ORIGINAL RESEARCH ARTICLE

Drug Safety 2005; 28 (11): 1057-1063
0114-5916/05/0011-1057/$34.95/0

Detecting Treatment Emergent Adverse Events in Clinical Trials

A Comparison of Spontaneously Reported and Solicited Collection Methods

Joachim F. Wernicke, *Douglas Faries*, *Denái Milton* and *Karen Weyrauch*

Lilly Research Laboratories, Indianapolis, Indiana, USA

Abstract

Background: The collection of adverse event data is an important component of clinical trials, but it is not clear whether solicited or unsolicited collection methods are better at distinguishing drug effects from the effects of placebo. The objective of this analysis is to compare the reporting rates and the ability to detect drug-placebo differences with spontaneous versus solicited adverse event collection methods.

Methods: Adverse events were collected by spontaneous (unsolicited) reporting and by structured questionnaires in three randomised, double-blind clinical trials. For both spontaneous and solicited adverse event collection methods, a drug/placebo (D/P) reporting ratio was computed by dividing the reporting rate for the experimental drug by the reporting rate for placebo for each adverse event. An index (Sp-So index) was calculated by dividing the spontaneous D/P ratio by the solicited D/P ratio. A number >1.0 indicates that the spontaneous adverse event collection method is more effective in distinguishing the drug from placebo and a number <1.0 suggests that the solicited adverse event collection method is more effective in distinguishing the drug from placebo.

Results: Reporting rates were greater when events were solicited than when the spontaneous reporting approach was used. The Sp-So index was >1.0 for 22 of the 29 (75.9%) events examined, suggesting that spontaneous collection of adverse events is more effective in distinguishing drug effect from placebo than the solicited approach. However, more statistically significant differences between drug and placebo were detected by the solicited method (nine events) than the spontaneous method (five events). This is due, in part, to the fact that differences in the percentages of adverse events between drug and placebo (rather than ratios of event rates) were more often greater when the solicited approach was used.

Conclusions: As expected, adverse events collected by solicitation leads to higher reporting rates. However, it is not clear that solicitation of events leads to greater ability to detect drug-placebo differences. By using a ratio to assess drug-placebo differences, spontaneous reporting provided larger drug-placebo differences more often than solicitation.

Background

One of the main goals of clinical trials is to establish the safety profile of the compound under investigation. Although collection and reporting of adverse events is a key feature in describing the safety profile of a drug, there is no general consensus as to the best way to do this. Most often, adverse events are collected in an unsolicited manner. Patients are asked if they have had any problems or if anything unusual has happened since their last visit. Any information reported by the patient in response to this question is considered a spontaneously reported adverse event. The spontaneous adverse event collection method provides the data for most pharmaceutical product labelling (basis for package inserts). However, this does not mean that it should be considered the 'gold standard' of adverse event assessment. Another approach is to solicit events with a structured checklist or questionnaire. A number of these have been employed. Examples are the Side Effects Checklist,[1] the Barkley Behaviour and Adverse Event Questionnaire-Modified (BBAEQ-M)[2] and the Association for Methodology and Documentation in Psychiatry-5:Somatic Signs (AMDP-5).[3] Systematic Assessment for Treatment Emergent Events (SAFTEE), an even more sophisticated method employing general and specific inquiries, has been developed and tested.[4] The value of implementing this rigorous approach has been questioned[5] and defended.[6]

Both the unsolicited and solicited approaches have advocates and critics. Those favouring a spontaneous collection method point out that patients should report what is really bothering them and that using a structured questionnaire may be suggestive. Proponents of spontaneous collection methods worry that the solicitation of events will artificially inflate reporting rates. This becomes an important issue when product labels are written. Historically, most approved product labels are based on spontaneous reports. A drug described with solicited events could be at a disadvantage if solicited event reporting systems lead to higher adverse event rates. The data presented here suggest that this is exactly what happens.

