Evidence from crossover trials: empirical evaluation and comparison against parallel arm trials

Dimitrios N Lathyris,1 Thomas A Trikalinos1,2 and John PA Ioannidis1,2*

Accepted 20 December 2006

Background We aimed to evaluate empirically how crossover trial results are analysed in meta-analyses of randomized evidence and whether their results agree with parallel arm studies on the same questions.

Methods We used a systematic sample of Cochrane meta-analyses including crossover trials. We evaluated the methods of analysis for crossover results and compared the concordance of the estimated effect sizes in crossover vs parallel arm trials.

Results Of 334 screened reviews, 62 had crossover trials. Of those, 33 meta-analyses performed quantitative syntheses involving two-arm two-period crossover trials. There was large variability on how these trials were analysed; only one of the 33 meta-analyses stated that they used the data from both the first and second period with an appropriate paired approach. Nine meta-analyses used the first period data only and 14 gave no information at all on what they had done. Twenty-eight meta-analyses had both crossover (n = 137, sample size n = 7162) and parallel arm (n = 132, sample size n = 11 398) trials. Effect sizes correlated well with the two types of designs (r = 0.72). Differences on whether the summary effect had a P < 0.05 or not were common due to limited sample sizes. The summary relative odds ratio for parallel arm vs crossover designs for favourable outcomes was 0.87 (95% CI, 0.74–1.02).

Conclusions Crossover designs may contribute evidence in a fifth of systematic reviews, but few meta-analyses make use of their full data. The results of crossover trials tend to agree with those of parallel arm trials, although there was a trend for more conservative treatment effect estimates in parallel arm trials.

Keywords Crossover studies, parallel studies, meta-analysis, empirical evaluation

Randomized trials may be conducted with either a parallel arm design or variants of crossover designs. Parallel arm trials are far more popular in the literature. They are simpler to design and analyse and are generally feasible. We also have well-established and accepted methods for their meta-analysis.1 Crossover trials are less common, but not rare. An appraisal of issue 1, 2001 of the Cochrane Database of Systematic Reviews found that 184 out of 1000 reviews referred to crossover trials and 151 included data from crossover designs.2 A Medline search of crossover studies published between 2000 and 2003 in five major medical journals found 40 crossover trials.3 Crossover trials are difficult or inappropriate to apply for many research questions. Moreover, they represent relatively complex designs. In a two-treatment crossover study each patient may be perceived as his own control: ideally one might simply compare the effect of treatment of the first period to that of the second period. However, treatment-period interaction and carry over effects jeopardize the validity of such simple inferences. Dealing with these problems poses analytical challenges.4,5 There are several different approaches for the analysis of data from crossover trials. Diverse options apply to
the primary analysis of the data; the problem is further accentuated in meta-analyses, where information to perform some types of analyses may be missing in published reports. One may use the first period results only, but this will waste all the data after crossover and will drastically diminish the power to detect any intervention effects. Alternatively, one may use the first and second period data as two different parallel studies, but this approach is biased, because it ignores treatment-period interaction. Finally, one may use outcomes from paired analyses. However, this information is often missing from published reports. Then one can try to approximate paired analysis by using imputed correlation coefficients from previous studies with sufficient amount of data.

This lack of unanimity and use of suboptimal methods may introduce bias in the analysis of crossover studies. We considered that it would be useful to obtain empirical evidence on the handling of the data and the influence of data from crossover trials in meta-analyses of randomized evidence. We tried to answer the following questions. What is the relative contribution of crossover trials to randomized evidence for diverse medical interventions? How are crossover trials handled in current meta-analyses? Do the results of crossover trials agree with those of parallel studies in the same meta-analysis, or do they find strong or weaker treatment effects? To answer these questions we analysed data from a systematic sample of meta-analyses.

