ORIGINAL RESEARCH ARTICLE

Development and Initial Validation of a Patient-Reported Adverse Drug Event Questionnaire

Sieta T. de Vries · Peter G. M. Mol ·
Dick de Zeeuw · Flora M. Haaijer-Ruskamp ·
Petra Denig

Abstract

Background Direct patient reporting of adverse drug events (ADEs) is relevant for the evaluation of drug safety. To collect such data in clinical trials and postmarketing studies, a valid questionnaire is needed that can measure all possible ADEs experienced by patients.

Objective Our aim was to develop and test a generic questionnaire to identify ADEs and quantify their nature and causality as reported by patients.

Methods We created a draft list of common ADEs in lay-terms, which were classified in body categories and mapped to the Medical Dictionary for Regulatory Activities (MedDRA®) terminology. Questions about the nature and causality were derived from existing questionnaires and causality scales. Content validity was tested through cognitive debriefing, revising the questionnaire in an iterative process. Feasibility and reliability were assessed using a Web-based version of the questionnaire. Patients received the questionnaire twice. Feasibility was assessed by the reported time needed for completion and ease of use. Reliability was calculated using Cohen's kappa and proportion of positive agreement (PPA) on: (1) any ADE at patient level; (2) similar ADEs at MedDRA® System Organ Class level; and (3) the same ADE at ADE-specific level.

Results In the development phase, 28 patients with type 2 diabetes or asthma/chronic obstructive pulmonary disease (COPD) participated. Questions and answer options were rephrased, layout was improved, and changes were made in the classification of ADEs. The final questionnaire consisted of 252 ADEs organized in 16 body categories, and included 14 questions per reported ADE. A total of 135 patients using a median of five different drugs completed the Web-based questionnaire twice. The median completion time was 15 min for patients not reporting any ADE, and 30 min for patients reporting at least one ADE. Three quarters of the patients found the questionnaire easy to use. Test–retest reliability was acceptable at patient level ($\kappa = 0.50$, PPA 0.64) and at MedDRA® System Organ Class level ($\kappa = 0.52$, PPA 0.54), but was low at ADE-specific level ($\kappa = 0.38$, PPA 0.38).

Conclusion We developed a generic patient-reported ADE questionnaire and confirmed its content validity. The questionnaire was feasible and reliable for reporting any ADE and similar ADEs at MedDRA® System Organ Class level. Additional work is, however, needed to reliably quantify specific ADEs reported by patients.

S. T. de Vries · P. G. M. Mol · D. de Zeeuw ·
F. M. Haaijer-Ruskamp · P. Denig (✉)
Department of Clinical Pharmacology, University of Groningen,
University Medical Center Groningen, PO Box 196,
9700 AD Groningen, The Netherlands
e-mail: p.denig@umcg.nl

1 Introduction

Today, patients are increasingly involved in information gathering and decision making at all levels of the healthcare system [1]. Patient self-reports of adverse drug events (ADEs) are an important additional source of information on the safety of drugs because they differ from healthcare professional reports [2–7]. Healthcare professionals often underestimate symptomatic ADEs experienced by patients [7, 8]. The added value of patient reports is acknowledged by the Food and Drug Administration (FDA) as well as the

△ Adis

European Medicines Agency [9, 10]. The FDA advises the use of patient-reported outcome (PRO) questionnaires for the measurement of outcomes that are best known by patients [9] (e.g., pain [11]). In PRO questionnaires, the patient is the direct source of information without interpretation of the responses by a healthcare professional [9, 12].

Patient-reported ADE questionnaires can be open-ended or checklist based. Compared to open-ended questionnaires, checklist-based questionnaires are more sensitive in identifying potential ADEs [13, 14]. However, these methods may lack specificity in the detection of true ADEs [13]. Adding questions per ADE on its nature and causality might solve this problem. To assess unknown ADEs of (new) drugs and comparing ADE profiles of different drugs, a generic PRO questionnaire is needed that can measure all possible ADEs [13, 15]. Most available patient-reported ADE questionnaires focus on specific ADEs, such as gastrointestinal ADEs [16] or ADEs specific for a drug class, such as inhaled corticosteroids [17] or chemotherapy [18]. Previously, a generic questionnaire was developed that contained approximately 600 symptoms classified by body category [19]. More recently, a questionnaire with 84 ADEs classified in 19 body categories was developed [3]. Although both questionnaires have been piloted, no explicit validation has been reported. Furthermore, both questionnaires lack questions supporting causality assessment and questions about the nature of the ADE such as those regarding seriousness, severity, frequency, and time course, which are relevant attributes in the evaluation of the ADE [20, 21].

The aim of our study was to develop and test a generic questionnaire for identifying ADEs and assessing their nature (e.g., frequency, severity) and causality as reported by patients. We tested the content validity and feasibility of the questionnaire as well as the reliability for reporting ADEs.

2 Method

The study consisted of three parts: (1) development of a draft ADE questionnaire, (2) content validation and revision of the questionnaire in an iterative process, and (3) feasibility and reliability testing of the revised questionnaire.

2.1 Questionnaire Development

The questionnaire consists of four sections with questions about: (1) general patient characteristics; (2) drug use in the past 4 weeks, diseases for which these drugs were used, whether the patient had other diseases; (3) ADEs experienced in the past 4 weeks using structured checklists; and (4) for each ADE a question to describe the ADE in the patient's own words with additional questions about its nature and causality. We expected that a period of 4 weeks would be sufficient for capturing a wide range of ADEs for which patients would be able to recall the relevant details. In the development phase, ADEs were selected, named, coded, and categorized into a body category, and questions were constructed to assess the nature and causality of the ADEs.

2.1.1 Adverse Drug Event (ADE) Selection and Naming in Lay-Terms

We aimed to include a wide range of common symptomatic ADEs. We identified possible ADEs from the Common Terminology Criteria for Adverse Events version 4.0 [22], and existing symptom and ADE checklists [3, 13, 18, 23–29]. Patient-reported data about ADEs from the Lareb Intensive Monitoring System of The Netherlands Pharmacovigilance Centre Lareb [30] were used to translate ADEs into lay-terms. We excluded ADEs based on laboratory results (e.g., hyperkalemia) and those related to specific devices (e.g., uncomfortable pressure of the mask). The first selection included 252 possible ADEs with an open-ended option for reporting "other" experienced ADEs.

2.1.2 Coding of ADEs

Two researchers (SdV and PD) independently coded each lay-term ADE to a lowest level term of the Medical Dictionary for Regulatory Activities (MedDRA®) terminology version 13.0, making use of codings suggested by pharmacovigilance experts from Lareb. MedDRA® is the international medical terminology developed under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Agreement between the codings existed in 74 % of the ADEs. Dissimilarities were resolved by discussion, and translation of the Dutch lay-terms into English by a professional translator was used to reach agreement on all MedDRA® terms. Two ADEs, "Bone fracture or fractures" and "Stroke," were classified at a higher hierarchical ADE group definition because of their nonspecific nature. One ADE (dry teeth) showed overlap in the MedDRA® terminology with another included ADE (dry mouth), and they were therefore combined.

