Homeopathy (2014) http://dx.doi.org/10.1016/j.homp.2014.11.002, available online at http://www.sciencedirect.com
This is the accepted version of the manuscript.

Data collection: treat every variable as a treasure

ALB Rutten

Abstract

Background: Collection of case history data is not common in homeopathy despite its great importance for this method. Information technology development progresses slowly and discussion about requirements is scarce.

Method: Two Dutch projects assessed Materia Medica of some homeopathic medicines and six homeopathic symptoms. The latter project especially relied heavily on data collection. In both projects much effort was spent on achieving consensus between participating doctors. There was much variance between doctors despite efforts to achieve consensus. Assessing causality was the most important source of bias, there is also much variance in assessing symptoms.

Conclusion: Data collection software should be developed step-by-step, guided by close monitoring and feedback of participating practitioners. *Homeopathy*

Key-words: data collection; causality; confirmation bias

Introduction

Collecting large volumes of data seems easy with modern Information technology (IT). Generally, more data is considered better data because with more data of the same variable we are more sure about the mean value, provided there is no bias. But we must be aware of the 'garbage-in-garbage-out' principle. It is tempting to collect data first and think about possible research later. But then there is considerable risk that the data do not suit the research questions. But if we keep thinking about data collection it could be a long time before we actually start collecting them. There are research projects collecting data for specific research questions, but at the moment, few practitioners use software that records treatment data, despite extensive use of software programs for repertorisation. The importance of data collection for homeopathy is obvious, so we should try to speed up the process of software development for homeopathic data collection. The purpose of this paper is to open up the discussion, based on experience of systematic data collection in Dutch homeopathic practices with simple programs recording only the most necessary data.

There are many theoretical considerations about data collection, but do they hold in daily practice? In the Netherlands a group of experienced homeopathic doctors started discussing successful cases retrospectively concerning specific homeopathic medicines in 1997.[1] The purpose was to validate existing Materia Medica by qualitative analysis of successful cases; hence is was called the Materia Medica Validation (MMV) project. A Dutch commercial database program (HARP) and three database programs developed by doctors for their own practice facilitated retrieval of successful cases. All programs were developed by homeopathic doctors following their own needs in collecting data. The programs were adjusted to new insights in the course of these projects.

This experience resulted in some hypotheses about the clinical decision process in selecting homeopathic medicines. The main hypotheses were that :

1) The choice of a homeopathic medicine is based on pattern recognition.
2) This pattern recognition is preceded by collecting symptoms and personal characteristics of the patient indicating a limited set of homeopathic medicines.

This 'differential diagnosis' is about prognosis. It comes after the conventional differential diagnosis about illness. As in most differential diagnoses about illness, the' differential prognosis' about successful homeopathic medicines is based on more than one symptom/characteristic.

The prognosis process can be described as sequential updating of probability that a medicine will work in the patient before you, based on experience in the past. Participants of these expert meetings concluded that a symptom indicating a specific homeopathic medicine occurred frequently in cases 'cured' by that medicine, more frequently than in other medicines. This is similar to diagnostic reasoning and can be described as a Bayesian process.[2] New information (symptom) changes the probability that a medicine will work: posterior odds = LR * prior odds. Likelihood ratio (LR) is the prevalence of the symptom in the medicine population divided by the prevalence in the remainder of the population.[3] The existing database programs were adjusted to these ideas and used for a subsequent prospective study (LR project) assessing the relationship between homeopathic symptoms and successful prescriptions.[4]

Figure 1: Constantin Hering's office

An important goal of data collection is to retrieve the information you need at any one time, nothing less, nothing more. In other words, the search has to be precise to avoid ignorance and too much work. We want to retrieve all relevant cases and provings if we study a particular medicine. Would a computer have helped Constantin Hering (Figure 1) in finding just that piece of information he needed when writing his Materia Medica of, say, *Lachesis*? We cannot be sure, because he probably had a detailed roadmap of his office in his brain, indicating the position of every piece of information ordered by a system only familiar to him. Nowadays he would have to create an electronic roadmap

on an information carrier the size of a finger nail that contained his whole library. Finding data on such an information carrier requires another, more explicit and less intuitive, way of organising data. In many cases we need the information to increase our knowledge about a specific medicine. After finding all related cases we may develop some qualitative ideas about the medicine.

