The Utrecht questionnaire (U-CEP) measuring knowledge on clinical epidemiology proved to be valid

Marlous F. Kortekaas*, Marie-Louise E.L. Bartelink, Esther de Groot, Helen Korving, Niek J. de Wit, Diederick E. Grobbee, Arno W. Hoes

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CX Utrecht, The Netherlands

Accepted 19 August 2016; Published online xxxx

Abstract

Objectives: Knowledge on clinical epidemiology is crucial to practice evidence-based medicine. We describe the development and validation of the Utrecht questionnaire on knowledge on Clinical epidemiology for Evidence-based Practice (U-CEP); an assessment tool to be used in the training of clinicians.

Study Design and Setting: The U-CEP was developed in two formats: two sets of 25 questions and a combined set of 50. The validation was performed among postgraduate general practice (GP) trainees, hospital trainees, GP supervisors, and experts. Internal consistency, internal reliability (item-total correlation), item discrimination index, item difficulty, content validity, construct validity, responsiveness, test—retest reliability, and feasibility were assessed. The questionnaire was externally validated.

Results: Internal consistency was good with a Cronbach alpha of 0.8. The median item-total correlation and mean item discrimination index were satisfactory. Both sets were perceived as relevant to clinical practice. Construct validity was good. Both sets were responsive but failed on test—retest reliability. One set took 24 minutes and the other 33 minutes to complete, on average. External GP trainees had comparable results.

Conclusion: The U-CEP is a valid questionnaire to assess knowledge on clinical epidemiology, which is a prerequisite for practicing evidence-based medicine in daily clinical practice. © 2016 Elsevier Inc. All rights reserved.

Keywords: Validation; Evidence-based medicine; Knowledge; Questionnaire; Clinical epidemiology; Assessment

1. Introduction

Knowledge of clinical epidemiology is crucial to be able to practice evidence-based medicine (EBM) in daily clinical practice [1]. Practicing EBM implies the ability to combine the best available evidence with the clinician’s expertise and the patient’s preferences [2]. Clinical epidemiology focuses on four important challenges clinicians are faced with. First, how to accurately diagnose a patient’s illness (diagnosis, D), second to determine what causes the disease (etiology, E), third how to predict the natural history of the disease in an individual patient (prognosis, P), and fourth to estimate effect of interventions on a patient’s prognosis (therapy, Th). In routine clinical practice, these four domains are incorporated in medical decision making, following the so-called DEPTh model [1]. Clinical epidemiology provides the framework and knowledge and skills for practitioners to critically appraise research evidence and translate outcomes of research into use in daily clinical practice. Given its importance for adequate evidence-based practicing in the future, monitoring theoretical knowledge on clinical epidemiology is important in the training of clinicians.

Testing knowledge needed to practice EBM is essential in clinicians [3] and should focus on those aspects useful in clinical practice. The second Sicily Statement pointed out that for a useful evaluation of EBM training it should be clear which aspect(s) such an assessment instrument intends to measure [4]. A number of questionnaires developed for testing knowledge needed to practice EBM exist already [5–7], but in our view, these do not prioritize clinical relevance, are time consuming to score, or assess therapeutic issues only. Importantly, developers of those questionnaires often provide only minimal data on validation [5–7].
What is new?

Key findings
- The Utrecht questionnaire on knowledge on Clinical epidemiology for Evidence-based Practice (U-CEP) is a questionnaire that assesses knowledge on clinical epidemiology covering three different aspects of evidence-based medicine (EBM: ask, appraise, and apply) and different aspects relevant to daily clinical practice (diagnosis, etiology, prognosis, therapy [DEPTh]).
- The U-CEP was shown to be valid among different clinicians, such as general practice trainees, hospital trainees, and general practice supervisors.
- The U-CEP is responsive to change and is therefore valid to monitor change in knowledge of clinicians on clinical epidemiology after EBM training.
- The U-CEP performs moderately on the test–retest reliability.

