

## EDUCATION AND DEBATE

# Is assessment of likelihood ratio of homeopathic symptoms possible? A pilot study

ALB Rutten<sup>1\*</sup>, CF Stolper<sup>1</sup>, RFG Lugten<sup>1</sup>, and RWJM Barthels<sup>1</sup>

<sup>1</sup> *Commissie Methode en Validering VHAN (Dutch Association of Homeopathic Physicians), Aard 10, 4813 NN Breda, The Netherlands*

**A pilot study was performed to investigate the possibilities and restrictions of likelihood ratio (LR) investigation using three symptoms. Qualitative vagueness and expectation bias is inherent in our method, but is, in part avoidable. It appears that experienced observers assess common homeopathic symptoms quite similarly. Clinical judgement is an essential part of our work and should be preserved during assessment of LR. The assessment does not influence clinical practice and can be maintained for a long period, provided the appropriate software is used. A limited range of symptoms seems most suitable for LR investigation.** *Homeopathy* (2003) 92, 213–216.

**Keywords:** likelihood ratio; vagueness; repertory

## Introduction

In the accompanying paper<sup>1</sup> we investigated the theoretical problems that could arise in assessing vague clinical and homeopathic symptoms. The likelihood ratio (LR) method allows some latitude in the quantitative interpretation of our symptoms, but we must be aware of expectation bias and qualitative misinterpretation of symptoms. In the editorial comment to our earlier article<sup>2</sup> Peter Fisher stated that LR investigation requires data collection ‘.. on a daunting scale, particularly for rarely used medicines and infrequent symptoms’.<sup>3</sup> In order to get more clarity about these problems and to prepare a protocol for the research of LR we performed a pilot study.

## Methods

We performed a prospective investigation of the prevalence of some symptoms in four homeopathic practices as a pilot study to trace methodological

problems researching LR of homeopathic symptoms. We assessed three kinds of symptoms with different kinds or degrees of vagueness: ‘desire for coffee’, ‘fear of snakes’ and ‘loquacity’. We expected ‘loquacity’ to be the vaguest symptom. This symptom had already been assessed in our materia medica validation.<sup>4</sup> For this symptom we recruited one extra observer.

We deliberately did not restrict participants in scoring the symptoms, in order to evaluate the differences in interpretation of the symptom based on the experience of each observer. One of the things we wanted to know was the inter-rater variability in the prevalence of the symptoms as observed by relatively unprepared (but experienced) observers; do they have the same notion of these symptoms? In [Table 1](#) this is expressed by the standard error in the mean prevalence of the symptom. The participants were not aware of each others’ data. We also tested three different ways to collect data. A spreadsheet was provided to two participants, one participant used his own spreadsheet, one used his own practice administration database program and one used a paper form.

The pilot study was carried out from July till December 2002. All new patients more than 2 years old were included. The experience of the investigators was noted in years.

\*Correspondence: L Rutten, Aard 10, 4813 NN Breda, The Netherlands.

E-mail: lexrtn@concepts.nl

Received 23 April 2003; revised 18 July 2003; accepted 11 August 2003

**Table 1** Results of pilot study

	LR	ES	PF	RB	SJ	Totals	Mean prevalence/ standard error
Experience	23	17	14	14	21		
Number of patients	114	104	65	77	149	509	
<i>Desire coffee</i>	9	11	9	4		33	
Prevalence <i>desire coffee</i>	0.079	0.106	0.138	0.052			0.065
Standard error <i>desire coffee</i>							0.02494
<i>Loquacity</i>	11	13	8	8	9	49	
Prevalence <i>loquacity</i>	0.096	0.125	0.123	0.104	0.060		0.096
Standard error <i>loquacity</i>							0.01202
<i>Fear snakes</i>	6	4	1	2		13	
Prevalence <i>fear snakes</i>	0.053	0.038	0.015	0.026			0.026
Standard error <i>fear snakes</i>							0.00915

Observers are: LR, ES, PF, RB and SJ. 'Number of patients' expresses the number of patient included in the investigation. Prevalence *desire coffee*=0.065 means: prevalence *desire coffee*=6.5%.