2.1.3 Categorization of ADEs

To increase the efficiency of completing the questionnaire, the ADEs were classified in body categories. By first checking body categories in which patients experienced

ADEs, they were directed to short checklists of specific ADEs within that body category. These lists with specific ADEs also include the option to report other ADEs. The body categories in the initial questionnaire were based on the classification used in the MedDRA® and in existing questionnaires [3, 19].

2.1.4 Assessing Nature and Causality of ADEs

Relevant known attributes of ADEs were duration, frequency, severity, and seriousness of the ADE; its impact on activities; and the patient's benefit–risk assessment of the drug [24, 30–32]. Existing questionnaires were screened for questions covering these topics [26, 27, 33–35]. Questions regarding causality were included, based on medical [36], and patient-reported considerations [37].

2.2 Content Validation

The draft questionnaire was subjected to cognitive debriefing interviewing to eliminate ambiguity in questions and answer options. Cognitive debriefing is a qualitative interview method in which the patient's understanding and interpretation of items and answer options of the questionnaire are assessed [38, 39]. A separate classification task was used to assess the appropriateness of the body categories.

2.2.1 Study Population

Patients included in the study were 18 years or older; diagnosed with type 2 diabetes, asthma, and/or chronic obstructive pulmonary disease (COPD); using drugs for these conditions; and able to speak, read, and write the Dutch language. Patients with these diagnoses were included to cover a population with a broad age range in which many different types of drugs are commonly used, both daily and as needed. Eligible patients were recruited by three general practitioners and two dieticians in the northern part of The Netherlands in 2011–2012.

2.2.2 Study Procedure

After signing informed consent, patients completed the questionnaire during which they were observed by a researcher (SdV) to detect any problems with completing the questionnaire. Immediately thereafter, a semi-structured interview was conducted using a topic list based on the "question-and-answer" model [38, 39]. A subset of patients was asked to do a classification task, for which all ADEs were randomly split into five lists. Patients were instructed to classify each ADE of one list into a body category. Each ADE was classified by at least four patients.

2.2.3 Analyses

The audio-recorded interviews were transcribed verbatim, and transcripts were screened by two researchers (SdV and PD) to identify problems in understanding the questions and answer options. The questionnaire was adapted in an iterative process by which changes were made addressing detected problems until no new problems were identified regarding understanding the questions and answer options (Fig. 1) [38].

Regarding the classification task, we considered an ADE classification as problematic when more than two patients classified the ADE in body categories different from our original classification, or when two patients were consistent in choosing a different category. These problematic ADEs were subsequently judged by four additional patients and a pharmacovigilance expert. Based on their judgements, revisions were made. This revised questionnaire was then translated from Dutch to English by a professional translator. The English version was screened for differences with the original Dutch version through informal back translation by the researchers, and final changes were made. A Web-based version of the content-valid questionnaire was then constructed using the Unipark Enterprise Feedback Suite 8.0 version 1.1 (http://www.unipark.de).

2.3 Feasibility and Reliability Testing

The Web-based version was used to assess the feasibility of completing the questionnaire, its ability to measure the ADEs in a consistent manner (test–retest reliability), and to assess the impact of using body categories on feasibility and ADE reporting.

2.3.1 Study Population

Included patients were aged 18 years or older, had been dispensed an oral glucose lowering drug, had an e-mail address, and were able to access the Internet. These patients were recruited via pharmacists in the northern part of The Netherlands in 2012.

2.3.2 Study Design and Procedure

In a test–retest design study, consenting patients received an e-mail message with the URL (uniform resource locator) to open the Web-based version. A personal login code was used to prevent multiple completions of patients [40]. After completion of the ADE part, questions were asked regarding feasibility, including self-reported time to complete the questionnaire and ease of use on a five-point Likert scale. In addition, the total time between opening

Fig. 1 Iterative process in adapting the developed questionnaire to a content-validated questionnaire
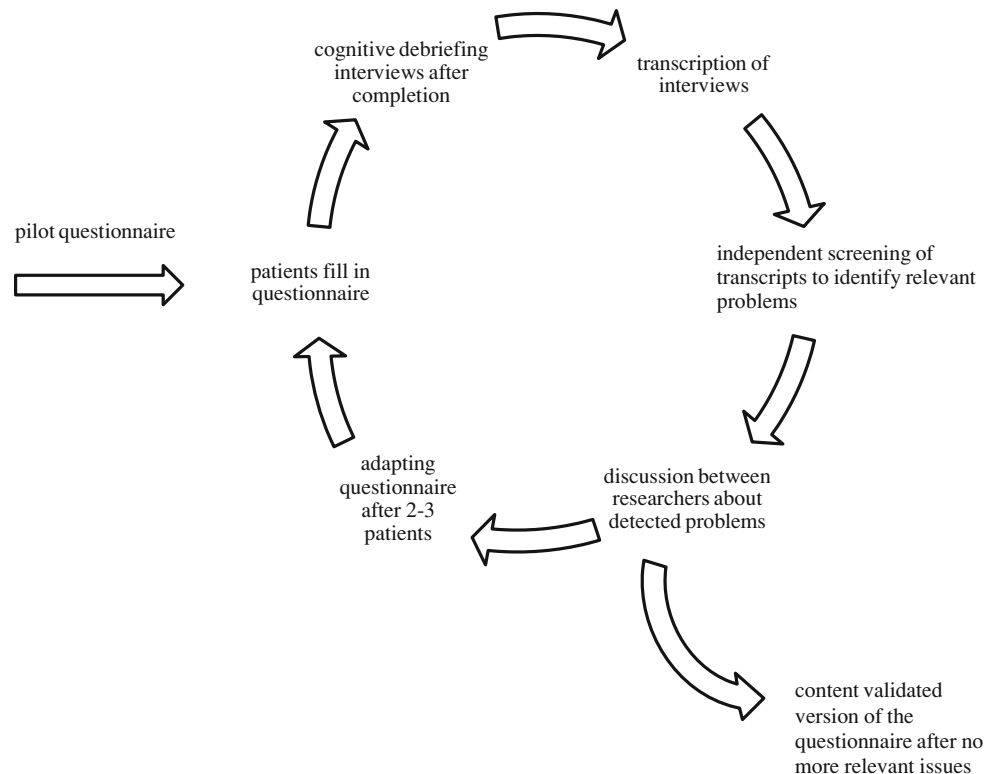
and closing of the digital questionnaire was logged (registered time), as well as the proportion of patients completing the questionnaires, and the number of ADEs reported in the "other" category. One week after completion, patients received an e-mail for the second questionnaire for the reliability analysis.

