JECH Online First, published on April 12, 2016 as 10.1136/jech-2015-206656

Selection of the primary end point in an observational cohort study

Guadalupe Gómez,[1] Oleguer Plana-Ripoll,[2] Urania Dafni[3,4]

INTRODUCTION

Any study exploring specific research questions requires to test well-defined hypotheses that efficiently translate these questions into measurable quantities. To attribute differences in outcomes between two or more groups to their respective different exposure levels, the flagship comparative approach is the randomised clinical trial (RCT).

An RCT is not always feasible, as is the case if the exposure of interest cannot be reasonably randomised, such as smoking, age, family or personal history of a disease. A cohort study is the next most rigorous approach to answer cause and effect questions, sharing with an RCT the advantage of prospective follow-up. In addition, observational studies are more suitable to answer certain important questions for an intervention such as detecting rare or late adverse effects of treatments.[1]

A comparative study that lacks the randomisation component should not also lag behind in other rigorous requirements that are customarily expected of randomised studies. The value of the scientific evidence produced by an observational study lies equally heavily on how explicit its design characteristics are in the protocol, and how faithfully the study implementation followed the study protocol design. Guidelines to that effect are available and should be respected.[2]

One of the most crucial parameters in a study is the choice of the end point that would best translate the objective and capture the effect of interest. It is often the case that several relevant end points are of comparable importance and it might be difficult to select the most appropriate one. In those situations, the union of several end points, a composite end point (CE), is used as the primary end point (PE) of interest. A key advantage of a CE is that it provides a better description of a disease process.[3] For example, in the cardiovascular literature, efficacy of interventions is often expressed as a composite of major cardiovascular events (MACE). In the HORIZONS-AMI clinical trial,[4] two primary 30-day end points were prespecified: major bleeding and net adverse clinical events, a composite of *major bleeding* and MACE. In this trial, MACE is composed of *death*, *reinfarction*, *target vessel revascularisation for ischaemia* and *stroke*. While *major bleeding* is the relevant end point (RE), the CE takes into account all other additional adverse clinical events including death.

Several authors have discussed the advantages and disadvantages of using a CE from a clinical or statistical perspective.[3 5–8] A CE could give an appropriate reflection of the clinical spectrum of important outcomes associated with the disease. It avoids the need to choose a single PE when many may be of equal importance. A CE could avoid interpretational problems associated with competing risks by preventing an apparent benefit being attributed to a given event when in fact the benefit appears to be due to the increased occurrence of a more serious outcome, such as death. Finally, a CE can eliminate multiplicity problems associated with comparing treatments for several distinct end points. However, a misleading impression of the beneficial effect of a treatment can occur for a CE with component end points of highly differing clinical importance since the treatment might only benefit the less important end points. Also, the lack of treatment effect on the CE does not imply that there is no treatment effect on some of the components.[9]

Assume we have two possible end points, labelled E_1 and E_2, of comparable importance and could satisfactorily answer the study's primary clinical question. Suppose that E_2 is a secondary end point. We refer to E_1 as the RE and E_2 as the additional end point (AE), and each can be a simple or CE. The CE, E_*, is formally defined as the occurrence of either E_1 or E_2, and for time-to-event end points, time-to E_* corresponds to the time for the first event of either E_1 or E_2 to occur.

At the planning phase of a cohort study, the choice of the PE is of crucial importance: sample size (SS) computations are based on this PE and primary analyses also focus on the PE. The more efficient PE is the one requiring a smaller SS for the specified significance level and power. Under the premises of having to choose between a *relevant* E_1 and *composite* E_*, a measure quantifying which of the two end points would be more efficient is of great help. Such a measure has been derived for RCTs[10] and is based on the evaluation of the Asymptotic Relative Efficiency (ARE) of the two end points.

The purpose of the present paper is to show that the ARE method is a useful tool to guide the choice of the most efficient end point. The methodology is described for the design of observational cohort studies and is particularly appealing to confirm or refute previous findings, framed in clearly defined hypotheses about the benefit of an intervention.

THE ARE METHOD

The ARE method for RCTs

In an RCT, the research question of interest is usually whether a new treatment has higher efficacy than the currently recommended treatment. Consider a two-arm randomised study with assignment to an active or control treatment, for example, new treatment versus standard of care. Efficacy is sometimes measured by a PE expressed as time from randomisation to the occurrence of an event. Assume that we are establishing the treatment effect in terms of the HR and that the difference between treatment groups is tested by means of a logrank test. Our dilemma is whether to use E_1 as PE or the CE E*. Note that two different logrank tests would be used for E_1 and E*.

