ON LINE SUPPLEMENTARY MATERIAL

Challenges of TES

High potential for selection bias may be introduced as would be observed if for example, all patients move onto one arm (open label drug), which is a frequent approach. Even if randomisation is maintained however (with an active comparator arm), loss of generalisability (and therefore external validity) is introduced – with losses to follow-up and patients that respond poorly to the experimental agent as well as those that experience major adverse events are often withdrawn from RCTs. RCT participants who have suboptimal response or experience mild intolerance may complete the RCT but are less likely to continue participation in a TES. Analysis of a TES does not usually include such patients but solely focuses on those who enter the open-label extension period. By not taking this into account in the method of analysis, researchers tacitly assume that patients not entering a TES have outcomes similar to the TES patients. In other words, this resembles a 'study completers' analysis, with the inappropriate assumption that the outcome of subjects not entering (or completing) the TES is 'missing at random'[2]. The end result is usually a favourable, but biased picture of the long-term assessment of benefit and harm of the study drug. In addition, the description of the population comprising and progressing in a TES report is frequently unclear and incomplete, making it difficult to be able to inform decisions about practice[3]. This includes combining of patients newly starting the drug (having previously been randomised to the comparator arm) with those who have already been on it for the duration of the RCT.

Methods

The first Delphi questionnaire (see appendix A) was accompanied by an explanation of the purpose of the exercise and sent to all the members of the task force who were asked to respond within 2 months. It included 21 items based on the 7 main domains for evaluation. Each item included a variable number of questions; some required a "yes" or "no" answer and others asked for a level of agreement of a statement on a scale from 0 to 10 to be chosen. All items had space for additional comments. (See appendix A for the questionnaire used.) We calculated the proportion of respondents who answered yes and no to each question included in the questionnaire. The steering group (MHB, MB, LC) arbitrarily *a priori* decided that group acceptance would be defined as items in which ≥70% respondents responded similarly (for questions requiring 'yes or no' answers). Statements requiring marking of a level of agreement were accepted if a mean score of ≥ 7/10 was recorded. Statements were rejected if a mean score of < 7 was noted and/or if more than 4 responses of the task force of 22 under 5 were recorded.

The initial set of items was amended based on the analysis and comments of the first round. A second Delphi was subsequently developed (by LS-F and MHB/LC/MB as before) that included 26 items for rating (see appendix B). This second questionnaire and the results from the first questionnaire were sent by e-mail to the task force members. Participants were asked to respond, to a second web-based Delphi survey within 6 weeks and after 4 weeks, e-mail reminders were sent to any non-responders.

Agreement of items formed the basis of the recommendations. A final and third round of discussion was undertaken electronically to modify any of the statements following which a voting round was undertaken to determine whether consensus had been reached. Finally, the task force also agreed upon the formulation of a research agenda.

Timelines

Stage 1 Delphi exercise

Stage 1 of the Delphi survey (appendix A) was mailed to all participants in April 2011. All responses were received by June 2011. This stage comprised 20 items, which included formally accepting the domains for evaluation summarised above. Opportunity for comments to raise additional points for consideration was also included at the end of each item.

Stage 2 Delphi exercise

A second Delphi was developed following on from the analysis and comments of the first Delphi survey (appendix B). This was sent to all participants in August 2011. All responses were received by December 2011 and reported to the Task Force in January 2012. These were analysed and a further round of modifications and amendments were made to 2 particular items following electronic discussion. The final set of recommendations was produced and a request for voting on these was sent in March 2012. All 21 task force participants voted by May 2012, accepting the recommendations.

The task force agreed to seek formal industry and regulatory body input before submitting these recommendations (triggered by agreement in one of the Delphi questions– see results). Obtaining agreement from several industry companies as well as European Medical Agency took some time; with the annual EULAR congress then seen as an appropriate setting to convene a meeting. As will be noted, some modifications were subsequently made but all responses were not collated until Autumn 2013. All members of the task force re-approved the recommendations by way of confirmation prior to document submission.

