Original Article

Effects and Interactions

Tamás Rudas

Department of Statistics, Faculty of Social Sciences Eötvös Loránd University, Budapest, Hungary

Abstract. If association and causation are different, as indeed they are, then it is not immediate that their strengths should be measured by the same quantities, as current practice does. This paper gives a few reasons why even effects need to be measured not one, but in several different ways. It is argued, what seems obvious but not reflected upon in current practice, that the research or policy question and the mode of data collection should determine the choice of the measure used. In particular, a measure of effect which avoids Simpson's paradox, the cross sum ratio, is discussed and compared to the cross product ratio. The results illustrate the need for further research to understand how effects should be measured in different situations.

Keywords: cross product ratio, cross sum ratio, directional collapsibility, Simpson's paradox

This paper deals with the problem of measuring effects in the context of categorical data or, more precisely, with comparing effects. The word causal, in spite of its popularity, is not added to effects, because policy decisions, which are in the focus of interest here, are often based on the available evidence, even if the underlying causal structure is not fully understood. Thus, effect in this paper may mean causal or evidential effect (Meek & Glymour, 1994). The paper is going to discuss various setups for making decisions regarding effects of treatments on potential responses.

In fact, the question regarding the measurement of effects is somewhat misleading. Different formulas do not measure effect differently, rather measure different concepts of effect. It will be illustrated in Data and Policy section, that the concept of effect one wishes to use, should depend on a number of aspects, including the kind of policy decision to be made, the nature of the response categories observed, and the way in which the data providing evidence were collected. In particular, measures of association (also called measures or parameters of interaction) do not appear to be automatically well suited to measure effects and their differences. The odds ratio or cross product ratio (CPR), which is the most widely used measure of association (Rudas, 1998), cannot measure an appropriate concept of effect in all cases, in spite of its central role in graphical modeling, Bayesian networks, and many other variants of causal modeling, see, for example, Lauritzen (2001). Instead, the cross sum ratio, CSR, proposed by Rudas (2010), is suggested to be used in certain (but not all) contexts.

The CPR and the CSR section discusses the main properties of CPR and of CSR from the aspects considered in Data and Policy section. The CPR is variation independent from the lower dimensional marginals. It will be argued, that in certain contexts, in particular when the treatments received reflect the choices of individuals, as in an observational study, variation independence from the marginals is not desirable when measuring effect. The CSR is not variation independent from the marginal distributions, and it never commits Simpson's paradox. For an early presentation of the paradox, see Yule (1903). Thus, the concept of effect measured by the CSR is free from the major problem of the concept of effect measured by the CPR. It was proved by Rudas (2010), that subject to mild assumptions, the comparison of effects by the CSR is the only one which avoids the paradox.

The CPR and the CSR section gives further comparisons of the concepts measured by the CPR and the CSR, respectively, with the goal of helping researchers and decision makers to decide which one to use. The no association/no effect situation is different in the two cases, and they optimize different loss functions, when used to choose the better treatment.

The paper is concluded with a brief discussion in the section Discussion. The current paper is different from many contributions, which identify a problem and offer a solution to it. Instead, this paper wants to convince the reader, that the widespread use of the odds ratio to measure effect and interaction alike, is problematic and suggests a few aspects, which each may call for different measures of association or effect to be developed. The CSR is not suggested as a universal alternative, but only one, which may be appropriate in certain situations. Indeed, the main message of the paper is that such a universally applicable measure of association or effect does not exist. The results presented here, rather call for more work in understanding what concepts of effect may be relevant in different setups, than provide an answer to this question. In particular, the results imply that in many situations, Simpson's paradox in decisions may not be avoided.

DOI: 10.1027/1614-2241/a000103

Data and Policy

In their classic paper, Goodman and Kruskal (1954) wrote the following:

> "The aim of this paper has been to argue that measures of association should not be taken blindly from the handiest statistics textbooks, but rather should be carefully constructed in a manner appropriate to the problem at hand.... This methodologically neutral position should not be carried on to an extreme."