What this adds to what was known?
- The U-CEP is the first questionnaire assessing knowledge on clinical epidemiology on all four clinically important challenges clinicians are faced with in daily clinical practice (DEPTh), addressing different aspects of EBM.

What is the implication and what should change now?
- Clinical epidemiology provides the framework, knowledge, and skills for practitioners to critically appraise research evidence and translate outcomes of research into use in daily clinical practice. Clinical epidemiology focuses on the four key questions in clinical practice (diagnosis, etiology, prognosis, and therapy; DEPTh). Given its importance for adequate evidence-based practicing in the future, monitoring theoretical knowledge on all aspects of clinical epidemiology is important in the training of clinicians.

We previously developed an EBM training program for the vocational training of general practitioners. Focus of the program is the decision process in primary care, and we aim to integrate the training as much as possible into daily clinical practice [8]. The EBM training is strongly based on dilemmas derived from clinical practice and focuses on relevant outcomes for patients. It covers all clinical domains because many clinical queries pertain not only to therapeutic but also to diagnostic or prognostic topics as well [8].

We report on the development and validation of the Utrecht questionnaire on knowledge on Clinical epidemiology for Evidence-based Practice (U-CEP), a questionnaire suitable for the evaluation of EBM training, with a focus on those aspects relevant to clinical practice.

2. Methods

2.1. Development of the U-CEP

We postulated that an optimal questionnaire should address the content of EBM training, cover as many different aspects of EBM (ask, acquire, appraise, apply, and assess) as possible, contain questions on clinically relevant aspects with an equal distribution across the different types of clinically relevant research (DEPTh), and test the minimal required methodological knowledge to be able to translate research results to clinical practice. At first, we used our experiences as teachers of EBM to include questions on difficulties clinicians frequently encounter in the interpretation of research findings and their use in daily clinical practice. We devised an initial set of 95 items based on the most relevant themes in clinical epidemiology [1]. Experts in the development of summative assessment helped to devise instructions, response options (e.g., not including the do not know option), and rules for scoring [9]. We exchanged opinions about the first drafts between four experienced teachers in EBM (M.E.L.B., G.v.d.H., M.F.K., and N.J.d.W.) and adapted elements of the questionnaire accordingly. Finally, the list of 95 items was judged and adapted by two senior clinical epidemiologists (A.W.H. and D.E.G.). Although there seems to be no agreement on the optimum length of a questionnaire, we aimed to develop a questionnaire that was both as long as needed and as short as possible. We reduced the number of items in the questionnaire on the basis of an item analysis using the scores of respondents. For this, we used the data from those who had followed EBM training or were expert on EBM (Table 1).