**Methods**

**Selection of meta-analyses, crossover studies and outcomes**

We perused every fifth review out of the 1669 included in the Cochrane Library Issue 2, 2003 (in the order listed in the Cochrane Database) to identify reviews which included crossover randomized trials in quantitative syntheses (meta-analyses). The spacing of the evaluated reviews was chosen arbitrarily. We excluded all non-randomized and pseudo-randomized trials; and selected only two treatments, two period crossover studies (AB/BA design). Comparisons of more than two treatments are rare and their interpretation is rather complicated anyhow. At a second step, we focused on meta-analyses with at least one parallel and at least one crossover trial so as to evaluate the concordance of results obtained with these two designs.

We examined both binary and continuous outcomes. In every systematic review we selected the primary outcome as stated by the reviewers. If a primary outcome was not clearly identified, we selected what we deemed to be the most clinically important outcome; if this was also not clear, we selected the one with the larger number of studies; if there were ties, we selected the one with the larger number of randomized patients; and then the meta-analysis with fewer zero cells in the 2×2 tables of the constituent studies (for binary outcomes). Whenever there were two or more independent meta-analyses from the same systematic review (e.g. on different, non-overlapping types of patients), we considered them as separate meta-analyses. Two investigators (D.L. and T.A.T.) independently selected eligible meta-analyses and crossover studies. Disagreements were referred to the third investigator (J.P.A.I.).

**Data extraction on primary trials**

For every eligible meta-analysis we extracted information on the specific comparison, and outcome, and the number of parallel and crossover trials. For each trial we recorded the first author and year of publication, the design (parallel vs. crossover), the way the meta-analyses utilized data from crossover trials [first period only, second period only, ‘combined’ data (and if so how combined), or no comment/unclear], the approach of the meta-analysis to the carry-over effect in each crossover trial, and information necessary for the calculation of the effect sizes (2×2 tables for binary outcomes, and number of patients, mean response and SD of the responses in each arm for continuous outcomes).

**Statistical methods**

**Meta-analyses**

Quantitative analyses were performed in meta-analyses that included both crossover and parallel studies. For binary outcomes we used the odds ratio (OR) as the metric of choice. Continuous outcomes were quantified using standardized mean differences (SMDs), as expressed by Hedges’s g metric. SMDs express the magnitude of the treatment effects relative to the within group standard deviations and can be used to synthesize studies that have quantified treatment effects in different scales.

All comparisons were coined so that the experimental treatment is compared vs the standard/older treatment or no treatment (or placebo) for a good, favourable outcome. For example, if the outcome used by the original trials and their meta-analysis was therapeutic success, we kept the data as is; however if it was therapeutic failure, then we took the complementary counts, i.e. we inversed the OR. If the outcome was measured on a continuous scale, where health was better with higher scores, we kept the data as is; if the scale meant worse health with higher scores, then we inversed the sign of the difference between arms in the calculation of the SMD. Therefore an OR > 1 or SMD > 0 suggests that the experimental treatment fares better than the comparator.

For every topic we calculated the summary effect size of crossover and parallel studies with both fixed and random effects inverse variance syntheses. Fixed effects analyses assume that the true effect of treatment is the same across synthesized studies, while random effects analyses allow for between-study heterogeneity (dissimilarity) and incorporate it in the calculations. For binary outcomes, if any trial arms in the analysed studies had zero observed events we added 0.5 to all cells of the pertinent two by two table.

We tested for between-study heterogeneity using the χ² distributed Q statistic and quantified its extent with the I² statistic. I² ranges between 0% and 100%, and values above 75% imply very high heterogeneity.

**Assessment of concordance**

We evaluated the agreement between parallel and crossover designs. Concordance was measured by assessing the Spearman
correlation coefficient for the effect size estimates between the two designs across all meta-analyses; how often the two designs showed effects in the same direction; how often the two designs agreed with respect to having $P < 0.05$ for the summary effect; and whether the effect size estimates of the two designs differed beyond chance.\textsuperscript{11} For the latter assessment, we calculated the relative odds ratio (ROR) by dividing the summary OR in parallel trials by the summary OR in crossover trials and also estimated the variance and 95\% confidence intervals (CIs) of the ROR.\textsuperscript{12} We transformed the weighted SMDs of the continuous outcomes to ORs\textsuperscript{13} using the following formulae: 
\[
\text{OR} = e^{(\sqrt{3}/3)\hat{g} + \text{se ln(OR)}} = (\sqrt{3}/3)\hat{\sigma} - \text{se}(g).
\]
When ROR > 1, parallel trials yield larger OR estimates than crossover trials, i.e. the parallel trials give a more favourable picture for the experimental intervention than the crossover trials, and vice versa when ROR < 1.