In the same spirit, in this section, aspects of decision making, which may influence the choice of the concept of effect to be used, are discussed. The general setup is that data are available in the form of a 2 × 2, Treatment by Response table,

$T =$

	Rp1	Rp2
Tr1	*a*	*b*
Tr2	*c*	*d*

and one wishes to choose the better treatment based on the measurement and comparison of the respective effects. Thus, measurement or comparison of effects in this paper refers briefly to a comparison of the differences in response patterns associated with the two treatments. Obviously, the two treatment options compared may also be a treatment and no treatment. Out of the two responses $Rp1$ and $Rp2$, one is preferable to the other and a comparison of the respective frequencies of preferable and less preferable outcomes is supposed to be the basis of the decision.

Distinct groups of such problems may be defined based on the following aspects:

(i) Whether, as the result of the policy decision, the selected treatment is imposed on or is only made available to the members of the population where the data came from;

(ii) Whether the data available came from a designed experiment or from an observational study;

(iii) Whether the responses are good versus neutral or neutral versus bad or good versus bad, and whether the bad outcome is as bad that it is to be avoided at any cost.

There are several other characteristics of problems when measurement and comparison of effects are relevant, but the ones above seem most important to be taken into account when the concept of effect is selected.

Instances of policy decisions, when a treatment is imposed, include treatment protocols for various medical conditions, often called medical guidelines. One example is the opioid treatment protocol in Ireland, see Farrell (2010). Other policy decisions do not make a certain treatment compulsory, rather make treatment options available. For example, a national health care system may decide to allow one or more smoking cessation aids, like e-cigarettes or nicotine replacement therapy, in the form of a patch or gum. See, for example, Brown, Beard, Kotz, Michie, and West (2014). In this case, it is not compulsory for a smoker to choose any of these smoking cessation aids, while in the first example, individuals classified into a certain patient group will, at least in theory, undergo the specified treatment. When choice is available, individuals may take into account a number of factors, which include cost, convenience, expected efficacy, lack, or presence of side effects. These two situations require different assessments of treatment effect.

In the first case, the average effects of the treatments on the members of the population are to be assessed and compared to choose the better treatment. In the second case, for both treatments, two quantities are of interest: what fraction of the affected population would choose it, and what would be its average effect on those who choose it. This difference points to the crucial role played by the mode of data collection in measuring the relevant effect. Data collection procedures, which assign individuals to the different treatments, are inappropriate to estimate the effect which is relevant in the case when the decision to be made is only to make a treatment available (as opposed to making it compulsory).

With a bit of simplification, one may conclude, that data collected from a designed experiment may be appropriate to estimate the effects the treatments would have, if applied to the entire population. Therefore, data from a designed experiment may be used to measure the effect of compulsory treatments and may be used to choose the better one. It has to be added, however, that the external validity of experiments involving human beings can often be questioned, limiting the appropriateness of the data collected for generalization to the entire population. It is data from observational studies, which give information about both the popularity and efficacy of the treatments, thus providing the relevant information when the better treatment to be made available is to be selected.

From this perspective, data collected via an observational study are not less valuable than data collected through a designed experiment, rather may be used to measure effect in a different context. Confounding, which is likely to be present in observational studies and prevents estimating the effect of applying the treatments to the entire population, becomes useful, when the task is to estimate the effect of making the treatment options available, as long as the observational study is based on a proper sample, and generalization to the entire population is possible. In the case of an observational study, the allocation into treatment categories, that is the fraction of individuals in the sample choosing one treatment or the other, is informative (Rudas, 2010), while in a designed experiment, this allocation is determined by the experimenter, and carries no information. When the presence of the other option distorts the fraction of those choosing a particular treatment, it is still reasonable to assume that if the less popular option is not available, the fraction of those choosing the more popular treatment will not be smaller.

