HYPOTHESIS

How the type of risk reduction influences required sample sizes in randomised clinical trials

K Bruynesteyn, A Wanders, R Landewé, D van der Heijde

Ann Rheum Dis 2004;63:1368–1371. doi: 10.1136/ard.2003.014035

To increase change between groups, randomised clinical trials (RCT) often include patients with high risk for a particular outcome, by inclusion criteria that select predictors for that outcome. This increases the statistical power, and fewer patients are required for that RCT. The way in which patient selection influences the power, and thus sample size required, depends on how an intervention reduces the individual risk: by an absolute or relative risk reduction model.

To increase the contrast between an intervention and a control group, randomised clinical trials (RCTs) often include patients with a high prior risk for the outcome of interest by selecting according to baseline predictors for that outcome. The general view is that such a strategy limits the number of patients required, by increasing the statistical power. In RCTs in rheumatoid arthritis (RA), one important outcome is progression of radiological damage. It is known that joint damage present at baseline (prior risk) is one of the strongest predictors for further progression of radiological joint damage.^{1–4} Therefore, we expected fewer patients would be needed for an RCT if patients were selected for the presence of joint damage at baseline. Somewhat surprisingly, power calculations showed that this expectation was false. To elucidate this counterintuitive observation, we performed a literature search, and found that statistical power is determined by a critical relationship between the prior risk on a particular dichotomous outcome (this is the baseline risk, independent of treatment), and on how a particular treatment exerts its effect with respect to this prior risk. We will explain this relationship step by step, and illustrate it with hypothetical and authentic data.

ABSOLUTE AND RELATIVE RISK REDUCTION

The risk of an individual patient for a particular outcome can be reduced in two different ways, according to the following two models: the absolute risk reduction (ARR) model, and the relative risk reduction (RRR) model. To understand these models it is necessary to be familiar with the terms ARR and RRR.

Consider an RCT comparing an active treatment with placebo. The outcome of interest is binomial (*event or no event*) and the active treatment can reduce the probability of that event. So, the event

rate in the treatment group (P_t) is supposed to be lower than the event rate in the placebo group (P_p). ARR is the difference in the event rates between the placebo and the treatment groups ($P_p - P_t$). Relative risk is defined here as the ratio of the event rate in the treatment group and that in the placebo group (P_t/P_p). RRR is defined as the reduction of the event rate in the treatment group, in proportion to that in the placebo group, or: $(P_p - P_t)/P_p$. This is mathematically similar to $(1 - \text{the relative risk})$ —that is, $(1 - P_t/P_p)$. RRR is usually expressed as a percentage—that is, $(1 - (P_t/P_p) \times 100\%)$.

If a treatment works according to an RRR model, the RRR remains constant, irrespective of the prior risk, whereas the ARR varies with the prior risk. If a treatment works according to an ARR model, the ARR remains constant irrespective of prior risk and then the RRR varies with different prior risks.

APPLICATION OF RRR AND ARR IN THREE SUBGROUPS

Suppose that we distinguish three subgroups, defined by the prior risk on a particular outcome: one group with a low prior risk, one group with an intermediate prior risk, and one group with a high prior risk. Note that prior risk does not directly refer to the event rate in the control group, but rather to patient characteristics (such as age, sex, disease aetiology, concomitant conditions, or disease status) measured at baseline, and known for their ability to influence the probability of the outcome. Suppose also that without any (adequate) treatment the outcome will occur in 20%, 60%, and 90% of the patients in each group, respectively. Assume that treatment works purely according to the RRR model (table 1.1), providing an RRR of 50%. So, treatment will lower the event rate from 20% to 10% (50% reduction) in the low risk group, from 60% to 30% (50% reduction) in the intermediate risk group, and from 90% to 45% (50% reduction) in the high risk group. The ARR is now 10%, 30%, and 45%, respectively. These figures can be used to calculate estimated sample sizes for future clinical trials. We calculated the samples sizes for two sided statistical testing using the power calculator of the UCLA department of statistics with the two sample arcsine approximation of the binomial distribution, with α set at 0.05 and β at 0.20.⁵ From table 1.1

Abbreviations: ARR, absolute risk reduction; CsA, ciclosporin A, DMARDs, disease modifying antirheumatic drugs; MTX, methotrexate; RA, rheumatoid arthritis; RCT, randomised controlled trial; RRR, relative risk reduction; SSZ, sulfasalazine

See end of article for authors' affiliations

Correspondence to: Professor D van der Heijde, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands; dhe@sint.azm.nl

Accepted 23 May 2004
Published online first
1 July 2004

(sample sizes in each group) it is obvious that the baseline risk importantly influences the appropriate sample size. In other words: in this scenario, selecting patients at high risk will increase the statistical power of a trial.

