ORIGINAL ARTICLES

Using a matrix-analytical approach to synthesizing evidence solved incompatibility problem in the hierarchy of evidence

Harald Walach*, Martin Loef

Institute of Transcultural Health Studies, European University Viadrina, Frankfurt (Oder), Germany

Accepted 23 March 2015; Published online 4 July 2015

Abstract

Objectives: The hierarchy of evidence presupposes linearity and additivity of effects, as well as commutativity of knowledge structures. It thereby implicitly assumes a classical theoretical model.

Study Design and Setting: This is an argumentative article that uses theoretical analysis based on pertinent literature and known facts to examine the standard view of methodology.

Results: We show that the assumptions of the hierarchical model are wrong. The knowledge structures gained by various types of studies are not sequentially indifferent, that is, do not commute. External validity and internal validity are at least partially incompatible concepts. Therefore, one needs a different theoretical structure, typical of quantum-type theories, to model this situation. The consequence of this situation is that the implicit assumptions of the hierarchical model are wrong, if generalized to the concept of evidence in total.

Conclusion: The problem can be solved by using a matrix-analytical approach to synthesizing evidence. Here, research methods that produce different types of evidence that complement each other are synthesized to yield the full knowledge. We show by an example how this might work. We conclude that the hierarchical model should be complemented by a broader reasoning in methodology. © 2015 Elsevier Inc. All rights reserved.

Key Words: Evidence-based medicine; Hierarchy of methods; Methodology; Placebo; Quantum theory; Matrix analysis

1. Introduction

The current consensus on how to create empirical evidence within medical research postulates a hierarchy of evidence. This means that research findings have to be considered in hierarchical order according to the methodologic strength of the design used on the basis of the internal validity of research methods [1]. Because internal validity is the highest in randomized controlled trials (RCTs), they are placed at the top of the hierarchy. And because meta-analyses normally combine effect size estimates from single RCTs, effect size estimates from such meta-analyses or summaries from systematic reviews are considered even more valid, forming the peak of the pyramid. Other types of evidence, for instance, evidence from cohort studies in naturally occurring cohorts and comparisons of such cohorts, evidence from case-control studies, case-series or single-armed observational studies, are normally considered weaker evidence because the internal validity of such studies is difficult to gauge.

This “received view” has some very important arguments in its favor. Randomized studies, if large enough [2], distribute potential confounders equally between groups and thus destroy their potential correlation with the outcome. Because many confounders of potential treatment effects are still unknown, any naturalistic, non-randomized study may potentially be influenced by covert confounders. Indeed, bias is introduced by a lack of randomization [3–6] or if randomization is compromised [7]. Thus, it seems intuitively clear to use “lower grade” evidence only as long as better evidence does not exist and to dismiss it as soon as better evidence, namely results from RCTs and, finally, meta-analyses, becomes available [8,9]. Although this rigid stance has recently been replaced by a more modest claim that demands a reasonable effect size ruling out confounders through any kind of

Conflicts of interest: None

Funding: The funding was provided by the Samueli Institute for work around that concept. M.L. was supported by a grant from the Oberberg Foundation, Berlin, while working on this article.

* Corresponding author. Institute for Transcultural Health Studies, European University Viadrina Frankfurt (Oder), Grosse Scharrnstrasse 59, 15230 Frankfurt (Oder), Germany. Tel.: +49-335-5534-2380; fax: +49-335-5534-2748.

E-mail address: walach@europa-uni.de (H. Walach).

http://dx.doi.org/10.1016/j.jclinepi.2015.03.027
0895-4356/© 2015 Elsevier Inc. All rights reserved.
comparative studies [10] and by the option to downgrade randomized evidence, if poor, and upgrade nonrandomized, naturalistic evidence, if valid [11], the hierarchical stance itself remains untouched.

Such a view suggests that the hierarchy of internal validity is not only necessary but also sufficient to make clinical decisions. However, because the hierarchy is one of internal validity only, external validity, or generalizability and applicability, is neglected [12]. In fact, as we will argue, external validity and internal validity are incompatible concepts: The more internal validity is emphasized in a study, the lower external validity tends to become (Fig. 1). Although this problem has been recognized by recent developments within the evidence-based medicine movement, it cannot be solved within the current thinking because it neglects the implicit incompatibility relationship between external and internal validity, which forms the basis of our proposal.

