Randomized Controlled Trials and the Flow of Information

Sherri Roush
Department of Philosophy
Group in Logic and the Methodology of Science
U.C., Berkeley

Nancy is ultimately most concerned about how to determine the relevance of evidence to implementation of evidence-based policy guidelines, in other words, the *transferability* of study results to a population different from the one that was studied and in which procedures or conditions are not the same as those in the study. And she is concerned about the privileged position Randomized Controlled Trials (RCTs) are given in the ranking schemes for evidence-based policy, because as she sees it RCTs do not address this question while other methods do.

RCTs are highly regarded because of their strength in ruling out confounding variables, but they can be weak on the transferability problem because the manipulations necessary for controlled experiment also guarantee that the setting and population are different from the situation and population targeted for application. However, both of these familiar points are simplifications that can be misleading. Some non-RCT type studies (e.g. soft interventions) can also be very good at ruling out confounding variables.[1] And, as I will explain, the problems leading to the transferability problem – interacting variables and a difference between study and target populations – are present in any study, not unique to RCTs. In addition,

[1] See F. Eberhardt and R. Sheines, "Interventions and Causal Inference" (2007), Philosophy of Science, 74, pp. 981–995.

there are well-known ways of addressing the transferability problem for RCTs, both by methods internal to such studies and by doing other studies and other types of study, as I will discuss.

So, Nancy is right in that the crude evidence ranking system that portrays some methods as always better than others, is definitely not faithful to what statisticians and scientists know. Simplified schemes can make sense practically, though, as tools to help those evaluating large bodies of studies to stay at least in the ballpark of sound assessment of quality. The question about the rankings used in formulation of policy guidelines today may be whether the particular simplification with RCTs univocally on top is causing more distortion than it avoids, in the application of scientific evidence to policy. Any simplified scheme would be a distortion, though, because one method will be superior to another only given a situation and a set of background knowledge. There are general things that can be said about how suitability of a type of study varies with the situation, but they are not simple. So, the task seems to be to determine which distortions are least damaging. I will not try to answer that practical question, but I will suggest some practical measures in guideline formulation that could supplement the methods scientists already have for addressing the transferability problem. My suggestions have to do with more efficient flow of information from actual applications back to the research community for use in defining the next round of studies.

A couple of preliminary observations: Nancy is concerned, as we should be, that in the focus on RCT we stop using knowledge that we got in other ways, knowledge that is not RCT-certified. However, the Scottish Guidelines for use of evidence in medicine are impressive, I think, in how

seriously they do take non-RCT studies despite the fact that they are not regarded as the gold standard. There seems also to be a high degree of respect for background knowledge not obtained through scientific studies, since typically every member of the Guidelines Committee who evaluates the scientific evidence is either a practitioner or patient representative, even if he or she is also a researcher. (This seems to be different from the U.S. Dept. of Education practice for committees formulating guidelines, where typically only one member of the team is a teacher and the rest pure researchers.) In the Scottish system, committee chairs are even given explicit instructions for insuring that the pre-existing hierarchy of status among health care professionals does not translate into disproportionate speaking time for high-status members, and suppression of valuable information from others.

These features maximize use of members' conscious knowledge, and also knowledge they might not even be able to articulate. Such background knowledge is brought to bear in the process of guideline formulation both in every participant's evaluation of every study, and, especially, in the "considered judgment" form each fills out at the end of the process. In other words, the Scottish System respects and uses both explicit and tacit background knowledge. This shows remarkable respect for common sense – you have, for example, nurses evaluating scientific papers[2] – and given the nationality, I suppose we should not be surprised. So, it seems to me that the Scottish are not throwing non-RCT evidence into the dustbin, although I have not seen similar features in other guideline formulation protocols.

[2] Evaluation of research studies requires, of course, some sophistication with statistics which pure practitioners and patient representatives may not have. For this reason the Scottish system has an Information Officer who gives tutorials about how to evaluate statistically presented evidence.

