Responder analysis for pain relief and numbers needed to treat in a meta-analysis of etoricoxib osteoarthritis trials: bridging a gap between clinical trials and clinical practice

R A Moore, O A Moore, S Derry, P M Peloso, A R Gammaitoni, H Wang

ABSTRACT
Background: Population mean changes from clinical trials are difficult to apply to individuals in clinical practice. Responder analysis may be better, but needs validating for level of response and treatment duration.

Methods: The numbers of patients with pain relief over baseline (≥15%, ≥30%, ≥50%, ≥70%) at 2, 4, 8 and 12 weeks of treatment were obtained using the WOMAC 100 mm visual analogue pain subscale score for each treatment group in seven randomised placebo-controlled trials of etoricoxib in osteoarthritis lasting ≥6 weeks. Dropouts were assigned 0% improvement from baseline from then on. The numbers needed to treat (NNTs) were calculated at each level of response and time point.

Results: 3354 patients were treated with placebo, etoricoxib 30 mg and 60 mg, celecoxib 200 mg, naproxen 1000 mg or ibuprofen 2400 mg daily. Response rates fell with increasing pain relief: 60–80% experienced minimally important pain relief (≥15%), 50–60% moderate pain relief (≥30%), 40–50% substantial pain relief (≥50%) and 20–30% extensive pain relief (≥70%). NNTs for etoricoxib, celecoxib and naproxen were stable over 2–12 weeks. Ibuprofen showed lessening of effectiveness with time.

Conclusion: Responder rates and NNTs are reproducible for different levels of response over 12 weeks and have relevance for clinical practice at the individual patient level. An average 10 mm improvement in pain equates to almost one in two patients having substantial benefit.

Clinical trials are performed usually for regulatory purposes, with outcomes typically reported as statistical comparisons between treatment group population means. The results of clinical trials can be difficult to translate into clinical practice. A report that an intervention shows an average 10 mm reduction more than placebo on a 100 mm visual analogue scale has little immediate impact.

Moreover, few of us are average. Most drugs provide a good response in half or fewer of the patients treated, true in postoperative pain, neuropathic pain, migraine and tumour necrosis factor antagonists in rheumatoid arthritis. An 80/20 rule seems to apply in osteoarthritis, with 80% of patients experiencing 20% pain relief but only 20% experiencing 80% relief; about half have their pain halved.

Genetic influences help determine the clinical response to analgesic drugs for non-specific anti-inflammatory drugs (NSAIDs), opioids and more generally, as well as the clinical response to methotrexate. Pain is driven by complex pathways of neural mechanisms which are likely to be different between individuals. Imaging reveals loss of grey matter in chronic pain above that found with age alone. Average data from skewed distributions can produce misleading results. Dichotomous responder analyses have been reported previously for acute and chronic pain. The validity of a dichotomous measure should be established before being widely used.

An added factor contributing to differences in treatment response observed in clinical practice compared with a clinical trial is the handling of dropouts. Commonly, a “last observation carried forward” technique is used in clinical trials, where data from patients with good pain control but intolerable adverse events will still be included in efficacy calculations using the population mean. In clinical practice, this same patient would be considered a treatment failure.

We used individual patient data from seven randomised placebo-controlled trials in osteoarthritis to investigate the effects of different levels of pain relief assessed at various time points on estimates of efficacy.

METHODS
Merck Research Laboratories provided pain response data from seven randomised placebo-controlled trials of etoricoxib in osteoarthritis lasting ≥6 weeks (protocols 007, 018, 019, 071, 073, 076 and 077). PDF copies of the company clinical trial reports were also available.

We calculated the number of patients in each treatment group in each trial achieving various response data from patients with good pain control but intolerable adverse events will still be included in efficacy calculations using the population mean. In clinical practice, this same patient would be considered a treatment failure.

We used individual patient data from seven randomised placebo-controlled trials in osteoarthritis to investigate the effects of different levels of pain relief assessed at various time points on estimates of efficacy.

In each study patients were asked, “During the last 48 hours, how much pain do you have? 1 (no pain) to 100 (extreme pain). The
Western Ontario and Mcmasters Universities (WOMAC). 100 mm visual analogue pain subscale score was calculated as the average of the responses to the five questions.

Criteria used in defining responders included:

- For patients who did not drop out, only actual measured values were used for calculations. Last observation carried forward was not used.
- For patients who withdrew for any reason, measurements made within 7 days of the last dose were used to calculate the response.
- Thereafter, patients were assigned 0% improvement.

