Data Mining in Homeopathic Materia Medica

Rainer Schäferkordt[1]

[1] Scientific Society for Homeopathy (WissHom), Department Practice, Koethen, Germany

Homeopathy

Address for correspondence Dr. med. Rainer Schäferkordt, Fritz-Reuter-Str. 23, D-19258 Boizenburg, Germany (e-mail: rainer@schaeferkordt.de).

Abstract

Introduction Data-driven research stems from the original idea of homeopathy, which can be transferred to the 21st century with modern statistical concepts, especially techniques of data mining.

Groundwork In preparing a statistical approach to Materia Medica, abstraction of symptoms is pivotal. The main works of Materia Medica were indexed, creating the requirements for analyzing existing data.

Goals A manifold range of objectives are conceivable for analysis of Materia Medica: e.g., checking the quality of the existing data; assessing the prevalence of symptoms; calculating correlations between symptoms; assessing the discriminating power of symptoms; handling of polar symptoms; analyzing cross-references between medicines; calculating domains for each medicine, such as spheres of action, organs and side localization; building a new repertory from scratch.

Findings As a first step, a comparison between data of Materia Medica, prognostic factor research (PFR) and repertories for six selected repertory rubrics was performed, showing moderately high correlations between Materia Medica and PFR.

Conclusion Methods of data mining applied to Materia Medica can help to analyze existing data to a maximum extent and contribute to the further development of the homeopathic method, both scientifically and practically.

Keywords
- data mining
- Materia Medica
- Bayes' theorem
- prognostic factor research

Introduction

When Hahnemann presented the concept of homeopathy in The Organon, it was his great concern not to rely on 'supernatural delusions', but on empiricism—both in the healthy and in the sick.[1] The result is well known: he left behind a collection of proving symptoms[2,3] and patient journals,[4] which we are still working through today. Hahnemann's contemporaries were overwhelmed by this empirical approach, which generated 'observations' on a large scale, and criticized his method as unscientific, censuring the 'heaps of symptoms' and 'orderless complexes', which they knew nothing about. Hering replied: 'In such a new field of research where, unavoidably, every observer is surrounded by uncertainties, nothing could be gained by theorizing or soaring into the clouds of opinions. The facts had first to be collected, one by one, and then arranged in some way, that they might be applied again and again, until the probabilities increased, and a scientific certainty was within reach'.[5]

Hering's words can be read as a stunningly exact description of what is nowadays called data mining. In modern definition, 'data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems'.[6] Data mining encompasses database techniques and statistical concepts that are ideal for dealing with large quantities of data —in our case homeopathic symptoms. It is true that every medicine can and should always be considered qualitatively, and many an individual symptom is worth meditating on. However, finding clusters, patterns, correlations and anomalies within and between medicines is the domain of statistics. In addition, in homeopathy, as everywhere else in empirical sciences, we have to deal with bias factors, the relevance of which is generally underestimated. Bias factors are primarily

received
December 26, 2024
accepted after revision
April 17, 2025

Georg Thieme Verlag KG, Oswald-Hesse-Straße 50, 70469 Stuttgart, Germany

DOI https://doi.org/10.1055/a-2591-4676.
ISSN 1475-4916.

psychological phenomena, distortions of perception and human sources of error. Statistical approaches should and can help to minimize these bias factors. And even more, they can help to discover things in our medicine data that are not visible to 'the naked eye'. In this way, we can gain a different, broader, more objective view of the medicines.

Groundwork

The existing homeopathic Materia Medica offers enormous research potential for data mining approaches. However, for applying statistical methods to Materia Medica, indexing—and therefore abstraction—of symptoms is pivotal. Only by back-tracing the linguistic diversity of provings and clinical symptoms to the meaning, independent of the respective formulation, can comparability and clustering of symptoms be assured. Thus, one examiner may have experienced him- or herself as 'irritable', another as 'angry', a third as 'aggressive', and a fourth as 'morose'—but what is meant is almost always the same. In the field of repertories, for example, it makes it difficult to use them if this abstraction is not achieved, and nearly identical information can be found in many different places. This is one of the shortcomings of Kent's repertory, while the works of Boger[7,8] and Boenninghausen[9] stressed this task much more. As repertories are indexes to the Materia Medica, they should offer less precision of the personal expression of each symptom, but more oversight of a combination of symptoms and a differentiation between several eligible medicines.