Another goal of data collection is quantitative analysis: finding incidence or prevalence of variables like symptoms and results and relationships between variables. Some symptoms are more related to specific medicines than to other medicines. In this case precision is essential to avoid false conclusions from the data. In this respect we must distinguish structured and unstructured data. Suppose we want to find all cases of loquacious patients in Hering's digitalised cases (unstructured data). Can we start calculating the prevalence of 'loquacity' in Hering's practice after simple text search in a word-processing program? Only partly, because in a number of cases the patient may be described as 'not-loquacious', noticed only after reading the corresponding sentence. We may also miss a number of cases where the patient is called 'talkative', 'chattering', 'verbose', or where 'logorrhoea' is perceived. And we are also not sure if Hering noticed every loquacious patient.

Although the computer can help us with a thesaurus, we must be careful in handling such unstructured data. The opposite case is collecting strictly structured data in prospective research. Here we ask the practitioner to record if the patient is loquacious in every consecutive new case. Thus, we avoid the problem of synonyms and we are sure the symptom is observed. Structured and unstructured data collection both have their advantages and disadvantages, but it makes sense to know possible bias in strictly structured data first, because probably such bias is also present in unstructured data. This kind of bias can also be present in partly structured data, as proposed by the CARE guidelines.[5]

This paper presents some reflections on data collection based on the two projects conducted in The Netherlands assessing homeopathic prescribing; one qualitative, the other quantitative. Both projects were facilitated by electronic data collection, but in different ways. In the first project electronic data collection played a very modest role, just retrieving names to find written documents, the second project depended heavily on electronic data for calculating prevalence of symptoms. Both projects involved intense discussions about what we were actually doing and about differences between doctors. Especially such differences appeared relevant in data collection. This experience may help to build effective databases, that are easy to use in daily practice.

Methods

The first project, Materia Medica Validation (MMV), comprised consensus meetings from 1997 to 2007. Twice a year experienced (> 5 years) Dutch homeopathic physicians were invited to bring in their best cases concerning two specified homeopathic medicines to reflect on how to improve homeopathic prescribing. The meetings were structured as open discussions, but with a pre-structured format for each case, describing reasons for prescribing the homeopathic medicine and the effects ascribed to the medicine. The cases should have a follow-up of at least one year considering the assessed medicine and the relation between effect and the medicine should be clarified. There were no other inclusion or exclusion criteria to allow an open discussion about different methods in homeopathy. All participants, however, were trained in classical homeopathy according to ideas set out by Hahnemann, Kent and Hering. These meetings were attended by 10-25 doctors, presenting in total between five (*Naja*) and 23 (*Sulphur*) best cases. Each case was discussed regarding causality (was the improvement really due to the medicine?), about the type of person and the symptoms present in the case. The Glasgow Homeopathic Hospital Outcome Scale (GHHOS) was used as an instrument to assist the discussion about causality. The participants had the patient's file with them, so there was the possibility to confirm symptoms that were not mentioned at first in

retrospect if the symptom occurred in other cases. An attempt was made to estimate the prevalence of the most important symptoms.

Participants used electronic data registration, if available, only to find the cases concerning the medicines to be discussed. Best cases had still to be hand-picked, only partly guided by the GHHOS score. Most cases were in hand-written files and were transcribed in structured forms before the meeting including sections like trigger symptoms (when the practitioner started to think about the medicine), improved symptoms, etcetera.

