Non-proportional Hazards for Time-to-event Outcomes in Clinical Trials: A practical guide to analysis strategies

John Gregson[a], Linda Sharples[a], Gregg W. Stone[b,c], Carl-Fredrik Burman[d], Fredrik Öhrn[d], Stuart Pocock[a]

[a] Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London

[b] Division of Cardiology, Columbia University Medical Center, New York Presbyterian Hospital, New York, New York;

[c] The Cardiovascular Research Foundation, New York, New York

[d] Statistical Innovation, Data Science and AI, R&D, AstraZeneca, Gothenburg, Sweden.

Word count: 4998

Disclosures: Drs. Burman and Öhrn are employees of Astra Zeneca. The remaining authors have nothing to disclose.

Corresponding author:
John Gregson
Department of Medical Statistics
LSHTM
Keppel Street
London
WC1E 7HT
John.gregson@lshtm.ac.uk

Acknowledgements:

We thank Daniel Jackson and Jonathan Wessen for their helpful comments on the article.

ABSTRACT

Most major clinical trials in cardiology report time-to-event outcomes using the Cox proportional hazards model so that a treatment effect is estimated as the hazard ratio between groups, accompanied by its 95% confidence interval and a log-rank P-value. But non-proportionality of hazards (non-PH) over time occurs quite often, making alternative analysis strategies appropriate. We present real examples of cardiology trials with different types of non-PH: an early treatment effect, a late treatment effect and a diminishing treatment effect. In such scenarios, we examine the relative merits of a Cox model, an accelerated failure time model, a milestone analysis and restricted mean survival time. We also present some post-hoc analyses for exploring any specific pattern of non-PH. We then make recommendations particularly regarding how to handle non-PH in pre-defined Statistical Analysis Plans, trial publications and regulatory submissions.

Key words: clinical trials, trial design, statistics, Cox proportional hazards, time-to-event outcomes, non-proportional hazards

INTRODUCTION

Clinical trials in cardiovascular disease often involve a time-to-event outcome, whereby patients are followed up from randomisation until the occurrence of a cardiovascular event or the end of the study. In such trials, a hazard ratio estimated from a Cox proportional hazards model is often reported as the main measure of treatment effect. The hazard ratio can be interpreted as the ratio of the event rate at any given time in the treatment group relative to the control group, and in the Cox model is assumed to remain the same throughout the duration of the study. However, in some instances, particularly where experimental and control treatments are very different, this is unlikely to be true. For example, in surgical trials a more aggressive or invasive strategy is sometimes associated with a higher early procedural risk but a lower long-term risk.[1,2] In drug trials, the effect of treatment may not materialize until several months, or even years after treatment initiation.[3,4] Non-proportional hazards describes situations like these where the hazard ratio is not constant over time. In such cases, an overall hazard ratio may not be the most informative summary of the treatment effect and alternative methods of analysis may be more suitable. However, there is a lack of practical guidance as to the best methods to assess the proportional hazards assumption, and which methods to use for analysis for the various types of non-proportional hazards. We therefore applied several methods for analysis to four clinical trials in cardiovascular disease. We describe the types of non-proportional hazards that occur and discuss the pros and cons of each method. We conclude with a set of recommendations on how to prepare for and then tackle non-proportional hazards in future trials.

METHODS

Data. We used data from four clinical trials with time-to-event outcomes, chosen as key examples of the different types of treatment effects over time that can occur.

The ASCOT study (Anglo Scandinavian Cardiac Outcomes Trial) was a randomized trial evaluating two experimental treatments using a 2x2 factorial design.[5,6] At baseline, 19,257 patients were randomized to one of two anti-hypertensive treatments: an amlodipine-based regime or an atenolol based regime. This part of the trial is called the ASCOT blood-pressure lowering arm (ASCOT-BPLA)[6] and our analysis of ASCOT-BPLA used the secondary endpoint of cardiovascular mortality. A subset of 10,305 patients with non-fasting total cholesterol concentrations of at least 6.5 mmol/L were also randomized to either atorvastatin or placebo. This part of the trial is called the ASCOT lipid-lowering arm (ASCOT-LLA).[5] Our analysis of ASCOT-LLA used the secondary endpoint of total coronary events.

The CHARM program randomized 7,599 patients with chronic heart failure to candesartan or placebo therapy in three randomized trials assessing the impact of candesartan on time to first heart failure hospitalization or cardiovascular death.[7] We analysed data from the CHARM-Overall program, which included patients from all three trials, had a primary outcome of all-cause death and a median follow-up of 3.1 years.

The EXCEL trial randomized 1,905 eligible patients with left main coronary artery disease to either coronary artery bypass graft (CABG group) or percutaneous coronary intervention (PCI group) with fluoropolymer-based cobalt–chromium everolimus-eluting stents.[1] The primary outcome, and the focus of our analyses was a composite outcome of death, stroke or myocardial infarction during the first 3 years of follow-up.

Statistical methods.

Identifying non-proportional hazards

To assess the proportional hazards assumption, Cox suggested that one can fit an interaction between the estimate of the (log) hazard ratio and time, with time modelled as a linear covariate.[8] We use this method to assess whether there is a statistically significant deviation from proportional hazards. An alternative is to compare the hazard ratio before and after a threshold time, i.e. by fitting an interaction between log hazard ratio and an indicator variable for time. This approach can have greater power to detect certain types of non-PH, but can increase the risk of falsely rejecting proportional hazards when the threshold time is chosen post-hoc. Several further methods (e.g. Grambsch-Therneau test) have been suggested for detection of non-PH, but none are clearly superior to using a time-treatment interaction variable.[9] Graphically assessing the extent of non-proportional hazards can also be useful. There are several informative methods for this purpose[10]; we used plots of the smoothed scaled Schoenfeld residuals against follow-up time, which show a smoothed estimate of the log hazard ratio against follow-up time. If the proportional hazards assumption is true then the underlying log-hazard ratio is constant over time, and so we expect the plot to show an approximately horizontal line.

Estimation methods

We used four statistical methods for the estimation of the overall treatment effect. They are summarised in the Central Illustration and illustrated in Figure 1. Our first analysis used the **Cox proportional hazards model** to estimate the hazard ratios associated with treatment.

Our second analysis used an **accelerated failure time model** to estimate time ratios associated with treatment. The time ratio describes the estimated delay until an event occurs with treatment relative to control. For example, a time ratio of 2 would mean that the time until an event occurs is twice as long in the treatment group relative to control, everything else being equal. There are several types of accelerated failure time models; we used the log-logistic model for the baseline hazard function, because unlike some other models (e.g. exponential) it is not restricted to proportional hazards.

Our third analysis estimated the difference in the percentage of patients with an event in the treatment group compared to the control group at a fixed time since baseline, known as the milestone time. We refer to this analysis as a **milestone analysis.** The percentage of patients with an event in each group was estimated using the Kaplan-Meier method, and the Greenwood formula was used to estimate standard errors. This method is similar to using logistic regression or calculating the odds ratio associated with treatment, but also accounts for loss to follow-up. The choice of milestone time is an important aspect in such analyses. In ASCOT-LLA, CHARM and EXCEL we chose a milestone time of 3 years, close to the median follow-up times in each of these studies (3.3 years, 3.1 years and 3.0 years respectively). A milestone time of 5.5 years was chosen in ASCOT-BPLA (median follow-up 5.5 years).