## Results

There is no clear relation between the description of remedies and the investigated symptoms, eg about 20 different remedies were prescribed for 49 loquacious patients. *Lachesis* was prescribed 10 times, three of these patients were loquacious. Then, of course, there is insufficient follow-up to assess results of treatment. After data collection the personal assessment criteria for each symptom were exchanged. Some participants asked every patient if he/she was talkative (LR and ES) and other participants merely noted their own impression during consultation. One observer (SJ) noted 'loquacity' if he was somewhat annoyed, or hampered in his usual history-taking, by the loquacity of the patient. Coffee intake was asked by every practitioner and then related to cultural and personal circumstances by clinical judgement by two raters (LR and RB). This symptom revealed some unexpected hits like a 4-year-old child stealing the leftovers of the coffee cups of the adults. One rater (PF) took more than five cups a day as desire for coffee and one rater (ES) noted desire for coffee as positive if coffee was among the five most wanted food items. Three observers enquired directly for fear of snakes. If the fear of snakes was confirmed, the additional question was if this fear was also present if the snake was behind glass or in a picture (television). One (PF) asked for fear of animals.

Time to register the presence of the symptoms with each new patient varied from 2 s when a database program was used to 30 s if it was noted on paper. It took 10 min to 1 h to prepare the data for transmission to the co-ordinator. It takes some hours to evaluate results of treatment, but these data have not yet been assessed, because follow-up was too short.

### Test validity

To assess the validity of our procedure we can use the ten questions for the assessment of diagnostic testing

proposed by Greenhalgh.<sup>5</sup>

*Is this test potentially relevant to my practice?*

Yes, the investigated symptoms are frequently used in homeopathic practice.

*Has the test been compared with a true gold standard?*

No, there is no true gold standard, we only have the results of the treatment that resulted (partially) from the symptom.

*Did this validation study include an appropriate spectrum of subjects?*

Yes, all new patients above 2 years were included, but there is still discussion about the assessment of the results of the treatment.

*Has workup been avoided, ie did everyone undergo the test?*

Yes, but we still are considering if we need 'hard evidence' if that the symptom was checked.

*Has expectation bias been avoided?*

No, this is a serious problem. Expectation bias should be well documented.

*Was the test shown to be reproducible?*

This will require replication.

*What are the features of the test as derived from this validation study?*

So far we have chosen symptoms that are recorded as keynotes for certain remedies.

*Were confidence intervals given?*

Yes, but we should consider other statistical techniques for vague symptoms.

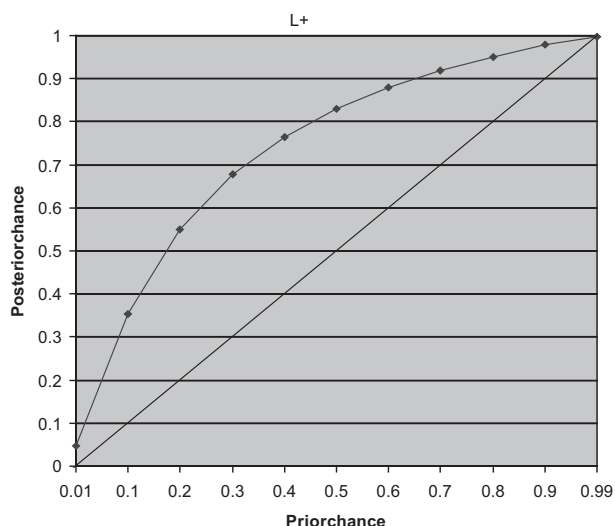
*Has a sensible 'normal range' been derived?*

Maybe, this is discussed elsewhere.<sup>1</sup> This should be well documented.

*Has this test been placed in the context of other potential tests in the diagnostic sequence?*

Not yet, after the assessment of many symptoms we can evaluate the meaning of the context.