There seem to be no straightforward methods to assess the magnitude of the lack of external validity of an experiment and the total survey error including the sampling and nonsampling errors. They reduce the appropriateness of the experimental or survey data to be generalized to the entire population. Depending on the methods applied and on the inhomogeneity of the population from the perspective of the effect or association studied, both may be substantial. While experiments usually produce more precise measurements or observations than surveys do, they are also more invasive, thus often there is no random selection of the individuals to participate, samples are often of the convenience type or self-selected. Overall, the analyst or the policy maker cannot trust data collected through one type of procedure better than data collected through the other type.

The implication is, that if the question is which treatment should be made compulsory, one needs data from a designed experiment, and the concept of effect to be measured and compared should not depend on allocation in the treatment categories. On the other hand, if the question is which treatment should be made available, one needs data from an observational study, and the concept of effect to be measured and compared should depend on allocation into treatment categories. To illustrate this, assume that a survey of 1,000 individuals returns the data in T_1 or the data in T_2. In both data sets, the odds of obtaining a preferable response to a less preferable one is 9:1 with Tr1 and is 3:2 with Tr2, but T_2 shows Tr1 so unpopular, that it may be a better decision to make Tr2 available. The decision does depend on the allocation in this case.

$T_1 =$

	Rp1	Rp2
Tr1	360	40
Tr2	360	240

$T_2 =$

	Rp1	Rp2
Tr1	9	1
Tr2	594	396

To further emphasize the point that in certain situations the relevant effect to be estimated from the Treatment by Response table should depend on the marginal distributions (and in other situations it should not depend on them), imagine a setup, when Tr1 was selected by 990 people, and Tr2 was selected by 10 people in the data available, and Rp1 (the preferable response) was observed 990 times, while Rp2 (the less preferable response) was observed 10 times. Which one would be the treatment of choice, based on these data? Would the decision maker be happy to conclude that about 99% of the population prefers Tr1 over Tr2 and the 99% preferable response shows that a huge majority of those choosing Tr1 give the preferable response? Would the decision maker, having seen these distributions, be ready to choose Tr1 to be made available, without having even seen the joint distribution of Treatment and Response, or would she worry that the joint distribution may look like the following?

$T_3 =$

	Rp1	Rp2
Tr1	980	10
Tr2	10	0

And even if the decision maker would consider the possibility that the data may look like above, would such a data set prompt her to decide to make Tr2 available, and not Tr1?

This question brings us to the third aspect mentioned above. This author is convinced, that if the responses are good versus neutral, then with the above data, it is the best option to make Tr1 available, even if only the marginal distributions are known, irrespective of the joint distribution. If Rp2 is bad, then the preferable concept of effect to be used should depend on how bad this response is. If it is catastrophic so that one wants to avoid it, then it does make a difference, whether the data are as in *T*3, or as in *T*4, and in this case a decision based on the marginals only, is not advisable.

$T_4 =$

	Rp1	Rp2
Tr1	990	0
Tr2	0	10

If Rp2 is to be avoided at any cost, then in the case of T_3, Tr2 is the right choice, in the case of T_4, Tr1 is the right choice, in spite of the two tables having the same marginal distributions, thus one needs to know the joint distribution to make the optimal decision. When the responses are good versus bad, the policy maker has to weigh these outcomes against each other and use this assessment in the decision.

To summarize, different decision making problems require using different concepts and measures of effect, and the correct choice needs to rely, at least, on the aspects (i), (ii), (iii) listed above. To choose the better treatment, experimental data are not necessarily more useful than observational data, even if the latter may be subject to confounding, rather are relevant for different decision making problems. In certain situations, the marginal distributions of the Treatment by Response table carry enough information to choose the better treatment, while this is not true in other situations. The measure of effect should not be variation independent from the marginal distributions of the Treatment by Response table, if the allocation is informative, which is desirable when the policy decision is to make one treatment available.