Our fourth analysis estimated the difference in **restricted mean survival time** (RMST difference) between groups, up until a fixed milestone time. Survival here refers to event-free survival, i.e. the absence of an outcome event, rather than simply continuing to be alive. The RMST in each group refers to the mean time spent free from an outcome event in each group up until the milestone time, after adjusting for loss to follow-up. The RMST difference can be represented as the difference in areas under the Kaplan-Meier plots for each group (Figure 1). Following the advice of Royston & Parmar,[11] we modelled event-free survival separately in each of the treatment and control groups using a flexible parametric survival model with 3 degrees of freedom (except in the EXCEL study wherein we used 2 degrees of freedom in order to achieve model convergence). We note that RMST can also be calculated using non-parametric methods.[12]

For each method, we estimated the appropriate treatment effect, its 95% confidence interval and a p-value from the corresponding hypothesis test.

In addition, we used some methods more suitable for post-hoc analyses. We used **piecewise hazards models**, whereby time since baseline was split into segments, and hazard ratios were calculated separately for each period of time by applying a Cox-proportional hazards model within each period.

Finally, we assessed the number of patients that would be required to achieve 80% power under each analysis method, assuming the observed time pattern of treatment difference is the truth. It is known that the standard error of each estimated treatment effect is approximately inversely proportional to the square root of the sample size. By using this relationship and the estimated standard error from each analysis it was possible to calculate the approximate sample size required to achieve 80% power (see Supplementary Appendix for further details). Analyses were done in Stata 15.1; flexible parametric models were implemented in the stpm2 package.

RESULTS

The cumulative incidence of events in each treatment group is shown for each of the four studies in Figure 2, with the pattern of treatment effect appearing to differ in each study. In ASCOT-LLA (Figure 2a) there was a steady divergence between cumulative incidence curves over time. This pattern is typical when proportional hazards is a reasonable assumption. ASCOT-BPLA (Figure 2B) showed a slightly different pattern. The cumulative incidence curves in each group were very similar for the first 2.5 years of follow up, and then gradually diverged, an example of a delayed treatment effect. Conversely, in CHARM (Figure 2c), the cumulative incidence curves diverged very early during follow up but then ran parallel to one another after 6 months. This pattern is referred to as an early effect. Finally, in EXCEL (Figure 2d), the curves diverged early on, but the early effect of treatment was not maintained, with the cumulative incidence curves converging later; a pattern we term a diminishing treatment effect.

We applied each of our estimation methods to each of these four studies.

ASCOT-LLA. In the ASCOT-LLA study total coronary events were compared amongst patients receiving atorvastatin or placebo. The treatment effect observed was consistent with proportional hazards; there was an approximately horizontal line throughout follow-up in the plot of Schoenfeld residuals and the test for a treatment-time interaction was non-significant (p=0.90, Supplementary Figure 1).

We found a statistically significant reduction in total coronary events regardless of the method used for analysis (Table 1). The hazard ratio of 0.71 from a Cox proportional hazards model and the time ratio of 1.51 from an accelerated failure time model have very similar interpretations. In the former the hazard for coronary events is 29% lower with atorvastatin and in the latter the time until a coronary event is delayed by 51%. In the milestone analysis, the 3-year event rate was 1.3% lower with atorvastatin and the RMST difference estimated that on average a patient was event-free 7.6 days longer with atorvastatin in the 3 years after randomization.

Note both the milestone analysis and RMST difference estimate an absolute effect of treatment whereas the Cox and accelerated failure time models estimate a relative effect (hazard ratios or time ratios, respectively). However, when the proportional hazards assumption is valid, milestone analysis and RMST have disadvantages. Data collected after the milestone time is ignored, resulting in a loss of power, and so more patients would be required in a trial using these methods as the primary analysis (Table 2). In addition, the choice of milestone time is somewhat arbitrary.

ASCOT-BPLA. Amlodipine was compared to atenolol in the ASCOT-BPLA trial. There was clear evidence of non-proportional hazards (p=0.0013; Supplementary Figure 1). A highly significant reduction in cardiovascular death was seen with all methods except for RMST. The hazard ratio comparing amlodipine to atenolol was 0.76 (p=0.0012), the time ratio was 1.29 (p=0.0010) and the estimated reduction cardiovascular deaths by 5.5-year was 0.8% (p=0.0019) (Table 1). However, the estimated RMST difference of 2.9 days was not statistically significant (p=0.31). When a delayed treatment effect is present, there is a large reduction in statistical power if RMST is chosen as the primary analysis, so that a trial using RMST would require many more patients than a trial using any of the other methods (Table 2). This is due to the lack of benefit in the early period of follow-up, and occurs despite less than 6% of total patient follow-up occurring after the milestone time. Additionally, while the RMST difference of 2.9 days seems like a small difference, a careful interpretation is warranted because it is plausible that the differences in event-free survival would continue to accrue beyond the milestone time. For example, suppose we were to take a longer-term perspective and consider RMST difference at 10 years. Suppose also that the risk of cardiovascular deaths between 5.5 years and 10 years were the same in both groups so that there is no further benefit or harm related to treatment after the milestone time. We would then expect the cumulative incidence curves for each group to run approximately parallel to one another between 5.5 and 10 years. The RMST difference (which can be visualised as the area between the cumulative incidence curves) would therefore continue to accrue and would be greater at 10 years than it was at 5.5 years (~16.2 days vs. 2.9 days; Supplementary Figure 2).

CHARM. In the CHARM-Overall study, in which the effect of candesartan was compared with placebo on all-cause death, there was strong evidence against proportional hazards (p-value for treatment-time interaction=0.009). There was an apparent early effect that lasted only for the first 6-18 months following randomisation (Figure 2c, Supplementary Figure 1). The evidence against proportional hazards was even stronger when comparing the HR before 6 months to the HR thereafter (0.59 vs 1.00, p-value for interaction=0.0001). In analyses of the effect of candesartan on mortality the p-value was close to 0.05 for all methods except for the RMST difference where there was strong evidence that mean 3-year survival was longer with candesartan (21.0 days, p=0.0008). The Cox model (hazard ratio 0.91, p=0.055) failed to demonstrate a treatment benefit, whereas results from an accelerated failure model (time ratio 1.11, p=0.032) and a milestone analysis (1.95% 3-year reduction in deaths, p=0.044) were both just statistically significant. RMST difference will generally be more statistically powerful than the other methods with an early treatment effect (Table 2). In general, any treatment difference in early events has a greater influence on the RMST difference than events occurring later. This can be visualised by considering the Kaplan-Meier plot in Figure 2c, in which a gap between the curves opens up before 6 months and the area between the two curves continues to accumulate with time. In contrast, the relative importance of early and late events is broadly similar with the other three methods so that the similar event rate later in follow-up had a greater diluting influence on the apparent early treatment benefit.