Nancy's big question is how to tell whether evidence is relevant to implementation in a population that was not the one studied. When should we think the results of a study are transferable to another population, and why? RCTs are not always inferior to other methods in virtue of having this problem, since any study, using any method, is done on a population which is not the same as the population targeted for treatment. The populations in RCTs can generally be expected to be more different from target populations than the populations of observational studies are, due to the contrivance required for the former. However, the question will always arise how similar the study and target population need to be for the results to transfer, and the study in question will not be able to answer it, no matter what kind of study it is. Other studies and information will always be needed for that question. RCTs are not unlike other studies in that their transferability problem can also be addressed through further studies. I will discuss one such path below.

Another problem for all studies that is relevant to transferability is the task of discerning interacting variables. I will describe how this problem arises for RCTs, but it is universal among methods. To get at the way the problem arises for RCTs I will first describe the advantages of this method. The crudeness of the picture I am about to draw will be evident to those acquainted with these things, but the points I need will, I think, survive this flattening.

In an RCT, you have a treatment (or intervention), T, and you want to know whether it has a significant ability to cause a desired effect (or an undesired side effect), O, in everybody. A group of people will be administered the treatment. Control means there is a group of the same size as the group of people who get the treatment, and the control group doesn't

get the treatment. If you didn't have that group, you wouldn't know whether improvement in the patients, say, was due to the treatment or would have happened anyway.

The following types of possibility remain, and illustrate a need for more than merely having a control group. It could happen that there is a variable, V, positively relevant to producing O, and going into the experiment more people in the control group have it than people in the treatment group. Then T could have the same positive effect in the treatment group as V has in the control group, and if we didn't know about V or didn't take it into account we would conclude that T has no effect on its own, even though it does.

Alternatively, the factor V could happen to be more widespread in the treatment group than in the control group, and we would wrongly end up thinking T has a bigger positive effect than it does. Some of that effect is from T, the rest is from V, but we don't know about V so we attribute all of the effect to T. Or, it is possible that T has no effect, but if V is more widespread in the treatment population than in the control population, then we would end up wrongly thinking T has an effect.

These possibilities are the main reasons why we need a randomization process (or a suitable functional equivalent)[3] for assigning subjects to experimental vs. control groups. Perfect randomization means there are no systematic differences between the treatment and control groups. It is a situation where equal numbers of people in each group have traits like age,

[3] It is a live issue between Bayesians and classical statisticians whether the process of choosing these groups by randomization has any benefit over choosing them by matching, that is, making sure for every (known) relevant trait, there are as many and like subjects in the control group as in the treatment group. Bayesians think the purpose of ruling out the possibilities in question is served equally well without a randomizing process to produce the similar profiles of the two groups. This dispute does not seem to me to make a difference to our questions here. So, when I say "RCT" I mean to be speaking also about Bayesian trials as far as possible.

gender, and health status with respect to particular conditions, etc. The bigger the groups studied the more likely that randomizing on these factors randomizes on all factors, known and unknown. Of course no particular study is ever perfect, but here we know how to take steps to improve the results, namely, use bigger populations, consider more variables, etc. Successful randomization will put equal numbers of subjects with the unknown V in both groups, so that V's effect on O is distributed the same way in each group, and any difference in effect can be attributed to T.

The problem relevant to our concerns here, and that it is hard to know how to resolve, comes even if we assume perfect randomization has been achieved. This means that for every trait, the two populations are the same, or the same on average, with respect to unknown variables like V. The problem is, what if variable V is a causal factor that enhances or is needed to enable T to have the desired effect? Then even though V is present in both treatment and control populations in equal measure, the treatment group will show a *higher* effect on O than T could have produced alone. (We assume that V alone cannot produce the effect without T.) T has the potential to produce an effect in such a case, but without knowing about V we would overestimate its effect. Thus we would conclude that T produces the effect seen in the experiment pretty much regardless of what other properties are present, when the experiment shows at most that under some circumstances T has a causal power.