Proponents of adverse event solicitation state that these methods lead to higher reporting rates and are, therefore, more sensitive in eliciting the true adverse event profile of drugs. They worry that patients will often not report events spontaneously. This may be because events are transient and have disappeared by the time of the visit. This is particularly problematic when visits are far apart and patients do not use a diary. One example of a situation where this becomes a significant issue is in studies of drugs for dementia, where patients may easily forget adverse effects unless prompted. Another important area of concern is in the reporting of sexual adverse effects, which patients are often reluctant to volunteer spontaneously. Proponents of soliciting events correctly point out that there is usually a comparator group (often placebo) that should mitigate the higher overall rates of event reporting. Although this is true, it is a point often lost in the competitive effort to market drugs.

There are pros and cons to both approaches, but very little data actually comparing the two methods in terms of reporting rates and ability to detect drug-placebo differences. This paper attempts to address this issue by examining data from three large clinical trials in which adverse event information was collected both in an spontaneous manner by open-ended questioning and in a solicited way by using structured questionnaires.

Methods

Data from three large, randomised, double-blind clinical trials, with a total of 653 patients, were used for this analysis. Study 1 was conducted in children (drug n = 109, placebo n = 110), study 2 included children and adolescents (drug n = 84, placebo n = 83) and study 3 was conducted in adults (drug n = 132, placebo n = 135).[7-9] Ethics review boards approved all of the trials, informed consent was obtained and studies were conducted in accordance with the Declaration of Helsinki. Since the question of interest relates to the ability to distinguish between the drug and placebo, the active drug is not identified and for all three studies is indicated by the word 'drug'. These studies were chosen because they employed both spontaneous and solicited adverse event collection methods on the same patients at the same visit. Only patients with both types of event collection data were included in this study.

Unsolicited adverse events were collected by open-ended questioning. That is, patients were asked to report experiences since the last visit in their own words and with whatever degree of elaboration they desired. The Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) III dictionary[10] was used to map the actual terms to standard terms. Rates of treatment emergent adverse events (i.e. those first reported or becoming more severe after baseline) are reported. Note that the COSTART dictionary that was in use when these studies were conducted has now been generally replaced by the Medical Dictionary for Regulatory Activities (MedDRA)[11] coding system.

To obtain solicited events, questionnaires were administered after spontaneous events were recorded. This was done so that patients would not be prompted by the questionnaires. Standardised questionnaires were used and there was no attempt to incorporate events offered spontaneously into the structured questionnaires. The structured questionnaires were: the Side Effects Checklist[1] (used for the child study), the BBAEQ-M[2] (used for the child/adolescent study) and the AMDP-5[3] (administered for the adult trial).

The Side Effects Checklist is a 30-item symptom checklist based on the Subjective Treatment Emergent Symptoms Scale developed by the US National Institute of Mental Health. The items on the checklist include general symptoms, such as trouble sleeping, diarrhoea, headaches and trouble eating. The patient was asked by the clinician if he/she had trouble with the symptoms on the checklist and chose from 'not at all', 'just a little', 'pretty much' and 'don't know'. The clinician also recorded the frequency of these symptoms based on conversation with the patient. An event is considered treatment emergent if it first occurs or worsens from baseline. The BBAEQ-M contains 24 items that are each rated on a 0–9 scale and treatment emergence was defined as any increase of at least two points from the maximum baseline score. The AMDP-5 contains 47 items that are each rated on a 0–3 scale and treatment emergent events were defined as items for which the maximum treatment period score was greater than the maximum baseline score.

For this analysis, adverse events that appeared to reflect the same symptom in the spontaneous and the solicited method were selected. For each event, the ratio between the rate reported by the group taking the drug and the rate for those in the placebo group was computed and is defined as the drug/placebo (D/P) ratio. This was done for the spontaneous and solicited event data. Thus, a D/P ratio >1.0 indicates that the event was reported more often by the drug than by the placebo group. The greater the ratio, the greater the ability to distinguish the drug from placebo. These D/P ratios were plotted with the solicited D/P ratio on the x-axis and the spontaneous D/P ratio on the y-axis. The diagonal line represents the point at which the ratio between drug and placebo adverse event reporting rates was the same for both collection methods. Points that fall above the diagonal line indicate that the spontaneous event collection system is more effective in distinguishing drug from placebo and points that fall below the line suggest that this is true for the solicited event system.