We calculated the number and percentage of responders for each level of response for each drug and time point and the number needed to treat (NNT) compared with placebo (with 95% CI).28 The relative risk with 95% CI was calculated using the z test, comparing different drug/dose combinations only in the trials in which they were used together.

Clinical trial reports were used to obtain, for each active treatment, the difference between active treatment and placebo for the WOMAC pain subscale score. This was defined in the company clinical trial reports as the mean time-weighted WOMAC treatment average) for the WOMAC pain subscale score. This was defined in the criteria used in defining responders included:

- For patients who did not drop out, only actual measured values were used for calculations. Last observation carried forward was not used.
- For patients who withdrew for any reason, measurements made within 7 days of the last dose were used to calculate the response.
- Thereafter, patients were assigned 0% improvement.

For patients who did not drop out, only actual measured values were used for calculations. Last observation carried forward was not used.

Clinical trial reports were used to obtain, for each active treatment, the difference between active treatment and placebo for the WOMAC pain subscale score. This was defined in the company clinical trial reports as the mean time-weighted WOMAC treatment average) for the WOMAC pain subscale score. This was defined in the criteria used in defining responders included:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Percentage of responders with treatment/placebo after different numbers of weeks</th>
<th>NNT (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Etoricoxib 30 mg; 5 trials, 1486 patients</td>
<td>&gt;15% 70/41 70/45 70/43 68/45</td>
<td>3.5 (2.9 to 4.2)</td>
</tr>
<tr>
<td></td>
<td>&gt;30% 54/27 56/30 59/31 59/36</td>
<td>3.7 (3.1 to 4.5)</td>
</tr>
<tr>
<td></td>
<td>&gt;50% 35/13 40/18 44/23 44/24</td>
<td>4.6 (3.8 to 5.6)</td>
</tr>
<tr>
<td></td>
<td>&gt;70% 17/5 22/9 26/13 26/17</td>
<td>8.9 (7.0 to 12)</td>
</tr>
<tr>
<td>Etoricoxib 60 mg; 3 trials, 711 patients</td>
<td>&gt;15% 77/46 76/51 73/44 71/47</td>
<td>3.3 (2.6 to 4.5)</td>
</tr>
<tr>
<td></td>
<td>&gt;30% 61/31 63/32 60/30 59/35</td>
<td>3.2 (2.6 to 4.6)</td>
</tr>
<tr>
<td></td>
<td>&gt;50% 43/16 46/19 47/20 44/23</td>
<td>3.7 (2.9 to 5.0)</td>
</tr>
<tr>
<td></td>
<td>&gt;70% 25/4 27/9 30/11 30/15</td>
<td>4.7 (3.9 to 6.0)</td>
</tr>
<tr>
<td>Celecoxib 200 mg; 2 trials, 714 patients</td>
<td>&gt;15% 64/42 67/45 64/40 64/40</td>
<td>4.6 (3.4 to 7.0)</td>
</tr>
<tr>
<td></td>
<td>&gt;30% 53/27 54/29 54/30 53/31</td>
<td>3.9 (3.0 to 5.3)</td>
</tr>
<tr>
<td></td>
<td>&gt;50% 30/12 39/18 39/22 39/22</td>
<td>5.6 (4.2 to 8.3)</td>
</tr>
<tr>
<td></td>
<td>&gt;70% 15/6 22/10 24/13 26/16</td>
<td>10.1 (7.1 to 18)</td>
</tr>
<tr>
<td>Naproxen 1000 mg; 2 trials, 531 patients</td>
<td>&gt;15% 78/54 75/57 70/48 64/40</td>
<td>4.2 (3.0 to 7.5)</td>
</tr>
<tr>
<td></td>
<td>&gt;30% 63/35 65/36 60/35 55/35</td>
<td>3.5 (2.6 to 5.5)</td>
</tr>
<tr>
<td></td>
<td>&gt;50% 44/20 47/22 45/22 44/23</td>
<td>4.1 (3.0 to 6.3)</td>
</tr>
<tr>
<td></td>
<td>&gt;70% 22/5 27/10 28/12 27/15</td>
<td>5.9 (4.4 to 8.8)</td>
</tr>
<tr>
<td>Ibuprofen 2400 mg; 2 trials, 618 patients</td>
<td>&gt;15% 69/43 65/47 60/48 58/49</td>
<td>3.8 (2.9 to 5.5)</td>
</tr>
<tr>
<td></td>
<td>&gt;30% 48/29 51/33 51/26 49/41</td>
<td>5.0 (3.6 to 8.1)</td>
</tr>
<tr>
<td></td>
<td>&gt;50% 28/14 35/20 40/24 39/27</td>
<td>6.5 (4.6 to 11)</td>
</tr>
<tr>
<td></td>
<td>&gt;70% 16/6 20/8 22/13 26/18</td>
<td>10 (7.0 to 21)</td>
</tr>
</tbody>
</table>