The CPR and the CSR

It would be nearly impossible to give a complete overview of the different measures (rather: concepts) of association and effect proposed in the literature. Goodman and Kruskal (1954) gave a review of the early history of the topic. Publications relevant for the argument in this paper include Bahadur (1961), Darroch (1974), Lancaster (1969), Rudas (1998, 2010), Rudas and Bergsma (2004). In current practice, the odds ratio (or cross product ratio, CPR) defined as

$$\mathrm{CPR} = \frac{ad}{bc}$$

is the most widely used measure of association and of effect. This is true across the different fields, including the social and behavioral sciences, economics and the health sciences, where associations and effects are investigated, although in the medical field the closely related risk ratio or relative risk seems to be used equally often. The hegemony of the CPR is related to its central role in a number of related techniques and models of statistical analysis, including log-linear and graphical modeling, Markov modeling, Bayesian nets, logistic regression, and survival analysis. Some of these techniques try to reveal associations and some try to measure effect, and, based on the argument in the previous section, it is not straightforward whether the same quantity, namely the CPR, may serve both goals equally well.

Indeed, a celebrated property of the CPR is its variation independence from the marginal distributions, see, for example, Rudas (1998) which, in the current context implies, that the value of the odds ratio carries no information with respect to the marginal distributions. As it was argued in the previous section, this may be problematic, when allocation in treatment categories is informative and the policy decision to be made is to make one of the treatments compared in the data, available to the population. For example,

$$\mathrm{CPR}(T_1) = \mathrm{CPR}(T_2) = 6,$$

indicating that Tr1 is better in both data sets, although in the case of T_2, one may well want to consider Tr2 better than Tr1. The CPR is not sensitive to how popular the two treatments are.

A way of conceptualizing and measuring effects, which is sensitive to allocation, was proposed in Rudas (2010). The cross sum ratio (CSR) is defined as

$$\mathrm{CSR} = \frac{a+d}{b+c},$$

and considers as evidence for Tr1 the positive responses to Tr1 (*a*) and the negative responses for Tr2 (*d*), while it considers in favor of Tr2 the positive responses to Tr2 (*c*) and the negative responses for Tr1 (*b*), and compares these amounts to decide whether the effect of Tr1 or that of Tr2 seems better. The CSR will deem Tr1 better than Tr2, if and only if

$$a - b > c - d,$$

that is, when CSR > 1. Otherwise, if CSR < 1, Tr2 is considered to have a better effect, and in case of equality, the two treatments are equally good. Obviously, such a way of comparing effects only seems useful, when the balance of positive and negative responses may be represented by the difference in the number of times they were observed. If the response Rp2 is catastrophic, and should be avoided, then the balances $a - b$ and $c - d$ should not be used. It has to be pointed out, however, that the same restriction applies to the application of the CPR. The CPR deems Tr1 better than Tr2, if and only if

$$a/b > c/d,$$

and if Tr2 is catastrophic and needs to be avoided, than the balances represented by *a/b* and *c/d* should not be used, either.

The respective meanings and suggested uses of the CPR and of the CSR will be further discussed in the next section, but before that, some of their important properties will be described.

The CSR is not variation independent from the marginal distributions, including allocation into treatment categories. For example,

$$\mathrm{CSR}(T_1) = \frac{600}{400}, \text{ and } \mathrm{CSR}(T_2) = \frac{405}{595},$$

so when T_1 is observed, Tr1 seems preferable, and for T_2, Tr2 seems preferable, and, as it was discussed in the previous section, when allocation into treatment categories is informative and the decision is to make the better treatment available, this is the right choice. This is in contrast with the decision suggested by the CPR, namely that Tr1 is better in the case of both data sets.

The allocation in treatment categories, and the distribution of the responses have such a strong influence on the possible values of the CSR, that if out of 1,000 individuals, 990 choose Tr1, and 990 preferable responses are observed then, irrespective of the joint distribution of Treatment and Response, the CSR always concludes that Tr1 is the preferable treatment, as intuition suggested in the previous section. In fact, the following holds:

Proposition 1. Suppose that a Treatment by Response table T cross-classifies N individuals, out of whom t_1 selected Tr1 and $t_2 = N - t_1$ selected Tr2, and that the preferable response Rp1 was observed r_1 times, while the inferior response Rp2 was observed $r_2 = N - r_1$ times. In this case,

$$\text{If } (t_1 - t_2) + (r_1 - r_2) > N, \quad \text{then } \mathrm{CSR}(T) > 1.$$

Proof. To see this, note that

$$(t_1 - t_2) + (r_1 - r_2) = a + b - (c + d) \\ + a + c - (b + d) = 2a - 2d,$$

so the condition means that $a > N/2$, thus $b + c < N/2$ and CSR = $(a + d)/(b + c) > 1$.