Phenomena

The concept of phenomena seems capable to negotiate our manifold Materia Medica. A phenomenon (from the ancient Greek φαινόμενον [fainómenon], 'one that shows itself, one that appears') means a delimitable unit of experience.

These considerations are crucial:

- A phenomenon refers primarily to what can be experienced with the senses (as opposed to thought- or theory-based clinical diagnostic entities) and thus coincides to a large extent with Hahnemann's epistemological ideas.
- Each phenomenon can be defined, delimited to other phenomena and provided with synonyms.
- All phenomena can be put into a poly-hierarchical structure, forming a thesaurus, so that any number of sub- and super-terms can be assigned to each phenomenon.
- Symptoms can be broken down into separate phenomena, and separate phenomena can be combined to symptoms again.

Phenomena refer to all areas of experience and can also be clinically characterized. For example, the anatomical structure 'leg' is a phenomenon, as is 'pain', 'inflammation' or 'stage fright'. The concept of phenomena is not strictly interpreted, so that also clinically diagnostic—not directly experienced, but in Materia Medica relevant—terms, such as 'eclampsia', can be taken into account.

Building a Thesaurus

Based on this concept, a bilingual thesaurus (English/German) was created by the present author using a bottom-to-the-top, inductive approach. The entire vocabulary of the most relevant Materia Medica works (see below) was indexed, and synonymous relationships between terms were established. Tools used in this process were WordNet,[10] the International Code of Symptoms,[11,12] MeSH,[13] AMDP manual,[14] and Dict.cc[15] and DeepL[16] for bidirectional translation. This resulted in a thesaurus with approximately 8,000 phenomena, comprising 24,000 English and 25,000 German synonyms (including all inflections).

Example: The phenomenon 'Anxiety' comprises the following synonym words found in Materia Medica: affright, affrighted, affrighting, affrights, afraid, alarmed, anguish, anxietas, anxieties, anxiety, anxious, anxiously, anxiousness, apprehend, apprehends, apprehension, apprehensions, apprehensive, apprehensiveness, coward, cowardice, cowardly, cowards, dread, dreaded, dreadful, dreadfully, dreading, dreads, fear, feared, fearful, fearfully, fearfulness, fearing, fears, fearsome, fright, frighten, frightened, frightening, frightful, frightfully, frightfulness, frights, horror, horrors, misgiving, misgivings, mysophobia, over-anxious, panic, panics, phobia, phobic, pusillanimity, pusillanimousness, scared, terrified, terrify, terrifying, timid, timidity, timorous, timorously, timorousness, trepid.

To identify the phenomena relevant for homeopathic work, the next step was to compare them with all the terms of the most important repertories (Kent,[17] Boger, Boenninghausen, Phatak[18]). The result of this analysis was approximately 2,500 phenomena whose definitions and synonyms have now been intensively revised. This was followed by a review of all associated symptoms, with extensive disambiguation of polysemic words (e.g., the word 'light' must be assigned to the phenomena 'sunlight', 'lightness' or 'bright', depending on the context). Beyond the repertorial vocabulary, approximately 1,000 further phenomena were identified as relevant and elaborated.

In the next step, all phenomena were assigned to a poly-hierarchical structure. A hierarchical structure (instead of alphabetical) is a prerequisite for the concept of abstraction and generalization, an (implicit) key feature in many approaches of homeopathic case analysis. For instance, a typical anatomical hierarchy is: Superior Extremity—Hand—Fingers—Fingertips. And it has to be poly-hierarchical to depict different hierarchical approaches to pathology and physiology. For example, the phenomenon 'liver' can be subordinated to both the phenomena 'glands' and 'abdomen'. In addition, each phenomenon was classified into one of 25 categories (anatomy, physiology, pathology, psyche, times, modalities, social, objects, physics, chemistry, biology, geography, etc.).

Indexing Materia Medica

This thesaurus was in turn used to index the following Materia Medica:

- Allen TF, *Encyclopedia of Pure Materia Medica*.
- Allen TF, *Handbook of Materia Medica and Homœopathic Therapeutics*.

- Boericke W, *Homœopathic Materia Medica*.
- Boger CM, *Synoptic Key*.
- Clarke JH, *Dictionary of Practical Materia Medica*.
- Hahnemann S, *Materia Medica Pura*.
- Hahnemann S, *Chronic Diseases*.
- Hering C, *The Guiding Symptoms of Our Materia Medica*.
- Lippe C, *Keynotes and Red Line Symptoms*.
- Lippe C, *Textbook of Materia Medica*.