2.2. Validation

2.2.1. Population and setting

The validation was performed among 219 postgraduate general practice (GP) trainees (180 first-year and 39 last [i.e., third]-year), 20 hospital trainees, 20 GP supervisors, and 8 expert academic GPs or clinical epidemiologists from the University Medical Center Utrecht (UMCU). Characteristics of participants, such as age, gender, time since graduation, PhD degree, and self-perceived knowledge on EBM, were collected through an online survey. The EBM training that the first-year GP trainees and hospital trainees received was a 2-day course in which essential skills, such as searching for evidence, critical appraisal of the literature for different research designs, and basic analytic skills were taught in accordance with the five steps of EBM training.
<table>
<thead>
<tr>
<th>Test property</th>
<th>Measure used</th>
<th>Participants</th>
<th>Acceptable results</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content validity</strong> (test covers entire topic of interest)</td>
<td>Expert opinion and survey</td>
<td>4 experts and 39 third-year GP trainees</td>
<td>Questionnaire covers all main aspects of EBM that are relevant to clinical practice [5]</td>
<td>Set A 25 Questions (Maximum score, 33)</td>
</tr>
<tr>
<td><strong>Construct validity</strong> (evidence that the test measures the construct it intends to)</td>
<td>Mean scores (95% CI) of experts, trainees, and supervisors; the latter two after EBM training, compared by ANOVA</td>
<td>76 participants: 8 experts, 49 trainees (29 first-year GP trainees and 20 hospital trainees), and 19 supervisors</td>
<td>Significant difference in mean scores, with experts performing better than trainees and supervisors performing better than supervisors</td>
<td>Mtrainees 25.7 (24.4–27.1)</td>
</tr>
<tr>
<td><strong>Internal consistency</strong> (degree to which all test questions on the test measure a single construct)</td>
<td>Cronbach alpha—average of all possible split half correlations</td>
<td>154 participants: 49 trainees (29 first-year GP trainees and 20 hospital trainees) and 19 supervisors after EBM training, 8 experts, and 78 GP trainees who studied medicine at the same university as the postgraduate training program</td>
<td>≥0.6–0.7 is considered acceptable, ≥0.7–0.9 is considered good, and ≥0.9 is considered perfect [7]</td>
<td>α = 0.79</td>
</tr>
<tr>
<td><strong>Internal reliability</strong> (the correlation between the question score and the overall score on the questionnaire)</td>
<td>ITC using Pearson product (correlation)</td>
<td>154 participants: 49 trainees (29 first-year GP trainees and 20 hospital trainees) and 19 supervisors after EBM training, 8 experts, and 78 GP trainees who studied medicine at the same university as the postgraduate training program</td>
<td>≥0.15–0.20 is considered satisfactory, ≥0.20–0.40 is considered good, and ≥0.40 is considered excellent [7]</td>
<td>Median, 0.22 (IQR, 0.14; range, 0.01–0.78)</td>
</tr>
<tr>
<td><strong>Item discrimination</strong> (ability of each item to discriminate between those with overall high scores and those with overall low scores)</td>
<td>IDI calculated for each item separately</td>
<td>154 participants: 49 trainees (29 first-year GP trainees and 20 hospital trainees) and 19 supervisors after EBM training, 8 experts, and 78 GP trainees who studied medicine at the same university as the postgraduate training program</td>
<td>&gt;0.20–0.40 is considered satisfactory, ≥0.40 is considered high [7]</td>
<td>Median, 0.35 (IQR, 0.23; range, 0.13–0.55)</td>
</tr>
<tr>
<td><strong>Item difficulty</strong> (relative)</td>
<td>Percentage of participants</td>
<td>76 participants: 8 experts, 49 trainees</td>
<td>Wide range (10%–90%) of</td>
<td>Ranged from 16% to 100% with 4 questions</td>
</tr>
</tbody>
</table>

(Continued)
### Table 1. Continued

<table>
<thead>
<tr>
<th>Test property</th>
<th>Measure used</th>
<th>Participants</th>
<th>Acceptable results</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>difficulty of each item</td>
<td>(who answer) answer a question correctly after EBM training, including experts</td>
<td>(29 first-year GP trainees and 20 hospital trainees), and 19 supervisors</td>
<td>difficulties allows a test to be used with both experts and novices [5]</td>
<td>Set A 25 Questions (Maximum score, 33)</td>
</tr>
<tr>
<td>Floor and ceiling effects (present when more than 15% of the participants achieve highest or lowest possible scores)</td>
<td>Percentage of participants achieving highest or lowest score after EBM training, including experts</td>
<td>76 participants: 8 experts, 49 trainees (29 first-year GP trainees and 20 hospital trainees), and 19 supervisors</td>
<td>&lt;15% achieves highest or lowest possible score [10]</td>
<td>Set B 25 Questions (Maximum score, 34)</td>
</tr>
<tr>
<td>Test–retest reliability (consistency of a test)</td>
<td>Pearson correlation of mean scores (±SD) of participants who filled in the questionnaire twice without EBM training in the meantime</td>
<td>18 first-year GP trainees</td>
<td>No significant differences between the mean scores (95% CI) of both groups</td>
<td>Set A and B combined 50 Questions (Maximum score, 66)</td>
</tr>
<tr>
<td>Responsiveness (questionnaire detects change over time in the construct to be measured)</td>
<td>Mean scores (±SD) of participants who filled in the questionnaire before and after EBM training, compared by paired-samples t-test</td>
<td>65 participants: 49 trainees (29 first-year GP trainees and 20 hospital trainees) and 16 supervisors</td>
<td>Significant better mean scores within groups after EBM training compared with before EBM training</td>
<td>Results are shown in Table 3</td>
</tr>
<tr>
<td>External validity (the capability of a questionnaire to be used in other settings as well)</td>
<td>Comparison of mean scores (±SD) of GP trainees from the UMCU with GP trainees from the AMC, both after their own EBM training, compared by independent-samples t-test</td>
<td>101 participants: 47 GP trainees from the UMCU and 54 GP trainees from the AMC</td>
<td>To be externally valid; no significant differences in mean scores (95% CI) between both groups</td>
<td>Results are shown in Table 3</td>
</tr>
</tbody>
</table>