Another way to compare the two adverse event collection systems is to calculate the ratio of the D/P ratios, referred to as the 'Sp-So index'. The Sp-So index was calculated by dividing the spontaneous D/P ratio by the solicited D/P ratio. A number >1.0 indicates that the spontaneous adverse event collection method is more effective in distinguishing drug from placebo and a number <1.0 suggests that the solicited method is more effective in distinguishing drug from placebo. To aid in the interpretation of the Sp-So index, 95% confidence intervals were computed. To avoid making distributional assumptions for the index, confidence intervals were calculated using non-parametric bootstrap resampling using the percentile method.[12] In addition, treatment differences (between drug and placebo) in percentages of patients with each adverse event were compared using Fisher's exact test for both solicited and spontaneous methods. Statistical tests and confidence intervals are not adjusted for multiplicity, as is the standard conservative approach for adverse events in order to avoid missing a safety signal. However, a Bonferoni approach may quickly be applied by multiplying the p-values by the number of comparisons in the study or the number of overall comparisons.

Table I. Percentage of patients experiencing adverse events with either study drug or placebo reported by spontaneous and solicited methods[a]

Spontaneous				Solicited				Sp-So
adverse event	drug (%)	placebo (%)	D/P ratio	adverse event	drug (%)	placebo (%)	D/P ratio	index (95% CI)
Study 1 (children)								
Headache (p = 0.017)[b]	30.3	16.4	1.85	Headache (p = 0.414)	45.9	40.0	1.15	1.61 (1.02, 3.00)
Rhinitis	22.0	19.1	1.15	Cold and sniffles	37.6	46.4	0.81	1.42 (0.90, 2.58)
Abdominal pain	15.6	15.1	1.03	Stomach aches	33.9	40.0	0.85	1.21 (0.67, 2.65)
Diarrhoea	10.1	10.0	1.01	Diarrhoea	22.9	19.1	1.20	0.84 (0.39, 1.93)
Dizziness	9.2	3.6	2.56	Dizziness	32.1	21.8	1.47	1.74 (0.65, 7.14)
Rash	9.2	3.6	2.56	Rashes	17.4	16.4	1.06	2.42 (0.84, 10.40)
Abnormal dreams	5.5	2.7	2.04	Bad dreams	23.9	25.5	0.94	2.17 (0.60, 9.02)
Dry mouth	3.7	4.5	0.82	Dry mouth and lips	29.4	34.5	0.85	0.96 (0.15, 4.26)
Study 2 (children and adolescents)								
Headache	23.8	22.9	1.04	Headaches	26.5	18.5	1.43	0.73 (0.37, 1.36)
Abdominal pain	14.3	10.8	1.32	Stomach aches	19.3	15.7	1.23	1.08 (0.41, 2.96)
Anorexia (p = 0.161)	11.9	4.8	2.48	Decreased appetite (p = 0.013)[b]	28.6	12.5	2.29	1.08 (0.44, 3.62)
Somnolence	7.1	3.6	1.97	Drowsiness	16.7	17.1	0.98	2.02 (0.47, 9.11)
Insomnia	6.0	6.0	1.00	Difficulty falling asleep	12.0	14.7	0.82	1.23 (0.24, 5.62)
Nervousness	6.0	4.8	1.25	Anxious/nervous	13.3	16.2	0.82	1.52 (0.30, 7.44)
Constipation	4.8	1.2	4.00	Constipation	11.0	3.7	2.97	1.35 (0.19, 3.37)
Dry mouth	2.4	1.2	2.00	Dry mouth	7.1	9.6	0.74	2.70 (0.00, 8.80)
Dizziness (p = 1.000)	2.4	1.2	2.00	Dizzy (p = 0.027)[b]	14.3	3.6	3.97	0.50 (0.00, 1.86)
Study 3 (adults)								
Dry mouth (p < 0.001)[b]	23.4	7.2	3.25	Dry mouth (p < 0.001)[b]	53.8	18.5	2.91	1.12 (0.59, 2.42)
Headache (p = 0.429)	19.1	15.2	1.26	Headaches (p = 0.023)[b]	37.1	23.7	1.57	0.80 (0.45, 1.43)
Insomnia (p = 0.009)[b]	17.0	6.5	2.62	Difficulty falling asleep (p = 0.355)	22.0	17.0	1.29	2.02 (0.92, 5.24)
Nausea (p = 0.042)[b]	13.5	5.8	2.33	Nausea (p = 0.030)[b]	25.0	14.1	1.77	1.31 (0.67, 3.12)
Constipation (p = 0.044)[b]	11.3	4.3	2.63	Constipation (p = 0.046)[b]	25.0	14.8	1.69	1.56 (0.72, 4.84)
Anorexia (p = 0.086)	9.2	3.6	2.56	Decreased appetite (p < 0.001)[b]	40.2	20.7	1.94	1.32 (0.52, 4.97)
Libido decreased	7.1	2.2	3.23	Decreased libido	18.2	11.1	1.64	1.97 (0.62, 7.56)
Dizziness (p = 0.173)	5.0	1.4	3.57	Dizziness (p = 0.040)[b]	19.7	10.4	1.89	1.89 (0.49, 5.83)
Palpitation	5.0	0.7	7.14	Palpitations (p = 0.036)[b]	15.9	7.4	2.15	3.32 (0.64, 5.91)
Abdominal pain	2.8	3.6	0.78	Gastric discomfort	23.5	21.5	1.09	0.71 (0.11, 3.03)
Tremor	2.1	0.7	3.00	Tremor	6.8	3.0	2.27	1.32 (0.00, 4.02)
Somnolence	0.7	4.3	0.16	Drowsiness	25.8	24.4	1.06	0.15 (0.00, 0.87)