The second project, the Likelihood Ratio project (LR project), was a prospective observational study in 10 Dutch practices from 2004 to 2007, assessing six homeopathic symptoms ('Diarrhoea from anticipation', 'Fear of death', 'Grinding teeth during sleep', 'Recurrent herpes of the lips', 'Loquacity' and 'Sensitivity to injustice'). The purpose of this project was to assess the relationship between these symptoms and positive results of respective medicines, avoiding influencing daily practice and minimising time necessary for recording data. There was no interference with usual practice, ethical approval was not required, but patients were informed that their data would be used anonymously for practice evaluation. Participating doctors were recruited from the doctors attending MMV meetings. All symptoms were checked in each consecutive new patient. All patients older than two years with checked symptoms were included. Patients with acute pathology and with no grounds for homeopathic therapy were excluded. The symptoms were specified in advance during a consensus meeting, e.g. 'Recurrent herpes of the lips' was defined by 'more than six times a year', 'Sensitivity to injustice' as 'resulting in subsequent behaviour, like writing letters to papers and politicians, participating in protest groups, etc.'. Scoring of results and establishing causal relationship was also discussed to obtain consensus. Only chronic cases were recorded. Results regarding prescribed medicines were frequently monitored and feedback was organised in newsletters and consensus meetings. In the end 4094 patients were included and 4074 prescriptions were evaluated. Results of treatment (per medicine) were recorded after at least three months using a modified GHHOS. If the last GHHOS score within a year for a specific medicine was >=2, with probable causal relationship between medicine and result, the patient was attributed to that medicine population. Statistical analysis was performed regarding relationships between symptoms and results and inter-doctor variance in results and prevalence of symptoms. Participants assessed their own results, there was no second opinion.

We used several software programs, three were developed by individual practitioners for their own practice using Access and Filemaker-pro. Most practitioners used a commercial administration program (HARP), originally recording prescribed medicines and results (GHHOS score). This program and the individual programs added six homeopathic symptoms for the LR-project. Results have discussed elsewhere, for this paper inter-rater variance is the most important outcome. This was analysed using Excel and SPSS software.

Results

Materia Medica Validation

After 11 years 24 medicines were validated by largely the same group of doctors. Discussing results with colleagues appeared to be a challenge. Opinions about causal relationship between medicine and improvement vary, even concerning best cases. Reconsidering cases sometimes casted doubt on cases with the highest GHHOS score. Positive life-events, like finding a new love or a better job, were often reasons to doubt the effect of the medicine. The necessity to repeat the medicine was one of most valued reasons to confirm causality. The GHHOS score describes the improvement, not

causality, but GHHOS>=2 is considered an indication of a homeopathic effect, because of improvement of more than just the presented complaint.

One of the outcomes that surprised the participants was the fact that some expected symptoms were not present in all cases. Only four out of ten *Causticum* patients were sensitive to injustice, five out of 12 *Stramonium* patients were afraid of the dark. Apparently, the absence of these symptoms is no absolute contra-indication for the respective medicines.

Results are described in detail elsewhere [3,4,7,8], in this paper results that are relevant for data collection are selected.

LR assessment

Recording results. All consecutive new patients entered the project, starting in June 2004. After six months the recording of results was monitored for the first time and differences between participants were discussed in a consensus meeting. Opinions appeared to differ about what results should be entered and when. Some participants thought that only good results were relevant, some recorded the first result three months after starting treatment, others after one month. Starting after one month resulted in some cases of worse condition (GHHOS -1 or -2), possibly due to initial aggravation of symptoms. If these patients did not come back after the first follow-up consultation the results remained in the database as end-results. This first consensus meeting concluded that all results should be recorded, starting after three months.

Despite several consensus meetings discussing interpretation of results and **the fact** that the participants were trained in assessing results in the MMV project variance was still considerable in the end-results, see the box-plot in Figure 2. Based on our experience with MMV we suspect that assessing causality causes this variance.

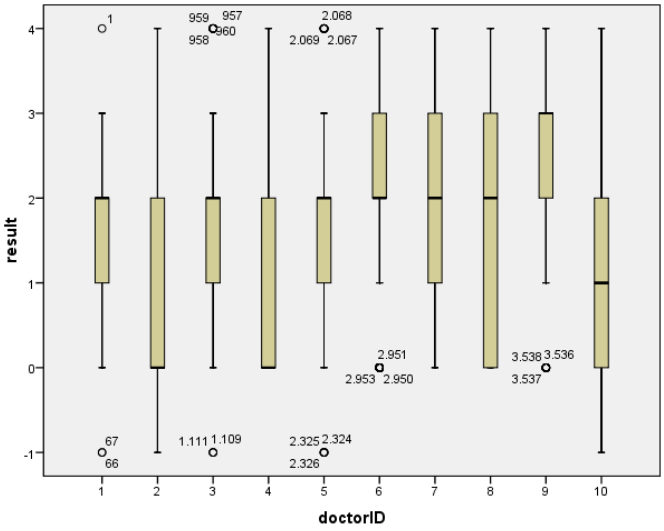

Figure 2: boxplot of results, ordered by doctor. Despite repeated consensus meetings, still much variance is present.