Now assume that a treatment works purely according to the ARR model (table 1.2), providing an ARR of 10%. Based on the same prior risk percentages, treatment will lower the event rate from 20% to 10% (10% reduction) in the low risk group, from 60% to 50% (10% reduction) in the intermediate risk group, and from 90% to 80% (10% reduction) in the high risk group. The RRR is now 50% $((1 - (10\%/20\%)) \times 100)$ in the low risk group, 17% $((1 - (50\%/60\%)) \times 100)$ in the intermediate risk group, and 11% $((1 - (80\%/90\%)) \times 100)$ in the high risk group, respectively. If again sample sizes are estimated using these figures, it becomes clear that the same sample sizes are required in low and high risk groups to statistically demonstrate an ARR of 10%. But to demonstrate an ARR of 10% for the intermediate risk group, a much larger sample size is needed. This can be explained by the fact that for intermediate risk groups the probability of the occurrence of a negative outcome is almost similar to the probability of the occurrence of a positive outcome. Note that an intermediate risk group may either include patients with an intermediate prior risk only, or a mix of patients with a low, high, and intermediate risk.

To summarise: depending on how a treatment reduces a person's risk of a particular negative outcome (according to an ARR or an RRR model), the patient selection for the trial with respect to the prior risk may influence the statistical power, and accordingly the sample size required for that trial. If a treatment (mainly) acts according to the RRR model, trials including patients with prognostic variables for a negative outcome (high prior risk) will yield more statistical power, and require a lower sample size. If a treatment acts mainly according to the ARR model, selection of patients according to prognostic variables has a different effect on sample size. From a statistical point of view, it would be wise

to avoid groups of patients with an average prior risk on the outcome of approximately 50%. Trials with such patient groups provide less statistical power than trials with patient groups with a lower or a higher prior risk.

In the sparse literature that have evaluated the relationship between treatment effect and the prior risk, it is suggested that the RRR is constant across the usual spectrum of prior risks in the vast majority of treatments (that is, following the RRR model).⁶⁻⁷ Nevertheless exceptions have been described. A large study evaluating the stroke risk in patients treated with aspirin, showed a decreasing RRR by increasing prior risk.⁸ However, the ARR also decreased by increasing baseline risk, so actually a mix of both models seemed to be operative. Obviously, treatments do not always act strictly according the RRR or to the ARR model: mixed models can also be found.

PRIOR RISKS, MODELS OF RISK REDUCTION, AND TREATMENT EFFECT

To our best knowledge, in rheumatology no research has been performed to examine the relationship of prior risks and models of risk reduction with treatment effect. Recently, radiographic data of an RCT comparing methotrexate (MTX) + ciclosporin A (CsA) versus CsA monotherapy have become available.⁹ We used these data to investigate this relationship. Treatment effect was defined here as the reduction in the proportion of patients with the outcome—namely, radiographic progression greater than or equal to the median group level, at 1 year. The prior risk for radiographic progression (baseline risk) was based on the radiological damage at baseline, above or below the median. Table 1.3 shows that the RRR was not similar in the two baseline risk groups (50% in the low risk group v 37% in the high risk group): The addition of MTX to CsA appeared not to follow a pure RRR model, but rather an ARR model (the ARR was approximately similar in both risk groups (25% v 29%)).

We further explored the radiographic data of the COBRA trial (table 1.4), in order to confirm the phenomenon of a

Table 1 Example of the absolute and relative risk model illustrated with data of clinical trials