The encyclopedic works of Allen, Clarke and Hering contain the greatest wealth of knowledge, so that these can be regarded as the most important foundation for data mining. Furthermore, works by renowned clinicians have been added, which more closely reflect the weighting of symptoms based on practical experience. In addition, Hahnemann's works have been incorporated. A total of approximately 20,000 book pages (~812,000 symptoms) of both English- and German-language Materia Medica was processed.

As symptoms of Materia Medica may be complex and consist of long sentences, confusion of context of phenomena may occur. To avoid this, a computer linguistic tool, coreNLP,[19] was used to split longer sentences into meaningful entities.

Statistics: Bayes' Theorem

Another pivotal aspect of data mining is the calculation of frequencies of occurrences of symptoms or phenomena. The statistical approach of Likelihood Ratio (LR), based on Bayes' theorem, which was introduced into homeopathy by Dr. Lex Rutten, among others,[20] may be an eligible instrument for this.

In medical research, calculation of LR is mostly used for diagnostic tests. It relies on the sensitivity and specificity of a diagnostic test and can be expressed as:

$$LR = \frac{Pr(T+D+)}{Pr(T+D-)}$$

where T is the test, and D is the diagnosis. So, LR is the probability of a person who has the disease testing positive divided by the probability of a person who does not have the disease testing positive.

Transferred to Materia Medica, it can be said that 'test' is the symptom (developed by a prover or a patient), and 'diagnosis' is the medicine (used in the proving or on a patient). So, we relate the count of a distinct symptom of a medicine to the count of all known symptoms of this medicine, and this to the prevalence of this distinct symptom in the remaining Materia Medica. Thus, a 'relative' count is performed, which provides valid results largely independent of the number of sources used. To avoid radical outliers, thresholds should be used for minimum (0.5) and maximum (99.0) values.[21]

For this statistical approach to Materia Medica, abstraction of symptoms, as described above, is crucial. LR values can be calculated for single phenomenon–medicine relationships, or for combination of a limited amount of phenomena, but not for an entire complex symptom, which are often found in Materia Medica, as these are too individual. For instance, the symptom 'Anxious dream, of numberless dogs and cats, with loud talking' (*Graphites*) is indexed like this:

Anxiety—Dreams—Dogs—Cats—Talking—Loud.

Looking at the whole Materia Medica, there are several symptoms including the phenomena 'Anxiety—Dreams—Dogs', but only for *Graphites* can the combination 'Anxiety—Dreams—Dogs—Talking' be found. So, statistical analysis is sensible only for combinations of phenomena occurring at least twice.

Goals

Having prepared Materia Medica in this way, the following goals and methods are conceivable as a starting point:

- By calculating correlations, similar or possibly largely identical phenomena can be identified (e.g., sensations, psychological phenomena).
- Statements on the discriminating power of symptoms can be made (which symptoms help to distinguish different drugs from each other?).
- In a similar way, polar symptoms can be identified or verified (or falsified) by negative correlations.
- By analyzing the prevalences, i.e. the frequencies, the phenomena or symptoms that are particularly relevant for the choice of medicine can be identified, in the sense of Aphorism §153: a high prevalence of a symptom shows that it is not 'special, unusual, peculiar'—but rather common or frequently occurring.
- For each medicine, domains can be calculated in terms of dominating organs, spheres of action, side localization, etc.
- Cross-references within the medicines (and the clinical experience often expressed therein) can be systematically and statistically evaluated.
- New possibilities open up for checking the quality of the data: in particular, by storing the identity of the provers (Hahnemann and Allen), matches or clusters in individual persons can be analyzed across different medicines, thus identifying any idiosyncrasies or other distortions.
- Differences between proving symptoms, poisonings and clinical symptoms can be systematically analyzed.
- Differences of data between different works, different eras and possibly also countries and cultures can be analyzed.
- The congruence of symptoms across different works, modifications or errors that individual symptoms have undergone over time and authors can be identified.
- Domains within medicine classes, biological systematics etc. can be calculated (for example, are animal medicines more aggressive than herbal ones?).
- Miasmatic questions can be answered using statistical methods.
- Data for clinical verification (e.g., PFR) can be compared with the existing Materia Medica.
- Repertories can be checked for their matching with Materia Medica.
- A new repertory can be built from scratch.

- Theorizing: In the sense of an inductive science, approaches to homeopathic theory can be tested and modified from the data of the Materia Medica.

