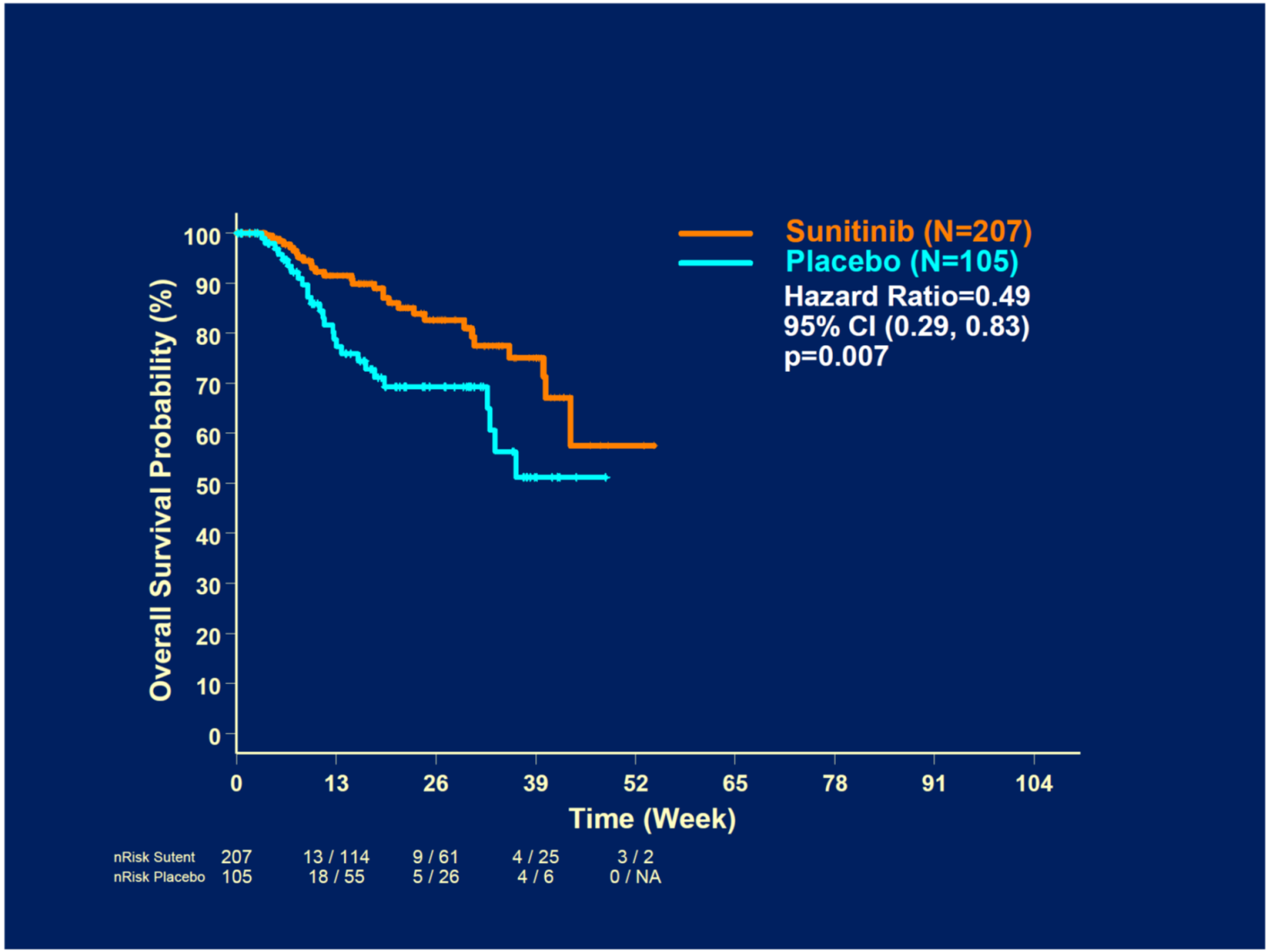
Number at risk

Sunitinib	207	106	67	53	34	18	5	1	0
Placebo	105	36	9	2	1	0	0	0	0

GD Demetri, AT van Oosterom, CR Garrett, et al. Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib: a randomised controlled trial. Lancet, **338**, 1329-1338. http://dx.doi.org/10.1016/S0140-6736(06)69446-4