In what follows we will argue that internal validity and external validity are, conceptually speaking, incompatible and therefore complementary concepts. They either cannot be maximized at the same time, or the sequence of knowledge produced—first internally valid knowledge from RCTs then externally valid knowledge from observational studies or the other way round—is not indifferent, or technically speaking, does not commute. Theoretically, such concepts cannot be mapped by any theoretical frame that assumes additivity and substitution of concepts. We need theoretical structures to map such relationships that, formally speaking, are non-Abelian. A non-Abelian algebra is a formal structure that is required whenever incompatible concepts have to be theoretically modeled. Unlike our normal, Abelian algebra, where the sequence of operations, for instance when multiplying, is irrelevant, a non-Abelian algebra takes care of situations where the sequencing of operations makes a difference, or, in other words, where concepts are incompatible. Such theoretical structures were introduced by quantum theory to model the strange behavior of nature. Such a model seems also to be necessary to understand the relationship between types of evidence. Before we continue, we hasten to add that we are completely in line with the impulse of enlightenment that evidence-based medicine has brought to medicine through its emphasis of empirical evidence over expert judgment and of research over authority. What we offer is not an argument against evidence-based medicine, but an argument to improve its theoretical foundations, its scope, and its clinical usefulness. To make this clearer, we will show by an example what this might mean in practical terms.

2. The mistaken assumption of linearity, hierarchy, and separability of types and elements of evidence

Assuming linearity, hierarchy, and separability of evidence means we can neglect types of evidence that are “weaker” because whatever can be learned from those “weaker” types of studies will either be part of the evidence that is “stronger” or, if it contradicts the latter, can be ignored because it is less reliable. This reasoning is used by meta-analysts and systematic reviewers and by guideline panels, using only such systematic reviews and RCTs, and thus, this thinking is pervasive in the field of clinical decision making. It infiltrates how research is conducted, how funding is allocated, and what types of studies are published in high-impact journals, and thus, ultimately, it determines what is considered “good scientific practice” by the public and by courts. It is important to realize that this whole edifice itself rests neither on proof, nor on empirical evidence, nor on a sound theory [13,14]. It is simply an assumption, a reasonable one. We show that this assumption is wrong when generalized using three arguments: (1) an argument from validity, (2) an argument from equipoise and agency, and (3) an argument from synergy.

2.1. An argument from validity

The received view rests on the assumption that internal validity is key, and that external validity is somewhat of
lesser importance. However, maximizing internal validity normally happens at the expense of external validity or generalizability [12,15,16] (Fig. 1). If a study is to be maximally internally valid, it has to use most or all the following strategies: It has to randomly allocate patients to groups and must not reveal the randomization code before all data are collected and analyzed. In addition, some measures of masking allocators, patients, outcome assessors, and statisticians are used. Therefore, it is an ethical requirement that patients have given informed consent. We know that many patients do not consent because they refuse randomization [17–21], and those who consent are different in many important respects from patients who do not consent [12,19,20,22,23]. Thus, in principle, it is very difficult to generalize from an RCT to the whole population. But in addition to that principal, theoretical limitation, it is also important to realize a practical limitation of external validity that is true for nearly all RCTs: From an experimental point of view, homogenizing the patient population tested within an RCT is a requirement because it reduces error variance and increases the signal-to-noise ratio. For this reason, regulatory and other trials use long lists of inclusion criteria. Thereby they can be successful with fewer patients—which is an ethical and economical requirement [24,25]. The practical downside is that the result is applicable only to a very small percentage of the full patient population. A typical example are trials of antidepressants that, very often, stipulate that only patients with major depression are treated, and patients with anxiety disorders, personality disorders, substance abuse, and so forth excluded. The results gleaned from such studies should only be applied to the same types of patients as were tested in the trial. However, they rarely are because those patients are rare. Sequenced Treatment Alternatives to Relieve Depression (STAR*D), a large outcomes study in realistic patient groups, found that 1-year remission rates are maximally 43% [26], more likely 38%, approaching what can be expected from a placebo effect [27], well below the usefulness expected from theory and trials [28]. Increasing internal validity will normally come at the cost of reduced external validity. Some exclusion criteria in some studies are posited because researchers want to homogenize the population and decrease error variance. But, other criteria are stipulated for ethical or insurance reasons. In a theoretically ideal trial with no exclusion criterion and only one inclusion criterion, a clear definition of the disease-treated, effect size estimates would drop because a lot of error variance would be introduced. Thus, maximizing internal validity and external validity at the same time would likely jeopardize the usefulness of the research in general. The problem is that we have no theoretical model about the relationship of these two types of validity.