Several these questions do not appear to be new and have possibly already been answered—e.g., the question of the sphere of action of medicines, the polarity of symptoms, or the differences between classes of medicines. However, we often rely on the old or even more recent masters ('expert opinions') without having had the tools to test their approaches independently with a data-based ('evidence-based') approach. Even if authors such as Boger have undoubtedly achieved great things, bias effects, misinterpretations, etc. can also be assumed here. This applies to an even greater extent to more recent, strongly interpretative approaches.

First Set of Findings: Comparison of Materia Medica and Prognostic Factor Research

In 2009, the Dutch group of VHAN (Vereniging van Homeopathische Artsen Nederland) presented a study in which 10 homeopaths recorded the prevalence of 6 defined symptoms in a group of 4,094 patients over a period of 3½ years.[22] The presence of these symptoms was actively assessed and documented for each patient and later compared with successful prescriptions. In other words, it was not cured symptoms that were recorded, but the presence of symptoms at the time of prescription, in the sense of prognostic factors (PFR). These data were used to calculate LR values, which in turn were used to calculate entries for the corresponding repertory rubrics (according to Kent's repertory).

These results now offer the possibility of checking the congruence of data from very different sources: to what extent do the prognostic results from clinical cases agree with Materia Medica? In addition, a comparison was made with the data from Kent's classical repertory, and also with the Complete Repertory, one of the most widely used and comprehensive repertories currently available. The question to what extent the continuous upgrading of the medicine entries is justified or reflects clinical reality can also be investigated.

To make data from different sources like these comparable, the grading of repertory entries was used as lowest common denominator. In the case of Materia Medica data, LR values were calculated for each symptom–medicine relation, and these were converted to repertory grades. Similar to the VHAN study, LR values from 1.5 to 2.9 were transformed to a grade 1 entry (plain type), 3 to 5.9 to grade 2 (italics) and 6 and above to grade 3 (bold). LRs below 1.5 were interpreted as a question mark for this entry.

In addition, a comparison of the accessible LR values (VHAN and Materia Medica) was undertaken. Pearson correlation coefficients were calculated (list-wise case exclusion, two-sided p) for repertory grades and LR values. The correlation analysis was done with PSPP.[23]

The results are shown in ►**Tables 1** to **6**. In summary, there is a surprisingly high degree of congruence between the data from Materia Medica and the clinical prognostic data (VHAN)—particularly surprising in respect of the fact that the data were collected in such different ways.

However, there are differences with regard to the individual symptoms, which are certainly also due to the nature of the symptoms. For example, the spectrum of symptoms chosen for the VHAN study deliberately includes symptoms that are difficult to collect and characterized by subjectivity (sensitive to injustice) or clinical entities (herpes labialis).

In the VHAN study, each symptom was clearly defined, but such definitions are largely missing for repertory rubrics, which limits comparability. In Materia Medica, semantic aspects can be clarified well using the referenced symptoms; however, the intensity of a symptom is generally clearly defined only in the VHAN study (e.g., recurrent herpes labialis: at least 6 times per year).

Whilst the correlations between the VHAN data and Materia Medica are mostly high, the correlations are lower (for both) to Kent's repertory, and even lower to Complete repertory. One possible explanation for this is that the values of the entries in the original Kent were determined more by the clinical-intuitive estimation of the authors, whereas the Complete uses the concept of absolute counting—an approach that inevitably leads to invalid values over time and thus shares the fate of any inflation. With regard to Materia Medica, it should be noted that more recent sources, especially from the second half of the 20th century onwards, are only included to a very limited extent. Whilst not much additional knowledge can be expected from newer sources with regard to established medicines, small and new medicines may be under-represented.

Both historical Materia Medica data and the concept of PFR are confirmed to large extents in this comparison. PFR can make an important contribution to evaluating and supplementing our knowledge of medicines. Its main limitations are that the medicines studied are somewhat random and, not least, reflect the prescription spectrum of the participating doctors—which may explain one or two small, unexpected medicines (e.g., *Calc-m*). This information requires further confirmation. Other bias factors, such as confirmation bias, are discussed by the VHAN authors. It also becomes clear that the number of medicines that can be evaluated in this way is limited, despite the considerable effort of the study.

In this respect, case data collected in this way should always be compared with the existing Materia Medica. However, drug provings, selective case reports and the entire Materia Medica are also subject to uncertainties, reproduction errors and distortions. Therefore, a valid Materia Medica—as well as repertories—must be fed by methodically different, statistically correctly analyzed sources. The most reliable symptoms lie at the intersection of all these sources.