More important is, however that, in addition to taking allocation into account, decisions based on the CSR never commit the so-called Simpson paradox. The discussion of Simpson's paradox has a long history, starting with Yule (1903). The paradox is related to the behavior of the CPR upon conditioning the Treatment by Response table on a Condition variable, or collapsing (marginalizing) a Condition by Treatment by Response table on the first variable. The fact, which is seen as paradoxical by many analysts, is that upon conditioning or marginalization the direction of association between Treatment and Response may change. In other words, the CPR is not directionally collapsible. To illustrate this, let the following data be available in the two categories of some conditioning variable

T (Cond = 1) =

	Rp1	Rp2
Tr1	1	8
Tr2	11	23

T (Cond = 2) =

	Rp1	Rp2
Tr1	20	21
Tr2	3	2

The CPR is 0.26 and 0.63 in the conditional tables, respectively, indicating that Tr1 is associated with Rp2 and Tr2 is associated with Rp1, under both conditions. If one marginalizes over the condition, one obtains

T (marg) =

	Rp1	Rp2
Tr1	21	29
Tr2	14	25

and the value of the CPR becomes 1.29, indicating that the direction of association changed. There are many published examples that such a reversal may occur in reality, see, for example, Wagner (1982). Even more disturbing is this reversal, if one believes/knows/assumes that the Treatment has effect on the Response. Many such real examples have been published, mostly in the setup of meta-analysis, see, for example, Hanley and Theriault (2004). In these cases, exposure to a potential risk factor seems to increase the odds of a disease developing in a number of unrelated studies, but when the data are combined, the "risk factor" appears to have a protective effect. This is, indeed, very disturbing. One way to get around this problem, often used in the medical literature, is to use, for meta-analysis, the Mantel-Haenszel odds ratio. This will give, when all conditional tables suggest the same direction of effect, the same for the combined table, irrespective of what direction is seen there. Although the application of the Mantel-Haenszel odds ratio will, thus, never report a paradoxical finding, it avoids the problem, rather than offering a solution to it. The main issue with applying the M-H procedure is that a different measure of effect is used for the marginal table than for the conditional ones, and in case of repeated conditioning or of repeated marginalization, these roles cannot be distinguished. This criticism is not going to be pursued in this paper.

The vast majority of the existing literature on Simpson's paradox describes the conditions under which the reversal may occur, and offers an "explanation" for the paradox in terms of deficiencies of the data. Rudas (2010) considered the fact that Simpson's paradox may occur with some sets of data, as a property of the CPR, and as an indication that the concept of effect measured by it may not be the right one. Some authors, mostly in the epidemiological context, claim that the risk ratio avoids Simpson's paradox. These authors refer to a different phenomenon under this name. The risk ratio does not avoid Simpson's paradox, as it is understood here. For the three tables above, the risk ratios are, respectively,

$$\mathrm{RR}(\mathrm{Cond}=1)=\frac{1/9}{11/34}=0.34$$

$$\mathrm{RR}(\mathrm{Cond}=2)=\frac{20/41}{3/5}=0.81$$

$$\mathrm{RR}(\mathrm{marg})=\frac{21/50}{14/39}=1.17,$$

illustrating that the reversal may occur with the risk ratio, as well.

A very attractive property of the kind of effects which are compared by the CSR, is that Simpson's paradox cannot occur, that is, the CSR is directionally collapsible. Rudas (2010) illustrated the application of the CSR to several real data sets, where published analyses based on the CPR had been controversial for leading to Simpson's paradox. It was argued in that paper that in all those cases the allocation was informative, and when a measure taking that information into account, the CSR, was used, the paradoxical conclusion disappeared.