The ARE value between E* and E_1 is a measure of the relative power of the two tests and can be interpreted as the ratio of efficiencies using each outcome[10] or the ratio of required SSs to achieve a desired power.[11] ARE values larger than 1 indicate that the CE E* is more efficient, leading to a design with a smaller SS than E_1 and it would be recommended as the PE for the investigation.

The ARE method would not be of practical use if it could not be computed on the basis of easily interpretable, intuitive and plausible anticipated parameters. As shown in Gómez and Lagakos, the ARE value depends on the probability p_1 and p_2 of observing the end points E_1 and E_2 in the control group during the follow-up and the relative treatment effects on E_1 and E_2 given by the HRs HR_1 and HR_2.

[1]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain; [2]Section of Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark; [3]Frontier Science Foundation-Hellas, Athens, Greece; [4]Laboratory of Biostatistics, Department of Nursing, School of Health Sciences National and Kapodistrian University of Athens, Athens, Greece

Correspondence to Dr Guadalupe Gómez Melis, Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Carrer Jordi Girona 1-3, Barcelona 08034, Spain; lupe.gomez@upc.edu

The ARE method further requires independent censoring, the specification of the expected behaviour of the risk of developing E_1 and E_2 (decreasing, constant or increasing over time) and the proportion of individuals allocated to each treatment group, all standard assumptions required for the computation of the SS in a study. The dependence between the times to E_1 and to E_2 is also needed. However, it has been proven that different dependence structures lead to similar results[12] and different degrees of association, given by Spearman's r, yield, in many situations, to similar recommendations on the choice of the end point. A free web platform CompARE (https://cinna.upc.edu/compare) has been developed to choose between E* and E_1, based on the ARE value, in the framework of an RCT.

The ARE method for observational studies

Specifying the SS or the statistical power to detect differences between groups, during the design stage of an RCT, is crucial for financial and ethical reasons. This also holds true for observational studies, since a smaller SS implies shorter recruitment periods, reducing the cost and duration of a study. The STROBE reporting guidelines requires reporting of the key elements of study design and how the study size was arrived at.[2] Hence, the choice of the PE is undoubtedly a cornerstone, equally so in an RCT and in an observational cohort study.

The ARE method for an RCT is based on the logrank test, established to test the efficacy of a new treatment in comparison with a control treatment for a time–to-event end point. When comparing the survival experience of two groups with different baseline exposures in a cohort observational study through the Cox model,[13] the logrank test could still be used to compute the SS. However, it can be proved that the power obtained on the basis of the logrank test formula[14] overestimates the true power of non-randomised comparisons when the main exposure (X_1) and the covariates ($X_2,\ldots,X_k$) are correlated. SS formulae for some special cases in which X_1 and $X_2,\ldots,X_k$ are correlated are provided.[15] These formulae essentially correspond to Schoenfeld's formula[14] divided by $1-r^2$, that is, inflated (in prospective studies) by the squared multiple correlation coefficient r^2 of regressing treatment X_1 on the covariates $X_2,\ldots,X_k$. While in an RCT the null hypothesis of no treatment difference is established as the equality of the two hazard functions, in an observational study, conceptualised as a conditionally randomised experiment,[16] the null hypothesis would depend on the covariates. If a proportional hazards model is assumed, then the hazard, given a specific distribution of covariates $X=(x_1,x_2,\ldots,x_k)$, can be modelled as $h(t;X)=h_0(t)\exp(\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k)$. Consequently, the HR between exposed ($x_1=1$) and unexposed ($x_1=0$) is $HR=\exp(\beta_1)$. The null and alternative hypotheses would thus be respectively stated as H_0: $(\beta_1,\beta_2,\ldots,\beta_k)=(0,\beta_2,\ldots,\beta_k)$ and H_1: $(\beta_1,\beta_2,\ldots,\beta_k)=(\beta_*,\beta_2,\ldots,\beta_k)$. In this situation, power or SS calculations might require additional considerations.[17] As is often the case when computing the SS in an observational study, the marginal hazard functions replace the conditional hazard for every set of covariates. The 'crude' null hypothesis is thus used instead and, for every confounding covariate, an increase in the SS of about 10% is added.[18 19] Taking this into account, the ARE method is directly applicable to observational studies.

The role that censoring could play with respect to any combination of covariates has to be taken into account as well. Several approaches have been developed to address these issues, such as the estimation of the conditional probability of being censored, given the individual's covariates by the inverse probability weighting method[20 21] and the rank preserving structural failure time models.[22] When applying the ARE method to an observational study, the same censoring mechanism in the exposed and unexposed groups, given a specific distribution of covariates, is required (see online supplementary files for more information). The ARE method could accommodate different censoring distributions; however, the web platform CompARE is by now only developed for administrative censoring.