Recording symptoms. The recording of symptoms was also frequently monitored and discussed in consensus meetings during this project. Here also considerable variance remained until the end of the projects in the symptoms 'Loquacity' and 'Sensitivity to injustice'. This is no surprise as these are the most vague symptoms, but we must be aware of the fact that this variance is still present in well-trained observers despite intensive consensus efforts to define symptoms. Figure3 shows the

prevalence of the symptoms 'Loquacity', 'Sensitivity to injustice' and 'Recurrent herpes of the lips' in 10 practices. In our Materia Medica and repertories most symptoms are not defined at all. On the other hand, defining symptoms caused other problems, e.g. age is influencing symptoms like 'Sensitivity to injustice'. The observers relied on clinical judgement in such cases. As can be expected inter-rater variance is more as the symptom is more subjective; 'Sensitivity to injustice' is the most subjective symptom.

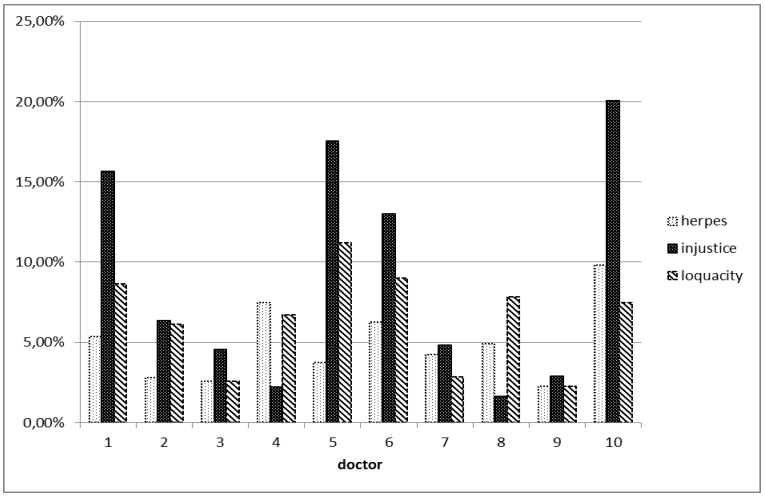

Figure 3: prevalence of three symptoms in 10 practices, see text

Confirmation bias. Observations can be influenced by prior knowledge. After six months all *Causticum* patients were sensitive to injustice, after two and a half years only 40%. Reconstructing this fact it appeared that in some patients the effect of *Causticum* subsided, and that in other patients *Causticum* was selected on other grounds. This result was consistent with the outcome of Materia Medica Validation of *Causticum*. We found this confirmation bias, to a lesser extent, in three other combinations: 'Grinding teeth during sleep' and *Mercurius*, 'Fear of death' and *Arsenicum* and 'Loquacity' and *Lachesis*.

In our project we found only four indications of confirmation bias in about 50 statistically significant relationships between symptoms and results. McKenzie also states that confirmation bias is mostly present in familiar situations.[6] Our data show that this bias could subside by longer follow-up.