Proposition 2. If in both of two Treatment by Response tables Tr1 is deemed better (worse) than Tr2 by the CSR, than in the sum of these tables, Tr1 will also be deemed better (worse) then Tr2 by the CSR.

Proof. If the two tables are

	Rp1	Rp2
Tr1	a_1	b_1
Tr2	c_1	d_1

and

	Rp1	Rp2
Tr1	a_2	b_2
Tr2	c_2	d_2

then their sum is

	Rp1	Rp2
Tr1	$a_1 + a_2$	$b_1 + b_2$
Tr2	$c_1 + c_2$	$d_1 + d_2$

So if

$$a_1+d_1>(<)b_1+c_1, and, a_2+d_2>(<)b_2+c_2,$$

then also

$$a_1+a_2+d_1+d_2>(<)b_1+b_2+c_1+c_2.$$

With the data in the previous example,

$$\mathrm{CSR}(\mathrm{Cond}=1)=\frac{1+23}{11+8}=1.26$$

$$\mathrm{CSR}(\mathrm{Cond}=2)=\frac{20+21}{3+21}=1.33$$

$$\mathrm{CSR}(\mathrm{marg})=\frac{21+23}{14+29}=1.02,$$

and no reversal occurs.

While the result that the CSR never commits Simpson's paradox was very straightforward, the proof of the following result from Rudas (2010) is much more involved. It is formulated here with the condition given in Rudas (2015). It essentially states, that from among a large class of comparisons of effects, decisions, which never commit the paradox, always deem the same treatment to have the better effect, as the CSR does. More precisely, assume that only comparisons of effects are considered that provide a numerical value, and if this is positive, then Tr1 is better than Tr2, if it is negative, then Tr2 is better than Tr1, and the treatments are equally good, if the value is zero. To define the class of comparisons, where the result holds, consider the following properties:

(a) If in the Treatment by Response table, $a = b = c = d$, then no treatment is better than the other;
(b) The value of the comparison of Tr1 to Tr2 is a monotone increasing function of a;
(c) If one treatment is better in the Treatment by Response table and the rows or the columns of the table are swapped, then the other treatment is deemed better.

These properties seem like fairly straightforward requirements. The meanings of (a) and (c) are obvious, and (b) means that if two Treatment by Response tables are identical, except that in one of them there are more positive responses to Tr1 than in the other one, then the first table provides more evidence in favor of Tr1 than the second table. Then, one has the following result:

Proposition 3. Suppose one wishes to compare the effects of two treatments, so that the comparison observes the rules (a), (b), (c) above. If the comparison is such that it never commits Simpson's paradox, then it always gives the same result as if the CSR was used.

Rudas (2015) proved a multivariate version of the above result.

In summary, not only is the CSR sensitive to allocation in treatment categories, which is a desirable property when this allocation is informative, but it also avoids Simpson's paradox and, further, represents all such comparisons from among those which observe the rules (a), (b), and (c).

What Do the CPR and CSR Measure?

Having seen several advantageous properties of the CSR relative to the CPR when measuring effect, in particular when allocation into treatment categories is informative, but also keeping in mind the good properties of the CPR when it comes to measuring association, this section takes a closer look at the question of what CSR and CPR measure and how they measure it.

The no-association/no-difference-in-effects situations are quite different under the CPR and the CSR. As it is easy to see, the CPR is equal to 1, that is, CPR detects no association or no difference in effects if and only if

$$d = \frac{bc}{a},$$

while the CSR finds no difference in effects, that is, its value is 0 if and only if

$$d = b + c - a.$$

It is instructive to compare the CSR to the logarithm of the CPR, which reports no association when its value is 0, and for which (a), (b), (c) above also hold. For the logarithm of the CPR, no association is found if and only if

$$\log(d) = \log(b) + \log(c) - \log(a),$$

thus, one may say that essentially, the CPR does with the logarithms of the frequencies what the CSR does to the frequencies themselves. Although the logarithm is a monotone function, so one may not expect a dramatic effect, the difference between the two conclusions may be substantial, for example, the lack or presence of Simpson's paradox.