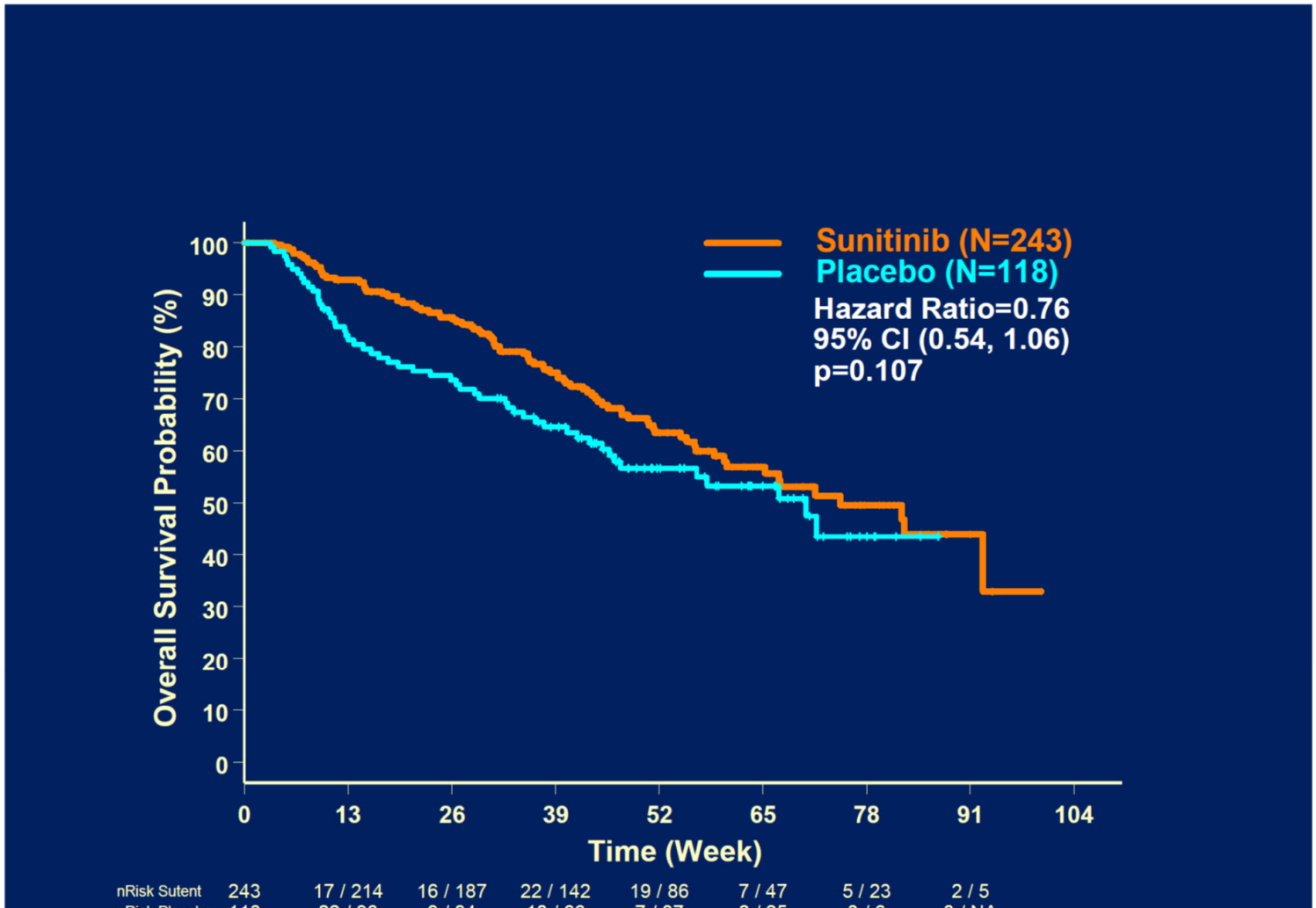
Sunitinib trial: Overall survival

The initial results of the study are shown in the following Figure[9].



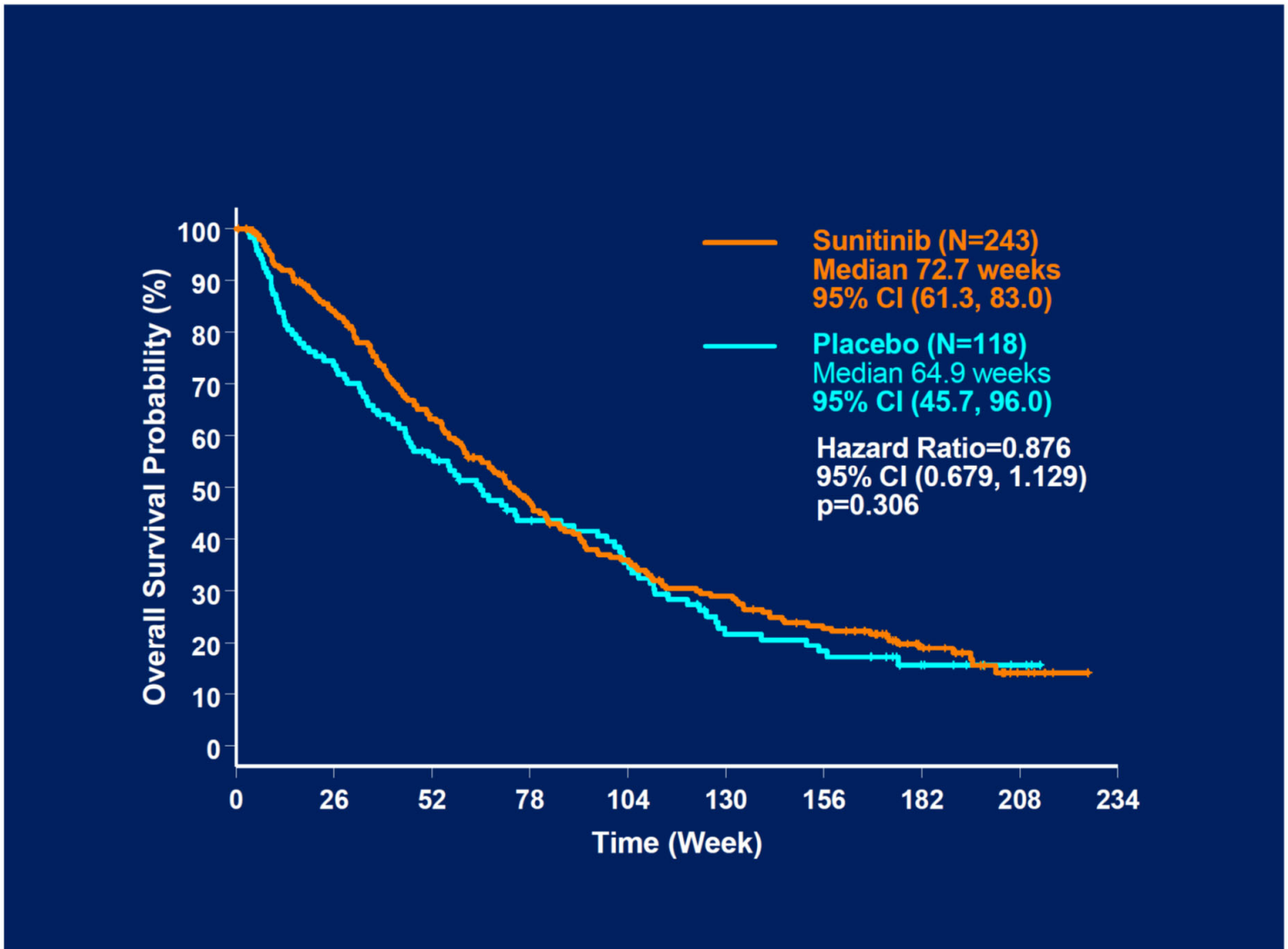
Sunitinib study: Extended follow-up

Extended follow-up of the study patients was presented in the following year's ASCO conference as shown below:



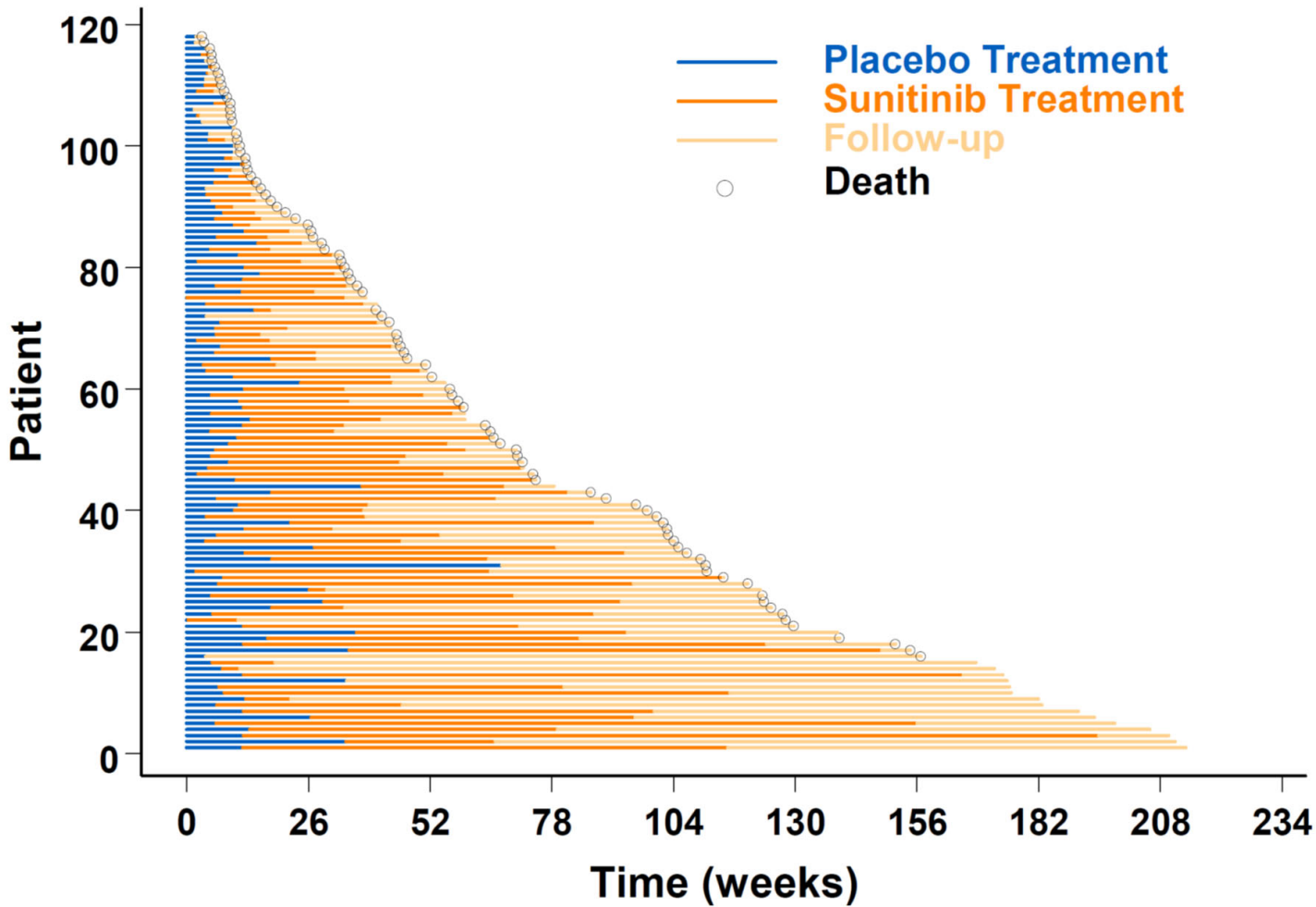
Sunitinib study: Final results

... and again in 2008:



Sunitinib trial: the impact of switching

The extent and impact of switching from placebo to sunitinib is shown in Figure 9.



Concern over the impact of cross-over

The main concern of course is that the decay in the survival advantage was due to the high proportion of patients crossing over at progression (Figure 9).

Out of 73 patients in the placebo group whose disease progressed, 69 crossed over to sunitinib. In fact, the median time to progression in the non-sunitinib arm was less than two months

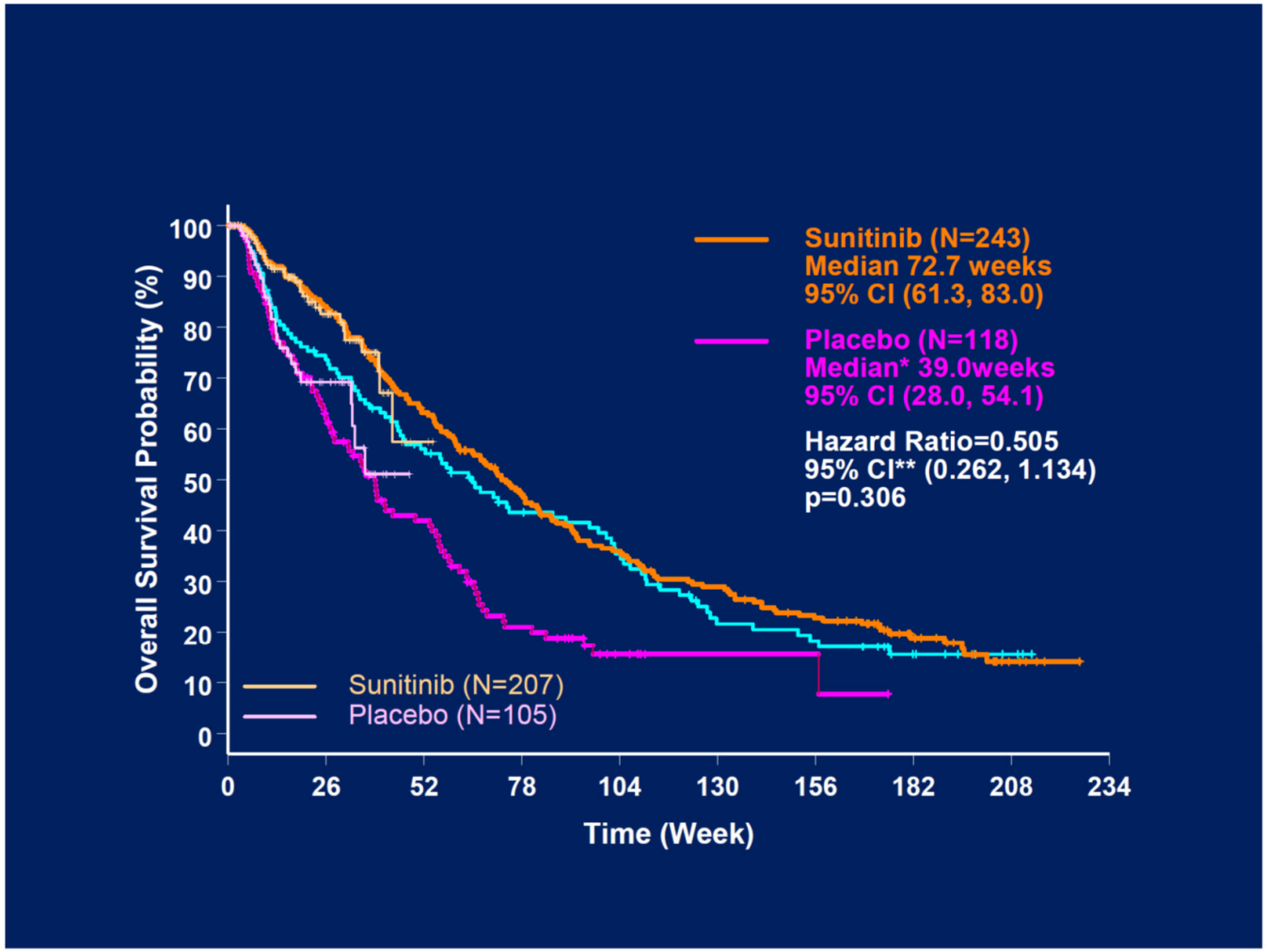
In an RPSFT model[10], the estimated value for $\psi = -0.656$ with the resulting hazard ratio was $\theta = 0.505$ (p=0.306).

This is to be compared with the hazard ratio of the usual ITT approach, where the hazard ratio of sunitinib versus placebo (which of course included many patients who switched) was $\theta_{ITT} = 0.876$ (p=0.306)

Note that the p-values in the ITT analysis and the RPSFT model are identical, even though the hazard ratio in the RPSFT model was substantially lower. This is by construction. The reason is the increased uncertainty induced by re-censoring.

The Sunitimib study: RPSFT model

The result of the RPSFT model is as follows:



Section 2

Estimation of treatment effects

Data analysis versus good design

Analysis from a trial data and estimation of treatment effects is the penultimate stage of the performance of a clinical study (the last being reporting of the results, which we will discuss in the next lecture).