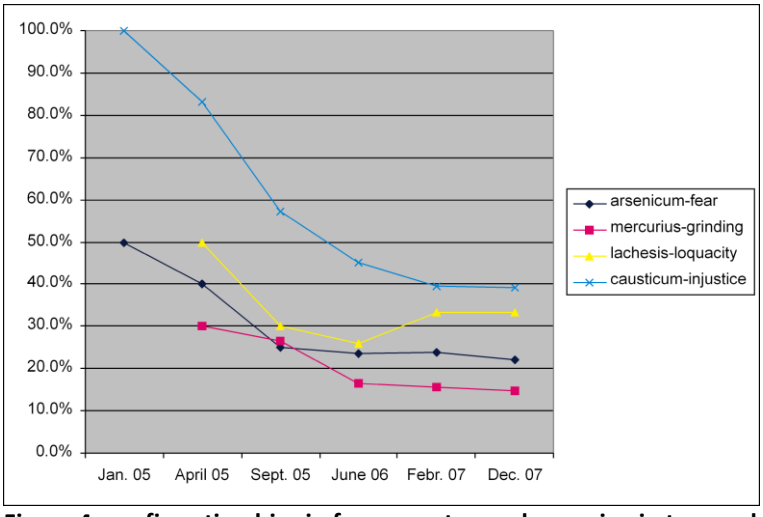

Figure 4: confirmation bias in four symptoms, decreasing in two and a half year

Discussion

Homeopathy is mostly data driven, but electronic data collection is still rare. Repertorisation software is much more common than data collection software. Understandably, software developers try to make a complete program right away, but this may cause delay. It is, however, not so difficult to collect a few structured variables, like the prescribed medicine and result of treatment. Collecting and analysing unstructured data, like verbatim recording of consultations, is quite different. It is not feasible for many practitioners and analysing such data may have problems we cannot foresee yet. Maybe we aimed too high and should progress in steps. Such small steps were made in data collection for two Dutch assessment programs. The programming was limited and not difficult because only a few structured variables were entered: the prescribed medicine, the result ascribed to each medicine and six symptoms.

Close monitoring of results in our prospective assessment of six homeopathic symptoms suggests that the human factor causes most bias. Therefore, even perfect software will not render reliable results if we neglect this. Most variance between observers was present in assessing results, the second most important source of variance is assessing symptoms, more so in vague symptoms. In our well-trained research group confirmation bias seemed limited, especially due to the long follow-up.

Assessing result is probably most influenced by assessing causality. If we calculate statistical parameters of symptoms, like Likelihood Ratio (LR), we must compare the data of patients responding well to a specific medicine with the remainder of the population. 'Responding well' is our reference or gold standard, but a rather weak one. In homeopathy a specific medicine population is only a small part of the whole practice population. In our prospective research the largest medicine population was the '*Natrium muriaticum* population' with 156 (4%) out of 4094 patients. The average medicine population consisted of 5 patients (0.12%). In a previous paper we calculated the influence of minor misjudgements of causality.[7] A simplified model is presented here.

Suppose a real population responding well to a specific medicine x consists of 1% (10 cases) of about 1,000 patients. Now, suppose that we recognise all these cases, but that we wrongly attribute 2% (20 cases) of all other cases to this medicine. Medicine x is prescribed and the result is good, but not due to the prescribed medicine. This would mean a very good assessment of causality, but it still has considerable effect on our results because of the much larger size of the remainder population, as shown in Figure 4. It results in 20 false positive cases, beside 10 real positive cases. Now, suppose that the prevalence of symptom A is 50% in the population responding well to medicine x and 10% in the remainder population. Then in our observed population of 30 patients seven (23.3%) will have symptom A, 5 out of the real (true positive) population, two out of the false positive population. If we compare LR in the real population and the observed population, LR is strongly under-estimated in the observed population: LR=2.33 instead of 5 in the real population.

LR = (prevalence target population) / (prevalence remainder population)	
LR (real population)= 50/10 = 5	LR (observed population) = 23.3/10 = 2.33

We see that only a small misjudgement of causal relationship has large influence on our calculations of symptom parameters and the amount of false positives might be much higher because of context effects. If we don't recognise all 'medicine x patients' in our population (false negatives) under-estimation of LR becomes even larger.

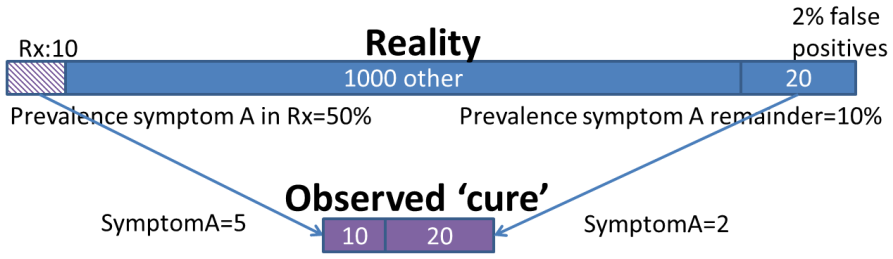

Figure 5: influence of weak gold standard on real data, see text

Bias caused by vagueness of symptoms will influence LR in the opposite direction, more so if the symptom is vague like 'Sensitivity to injustice'.[8] Considering the fact that in randomised controlled trials the placebo effect is surprisingly high we must beware of the possibility that false positives in our medicine populations are much more than 2%. This could imply that many LRs as we calculate them are too low and therefore part of them not statistically significant.