Patients were randomly assigned to three groups using simple randomization [41] to receive: (A) the same questionnaire twice (the "test–retest group"); (B) a questionnaire with the body category structure at the first measurement (T1) and without these categories at the second measurement (T2) [the "group with body categories at T1"]; or (C) reversing the order used in B (the "group with body categories at T2").

One reminder was sent to the patients who did not complete the first questionnaire within a month. Patients who did not complete the second questionnaire were send a reminder twice. We aimed to include about 50 patients per group, which has been reported as a reasonable number for reliability studies [42].

2.3.3 Analyses

Differences in sex and age between responders and non-responders were assessed using Chi-square and Mann–Whitney U tests. Descriptive statistics were used for the feasibility parameters, including self-reported completion time, ease of use, proportion of patients completing the

questionnaires, and number of ADEs reported in the "other" category. ADEs that were reported as "other" were evaluated and, if possible, classified by the researchers within the provided ADE lists. To assess the number of chronic diseases, we classified each self-reported disease in 1 of 12 chronic diseases, excluding conditions of normal ageing (e.g., loss of hearing).

We measured the agreement between ADE reporting at T1 and T2 at three levels: any ADE at "patient level," similar ADEs at primary System Organ Class "MedDRA® level," and the same ADE at the lowest description "ADE specific level." Cohen's kappa coefficient and proportion of positive agreement were calculated as measures of agreement. Especially at the lowest level, where specific ADEs will be checked by few patients, the kappa statistic is negatively affected by the skewed distribution and proportion of positive agreement has been proposed as an alternative [43]. The proportion of positive agreement was calculated by the formula $2a/[N + (a - d)]$, in which N is the total number of observations, a is the number of patients reporting ADE at T1 and T2, and d is the number of patients not reporting ADE at T1 and T2 [44]. Kappa and proportion of positive agreement values of >0.5 were considered to be acceptable [45]. We conducted additional analyses aggregating experienced ADEs using the patients' own description of the ADEs. Based on these descriptions, two researchers (SdV and PD) clustered ADEs that were checked as separate ADEs but described by the patients as

being one problem. Although one might expect that this clustering is similar to the aggregation at MedDRA® level, it is possible that patients use terms from different MedDRA® classes to describe one problem. For instance, goose bumps, shivering, and cold limbs can be seen as one problem by the patient but are coded in different primary MedDRA® System Organ Classes. Misclassification can also occur when patients check similar but not the same symptomatic ADEs at T1 and T2. Finally, we calculated how often patients checked a symptom only as a symptom at one time point but as a possible ADE at another time.

The effect of including body categories was tested by comparing feasibility parameters and the number of reported ADEs between the questionnaire with body categorization and without at baseline, using Chi-square and Mann–Whitney U tests. Additionally, the agreement values of the *group with the body categories at T1* and the *group with the body categories at T2* were compared using the normal curve deviate statistic (Z value) [46].

Sensitivity analyses were conducted to investigate whether the number of days between completing the first and second questionnaire influenced the agreement values. All analyses were conducted using IBM SPSS Statistics version 20 (Armonk, New York, USA). *P*-values of <0.05 were considered to be statistically significant.

3 Results

3.1 Questionnaire Development

The initial version of the questionnaire contained 252 ADEs categorized in 21 body categories, and 11 questions regarding the nature and causality assessment for every ADE identified.

3.2 Content Validation

Twenty-eight patients, 54 % of them women, participated (Table 1). Ages ranged from 22 to 90 years, with a median of 61 years. Almost all patients used more than one drug.

3.2.1 Content Validation, Cognitive Debriefing Interviews

Based on the cognitive debriefing interviews, the questionnaire was revised 14 times. This included a revision of the general structure of the questionnaire, and a major revision by asking for ADEs as well as symptoms. The final revision was tested in five patients and no major problems in the interpretation of questions and answer options were detected. Problems detected in the questionnaire are presented according to the domains of the question-and-answer model, with examples given in Table 2.

Table 1 Patient characteristics of content validation

Total number of participants	28 (15 women)
Median age in years (range)	61 (22–90)
Education	
Low[a]	14
Middle[b]	10
High[c]	4
Diagnosis	
Type 2 diabetes	16
Asthma/COPD	10
Type 2 diabetes and asthma/COPD	2
Multiple- versus single-drug users	
Multiple drug users	26
Single drug users	2
Median number of self-reported prescription drugs (range)	5 (1–14)
Median number of self-reported chronic diseases, including asthma/COPD, diabetes, and cardiovascular diseases (range)	3 (1–6)

COPD chronic obstructive pulmonary disease

[a] No education; elementary school; junior secondary vocational education

[b] Junior general secondary education; senior secondary vocational education

[c] Senior general secondary education; higher professional education; university education

Wording of the body categories and ADEs was generally clear for the patients (Table 2: "Comprehension"). Several ambiguous interpretations, reading difficulties, and vague statements were reported by patients regarding specific question and answer options, which were subsequently changed. Eight patients reported that the recall period of 4 weeks for the experienced ADEs was short (Table 2: "Retrieval"). Because this did not reflect the content validation, no changes regarding the recall period were made during the study period.

The initial questionnaire asked to indicate "experienced ADEs." However, it became clear that patients, when confronted with a checklist of possible symptomatic ADEs, incorrectly started to check symptoms that they actually did not see as ADEs (Table 2: "Judgement"). Asking to check both experienced symptoms and ADEs solved this problem. The answer option "do not know" was added because some patients were not sure whether the experienced symptom was related to a drug they used. Almost half of the patients either skipped the body categories to go directly to the specific checklists (navigation) or had difficulties in deciding which body category their symptom might be classified into. Other patients who used the body categories found them helpful and easy to use. As a result, we kept the body category structure as a supportive step in

Table 2 Examples of issues mentioned by patients per topic during the cognitive debriefing