So, a problem with RCTs is that although by isolating T as the treatment the RCT can show that T has causal potential,[4] that *particular* RCT cannot show that T is a sufficient cause for bringing about the effect

[4] This is so under certain assumptions about the functional form of the causal structure, e.g., linearity, as discussed below.

we see on O. There could be features present in the same distribution in both treatment and control populations (properties of the subjects, the implementation, or the background situation) that had a role, in conjunction with T, in bringing about O, and we are not going to see that in that study that showed a potential causal role for T. Randomization is a strong tool for ruling out confounders, but it does not enable us to see those unknown enabling and enhancing factors in the same experiment, the interaction effects.[5]

This is one clear aspect of the problem involved in using experimental results to draw conclusions about what will work in real-life situations. For example, maybe the study takes place in a state where the drinking water contains unusually high levels of fluoride and that happens to make the drug work better on anyone who has the condition in question. Maybe it is even a necessary condition for the drug to work. In that case both treatment and control group will have an advantage that the target population won't have. Or, it could be that administration of the medicine is a delicate enough affair that it requires skills that practitioners in many other regions won't have. Or maybe longer classes were only able to raise test scores in the study population because good books were used, whereas the extra hours would be useless in other regions of the country where you don't have anything but the same crappy books to read in the added time.

This is one way that the transferability problem arises for RCTs. An RCT on T can provide evidence that T is a causal factor because it isolates T and rules out confounding variables. But that RCT doesn't show T is sufficient for the degree of effect seen. And the randomization within that study can hide the other interacting factors, if there are any.

[5] The same argument can be made, of course, for detracting factors.

However, we shouldn't think that no consideration is given to the transferability problem of a RCT within that very study. In processing the data scientists can check subgroups of their subjects that have properties that have been measured and see if unexpected correlations show up. To check for unknown variables they can look for unexpected clusters in the data. Of course, only variables that are actually present in the experimental population can be found in these ways, but again, non-RCT study populations are also distinct from their target populations. One further way that RCTs can address this issue is to form their experimental population by taking random samples of the target population. There are a lot of tools for addressing the transferability problem, and probably their existence and importance should be stressed to all who carry out and use studies.

Before discussing a positive suggestion, I will explain briefly why although Nancy's language of capacities is a good way to describe what the transferability problem is (in a common special case), and how hard the problem is, I think that changing from the concept of cause to that of capacity will not give us tools to solve this problem. (I am not sure Nancy was claiming that, in any case.) T has a *capacity* to affect outcome O if under all conditions T has the *ability* to affect O. What this means, I take it, is that there is a set of conditions under which T will have some degree of effect on O, in the following way: T in conjunction with each and several of a certain set of other variables will have a range of distinct and distinguishable effects on O. Nancy wants to use this way of carving up the space to take knowledge of capacities and then plug the actual conditions at, say, different hospitals or schools, into the place of "other variables" and see

what the effect is that pops out. This will say whether the evidence applies to their situation or not.

This conception of the situation explains intuitively what the problem of evidence-based implementation is. However, to take it as a solution requires supposing we *have* knowledge of capacities, and this is as hard as the problem was in the first place. Verification of a claim that T has a capacity in this sense looks to me to be roughly the same problem as verifying that T is a causal factor, since the claim looks to me equivalent to saying that there are A, .., F, such that T together with A causes degree x effect on O, and T together with B causes degree y effect on O, and etc. It looks like the only way you could verify that T has the capacity in question would be to verify that T and A together *cause* degree x effect on O, and T and B together *cause* degree y effect on O, etc.

It is true, as just explained, that one RCT will not establish one of those conjuncts, but it can give evidence that *there exist* a range of conditions under which T causes some degree of effect on O. If capacity means that weaker existence claim, then it seems to me that an RCT on T would be one appropriate way to establish a capacity (not the only one). The natural way to establish a specific list of causal claims that constitutes the stronger claim of capacity would in this approach be to do an RCT on a new treatment, T', which is composed of T in conjunction with another factor, A, and so on with all permutations of T with other factors. The question we need to answer for a capacities claim is the same as many instances of the one we need to answer for the causal formulation: What degree of effect does T' (T plus A) have on O? and so on for T'', T''', etc. In each step there will be those unknown variables discussed above with RCTs, but after the

trial on T' that is a set now reduced by one, A. One RCT cannot solve its own transferability problem, but further RCTs can make headway on it.