Data from the CHARM study also demonstrate the sensitivity of findings to the choice of milestone time. Although there was some evidence of a difference in 3-year survival, had we chosen a different milestone time, for example 34 months or 38 months the difference would not have been statistically significant (p=0.093 and p=0.096 respectively). On the other hand, if one takes a short-term perspective with the milestone set at 6 months, then the mortality difference is highly significant (4.9% vs 2.9%, p<0.0001). It is also worth noting that for treatment effects that decrease over time, the exclusion of data after the milestone time does not lead to a loss in statistical power, because including later deaths would further dilute the significance of the early treatment effect.

EXCEL. The EXCEL study compared a composite outcome of stroke, MI or death in patients with left main coronary artery disease treated with PCI or CABG. There was a much higher procedural risk of the composite outcome with CABG (7.9% within 30 days) compared to PCI

(4.9%), but by 3-years the proportions of patients with the composite outcome were similar for CABG (14.7%) and PCI (15.4%) (Figure 2d). Unsurprisingly therefore, there was very strong evidence against the proportional hazards assumption with the estimated log hazard ratio differing markedly over follow-up (p for treatment-time interaction=0.003, Supplementary Figure 1).

None of the main methods for estimation demonstrated a clear treatment benefit for either intervention, although there were notable differences between the methods. Results from a Cox model (hazard ratio for PCI vs CABG: 1.01, p=0.97) or from an accelerated failure time model (time ratio: 1.06, p=0.88) do not provide much insight. Naively interpreted, these estimates appear to indicate a lack of difference between groups, whereas the two treatments clearly differ in the timing of the risk of outcomes occurring with each intervention. The underlying assumption of proportional hazards used in a Cox model and assumption of a constant time ratio in the accelerated failure time model were clearly not satisfied (treatment-time interaction p<0.001 for both). The milestone analysis at 3 years for the percentage with the primary outcome (treatment difference +0.5%, 95% CI -2.7% to 3.7%) is readily interpreted and does not make any modelling assumptions, but it does not take into account the difference in the timing of events during follow-up. The RMST difference is perhaps the most useful of the four methods for summarising the data from the EXCEL study. It takes into account the fact that although the total number of events was similar in the two groups, they tended to occur later in the PCI group thereby lengthening the time a patient was event-free. The estimated gain in event-free survival up to 3 years is 18.3 days (95% CI, -11.1 days, 47.8 days), but the difference is not statistically significant.

In order to further understand the results from EXCEL, we performed 3 sets of post-hoc analyses. First, because the primary outcome was a composite of clinically heterogeneous events, we present cumulative incidence curves separately for each component (Supplementary Figure

3). We find that the lower procedural risk with PCI is largely due to a reduction in MI and stroke, rather than death. If considered alongside the patterns of the individual events through time, this analysis may be helpful in suggesting how future event rates might differ in the two groups. For example, it may suggest whether future Kaplan-Meier curves will continue to converge, crossover or progress in parallel. In a second analysis we used piecewise hazards models, where we split follow-up time into three segments representing procedural, mid-term and long-term follow-up, calculating hazard ratios separately within each segment. The hazard ratios for PCI versus CABG were 0.61 (95% CI 0.42 to 0.88) within 30 days of randomisation, 1.05 (95% CI, 0.64 to 1.70) from 30 days to 1 year, and 1.93 (95% CI, 1.25 to 2.97) from 1 year to 3 years (Figure 3a). This simple approach can provide useful insight into the underlying patterns of risk. In a third approach, we generated additional diagnostic graphical displays (Figures 4b & 4c). Figure 3b shows the difference in event-free survival estimates and 95% confidence intervals throughout follow-up, equivalent to performing a milestone analysis at each day during follow up. This visually demonstrates that the early benefit of PCI is gradually eroded over time by an increased post-procedural risk. Figure 3c shows the difference in mean event-free survival time over study follow-up. The upward trend of the curve shows that the early benefit due to reduced procedural PCI risk has an effect on RMST out to nearly 3 years. For any choice of milestone time in the range up to 2 years the treatment difference in RMST is statistically significant. The greater number of primary events after PCI thereafter reduces the apparent benefit, whilst the confidence interval increases in width so that the treatment effect is no longer statistically significant.

One caution in all these post-hoc analyses is that no correction is made for multiple testing, as they need to be perceived as exploratory analyses.

DISCUSSION

For clinical trials of time-to-event outcomes, it has become standard practice to use Cox proportional hazards models both for trial design (e.g. power calculations) and statistical analysis. However, this may not be the best approach when the effect of treatment varies over time. Our analyses of four major cardiology trials demonstrate some alternative approaches, and outline some of their advantages and disadvantages under various patterns of treatment effect.

When proportional hazards are satisfied, the Cox PH model is the most statistically powerful method, and hazard ratios are readily understood by clinicians. We therefore see little practical reason to use alternative analysis strategies as the pre-specified primary analysis when deviation from PH is not expected, despite recent critiques of the hazard ratio for estimating treatment effects.[13] However, when major deviations are anticipated, it may be possible to adapt the design. In studies where an early treatment effect is anticipated, it may be possible to recruit fewer patients whilst maintaining adequate power by using RMST differences as the primary method of analysis rather than the Cox model. In contrast, when a delayed treatment effect is likely RMST difference is best avoided. In addition, with a delayed effect the sample size may need be inflated to allow for the extra variability caused by events occurring at the beginning of the trial when there is no difference between treatment groups, as was done in the CORONA trial of rosuvastatin.[14]

In most cases the type of treatment effect is unknown in advance but the analysis method needs to be pre-specified. Unfortunately, there is no clear “best” method across all types of treatment effect. Although we are aware of several tests-based methods that maintain good statistical power to detect differences between treatments across a range of types of non-PH,[15,16] these methods only provide a p-value without an accompanying estimation method linked to the test. An example is a test based on a series of weighted log-rank tests, where some of the tests counter-intuitively weight events occurring later in follow-up as more important than those occurring earlier.[16] The p-value from such tests is informative only in that it shows that the pattern of survival differs between treatment groups; it does not identify which treatment is "better" nor does it quantify how the difference between groups affects patient outcomes. Therefore, we do not recommend methods based only on hypothesis testing.

Non-PH can have important implications for trial design beyond the choice of analysis strategy. When treatment is associated with a lower (or higher) short-term risk which later reverses, it is important that the trial continues for sufficient duration so that the long-term effects of the treatment can be fully understood. Longer-term (i.e. 5-year) results from the EXCEL study will therefore be helpful to further understanding of the risks and benefits of PCI relative to CABG in patients with left main coronary disease. A second implication for trial design is that the stopping criteria used by data monitoring committees should take into account potential non-PH patterns of treatment effect. For instance, the CHARM program DSMB did not recommend stopping early even though a planned interim analysis of short-term mortality showed a highly significant reduction in mortality on candesartan.[17] Conversely, caution would be required when stopping a trial early for futility if a delayed effect was anticipated.