Under these premises, and assuming that the ARE value can be interpreted as a ratio of needed SSs,[11] an ARE value equal to 1.2 would mean that E_1 requires an SS, which is 20% higher than for E_*, to achieve the same power.

APPLICATION OF THE ARE METHOD

The application of the ARE method to observational cohort studies is illustrated using the results of a cohort published study as if they were the anticipated values the investigators would have had available when designing the study. Alcántara *et al*[23] recently aimed to study the effect of concurrent depression and stress in adults with coronary heart disease on the risk of suffering myocardial infarction (MI) or death. In a cohort of 4487 patients, 274 (6.1%) were exposed to high stress and high depressive symptoms and 3613 (80.5%) to low stress and low depressive symptoms. The remaining 600 individuals (13.4%), who were exposed to high stress and low depressive symptoms or *vice versa*, are not included in this example. Among these 3887 patients followed for 2.5 years, 408 (10.5%) experienced the CE of either MI or death. Specifically, 219 (5.6%) suffered an event of MI and 279 (6.8%) died during the follow-up.

The authors' main analysis was based on the results on the CE of MI or death. After adjusting for several potential confounders, participants with concurrent high stress and high depressive symptoms had a 48% higher risk of experiencing the CE compared to the unexposed group (HR=1.48, 95% CI (1.08 to 2.02)). Given the SS and characteristics of the study population, and considering a level of significance of 5%, the power to

Information about all the candidate endpoints for your trial

Candidate endpoint *E*	Terminating? (click if yes)	Probability of observing *E* in control group	Hazard Ratio	Type of endpoint	Definition of the composite
Myocardial Infarction	☐	0.055	1.12	Relevant component	☑
Death	☑	0.068	1.1	Additional component	☑

Figure 1 Screenshot from the free web platform CompARE (https://cinna.upc.edu/compare) with information from the relevant end point E_1 (myocardial infarction) and additional end point E_2 (all-cause death), its probabilities in the unexposed group (p_1 and p_2) and the relative treatment effect on E_1 given by HR_1.

consider this difference as statistically significant is 53%.[14]

In this study, the observed effect on the CE of MI and death was found to be significant. Let us explore the result if the initial plan of the authors would have been to investigate the effect of concurrent stress and depression on either MI or death alone. When using death as the main end point, the results would have been similar to the original ones. In fact, the authors found that exposed individuals had a 52% higher risk of death compared to those unexposed (HR=1.52, 95% CI (1.05 to 2.21)) and the corresponding power is 42%. On the other hand, the main finding of this study would have been of no association when considering MI alone. The authors found that the exposed group had a 12% higher risk of MI than the unexposed group (HR=1.12, 95% CI (0.70 to 1.79)) and the corresponding power is only 6%.

Using the ARE method, it is possible to anticipate whether the addition of death to MI would generally require a smaller SS to detect a specific difference, compared to considering MI alone or *vice versa*. In other words, the ARE method provides the outcome achieving a higher power, assuming a given SS, and thus allows to specify in advance which should be the main outcome of the study.

We have used the CompARE platform to illustrate these computations. To this end, we provide the expected anticipated risk for the RE (MI) of p_1=5.5% and AE (all-cause death) of p_2=6.8% in the unexposed group during the follow-up time, and the effect of the exposure on MI (HR=1.12) (figure 1). We obtain a plot for different correlations between MI and all-cause death and for different HRs on the all-cause death, illustrating different potential effects of the exposure on all-cause death (figure 2). The ARE method shows that it would have been more efficient to consider MI alone (ARE≤1) if the exposure would increase the death risk by less than 5%, that is, HR≤1.05, while it would have been more efficient to consider the CE of MI and death (ARE>1) for an increase in risk death of more than 5%, that is, HR>1.05. Since in this study the authors observed an increased death risk of 52% in the exposed group compared to the unexposed group, their consideration of the CE as the PE was the right decision, requiring a smaller SS than MI alone.

Analogously, it is possible to explore whether it is more efficient to use death alone or consider the CE of death or MI, obtaining a plot for different HRs on MI (figure 3). It would have been more efficient to consider death alone (ARE≤1) whenever the exposure increases the risk of MI by less than 20% (HR≤1.20).