**Abbreviations:** U-CEP, Utrecht questionnaire on knowledge on Clinical epidemiology for Evidence-based Practice; GP, general practice; EBM, evidence-based medicine; 95% CI, 95% confidence interval; ANOVA, analysis of variance; ITC, item-total correlation; IQR, interquartile range; IDI, item discrimination index; SD, standard deviation; UMCU, University Medical Center Utrecht; AMC, Academic Medical Center.
EBM training in the development of the questionnaire. We also assessed content validity by asking a convenience sample of 39 third-year GP trainees for their opinion on difficulty and relevance (all on a five-point Likert scale ranging from 1 [very easy/poor] to 5 [very difficult/good]) and the score they expected to have (ranging from 1 [very poor] to 10 [very good]). Correlations between expected and achieved scores were calculated using Pearson correlation. Floor and ceiling effects were assessed as well, defined as the range and distribution of scores [10]. We considered the questionnaire valid on this item when less than 15% of participants achieved the highest or lowest score after EBM training because otherwise no improvement or deterioration of knowledge can be assessed [10].

2.3.2. Content validity
Content validity was ascertained by involving experts in EBM training in the development of the questionnaire. We considered the questionnaire valid, according to the experts (denominator) who answered a question, we calculated the proportion of all participants who correctly answered each item from the proportion of participants in the bottom quartile who correctly answered the item [15]. It ranges between −1 and +1. The ITC was defined as the correlation between the question score and the overall score on the questionnaire [5]. The item difficulty was defined as the relative difficulty of an item. Per question, we calculated the proportion of all participants including the experts (denominator) who answered a question correctly after EBM training (numerator).

We considered the questionnaire valid, according to the individual item analysis, when internal consistency was good (Cronbach alpha > 0.7), containing questions with a positive ITC (>0.15) and a positive IDI (>0.2), and no questions were answered correctly by less than 10% or more than 90% of the participants after EBM training [10].

2.3.3. Construct validity
Construct validity was defined as the degree to which the scores of the questionnaire were consistent with hypothesized scores [16]. Construct validity was measured by comparing the mean scores of groups of participants whom we expected to score differently on the questionnaire; trainees, supervisors, and experts, both before and after EBM training. Because no differences between GP and hospital trainees were expected, we analyzed them as one group. The three groups were compared using the analysis of variance test. Experts were expected to perform best. Trainees were expected to perform significantly better than supervisors because they are used to the concept of EBM during their undergraduate training programs in contrast to most GP supervisors. Moreover, we expected GP trainees to have higher perceived need to learn.