Response with placebo is for placebo groups from trials in the particular comparison being made.
NNTs calculated for $\geq 15\%$, $\geq 30\%$ and $\geq 50\%$ pain relief were very similar with etoricoxib 60 mg and 30 mg, celecoxib 200 mg and naproxen 1000 mg (table 1, fig 2). NNT values were between 3 and 5 over the 12 weeks of measurement. For $\geq 70\%$ pain relief the NNT was distinctly higher with values between 6 and 10. The pattern for ibuprofen 2400 mg was different, with NNTs generally much higher (worse) at longer study duration and less consistency between the various levels of response.

There were three direct comparisons, each in two trials, of daily doses of etoricoxib 60 mg and naproxen 1000 mg, etoricoxib 30 mg and celecoxib 200 mg, and etoricoxib 30 mg and ibuprofen 2400 mg. No statistically significant differences were found between them at any level of response or any duration of treatment.

The additional reduction in the WOMAC pain subscale score for each active treatment over placebo between the flare/randomisation visit and end of treatment is shown in fig 3. The smallest mean difference above placebo was 8 mm for ibuprofen 2400 mg daily and the largest was 15 mm for etoricoxib 60 mg daily. This shows that about 60% of patients have moderate benefit (fig 1) while average reductions in pain over placebo appear modest.

**DISCUSSION**

Population mean changes have no easy resonance outside a clinical trial. An average of 10 mm out of 100 mm (10% improvement; fig 3) conveys no expectation of great benefit, with little to balance against known risks. A different approach is needed.

A pilot study of responder analysis in a single trial indicated that it might be a useful way of reporting pain results in osteoarthritis. It suggested that at least 50% pain relief after 6 weeks of treatment could be a useful discriminator between interventions of greater and lesser efficacy. This needs validation with regard to different levels of pain relief, and especially duration, given that arthritis treatments are used in the medium to long term.

IMMPACT provided recommendations for interpreting the clinical importance of treatment outcomes in clinical trials of chronic pain. It suggested that a 10–20% decrease in pain intensity was minimally important, $\geq 50\%$ moderately important and $\geq 70\%$ substantial. This meta-analysis used these three discriminator points, together with the even higher discriminator of $\geq 70\%$ pain relief, to perform a responder analysis, calculate NNTs and examine the effects of trial duration.
Response rates declined as the discriminator level increased for all five active drugs and placebo (fig 1, table 1). Using the IMMPACT descriptions for commonly used NSAIDs, 60–80% of patients with osteoarthritis can expect minimally important pain relief, 50–60% moderate pain relief, 40–50% substantial pain relief and 20–30% extensive pain relief.

With placebo and active drugs the proportion achieving higher levels of pain relief increased over time, perhaps due to a natural waning of pain inherent in the “flare” design. The tendency to less response at lower levels of pain relief over time with naproxen and ibuprofen may reflect higher withdrawal rates for traditional NSAIDs over cyclooxygenase-2 selective inhibitors. Patients may be balancing benefit and harm, with lower levels of relief perhaps not worthwhile in the face of adverse events. In the responder approach, dropouts contribute to the denominator only, whereas with “last observation carried forward” they appear to continue to benefit after withdrawal.