Analysis of trial data requires a number of statistical methods and models and is considered the most important part of a study's implementation. This is because analysis appears closer to the results of the study.

However, the design of a study is much more important than the analysis of trial data and the latter cannot supplant the former.

In this section we discuss analytical approaches having to do with a number of contexts of clinical trials as well as context *within* a single trial (e.g., efficacy versus toxicity considerations).

Dose-finding and pharmacokinetic (PK) studies

Dose-finding studies have the following main outcomes of interest:

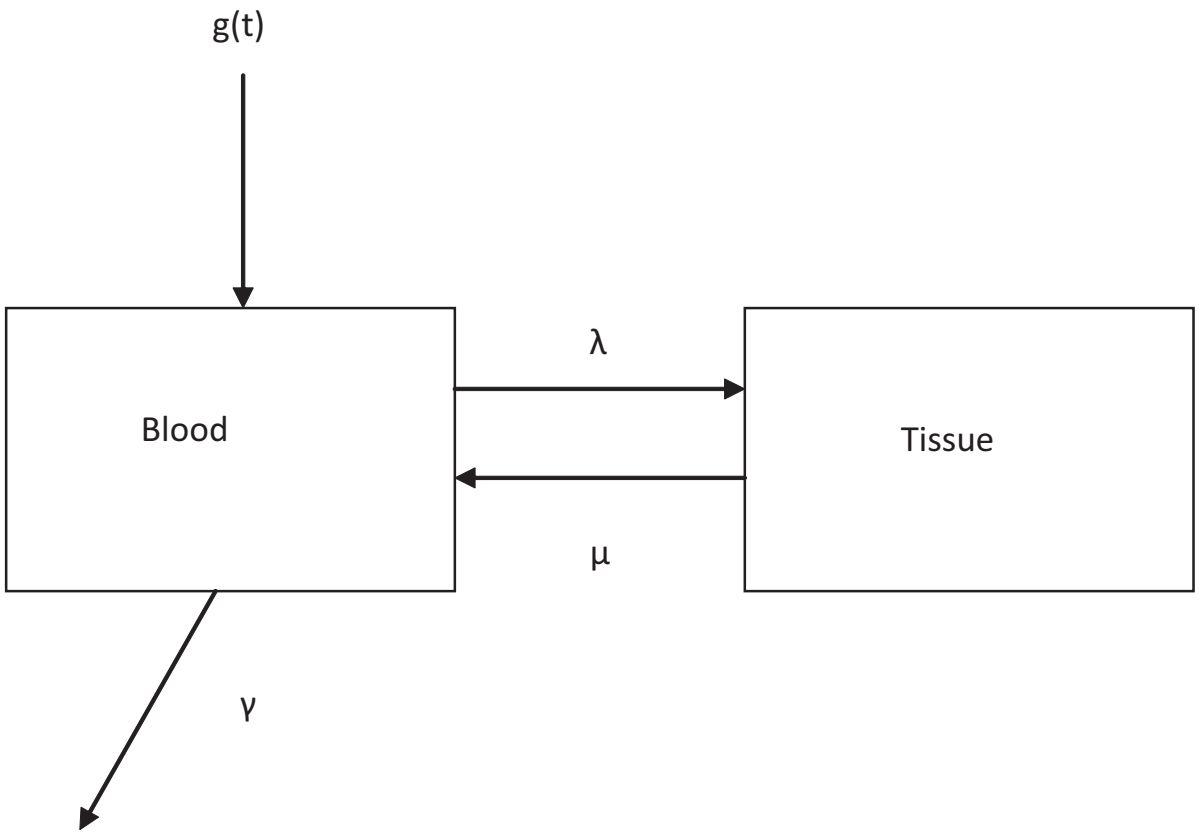
- Maximal tolerated dose (MTD)
- Absorption rate
- Elimination rate
- Area under the (drug concentration) curve
- Peak concentration
- Half life
- Correlation between plasma drug levels and side effects
- Proportion of patients who demonstrate evidence of efficacy

PK studies are instrumental in permitting investigators to address all of these outcomes.

A two-compartment PK model

Without going in too much detail about PK studies, we review here the basic two-compartment model.

Figure: The basic two-compartment PK model



A two-compartment PK model

In this model, a drug is infused into compartment X at a rate $g(t)$. The drug is transported from compartment X (e.g., blood) to Y (e.g., tissues) at a rate λ and back to X at a rate μ and is eliminated from X at a rate γ.

Mathematical modeling of the two-compartment PK model

The mathematical analysis of the two-compartment PK model is based on a system of differential (rate) equations such as

$$\frac{dX(t)}{dt} = \underbrace{-(\lambda+\gamma)X(t)}_{\text{levels leaving } X} + \underbrace{\mu Y(t) + g(t)}_{\text{levels returning to } X}$$

The solution to this system of equations is given by the following formulas:

$$X(t) = c_1(t)e^{\xi_1 t} + c_2(t)e^{\xi_2 t}$$

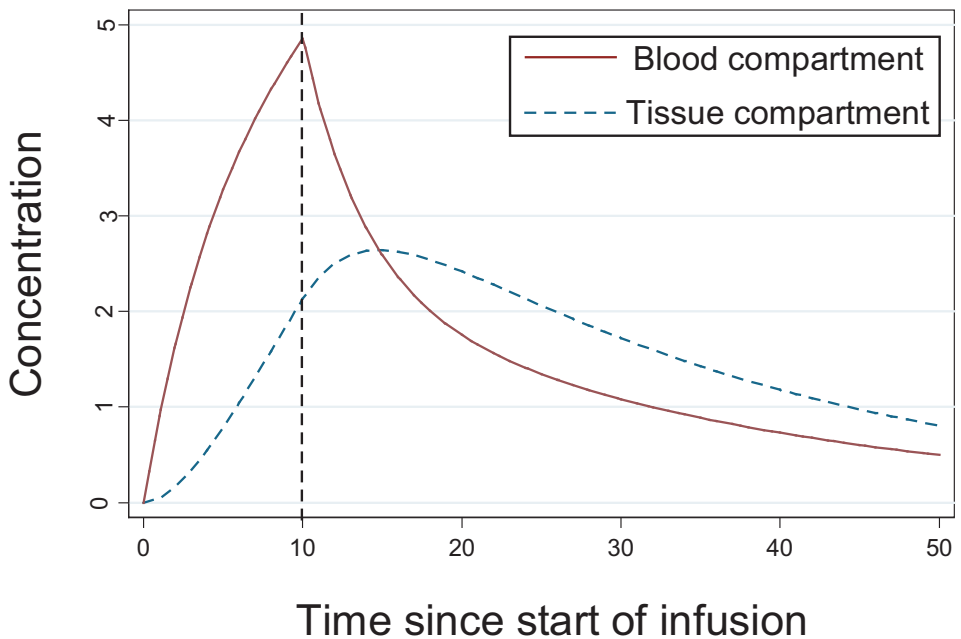
$$Y(t) = c_1(t)\frac{\xi_1+\lambda+\gamma}{\mu}e^{\xi_1 t} + c_2(t)\frac{\xi_2+\lambda+\gamma}{\mu}e^{\xi_2 t}$$

for appropriate ξ_1, ξ_2, and functions $c_1(t)$ and $c_2(t)$.