This means that we must maximise our efforts to assess causality in our cases. The GHHOS scale seems insufficient, only the improvement of more than just the complaint (GHHOS>=2) is an indication of a homeopathic effect. We might use an adaptation of the Naranjo algorithm used to assess causality in adverse effects of medicines.[9]Since recently this algorithm is discussed by the Clinical Data Working Group of the Homeopathic Pharmacopoeia of the United States. For a preliminary adapted version see Table 1. In this algorithm other indications for causality, like initial aggravation, necessity to repeat the medicine and absence of other factors, are taken into account. This adaptation has not yet been validated.[10]

Variance in assessment of symptoms is unavoidable, even in prospective research of well-defined symptoms. This is an indication for the considerable problems we can encounter in retrospective analysis of unstructured data. The problem of semantics, e.g. synonyms, might be solved by software, but not the human factor.

The limitations of our projects was that they were not designed to evaluate data collecting programs, the programs followed the needs of the participants in the projects. This paper is based on experience in the Netherlands. We are not aware of other groups with a comparable combination of projects. We realise that other groups have other needs, but that is just the point we want to make: think small first before trying to make programs that fit everybody's needs. Then, discuss results and feasibility of data collection.

How to advance in data collection?

We could think of hundreds of research questions when designing a homeopathic database, like 'Which method of analysis of symptoms is the best?'. But is the outcome reliable if we make mistakes in assessing results? Our experience with collecting just a few variables for quantitative analysis shows that the human factor should be closely monitored. Our prospective research of a small number of variables suggests in this respect:

- Start with a research protocol
- Meet your group of observers and keep in contact
- Limit the number of research questions and variables
- Define research questions and variables and obtain consensus about all variables
- Monitor the outcome frequently
- Discuss results with the observers and renew consensus if necessary

These prerequisites are more feasible for structured data. For unstructured data the possibility to define variables is limited, but the variable 'result' can and should always be defined. We can partly

structure other variables following the CARE guidelines, but the topics of the CARE checklist are mostly free text fields if stored in a computer database. We must realise that recording cases according to the CARE guidelines is time consuming. It might be necessary to adjust these guidelines after building up a considerable body of cases from a large number of practitioners. The assessment of causality might be more explicit in these guidelines.

For most practitioners available time is limited, especially if we want to record data of all cases. It would be a great progress if every practitioner could find all his successful cases ordered by medicine. This is also of direct advantage for the practitioner. This purpose is satisfied by a database with two structured variables: one containing a list of all homeopathic medicines and one containing a result-score like the GHHOS. Every practitioner can collect these data for his own use. Another incentive for collecting treatment data is the possibility to use recorded data for administrative goals, like billing and correspondence to colleagues.

The second step is sharing data between practitioners. Then consensus about assessing result becomes necessary. This step could be combined with a third variable: the medical diagnosis, to be selected from a predefined list, like ICD.[11] Then we can assess the prevalence of diseases in homeopathic practice, and the prevalence of the disease in each medicine population. Comparing the prevalence of a disease in separate medicine populations with the prevalence in the remainder of the population renders the likelihood ratio (LR) of the disease for each medicine, so we can see what medicines are most successful for the disease. Monitoring of results and feedback to the observers is also required.

What goes for medical diagnosis, also goes for homeopathic symptoms: we can assess prevalence and LR for any symptom. But we need a precise protocol, a coherent group of observers and consensus about the definition of the symptoms. And, of course, monitoring and feedback. This requires a limited number of extra variables in the database, depending on the number of symptoms we want to assess.

The next step could be adding free text fields for unstructured data. For unstructured electronic data CARE guidelines can ensure that all information for reproducibility is in the case description. This requires at least 12 (mostly free text) fields in the database.

For analysis of unstructured data we should develop a protocol. Such a protocol could be endorsed by consensus meetings assessing best cases, like our Materia Medica Validation.

Conclusion

Data collection by computer could be empowered by proper software if we develop this software step by step in close communication with users. The success depends on feasibility of data collection in daily practice and validity of outcome. We need more pilot studies from different perspectives.