Finally, we assessed whether there was a systematic difference in the estimated summary effects between the two designs across all topics. The summary ROR was calculated across all topics with random effects syntheses and heterogeneity in the ROR estimates across meta-analyses was measured with the $Q$ and $I^2$ statistics.

**Subgroup and sensitivity analyses**

The main analyses included all topics, regardless of the way data from the crossover studies were reported to be analysed. In separate analyses we estimated the summary ROR across the two designs considering only the crossover studies where only data from the first period were used; and estimated the summary ROR considering all other crossover studies. We also performed a sensitivity analyses by excluding the largest meta-analysis that included almost half of the crossover studies in our empirical evaluation.

Analyses were conducted in Intercooled Stata 8.2 and in SPSS 13.0. All $P$-values are two tailed. Hypothesis-testing for between-study heterogeneity is done at $\alpha = 0.10$.\textsuperscript{14}

**Results**

**Eligible meta-analyses**

Figure 1 shows the flow of the systematic screening of the Cochrane Library. We screened 334 systematic reviews and excluded 272 of them, as they did not have any crossover trials. Of the remaining 62 systematic reviews, 26 were finally selected.\textsuperscript{15-40} These pertained to 28 independent meta-analyses, because one review contained eligible data for three different meta-analyses (on three different types of study populations).

**Analysis of data from crossover trials**

A variety of approaches were used to incorporate the results of crossover trials in the 28 meta-analyses that included both types of designs and 12 out of these 28 meta-analyses did not mention at all their approach towards crossover trials results. Nine meta-analyses used only the first period results of crossover studies. Three meta-analyses combined data from first and second period, but only one of them described the method of combination, which was the paired analysis proposed by the Cochrane group. One meta-analysis used data from the second period only. Finally, three meta-analyses did not have a consistent approach towards analysis of crossover trials, using first period data in some trials and data from both periods in others. Only four of the 28 meta-analyses discussed the problem of carry over effect: three of them assumed that there was no carry-over effect in the analysed crossover studies and one relied on the carry over test results of crossover studies as reported in their initial publication.

In the five meta-analyses that had included only crossover trials, two meta-analyses did not mention at all what they had done, one approximated a paired analysis as described in the Reviewers’ Handbook of the Cochrane Collaboration, one used data from the second period, and one combined data from both periods without clarifying precisely the analysis methods. Only one of these meta-analyses stated that it assumed that there was no carry-over effect to the analysed crossover studies, while the remaining four meta-analyses did not comment at all on the potential problem of carry-over effect.

**Characteristics of crossover and parallel arm trials**

In the 28 meta-analyses that included both types of designs (Table 1), there were 142 crossover trials. We excluded five crossover trials from two meta-analyses (four trials analysed...
### Table 1 Characteristics of included meta-analyses