Limitations and Problems

Analyzing Materia Medica in the prescribed way, the following shortcomings should be mentioned first and foremost:

Table 1 Symptom/repertory rubric 'Fear of death'

Medicine	VHAN	MM	Kent	Complete	LR VHAN	LR MM
Acon	3	3	3	4	10.6	9.4
Am-c	2	1	1	3	5.82	1.1
Anac	3	0.5	1	3	11.1	0.7
Arg-n	1	0.5	2	3	2.01	0.9
Ars	2	2	3	4	5.95	3.9
Calc	1	1	3	4	1.39	1.3
Carc	1			1	2.45	
Ign	1	0.5	1	3	2.38	0.5
Kali-p	2	1	1	3	3.27	2.1
Lac-c	3	1	3	4	6.55	1.8
Lach	1	1	2	4	2.51	1.3
Lyc	1	0.5	2	3	1.21	0.9
Mag-c	1	0.5		1	2.75	0.8
Nat-m	0.5	0.5	1	3	0.49	0.3
Nux-v	1	0.5	2	4	1.3	0.7
Phos	1	1	3	4	1.37	1.1
Puls	0.5	1	2	4	0.88	1.6
Sep	1	0.5	1	3	1.7	0.1
Sil	1	0.5		1	1.58	0.2
Sulf	0.5	0.5	1	3	0.29	0.7
Verat	3	1	2	4	8.74	2.0

VHAN, MM: $r = 0.517$ $p \leq 0.02$.
VHAN, Kent: $r = 0.232$ $p \leq 0.2$.
VHAN, Comp: $r = 0.288$ $p \leq 0.2$.
MM, Kent: $r = 0.374$ $p \leq 0.001$.
MM, Comp: $r = 0.179$ $p \leq 0.05$.
LR VHAN, LR MM: $r = 0.59$ $p \leq 0.01$.
VHAN = Rutten et al 2009[22]; MM = Materia Medica; Comp = Complete Repertory.
The repertory grades are transformed to numbers: Plain = 1, italics = 2, bold = 3, bold & italics = 4. For the VHAN and Materia Medica data, a grade of 0.5 is assumed when the LR value is < 1.
$r < 0$: negative correlation; $r < 0.1$: no correlation; $r < 0.3$: weak correlation; $r < 0.6$: medium correlation; $r \geq 0.6$: strong correlation.
$p \leq 0.01$: highly significant; $p \leq 0.05$: significant; $p \geq 0.05$: non-significant.

- Sources: The focus of the analyzed Materia Medica was initially placed on encyclopedias and classical authors. More recent works have not been included due to copyright restrictions on the one hand, and the controversial validity of recent provings and hypothetical Materia Medica on the other. To what extent the conservative source situation is a disadvantage can hardly be considered at this point in time, since the different methods of homeopathy have not yet been evaluated against each other. Even with the works now incorporated, heterogeneous quality can be assumed, so that the demand for a better, empirical evaluation of the Materia Medica that meets today's standards remains unchanged. Likewise, clinical cases were not included.

Table 2 Symptom/repertory rubric 'Diarrhea from anticipation'

Medicine	VHAN	MM	Kent	Complete	LR VHAN	LR MM
Arg-n	3	3	2	3	11.1	14.7
Bell	1				2.17	
Calc	1	1			1.84	2.0
Caust	1				1.48	
Cimic	3				6.52	
Elaps	2				5.71	
Gels	3	3	2	4	14.5	24.7
Merc	1	1			1.69	2.1
Ph-ac	3	2	2	3	7.12	4.9
Staph	1				2.14	

VHAN, MM: $r = 0.913$ $p \leq 0.02$.
VHAN, Kent: $r = \text{NaN}$.
VHAN, Comp: $r = \text{NaN}$.
MM, Kent: $r = \text{NaN}$.
MM, Comp: $r = 0.899$ $p \leq 0.1$.
LR VHAN, LR MM: $r = 0.953$.
Abbreviation: NaN, not enough data.