One may also ask, whether the two approaches may be reconciled in the sense of the existence of a measure of (difference in) effects, which, on the one hand, never commits Simpson's paradox, and, on the other hand, reports no difference in effect for an independent table. The answer is negative, as long as properties (a), (b), and (c) are assumed, in the sense that this can only happen for tables with identical rows or with identical columns.

Proposition 4. Consider a measure of effect M, for which (a), (b), (c) hold, and which never commits Simpson's paradox. Also consider all frequency distributions, which are independent, that is, have the following structure

$$T = \begin{array}{|c|c|c|} \hline & \text{Rp1} & \text{Rp2} \\ \hline \text{Tr1} & a & \alpha a \\ \hline \text{Tr2} & \beta a & \alpha\beta a \\ \hline \end{array}$$

for positive a, α, and β. Then, M reports no effect in T, if and only if $\alpha = 1$ or $\beta = 1$.

Proof. By Proposition 3, M finds no effect, if and only if, CSR finds no effect, which happens if and only if

$$a + \alpha\beta a = \alpha a + \beta a,$$

that is, if and only if

$$1 + \alpha\beta = \alpha + \beta.$$

This condition can be written as

$$1 - \alpha = \beta(1 - \alpha),$$

and this holds if and only if $\alpha = 1$ or $\beta = 1$.

From the perspective of the policy maker, the choice needs to be made not only between two possible treatments, but rather between different measures to choose one. When it comes to choosing from among the CPR or the CSR, in addition to their different properties relative to the characteristics of the data collection procedure leading to the data upon which the selection of the treatment has to be based, also the concept of loss minimized by choosing the treatment suggested by the selected measure is relevant. This is related to the kind of policy decision to be made: whether the selected treatment will be made compulsory or will be made available only.

In the former case, if the data available, based on the way they were collected, make it reasonable to believe that every member of the population would react to the respective treatments as those who have actually received them did, the treatment selected based on the CPR maximizes the ratio of positive to negative responses out of the two treatments compared. If the negative response is to be avoided, this ratio is not a good measure of success. Unfortunately, the decision maker runs the risk of Simpson's paradox, which may make the decision catastrophic, or, at least, will make the decision questionable. With such a data collection procedure, the allocation into treatment categories cannot be informative (otherwise the generalization to the responses of the entire population would not be justified), and the application of the CSR would bias the comparison by being dependent on allocation, and thus is not recommended.

In the latter case, if the available data, based on the way they were collected, make it reasonable to believe, that either treatment would be selected by the same fraction of the population as of those who were observed, then the selection based on the CSR maximizes the difference between the numbers of positive and negative responses. For this to be feasible and desirable, the choices of a treatment need not be influenced by the lack or presence of the other option, and if this cannot be assumed, more complex data collection procedures may be needed. Further, the difference in the frequencies of positive and negative responses needs to be an acceptable measure of the success of the treatment (which may not be true, if one response has to be avoided because of being very bad). In exchange, Simpson's paradox will never be committed. If, in such a situation, the CPR was used, then the selection of the treatment would disregard the effect of the popularity of the treatments and may end up choosing an option, which would only be selected by very few and, thus, one with very small overall effect.

If the negative outcome should be avoided and this is seen as the most important goal of the decision making, then simply $b - d$ may be used, illustrating again, that the best measure does depend on several aspects of the decision making problem and the CPR and the CSR investigated here do not represent all possible measures that may have to be used.

Discussion

This author sees the results presented in the paper largely negative or, at least, calling for significant further work.