Domain of the question-and-answer model	Topic in topics list	Examples of issues mentioned by patients
Comprehension	Text	"Difficult question. I do not understand it entirely. A bit difficult question. I am reading it 7 times and still do not know what they mean." (female, 58 years)
	Adverse drug events	"Most of this is common language, no medical terms, and otherwise it is explained." (female, 44 years)
Retrieval	Recall	"In my opinion, the period of 4 weeks is quite short." (male, 61 years)
Judgement	Symptom or adverse drug event	"I thought, I experience all kind of things. But if you read further, it is about medication, then you say, no that thing has nothing to do with it. But I experience that symptom but it has nothing to do with medication." (male, 80 years)
		"I find it difficult to say which are side effects. I do experience symptoms but are they symptoms or side effects. No idea. And I think I still reported it [the symptom] because I do not know and because perhaps you may think when it is reported by everyone, it can be a side effect." (female, 61 years)
	Body categories	"I think of only some things with that [body category] and then later I had to go back, no, this fits with that one [body category]." (female, 71 years)
		"Do you have problems with your eyes? Yes. Sometimes I have a blurred vision, I cannot tolerate sunlight very well, so in that case you check eyes. Bladder, I use that tolbutamide from which I have to pee a lot, so the bladder. The skin, I have quite a dry skin lately. Often, my back or my hands are itching, so then you check skin. So, you just go by this [list]."(female, 53 years)
Response	Answer options	Regarding "how often" ADE is experienced: "Almost every day and that for a period of 14 days." (female, 58 years)
	Lack of…	"I would include whether the side effect is treated or whether it disappeared spontaneously." (female, 44 years)
Respondent burden	Structure	"It can be confusing, at one time you are asked for the drug. And the next time not. Then again about side effects, and then again about drugs." (male, 58 years)
	Layout	"For me it [the size of the letters] is a little bit small." (male, 90 years) "At a certain point I found it [the navigation in the questionnaire] a bit chaotic." (female, 44 years)
	Relevance	Regarding the number of answer options: "Somewhat less. Maybe half of it can go. I think, everything a little bit more concise." (female, 53 years)

the questionnaire, but patients no longer needed to check body categories before going to the specific checklists.

Answer options that did not fit with the judgements of the patients were detected and adapted, and answer options were added (Table 2: "Response"). The answer options of the question "how often did you experience this side effect during the past 4 weeks (on how many or which days)?" were changed multiple times. Problems remained especially for intermittently occurring ADEs, and this question was therefore adapted into an open-ended question (Table 3).

Two questions were added to the initial questionnaire because they yielded additional information regarding causality (Table 3). One question was added to cover an additional attribute, namely actions taken (Table 3).

One patient reported difficulties with the sequence of the questions per ADE (Table 2: "Respondent burden"). This was improved by clustering the topics of the questions. One patient had some problems with the size of the letters in the questionnaire (font size 11, Arial), but none of the other patients reported such reading difficulties. Problems

regarding navigation in the questionnaire, especially due to layout issues were detected and resolved. After seven interviews, the questionnaire was split into two distinct parts, separating the specific questions about the ADEs from the first part of the questionnaire. Two patients mentioned that they felt many questions per ADE were included but that this was not a problem for them. Comments on the length and number of answer options of a causality question led to shortening these phrases (Table 3).

3.2.2 Classification Task

Based on the classification task, where the patients had to assign ADEs to body categories, 51 problematic ADEs (20 %) were detected. As a consequence, we made the following adaptations: shifting the ADE to a more fitting body category (5 ADEs), renaming the ADE (2 ADEs), a combination of shifting and renaming of the ADE (2 ADEs), renaming a body category (8 ADEs), combining body categories (16 ADEs), and creating a new body category (6 ADEs). For 12 ADEs, no changes were made.

Table 3 Comparison of questions regarding nature and causality of ADEs between initial and last revision

Attribute assessed	Initial question[a]	Question in last revision	Answer options
Duration/timeline	Since when have you experienced this side effect? Try to be as specific as possible (for example 1 March 2009 or June 2010 or 2006)	When did you **first** experience this side effect of your medication?	Adapted to checklist-based
Duration/timeline	–	*New*: Has this side effect gone away by now or improved?	–
Frequency	How often did you experience this side effect during the past 4 weeks?	How often did you experience this side effect during the past 4 weeks (on how many or which days)?	Adapted to open-ended
Severity	How much did this side effect bother you in the past 4 weeks (how bad or intense was it)?	On the days that you experienced this side effect, how much did it bother you (how bad or intense was it)?	No changes
Severity + impact on activities	How much influence did this side effect have on your daily functioning in the past 4 weeks?	On the days that you experienced this side effect, how much influence did it have on your daily functioning?	No changes
Actions taken	–	*New*: What action did you take in relation to this side effect during the past 4 weeks?	–
Seriousness	Did this side effect result in serious medical situations for yourself during the past 4 weeks?	*Unchanged*	One answer option excluded
Importance to patient + benefit–risk assessment	How satisfied are you with the drug/drugs described in question IV-8 when you consider both **this particular** side effect and the effect of this drug/these drugs?	How satisfied are you with the drug (or drugs) described in question 38 when you consider both **this particular** side effect and the effect of the drug or drugs?	No changes
Causality	Which drug(s) do you think caused this side effect?	Which drug or drugs do you think caused this side effect?	No changes
	Why do you think this side effect is caused by this drug/these drugs (several answers possible)?	Why do you think this symptom was caused by your medication (you may give more than one answer)?	Length of answer options has been shortened
	How sure are you that this side effect is caused by this drug?	How sure are you that this side effect is caused by this drug or these drugs?	No changes
	Do you think there are possible other factors for your experiencing this side effect (other than your medication)?	Do you think there are other reasons for your experiencing this side effect (other than your medication)?	No changes
	Have you experienced this symptom in the past in combination with other medication?	Have you experienced this side effect in the past in combination with **other medication**?	No changes
Causality (also timeline)	–	*New*: How long had you been using this drug or these drugs before this side effect started occurring?	–

[a] The term "side effect" was used in the questionnaire as a lay-term for adverse drug event

3.2.3 Final Revision

Based on the English translation, one ADE was detected that was considered ambiguous in the original Dutch version. To solve this, two ADE descriptions instead of one were introduced ("blood with feces" and "blood in feces"). The comparison of the English version with the Dutch version resulted in a few minor changes regarding the wording in both versions. Finally, after combining the ADEs with an overlapping MedDRA® term (dry teeth/mouth), the final questionnaire contained 252 ADEs categorized in 16 body categories with 14 questions per ADE regarding its nature and causality (Appendix I ADE questionnaire, Appendix II Questions per reported ADE, and Table 3).

3.3 Feasibility and Reliability

In total, 187 patients gave informed consent in response to an invitation that was mailed to 958 patients. These 187 patients were slightly younger (65 vs 67 years, $Z = -2.653$, $P = < 0.01$) than patients not responding. There was no significant difference regarding sex (39.6 vs 44.7 % women, $\chi^2 = 1.638$, $P = 0.20$). Of the consenting patients,

152 started with the study by opening the questionnaire, and 137 completed both questionnaires (73.3 %). Four times, a patient reported an ADE in the "other" box, which could all be classified to one of the listed ADEs by the researchers. One patient reported in the comments that the reported ADE was probably not due to a drug but due to surgery. This ADE was excluded from further analysis. One patient was excluded from the test–retest analysis for reporting to have experienced the "same symptoms" at T2 as at T1, instead of checking the symptoms again. Another was excluded because of this patient's comment that several symptoms had been wrongly checked. Further analyses were thus based on 135 patients, 45 in each group. The median age of this population was 65 years; on average, they used five prescription drugs (Table 4). The median number of days between completing the first and second questionnaires was 8 days (SD 4).