Table 3 Symptom/repertory rubric 'Sensitive to injustice'

Medicine	VHAN	MM	Kent	Complete	LR VHAN	LR MM
Am-m	2			1	4.34	
Ambr	1				2.71	
Anac	2			3	5.47	
Aur	1			1	1.67	
Bell	1			1	2.07	
Bor	2				4.34	
Calc	1			1	1.01	
Calc-m	1				1.97	
Carc	1			2	2.29	
Caust	2	2		3	4.39	3.0
Chin	1			1	1.55	
Cocc	2	2		3	4.2	4.7
Cupr	1			1	1.67	
Ign	1			1	1.98	
Kali-bi	1				1.55	
Med	1			1	2.27	
Merc	1			1	1.41	
Nat-m	1	0.5		1	1.04	0.22
Nux-v	0.5	1		3	0.81	2.1
Ph-ac	1			1	1.89	
Sep	0.5			2	0.81	
Staph	1	3		3	1.01	20.4

VHAN, MM: $r = 0.363$ $p \geq 0.2$.
VHAN, Kent: $r = \text{NaN}$.
VHAN, Comp: $r = 0.334$ $p \leq 0.2$.
MM, Kent: $r = \text{NaN}$.
MM, Comp: $r = 0.451$ $p \geq 0.2$.
LR VHAN, LR MM: $r = -0.237$ $p \geq 0.2$.

Table 4 Symptom/repertory rubric 'Herpes lips'

Medicine	VHAN	MM	Kent	Complete	LR VHAN	LR MM
Aloe	3				8.06	
Bar-c	2			3	3.66	
Bor	2	1	1	1	8.06	1.15
Bry	2			1	3.09	
Caust	1		1	1	2.65	
Chin	1				2.87	
Gels	2				3.09	
Lach	1	1	1	2	1.92	2.5
Lyc	1			1	1.89	
Nat-m	2	3	3	4	3.35	16.7
Rhus-t	1	3	3	4	2.11	12.5
Sep	1	3	3	4	2.21	10.0
Sil	1	2	1	1	1.83	3.8
Staph	2				3.17	
Stram	2				4.47	
Sulf	1	0.5	1	1	1.37	0.2
Thuj	1	0.5		1	1.67	0.45

VHAN, MM: $r = 0.136$ $p \geq 0.2$.
VHAN, Kent: $r = 0.149$ $p \geq 0.2$.
VHAN, Comp: $r = 0.137$ $p \geq 0.2$.
MM, Kent: $r = 0.513$ $p \leq 0.02$.
MM, Comp: $r = 0.334$ $p \leq 0.1$.
LR VHAN, LR MM: $r = -0.069$ $p \geq 0.2$.

Although there is no lack of case reports in the homeopathic literature, these are fraught with several problems.[24] Despite various approaches,[25,26] a systematic review of literature cases, and in particular the prospective collection of cases documented according to specific criteria, has not yet taken place to any significant extent.

- Thesaurus: Even though the thesaurus has been compiled with great care and with the help of various reference works, *the* thesaurus will never exist. The definition and delimitation of terms and the assignment of synonyms always remain to some extent a subjective process or the task of lengthy clinical–empirical research.[27,28]
- Indexing of the symptoms: Using the thesaurus, the indexing was primarily performed in an automated way, with subsequent manual checking of all assignments. Due to the large number of symptoms, mis-classifications cannot be completely ruled out here.
- Syntactic analysis of the symptoms: This step was performed with the help of the coreNLP module. This neural network-based technique is adaptive, but the 'training' of a specific 'model' is a very time-consuming process that is never complete. Moreover, even for experienced homeopaths, the syntactic semantics of a complex symptom is not always clearly comprehensible or abstractable.

Table 5 Symptom/repertory rubric 'Loquacity'

Medicine	VHAN	MM	Kent	Complete	LR VHAN	LR MM
Ambr	2	1	1	3	5.8	2.8
Anac	1	1	1	3	2.57	1.6
Bell	1	1	2	4	2.95	1.9
Calc	0.5	0.5	1	3	0.4	0.05
Calc-m	1			1	2.8	
Caust	1	0.5	1	2	1.0	0.3
Cimic	2	2	2	4	4.41	4.9
Dig	2			1	10.3	
Hyos	2	3	3	4	5.19	17.8
Lach	2	3	3	4	5.34	11.6
Lyc	0.5			3	1.63	
Med	1				1.93	
Nat-m	0.5		1	1	0.38	
Nux-v	0.5	0.5	1	4	0.76	0.4
Phos	0.5	0.5	2	3	0.8	0.6
Sacch	2				4.41	
Sep	1				1.33	
Stram	2	3	3	4	3.43	16.3
Sulf	0.5		1	3	0.69	
Tarent	2	0.5	1	3	3.85	0.02
Tub	1	0.5		3	1.92	0.3
Verat	2	2	2	4	7.74	4.4

VHAN, MM: $r = 0.718$ $p \leq 0.005$.
VHAN, Kent: $r = 0.561$ $p \leq 0.05$.
VHAN, Comp: $r = 0.274$ $p \geq 0.2$.
MM, Kent: $r = 0.342$ $p \leq 0.005$.
MM, Comp: $r = 0.208$ $p \leq 0.05$.
LR VHAN, LR MM: $r = 0.484$ $p \leq 0.1$.