<table>
<thead>
<tr>
<th>Meta-analysis (year)</th>
<th>Reference number CDSR</th>
<th>Interventions and disease</th>
<th>Outcome used in the meta-analysis</th>
<th>Nstudies (Patients)</th>
<th>Analysis of crossover trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Definition</td>
<td>ES &gt; 0</td>
<td>OR &gt; 1 favours experimental intervention</td>
</tr>
<tr>
<td>Continuous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gotzsche (2002)</td>
<td>15 CD000189</td>
<td>Short-term corticosteroids vs placebo for RA</td>
<td>Joint tenderness</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Suarez-Almazor (1997)</td>
<td>16 CD000957</td>
<td>MTX vs placebo for RA</td>
<td>Number of tender joints</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Ferreira (2001)</td>
<td>17 CD000998</td>
<td>Nutrition vs placebo/usual diet for stable COPD</td>
<td>Weight loss</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Pinelli (2001)</td>
<td>18 CD001071</td>
<td>Non-nutritive sucking vs control in preterm infants</td>
<td>Heart rate change</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White (2001)</td>
<td>19 CD001106</td>
<td>CPAP vs placebo for obstructive sleep apnea</td>
<td>Epworth Sleepiness Scale</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Pouste(2003)</td>
<td>20 CD001304</td>
<td>Low PA diet vs no diet for PKU</td>
<td>Blood PA level</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Suarez-Almazor (2000)</td>
<td>21 CD001461</td>
<td>Azathioprine vs placebo for RA</td>
<td>Number of tender joints</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Brton (2001)</td>
<td>22 CD001817</td>
<td>Diuretics vs placebo for non-intubated preterms with CLD</td>
<td>Change in compliance</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Lemyre (2001)</td>
<td>23 CD002272</td>
<td>NIPPV vs NCPAP for apnea of prematurity</td>
<td>PCO₂ at 4–6h</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Lonergan (2002)</td>
<td>24 CD002852</td>
<td>Haloperidol vs placebo for agitation in dementia</td>
<td>Change in agitation</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Jones (2001)</td>
<td>25 CD003537</td>
<td>Inhaled corticosteroid vs placebo for asthma and mild COPD</td>
<td>Osteocalcin</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Green (2001)</td>
<td>26 CD003686</td>
<td>NSAID vs placebo for lateral elbow pain</td>
<td>Pain VAS</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Riemsma (2003)</td>
<td>27 CD003688</td>
<td>Patient education vs control in RA</td>
<td>Change in pain score</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>McCrory (2003)</td>
<td>28 CD003900</td>
<td>Anticholinergics vs 2-agonists for COPD exacerbation</td>
<td>Change in FEV₁ at 90min</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Jurgens (2003)</td>
<td>29a CD004022</td>
<td>Low vs high Na diet (Caucasians, normal SBP)</td>
<td>Change in SBP</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Jurgens (2003)</td>
<td>29b CD004022</td>
<td>Low vs high Na diet (Caucasians, elevated SBP)</td>
<td>Change in SBP</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Jurgens (2003)</td>
<td>29c CD004022</td>
<td>Low vs high Na diet (African-Americans, elevated DBP)</td>
<td>Change in SBP</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Toweed (2002)</td>
<td>30 CD004257</td>
<td>NSAIDS vs acetaminophen for OA</td>
<td>Rest pain</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Binary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hughes (1999)</td>
<td>31 CD000057</td>
<td>Clomiphene citrate vs placebo for unexplained subfertility</td>
<td>Pregnancies per patient</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Cheine (2001)</td>
<td>32 CD000234</td>
<td>-blocker supplementation vs placebo in schizophrenia</td>
<td>Leaving the study early</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Clarke (2001)</td>
<td>33 CD000236</td>
<td>Pergoline vs bromocriptine for L-Dopa complications in PD</td>
<td>Proportion improved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Ortiz (1999)</td>
<td>34 CD000951</td>
<td>Folinic acid vs placebo for MTX side effects in RA</td>
<td>GI side effects</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Higgins (2003)</td>
<td>35 CD001015</td>
<td>Lecithin vs control for AD</td>
<td>Proportion deteriorated</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Wilfen (2000)</td>
<td>36 CD001133</td>
<td>Anticonvulsants vs placebo for all neuropathic pain</td>
<td>Proportion improved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Melhart (2000)</td>
<td>37 CD001218</td>
<td>True vs sham acupuncture for idiopathic headache</td>
<td>Proportion improved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Sultana (2000)</td>
<td>38 CD001944</td>
<td>Thioridazine vs typical neuroleptics for schizophrenia</td>
<td>Leaving the study early</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Proctor (2001)</td>
<td>39 CD002123</td>
<td>TENS vs placebo for primary dysmenorrhoea</td>
<td>Proportion with pain relief</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Hood (2001)</td>
<td>40 CD002901</td>
<td>Digitalis vs control for CHF (patients with sinus rhythm)</td>
<td>Clinical deterioration</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