Mathematics of the two-compartment PK model (continued)

In the special case where the drug is infused at a constant rate $g(t) = g_0$ over time t_0 and the initial concentration in the X and Y compartment is $X(0) = 0$ and $Y(0) = 0$ we obtain two models describing the concentration in the two compartments. The concentration in the two compartments is shown in the following Figure.

Figure: The basic two-compartment PK model



Area under the (drug concentration) curve (AUC)

The area under the concentration curve for compartment X is given as

$$AUC_x = \underbrace{\int_0^{t_0} X(t)dt}_{\text{drug up to } t_0 \text{ in } X} + \underbrace{\int_{t_0}^{\infty} X(t)dt}_{\text{drug after time } t_0 \text{ in } X}$$

So that,

$$AUC_x = \frac{g_0 t_0 + X(0)}{\gamma}$$

Under the special circumstances mentioned earlier, $AUC_x = \frac{g_0 t}{\gamma}$. Similarly,

$$AUC_y = \frac{\lambda g_0 t_0 + \lambda X(0)}{\mu\gamma} = \frac{\lambda}{\mu} AUC_x$$

and under the special circumstances mentioned earlier, $AUC_y = \frac{\lambda g_0 t_0}{\mu\gamma}$.

Analysis of SA studies

SA studies are concerned with both efficacy and toxicity.

Often the efficacy and toxicity outcomes are expressed in terms of dichotomous (yes/no probabilities). Often, analyzing these data involves the estimation of absolute probabilities (proportions).

Case study: Study of Alimta for thymoma

For example, consider the following two-stage cancer trial of pemetrexed (Alimta) in thymoma, a rare cancer involving the thymus. The study was designed as follows:

- *First stage*
  Eighteen patients were to be accrued at the first stage. If one or more partial or complete response (based on RECIST criteria) were observed, the study would be continued to the second stage.
- *Second stage*
  Nine more patients were to be accrued in the second stage. If four or more responses (defined above) were observed the study would be considered successful.

Design of the thymoma study
Efficacy

The desired response (alternative hypothesis) was $p_A = 0.2$ while, a response below $p_0 = 0.05$ would be considered of no interest. The above design is not optimal in the sense of Simon but has the following characteristics:

- Ensures that the probability of early termination (PET) under the alternative hypothesis (i.e., under the assumption that $p = p_A = 20\%$) is less than 2%.
- Generates power of about 80% (actually, power is 81.6%).
- The exact type I error is $< 5\%$

Design of the thymoma study
Safety

The estimate of the upper and lower limit of the toxicity is based on all evaluable patients ($n = 27$ in this study). The confidence intervals are given at the 90% level, they are two-sided and the exact binomial distribution is used instead of the normal approximation. Given these considerations the 90% confidence interval for various scenarios is as given in the following table:

Number of toxicities [1]	90% CI
0	(0.000, 0.105[2])
1	(0.002, 0.164)
2	(0.013, 0.215)
3	(0.031, 0.263)
4	(0.052, 0.308)
.	.
.	.
.	.

[1] Grade 3 or higher toxicities (grade 3: Severe AE, 4: life-threatening, 5: death related to AE)

[2] 95% one-sided upper limit

Analysis of the thymoma study

Eighteen patients were accrued in the first stage. There were four responses observed (two partial and two complete). The study was continued and nine more patients were accrued with an additional partial response observed among the latter nine patients.

Clearly the study was successful. Now we need to figure out what the estimate and confidence interval of the response rate is.

We note that we cannot simply generate binomial confidence interval based on the final number of patients, but we should account for the fact that an interim analysis was performed.

Analysis of the thymoma study
Estimation of response

To generate confidence intervals, we use the program `KSTAGE` by Barry Brown.

The program essentially sums up (binomial) probabilities of all possible scenarios that can lead to the current state of affairs (Atkinson & Brown, *Biometrics*, 1985).

The output from this software is as follows:

```
Enter Number of Stages  and  Cumulative Number of  Trials for each Stage:
?
2 18 27

Enter Lo and Hi stopping values starting with stage  1:
 (-1 indicates no stopping)
?
0 -1
```

Note that we entered -1 for the upper limit (of response events) because the study will not stop regardless if the total number of responses required by the design is reached during the first stage.

Analysis of the thymoma study
Estimation of response (continued)

```
Enter  Stage Number  and  Event Number:
 (from which C.I. is calculated)
?
2 5
        K-Stage Design:

Number of Stages =  2

 Stage #    # of Trials    Cum # of Trials    Lo Quit   Hi Quit

    1           18                18             0        -1
    2            9                27

Kstg=  2  Kevt=   5

The 94% Confidence Interval is ( .0630344, .3808469)
```

Thus, the 95% confidence interval for efficacy is between 6.3% and 38%, which excludes 5% (the lower limit of efficacy). Thus, the study is a success!

Analysis of the thymoma study
Estimation of toxicity

There were 8 patients out of 27 that experienced at least one grade-3 toxicity during the study.

Disregarding the sequential nature of the study (which was not based on toxicity criteria anyway), the exact 90% binomial confidence interval for toxicity is 15.7%-47.1%.

On the other hand, no grade-4 or higher toxicities were observed, so the upper bound of the 95% one-sided confidence interval for the rate of grade-4 or higher toxicity is 10.5%.

Case study: The mesothelioma Phase II study

Rusch, Piantadosi and Holmes (*J Thorac Cardiovasc Surg. 1991*), report on a study of mesothelioma, a rare form of lung cancer associated with asbestos exposure. In that study, three approaches, biopsy, limited resection or extrapleural pneumonectomy (EPP) were attempted on 83 patients suffering from mesothelioma.

The complete data are given at

http://www.cancerbiostats.onc.jhmi.edu/Piantadosi_clinicaltrials/Software/Data%2BPrograms.zip.

The survival of patients in the three groups is given in the following table:

	age	sex	ps	hist	wtchg	surg	ptime	prog	stime	dead	X_st	X_d	X_t	X_t0
1	69	1	0	136	1	1	175	1	725	0	1	0	725	0
2	61	1	0	131	2	1	61	1	294	1	1	1	294	0
3	71	1	0	136	1	1	133	1	316	1	1	1	316	0
4	68	1	0	136	1	1	1009	1	1029	0	1	0	1029	0
5	65	0	0	NA	2	1	117	1	545	0	1	0	545	0
6	68	1	1	136	1	1	20	1	122	1	1	1	122	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Case study: The mesothelioma Phase II study
Descriptive summaries

The survival status in the three groups is given in the following table:

	Result		
Group	Dead	Alive	Total
Biopsy	32	5	37
Limited	21	5	26
EPP	15	5	20

We should observe the important fact that, in each group, five patients did not die by the end of the study. Their survival was not observed fully (we know simply that they did not die by the end of the study). These are "censored" observations.