Results

Definition of a TES

The initial definition, which achieved the highest mean (SD) score 7.7 (2.55); median (range) 9 (1-10), was selected and subsequently modified based on additional comments made by the participants. In the second Delphi survey, the following revised definition had final 100% agreement, "A TES is a study that follows all patients beyond a pre-specified trial period whether the trial was a) a placebo-controlled RCT with the possibility to cross-over to open-label experimental drug or b) a placebo-controlled RCT with the possibility to cross-over to usual care or c) an active comparator trial."

Industry input highlighted that this definition may exclude studies where cross-over to other treatments are included and was therefore modified to the following final definition of a TES: "*A TES is a study that follows all patients beyond a pre-specified trial period whether the trial was a) a placebo controlled RCT with or without the possibility to cross-over to open-label experimental drug or b) an active comparator trial.*"

Definition of the start of a TES

The starting point of a TES should be stated in the pre-specified protocol with clear justification; and should be at the point of exposure to the experimental drug of interest (100% acceptance). For the experimental randomised arm this will be start of the original RCT; whilst for those randomised to placebo/active comparator arm, this point will be on switching to experimental treatment (during RCT or at start of TES, see later and figure 1).

Minimum duration of a TES:

The committee could not reach consensus on whether or not a minimum length of a TES should be defined (68% agreement to define), as this would be determined by the research question. The task force therefore agreed not to define a minimum duration; nevertheless, the rationale for the length chosen should be stated in the pre-defined protocol with adequate justification (100% agreement).

Population for inclusion in a TES:

All but one of respondents agreed that the population of TES should not be stipulated in guidelines, as this would be determined by the individual research question. Ideally, however, it should include all patients included in the RCT, with the ability to separately report on patients who are of specific interest, for example, those in remission or low disease activity.

Minimal data items/outcomes

Table 2 from the main manuscript is included below, with the addition of individual mean/median agreement scores.

Nature of information	**Agreement, mean (SD) score, 1-10**	**Agreement, median (range) score, 1-10**
Progression from RCT to TES		
Progress of subjects at each stage from RCT ^start to TES* completion with:	8.7 (2.2)	10 (2-10)
A flow diagram detailing **absolute** numbers of subjects at each relevant time-point	9.9 (0.36)	10 (9-10)
Duration of active treatment	9.5 (0.65)	10 (8-10)
Time of last observation	9.5 (0.94)	10 (7-10)
Patient drop-outs		
All drop-outs detailed	9.1 (1.46)	10 (5-10)
The drop-out rates from each arm during the original RCT and the cross-over groups	9.3 (1.07)	10 (7-10)
Reason for exclusion from the TES if the patient discontinues the drug	9.5 (1.02)	10 (7-10)
Reason for cessation of follow-up	9.4 (1.0)	10 (7-10)
Specification of reasons for cessation of follow up other than adverse event or inefficacy as above, e.g. geographical or doctor related reasons	8.7 (1.82)	9.5 (4-10)
Outcomes		
Functional status at the time of inclusion in the TES if applicable	8.8 (1.67)	9.5 (4-10)
Functional status at last observation if applicable	8.3 (1.73)	8 (4-10)
Disease activity at the time of inclusion in the TES if applicable	9.3 (0.91)	10 (8-10)
Disease activity at last observation if applicable	9.4 (0.85)	10 (8-10)

For those patients entering the TES having achieved low disease activity or remission during the RCT, the sustainability of such disease states should be evaluated and made available	8.6 (1.22)	8 (7-10)
For those subjects that enter a TES not having achieved remission/acceptable disease activity state following the RCT, the number that achieve this during the TES should be reported – to determine whether longer drug exposure has the potential to improve disease state of such subjects further	8.3 (1.59)	8 (5-10)
Treatment		
The disease –related co-medication (DMARD#, corticosteroid) at each stage from RCT start to TES completion	7.2 (2.72)	8 (0-10)
Safety		
The serious adverse events and any outcome related to safety at each stage from RCT start to TES completion	8.3 (2.08)	8.5 (2-10)

Safety

- TES may identify new adverse effects that the original RCT was not able to detect due to greater cumulative drug exposure; mean (SD) 8.4 (1.65); median (range) 9 (4-10).
- TES may identify whether the *incidence* of known adverse effects changes with longer-term drug exposure; mean (SD) 7.5 (1.61); median (range) 7.5 (4-10).