AD, Alzheimer’s disease; CDSR, Cochrane database of systematic reviews number; CHF, congestive heart failure; CLD, chronic lung disease; COPD, chronic obstructive pulmonary disease; (N)CPAP, (nasal) continuous positive airway pressure; DBP, diastolic blood pressure; ES, effect size; h, hours; GI, gastrointestinal; MTX, methotrexate; NIPPV, nasal intermittent positive pressure ventilation; NSAID, non-steroid anti-inflammatory drugs; Nstudies, number of studies; OA, osteoarthritis; OR, odds ratio; PA, phenylalanine; PKU, phenylketonuria; PD, Parkinson’s disease; RA, rheumatoid arthritis; SBP, systolic blood pressure; TENS, transcutaneous electrical nerve stimulation; VAS, visual analogue scale.

Note: All meta-analyses are included in the Cochrane Library, Issue 2, 2003. The year next to the authors name refers to the last amendment of each meta-analysis. Outcomes are described as defined in the systematic reviews. For the quantitative analyses, all comparisons were coined so that the experimental treatment is compared vs the standard/older treatment or no treatment (or placebo) for a good, favourable outcome. Thus outcome definitions were reversed for entries with ‘No’ in the fifth column.
more than two groups and one was a partial crossover trial where many patients continued the first period’s treatment in the second period). Finally, we analysed 137 crossover studies with a total sample size of 7162 included in the analyses and 132 parallel studies with 11398 study subjects. In one topic (effects of low sodium diet vs high sodium diet on blood pressure, and biochemical indices) three different outcomes were analysed that included 91 crossover analyses with a total sample size of 4518 and 30 parallel studies with 4467 subjects. These trials represented a large proportion of the overall database and thus we performed sensitivity analyses excluding this topic.

A wide variety of diseases and interventions were represented. The most common targeted categories of diseases were joint diseases (n = 7, of which n = 5 for rheumatoid arthritis) and respiratory (n = 6, of which n = 3 for chronic obstructive pulmonary disease). Parallel trials had a larger analysed sample size than crossover trials (median 1990 vs 1991, P = 0.006). The crossover trials contributed anywhere between 5% and 79% of the total sample size of analysed subjects across these meta-analyses. In 16 of 28 meta-analyses, the total sample size in the parallel trials was larger than the sample size analysed in the crossover trials. There was no difference in the year of publication of crossover vs parallel trials (median 1990 vs 1991, P = 0.36). Of the 28 meta-analyses, 18 used continuous outcomes (Table 1).

Treatment effects and between-study heterogeneity
Summary SMDs expressed by Hedges’s g for continuous outcomes and summary OR for binary outcomes and I² for the extent of heterogeneity are shown in Table 2. These data refer to all studies and separately to crossover and parallel studies. Between-study heterogeneity did not seem to be clearly explained by differences between crossover and parallel trials. Only one meta-analysis found very-large between-study heterogeneity overall; one found very large between-study heterogeneity when separate analyses were performed for crossover trials; and two when separate analyses were performed for parallel arm trials.

Concordance between crossover and parallel trial results
The summary effect sizes for crossover trials were highly correlated to the summary effects in parallel trials across all 28 meta-analyses (Spearman correlation coefficient 0.72, P < 0.001). The results were similar, but based on more sparse data, separately for meta-analyses of continuous (ρ = 0.68, P = 0.002) and binary (ρ = 0.66, P = 0.030) outcomes.

In 23 of the 28 meta-analyses, the point summary estimates were in the same direction with both designs (15/18 continuous outcomes, 6/8 binary outcomes). When the overall results had P < 0.05, the two types of designs always found estimates of effects in the same direction.