Some questions are related to each other (2,5,8,9). Vagueness and the absence of a gold standard hamper



**Figure 1** Loquacity and *Lachesis*; LR+=4.9.

the validity of the test, but if we try to avoid these biases our test becomes irrelevant.

#### **Lachesis and loquacity**

This study revealed the prevalence of ‘loquacity’ in the population seeking homeopathic help. There were 10 *Lachesis* prescriptions in a population of 509 patients, so we think that the prevalence in the whole population is a tolerable approximation of the prevalence in the non-*Lachesis* population. From our materia medica validation meetings we also have retrospective data on loquacity and *Lachesis*, 16 cases were assessed with a score of +3 or +4 on the GHOS scale. Of these cases, seven were marked as loquacious. We expect no under-reporting because the symptom loquacity is well known for *Lachesis* and all participants had the opportunity to complete their data while discussing the cases. In our validation procedure 3–6 colleagues assess the quality of the case. In both the retrospective and the prospective data ‘loquacity’ was ill defined. Based on a combination of the prospective and the retrospective data we estimate the LR+ of loquacity for *Lachesis* to be 4.9. The 95% confidence interval (Sime<sup>6</sup>) is 3.3–6.5.

Based on this LR+ we draw the graph in Figure 1. On this graph with LR+ = 4.9 we can see that a prior chance of 1% goes to 4.7% if loquacity is present. A prior chance of 4.7% goes to 19%; 19% goes to 54%; 54% goes to 85%. These figures suggest that we need four symptoms of this strength to make a nearly certain prescription, assuming that the prior chance at the beginning of the consultation is 1%. This is of course still very hypothetical and depends on several conditions such as the symptoms being mutually independent. The estimation of LR of ‘loquacity’ for *Lachesis* is better than our current bold typeface in the repertory, but still halfway towards our final goal, data based on prospective research.

## **Discussion**

Three main points emerge from this pilot study: (1) handling of vagueness. (2) the feasibility of LR research and (3) the validity of our calculation of LR of loquacity for *Lachesis*.

1. It appears that inter-rater variability (expressed in standard error) is low for all three symptoms. The standard error for ‘loquacity’ is lower than that for ‘desire for coffee’. The higher standard error for ‘desire for coffee’ might be caused by the experimental situation, for it is tempting to use a threshold value (such as number of cups per day) for such a symptom. One might discuss the validity of this threshold in practice, for the real meaning of ‘desire for coffee’ in homeopathic practice is that there is a constitutional need for coffee. Age, profession, culture and so will also influence the intake. The low standard error in ‘loquacity’ seems to confirm that experience provides implicit consensus about the meaning of such a symptom. In this pilot study we see no signs of expectation bias between the choice of the medicine and assessment of symptoms yet, but numbers may be too low. We should again be suspicious of expectation bias when outcome results are available. For this reason the assessment of symptoms should only take place during the first consultation.

2. Of the three tested systems of data collection (practice administration software, separate spreadsheet and paper) practice administration software is highly preferable. Then collection of data and administration of results is integrated in administrative procedures (such as billing) performed for each consultation. This way the LR research causes no extra time consumption and can be maintained for many years. Necessary data can easily be made available at any time. Ten doctors can enter 2000 patients each year. A hundred groups could evaluate about 500 symptoms in about five years if each group assessed 5 symptoms. This indicates that we can replace many questionable rubrics (caused by unreliable entries of ‘small’ and ‘large’ remedies) by trustworthy rubrics. If we assess the symptoms that are valued as indicative for certain remedies we will be more certain about the expected success of those remedies.

3. To estimate the LR+ of loquacity for *Lachesis* we combined prospective and retrospective data. This is not ideal. We assumed that the symptom loquacity is sufficiently well known and that underreporting is unlikely. But on the other hand there is still the possibility of expectation bias. The retrospective data were obtained during consensus meetings where results of treatment were assessed by peer review. This is the most elaborate procedure yet attempted for symptom validation, but still questionable as ‘gold standard’. We should aim at prospective research as the gold standard.