At T1, 25.2 % ($N = 34$) of the 135 patients reported one or more ADEs, and 27.4 % ($N = 37$) at T2. In total, 173 ADEs were reported at T1, and 146 ADEs at T2. The most common type of ADEs were gastrointestinal disorders (Table 5). Less than 1 % of the questions about the nature and causality of the ADE were not completed (0.4 % missing at T1, and 0.2 % at T2). For most ADEs (124 at T1 and 96 at T2), patients checked only one reason for suspecting the ADE. The most common reason was that they did not experience the symptom before they took the drug. In three quarters of the cases, the patients indicated which drug they thought caused the symptom, and in most of these cases they were quite sure about the relationship between the drug and the ADE (Table 5). Finally, there were 51 cases where a symptom was reported only as a symptom at one point but as a possible ADE at another time (22 times as symptom at T1 but ADE at T2, and 29 times as ADE at T1 but symptom at T2).

Self-reported time for questionnaire completion was in general lower than the registered time (Table 6). On average, the median self-reported time was 15 min for patients not reporting any ADE (with three patients reporting >30 min), and 30 min for those reporting one or more ADEs (with four patients reporting >60 min). Differences observed in completion time between the questionnaire with and without body categorization were not significant (Table 6). Most of the patients agreed that the questionnaire was easy to use (74.4 % for the questionnaire with body categories; 75.6 % for the questionnaire without body categories), which did not significantly differ between the two versions of the questionnaire ($\chi^2 = 0.028$, $P = 0.986$). Overall, this percentage was lower for patients reporting one or more ADEs than for patients not reporting any ADE (52.9 vs 82.2 %, $\chi^2 = 12.791$, $P = 0.002$).

The agreement of reported ADEs regarding the test–retest reliability was acceptable at patient level and at

MedDRA$^{®}$ level ($\kappa > 0.5$, proportion of positive agreement >0.5). At ADE specific level, the agreement was lower ($\kappa = 0.38$, proportion of positive agreement = 0.38, Table 7). By aggregating separately checked but related ADEs according to the patient's own description, the 64 ADEs reported at T1 were reclassified as 34 distinct ADEs, and the 51 ADEs at T2 as 31 distinct ADEs. There was agreement for 16 of these ADEs and the proportion of positive agreement was 0.49.

Agreement between the two measurements was slightly higher for patients who completed the questionnaire including body categories at first measurement in comparison to those who first completed the questionnaire without this categorization. However, kappa values did not significantly differ between the *group with the body categories at T1* and the *group with the body categories at T2* (Table 6). The two-by-two tables of the agreement analyses are presented in Appendix III. The number of reported ADEs was similar between the questionnaire with and without body categories ($Z = -0.049$, $P = 0.961$). Sensitivity analyses including only those patients who completed the second questionnaire within 10 days did not lead to significant differences in agreement measures (Appendix IV).

4 Discussion

We developed and tested a generic questionnaire for patient reporting of ADEs. The questionnaire adds to the available questionnaires in that it is both generic and checklist-based and includes specific questions about causality, severity, duration, seriousness, and frequency of each experienced ADE. The questionnaire is intended for use in postmarketing studies and clinical trials.

Through cognitive debriefing interviews, significant problems were detected in several domains of the question-and-answer model that needed to be resolved. After initial adaptations, some problems reoccurred, underlining the relevance of an iterative process. The input of patients was found to be vital for the development and content validation. It became clear that directly asking for ADEs can lead to over-reporting because some patients accidently checked symptoms as well as ADEs when confronted with a list of symptomatic ADEs. While going through the lists, patients sometimes forgot that they should only check symptoms perceived as being ADEs. This happened even while patients were able to distinguish ADEs from symptoms, as has been established before [37, 47]. Some of the available checklist-based ADE questionnaires use terms such as symptoms, problems, and ADEs interchangeably (e.g., see [18, 27]). We recommend clear differentiation between symptoms that could be related to the underlying disease

Table 4 Patient characteristics, number of adverse drug events (ADEs) reported per group (P-values for differences among the three groups)

	Total	Test–retest group (T1 and T2 with body categories)	Group with the body categories at T1	Group with the body categories at T2	P-value[a]
Number of participants	135	45	45	45	
Women (%)	49 (36.3)	16 (35.6)	14 (31.1)	19 (42.2)	0.544
Median age in years (range)	65 (41–86)	64 (44–86)	67 (47–82)	63 (41–83)	0.210
Education (%)					0.796
Low[b]	38 (28.1)	15 (33.3)	11 (24.4)	12 (26.7)	
Middle[c]	50 (37.0)	13 (28.9)	17 (37.8)	20 (44.4)	
High[d]	40 (29.6)	14 (31.1)	15 (33.3)	11 (24.4)	
Other	7 (5.2)	3 (6.7)	2 (4.4)	2 (4.4)	
Median number of self-reported prescription drugs (range)	5 (2–18)	5 (2–14)	5 (2–13)	6 (3–18)	0.095
Median number of self-reported chronic diseases, including diabetes and cardiovascular diseases (range)	3 (1–10)	3 (2–8)	3 (1–5)	3 (2–10)	0.367
Number of patients reporting an ADE at T1 and T2 (%)	T1: 34 (25.2)	T1: 12 (26.7)	T1: 11 (24.4)	T1: 11 (24.4)	T1: 0.961
	T2: 37 (27.4)	T2: 13 (28.9)	T2: 11 (24.4)	T2: 13 (28.9)	T2: 0.862
Number of ADEs reported (range)	T1: 173 (1–19)	T1: 64 (1–15)	T1: 35 (1–10)	T1: 74 (1–19)	T1: 0.339
	T2: 146 (1–11)	T2: 51 (1–10)	T2: 34 (1–9)	T2: 61 (1–11)	T2: 0.394

T1 first measurement, *T2* second measurement after 1-week period

[a] For differences among the three groups

[b] No education; elementary school; junior secondary vocational education

[c] Junior general secondary education; senior secondary vocational education

[d] Senior general secondary education; higher professional education; university education

and ADEs, as is done in other checklists [23], ensuring that respondents maintain the distinction while completing the questionnaire. This mechanism may explain in part why more ADEs are reported in checklists than in open-ended questionnaires [13].