Thus, internal validity and external validity are incompatible from at least two perspectives: Principally because, by definition, results from RCTs do not apply to all those patients unwilling to be randomized, and practically, because most trials have to introduce exclusion criteria to be feasible, restricting the generalizability of their results. Maximizing and emphasizing internal validity, therefore, normally, comes at the expense of external validity. Indeed, a review of external validity in trials mentions only two examples of 149 references where internal validity was high and external validity was also good [12]. Evidence that is internally valid is not simply something that can replace evidence that is externally valid, as is done when RCT evidence replaces “lower grade evidence.” Internal validity and external validity are, at least partially, incompatible concepts, and yet, they have to be applied conjointly.
2.2. An argument from equipoise and agency

The fact that patients have to consent to be randomized and thus take the chance of receiving a—often unwanted—control treatment means that they are restricted in their agency. If the treatment was freely available and if they had the money to buy it, many patients would likely just do that, and in fact do. RCTs, conceptually speaking, treat patients as passive recipients of interventions. This might be practically negligible for new interventions where patients and doctors are in equipoise, and for pharmacological or surgical interventions, where patients are indeed passive receivers of interventions. However, this fact cannot be neglected for all those interventions where patients either have to invest an effort to change, or where doctors or patients have high a priori preferences [29,30]. If patients have to invest efforts, such as in psychotherapy, meditation, lifestyle, or other complex and behavioral interventions, patients who are seeking such therapies by choice will very likely differ from those who are offered such treatments as a potential option within a trial. Patients with strong preferences are likely to be unwilling to be randomized, and many aborted studies, for instance in complementary medicine, are due to this fact [31]. Thus, by default, randomized studies in such areas can only produce evidence of limited external validity. But, we might also have a problem with model validity here. If an intervention, by definition, needs the active collaboration of patients, then studying such an intervention in an RCT will not produce a valid estimate of potential effects. It might produce estimates of minimally achievable effect sizes with patients that have no clear preference, or introduce other kinds of bias [32]. However, such a result is not representative of those patients the treatment has been designed for initially, and thus, the results also lose their model validity. Model validity refers to the question, whether a study represents a fair trial and models the treatment in a way as it is used in practice. This is a point for discussion around the recent question whether meditation-based treatments are effective for depression and anxiety [33]. As we have pointed out, RCTs produce biased estimates of potential effects [34].

The problem becomes more pertinent when parts of the patient population have such strong preferences that they will not participate in trials. In such a situation, it is not possible to generalize from such results.

This is very often the case with treatments that have already been in use for some time because they are part of traditional healing systems [30]. Observational knowledge about real-world effects already exists, and thus, patients and doctors are often unwilling to perform trials because they “know that it works” anyway. This is a variety of the Philip’s paradox [10]: Often we do not have scientific evidence from well-controlled randomized studies for those interventions that are most effective and have been around for quite some time, and where effects are very obvious, it is impossible to conduct a masked trial [35]. Although this is obvious for some standard procedures in medicine, such as defibrillation or anesthesia, where randomized trials would not be required because of the sheer size of the effect [36], it is also true for many treatments of traditional medicine [37]. In such a case, equipoise is missing, and valid RCTs are near impossible to perform.

Thus, there is a double problem: Not only do we observe an incompatibility between internal and external validity, we also see that the sequence of the types of evidence gleaned is important. In the case that an intervention is new, such as with introducing new drugs, there is equipoise, and RCTs are not a problem. What is a problem is generating evidence that is externally valid. Therefore, in such situations, RCTs normally come first and large observational cohort studies in clinical practice that determine the acceptability, applicability, real-world effects and side effects of the new intervention follow. In traditional interventions that have been used for quite some time, such as in traditional medicine, or in surgery, followers and providers have certain opinions based on their experience. Observational evidence already exists, and therefore, evidence from RCTs is often either more difficult or sometimes even impossible to garner, and if produced, very often discarded as meaningless by patients and practitioners. Internal validity and external validity are neither in a linear, nor in a sequentially indifferent relationship, nor do they hierarchically implicate each other in the sense that one has to precede the other. Rather, the temporal sequence determines the knowledge structure that is created which in turn reacts back on potential evidence. This means that the sequence in which these types of evidence are generated does make a difference, and what is known cannot be undone. Technically speaking, the types of evidence do not commute. Let us bear this in mind, for this situation calls for a different type of theoretical model than that which has been hitherto applied.