2.3.4. Test—retest reliability
Test—retest reliability was defined as the stability of the scores between two moments of filling in the questionnaire. To calculate the test—retest reliability, we asked 18 third-year GP trainee participants to fill in the questionnaire twice before EBM training; the second time 3 months after the first to avoid recall bias. Because of logistical reasons, the first time the questionnaire was filled in on paper. We compared mean scores on the two questionnaires using the paired-samples t-test.

2.3.5. Responsiveness
Responsiveness was defined as the ability of the questionnaire to detect change over time in the construct to be measured [16]. We compared the mean scores of participants before EBM training with their mean scores after EBM training. Responses of participants were analyzed as one group and separately per group (GP trainees, hospital trainees, and GP supervisors) using the paired-samples t-test.

2.3.6. Feasibility
Feasibility of a questionnaire depends on time needed to complete and score a questionnaire. We assessed time to completion (in minutes) and time needed to check and...
score (in minutes) among a convenience sample of 39 third-year GP trainees.

2.3.7. External validation

The questionnaire was externally validated among GP trainees from another university medical center, the Academic Medical Center (AMC) in Amsterdam. Mean scores of GP trainees from the AMC were compared with mean scores of GP trainees from the UMCU using the independent-samples t-test.

Statistical analyses were performed using Statistical Package for the Social Sciences V.20.0 (SPSS, Chicago, Illinois, USA). Results were considered statistically significant at the $P < 0.05$ level. Corrections for multiple testing were introduced to reduce the risk of bias.

3. Results

3.1. Final format of the U-CEP

The shortening process was based on the results of the individual item analysis on internal reliability and consistency, derived from the scores of the respondents ($N = 154$; 49 trainees [29 first-year GP trainees and 20 hospital trainees] and 19 supervisors after EBM training, 8 experts, and 78 GP trainees who studied medicine at the same university as the postgraduate training program). We started with a 95-item questionnaire and removed the question with the lowest ITC first ($-0.073$). We then recalculated the Cronbach alpha to check whether the internal consistency remained the same or became better (in our case, Cronbach alpha changed from 0.858 to 0.860). We then removed the next question with the (then) lowest ITC ($-0.048$). This procedure continued until we had to remove a question that would result in an unequal distribution of the questions across the four different clinical domains (DEPTHs). Because we aimed at developing two questionnaires and we wanted to have at least two questions per domain per set, the minimum number of questions per domain was four. When the total number of questions per domain would become less than four after removal as based on the (lowest) ITC, we did not remove the question but continued and removed the next worst question (based on the ITC). This resulted in a removal of 55 questions with ITCs varying between $-0.13$ and 0.19, resulting in a 40-item questionnaire (Cronbach alpha, 0.90) with three questions with an ITC <0.15. For two domains, the minimum amount of four questions was reached. We then looked which questions were removed and reinserted 10 questions we believed were essential to know. The 50 questions were divided over two comparable sets of 25 items each as this was regarded useful when different sets of questions before and after EBM training were necessary. Results on psychometric properties of the two separate, but comparable, sets of 25 questions each and the combined set (50 questions) are shown in Table 1.

Later, we report on the two sets of 25 questions only. Both sets of the U-CEP were developed and validated in Dutch. These two versions were subsequently translated into English using forward—backward translation, and the English versions were checked for inconsistencies by two senior clinical epidemiologists. The English versions have not yet been applied in a teaching setting.

As shown in the supplementary, both sets of 25 questions have 6 open-ended and 19 (single-best only) multiple-choice questions. Most questions are scored one for a correct answer and zero for an incorrect answer, four open-ended questions are scored zero to three with one for each part of the question; domain (one point), determinants (one point), and outcome (one point). The maximum score for set A is 33 and set B 34. Both sets cover the different aspects of EBM well, with set A containing 7 questions on ask, 14 on appraise, and 4 on apply. Set B contains 6 questions on ask, 13 on appraise, and 6 on apply. The questions are equally distributed across the different clinical domains in both set A (2 diagnostic, 3 etiologic, 3 prognostic, and 4 therapeutic questions) and set B (4 diagnostic, 1 etiologic, 2 prognostic, and 4 therapeutic questions). Both sets contain calculation questions and general questions about clinical epidemiology (e.g., “What is the most important reason to ask patients for informed consent?”).