a p-Values refer to the comparison between drug and placebo.

b Events for which a statistically significant difference (p < 0.05, Fisher's exact test) was detected by either event collection method.

D/P ratio = frequency of reports in the drug group divided by frequency in the placebo group; **Sp-So index** = the D/P ratio for events obtained spontaneously, divided by the D/P ratio obtained by solicited event collection methods.

Results

Table I shows the reporting rates (percentage of patients reporting each event) and the D/P ratio of reporting frequency for spontaneous and solicited adverse events.

As expected, reporting rates were greater when events were solicited than with the spontaneous reporting approach. As indicated by Sp-So indexes >1.0, for most adverse events the spontaneous collection system was found to be more effective in detecting a difference between drug and placebo compared with the solicited adverse event collection method. Of the 29 events listed, this was true for 22 (75.9%) events. However, the Sp-So index confidence intervals were wide and covered the value of 1.0 for all but two of the events, which indicates that the differences were not statistically significant. It is also interesting to note that more statistically significant differences between drug and placebo were detected by the solicited method (nine events) than the spontaneous method (five events). This is due, in part, to the fact that differences in the percentages of adverse events between drug and placebo (rather than ratios of event rates) were more often greater when the solicited approach was used.

The results described previously can also be represented graphically. Figure 1 compares ratios for events from the three studies. For each event, the D/P ratio for solicited events is plotted on the x-axis and the D/P ratio for spontaneous events provides the y-axis value for that point. Points above the diagonal line indicate that the D/P ratio was higher when adverse events were collected by spontaneous reports. As noted previously, for most events the ratios were higher using the spontaneous collection approach.