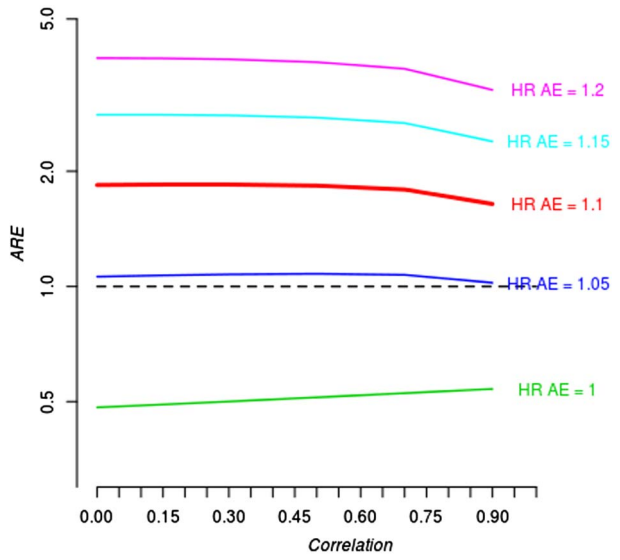

Figure 2 Asymptotic Relative Efficiency (ARE) of the composite end point E* (MI or all-cause death) versus the relevant end point E_1 (MI) for different values of Spearman's correlation coefficient and different effects of the exposure on the additional end point E_2 (all-cause death) given by HR_AE. The plots correspond to: expected anticipated risk for MI in the unexposed group during the follow-up time, p_1=5.5%; expected anticipated risk for all-cause death in the unexposed group during the follow-up time, p_2=6.8%; effect of the exposure on MI, HR_RE=1.12. MI, myocardial infarction; RE, relevant end point.

DISCUSSION

The choice of the PE is crucial during the design stage of any type of study. When several potential end points are of comparable importance to answer the research question of interest, other factors might be considered in order to choose the most suitable one among them. The ARE method provides a tool to make a more informed choice of the PE, given a small set of easily interpretable and anticipated parameters. This paper has shown how to use the ARE method in the design

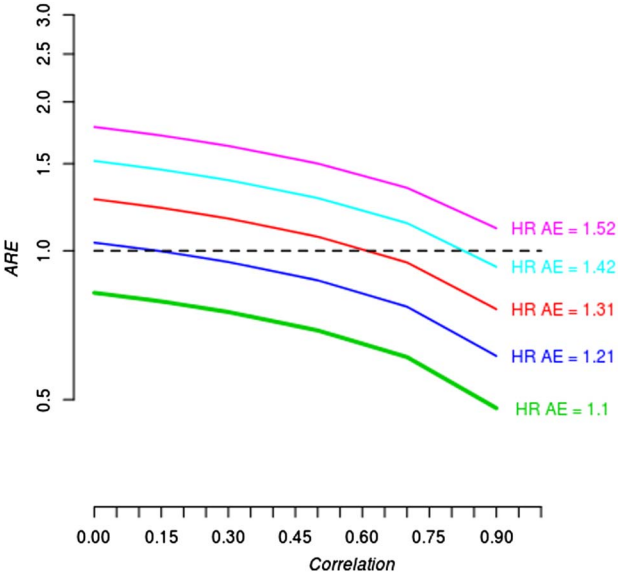

Figure 3 Asymptotic Relative Efficiency (ARE) of the composite end point E* (MI or all-cause death) versus death for different values of Spearman's correlation coefficient and different effects of the exposure on MI given by HR_AE. The plots correspond to: expected anticipated risk for MI in the unexposed group during the follow-up time, p_1=5.5%; expected anticipated risk for all-cause death in the unexposed group during the follow-up time, p_2=6.8%; effect of the exposure on death, HR_RE=1.52. MI, myocardial infarction; RE, relevant end point.

stage of an observational cohort study to choose the end point requiring a smaller study size. The CompARE platform could be used to explore the relative efficiency among multiple combinations of equally aligned end points.

The application of the ARE method to observational studies has limitations compared to RCT. First, the loss of randomisation makes the two treatment groups unlikely to be exchangeable and leads to the necessity to consider potential confounders. Second, the censoring mechanism in the exposed and the unexposed groups, given any combination of covariates, plays an important role in the computations of the ARE. Ideally, these factors would have to be taken into account when designing a study and we are working in this direction. Nevertheless, the most common approach when computing the necessary SS is to assume no confounding and therefore we are not imposing new limitations on the study design.

To summarise, researchers can use CompARE for observational studies to obtain plots, similar to the one reproduced in figure 2, to make a more informed decision on the choice of the PE in the study.