Case study: The mesothelioma Phase II study
Analysis of survival data

Survival data are unique in that not all events (deaths) are observed. We analyse these data by breaking up the time scale in intervals according to observed deaths. The probability of death is given by

$$p_i = \frac{d_i}{n_i}$$

where d_i is the number of deaths observed in that interval and n_i the number of persons that were alive at the start of the interval. The probability of surviving that interval is, therefore, $1 - p_i = 1 - \frac{d_i}{n_i}$. The probability of surviving past the time of the kth failure t_k is

$$\hat{S}(t_k) = \prod_{i=0}^{k} \left(1 - \frac{d_i}{n_i}\right)$$

Case study: The mesothelioma Phase II study
Analysis of survival data (continued)

The results of the analysis are given in the following table:

Event Time t_i	Beg. Total	Number Fail	Number Lost	Survival Probability	Std. Error
4	82	1	0	0.9878	0.0121
6	81	1	0	0.9756	0.0170
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
475	28	1	0	0.3293	0.0519
499	27	0	1	0.3293	0.0519
503	26	1	0	0.3166	0.0514
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1265	2	1	0	0.0585	0.0502
1338	1	0	1	0.0585	0.0502

Case study: The mesothelioma Phase II study
Analysis of survival data (continued)

From the previous table we see how survival probability estimates are generated.

Starting with 100% probability of survival at time $t = 0$, and excluding the individual with zero survival, we drop to $1 - \frac{1}{82} = 0.9878$ after the first failure at time $t_1 = 4$ days.

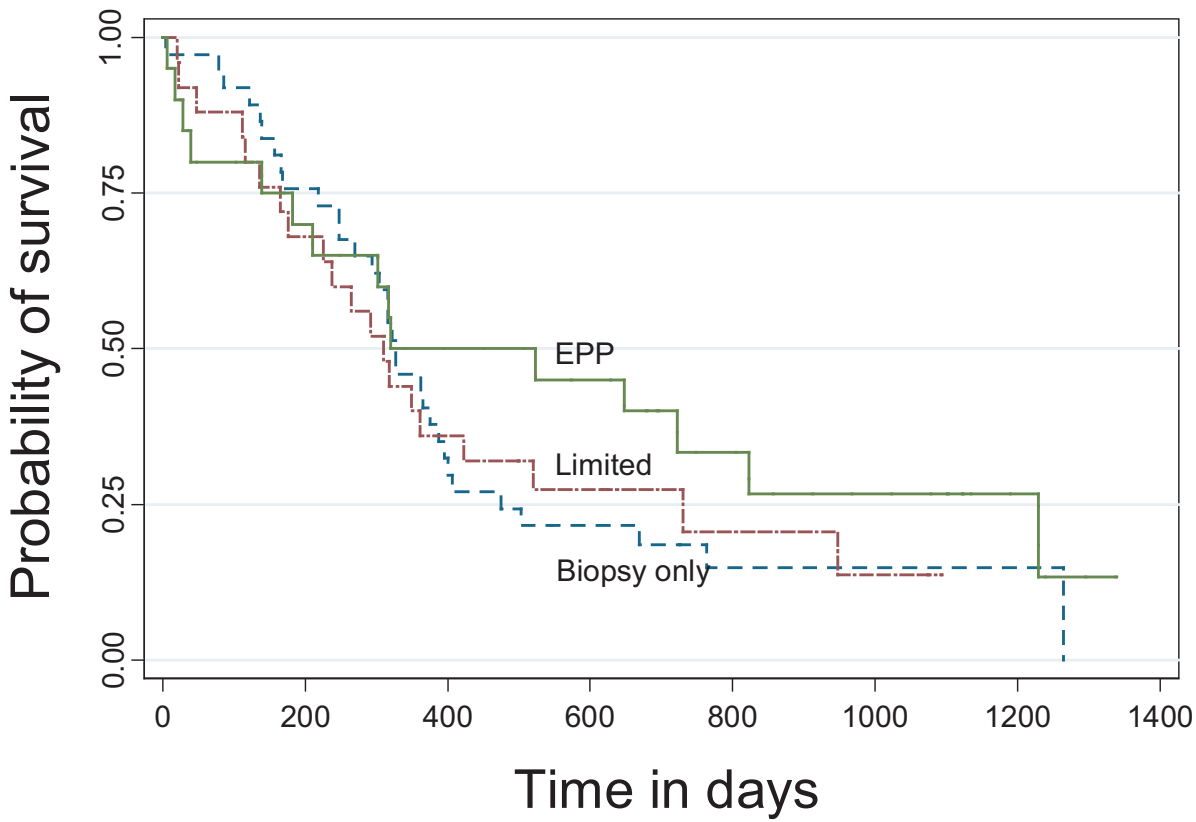
The next failure occurs at time $t_2 = 6$ at which point, the probability of survival is $p_2 = \left(1 - \frac{1}{82}\right)\left(1 - \frac{1}{81}\right) = 0.9756$.

By contrast, when an observation is censored (at time $t = 499$ days, the probability of survival through that interval is 100% so there is no difference in the probability for that subject. We see that the probability remains the same as that of the 55th failure at time $t_{55} = 475$ days, i.e., $\hat{S}(t_{55}) = 0.3293$.

Case study: The mesothelioma Phase II study

The Kaplan-Meier estimator of survival

We plot the estimate of survival over time by drawing a horizontal line between successive failures and a vertical line of length d_i/n_i at each event i. Censored observations are ignored. This produces the so called "Kaplan-Meier" estimate of survival.



Case study: The mesothelioma Phase II study
Comparisons between groups

Summaries of the survival experience in the three groups are given in the following Table:

surg	time at risk	incidence rate[3]	no. of subjects	Survival time 25%	50%	75%
biopsy	15042	.0021274	37	218	327	475
limited	9678	.0020665	25	165	310	730
EPP	10441	.0014366	20	139	320	1229
total	35161	.0019055	82	168	320	722

For example, the median survival in the three groups is 327, 310 and 320 days respectively.

[3] Equals, number of deaths divided by time at risk. For example, in the biopsy group this is $32/15042 \approx 0.00212$.

Case study: The mesothelioma Phase II study
Comparisons between groups:The log-rank test

To compare the survival in the three groups, we consider the so-called "log-rank" test. This test is based on the inherent ordering of the deaths by the time they occurred. At each death, we can construct a 3×2 table. The table will look as follows for the first failure that occurred at $t = 4$ days in the biopsy group:

	Result		
Group	Dead	Alive	Total
Biopsy	1	36	37
Limited	0	25	25
EPP	0	20	20

After generating these tables, we perform a Mantel-Haenszel test of association between surgical group and survival status. This measures whether, on average, the proportion of deaths falls inordinately on one or more of the three groups.

Case study: The mesothelioma Phase II study
The log-rank test (continued)

Carrying out the log-rank test analysis we obtain the following output:

```
Call:
survdiff(formula = Surv(stime, dead) ~ group, data = mesoth)

                N Observed Expected (O-E)^2/E (O-E)^2/V
group=biopsy   37       32     30.2     0.107     0.195
group=limited  26       21     18.5     0.346     0.483
group=EPP      20       15     19.3     0.969     1.401

Chisq= 1.5  on 2 degrees of freedom, p= 0.479
```

The p value of the log-rank test is 0.479 suggesting that there is no evidence for a difference in survival across the three surgical procedures.