Post-hoc analyses of trials with non-PH can sometimes provide useful insights. A first step is to assess whether the proportional hazards assumption is reasonable. Using graphical displays such as scaled Schoenfeld residuals can help determine whether a single hazard ratio captures the effect of treatment with a reasonable degree of accuracy across the entire follow-up period. Formal statistical testing of proportional hazards is sometimes useful, but such tests can miss clinically important deviations from proportional hazards in small studies while detecting clinically unimportant deviations from proportional hazards in very large studies.

When the proportional hazards assumption is not reasonable, using a piecewise hazards model can be useful. In our analysis of EXCEL it helped identify periods during which the hazard with PCI was less than, similar to, or greater than the risk with CABG. A limitation of this methodology is that the post-hoc selection of time periods that appear visually different may exaggerate the real differences in hazard ratios over time. The hazard ratios calculated for later time periods also only include survivors of earlier time periods, and so are not truly randomised comparisons. A further post-hoc analysis not considered here is to explore whether non-PH has arisen by the combination of clinically distinct subgroups of patients in whom the effect of treatment is different. It is possible to have non-PH overall even though the PH assumption is satisfied within subgroups. Examples where this may have occurred are present in the medical literature[18,19], although we are unaware of any convincing examples in cardiovascular trials to date. In such scenarios, subgroup analyses and/or stratified Cox proportional hazards models may be useful.

One major concern is how one incorporates potential non-PH into the predefined statistical analysis plan for a major trial. It can be hard to anticipate the existence and pattern of non-PH in a trial, so in most circumstances the Cox PH model and associated log-rank test will be the predefined primary analysis. However, we would encourage SAPs to document contingency plans for an alternative primary analysis should clear evidence of non-PH be detected when the trial is unblinded for the final analysis. For instance, if clear evidence of a pattern of early treatment effect is reported, then the PH assumption is violated. An analysis using RMST could then be performed (as could have been applied to the CHARM program). Another example is in a meta-analysis of trials for oseltamivir treatment in influenza, where the pre-defined intent of Cox models was replaced by accelerated failure time models, since the former did not "fit the data" whereas the latter did.[20] However, if not pre-specified, to change the primary analysis methodology of the primary outcome in light of lack of model fit is a radical step. Further debate is needed as to when such a step is truly acceptable in the primary publication and/or regulatory submission of a major clinical trial, and how this can be done in a way that does not result in an increased probability of false positive findings. A key question is "how great does the departure from the PH assumption need to be?" in order to replace the focus on a Cox PH model with emphasis on an alternative more appropriate technique.

Our review has limitations. First, with only one study for each pattern of treatment effect, the generalisability of our study may be questioned. However, our analyses are meant to be illustrative, with some of the findings, for example those relating to the power of RMST relative to a Cox model under various patterns of treatment effect, already established in the statistical literature.[21] Second, we did not present data on a fifth type of treatment effect, wherein the Kaplan-Meier curves cross during follow-up ("crossover pattern"), as was observed in the STICH trial of CABG versus medical therapy.[22] In such situations, a single effect estimate is unlikely to accurately capture the effect of treatment so the choice of appropriate statistical analyses would require careful consideration. One would need to consider the relative importance of later versus earlier events, and ensure that the study continues for long enough to allow a full understanding of the effect of treatment over time.

In conclusion, serious attention needs to be given to appropriate analysis strategies when non-PH are evident in time-to-event outcomes. It is important to detect the type of non-PH that is present and select the analytical technique most appropriate to that situation. The consequences for more thorough Statistical Analysis Plans, trial publications and regulatory submissions need a further collective clarity of thought.

REFERENCES

1. Stone, G. W. *et al.* Everolimus-Eluting Stents or Bypass Surgery for Left Main Coronary Artery Disease. *N. Engl. J. Med.* **375,** 2223–2235 (2016).
2. Warlow, C., Farrell, B., Fraser, A., Sandercock, P. & Slattery, J. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: Final results of the MRC European Carotid Surgery Trial (ECST). *Lancet* **351,** 1379–1387 (1998).
3. Bhatt, D. L. *et al.* Cardiovascular Risk Reduction with Icosapent Ethyl for

Hypertriglyceridemia. *N. Engl. J. Med.* **380,** 11–22 (2019).

4. Schwartz, G. G. *et al.* Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome. *N. Engl. J. Med.* **379,** 2097–2107 (2018).

5. Sever, P. S. *et al.* Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial - Lipid Lowering Arm (ASCOT-LLA): A multicentre randomi. *Lancet* **361,** 1149–1158 (2003).

6. Dahlöf, B. *et al.* Prevention of cardiovascular events with an antihypertensive regimen of amlodipine adding perindopril as required versus atenolol adding bendroflumethiazide as required, in the Anglo-Scandinavian Cardiac Outcomes Trial-Blood Pressure Lowering Arm (ASCOT-B. *Lancet* **366,** 895–906 (2005).

7. Pfeffer, M. A. *et al.* Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme. *Lancet* **362,** 759–766 (2003).

8. Cox, D. R. Models and Life-Tables Regression. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **34,** 187–220 (1972).

9. Ng'andu, N. H. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat. Med.* **16,** 611–626 (1997).

10. Hess, K. R. Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Stat. Med.* **14,** 1707–1723 (1995).

11. Royston, P. & Parmar, M. K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med. Res. Methodol.* **13,** 152 (2013).

12. Andersen, P. K., Hansen, M. G. & Klein, J. Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations. *Lifetime Data Anal.* **10,** 335–350 (2004).

13. Stensrud, M. J., Aalen, J. M., Aalen, O. O. & Valberg, M. Limitations of hazard ratios in clinical trials. *Eur. Heart J.* **40,** 1378–1383 (2019).

14. Kjekshus, J. *et al.* Rosuvastatin in Older Patients with Systolic Heart Failure - The CORONA Trial. *N. Engl. J. Med.* **357,** 2248–2261 (2007).

15. Royston, P. & Parmar, M. K. B. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Med. Res. Methodol.* **16,** (2016).

16. Uno, H., Tian, L., Claggett, B. & Wei, L. J. A versatile test for equality of two survival functions based on weighted differences of Kaplan-Meier curves. *Stat. Med.* **34,** 3680–3695 (2015).

17. Pocock, S., Wang, D., Wilhelmsen, L. & Hennekens, C. H. The data monitoring experience in the Candesartan in Heart Failure Assessment of Reduction in Mortality and morbidity program. in *Data Monitoring in Clinical Trials: A Case Studies Approach* 166–175 (2006). doi:10.1007/0-387-30107-0_15

18. Mok, T. S. *et al.* Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma. *N. Engl. J. Med.* **361,** 947–957 (2009).

19. Ford, I., Norrie, J. & Ahmadi, S. Model inconsistency, illustrated by the cox proportional hazards model. *Stat. Med.* **14,** 735–746 (1995).

20. Dobson, J., Whitley, R. J., Pocock, S. & Monto, A. S. Oseltamivir treatment for influenza in adults: A meta-analysis of randomised controlled trials. *Lancet* **385,** 1729–1737 (2015).

21. Royston, P. & Parmar, M. K. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials* **15,** 314 (2014).

22. Velazquez, E. J. *et al.* STICH 1: Coronary-Artery Bypass Surgery in Patients with Left Ventricular Dysfunction. *N. Engl. J. Med.* **364,** 1607–1616 (2011).