In researching our instruments we will detect several weaknesses. One of them is insufficient inquiry about

symptoms during consultation and bias in interpretation of symptoms. Many symptoms in the materia medica are not clearly defined. Our pilot study indicated that experienced practitioners have a kind of intuition about symptoms that overrides descriptions like the amount of coffee or words spoken. Maybe bias is unavoidable, but we are insufficiently aware of it at the moment. We must make it clear that we are researching 'Loquacity, changing the subject frequently' as observed during consultation and not 'Loquacity' in a broader sense. If we research loquacity in the broader sense we must reach consensus about what we understand by loquacity and about the way we investigate it, eg about the questions we use. The use of LR allows quantitative vagueness only if we can avoid expectation bias. This kind of bias can be detected by analysing the data on variation from randomness related to different observers.

We must develop means to detect misclassification and to handle it, including:

- Adequate description of measurements. Symptoms must be assessed during the first consultation and may not be altered without compelling reason.
- Description of possible bias.
- Aggregation of multiple raters and description of inter-rater variance.

If the process is well described, the user of the data can estimate its applicability. If, say, the prescriber knows the assessed interpretation and prevalence and inter-rater variance of the symptom loquacity and the prevalence in his own practice it is possible to estimate the validity of the LR in that particular practice.

We think that the research for LR is most suited for symptoms with a prevalence between 2 and 15% and a reputation as keynotes. There are only a few hundred symptoms that meet these conditions.

## Conclusion

This pilot study indicates that the prevalence of symptoms in homeopathic practice can be researched, even for vague symptoms. The next step is to investigate our 'gold standard', constituting the 'a and c groups' of the  $2 \times 2$  diagram (see accompanying paper<sup>1</sup>). We have no hard 'gold standard' like the inflamed appendix, excised from the abdomen of the patient. But we must be practical, the aim is to improve our method by learning from our best cases. If we regard our best cases as 'gold standard' we can investigate the symptoms that led us to these. We are now developing criteria to assess the best cases.

Adequate description of the assessment is imperative, including inter-rater variance. Using the combination of retrospective research for the prevalence of the symptom in the remedy group and prospective research for the prevalence in the rest of the population might give an indication of LR. The validity of this procedure seems questionable (but still better than the present representation in the repertory).

Furthermore our pilot study indicates that collecting data, necessary to investigate LR, is not time-consuming so it can be maintained for years, especially with help of proper software.

The question is not only if the test is valid, but also if the investigator did what the average clinician does (or should do). After the test our concern is if the test (symptom) can be applied by every clinician, in this way limiting variance between clinicians.

We think that a few hundred rubrics of the repertory will benefit greatly from assessment of their LR. This will not only improve the efficacy of the repertory but it will enable us to investigate our methodology further. Questions like 'What does LR mean in our practice?' and 'How many symptoms do we need to be sure of our prescription?' can be formalised once the strength of symptoms is better described. In a subsequent paper we will describe the possible changes to the repertory if LR were included and some methodological consequences.

## Acknowledgements

We thank the other participants in the pilot-study, Paul Fruijtjer and Stan Jesmiatka, for gathering the data and active participation. We thank Dick Bezemer for his comments.

## References

- 1 Rutten ALB, Stolper CF, Lugten RFG, Barthels RJWM. Assessing likelihood ratio of clinical symptoms: handling vagueness. *Homeopathy* 2002; **92**: 182–186.
- 2 Stolper CF, Rutten ALB, Lugten RFG, Barthels RJWM. Improving homeopathic prescribing by applying epidemiological techniques: the role of likelihood ratio. *Homeopathy* 2002; **91**: 230–238.
- 3 Fisher P. Editorial. Send in your cases, or the lost art of the concise case report. *Homeopathy* 2002; **91**: 195–196.
- 4 Stolper CF, Lugten RFG, Rutten ALB. Materia Medica Validering: Lachesis en Tarentula. *Similia Similibus Curentur* 2000; **30**: 18–19.
- 5 Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. *BMJ* 1997; **315**: 540–543.
- 6 Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991; **44**: 763–770.