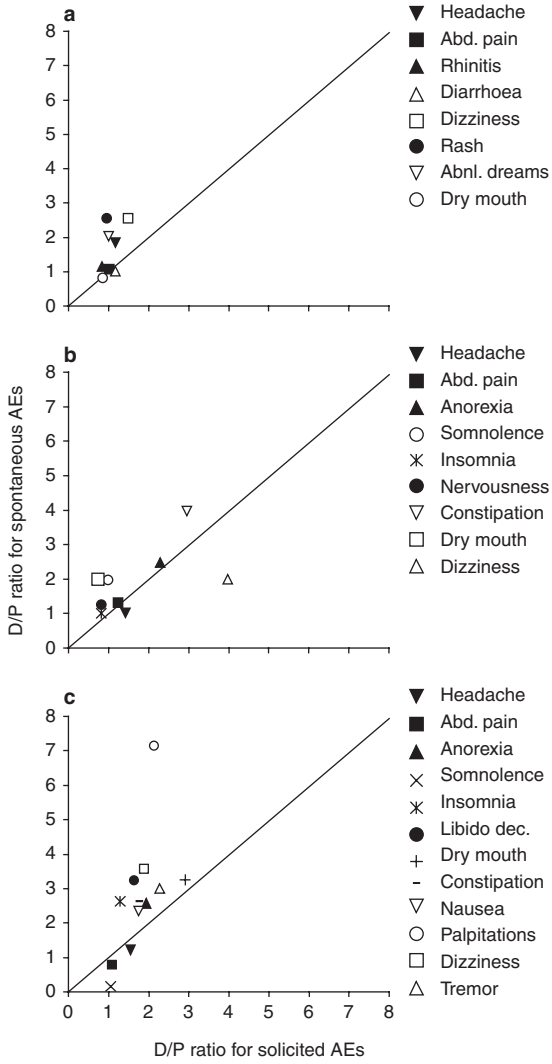

Fig. 1. Ratios of drug/placebo (D/P) event reporting rates obtained by spontaneous versus solicited adverse event (AE) collection methods. (**a**) Child study. (**b**) Child/adolescent study. (**c**) Adult study. **Abd. pain** = abdominal pain; **Abnl. dreams** = abnormal dreams; **Libido dec.** = libido decreased.

Discussion

One interesting comparison between the spontaneous and solicited methods of adverse event data collection was provided by the assessment of 'libido decreased/decreased libido' in the adult study. Although reporting rates were lower with the spontaneous collection method, this approach was almost twice as effective in detecting a difference between drug and placebo than the solicited method (Sp-So index of 1.97). It is often presumed that patients are reluctant to report sexual adverse effects spontaneously and that the best way to elicit such events is with a questionnaire. This presumption is supported by the raw reporting rates shown previously, but the spontaneous collection method appears to be more effective in distinguishing drug from placebo.

If an Sp-So index >1.0 is accepted as evidence that spontaneous adverse event collection is more

effective than solicited approaches in distinguishing drug from placebo, the question remains as to why this is so. A likely explanation is that events that are more bothersome are more likely to be reported spontaneously. When patients are asked what they have experienced since the last visit, they are more likely to remember events that were more problematic, whereas if asked about an event they may remember that indeed they had the experience, but had forgotten it. This could explain why so many of the events, including somnolence, dry mouth, insomnia, palpitations, rash and abnormal dreams had such high Sp-So ratios (>2).

Somnolence/drowsiness in the adult study was an event where detection of a drug-placebo difference was greater with the solicited collection method. For this event, the Sp-So index was 0.15, suggesting that the solicited method is more effective in distinguishing drug from placebo. This result is driven by the near zero spontaneous reporting rate for the drug group (0.7%), leading to a small D/P ratio of 0.16. This finding may be an artifact, the cause of which is not known, as this result was not consistent with the results from the same drug in the child/adolescent study. The spontaneous method was found to be more effective in distinguishing drug from placebo, with a D/P ratio of 2.02, but neither method detected a statistically significant difference between treatments. There is no reason to believe that these drugs behave differently in adults versus children, although this possibility has not been studied systematically.

Limitations

Although this research provided data on solicited and spontaneous reported events from three randomised, placebo-controlled clinical trials,[7-9] there are limitations. The greatest of these is the fact that spontaneous terms (mapped with an event dictionary) were compared with what appeared to be similar events on a questionnaire. 'Dry mouth' and 'dry mouth' probably refer to the same phenomenon, but is 'insomnia' equivalent to 'difficulty falling asleep'? Use of the COSTART dictionary may have led to some lack of precision. The MedDRA system is believed to provide a more accurate description of adverse events. Such methodological questions cannot be addressed by this analysis, but require prospective studies designed to answer such questions.