Resampling methods

A very powerful methodology to generate distributions of various statistics is through resampling. The "bootstrap" as it's called, involves generating repeated analyses by resampling out of the dataset with replacement.

We can run a bootstrap analysis of the previous survival analysis to obtain the distribution. The median and the associated 95% confidence interval, based on the normal distribution, is 320 days and (292, 387) days respectively. The bootstrap estimate of the median and the 95% confidence interval is 320 days and (276.1, 363.9) days respectively.

The bootstrap, in this case, merely validated a known distributional result. The true power of the bootstrap is that it can generate similar distributions more difficult to calculate (e.g., the difference between two median survivals).

Comparative efficacy trials (Phase III)

While developmental studies such as DF and SA studies use mainly descriptive means to present the treatment effects, comparative trials describe data, quantify possible treatment differences and assess extraneous influence.

The usual approach is a test of statistical significance, i.e., determining to what extent the observed differences are attributable to random variation.

Case study: FAP prevention study

The following is the Familial Adenomatous Polyposis (FAP) dataset (Giardiello et al., *NEJM*, 1993):

id	sex	age	polyp number at month 0	polyp number at month 12	polyp size at month 0	polyp size at month 12	rx
1	0	17	7	–	3.6	–	1
2	0	20	77	–	3.8	–	0
3	1	16	7	4	5.0	1.0	1
4	0	18	5	26	3.4	2.1	0
5	1	22	23	16	3.0	1.2	1
6	0	13	35	40	4.2	4.1	0
7	0	23	11	14	2.2	3.3	1
8	1	34	12	16	2.0	3.0	0
9	1	50	7	11	4.2	2.5	0
10	1	19	318	434	4.8	4.4	0
11	1	17	160	26	5.5	3.5	1
12	0	23	8	7	1.7	0.8	1
13	1	22	20	45	2.5	3.0	0
14	1	30	11	32	2.3	2.7	0
15	1	27	24	80	2.4	2.7	0
16	1	23	34	34	3.0	4.2	1
17	0	22	54	38	4.0	2.9	0
18	1	13	16	–	1.8	–	1
21	1	34	30	57	3.2	3.7	0
22	0	23	10	7	3.0	1.1	1
23	0	22	20	1	4.0	4.0	1
24	1	42	12	8	2.8	1.0	1

Case study: FAP prevention study
Comparisons of month-12 polyp number and size

The baseline (month-0) and month-12 number and size of polyps in each treatment arm is shown in the following Table:

	Treatment arm		
	Treatment 0	Treatment 1	
Time point	Mean (± SD)	Mean (± SD)	p-value[1]
Number of polyps			
Month 0	53.9 (90.2)	28 (44.5)	0.403
Month 12	77.9 (126.7)	13 (10.8)	0.145
Polyp size			
Month 0	3.3 (0.93)	3.2 (1.22)	0.816
Month 12	3.1 (0.73)	1.8 (1.41)	0.022

[1] T test

This suggests that, while there is no significant reduction in the number of polyps in month 12, there might be a reduction of their size due to therapy (active treatment=1, standard treatment=0).

Case study: FAP prevention study
Using differences from baseline

Instead of comparing the month-12 number of polyps or polyp size, we can compare the *difference* between month-12 and month-0 in the number and size of the polyps. The revised analysis is given in the following Table:

	Treatment arm		
	Treatment 0	Treatment 1	
Time point	Mean (± SD)	Mean (± SD)	p-value
Number of polyps			
Month 12 difference	26.3 (36.9)	-18.7 (43.7)	0.026
Polyp size			
Month 12 difference	-0.2 (0.90)	-1.5 (1.8)	0.053

This analysis shows how much the variability of the measures under comparison has been reduced by removing the biological effect, which is the largest component of the variability. Now, there is both a reduction in the number and, possibly, in the size of the polyps associated with active treatment.

Case study: FAP prevention study
Analysis of covariance (ANCOVA) analysis

Another way to do this analysis is to adjust for the baseline number or size of the polyps. This involves a model for each subject i as follows:

$$Y_{12,i} = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1 Y_{0i}}_{\text{baseline quantity}} + \underbrace{\beta_2 T_i}_{\text{treatment effect}} + \underbrace{\epsilon_i}_{\text{error term}}$$

The results are given in the following table:

Dependent	Model Terms	Parameter Estimate	Standard Error	p-value
Number of polyps	β_0	21.5	14.2	–
	β_1	1.1	0.1	<0.0001
	β_2	-43.2	18.9	0.037
Polyp size	β_0	1.13	0.90	–
	β_1	0.21	0.1	0.403
	β_2	-1.29	0.51	0.023

The ANCOVA analysis shows that the size *and* number of polyps are significantly lowered in relation to the active treatment.

Case study: NSCLC lung cancer trial

Lad, Rubinstein, Sadeghi, et al. *J Clin Onc*, 1988) report a randomized trial of CAP (a combination of cytoxan, doxorubicin and platinum chemotherapy as adjuvant treatment to radiotherapy in non-small-cell lung cancer (NSCLC).

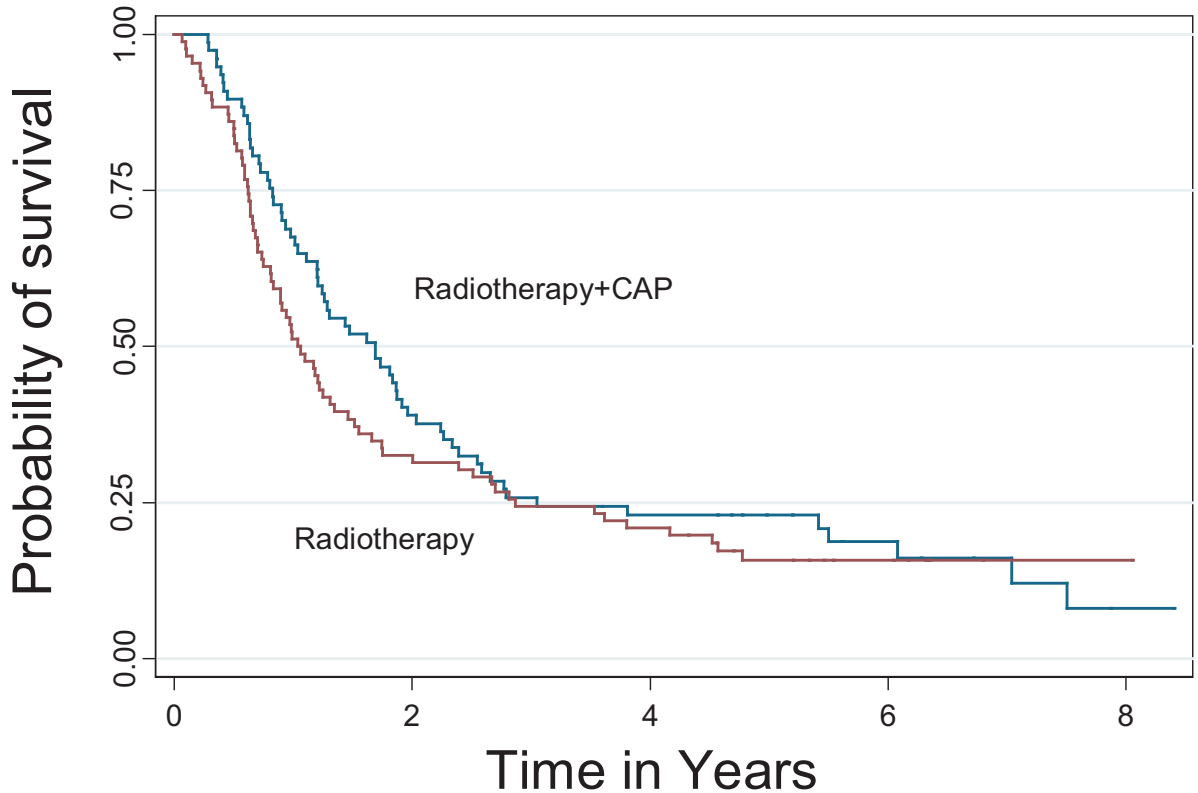
The data are given in the following list:

	celltype	karn	t	n	treat	surv	dead	dfs	event	age	race	elig	wtloss	sex	survyear
1	1	2	1	0	1	1046	1	413	1	70	1	0	1	1	2.8657530
2	1	2	2	2	0	342	1	342	0	67	1	0	1	1	0.9369863
3	1	2	2	2	1	54	1	18	1	61	1	0	0	1	0.1479452
4	1	2	2	2	0	303	1	264	1	52	1	0	0	1	0.8301370
5	1	2	1	2	1	295	1	248	1	59	0	0	0	1	0.8082192
6	2	1	3	2	1	88	1	59	1	39	0	0	0	0	0.2410959
7	2	2	2	2	1	241	1	241	0	46	0	0	0	0	0.6602740
8	2	2	1	2	1	567	1	252	1	44	0	0	0	1	1.5534250
9	2	2	2	2	0	286	1	211	1	38	0	0	1	0	0.7835616
10	2	2	2	1	0	265	1	262	1	62	1	0	0	1	0.7260274

Case study: NSCLC lung cancer trial: *Kaplan-Meier analysis*

The Kaplan Meier plot is given in the following Figure:

Figure: Survival by treatment group in the lung-cancer trial



Case study: Lung cancer trial
The Cox proportional hazards model

An approach to assess the effect of various factors on survival is through the Cox proportional hazards model. This model asserts that the *hazard* of death dependent on a number of predictors $\mathbf{X}$ is given by

$$\lambda(t;\mathbf{X}) = \lambda_0(t)e^{\beta'\mathbf{X}}$$

in other words, the predictor effect is multiplicative and is constant over time. This model is called proportional because the hazard ratio is indepedent of time. The implication of this model is that

$$\log\left\{\frac{\lambda(t;\mathbf{X})}{\lambda_0(t)}\right\} = \beta_1 X_1 + \beta_2 X_1 + \cdots$$

Case study: Lung cancer trial

Analysis via the Cox proportional hazards model

The analysis of the Cox model is given in the following table:

Factor	Haz. Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
treat="2"	1.30616	0.233470	1.49	0.135	0.920127	1.854151
cell type="2"	1.31154	0.241337	1.47	0.141	0.914435	1.881098
t="2"	0.91412	0.247078	-0.33	0.740	0.538186	1.552655
t="3"	1.16275	0.362530	0.48	0.629	0.631092	2.142298
n="1"	0.95181	0.356827	-0.13	0.895	0.456500	1.984541
n="2"	1.26627	0.448955	0.67	0.506	0.632026	2.536994
age	1.00383	0.010218	0.38	0.707	0.984000	1.024057
sex	1.06313	0.222285	0.29	0.770	0.705692	1.601625
weight loss	1.10107	0.346521	0.31	0.760	0.594196	2.040323
race	1.28824	0.356660	0.91	0.360	0.748749	2.216457

For example, the hazard among subjects in treatment 2 (radiotherapy) is 30.6% higher than treatment 1 (radiotherapy + CAP), or radiotherapy+CAP reduces the hazard to 76% ($\approx 1/1.31$) regardless of the length of the survival. This is not a statistically significant difference.

Other analyses

The spectrum of analysis of data generated from clinical trials is extensive. Some additional analyses used are:

- *Longitudinal analyses*
  These are analyses involving repeated measurements on the same subjects.
- *Time-dependent covariates*
  While many predictors are fixed at baseline and are assumed to have a constant effect over time, time-updated factors attempt to model factors that change over time.
- *Measurement error*
  While most analyses assume that all factors are measured exactly, a number of analyses have been introduced that allow for some error in the measurement of covariate predictors.

Other analyses (continued)

- *Random versus fixed effects*
  Most statistical models assume that the effects of predictors are fixed (non-random).

  For example, in assessing the effect of institution in a multi-center study, a fixed-effect analysis considers the institutions participating as fixed, while a random-effect analysis considers these as a random sample from all possible institutions.

  This has also been applied in longitudinal models that increase the flexibility of the statistical model by allowing different slopes or intercepts for each subject thus more realistically modeling response.

References

1. Blay JY. Pharmacological management of gastrointestinal stromal tumours: an update on the role of sunitinib. *Ann Oncol* 2010, **21**:208–215.
2. Demetri GD, Garrett CR, Schöffski P, et al. Complete longitudinal analyses of the randomized, placebo-controlled, phase III trial of sunitinib in patients with gastrointestinal stromal tumor following imatinib failure. *Clin Cancer Res* 2012; **18**: 3170–3179
3. Huang X, Xu Q. Adjusting the crossover effect in overall survival analysis using a rank preserving structural failure time model: The case of sunitinib GIST trial. 2nd Annual Pacific Coast Statisticians and Pharmacometricians Innovation Conference. San Luis Obispo, CA, July 13-15 2011.
4. Jönsson L, Sandin R, Ekman M, et al.Comparative effectiveness research/health technology assessment (HTA). Analyzing overall survival in randomized controlled trials with crossover and implications for economic evaluation. *Value in Health* 2014; **17**: 707–713.
5. Motzer RJ, Hutson TE, Tomczak P, et al. Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. *N Engl J Med* 2007;**356**: 115–124.

References (cont.)

- Robins JM, Finkelstein DM. Correcting for non-compliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**:779–788.
- Mark SD, Robins J. A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial. *Cont Clin Trials*; 1993; **14**: 79–97.
- Sterne JAC, Tilling K. G–estimation of causal effects, allowing for time–varying confounding *The Stata Journal***2**: 164–182
- Tsiatis A. ST 520: Statistical Principles of Clinical Trials. Lecture notes; chapter 8. Retrieved on 17 November, 2015 from www4.stat.ncsu.edu/d̃zhang/st520/520notes.pdf.
- White I. Methods for handling treatment switching: rank-preserving structural nested failure time models, inverse-probability-of-censoring weighting, and marginal structural models. Presented at the HTMR network workshop on Methods for adjusting for treatment switches in late-stage cancer trials. London, 20th February 2012.