Several patients reported that a recall period of 4 weeks was quite short, for instance, to capture ADEs that fluctuate over time, as has been identified before [48]. On the other hand, the period should not be too long when the aim is to collect information on symptomatic ADEs that can be mild in nature. The optimal recall period may depend on the nature of the ADE [48]. Although a recall period of 4 weeks is quite common, and even shorter recall periods have been used in ADE questionnaires [17], the reliability of various recall periods needs to be tested in further studies.

Reducing respondent burden is relevant for the feasibility of using the questionnaire. We identified problems in navigating the questionnaire and these were solved by formatting the questionnaire along principles of cognitive design [49]. Around half of the patients found the body category structure helpful, but we detected some difficulties with our initial ADE classification based on the MedDRA® System Organ Classes. We thus adapted this to a more patient-based classification system. The feasibility

test showed, however, that the categorization structure only marginally decreased the time to complete the questionnaire for patients reporting at least one ADE. Only four ADEs were reported as "other," indicating that most patients were able to identify their experienced ADE within the provided lists. For most of the patients reporting at least one ADE, the time needed to complete the questionnaire was <60 min. In our opinion, this time is acceptable for a questionnaire intended for research purposes, in which questions about general characteristics and drug use were included. It should, however, be noted that only a quarter of patients reported at least one ADE. The majority of the patients agreed that the questionnaire was easy to use, but this number was lower for those reporting an ADE than those reporting no ADEs. Of the patients who opened the questionnaire, around 10 % were lost to follow-up.

Although the test–retest reliability of the patient-reported ADE questionnaire was considered acceptable at patient level and at MedDRA® level, it was below the threshold of 0.6–0.8 recommended for reliability coefficients [50]. For ADE reporting, however, a skewed distribution is observed where many patients report no ADEs on both measurements, which decreases the kappa values used for the reliability assessment [51, 52]. Formulas to adjust for such effects have been proposed, for example, the

Table 5 Nature and causality reported at adverse drug event (ADE) level for the three groups at first (T1) and second (T2) measurement

	T1 (N = 173)	T2 (N = 146)
MedDRA® System Organ Class (%):		
Cardiac disorders	1 (0.6)	0 (0.0)
Ear and labyrinth disorders	1 (0.6)	1 (0.7)
Eye disorders	17 (9.8)	6 (4.1)
Gastrointestinal disorders	53 (30.6)	37 (25.3)
General disorders and administration site conditions	11 (6.4)	10 (6.8)
Infections and infestations	2 (1.2)	1 (0.7)
Injury, poisoning, and procedural complications	0 (0.0)	1 (0.7)
Investigations	4 (2.3)	9 (6.2)
Metabolism and nutrition disorders	5 (2.9)	6 (4.1)
Musculoskeletal and connective tissue disorders	4 (2.3)	5 (3.4)
Nervous system disorders	31 (17.9)	17 (11.6)
Psychiatric disorders	9 (5.2)	22 (15.1)
Renal and urinary tract disorders	2 (1.2)	2 (1.4)
Reproductive system and breast disorders	8 (4.6)	8 (5.5)
Respiratory, thoracic, and mediastinal disorders	6 (3.5)	7 (4.8)
Skin and subcutaneous tissue disorders	14 (8.1)	10 (6.8)
Vascular disorders	5 (2.9)	4 (2.7)
Number of times a reason for a relationship among drug and ADE was reported		
I did not experience this symptom before I started taking the drug	97	87
The symptom started soon after I started taking the drug	49	33
I experienced this symptom less often before I started taking the drug	20	17
The symptom was less serious before I started taking the drug	8	8
The symptom went away when I stopped taking the drug and came back when I started taking it again	2	2
The symptom went away when I stopped taking the drug	5	1
The symptom started or grew worse when the drug dosage was increased	14	9
The symptom decreased or went away when the drug dosage was decreased	2	1
A healthcare professional (a doctor or pharmacist) confirmed this	20	22
The symptom is described in the patient leaflet	34	22
Other	11	12
Number of times it was not known by the patient which drug caused the ADE (%)	75 (43.4)	61 (41.8)
Patients' certainty about the relationship among the reported ADE and reported drug (%)[a]		
Very sure	30 (17.3)	21 (12.1)
Quite sure	48 (27.7)	50 (28.9)
Not very sure	14 (8.1)	7 (4.0)
Very unsure	2 (1.2)	4 (2.3)

[a] Percentages do not sum to 100 % due to roundings or not indicating a causal drug

as an alternative agreement measure, which showed similar results. Future studies assessing the reliability of ADE reporting are advised to recruit a more balanced group of patients experiencing and not experiencing ADEs [51]. Based on a combined approach, that is, looking at kappa values, alternative agreement measures, and additional analysis of ADEs at patient level, we conclude that our questionnaire was not sufficiently reliable at the ADE-specific level. This result implies that the distinct symptoms reported by patients as ADEs using these checklists should not be used blindly to quantify rates at the lowest ADE-specific level. Part of the lack of reliability might be solved by improving the questionnaire, but some lack of reliability at the lowest ADE level could be inherent to patient reporting.

One can expect that uncertainty by patients about a symptom being an ADE may lead to inconsistent answers. The finding that some patients checked a symptom as an ADE on one measurement but not on the other indicates such uncertainty. Furthermore, in around half of the cases the patients did not mention a potential drug that they believed was causing that specific ADE or were not very sure about the causal relationship. On the other hand, some of the inconsistency was caused by using a checklist that does not require differentiation between related and disparate ADEs. Patients often checked multiple related ADEs, but not exactly the same ADEs on the two measurements. When aggregated at MedDRA® level or using the patient's own descriptions, patients were therefore found to be more consistent. This problem could be a consequence of direct patient reporting; that is, reporting without involvement of a healthcare professional who can interpret and cluster specific symptoms to a more general ADE description. However, a more intelligent questionnaire flow or an interactive questionnaire, might solve this problem. For instance, using an interactive questionnaire requiring patients to cluster related symptoms that are considered as one problem before they move to answer more detailed questions. Such a questionnaire should incorporate a more flexible linkage to the MedDRA® System Organ Class by not only focusing on the primary MedDRA® class. This prevents symptoms with different primary MedDRA® classes used to describe one ADE being classified in different MedDRA® classes. Notwithstanding these possible improvements to the questionnaire, some patients clearly checked totally different ADEs at the two measurements. We chose a period of 1 week between the measurements to exclude memory effects, but this period may have been too long to exclude true changes in the experience of ADEs in the previous 4 weeks, especially for ADEs that might change from day to day [12].