The foregoing considerations show, that when one wishes to choose the better treatment out of two options based on data, a number of aspects need to be taken into account, and for several realistic decision making tasks, there exists no procedure with any optimality property, at least, as of now. It should be clear, however, that in sharp contrast to current practice, the CPR cannot be automatically used to choose the better treatment, because it disregards the potentially different levels of popularity or acceptability of the treatments, and may also commit Simpson's paradox. Use of the CSR is not a general solution, either, because it depends heavily on allocation, and this is not always desirable. However, when applicable, the CSR avoids Simpson's paradox.

It has also been demonstrated, that when it comes to choosing the better treatment based on data, it is not sufficient to take into account the traditionally existing dichotomy between data arising from a designed experiment or from an observational study. To find the right method to choose the better treatment, further aspects, including the kind of policy decision to be made, and the relative meanings of the positive and negative responses need to be taken into account.

This underlines again, that analyzing effects, including causal effects, based on observational data does have an intrinsic limitation, as long as it is based on graphical Markov models, as is the case in a large body of current literature, in particular using Bayesian networks. If the CPR cannot be seen as appropriate to choose the treatment with the better effect, then it is even less appropriate to be used as the basis of complex causal modeling.

Further, if under mild assumptions, Simpson's paradox may only be avoided by using the CSR, and if CSR is only appropriate under very specific conditions, then in many decision making situations, the paradox cannot be avoided. Perhaps, one may want to conclude, that "which one is the better treatment" is such a very loosely formulated question, that no logically consistent answer to it exists, and decision makers need to formulate more specific questions, and for each of these, the optimal decision making procedures need to be found.

Acknowledgments

This research was supported in part by Grant K-106154 from the Hungarian National Scientific Research Fund (OTKA). An earlier version of this material was presented as a plenary talk at the 6th European Conference of Methodology in Utrecht. The author thanks Prasanta Bandyopadhyay, Antonio Forcina, Jacques Hagenaars, Anna Klimova, Renáta Németh, Donald Rubin, and two anonymous reviewers for several helpful comments and questions.

References

Bahadur, R. (1961). A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 158–168). Stanford, CA: Stanford University Press.

Brown, J, Beard, E., Kotz, D., Michie, S., & West, R. (2014). Real-world effectiveness of e-cigarettes when used to aid smoking cessation: A cross sectional population study. *Addiction, 109*, 1531–1540. doi: 10.1111/add.12623

Darroch, J. N. (1974). Multiplicative and additive interaction in contingency tables. *Biometrika, 61*, 207–214.

Farrell, B. M. (2010). *The introduction of the opioid treatment protocol*. Dublin, OH: Health Service Executive.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.

Hanley, J. A., & Theriault, G. (2000). Simpson's paradox in meta-analysis. *Epidemiology, 11*, 613–614.

Lancaster, H. O. (1969). *The chi-squared distribution*. London, UK: Wiley.

Lauritzen, S. L. (2001). Causal inference from graphical models. In O. E. Barndorff-Nielsen, D. R. Cox, & C. Klüppelberg (Eds.), *Complex stochastic systems* (pp. 63–107). London, UK: Chapman and Hall.

Meek, C., & Glymour, C. (1994). Conditioning and intervening. *The British Journal for the Philosophy of Science, 45*, 1001–1021.

Rudas, T. (1998). *Odds ratios in the analysis of contingency tables*. Newbury Park, CA: Sage.

Rudas, T. (2010). Informative allocation and consistent treatment selection. *Statistical Methodology, Special Issue on Statistics in the Social Sciences, 7*, 323–337.

Rudas, T. (2015). Directionally collapsible parameterizations of multivariate binary distributions. *Statistical Methodology, 27*, 132–145.

Rudas, T., & Bergsma, W. P. (2004). Reconsidering the Odds Ratio as a measure of 2 × 2 association in a population. *Statistics in Medicine, 23*, 3545–3547.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician, 36*, 46–48.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2*, 121–134.

Published online December 23, 2015

Tamás Rudas

Department of Statistics
Faculty of Social Sciences
Eötvös Loránd University, Budapest
Pazmany Peter setany 1/A
1117 Budapest
Hungary
Tel. +36 1 372 2500/6871
E-mail rudas@tarki.hu