TABLES AND FIGURES

Central illustration

Analysis	**Cox proportional hazards model**	**Accelerated failure time model**	**Milestone analysis**	**Restricted mean survival time (RMST)**
Effect measure	Hazard ratio	Time ratio	Difference in proportion with event at milestone time	Difference in event-free survival up to milestone time
Improved survival when effect measure	<1	>1	<0	>0
Pros	Uses all study data Powerful with proportional hazards	Uses all study data Powerful with a constant time ratio	No assumptions required	No assumptions required Powerful for early treatment effect
Cons	Lacks power for early effects May be difficult to interpret or uninformative when assumptions not met	Lacks power for early effects May be difficult to interpret or uninformative when assumptions not met	Data after milestone time ignored	Data after milestone time ignored Lacks power for delayed effects
Recommended with				
Proportional hazards	✓	✓	+ / -	+ / -
Early effect	+ / -	+ / -	+ / -	✓
Delayed effect	+ / -	+ / -	+ / -	✗
Diminishing effect	✗	✗	✗	✓

Table 1. Results of main estimation methods.

Study	**Treatments compared**	**Outcome studied**	**Cox proportional hazards**		**Accelerated failure time**		**Milestone analysis***		**Restricted mean survival time* (RMST)**	
			Hazard ratio (95% CI)	P-value	Time ratio (95% CI)	P-value	Difference in frequency of event (95% CI), %	P-value	Difference in RMST (95% CI), in days	P-value
ASCOT lipid lowering	Atorvastatin vs Placebo	Total coronary events	0.71 (0.59 to 0.86)	0.00058	1.51 (1.19 to 1.92)	0.00067	-1.29 (-2.06 to -0.52)	0.00096	7.6 (2.5 to 12.7)	0.0037
ASCOT BP lowering	Amlodipine vs Atenolol	Cardiovascular death	0.76 (0.65 to 0.90)	0.0012	1.29 (1.11 to 1.50)	0.001	-0.79 (-1.29 to -0.29)	0.0019	2.86 (-2.62 to 8.35)	0.31
CHARM	Candesartan vs Placebo	All-cause death	0.91 (0.83 to 1.00)	0.055	1.11 (1.01 to 1.21)	0.032	-1.95 (-3.85 to -0.05)	0.044	21.0 (8.7 to 33.4)	0.00084
EXCEL	PCI vs CABG	Death, stroke or MI	1.01 (0.80 to 1.28)	0.97	1.06 (0.50 to 2.26)	0.88	0.48 (-2.76 to 3.72)	0.77	18.3 (-11.1 to 47.8)	0.22

*Milestone and restricted mean survival time were performed at 3 years in the ASCOT-LLA, CHARM and EXCEL studies, and 5 years in the ASCOT-BPLA trial.

Table 2. Number of patients required to provide 80% power for each method, assuming the same pattern of survival observed in each study

		Patients required for 80% power using each analysis method*			
Study	**Patients actually recruited in each trial**	**Cox proportional hazards**	**Accelerated failure time**	**Milestone analysis**	**Restricted mean survival time (RMST)**
ASCOT lipid lowering*	10305	6825	6987	7416	9579
ASCOT-BP lowering*	19,257	13,948	14,484	15,743	>100,000
CHARM	7599	16,208	12,964	14678	5350
EXCEL	1905	>100,000	>100,000	>100,000	10,068

*Power calculated is for the major outcomes studied in this article, which were secondary outcomes in ASCOT-BPLA and ASCOT-LLA

Figure 1. Illustration of how each of the four primary analysis strategies relate to the Kaplan-Meier cumulative incidence curves.

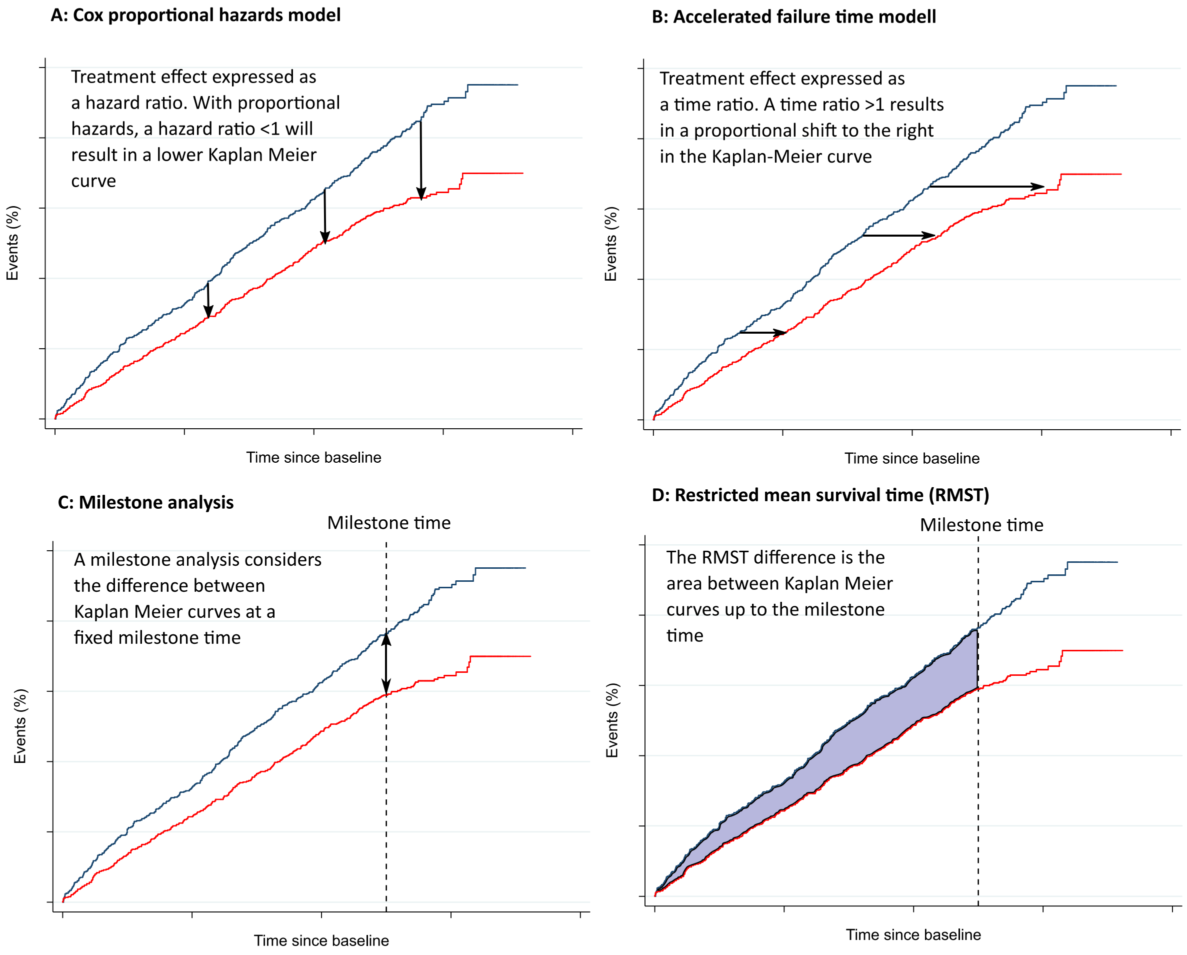


Figure 2. Kaplan-Meier cumulative incidence curves from each of the four cardiovascular trials