NNTs were comparable for pain relief of $\geq 15\%$, $\geq 30\%$ and $\geq 50\%$, with higher (worse) NNTs for $\geq 70\%$ for both doses of etoricoxib, celecoxib and naproxen. For these three drugs, NNTs were reasonably stable over 2–12 weeks. Ibuprofen 2400 mg daily was different, with NNT values generally increasing (worsening) with longer duration. The longitudinal responder analysis provides more insight than population average differences (fig 3). This apparently different behaviour with ibuprofen 2400 mg daily did not translate into a statistically significant difference in NNTs between ibuprofen and etoricoxib 30 mg, nor was a significant difference found between
etoricoxib 30 mg and celecoxib 200 mg, or etoricoxib 60 mg and naproxen 100 mg. Establishing a dose-response in analgesic trials is known to be difficult even where direct comparisons are available in relatively simple models such as postoperative pain. Demonstrating an absolute difference in response of 10% requires a substantial number of trials and patients. Confirming statistically that ibuprofen 2400 mg is inferior to other NSAIDs at commonly used doses in osteoarthritis will require more data. A range of ±0.5 NNT was proposed to determine whether an NNT has “clinical relevance”—whether the NNT is within acceptable bounds of clinical “accuracy”. A subsequent proposal was that ±0.5 NNT could be used to determine that NNTs were different. If a numerical difference between NNTs of 1 (e.g. 3.5 vs 5.0) was taken to be important, then the application to NNTs in table 1 would begin to differentiate between drugs, with ibuprofen particularly being judged less effective.

The results of the responder analysis achieve the same global conclusions as the original trials—namely, that etoricoxib and its comparators have useful analgesic properties in osteoarthritis. Arguably the most important outcome from these analyses is that, for osteoarthritis, patients and professionals can be provided with trial data that translate into clinical practice by using realistic estimates of the chance of achieving a particular level of benefit. For pain, a mean difference of 10 mm over placebo translates into about 40% having substantial benefit and 50% not having even minimal benefit. Most people with osteoarthritis treated with an NSAID at an appropriate dose can expect to get at least a minimal benefit (though 1 in 5 will not), almost 1 in 2 can expect a substantial benefit and about 1 in 5 an extensive benefit. The prospect of a 1 in 2 chance of substantial benefit has considerably more impact than an average 10% improvement in pain. Moreover, the information is conveyed in terms of both the likelihood of benefit (1 in 2) and the extent of the benefit (substantial).

Patients and professionals are interested in the known associated risks of both common and rare harm. These, too, can be conveyed in terms of likelihood of harm and its consequences. Balancing benefit and harm is easier when common language describes both. Providing information about the chance of response at various thresholds might produce a more realistic appreciation of benefits and risks of treatment. There is, of course, the caveat that these results come from flare designs in clinical trials where patient selection criteria may make the population different (less comorbidity, perhaps) than a clinical practice population.

Responders are defined not just by the level of response but by the outcome used to define response. We chose the WOMAC pain subscale, combining pain associated with walking, climbing stairs, sitting, lying down and at night while in bed. Whether pain is the most appropriate outcome for responder analysis or whether function, sleep, quality of life or compound outcomes like the OMERACT-OARSI set of responder criteria are preferable remains to be examined. Finally, this type of analysis may differentiate between treatments in a manner that is not possible with population mean changes in pain intensity, and discriminatory power may reside in different outcomes.

Responder analysis looks promising and much more helpful than an average change of a few millimetres based on populations of responders combined with non-responders.

CONCLUSION

Population mean change in pain intensity reported in clinical trials may be difficult to translate into clinical decision-making and patient expectations of benefit. Responder analyses and NNTs calculated from them are reproducible for different levels of response and over at least 12 weeks of treatment with effective drugs. This offers the possibility of providing patients and professionals with information on the chance of achieving particular degrees of pain relief, improving clinical decision-making and patient communication.

Acknowledgements: The authors are grateful to Merck Research Laboratories for making available data from the original trials.

Funding: Pain Research is supported in part by the Oxford Pain Research Trust. Neither the Trust nor Merck Research Laboratories had any role in the design, planning, execution of the study or in writing the manuscript. No financial support was received from Merck Research Laboratories for this work. RAM is funded by NIHR Biomedical Research Centre Programme.

Competing interests: RAM has received research grants, consulting or lecture fees from pharmaceutical companies including Pfizer, MSD, GSK, AstaZeneca, Grunenthal, Menarini, Futura and others. RAM and SD have also received research support from charities and government sources at various times. RAM is the guarantor. RAM, OAM and SD have no direct stock holding in any pharmaceutical company. PMP, ARG and HW are employees of Merck Inc.

Contributions: RAM, ARG and PMP were involved with the original concept, planning the study, searching, writing it, analysis and preparing the manuscript. HW provided the data; OAM and RAM performed calculations and analysis; OAM and SD were involved with planning and writing. All authors read and approved the final manuscript.

Provenance and peer review: Not commissioned; externally peer reviewed.

REFERENCES