	Control (%)	Treatment (%)	RRR (%)	ARR (%)	Sample size in each group
<i>1 Example of the relative risk reduction model</i>					
Low baseline risk	20	10	50	10	195
Medium baseline risk	60	30	50	30	41
High baseline risk	90	45	50	45	14
<i>2 Example of the absolute risk reduction model</i>					
Low baseline risk	20	10	50	10	195
Medium baseline risk	60	50	17	10	387
High baseline risk	90	80	11	10	195
	CSA + placebo† (%)	CSA + MTX† (%)	RRR (%)	ARR (%)	Sample size in each group
<i>3 CSA + MTX compared with CSA alone</i>					
Low baseline risk*	50 (n = 28)	25 (n = 32)	50	25	57
High baseline risk	79 (n = 29)	50 (n = 26)	37	29	40
	SSZ† (%)	COBRA† (%)	RRR (%)	ARR (%)	Sample size in each group
<i>4 COBRA treatment compared with SSZ alone</i>					
Low baseline risk*	40 (n = 30)	18 (n = 38)	54	22	68
High baseline risk	82 (n = 34)	58 (n = 33)	30	25	51

RRR = $(1 - (\text{outcome rate treatment group divided by outcome rate controls}) \times 100)$; ARR = $(\text{outcome rate control group} - \text{outcome rate treatment group})$.
Sample size calculation based on $\alpha = 0.05$, $\beta = 0.20$, two sided.
CSA, ciclosporin; MTX, methotrexate; SSZ, sulfasalazine; COBRA treatment, step down prednisolone + MTX + SSZ.
*Risk groups based on baseline radiographic damage (below or above the median); †percentage of patients with radiographic progression larger than the median.

Table 2 CsA + MTX and COBRA studies

	Low baseline risk		High baseline risk	
	CsA+MTX	CsA+placebo	CsA+MTX	CsA+placebo
1 CsA + MTX study				
Progression				
Yes	8 (A)	14 (B)	13 (A)	23 (B)
No	24 (C)	14 (D)	13 (C)	6 (D)
Total	32 (A+C)	28 (B+D)	26 (A+C)	29 (B+D)
Relative risk (RR) for CsA+MTX v CsA+placebo	$= (8/32) : (14/28) = 0.50$		$= (13/26) : (23/29) = 0.63$	
Absolute risk reduction (ARR)	$= (14/28) - (8/32) = 0.25$		$= (23/29) - (13/26) = 0.29$	
Relative risk reduction (RRR)	$= (1 - (8/32) : (14/28)) = 0.50$		$= (1 - (13/26) : (23/29)) = 0.37$	
	Low baseline risk		High baseline risk	
	SSZ	COBRA	SSZ	COBRA
2 COBRA study				
Progression:				
Yes	12 (A)	7 (B)	28 (A)	19 (B)
No	18 (C)	31 (D)	6 (C)	14 (D)
Total	30 (A+C)	38 (B+D)	34 (A+C)	33 (B+D)
Relative risk (RR) for COBRA v SSZ	$= (7/38) : (12/30) = 0.46$		$= (19/33) : (28/34) = 0.70$	
Absolute risk reduction (ARR)	$= (12/30) - (7/38) = 0.22$		$= (28/34) - (19/33) = 0.25$	
Relative risk reduction (RRR)	$= (1 - (7/38) : (12/30)) = 0.54$		$= (1 - (19/33) : (28/34)) = 0.30$	

decreasing RRR by an increasing baseline risk. In the COBRA trial,¹⁰ combination treatment with step down prednisolone, MTX, and sulfasalazine (SSZ) was compared with SSZ monotherapy. We again defined treatment effect as the reduction in the number of patients with radiographic progression above the median level in 1 year.

Table 1.4 shows that the ARR (combination treatment, compared with SSZ alone) was similar in the two risk groups (22% v 25%), and that the RRR decreased by increasing baseline risk (54% v 30%).

We statistically compared the treatment effects in both subgroups of baseline damage, in order to test the null hypothesis that the relative risk reductions were similar (test of interaction) as recommended by Matthews and Altman, and Altman and Bland.^{11, 12} In neither of the studies could a significant interaction be demonstrated ($p = 0.29$ for the difference in treatment effects in the MTX + CsA study, and $p = 0.17$ for the difference in treatment effect in the COBRA study). For the test of interaction, as well as for an interpretation and discussion of the lack of statistical interaction, we refer to appendix 1.

These observations cannot of course be generalised to the effect of all disease modifying antirheumatic drugs (DMARDs) on radiological joint damage or on other outcome measures. However, if it is true that some DMARDs decrease radiological progression by an ARR rather than an RRR model, selection of patients with an average prior risk on radiological progression of about 50% should be avoided when an RCT with such a DMARD is carried out (table 1.2). But only a proportion of all patients with RA who eventually show radiological progression will have radiological joint damage at inclusion. This makes it difficult to selectively enrol patients with a high baseline risk, and the actual patient accrual may include patients with intermediate (~50%) rather than with high baseline risk.