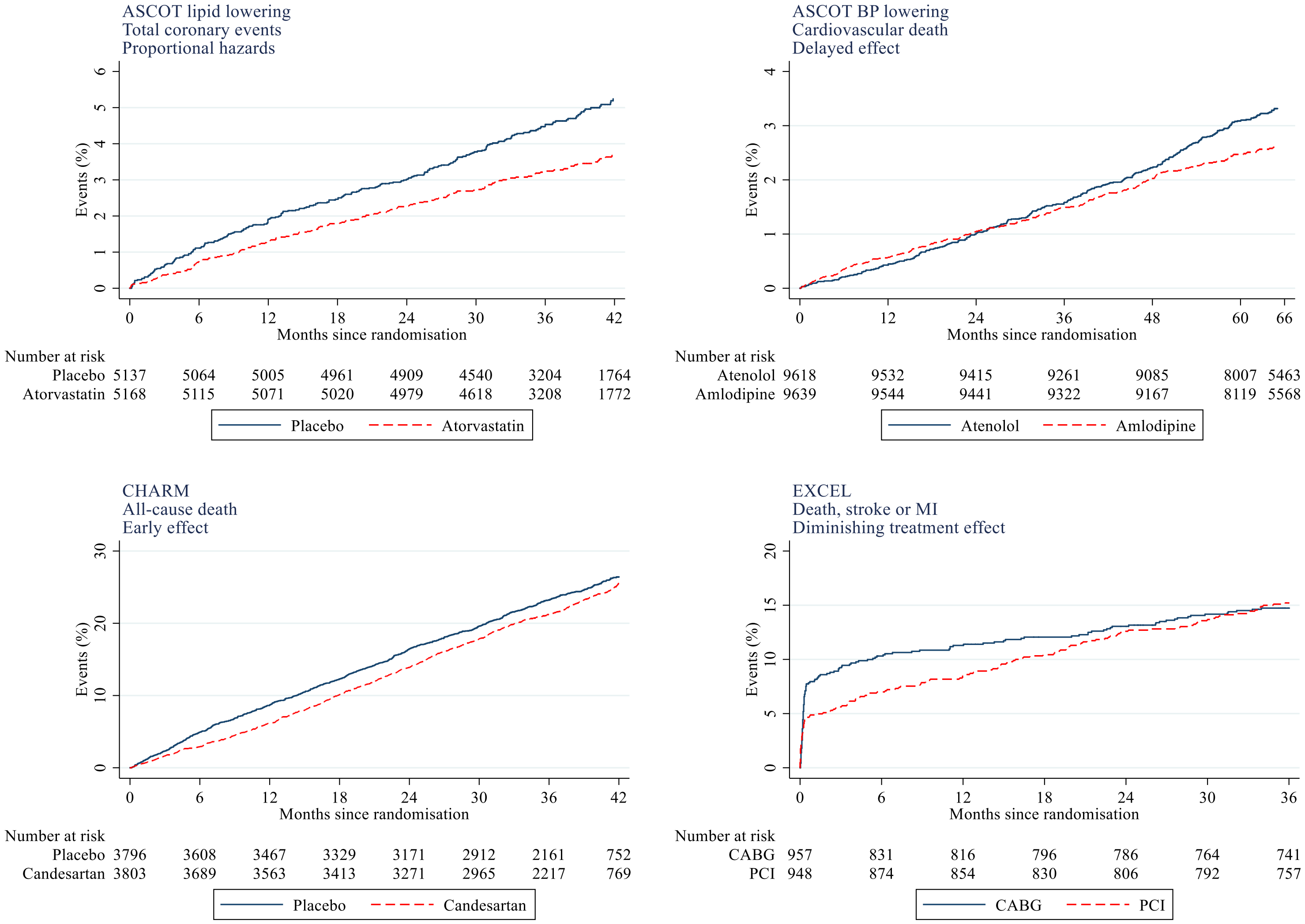
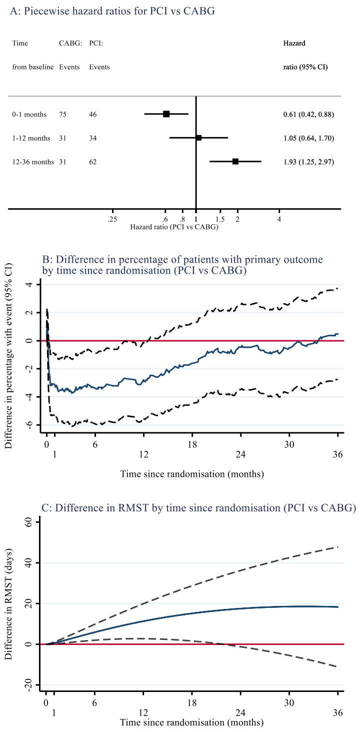


Figure 3: Hazard ratios from piecewise hazards models (A), difference in probability of event-free survival over time (B), and difference in restricted mean event-free survival (C) in the EXCEL trial.