- TES may confirm whether the *nature* of known adverse effects identified from the RCT changes with longer-term exposure; mean (SD) 7.6 (1.5); median (range 7.5 (5-10).
- TES are sub-optimal to detect rare safety events because they are not powered for this; mean (SD) 7.6 (2.71); median (range) 8 (0-10).
- TES are sub-optimal to detect rare safety events because they include a selected population (responders with likely no previous serious adverse events); mean (SD) 7.0 (2.75); median (range) 8 (1-10).

Efficacy: The task force agreed that the greater cumulative exposure of the active drug per patient in a TES might identify additional information on the drug's efficacy; mean (SD) 7.0 (2.14); median (range) 7.5 (3-10). Whilst definitions of relapse are currently not available and requires working on, if/when validated, a TES might allow evaluation of relapse including time to relapse and therefore sustainability of original disease control; mean (SD) 7.8 (1.42); median (range) 8 (5-10).

Additional data/outcomes

Possible additional outputs to safety and efficacy were explored. Economic evaluation of long-term treatment with the active drug may be possible if appropriate measures are recorded in the TES; mean (SD) 7.2 (2.33); median (range) 7.5 (0-10). The committee did not accept that a TES could accurately evaluate health-related quality of life; mean (SD) 6.6 (2.35); median (range) 7 (0-10), risk-benefit ratio and therefore overall advantage of the drug; mean (SD) 6.2 (2.96); median (range) 7 (0-10), or compliance; mean (SD) 6.1 (3.09); median (range) 7 (0-10).

Method of data analysis

Table 3 from the main manuscript is included below, with the addition of individual mean/median agreement score

Statement	Agreement, mean (SD) score	Agreement,median (range)
The null hypothesis should be stated at the start where appropriate	7.9 (2.16)	8.5 (4-10)
Multiple comparisons should be taken into account when determining the level of statistical significance	8.1 (1.7)	8.5 (5-10)
The null hypothesis should take account of the results of the original RCT^. Depending on the research question, the results of a RCT should be accommodated in the TES*	7.5 (2.19)	8 (1-10)
The report should comment on cumulative outcome analysis (beneficial and adverse events) maintaining the original trial groups i.e. from RCT start, not TES start to avoid reporting of only the sub-selected patient group that proceeds onto the TES	8.6 (1.54)	9 (5-10)
The selection bias associated with a TES population means meaningful non-inferiority/ superiority analysis would not be reliable. The report should focus on how data for sustained effect from the start to the end of TES period, within a single group or the difference between groups was analysed and whether there was any suggestion of increased effect (although this could not be subject to formal statistical testing).	9.4 (0.84)	10 (8-10)
The plan for subjects that drop out of a TES	8.9 (0.91)	9 (8-10)

should be specified to demonstrate sustained effect from the start to end of TES period. With reducing number of participants (the denominator), the proportion responding will artificially increase if/when the number of patients (numerator) responding stays the same.		
The analysis should include survival/retention rates on therapy explicitly reporting the number of patients at each milestone with reasons for change detailed.	8.9 (1.29)	9.5 (7-10)
A plan on how to analyse this should be included with both intent-to-treat (ITT) (denominator as original number entering RCT) and completer (those entering TES only) population analyses reported. A completer analysis should always be reported together with an ITT analysis.	9.3 (0.99)	10 (7-10)
The repeated measures analysis of the data from a TES in rheumatology should include the area under the curve of absolute disease activity (i.e. not dichotomous response/change) preferentially expressed as a score (e.g., DAS, SDAI, etc.)	7.3 (2.55)	8 (1-10)
A TES should preferably include hard endpoints (e.g. death, work disability, joint replacement surgery, hospital admission) from TES +/- linkages with other data sources	8.6 (1.28)	8 (7-10)