In six meta-analyses, evidence from both designs gave P < 0.05 in the summary results, and in 12 meta-analyses both designs gave P’s ≥ 0.05. However, in 10 meta-analyses the two designs differed in the presence or not of P < 0.05 (in seven only the crossover data resulted in P < 0.05, in 3 only the parallel trials data resulted in P < 0.05). Overall the kappa coefficient was 0.27 (95% CI, −0.07 to 0.62).

ROR analyses
ROR analyses showed that the difference in the effect size observed in the two types of designs was beyond chance in only one of the 28 meta-analyses (Figure 2). This meta-analysis evaluated distal loop diuretics in preterm infants with chronic lung disease: there was only one small parallel trial (n = 21 infants) and one even smaller crossover trial (n = 10 infants, counted twice in the analysis [for the two periods]) that found opposite results. The crossover trial found even formally statistically results, despite its small sample size.

There was considerable variability in the point estimates of the ROR per meta-analysis and 12 of the ROR point estimates were either larger than two or smaller than 0.5 (Figure 2). However, the large majority of RORs had wide CIs, because usually the amount of evidence was not extensive with either or both designs. Overall, there was no between-meta-analysis heterogeneity in the ROR estimates (I² = 0.46, I² = 0%). The summary ROR was 0.87 and its 95% CI marginally encompassed 1.00 (ROR = 0.87, 95% CI 0.74–1.02), suggesting a trend for more conservative treatment effects in parallel arm trials than in crossover trials.

Exclusion of the largest meta-analysis yielded similar estimates, with a summary ROR of 0.88. The summary ROR estimates were similar in meta-analyses that clearly specified that they had used consistently only the first period data (ROR = 0.87) and the other meta-analyses (ROR = 0.87). The ROR estimates were 0.94 for binary and 0.86 for continuous data, respectively and these did not differ beyond chance (P = 0.74). Finally, the ROR was still 0.91 (95% CI 0.73–1.13) when limited to meta-analyses where the analysed sample size of the crossover designs was smaller than the sample size of the parallel arm trials (Table 3).

Discussion
Our empirical evaluation shows that crossover designs may contribute evidence on the effectiveness of medical interventions in about a fifth of systematic reviews. We found that their results correlate with those of parallel arm trials in broad terms. However, if anything, there was a trend for more conservative treatment effect estimates in parallel arm trials. This trend should be seen with caution. Perhaps more importantly, the majority of meta-analyses do not mention at all their approach towards the carry-over effect problem and none of them proceeded to reanalysis of data from all crossover periods with testing for carry-over effect. Systematic reviewers usually do not state how exactly they have handled the results of such studies, while a common practice is to use only the first period results, thus wasting the second period information.

For some disciplines, such as rheumatology, respiratory medicine and dermatology, crossover trials may be encountered quite commonly in the literature. This has been demonstrated also in empirical evaluations of large samples of trials in these disciplines.41–44 Sometimes medical interventions may have been evaluated only with crossover trials. It is well appreciated...
that crossover designs should not be performed for interventions that have substantial carry-over effects. One might reasonably question whether this prerequisite is fulfilled for several of the topics in Table 1, like for example the use of clomiphene citrate vs placebo for unexplained subfertility. Therefore, the caveats associated with the design and conduct of these designs should be better understood. This applies both to the design and conduct of the original trials and to their incorporation into systematic reviews and meta-analyses.

Although treatment effects correlated well between parallel arm and crossover trials, there was a trend for crossover trials to give more favourable results for the experimental intervention. The trend could be due to chance. Alternatively, it may reflect the fact that crossover trials tended to be smaller and it is common for small trials to give more impressive estimates of treatment effects than larger trials on the same topic.45 However, the number of trials per meta-analysis was too small to allow examining small study effects within each meta-analysis in a meaningful way. A second possibility is that crossover trials may have shorter follow-up than parallel trials and sometimes treatment effects may wane over time with longer follow-up.46 However, follow-up data were not available in sufficient detail to allow us to explore this explanation further. Overall, the difference between the two designs, even if truly present, does not seem large enough that it would suggest
that one type of design in particular yields consistently biased results.