The comparison between the questionnaire with and without the body category structure showed no significant

prevalence-adjusted bias-adjusted kappa [53], but their inappropriateness has also been demonstrated [51]. We therefore calculated the proportions of positive agreement

Table 6 Time in minutes needed to complete the questionnaire with and without body categories for reporting no adverse drug event (ADE) and one or more ADEs

	With body categories (group with body categories at T1 + test–retest group, N = 90)	Without body categories (group with body categories at T2, N = 45)	P-value
No ADE reported			
Median self-reported time needed to complete questionnaire (range; SD)	15 (3–40; 7.7)	13 (2–60; 10.7)	0.377
Median registered time needed to complete questionnaire (range; SD)	17 (5–48; 10.1)	16 (7–82; 16.2)	0.720
One or more ADEs reported			
Median self-reported time needed to complete questionnaire (range; SD)	23 (15–240; 63.6)	40 (20–120; 29.0)	0.166
Median registered time needed to complete questionnaire (range; SD)	54 (22–96; 24.1)	71 (32–147; 36.5)	0.115

SD standard deviation, *T1* first measurement, *T2* second measurement after 1-week period

Table 7 Kappa values and proportion of positive agreement for test–retest reliability and body categories at patient level, MedDRA® level, and ADE-specific level

	Test–retest group		Group with body categories at T1		Group with body categories at T2		Differences in kappa values between group with body categories at T1 and T2
	Kappa values (95 % CI)	PPA	Kappa values (95 % CI)	PPA	Kappa values (95 % CI)	PPA	Z-value
Patient level (N = 45)	0.502 (0.21–0.79)	0.64	0.639 (0.37–0.91)	0.73	0.433 (0.12–0.74)	0.58	0.387
MedDRA® level (N = 810)	0.521 (0.35–0.69)	0.54	0.395 (0.19–0.60)	0.42	0.264 (0.12–0.40)	0.30	0.330
ADE specific level (N = 11,340)	0.380 (0.24–0.52)	0.38	0.259 (0.06–0.46)	0.26	0.158 (0.003–0.31)	0.16	0.301

PPA calculated by the formula $2a/[N + (a - d)]$, in which N = total number of observations, a = patients reporting ADE at T1 and T2, and d = patients not reporting ADE at T1 and T2

T1 first measurement, *T2* second measurement after 1 week period, *PPA* proportion positive agreement, *CI* confidence interval

differences in the number of reported ADEs or in agreement measures. From this we conclude that including a body category system did not influence the reliability of the ADE reporting. Because the cognitive debriefing showed that the body categories were helpful and increased the feasibility for some patients, we still recommend the use of such a categorization as a supportive element.

To our knowledge, this is the first study to validate a generic patient-reported questionnaire intended for systematic data collection of ADEs. We conducted a broad search for symptomatic ADEs, which we translated in lay-terms and linked to MedDRA® terms. The use of these standard terms makes it possible to compare ADE data across different studies, which is important in the evaluation of drug safety [54]. We included a heterogeneous population with respect to age and education level in the content-validation study. Patients were selected for having type 2 diabetes, asthma, or COPD, but many of them used multiple drugs, and also used drugs for other diseases. We expect that the questionnaire is suitable for adult patients on a steady drug regimen who are able to read and write. We cannot, however, guarantee that all ADE terms are content valid. In addition, we tested the Dutch version of the questionnaire. The use of the questionnaire in other languages requires additional testing [55]. We expect that the reliability for ADE reporting of the Web-based version is comparable to the paper-based version. The navigation through the questionnaire and the time needed to complete, however, may differ between the Web-based and paper-based versions [56]. We tested the questionnaire in an observational, postmarketing setting. We expect that the questionnaire is also applicable in clinical trials in which patients are initial drug users, but this should be confirmed in future studies. Further validation studies are needed (e.g., establishing the probability of a causal relationship between the reported ADEs and the drugs using an external reference) because content validation is an essential but only first step in providing evidence of full validity [12, 57, 58].

5 Conclusions

Participants in postmarketing studies and clinical trials can use multiple drugs that may interact and cause unexpected ADEs. Using a generic questionnaire in which all experienced ADEs can be reported by patients is therefore important. In terms of content validity, our patient-reported ADE questionnaire can be used for assessing the nature and causality of symptomatic ADEs as experienced by patients undergoing chronic drug therapy. The questionnaire is feasible for research purposes, and reliable to identify numbers of patients experiencing ADEs in general and at MedDRA® System Organ Class level. To quantify specific patient-reported ADEs, improvements to the structure of the questionnaire are required.

References

1. Meadows KA. Patient-reported outcome measures: an overview. Br J Community Nurs. 2011;16(3):146–51.
2. Zhu J, Stuver SO, Epstein AM, et al. Can we rely on patients' reports of adverse events? Med Care. 2011;49(2):948–55.
3. Jarernsiripornkul N, Krska J, Capps PA, et al. Patient reporting of potential adverse drug reactions: a methodological study. Br J Clin Pharmacol. 2002;53(3):318–25.
4. Weingart SN, Gandhi TK, Seger AC, et al. Patient-reported medication symptoms in primary care. Arch Intern Med. 2005;165(2):234–40.
5. Basch E, Jia X, Heller G, et al. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. J Natl Cancer Inst. 2009;101(23):1624–32.
6. Wetzels R, Wolters R, van Weel C, et al. Mix of methods is needed to identify adverse events in general practice: a prospective observational study. BMC Fam Pract. 2008;9:35.
7. Blenkinsopp A, Wilkie P, Wang M, et al. Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. Br J Clin Pharmacol. 2007;63(2):148–56.
8. Hakobyan L, Haaijer-Ruskamp FM, de Zeeuw D, et al. A review of methods used in assessing non-serious adverse drug events in observational studies among type 2 diabetes mellitus patients. Health Qual Life Outcomes. 2011;9:83.
9. US Department of Health and Human Services, FDA, Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Center for Devices and Radiological Health. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims [online]. http://www.fda.gov/downloads/Drugs/GuidanceCompliance RegulatoryInformation/Guidances/UCM193282.pdf. Accessed 14 Nov 2012.
10. European Medicines Agency. Fourth report on the progress of the interaction with patients' and consumers' organizations (2010) and results/analysis of the degree of satisfaction of patients and consumers involved in EMA activities during 2010. EMA/632696/2011 [online]. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/10/WC500116866.pdf. Accessed 14 Nov 2012.
11. European Medicines Agency. Note for guidance on clinical investigation of medicinal products for treatment of nociceptive pain [online]. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003525.pdf. Accessed 14 Nov 2012.
12. Frost MH, Reeve BB, Liepa AM, et al. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health. 2007;10(Suppl 2):S94–105.
13. Bent S, Padula A, Avins AL. Brief communication: better ways to question patients about adverse medical events: a randomized, controlled trial. Ann Intern Med. 2006;144(4):257–61.
14. Sheftell FD, Feleppa M, Tepper SJ, et al. Assessment of adverse events associated with triptans—methods of assessment influence the results. Headache. 2004;44(10):978–82.
15. Ruiz MA, Pardo A, Rejas J, et al. Development and validation of the "Treatment Satisfaction with Medicines Questionnaire" (SATMED-Q). Value Health. 2008;11(5):913–26.
16. Bytzer P, Talley NJ, Jones MP, et al. Oral hypoglycaemic drugs and gastrointestinal symptoms in diabetes mellitus. Aliment Pharmacol Ther. 2001;15(1):137–42.
17. Foster JM, van Sonderen E, Lee AJ, et al. A self-rating scale for patient-perceived side effects of inhaled corticosteroids. Respir Res. 2006;7:131.
18. Sitzia J, Dikken C, Hughes J. Psychometric evaluation of a questionnaire to document side-effects of chemotherapy. J Adv Nurs. 1997;25(5):999–1007.
19. Corso DM, Pucino F, DeLeo JM, et al. Development of a questionnaire for detecting potential adverse drug reactions. Ann Pharmacother. 1992;26(7–8):890–6.
20. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH harmonised tripartite guideline pharmacovigilance planning E2E [online]. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2E/Step4/E2E_Guideline.pdf. Accessed 14 Nov 2012.
21. Harmark L, Puijenbroek E, Grootheest K. Longitudinal monitoring of the safety of drugs by using a web-based system: the