Frequency and nature of split reporting

The committee agreed that reporting frequency should not be specified for all TES since this depends on the research question. In addition, industry representatives highlighted the fact that data cuts and reports may be undertaken in response to regulatory considerations that may not be foreseen. It was agreed that the protocol of each TES should pre-specify the minimum frequency of reports to be written and the basis for them (purpose, outcomes, length of RCT); mean (SD) 8.2 (2.28); median (range) 9 (1-10). Regarding the nature of reports, it was agreed that the results of efficacy and safety of a TES should be reported together when all patients have reached the specified time point as applicable; [mean (SD) 8.8 (1.36); median (range) 9 (5-10). The credibility of split reporting (for example, one abstract on efficacy, one on safety, one on quality of life outcomes) is questionable and should be discouraged by abstract selection committees and journal editors; mean (SD) 8.7 (1.75); median (range) 10 (5-10). It was acknowledged following stakeholder feedback however that abstract length limitations can be significant, and that certain complex safety communications focusing on specific events of interest would suffer from requirement to include efficacy information. The above recommendation was modified to, "*the results of efficacy and safety of a TES should generally be reported together; abstract selection committees and journal editors should carefully consider reporting of efficacy alone before acceptance*".

Consent:

The group accepted that all of the subjects undergoing a RCT should be informed of the importance of long-term surveillance and be given the opportunity of entering in the long-term follow-up; mean (SD) 9.4 (0.85); median (range 10 (8-10). Subjects included in a TES should sign a new informed consent form for continuation of data collection; mean (SD) 7.6 (2.87); median (range) 8.5 (1-10). The patient representative emphasised the importance of the patient knowing when they have transitioned from the RCT to TES; hence would be in favour of them signing a new consent form both for continuation of the drug and for data collection at that time point; mean (SD) 7.6 (2.87); median (range) 8.5 (1-10). Although it was agreed that annually updating the consent of patients included in a TES was not necessary; mean (SD) 3.7 (4.4); median (range) 1.5 (0-10); particularly since each additional consent runs the risk of additional drop-out, the statement in the second survey that this need not occur only achieved a mean (SD) score of 6.2 (3.79); however, median (range) score of 8 (0-10) was noted. Nevertheless, the patient representative advised we should not recommend the need for consent to be annually updated.

These comprised: AbbVie, Bristol-Myers Squibb, Merck-Sharp & Dohme, Pfizer, Roche, UCB. Open discussion was subsequently held at a EULAR annual congress 2013 session (see discussion) where Dr Cesar de la Fuente Honrubia (Spanish Agency of Medicines and Medical Devices, European Medicines Agency) and Professor Paul Peter-Tak (Senior Vice President/Head, ImmunoInflammation, Glaxo SmithKline) provided regulatory and industry perspectives respectively. The following companies attended EULAR +/- actively participated in the discussion of the recommendations and provided feedback to the final document: AbbVie, Bristol-Myers Squibb, Merck-Sharp & Dohme, Pfizer, Roche.

Discussion

Outcome following interactive session at EULAR, Madrid 2013: Consensus on steps going forward

Whilst the participating stakeholders approved the recommendations, a fundamental question was raised for future consideration - should industry undertake open-label extension studies in the first place? Central to this is establishing what the objectives of a TES are and whether the TES is the most appropriate method and best use of resources. Similar issues to those detailed earlier questioning the validity of TES and the evidence that switching from placebo to active improves disease control were raised. The notion that it is important for those who have benefitted to stay on drug perhaps contradicts the justification for a study that assumes equipoise between the 2 treatment options. It may be more appropriate to offer the trial participant that benefitted to continue allocated treatment in a blinded fashion – thereby preserving the rigour of efficacy and safety data. The expectation of regulatory authorities was highlighted as a crucial factor driving the conduct of such studies; a large proportion of industry-sponsored TES are conducted as a result of requests made by the regulatory authorities during the registration process. This is not only to acquire longer-term efficacy data but often mainly to address specific safety concerns.

The meeting closed with a general agreement to pursue this field further; with the acknowledgment that any future discussion and consideration amongst the clinical academic community for change will rely on the engagement of both industry and regulatory authorities. This report is therefore seen as an initial phase of a wider initiative and a springboard for further development.

References

1. Balshem, H., Helfand M, Schunemann HJ, et al., GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology, 2011. **64**(4): p. 401-6.
2. Rubin, DB. Inference and missing data. Biometrika, 1976. **63**(3): p. 581-592.
3. Zwarenstein, M., Treweek S, Gagnier JJ, et al., Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ, 2008. **337**(nov11 2): p. a2390-a2390.
4. Jones, J. and D. Hunter, Consensus methods for medical and health services research. BMJ, 1995. **311**(7001): p. 376-80.