Most tumour necrosis factor inhibitor trials in rheumatology are undeservedly called ‘efficacy and safety’ trials: a survey of power considerations

Y. Yazici1, N. M. Adler2 and H. Yazici3

Objectives: Many randomized clinical trials (RCTs) are labelled efficacy and safety while due consideration for power is provided only for efficacy outcomes. This in turn necessitates a discussion of the inadequacy of sample size (type II error) for identifying harm. This is particularly important in RCTs of TNF inhibitors as harm related to these agents is still a matter of debate.

Methods: Pubmed was searched for all RCTs published examining TNF inhibitors in RA, PsA and AS. Only original study reports were surveyed for whether: (i) they were labelled as efficacy, safety or both; (ii) the methods sections included safety as a primary or secondary end point; (iii) power calculations were adequately explained; (iv) statistical tests of significance were given for harm; and finally (v) any discussion of type II error for harm was present.

Results: Of the 34 articles surveyed, 24 (71%) were labelled as efficacy and safety. Among these, 23 (96%) did not include safety as a formal primary or secondary end point. In only 2/24 (8%) power calculations were given for safety. Finally, in only 3/22 (14%) any discussion about the inadequate sample size (type II error) for detecting harm could be found.

Conclusions: Most reports of RCTs of TNF inhibitors in rheumatological diseases are inappropriately labelled as addressing efficacy and safety. Their lack of power in detecting harm is not adequately discussed, either.

KEY WORDS: Tumour necrosis factor inhibitors, Randomized clinical trial, Power calculation, Type II error, Harm.

Introduction

There is a trend for labelling randomized clinical trials (RCTs) as addressing ‘efficacy and safety’ at the same time while formal power calculations are provided only for efficacy outcomes. Safety information gathered from such trials may be incomplete and should thus be vocalized in the manuscripts. This consideration is particularly important in RCTs of TNF inhibitors as harm related to these agents is still a matter of debate.

The statistical power and the related sample size calculation needed to detect differences in outcome at defined levels of significance, if such differences truly exist, is an important consideration before embarking on an RCT. This holds equally true for efficacy and harm if a trial aims formally to look at both. Its calculation is based on the investigator’s pre-trial estimation of the degree of possible efficacy or harm of the drug to be tested and its comparator based on previous RCTs, observational studies and the like and should be briefly but explicitly described in the methods section of the related manuscript.

Another methodological consideration is the ‘type II error’, which is concluding that an association does not exist, which in reality does, between two parameters, but which cannot be statistically discerned due to small sample size. To conclude that no efficacy or harm has been noted during a trial should be discussed in the light of the possibility of this error.

Authors reviewed all RCT of TNF inhibitors to determine the frequency of labelling such trials as ‘efficacy’, ‘safety’ or ‘efficacy and safety’ trials, to review the information provided regarding power calculations for efficacy and safety, to determine whether safety was a formal primary or secondary trial end point and finally whether a type II error discussion in reporting adverse events (AEs) was present when necessary.

Methods

Pubmed was searched for all RCTs studying etanercept, infliximab and adalimumab in RA, PsA and AS patients in the English language, published before 31 December 2006. All extension studies, secondary analyses of previously published data were excluded and only original study reports were included. Two authors then collected the following information for each article. When there were discrepancies among the authors regarding the nature of the data extracted a consensus was reached by reviewing the study together. The following were tabulated:

(i) Titles, abstracts and the introduction sections of all articles were searched to determine if the authors referred to their RCT as a ‘safety’, ‘efficacy’, or ‘efficacy and safety’ trial.
(ii) The methods sections were surveyed to see whether ‘harm’ or ‘safety’ was designated as a primary or secondary end-point.
(iii) The components of a power calculation (as explained in [1]) were tabulated for each article:
   (a) Type I (α) error (the probability of falsely concluding a difference exists).
   (b) Type II (β) error (the probability of falsely concluding no difference which when subtracted from unity, gives statistical power).
   (c) The assumptions made regarding the expected differences in response among the active treatment arm and placebo/control arm (treatment effect).

The components (a) and (b) were the same for efficacy and safety issues while (c) was slightly different for the equivalency trials [2] where a non-inferiority or equivalence CI designation replaced treatment effect.

When a specific α-error was not provided for the power calculations, a broad statement like ‘a P-value of <0.05 was used in all statistical analysis’ was also considered as part of the power
calculations. Similarly, in analysing powering for safety, the designation of a minimum sample size to observe a single occurrence of an AE, as distinct from no occurrence, at a certain probability was also taken as giving the power calculation.

(iv) We listed whether each article provided formal statistics when giving AEs. These were tabulated separately for serious AEs (SAEs) and other AEs. An AE was tabulated as an SAE if so designated in the related manuscript. When an article tabulated ‘serious infections’ separately this was added to the otherwise SAEs.

(v) We finally sought if the discussion section of the article included consideration of a type II error as a reason for lack of differences in AE rates between active treatment and placebo arms of the RCT. Any statement alluding to the need for a larger study group to further assess safety was taken as a discussion of type II error.

Results

There were 34 articles fulfilling the search criteria (RA = 21, PsA = 5, AS = 8) [3–36]. These were published in leading rheumatology or general medicine journals with an impact factor >6.0 in 32/34 of the publications (ISI web of knowledge, 2006).

Twenty-four articles (71%) called themselves ‘efficacy and safety’ trials either in the title, abstract or introduction sections. In 5/24 (21%), this double labelling appeared in the title (Table 1). Of these 24 trials, only 1 had given safety as a primary or secondary trial end point. One more study, this time labelled as a safety trial, had given safety as a primary end point [36].