Another limitation of this analysis is that all patients are considered as one group. Gender, diagnosis, dose of drug and treatment duration may all affect the perception and reporting of adverse events. Although the studies described were fairly large, they were not powered to detect such differences. Also, the drug studied was not the same for all trials, which could have affected the outcome.

Conclusion

Although collection of adverse events is accepted as an important component of all clinical trials, the best method for accomplishing this is less clear. Spontaneous (unsolicited) collection of adverse event data is used in most pharmaceutical trials. This methodology has been criticised on the premise that it leads to under-reporting. That is, it is hypothesised that patients do not report all events unless they are prompted. Solicitation of adverse events is thought to be more sensitive since questionnaires or checklists prompt patients. Clearly, solicitation of events leads to higher reporting rates. However, in clinical trials of active drugs versus placebo the more important question is 'which method is better in distinguishing drug from placebo?'

The data in this study, which includes three trials where both solicited and spontaneous collection of events was used, confirm that solicitation of events leads to higher reporting rates. However, it is not clear that solicitation of events leads to better detection of drug-placebo differences. Indeed, more statistically significant treatment differences were detected when comparing differences in percentages between treatments using solicited events than spontaneously reported events. However, for most events the drug/placebo ratios were actually greater when spontaneous reporting was used. Thus, the impact of the collection method used depends somewhat on the statistical method selected and the overall incidence of the events.

Acknowledgements

This study was conducted as part of the authors' employment at Eli Lilly and Company.

References

1. National Institute of Mental Health (NIMH). STESS: Subjective Treatment Emergent Symptoms Scale. Psychopharmacology Bulletin Vol 21 (Issue No. 4). Rockville (MD): National Institute of Mental Health, 1985: 1073-5
2. Barkley RA, McMurray MB, Edelbrock CS, et al. Side effects of methylphenidate in children with attention deficit hyperactivity disorder: a systemic, placebo-controlled evaluation. Pediatrics 1990; 86 (2): 184-92
3. Stieglitz RD, Fahndrich E, Helmchen H. The AMDP system. In: Mezzich JE, von Cranach M, editors. The international classification in psychiatry: unity & diversity. New York: Cambridge University Press, 1988: 180-204
4. Levine J, Schooler NR. Strategies for analyzing side effect data from SAFTEE: a workshop held fall 1985 in Rockville, Maryland. Psychophamacol Bull 1986; 22 (2): 343-81
5. Rabkin JG, Markowitz JS, Ocepek-Welikson K, et al. General versus systematic inquiry about emergent clinical events with SAFTEE: implications for clinical research. J Clin Psychopharmacol 1992; 12 (1): 3-9
6. Levine J, Schooler NR. General versus specific inquiry with SAFTE [letter]. J Clin Psychopharmacol 1992; 12 (6): 448
7. Michelson D, Faries D, Wernicke J, et al. Atomoxetine in the treatment of children and adolescents with attention-deficit/hyperactivity disorder: a randomized, placebo-controlled, dose-response study. Paediatrics 2001; 108 (5): e83
8. Michelson D, Adler L, Spencer T, et al. Atomoxetine in adults with ADHD: two randomized, placebo-controlled studies. Biol Psychiatry 2003; 53 (2): 112-20
9. Data on file (unpublished), Eli Lilly and Company
10. US Department of Health and Human Services. Coding symbols for thesaurus of adverse reaction terms. 3rd ed. Rockville (MD): Department of Health and Human Services, Food and Drug Administration, Center for Drug and Biologics, Division of Drug and Biological Experience, 1995
11. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf 1999; 20 (2): 109-17
12. Chernick MR. Bootstrap methods: a practitioner's guide. New York: John Wiley & Sons Inc., 1999: 53

Correspondence and offprints: Dr *Joachim F. Wernicke*, Lilly Corporate Center, Indianapolis, IN 46285, USA.
E-mail: jfwernicke@lilly.com