However, there is an important limitation on this picture of causal structure as a matter of capacities. As is clear from the capacity claim as a conjunction of causal factors, and from the analogy of the force diagram from physics, this picture assumes that the causes combine in a linear fashion. This is often the case, and it is often not the case; it could instead, for example, have a threshold structure, where one factor kicks in only if another goes above a certain value. In order to draw sound conclusions from RCTs or other studies, even about T being a causal factor in an outcome, we must make some assumption about the functional form of the combined causes in such a system. If we have reason to assume linearity -- and we may get such reason from studies that are not RCTs, -- then the studies will tell us more. If not, things are harder.

We should, I think, resist the expectation that there will be a single time when the transferability problem for a study result is solved. But we can do better and worse depending on how much information is taken into account, and how wisely. Over time, with more information, we should be able to improve the soundness of our conclusions, and it is worth making explicit how that process goes, in order to see how it might be made more efficient. To understand what I will propose, consider an example, again involving RCTs: Suppose the treatment, T, is birth control pills (BCP), and the study population contains only women randomized between treatment and control groups over race, age, diet, smoking, and exercise. Suppose we find that the treatment group has an 11% higher chance of blood clot than the control group. What are we allowed to conclude for *all* women in the age range studied?

Suppose randomization was full over every trait, not just those mentioned, but there is a factor we don't know about, call it FVL, equally present in treatment and control groups, that is (suppose an extreme case) necessary for BCP to cause blood clots and the two factors are mutually enhancing. There are two joint causes of the blood clots in the treatment group. In this case, the result 11% is definitely an overstatement of the blood clot risk for those women without FVL, and an understatement for those with FVL. Going with this for a policy warning to women might needlessly convince many women not to use BCP, and insufficiently alarm those in real danger.

This kind of case happens all the time, and the recipe for imagining them is clear. Another such case: C and T are both populations with all, or average same per cent, unknown or unconsidered trait that is relevant to outcome. Suppose the T is longer school hours, GB is good books, and the outcome is higher test scores. Suppose you get an effect from T, longer school hours, on test scores, but GB is actually necessary for T to have that level of effect on O. If the two factors are mutually enhancing, then randomizing for GB will hide the fact that GB was necessary for T to produce that degree of effect on O, and so it will hide the fact that this study that has everyone or average with GB, and shows on average an effect from T, is not generalizable to populations without GB. In this case the positive effect of longer hours is definitely overstated for populations without good books, and so for the general population.

To address the transferability problem we can do another study taking T + GB as treatment, with three control groups: T only, GB only, and neither. In the other case, we could do another RCT in which T = BCP + FVL is tested against the three control groups. When you do that, you find

the FVL group with, say, 2% blood clots, the BCP group with, say, 0.5% blood clots, the group with neither factor with 0.1 % blood clots, and the BCP +FVL group at 20%.

→ If you apply BCP to women with no FVL then there's almost no worry about blood clots.

→ If you apply BCP to women with FVL, you are courting disaster, and probably need a compensating treatment, if you prescribe at all.

Guideline: Test women for FVL before prescribing BCP.

Notice that this is not a definitive answer to the question under what specific set of circumstances, will BCP cause a significant rate of blood clots. There will not be a point in time when that is established, since there always may be other unknown "V"s. But no method can offer exhaustion of the possible relevant circumstances. This is not a distinctive problem of RCTs but of the finitude of human resources and time. However, studies over time, each intelligently related to all the previously gleaned information, do yield a process by which we can improve our epistemic situation in each new round of studies relative to the previous. This point is more general than RCTs; the studies you start with could be of any suitable type, and the studies that try to improve on it could surely be of any type that gave relevant information. The point is to commit to a process that uses the results of each round to improve on those of the next.