FURTHER COMMENTS

We should like to make a few additional remarks. Firstly, a person who is developing a trial has to make a choice between aiming at a mixture of high, intermediate, and low risk patients, and focusing on just one category. For generalisability one may choose to include patients at all

Table 3 Comparing relative risks in two risk groups

	COBRA trial		CSA+MTX trial	
	Low baseline risk	High baseline risk	Low baseline risk	High baseline risk
1. RR	0.46	0.70	0.50	0.63
2. log (RR)*	-0.7765 (E ₁)	-0.3567 (E ₂)	-0.6931 (E ₁)	-0.4620 (E ₂)
3. SE (log (RR))†	0.4081	0.1692	0.3598	0.2178
<i>Difference between the logarithms of relative risks:</i>				
4. d (=E ₁ - E ₂)	-0.4198		-0.2311	
5. SE (d)	= $\sqrt{(0.4081^2+0.1692^2)}$ = 0.4419		0.4206	
6. CI (d)‡	= -0.4198 ± 1.96 × 0.4419, or: -1.2859 to 0.4463		-1.0555 to 0.5933	
7. Test of interaction	z = -0.4198/0.4419 = 0.95 (p = 0.17)		z = -0.55 (p = 0.29)	
<i>Ratio of relative risks (ratio):</i>				
8. Ratio = exp(d)	= exp (-0.4198) = 0.66		0.79	
9. CI (ratio)	= exp (-1.2859) to exp(0.4463), or: 0.28 to 1.56		0.35 to 1.81	

*Values obtained by taking the natural logarithm; †the SE (log (RR)) can be calculated as follows: $SE(\log(RR)) = \sqrt{(1/A - 1/(A+C) + 1/B - 1/(B+D))}$, in which A, B, C and D refer to the raw numbers in the 2×2 tables (tables 2.1 and 2.2); ‡CI, confidence interval.

types of risk. However, we showed here that this might lead to larger sample sizes. On the other hand, one should consider whether the preferred inclusion of high risk patients is feasible. If high risk patients are difficult to include for any reason, the argument of an appropriate recruitment rate may outweigh the argument of limited sample sizes by the selective inclusion of high risk patients.

Patient selection in RCTs is often based on characteristics that are predictive of a certain outcome. The aim of this report was partly to show that statistical power is dependent on the level of that prior risk, as well as on how treatment actually reduces that risk. This is a different approach from selecting patients on the individual likelihood of *responding* to a particular treatment. Selecting patients with a high probability of responding to a particular treatment will also result in a larger treatment effect and thereby an increasing ARR and RRR. This means more statistical power, and as a consequence, a smaller required sample size. Increased responsiveness at a group level in RCTs can be promoted in two different ways.¹³ The first is by taking measures that result in a general increase of a patient's compliance. Those who take their medicine (appropriately) might respond better than those who do not take their medicine (appropriately). The second way to promote responsiveness is by identifying subgroups of patients who are intrinsically responsive to the particular treatment. But patient characteristics that are predictive of a high response to any treatment have hardly been identified in rheumatology up to now. This will be an important research area in the future.

Our final remark refers to the models used to describe the relationship between the prior risk and the treatment effect. These models are based on discrete binomial outcome measures. The primary outcome measure of a trial can also be a continuous variable. Although the models cannot simply be translated to continuous measures, there are no arguments as to why treatment effects should act differently when measured on a continuous scale as compared with a dichotomous one.

We here conclude that patient selection according to factors predictive of the outcome that should be influenced by treatment has an impact on the statistical power of RCTs with dichotomous outcome measures. The precise direction of that impact depends not only on the level of prior risk but also on whether the ARR or the RRR remains constant irrespective of the prior risk. As a rule of thumb, statistical power is best guaranteed by selecting high risk patients, because this scenario omits the dependency of the type of risk reduction. Better insight into prediction of individual responsiveness may further increase statistical power and decrease required sample size.

Authors' affiliations

K Bruynesteyn, A Wanders, R Landewé, D van der Heijde, Department of Internal Medicine, Division of Rheumatology, University of Maastricht, Maastricht, The Netherlands

APPENDIX I

The statistical proof of whether an ARR or RRR model is operative is difficult and circumstantial. The statistical inference refers to the comparison of treatment effects across subgroups of an RCT. The null hypothesis is that the difference of relative risks in both subgroups is zero (the so-called test of interaction).