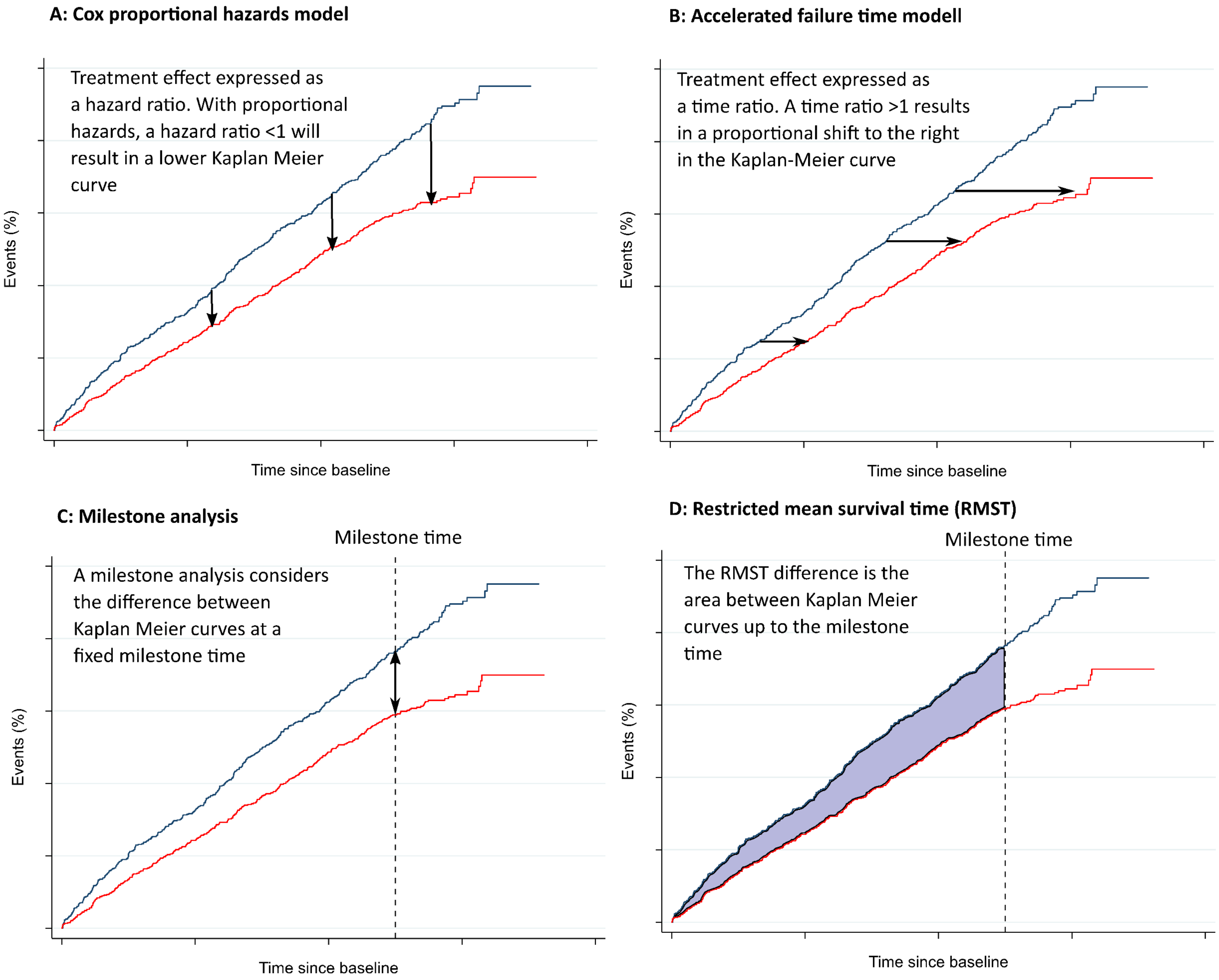


In panels B and C, the solid blue line represents the estimate and the black dashed lines are 95% confidence intervals.

TABLES AND FIGURES

Central illustration

Analysis	**Cox proportional hazards model**	**Accelerated failure time model**	**Milestone analysis**	**Restricted mean survival time (RMST)**
Effect measure	Hazard ratio	Time ratio	Difference in proportion with event at milestone time	Difference in event-free survival up to milestone time
Improved survival when effect measure	<1	>1	<0	>0
Pros	Uses all study data Powerful with proportional hazards	Uses all study data Powerful with a constant time ratio	No assumptions required	No assumptions required Powerful for early treatment effect
Cons	Lacks power for early effects May be difficult to interpret or uninformative when assumptions not met	Lacks power for early effects May be difficult to interpret or uninformative when assumptions not met	Data after milestone time ignored	Data after milestone time ignored Lacks power for delayed effects
Recommended with				
Proportional hazards	✓	✓	+ / -	+ / -
Early effect	+ / -	+ / -	+ / -	✓
Delayed effect	+ / -	+ / -	+ / -	✗
Diminishing effect	✗	✗	✗	✓



A: Cox proportional hazards model

Treatment effect expressed as a hazard ratio. With proportional hazards, a hazard ratio <1 will result in a lower Kaplan Meier curve

B: Accelerated failure time modell

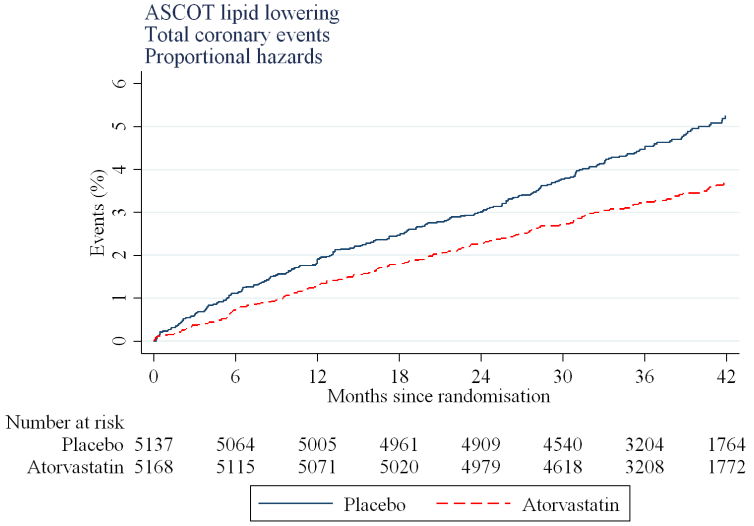
Treatment effect expressed as a time ratio. A time ratio >1 results in a proportional shift to the right in the Kaplan-Meier curve

C: Milestone analysis

A milestone analysis considers the difference between Kaplan Meier curves at a fixed milestone time

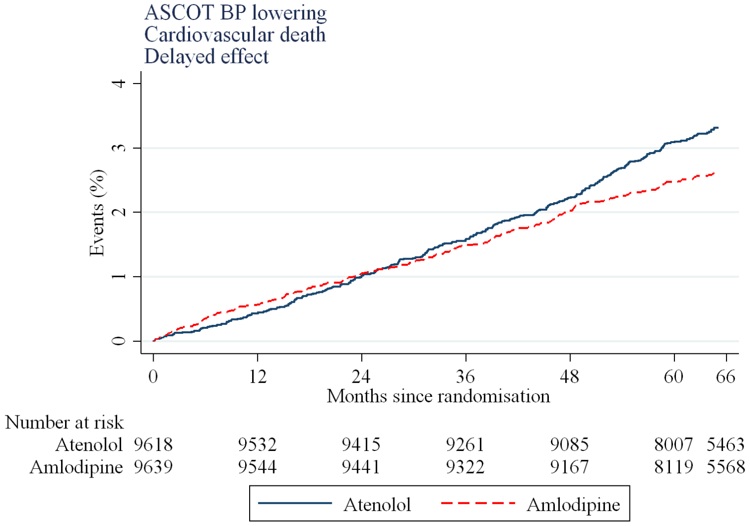
D: Restricted mean survival time (RMST)

The RMST difference is the area between Kaplan Meier curves up to the milestone time



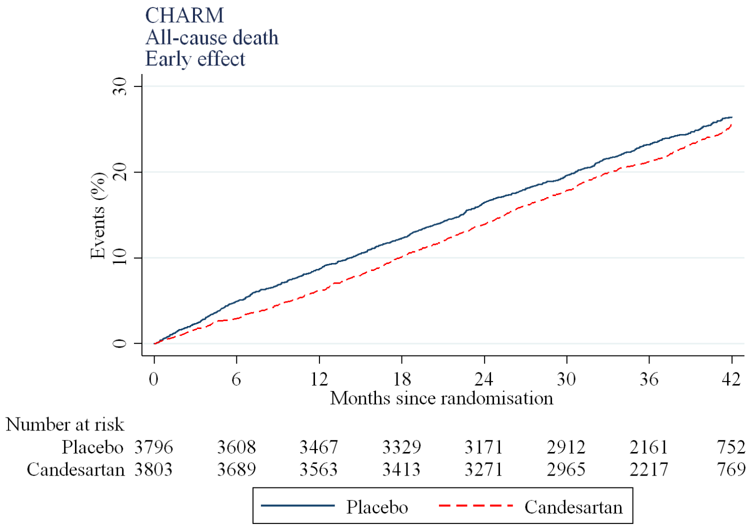
ASCOT lipid lowering
Total coronary events
Proportional hazards