case of pregabalin. Pharmacoepidemiol Drug Saf. 2011;20(6): 591–7.

22. US Department of Health and Human Services, National Institutes of Health National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE) version 4 [online]. http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_Quick Reference_8.5x11.pdf. Accessed 14 Nov 2012.

23. Uher R, Farmer A, Henigsberg N, et al. Adverse reactions to antidepressants. Br J Psychiatry. 2009;195(3):202–10.

24. Welch V, Singh G, Strand V, et al. Patient based method of assessing adverse events in clinical trials in rheumatology: the revised Stanford Toxicity Index. J Rheumatol. 2001;28(5): 1188–91.

25. Basch E, Iasonos A, Barz A, et al. Long-term toxicity monitoring via electronic patient-reported outcomes in patients receiving chemotherapy. J Clin Oncol. 2007;25(34):5374–80.

26. Brostrom A, Stromberg A, Martensson J, et al. Association of Type D personality to perceived side effects and adherence in CPAP-treated patients with OSAS. J Sleep Res. 2007;16(4): 439–47.

27. Day JC, Wood G, Dewey M, et al. A self-rating scale for measuring neuroleptic side-effects. Validation in a group of schizophrenic patients. Br J Psychiatry. 1995;166(5):650–3.

28. Brown V, Sitzia J, Richardson A, et al. The development of the Chemotherapy Symptom Assessment Scale (C-SAS): a scale for the routine clinical assessment of the symptom experiences of patients receiving cytotoxic chemotherapy. Int J Nurs Stud. 2001;38(5):497–510.

29. The Ohio State University Medical Center. Checklist for reporting side effects [online]. https://patienteducation.osumc.edu/Pages/search.aspx?k=checklist+for+reporting. Accessed 14 Nov 2012.

30. Harmark L, van Grootheest K. Web-based intensive monitoring: from passive to active drug surveillance. Expert Opin Drug Saf. 2012;11(1):45–51.

31. DeWitt JE, Sorofman BA. A model for understanding patient attribution of adverse drug reaction symptoms. Drug Inf J. 1999;33:907–20.

32. De Smedt RH, Jaarsma T, Ranchor AV, et al. Coping with adverse drug events in patients with heart failure: exploring the role of medication beliefs and perceptions. Psychol Health. 2012;27(5):570–87.

33. Harmark L, van Puijenbroek E, Straus S, et al. Intensive monitoring of pregabalin: results from an observational, web-based, prospective cohort study in the Netherlands using patients as a source of information. Drug Saf. 2011;34(3):221–31.

34. Grootenhuis PA, Snoek FJ, Heine RJ, et al. Development of a type 2 diabetes symptom checklist: a measure of symptom severity. Diabet Med. 1994;11(3):253–61.

35. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30(6):473–83.

36. Naranjo CA, Busto U, Sellers EM, et al. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther. 1981;30(2):239–45.

37. Krska J, Anderson C, Murphy E, et al. How patient reporters identify adverse drug reactions: a qualitative study of reporting via the UK Yellow Card Scheme. Drug Saf. 2011;34(5):429–36.

38. Ploughman M, Austin M, Stefanelli M, et al. Applying cognitive debriefing to pre-test patient-reported outcomes in older people with multiple sclerosis. Qual Life Res. 2010;19(4):483–7.

39. Collins D. Pretesting survey instruments: an overview of cognitive methods. Qual Life Res. 2003;12(3):229–38.

40. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. J Am Med Inform Assoc. 2000; 7(4):416–25.

41. Urbaniak GC, Plous S. Research Randomizer (version 3.0) [online computer software]. http://www.randomizer.org/. Accessed 14 Nov 2012.

42. De Vet H, Terwee C, Mokkink L, et al. Measurement in medicine: a practical guide. Cambridge: University Press; 2011.

43. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6):551–8.

44. St Sauver JL, Hagen PT, Cha SS, et al. Agreement between patient reports of cardiovascular disease and patient medical records. Mayo Clin Proc. 2005;80(2):203–10.

45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

46. Foody GM. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. Photogramm Eng Remote Sensing. 2004;70(5):627–33.

47. De Smedt RH, Denig P, Haaijer-Ruskamp FM, et al. Perceived medication adverse effects and coping strategies reported by chronic heart failure patients. Int J Clin Pract. 2009;63(2):233–42.

48. Stull DE, Leidy NK, Parasuraman B, et al. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Curr Med Res Opin. 2009;25(4):929–42.

49. Mullin PA, Lohr KN, Bresnahan BW, et al. Applying cognitive design principles to formatting HRQOL instruments. Qual Life Res. 2000;9(1):13–27.

50. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96–106.

51. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol. 2000; 53(5):499–503.

52. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85(3):257–68.

53. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423–9.

54. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for industry Premarketing risk assessment [online]. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126958.pdf. Accessed 14 Nov 2012.

55. Coyne KS, Tubaro A, Brubaker L, et al. Development and validation of patient-reported outcomes measures for overactive bladder: a review of concepts. Urology. 2006;68(2 Suppl):9–16.

56. Dillman DA, Gertseva A, Mahon-Haft T. Achieving usability in establishment surveys through the application of visual design principles. J Off Stat. 2005;21(2):183–214.

57. Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: developing best practices based on science and experience. Qual Life Res. 2009;18(9):1263–78.

58. Rothman M, Burke L, Erickson P, et al. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report. Value Health. 2009;12(8): 1075–83.