3.2. Population

One hundred fifty participants filled in the questionnaire before EBM training only (149 first-year GP trainees and 1 GP supervisor), 3 GP supervisors filled in the questionnaire after EBM training only, and 65 participants filled in the questionnaire both before and after EBM training (29 first-year GP trainees, 20 hospital trainees, and 16 GP supervisors). Eight experts and 39 third-year GP trainees filled in the questionnaire once. Characteristics of participants are shown in Table 2.

3.3. Validity measures

3.3.1. Individual item analysis

The Cronbach alpha was 0.79 for set A and 0.80 for set B. For set A, the median ITC was 0.22 (interquartile range [IQR], 0.14; range, 0.01–0.78) and for set B 0.26 (IQR, 0.14; range, 0.01–0.74). No questions had a negative ITC. In both sets, four questions (16%) had an ITC below 0.15. The median IDI for set A was 0.35 (IQR, 0.23; range, 0.13–0.55) and set B 0.43 (IQR, 0.25; range, 0.17–0.60) with four questions in set A below 0.2 and two in set B. Four questions in each set had answer scores >90%, and one question in set A had a score of 100%. No question had a score below 10%.

3.3.2. Content validity

The response rate to the online survey on content validity was 74% (29 third-year GP trainees from the
Mean scores on difficulty were 3.5 (standard deviation [SD], 0.2) and 3.8 (SD, 0.2) for set A and B, respectively. A total of 54% participants considered set A neither easy nor difficult (score 3) and 39% considered it difficult (score 4). For set B, 38% considered it neither easy nor difficult (score 3) and 50% difficult (score 4). A total of 77% and 63% considered respectively set A and set B relevant to clinical practice (score 3), with mean scores of 3.9 (SD, 0.2) for set A and 2.8 (SD, 0.2) for set B. Trainees expected to score on average a 6.2 (SD, 0.3) for set A and 4.7 (SD, 2.6) for set B. These expected scores both correlated positively very strong with the participants scored the highest or lowest possible score. Therefore, the questionnaire is valid to assess improvement or deterioration of knowledge.

### 3.3.3. Construct validity

Both sets showed satisfactory construct validity with both before and after EBM training trainees scoring significantly better than supervisors. Comparisons between groups before EBM training showed a significant difference in mean score between trainees (N = 200) and supervisors (N = 17). The 200 trainees had a mean score (95% confidence interval [95% CI]) of 22.1 (21.4–22.9) on set A and 21.1 (20.3–21.9) on set B, compared with the mean scores of the 17 supervisors of 16.5 (13.2–19.8) and 15.1 (12.3–18.0), respectively. When experts’ mean scores on set A and B were compared with mean scores on both sets of participants after EBM training, experts (mean score, 30.3 [95% CI, 28.9–31.7] and 30.0 [95% CI, 28.3–31.8]) scored significantly better than trainees (mean score, 25.7 [95% CI, 24.4–27.1] and 25.4 [95% CI, 24.0–26.7]) and supervisors (mean score, 19.3 [95% CI, 16.1–22.1] and 18.7 [95% CI, 15.9–21.6]) (Table 1).

#### 3.3.4. Test–retest reliability

Results showed a modest performance on test–retest reliability with significantly different mean scores (±SD) between the two measurements for both set A (18.3 ± 3.0 vs. 24.7 ± 4.5, P = 0.001, Pearson correlation –0.70, P = 0.001) and set B (17.7 ± 3.5 vs. 22.9 ± 4.9, P = 0.004, Pearson correlation –0.24, P = 0.33).

#### 3.3.5. Responsiveness

Both sets proved to be responsive with significant higher mean scores after EBM training than before EBM training for all groups (GP trainees, hospital trainees, and supervisors; see Table 3).