Acknowledgements

The Committee for Methods and Validation organised the projects mentioned in this paper and consists of Erik Stolper, Rob Barthels, Roland Lugten en Lex Rutten. We thank all participants in the Materia Medica Validation and LR project for their contributions.
The data collection programs were developed by Jan Scholten (HARP), Jean Pierre Jansen, René van der Reijden and Lex Rutten.
Members of the Clinical Data Working Group of the Homeopathic Pharmacopoeia of the United States are: Joyce Frye, Todd Hoover, Peter Smith, David Riley, and Robbert van Haselen.

Table 1: adapted Naranjo algorithm (version June 2014), with permission of Clinical Data Working Group of the Homeopathic Pharmacopoeia of the United States

	Yes	**No**	**Not Sure or N/A**
1. Was there an improvement in the main symptom or condition for which the homeopathic medicine was prescribed?	+1	-1	0
2. Did the clinical improvement occur within a plausible timeframe relative to the drug intake?	+1	-2	0
3. Was there an initial aggravation of symptoms? (Need to define in glossary)	+1	0	0
4. Did the effect encompass more than the main symptom or condition, i.e. were other symptoms ultimately improved or changed?	+1	0	0
5. Did overall wellbeing improve? (suggest using validated scale)	+1	0	0
6. In terms of direction of cure, did some symptoms improve in the opposite order of the development of symptoms of the disease?	+1	0	0
7. In terms of direction of cure, did *at least two* of the following aspects apply to the order of improvement of symptoms: - from organs of more importance to those of less importance - from deeper to more superficial aspects of the individual - from the top downwards	+1	0	0
8. Did "old symptoms" (defined as non-seasonal and non-cyclical symptoms that were previously thought to have resolved) reappear temporarily during the course of improvement?	+1	0	0
9. Are there alternate causes (other than the medicine) that –with a high probability- could have caused the improvement? (Consider known course of disease, other forms of treatment, and other clinically relevant interventions)	-3	+1	0
10. Was the health improvement confirmed by any objective evidence? (e.g. lab test, clinical observation, etc.)	+2	0	0
11. Did repeat dosing, if conducted, create similar clinical improvement?	+1	0	0

References

[1] Stolper CF, Rutten ALB, Lugten RFG, Barthels RJWMM. Materia medica validation and meta-analysis: A post-graduate course combining learning and research. *Homeopathic Links* 2004;17(3):186-188
[2] Gill CJ, Sabin L, Schmid CH. Why clinicians are natural bayesians. *BMJ* 2005;330:1080-3
[3] Stolper CF, Rutten ALB, Lugten RFG, Barthels RJWMM. Improving homeopathic prescribing by applying epidemiological techniques: the role of likelihood ratio. Homeopathy 2002;**91**:230-238
[4] Rutten ALB., Stolper CF, Lugten RF, Barthels RJ. Statistical analysis of six repertory-rubrics after prospective assessment applying Bayes' theorem. *Homeopathy* 2009;98:26–34. Doi 10.1016/j.homp.2008.11.012
[5] Gagnier JJ, Kienle G, Altman DG,Moher D, Sox H,Riley D. The CARE guidelines: consensus-based clinical case reporting guideline development. Global Adv Health Med 2013;2(5):38-43. DOI:10.7453/gahmj.2013.008
[6] McKenzie C.R. Increased sensitivity to differentially diagnostic answers using familiar materials: implications for confirmation bias. Memory &.Cognition. 2006;34(3), 577-88.
[7] Rutten ALB, Stolper CF, Lugten RF, Barthels RJ. 'Cure' as the gold standard for likelihood ratio assessment: theoretical considerations. Homeopathy 2004;93:78-83
[8] Rutten ALB., Stolper CF, Lugten RF, Barthels RJ. Assessing likelihood ratio of clinical symptoms: handling vagueness. Homeopathy. 2003;92:182-186.
[9] Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA,. Janecek E, Domecq C, Greenblatt DJ. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 1981; 80: 289–95
[10] Rutten L (ALB). The importance of case histories for accepting and improving homeopathy. Comp Ther Med 2013;21:565-570
[11] www.who.int/classifications/icd/en/ accessed on 2014 April 3