Number at risk								
Placebo	5137	5064	5005	4961	4909	4540	3204	1764
Atorvastatin	5168	5115	5071	5020	4979	4618	3208	1772



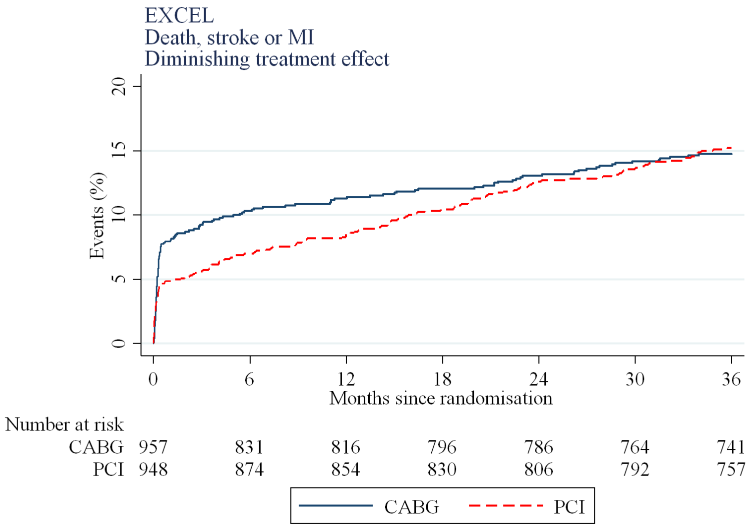
ASCOT BP lowering
Cardiovascular death
Delayed effect

Number at risk							
Atenolol	9618	9532	9415	9261	9085	8007	5463
Amlodipine	9639	9544	9441	9322	9167	8119	5568



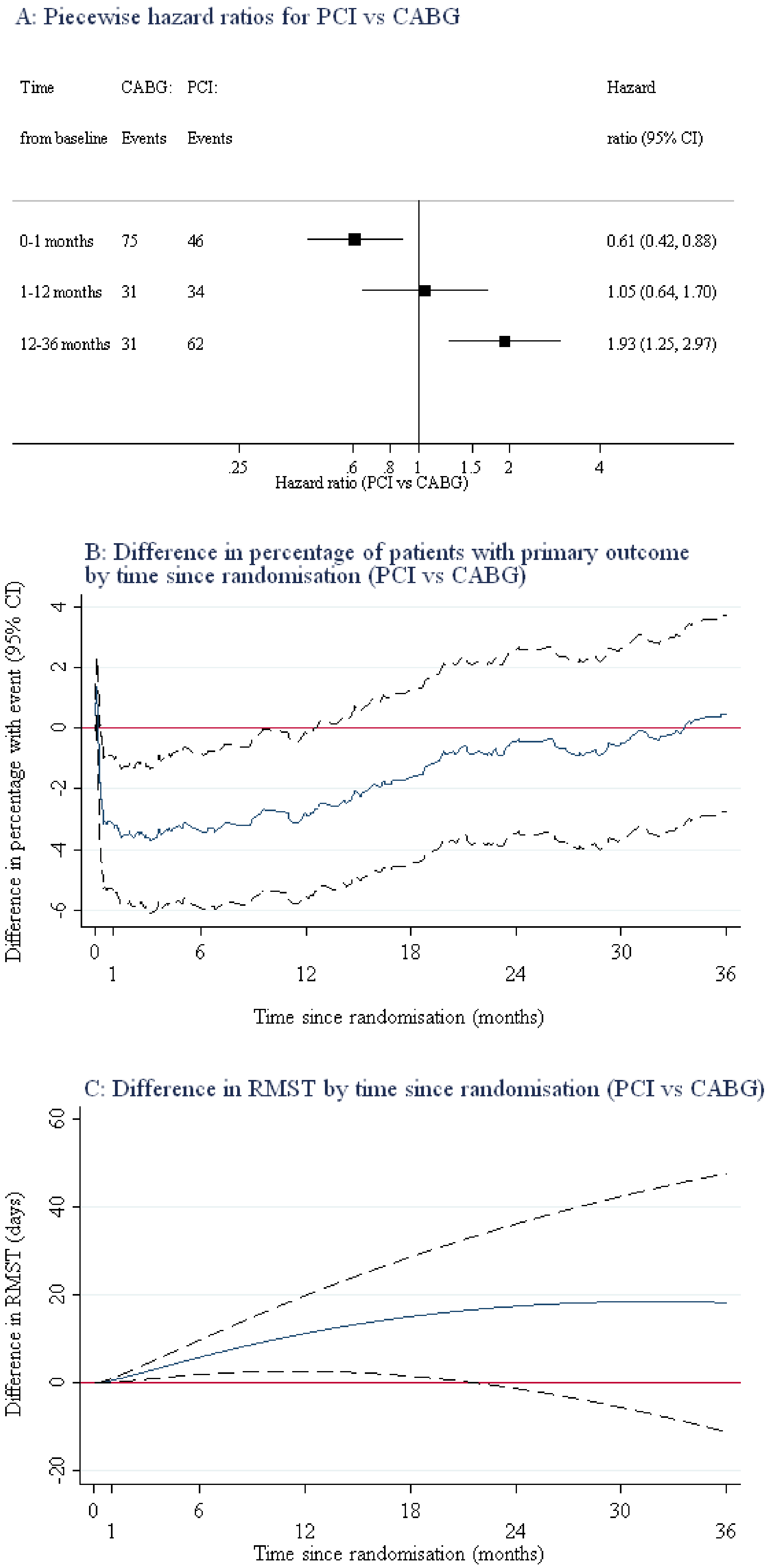
CHARM
All-cause death
Early effect

Number at risk								
Placebo	3796	3608	3467	3329	3171	2912	2161	752
Candesartan	3803	3689	3563	3413	3271	2965	2217	769



EXCEL
Death, stroke or MI
Diminishing treatment effect

Number at risk							
CABG	957	831	816	796	786	764	741
PCI	948	874	854	830	806	792	757



A: Piecewise hazard ratios for PCI vs CABG

Time from baseline	CABG: Events	PCI: Events	Hazard ratio (95% CI)
0-1 months	75	46	0.61 (0.42, 0.88)
1-12 months	31	34	1.05 (0.64, 1.70)
12-36 months	31	62	1.93 (1.25, 2.97)

B: Difference in percentage of patients with primary outcome by time since randomisation (PCI vs CABG)

C: Difference in RMST by time since randomisation (PCI vs CABG)