Treatment effects are represented here as relative risks (for example, the risk of having radiographic progression above the median when treated with CsA + MTX in relation to that risk when treated with CsA + placebo). The two subgroups are the low baseline risk group and the high baseline risk group.

In an ARR model, the absolute risk reductions are similar in both subgroups, whereas the relative risk reductions in both subgroups are different. The latter effect refers to interaction, and can be tested statistically, as shown below. The equivalence of absolute risk reductions is difficult to prove. Theoretically, the proof of different relative risk reductions does not suffice, because this does not prove that absolute risk reductions in both subgroups are similar.

Here we show the test of interaction performed in both clinical trials that we have used to corroborate the argument of the ARR model in our article. Tables 2.1 and 2.2 show the raw data of both trials, which are also shown in the article, but now tabulated in a different order, in order to improve readability with respect to the inferences below. Note that we have calculated the relative risk here as a starting point for the inferences. We have also added, as an illustration, the absolute and relative risk reductions calculated in the article.

Table 3 shows the statistical inferences necessary to test for interaction in both trials. The inference is complicated because of the logarithmic transformation (and the necessary back-transformation). The test of interaction is not statistically significant in either of the trials, indicating that the null hypothesis that both relative risk reductions are similar cannot be rejected (*or*: it is not proved that the relative risk reductions in both subgroups are different). Because the subgroup sizes are small (lack of statistical power), because a test of interaction is quite conservative, and because the difference of relative risk reductions is considerably higher than the difference of absolute risk reductions in both trials, the failure to prove interaction here does not disqualify our statement that an AAR rather than an RRR model is operative here. (Note in table 3 that the ratio of relative risks (sometimes abbreviated as RRR) is not similar to the relative risk reduction in this article (abbreviated by us as RRR).)

REFERENCES

- Uhlir T, Smedstad LM, Vaglum P, Moum T, Gerard N, Kvien TK. The course of rheumatoid arthritis and predictors of psychological, physical and radiographic outcome after 5 years of follow-up. *Rheumatology (Oxford)* 2000;**39**:732–41.
- Boers M, Kostense PJ, Verhoeven AC, van der Linden S. Inflammation and damage in an individual joint predict further damage in that joint in patients with early rheumatoid arthritis. *Arthritis Rheum* 2001;**44**:2242–6.
- Combe B, Dougados M, Goupille P, Cantagrel A, Eliaou JF, Sibilia J, et al. Prognostic factors for radiographic damage in early rheumatoid arthritis: a multiparameter prospective study. *Arthritis Rheum* 2001;**44**:1736–43.
- Drossaers-Bakker KW, Zwiderman AH, Vlieland TP, Van Zeelen D, Vos K, Breedveld FC, et al. Long-term outcome in rheumatoid arthritis: a simple algorithm of baseline parameters can predict radiographic damage, disability, and disease course at 12-year followup. *Arthritis Rheum* 2002;**47**:383–90.
- <http://calculators.stat.ucla.edu/powercalc/> (accessed July 2004).
- Osiri M, Suarez-Almazor ME, Wells GA, Robinson V, Tugwell P. Number needed to treat (NNT): implication in rheumatology clinical practice. *Ann Rheum Dis* 2003;**62**:316–21.
- McAlister FA. Commentary: relative treatment effects are consistent across the spectrum of underlying risks.usually. *Int J Epidemiol* 2002;**31**:76–7.
- Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet* 1995;**345**:1616–19.
- Gerards AH, Landewé RBM, Prins APA, Bruijn GAW, Goei Thè HS, Laan RFJM, et al. Cyclosporin A monotherapy versus cyclosporin A and methotrexate combination therapy in patients with early rheumatoid arthritis: a double blind randomised placebo controlled trial. *Ann Rheum Dis* 2003;**62**:291–6.
- Boers M, Verhoeven AC, Markuse HM, van de Laar M, Westhovens R, van Denderen JC, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet* 1997;**350**:309–18.
- Matthews JN, Altman DG. Statistical notes: Interaction 3: How to examine heterogeneity. *BMJ* 1996;**313**:862.
- Altman DG, Bland JM. Statistical notes: Interaction revisited: the difference between two estimates. *BMJ* 2003;**326**:219.
- Sackett DL. Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). *CMAJ* 2001;**165**:1226–37.