2.3. An argument from synergy, or harnessing the placebo effect

The linear model of additivity of treatment effects used for the interpretation of placebo-controlled RCTs assumes that a treatment effect is a combination of artifacts, natural course of the disease, regression to the mean, nonspecific, psychologically mediated treatment effects, and specific effects. It uses a subtractive rationale: By comparing a true treatment group to a sham group, it factually subtracts the effect of the true treatment from those of the control treatment and only looks at the difference, neglecting everything else [10,38]. This procedure makes a vital assumption that has rarely been tested and if tested has been found to be wrong [39]: Namely that the treatment components are additive and that the nonspecific effects are roughly stable and uncorrelated with treatment effects across trials. The latter assumption has been shown to be wrong: Treatment and placebo effects are highly correlated...
at $r = 0.78$ in a variety of studies with treatment duration of more than 3 months [40]. And the former assumption of stability of nonspecific effects across trials, at least within domains, has never been tested. However, given all we know about placebo effects, it is a highly unreasonable assumption [41]. Placebo effects are dramatic and strong in studies where researchers aim to maximize them [42], although they are comparatively small in trials where researchers aim to minimize noise [43]. They are dependent on the individual meaning patients attribute to an intervention and to the quality of interaction [44] and hence cannot be easily predicted [45,46].

Thus, focusing only on specific efficacy is misleading and leads to what has been termed the efficacy paradox [47,48]. This paradox describes a situation whereby a treatment can be very effective in general terms but ineffectual vis-a-vis its own placebo because the nonspecific effects are very strong while the specific effects are very small. And yet, a different intervention, which might produce small nonspecific effects but very strong specific effects, may be proved efficacious although its general effectiveness might be comparatively small. Such a situation is depicted in Fig. 2.

An example of the first situation can be found in antidepressant medications. Although critics point out that the specific effect over and above placebo of antidepressants is quite small if the publication bias is removed [49], in fact about $d = 0.32$, substantially less than the $d = 0.5$ that is stipulated by National Institute for Health and Care Excellence as the benchmark for antidepressant treatment, the generic effect of current antidepressant treatment, including the placebo effect produced through the powerful myth that has been created around those drugs, can be quite strong [50], reaching more than two standard deviations against baseline [51].

The clear experimental demonstration of the efficacy paradox would require at least a four-armed clinical experiment, which to our knowledge has not been conducted. However, the three German Acupuncture trials (GERAC) come closest to illustrating the reality of the efficacy paradox. In those studies with about 1200 patients in each trial, acupuncture was tested against a sham-acupuncture procedure—shallow needling in nontherapeutic points—and against best conventional practice. The latter contained all types of guideline-derived therapies, including medications, all of which had already been tested against placebos. None of the studies succeeded in showing any superiority of acupuncture against the sham-acupuncture procedure. But in two of the three studies—in osteoarthritis and low back pain—both the real and the sham acupuncture were nearly twice as effective than the best-practice conventional treatment [52,53], whereas in the third, all three arms were roughly equal [54]. Thus, in two studies, a placebo procedure was statistically significant and clinically more effective than a proven active treatment, illustrating the reality of the efficacy paradox.

An example of the second situation of clear, significant specific effects that are, however, clinically and overall speaking relatively small, can be seen in modern anticancer or antidementia drugs, which all have to prove their efficacy, but whose nonspecific placebo component is small because dementia impairs the capacity to form expectations and hope, thus diminishing potentially beneficial effects of nonspecific elements of treatments, and because cancer in its end stage is difficult to influence by psychological means [55–57].

Thus, the efficacy paradox describes, and is due to, a lack of linear additivity of therapeutic components. Very small specific effects can be therapeutically extremely valuable because they not only add, but synergistically boost nonspecific treatment effects. And comparatively strong specific effects can be clinically useless because they fail to activate some kind of internal response.