#### 3.3.6. Feasibility

The mean time (±SD) needed to complete set A was 24 (SD, 11) minutes and 33 minutes for set B (SD, 11). For both sets, checking and scoring took between 1 and 3 minutes per participant as only the open-ended questions (i.e., calculations and questions in which participants had to identify the research question) had to be checked.

### 3.3.7. External validation

Both sets proved to be externally valid with no significant differences in mean scores (±SD) when 54 GP trainees from the AMC were compared with 47 GP trainees from the UMCU (set A: 23.3 ± 6.3 vs. 24.9 ± 4.6 [P = 0.07], set B: 23.0 ± 4.2 vs. 24.1 ± 4.7 [P = 0.22], respectively).

### 4. Discussion

The U-CEP measures knowledge on clinical epidemiology, focusing on aspects relevant to daily clinical practice. This is important because it has become clear that practitioners face challenges in incorporating biomedical
knowledge from research into practice [17–19]. Other questionnaires measuring this knowledge focus on calculating and interpreting biostatistics (Berlin questionnaire), are more time consuming to fill in and score, while they consisted of open-ended questions only (Fresno), or focus on the critical appraisal of therapeutic research only (‘Assessing Competency in EBM’ tool, ACE tool) [5–7]. However, clinical epidemiology involves four clinically important challenges clinicians are faced with in daily clinical practice (DEPTh) [1,5–7]. The importance of knowledge about the differences between these four is emphasized by the presence of different guidelines for complete and accurate reporting for the different types of research [20,21]. Therefore, the U-CEP measures knowledge on all clinical epidemiology aspects. Moreover, the U-CEP is responsive to change. As a result, it is valid to monitor changes in knowledge of clinicians on clinical epidemiology after EBM training. For two other questionnaires, the Fresno and ACE tool, this has not been assessed [5,7].

Several aspects of the performance of the U-CEP must be considered critically. First, the third-year GP trainees considered set B of the U-CEP to be of only moderate clinical relevance. As no data on clinical relevance are available for earlier EBM questionnaires, interpretation of these results is difficult [5–7]. A possible explanation for this somewhat disappointing score is a different interpretation of the relevance for clinical practice between GP trainees and experienced clinicians and teachers. Possibly, the third-year GP trainees considered the topics dealt with (focusing on diagnosis, prognosis, and therapy) in the questionnaire clinically relevant, but the questions themselves (including clinical epidemiology methodology and jargon) were not considered to be clinically relevant. We, however, did not test whether this indeed was the case. Second, the U-CEP performs modest on the test–retest reliability with scores at the second measurement comparable to those of GP trainees who received training in EBM. This can be explained by the fact that (for external reasons) the first time distribution was on paper and the second time online, whereas it is recommended that the administration format should remain the same across assessment points [22]. Also, the time interval of 3 months may have been too long. In these 3 months, participating first-year GP trainees may have gained knowledge on clinical epidemiology by self-study or in clinical practice. All GP trainees are supposed to have a textbook on EBM at the start of their GP training program, and some may have studied this book in preparation of the EBM training, especially in case of disappointing scores at the baseline measurement (explaining the negative correlation in set A). Third, the internal reliability (ITC) and ability to discriminate between participants (IDI) is quite low, albeit this was the case for a small number of questions only. Removal of these questions did not change the internal consistency of the questionnaire, however, and we consider content validity to be better with those questions included in the questionnaire. Furthermore, although external validation among GP trainees from another university medical center in the Netherlands proved the U-CEP to be externally valid, further validation of the U-CEP across different undergraduate and postgraduate training programs among other health care professionals and trainees. Validation after translation into other languages will provide valuable information on the validity of the questionnaire when used in other countries as results may be different [10]. Finally, although it is custom to refer to questionnaires like the U-CEP as questionnaires on EBM knowledge [5–7], these instruments only assess knowledge and do not measure whether this knowledge is indeed used in daily clinical practice (and thus leads to evidence-based practice by integrating the three aspects of EBM). Other instruments are needed to assess the latter.