Acknowledgements The authors thank the Deputy Editor and the referees for all their comments which have resulted in a clear improvement of the manuscript. This work has been partially funded by Grants 2014 SGR 464 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya and MTM2012-38067-C02-01 and MTM2015-64465-C2-1R from Ministerio de Economia y Competitividad, Spain. Some of this work has been done during the Intensive Research Program in Statistical Advances for Complex Data held at Centre de Recerca de Matemàtica, Bellaterra.

Contributors GG is responsible for the overall content of the article, as well as for submitting and acting as the correspondent. GG, OP-R and UD are equal contributors to the planning, conduct, and reporting of the work described in the article, as well as to the manuscript writing.

Funding The study sponsors have not had any involvement in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; and in the decision to submit the paper for publication.

Competing interests None declared.

Provenance and peer review Commissioned; externally peer reviewed.

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jech-2015-206656).

To cite Gómez G, Plana-Ripoll O, Dafni U. *J Epidemiol Community Health* Published Online First: [*please include* Day Month Year] doi:10.1136/jech-2015-206656

J Epidemiol Community Health 2016;**0**:1–4. doi:10.1136/jech-2015-206656

REFERENCES

1 Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. *BMJ* 2004;328:39–41.
2 Elm EV, Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.
3 Freemantle N, Calvert M, Wood J, *et al*. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA* 2003;289:2554–9.
4 Stone GW, Witzenbichler B, Guagliumi G, *et al*. Bivalirudin during primary PCI in acute myocardial infarction. *N Engl J Med* 2008;358:2218–30.
5 Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A, *et al*. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;334:786.
6 Montori VM, Permanyer-Miralda G, Ferreira-González I, *et al*. Validity of composite end points in clinical trials. *BMJ* 2005;330:594–6.
7 Ferreira-González I, Permanyer-Miralda G, Busse JW, *et al*. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *J Clin Epidemiol* 2007;60:651–7.
8 Huque MF, Alosh M, Bhore R. Addressing multiplicity issues of a composite endpoint and its components in clinical trials. *J Biopharm Stat* 2011;21:610–34.
9 Gómez Melis G. Some theoretical thoughts when using a composite endpoint to prove the efficacy of a treatment. *In Proceedings of the 26th International Workshop on Statistical Modelling*, Edited by D. Conesa, A. Forte, A. López-Quílez and F. Muñoz. International Workshop on Statistical Modelling, 2011. http://hdl.handle.net/2117/22571
10 Gómez G, Lagakos SW. Statistical considerations when using a composite endpoint for comparing treatment groups. *Stat Med* 2013;32:719–38.
11 Gómez G, Gómez-Mateu M. The asymptotic relative efficiency and the ratio of sample sizes when testing two different null hypotheses. *SORT-Stat and Operations Res Transactions* 2014;38:73–88.
12 Plana-Ripoll O, Gómez G. Selecting the primary endpoint in a randomized clinical trial. The ARE Method. *J Biopharm Stat* 2015; doi:10.1080/10543406.2015.1094808
13 Tsiatis AA, Rosner GL, Tritchler DL. Group sequential tests with censored survival data adjusting for covariates. *Biometrika* 1985;72:365–73.
14 Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981;68:316–19.
15 Bernardo MVP, Lipsitz SR, Harrington DP, *et al*. Sample size calculations for failure time random variables in non-randomized studies. *J R Stat Soc Series D (The Statistician)* 2000;49:31–40.
16 Hernán MA, Robins JM. Causal Inference. http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/ (accessed 10 Dec 2015).
17 Velentgas P, Dreyer NA, Nourjah P, *et al*, eds. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. www.effectivehealthcare.ahrq.gov/Methods-OCER.cfm.
18 Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;13:356–65.
19 Breslow NE, Day NE. *Statistical methods in cancer research. Volume II—The design and analysis of cohort studies*. Lyon: International Agency for Research on Cancer, 1987.
20 Robins JM. *Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. Statistical Models in Epidemiology. The Environment and Clinical Trials*. New York: Springer-Verlag, 1999.
21 Robins JM. *Marginal structural models. In 1997 Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, 1998:1–10.
22 Tilling K, Sterne JAC, Didelez V. G-estimation for Accelerated Failure Time Models. In: Tu Y-K, Greenwood DC, eds. Modern Methods for Epidemiology. Dordrecht: Springer Science+Business Media; 2012.
23 Alcántara C, Muntner P, Edmondson D, *et al*. Perfect storm: concurrent stress and depressive symptoms increase risk of myocardial infarction or death. *Circ Cardiovasc Qual Outcomes* 2015;8:146–54.