Note that in all those cases, the sequence of events is very important: Patients need to first believe, through previous studies and their public perception or through a popular myth, that an intervention is effective, before a strong placebo response can be produced. If one were to take away the specific effects and trust the placebo component only, the latter would collapse as well. This shows that these effects are synergistic. Synergistic systems are not additive but have to be described by multiplicative operations. The hallmark of synergy is that all components working together in synergy can achieve things that each by itself and added up will not achieve. An example for a synergistic system is a skilled child rider on a powerful horse. The child can make the horse do things that the horse would not do by itself, unless in danger, and the child will become quicker and can jump higher with the horse.

Thus, any therapeutic system optimizes nonspecific effects and capitalizes on the knowledge or rumor of its
specific efficacy, producing a greater clinical overall efficacy than each and every component would produce on its own. Additivity of components is a formal abstraction that is wrong in practice and dangerous conceptually. It is dangerous because it suggests we can treat evidence about treatment effects as systems of linear components, although they are synergetic and multiplicative, rather than additive.

This analysis has led to the conclusion that internal validity and external validity are partially incompatible, that agency and the lack of equipoise also suggest incompatibility of internal and external validity, and moreover an incompatibility of sequencing or a lack of commutativity. This latter point is also emphasized by the argument from synergy stating that components of therapeutic interventions are not additive. Taken together, these arguments call into question the received view of a linear hierarchy of evidence that can look at internally valid evidence only and neglect the rest. What are the conceptual consequences?

3. Consequences: rethink hierarchies and think quantum theoretically

This implies that methods are not arranged in a hierarchical order but are mutually complementing each other in their strengths and weaknesses. Thus, a summary of empirical evidence has to be, by necessity, a multifaceted, multimethod review as is advocated by realist synthesis [58–62], and other recent models [60,63]. Conceptually, we need to think about the theoretical foundation of methodology in terms of a theoretical model that is following quantum theoretical reasoning. This does not mean to apply quantum physics, which would be silly, but to apply theoretical structures that have been shown to be useful for modeling relationships within the analysis of matter.

3.1. The quantum theoretical structure of the methodological problems of evidence

Normally and implicitly, we model our concepts along classical lines, following classical rules of thinking. Whenever we have additive components, we do that, as in

\[ y = a + b + c + e. \] (1)

This is a formal description of a linear system where four components—for example, specific effect, nonspecific effect, regression, random error—add up to yield the full effect. We have seen that this is likely wrong for the understanding of components of efficacy.

In methodology, we also implicitly assume that we can use Abelian thinking to model the concepts in question. An Abelian algebra is one where the concepts we are dealing with are all compatible, and therefore, it is irrelevant which operation is performed first. A methodologic example of indifference of the sequencing of operations would be if it is irrelevant whether we have first evidence from RCTs and then from observational studies or the other way round.

A formal expression of this situation is

\[ a \times b = b \times a \] (2)

or

\[ (a \times b) - (b \times a) = 0 \] (2a)

As soon as we substitute “2” and “3” for “a” and “b” in Equation (2a), we can immediately see that this is how we normally operate. The terms “a” and “b” are therefore called “compatible” or “commutative” because we can change their sequence and all stays the same.

We have seen in Section 2.1 that external validity and internal validity are partially incompatible concepts, and in Sections 2.2 and 2.3 that sequencing comes into play when dealing with issues of external and internal validity and various traditions of knowledge. It does make a difference for some interventions whether we know first about specific efficacy and only later learn about their effectiveness, or the other way round. The scandal about the lethal side effects of Cox2-inhibitors is a typical example of the sequencing effects of the types of evidence. The legal requirements necessitate that efficacy studies and RCTs come first, and long-term observational trials that teach us about acceptability and side effects only come later. Clearly, this sequence of procedures does not commute: Had we learned about problematic side effects, we would not test for efficacy. Once we have tested for efficacy, data about side effects and acceptability have difficulty in being taken seriously. The whole issue of traditional or well-known classical medical interventions or surgery shows the same situation from the reverse viewpoint: Where effectiveness is commonly assumed, studies about efficacy are either impossible, useless, or too late. Here also, we see a sequence that does not follow the postulate of commutativity expressed in Equation (2) or Equation (2a).