Summary

Data mining enables us to extract much more out of the Materia Medica than previous approaches did. Going beyond the results presented, this paper sketches a research agenda to support a data-driven homeopathy by analyzing existing data. An important task in this is to measure the quality of the Materia Medica data. Most of these data are at least 100 years old, and most of homeopathic practice relies on it. But the circumstances of how these historical data were acquired are often unclear.[29] Basically, drug provings, intoxications and clinical patient data are the most important sources. In some works, such as Allen's Encyclopedia, these sources are given, at least partly, but most symptoms stand for themselves, often remodelled by different authors over the years, reflecting clinical experience as well as copying errors.[30] By categorizing and comparing symptoms according to their origin, to authors, to different provers, or to ages, bias signals may be found. By comparing historical data to contemporary (e.g., PFR) data, systematic errors may be identified. Comparison

Table 6 Symptom/repertory rubric 'Grinding teeth during sleep'

Medicine	VHAN	MM	Kent	Complete	LR VHAN	LR MM
Am-m	3				7.54	
Arg-n	1				1.44	
Bell	2	3	3	4	5.46	7.4
Calc	0.5	1	1	3	0.74	1.2
Calc-m	2				3.42	
Calc-p	1				2.7	
Carc	2			1	3.11	
Cench	3				9.42	
Cocc	2			1	4.36	
Ign	1	1	2	3	2.29	2.6
Merc	1	1	2	3	2.47	1.1
Ph-ac	1				1.63	
Psor	1	2	1	1	2.69	3.0
Puls	1				1.6	
Sep	1	0.5	1	1	1.63	0.8
Staph	1				1.76	
Thuj	1	0.5	1	1	1.56	0.04

VHAN, MM: $r = 0.759$ $p \leq 0.05$.
VHAN, Kent: $r = 0.807$ $p \leq 0.05$.
VHAN, Comp: $r = -0.09$ $p \geq 0.2$.
MM, Kent: $r = 0.38$ $p \leq 0.05$.
MM, Comp: $r = 0.39$ $p \leq 0.05$.
LR VHAN, LR MM: $r = 0.925$ $p \leq 0.002$.

of different sources may minimize different kinds of errors. Comparisons of symptoms that emerge under different circumstances can disclose the most reliable data. The results of the correlation analysis of selected symptoms showed that different approaches can produce similar results. Every kind of data collection has its advantages and its disadvantages. Whilst collection of prospective data often requires high levels of effort and participants, many approaches of data mining touch questions of ethics and data privacy.[31] In dealing with long-published data, these tasks do not arise.

Conclusion

This kind of research ties in with the original ideas of Hahnemann, who based the invention of homeopathy on a broad base of empirical data and only to a lesser extent on theory. It also connects to contemporary approaches in medical, data-driven research in general. It may help to amplify the scientific identity of homeopathy as a data-based method, overcoming speculative, eminence-based statements and theories.[32] In addition to research on effectiveness (especially randomized controlled trials) and mechanisms of action (especially laboratory studies), the further development of therapeutic knowledge and study methods can be considered the most important step toward a scientific and improved homeopathy.

Highlights

- Methods of data mining continue the empirical idea of homeopathy.
- Indexing of symptoms is pivotal for further analysis.
- Using Bayes' theorem as statistical approach assures valid outcomes in processing large amounts of data.
- Comparison of PFR data to Materia Medica shows good correlations: much better than comparison to repertories.
- A broad research agenda is conceivable for data mining approaches.

Funding
None.

Conflict of Interest
R.S. is CEO of Analogon Enterprise GmbH, which develops and distributes the software Analogon.