The U-CEP was shown to be valid among different clinicians, such as GP trainees, hospital trainees, and GP supervisors. The online availability allows easy distribution with questions in random order. We have shown that the questionnaire can be used in a valid and reliable way as an evaluation tool for EBM training. Different sets of 25 questions may be used before and after EBM training or one larger set of 50 questions can be used when one wants to use the questionnaire for a formative assessment to help clinicians learn.

### Table 3. Responsiveness: comparisons between scores on the U-CEP before and after EBM training of participants (N = 65) who filled in the questionnaire before and after EBM training

<table>
<thead>
<tr>
<th>Variables</th>
<th>Before Score (SD)</th>
<th>After Score (SD)</th>
<th>Difference Score (SD)</th>
<th>95% CI</th>
<th>Before Score (SD)</th>
<th>After Score (SD)</th>
<th>Difference Score (SD)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP trainee (N = 29)</td>
<td>14.4 (6.9)</td>
<td>24.9 (4.7)</td>
<td>10.5 (7.1)</td>
<td>7.8–13.2</td>
<td>13.9 (7.3)</td>
<td>24.8 (4.6)</td>
<td>10.9 (7.8)</td>
<td>7.8–13.8</td>
</tr>
<tr>
<td>Hospital trainee (N = 20)</td>
<td>24.4 (3.6)</td>
<td>26.9 (4.6)</td>
<td>2.5 (2.6)</td>
<td>1.2–3.7</td>
<td>23.6 (3.9)</td>
<td>26.3 (4.9)</td>
<td>2.7 (2.4)</td>
<td>1.5–3.8</td>
</tr>
<tr>
<td>Supervisor (N = 16)</td>
<td>16.8 (6.6)</td>
<td>20.4 (5.3)</td>
<td>3.7 (5.6)</td>
<td>0.7–6.7</td>
<td>15.3 (5.6)</td>
<td>20.1 (4.5)</td>
<td>4.8 (4.9)</td>
<td>2.2–7.5</td>
</tr>
</tbody>
</table>

**Abbreviations:** U-CEP, Utrecht questionnaire on knowledge on Clinical epidemiology for Evidence-based Practice; EBM, evidence-based medicine; 95% CI, 95% confidence interval; SD, standard deviation; GP, general practice.
Acknowledgments

The authors thank all the participants who filled in the questionnaire(s) as well as Pieter Jan van der Schoot and Erwin Veltman for their technical support with the online version of the questionnaire.

Authors’ contributions: M.F.K. was the primary investigator of this study and designed the study, acquired the data, analyzed and interpreted the data, drafted the article, and approved the final version to be published. M.E.L.B. was the supervisor of the study and designed the study, assisted in the interpretation of the data, revised the article critically for important intellectual content, and approved of the final version to be published. M.E.L.B. was the supervisor of the study and designed the study, as well as Pieter Jan van der Schoot and Erwin Veltman for their technical support with the online version of the questionnaire.

D.E.G. revised the article critically for important intellectual content, and approved of the final version to be published. H.K. analyzed the data, and interpreted the data, revised the article critically for important intellectual content, and approved of the final version to be published. H.K. analyzed the data, and approved of the final version to be published. N.J.d.W. was the supervisor of the study and designed the study, revised the article critically for important intellectual content, and approved of the final version to be published. N.J.d.W. was the supervisor of the study and designed the study, revised the article critically for important intellectual content, and approved of the final version to be published. D.E.G. revised the article critically for important intellectual content, and approved of the final version to be published. D.E.G. revised the article critically for important intellectual content, and approved of the final version to be published. A.W.H. was the supervisor of the study and designed the study, revised the article critically for important intellectual content, and approved of the final version to be published.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.08.009.

References