There is a hard problem about how to engage in ever new rounds of the experimental process: though we know how to do an RCT on BCP +

FVL, how did we, and how are we going to, *discover* factors like FVL and good books as variables worthy of the next RCT? It would take infinite time to discover all of them, but we can do better or worse with the finite time that we have, and try to maximize the efficiency of this process of discovery. The question is How? I think that practical steps could be added to the guideline formulation procedure to encourage particular types of information flow in the system that would encourage more and faster discovery and use of such variables.

There are at least two aspects of the problem of discovering new variables to study. One has to do with identification of fruitful variables, the other with transmission of that information to research scientists, since information about these potential variables will not only be discovered by them. Even if one thinks that RCTs are superior for certifying the causal potency of a factor, it cannot be denied that other types of study bring information as to which variables have a potential role and are plausibly worth further study. There are an infinite number of factors we could possibly test, but if, for example, we see in a quasi-experiment that one variable shows a trend suggesting causation of an effect, it would only be rational to prefer to do RCTs on this factor, taken as treatment, rather than to do such a trial on a randomly chosen variable we have no such information about. Non-RCT studies and background knowledge do well at identifying trends and potential causal factors, the most likely variables out of the infinite sea of factors most of which are irrelevant.

What kind of evidence is there besides RCTs? There is common sense, practitioner experience, patient and patient representative testimony at the open session that the Scottish do every round, narrative evidence, observational studies (cohort studies, comparison studies), quasi-

experiments. Notice that common sense would have told you, once you thought of it, that GB was really plausible as a causal factor. But only once you thought of it. It would be helpful for those evaluating evidence for policy guidelines to be *looking* for further factors to test, and be receptive to information from any relevant source.

There are already several stages in the Scottish guideline process where this kind of information that could be useful for guiding future studies is collected from many sources.

-- Knowledge of committee members

-- Results of systematic literature survey on question (includes non-RCT studies)

-- Peer Review of formulated guidelines

-- National Open Session about formulated guidelines

And though they are expensive, observational studies of actual implementation would provide information about how well the previous round of RCTs transferred, and which variables might be good for future RCTs. The Scottish System might think of seeking funding for this kind of research.

These aspects have a role in collecting information about new possibly relevant variables. But the information also needs to go to the right places. There are some feedback loops already in the Scottish process, for taking that information back to the research community. But it would be useful to have more explicit methods in place for making sure that information has a direct path back to researchers. My proposal would be for more emphasis on those feedback loops.

First, I would suggest that the information officer instruct those evaluating the evidence that and why noticing and feeding back such information is crucial to improving the successful implementation of guidelines in the current step and the next. This connection could be emphasized to reviewers, to encourage them to leave no information about a potential factor behind. Also, at the peer review stage, the reviewers could not only evaluate the soundness of the guidelines relative to existing evidence, but also use their background knowledge to identify new potential variables, and have an efficient route already set up by which that gets fed back to the research community. Those reviewing studies could be encouraged to evaluate the transferability of the results of each study not only singly, but also by what the other current studies suggest on whether they will transfer. Additionally, information about potential new variables may arise in the National Open Meeting; having a process in place whereby this is noted and routed back – perhaps one committee member designated to do this – would bring more efficient use of the information. None of these additions to the process seem very costly, and information of the sort needed seems already to be present, latently, in the Scottish process, so the dividend for adding such procedures would seem to be high.

In summary, it seems to me that the transferability problem for a given RCT can be addressed through more RCTs, and through other types of study. All studies have a transferability problem, and all kinds of studies can potentially be used to address it. Though we cannot expect at any given time, now or in the future, to have achieved the full resolution of a transferability problem, we can do better in each round. The continuous cycle of studies that could over time improve our understanding of the transferability of a given result depends on identifying the next variables to do RCTs or other

studies on, and getting that information back to researchers. We already have latent information about this that probably goes unused or is used too slowly, and there are simple and low-cost procedural feedback loops that could help us use that information more efficiently.