References

1 Hahnemann S. Organon der Heilkunst. Textkritische Ausgabe der sechsten Auflage. Heidelberg: Haug; 1999
2 Hahnemann S. Reine Arzneimittellehre. Dritte Auflage von 1830. Heidelberg: Haug; 1995
3 Hahnemann S. Chronische Krankheiten. Zweite Auflage von 1835. Heidelberg: Haug; 1995
4 Hahnemann S. Die Krankenjournale. Digitale kritische Datenbank-Edition. Institut für Geschichte der Medizin der Robert-Bosch-Stiftung. Available at: https://www.igm-bosch.de/krankenjournale.html. Accessed November 26, 2024
5 Hering C. Analytical repertory of the symptoms of the mind. Philadelphia: American Homoeopathic Pub. Society; 1881
6 Association for Computing Machinery. SIGKDD. Data Mining Curriculum: A Proposal. Available at: https://www.kdd.org/curriculum/index.html. Accessed November 26, 2024
7 Boger CM. General Analysis. 6th edition. Mumbai: Roy & Company; 1939
8 Boger CM. A Synoptic Key of the Materia Medica. 4th edition (1931). New Delhi: B. Jain Publishers; 2000 (Reprint)
9 Boenninghausen von C. Therapeutisches Taschenbuch. Leipzig: Marggraf's homöopathische Officin; 1897
10 Princeton University. About WordNet. Princeton University: WordNet, 2010. Available at: https://wordnet.princeton.edu/. Accessed April 17, 2025
11 Bleul G. Systematik von Repertorien – grundsätzliche Überlegungen. AHZ 2011;256:12–15
12 Waldvogel L, Schäferkordt R. Revision der repertorialen Systematik der Gemütssymptome. AHZ 2017;262:12–17
13 Deutsches Institut für Medizinische Dokumentation und Information, Ed. Medical Subject Headings—MeSH (German edition). Cologne, 2015
14 Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie, Ed. Das AMDP-System. Manual zur Dokumentation psychiatrischer Befunde. 8th Ed. Göttingen Hogrefe; 2007
15 Hemetsberger P. dict.cc. Available at https://www.dict.cc. Accessed April 17, 2025
16 Kutylowski J. DeepL. Available at https://www.deepl.com. Accessed April 17, 2025
17 Kent JT. Repertory. 6th edition (1945). New Delhi: B. Jain Publishers; 1991 (Reprint)
18 Phatak SR. A Concise Repertory of Homoeopathic Medicines. 4th edition. New Delhi: B. Jain Publishers; 2005
19 Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In:

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics 2014: System Demonstrations: 55–60
20 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
21 Rutten AL, Stolper CF, Lugten RFG, Barthels RW. A Bayesian perspective on the reliability of homeopathic repertories. Homeopathy 2006;95:88–93
22 Rutten ALB, Stolper CF, Lugten RFG, Barthels RWJM. Statistical analysis of six repertory rubrics after prospective assessment applying Bayes' theorem. Homeopathy 2009;98:26–34
23 GNU pspp 1.6.2. Free Software Foundation, Inc.; 2007. Available at https://www.gnu.org/software/pspp/. Accessed April 17, 2025
24 Schäferkordt R, Kösters C. Das WissHom-Projekt 'Empirium': Forschung und Qualitätssicherung durch Falldokumentation. AHZ 2015;260:1–5
25 Baas C. The pitfalls of clinical case research: lessons from the Delphi Project. Homeopathy 2004;93:21–26
26 Cámpora CN. Homeopathic case documentation: a concrete case study from the Argentinian database BRECHA (Banco de Reporte y Estudio de Casos Homeopáthicos de Argentina). ICE 12, conference transcript. Available at: https://www.wisshom.de/whwp/wp-content/uploads/2020/03/b9826_ice12_kongressband_gesamt_aktualisiert.pdf. Accessed November 29, 2024
27 Rutten LA, Frei H. Frequently occurring polar symptoms assessed by successful cases. Homeopathy 2012;101:103–111
28 Rutten L. Is the doctor who cures right? Or should we look for black swans? Complementary aspects of homeopathy's scientific identity. Homoeopath Links 2023;36:103–111
29 Lucae C, Wischner M. Rein oder nicht rein? Zur Quellenlage von Hahnemanns Arzneimittellehre. ZKH 2010;54:13–22
30 Allen TF. The Encyclopedia of Pure Materia Medica. Introduction. New Delhi: B. Jain Publishers; 1997 (Reprint)
31 Haserück A. Chancen der Künstlichen Intelligenz mitgestalten. Dtsch Arztebl 2023;43:A1768–A1770
32 Manchanda RK, Khurana A, Van Wassenhoven M, et al, eds. Scientific Framework of Homoeopathy. Central Council for Research in Homoeopathy, Liga Medicorum Homoeopathica Internationalis; European Committee of Homoeopathy. New Delhi, April 2021