Some other caveats should be discussed. The number of meta-analyses that we included in the ROR estimations is not large, so there was considerable accompanying uncertainty in the summary ROR estimates. Moreover, for the large majority of the meta-analyses included here, the amount of evidence was limited both for parallel and crossover studies. This is not atypical of randomized evidence in general in the medical sciences and it leaves considerable uncertainty about the credibility of the results obtained with either study design. Finally, we only examined meta-analyses published in the Cochrane Library. This may be an advantage, as these meta-analyses follow specific instructions and standards, but also a disadvantage, as the RevMan statistical package used in the Cochrane reviews cannot easily analyse crossover trials data.

Table 3 Summary ROR estimates

<table>
<thead>
<tr>
<th>Overall</th>
<th>N</th>
<th>ROR (95% CI)</th>
<th>I²% (P)Het</th>
</tr>
</thead>
<tbody>
<tr>
<td>All meta-analyses</td>
<td>28</td>
<td>0.87 (0.74–1.02)</td>
<td>0 (0.46)</td>
</tr>
<tr>
<td>Excluding CD004022</td>
<td>25</td>
<td>0.88 (0.65–1.18)</td>
<td>5 (0.40)</td>
</tr>
<tr>
<td>Outcome type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>18</td>
<td>0.86 (0.67–1.10)</td>
<td>20 (0.22)</td>
</tr>
<tr>
<td>Binary</td>
<td>10</td>
<td>0.94 (0.57–1.55)</td>
<td>0 (0.75)</td>
</tr>
<tr>
<td>Crossover trial analyses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First period only</td>
<td>9</td>
<td>0.87 (0.69–1.10)</td>
<td>13 (0.30)</td>
</tr>
<tr>
<td>Other/unclear</td>
<td>19</td>
<td>0.87 (0.57–1.34)</td>
<td>0 (0.60)</td>
</tr>
<tr>
<td>Total analysed sample size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larger in crossover trials</td>
<td>12</td>
<td>0.91 (0.73–1.13)</td>
<td>0 (0.66)</td>
</tr>
<tr>
<td>Larger in parallel trials</td>
<td>16</td>
<td>0.91 (0.66–1.30)</td>
<td>18 (0.25)</td>
</tr>
</tbody>
</table>

N: Number of trials.

Figure 2 Relative odds ratios (ROR) between crossover and parallel arm trials. RORs and 95% confidence intervals for each meta-analysis for the comparison of effect sizes in the two types of designs. Separate summary ROR estimates are provided for the continuous outcomes (top) and binary outcomes (bottom). The grand summary pertains to the summary ROR estimate across all meta-analyses. All comparisons were coined so that the experimental treatment is compared vs the standard/older treatment or no treatment (or placebo) for a good, favourable outcome. Confidence intervals that extend beyond the border of the graph have been marked with to crossing lines (like an ‘x’)
Surprisingly, only one of the analysed meta-analyses in our study used the approach proposed by Reviewers’ Handbook of the Cochrane Collaboration for incorporation of data of crossover studies in meta-analyses. It is difficult to extrapolate these findings to meta-analyses that have been published in scientific journals. Non-Cochrane meta-analyses may be quite differently reported compared with Cochrane reviews, partly because of space limitations. It would be inappropriate to consider the results of parallel arm trials as the perfect gold standard against which the crossover trials are judged—or vice versa. Thus agreement between the two study designs does not guarantee that the estimates are unbiased. In particular for small trials, there may be room for biased results with either type of designs. In all, our empirical evaluation highlights the need for better understanding and improved use of crossover designs and their data.

Conflict of interest: None declared.

KEY MESSAGES
- Crossover designs contribute evidence in about a fifth of systematic reviews in the Cochrane Library.
- Full use of crossover data is uncommon in meta-analyses.
- The observed effect sizes in crossover and parallel trials do not disagree beyond chance across different topics.
- However, there was a trend for more conservative treatment effect estimates in parallel arm trials.

References