There were no power calculations to look at AEs in 22/24 of the so-called ‘efficacy and safety’ trials. Of these, 3/22 (14%) included any discussion of a type II error as a possible reason for lack of noted harm (Table 1).

Eleven articles gave statistics for SAEs. (Table 1). In 2/11 power calculations were available. Seven of these 11 had been labelled as efficacy and safety, three as efficacy and one as a safety trial. Out of the two articles that had included safety as a main outcome measure one gave statistics for the SAE observed. Out of the same 11 articles that gave statistics for SAE, in two papers this statistic showed a significant difference between the treatment groups. Of the remaining nine with no significant differences in SAE, only two provided discussion of a type II error for a possible reason for not observing a significant difference in the number of SAEs.

Twenty-four articles gave at least partial statistics for all AEs. In 18/24 (75%) discussion of a type II error was not available.

Finally, 6/33 (18%) of the manuscripts designated as efficacy or ‘efficacy and safety’ studies by their authors did not give power calculations related to the efficacy outcomes (Table 1). Of the 28 that gave information regarding power calculations, 19 (68%) articles gave all three necessary components [1].

Discussion

In this study, our focus was on the RCTs about TNF inhibitors mainly because they fall in the authors’ specialty of practice and there has been considerable debate about potential harm with these agents. On the other hand, there is no reason to suspect that the issues brought up are unique to RCTs with TNF inhibitors. The authors, however, are unaware of similar surveys as they relate to other agents neither in rheumatology nor in other specialties.

Even if published in prestigious journals, 71% of the reports on RCTs of TNF inhibitors in RA, PsA and AS patients called
themselves ‘efficacy and safety’ trials, while only one listed safety as a primary or secondary end point [13]. Only two RCTs gave power considerations for safety [13, 20].

The majority, 21/24 of the articles that were labelled as reporting both efficacy and safety neglected to mention type II error as an explanation for the lack of increased AEs. This paucity of a type II error discussion remained true also when it was specifically sought among the articles that had given formal statistics for both AE and SAE.

It can be said that an RTC can have an adequate power to scrutinize safety but the authors might have not performed formal power calculations. For the purposes of this discussion we tried to look into this issue more closely by doing post hoc power calculations (http://stat.ubc.ca/~rollin/stats/ssize/) among seven articles [9, 13, 14, 21, 25, 33, 35] with no significant differences in the number of SAEs between the study arms and with no type II error discussion. We estimated the minimum number of patients that needed to be studied to see any statistically significant difference in the number of said events. These calculations were based on the frequency of SAEs observed in the comparator/placebo arm with an arbitrarily taken minimum difference of ≥20% in the number of SAEs in the study drug arm, an α-error of 0.05 and a power of 80%, equal sample sizes and they were one tailed. In this scheme, the minimum number of patients needed to be studied per arm ranged from 105 to 5313 (median 1007). The lowest value of this range comes from Ref. [9] where no SAEs were reported in the comparator/placebo arm of 35 patients. For the remainder, the numbers needed to be studied were 608 or higher in either arm and obviously much outside the range of the patient numbers in the manuscripts surveyed. It is also to be noted that the sample size calculations in these analyses have been rather conservative. Smaller effect differences that would still be clinically quite important as an increase in SAEs and doing two-tailed analyses would further increase the quoted required sample sizes. In short, it was apparent that these seven articles that gave formal statistics for SAEs (all non-significant), and no type II error discussion were in reality not powered for safety as related to SAEs.

It was interesting to note that power calculations were also wanting for efficacy analyses where only 19/34 (56%) of the studies gave full information about a power calculation while another 6/34 (18%) gave no information.

It should further be noted that RCT cohorts are highly selective for lack of major comorbidities. Hence, even if these RCTs had been properly powered to look at safety as a primary outcome, the results would have to be interpreted with caution until Phase 4 post-marketing data were available due to the highly selective nature of the populations in RCTs [37, 38]. This is what has happened with infections, tuberculosis in particular, with TNF inhibitors. It was only from Phase 4 data that these problems became apparent and the necessary precautions to prevent them were developed [39].

RCTs of course report on the AEs seen during the trial and they should. However, these are basically efficacy trials and no conclusive interpretation about the safety of the medications tested can be drawn until post-marketing data or trials powered to look at safety issues are available. While it is incorrect to double label these RCTs as ‘safety’ trials, the authors also propose that the issue of the potential presence of a type II error regarding observed harm should be clearly brought out in the manuscripts. It is sobering to note that this issue was addressed in only three of 24 articles that did realise a type II error discussion were in reality not powered for safety as related to SAEs. Problems range from overestimating positive predictive value to innumeracy about numbers needed to treat [40]. What appears to some as self-evident probably is not so to many. Physicians cannot solely depend on what authors or sponsors or both tell them what numbers mean. These almost self-evident but frequently neglected, as this survey showed, particulars of an RCT if consistently given their due importance would help all involved parties to have a much clearer understanding of the true clinical benefits and harms of any new remedy at hand.

References


Disclosure statement: H.Y. has received unrestricted educational grants from Schering-Plough and travel support from Wyeth. Y.Y. has received consultancies and honoraria from Roche, Bristol-Myers-Squibb, Pfizer, Amgen, Centocor, Genentech and Celgene over the last 2 yrs. The other author has declared no conflicts of interest.
38. Sokka T, Pincus T. Most patients receiving routine care for rheumatoid arthritis in 2001 did not meet inclusion criteria for most recent clinical trials or american College of rheumatology criteria for remission. J Rheumatol 2003;30:1138–46.