To model such cases where concepts are incompatible and/or sequencing does play a decisive role, we need an algebra of noncommuting operations, or a non-Abelian algebra as the underlying theoretical structure. This is formally written as

\[ a \times b < > b \times a \] (3)

or

\[ (a \times b) - (b \times a) > 0 \] (3a)

If we again insert “2” and “3” for “a” and “b” as in Equation 2a mentioned previously, we can immediately see how strange this situation is and how different from our “normal” theoretical reasoning. Such a formal structure is at the core of the quantum formalism and is a generalized form of the famous uncertainty relation. In other words, if one neglects all things that are typical for physical quantum theory proper and asks the question: What is the
indispensable difference between a classical and a quantum theory? Then exactly this way of dealing with noncommuting or incompatible operations turns out to be the decisive difference [64—66]. It is important to realize that this has nothing to do with quantum physics proper. It is simply a theoretical structure that is necessary whenever we deal with incompatible operations. Quantum physics turned out to be the first discipline to have noted this. It is therefore at least reasonable to assume that such a structure might also be pertinent in other areas. And as we have seen, it is indeed the case for problems of validity, additivity of effects, or sequencing of types of evidence. It has been pointed out that such structures might be important in more areas of human inquiry [67,68]. For instance, cognitive science researchers have discovered that using quantum formalisms to model the results they find in cognitive research is more powerful and matches reality more closely than using classical procedures [69]. By the same token, it is not surprising that methodology and the structure of empirical evidence seem to also follow theoretical concepts that can more easily be modeled using quantum theoretical descriptions.

3.2. Abolish hierarchies

An immediate consequence is that hierarchical and linear thinking in methodology should be abolished [47]. There is no such thing as an ever-valid hierarchy of evidence. There is only evidence that has a higher level of internal validity and evidence that may have a higher level of external validity [70]. And one cannot be dispensed of at the expense of the other. Incompatible or complementary concepts, as internal and external validity, have to be used conjointly, although they are mutually incompatible. This is in fact the conceptual definition of incompatibility or complementarity that was introduced by Niels Bohr [71,72]. It is part of the definition of incompatible operations that they cannot be applied at the same time, and that the sequence of their application does make a difference. This is why they are called noncommuting operations. Hierarchies suggest that we can neglect evidence that is of lesser “dignity.” The model advocated here suggests that there is no such thing as evidence of a lesser dignity—apart from evidence that is useless because the methodology used was not applied properly or the implementation or organization of a study was flawed. Provided evidence is produced properly, evidence from cohort studies or case series is also valuable. It is valuable because it is different in type and scope from the evidence produced by RCTs [73,74].

Therefore, systematic reviews should widen their remit and include not only RCT-level evidence but complement and qualify the findings from RCTs by evidence from other types of studies [60,63,75,76]. And it should be considered bad practice for guideline panels and reviewers to dismiss evidence from nonrandomized studies. Various authors have pointed this out already [47,77,78].

One major criticism of such proposals has been that no one knows how to put this into practice. We think that a combination of qualitative and quantitative approaches might be useful here. We will demonstrate this by an example.

3.3. Example: combining realist synthesis and matrix analysis in systematic reviews

Realist synthesis is a method of synthesizing evidence in an informed qualitative—narrative way [59—62]. It produces suggestions for decision makers of the type “For patient A with disease Y under condition Q and in circumstances C do X.” All available evidence is brought together to inform practical decisions. This is very close to what is envisaged here, as the method also turns away from a hierarchical understanding of methodology, albeit in a more pragmatic stance and less formally.

Matrix analysis is an instrument from qualitative research [79]. Here, different outcomes and results from various studies, addressing different questions, are presented in a matrix that allows the researcher to have a clear overview.

Our proposal is to combine these methods. This would respect the inherent incompatibility of methodologic concepts. Research methods are ways to answer different questions. Placebo-controlled RCTs answer the question about specific efficacy. Pragmatic RCTs answer the question of comparative effectiveness. Nonrandomized comparative cohort studies answer the question of differential effectiveness of interventions under real-world conditions for different types of patients. Long-term observational studies answer the question of general effectiveness in clinical practice and in the general population, and document side effects, and so forth. None of these questions is “better” or “more valuable” or “stronger” than the other, nor do the answers to these questions produce “better” or “stronger” evidence. They simply produce different answers and diverse knowledge. All this information together yields a matrix that, in some cases, may be broadened out by fundamental research producing mechanistic knowledge [80], in vitro results, or still others. The research types answering certain questions form the rows of the matrix.

Depending on the fine tuning required, one can produce columns for answers such as positive, negative, or unclear results; effect sizes; or patient populations studied. If required, the matrix can be three dimensional including measures of quality of the studies. Thus, the layout of the matrix will change according to the research question. The matrix is then populated with, for instance, the vote counts of the studies or effect sizes, depending on the question. That way, all research methods are treated equally, no information is dropped, systematic reviewers produce an overall pattern, and readers can then form their own opinion. Similar approaches have been used in applied health systems analysis to model trajectories of practice and patient care [81].
We have applied this method in various systematic reviews that have produced meaningful results which, had we looked at certain types of evidence only, would have remained completely devoid of any meaning [82–85].

In one study, we collated evidence in a systematic review about the potential causative role of metallic mercury for Alzheimer disease [86]. The evidence found can be rearranged in the following matrix (Table 1).

One can see immediately that one row of the matrix is empty: There are no RCTs on the influence of metallic mercury on cognitive decline or cognitive measures, simply because it would have been unethical to conduct such studies. Any systematic review looking only at randomized trial evidence would have concluded that there are no studies; hence, the question cannot be answered. However, looking at various types of evidence, one can see a different picture. It is very clear for basic research, comparatively clear for cohort studies and case studies, less clear for autopsy studies. This demands an explanation. The explanation can be found in a modifying theoretical model, published with the original review, stating that mercury is a selenium scavenger, thus aggravating a low-grade selenium deficit in certain populations which will only become apparent long term by compromising the brain’s redox systems. Thereby, the matrix approach has helped in two ways: It has broadened the evidence base, and it has made obvious where inconsistent evidence has to be explained by a theoretical model.

Matrix analysis is a practical realization of the circular understanding of mutual dependence of methods on knowledge from other types of studies: It leads to an immediate understanding of what types of evidence are still necessary and thus creates the argumentative and conceptual basis for such research.

The matrix approach could be refined by a Bayesian type of reasoning: If one makes explicit—for instance by providing theoretical rationales—the prior probability for the effectiveness of an intervention, one could estimate the number of studies that are needed to shift probability toward virtual certainty. One could also use mechanistic knowledge from animal and in vitro models to construct a prior probability that is then contrasted or enriched with findings from clinical studies of various types.

This proposal has an obvious downside: It would be the end of loads of “quick-and-dirty” systematic reviews that make their authors’ lives easy by neglecting everything except RCTs, thus reducing complexity and cognitive load [87].

4. Conclusion

We advocate a reconsideration of the current thinking about methods and types of evidence and reconsidering the notion of a hierarchy of increasing “levels of evidence.” We have shown that the conceptual foundations of this type of reasoning are weak at best. Although a hierarchy is useful within the domain of internal validity, it is counterproductive for assembling the totality of evidence across domains. Instead, research methods should be seen as complementing each other because the types of validity they produce are incompatible and complementary. This situation calls for an alternative theoretical structure. This view is inspired by the quantum theoretical treatment of incompatible operations. One way of putting this into practice is a matrix-analytical approach to reviewing extant evidence. Here, an overview of all types of evidence produces a full picture of the current situation and will also inform further developments.

Acknowledgments

This article would not have been possible without the understanding that H.W. gleaned from many discussions and other collaborations with Hartmann Römer, who is the theoretical father of Generalized Quantum Theory. The studies that formed the basis of our matrix analysis were supported by the Samueli Institute’s Brain, Mind, and Healing program.

References


Table 1. Example: matrix analysis of the evidence pertaining to the question whether metallic mercury is a causative factor for Alzheimer disease

<table>
<thead>
<tr>
<th>Type of study</th>
<th>Positive</th>
<th>Negative</th>
<th>Undecided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic research</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Human studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural cohorts</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cross sectional</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Randomized studies</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Single cases</td>
<td>18</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Autopsy studies</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Adapted from Mutter et al. (2010) [86].