Assessing health system interventions: key points when considering the value of randomization
Mike English, a Joanna Schellenbergb & Jim Todd c

Abstract Research is needed to help identify interventions that will improve the capacity or functioning of health systems and thereby contribute to achieving global health goals. Well conducted, randomized controlled trials (RCTs), insofar as they reduce bias and confounding, provide the strongest evidence for identifying which interventions delivered directly to individuals are safe and effective. When ethically feasible, they can also help reduce bias and confounding when assessing interventions targeting entire health systems. However, additional challenges emerge when research focuses on interventions that target the multiple units of organization found within health systems. Hence, one cannot complacently assume that randomization can reduce or eliminate bias and confounding to the same degree in every instance. While others have articulated arguments in favour of alternative designs, this paper is intended to help people understand why the potential value afforded by RCTs may be threatened. Specifically, it suggests six points to be borne in mind when exploring the challenges entailed in designing or evaluating RCTs on health system interventions: (i) the number of units available for randomization; (ii) the complexity of the organizational unit under study; (iii) the complexity of the intervention; (iv) the complexity of the cause–effect pathway; (v) contamination; and (vi) outcome heterogeneity. The authors suggest that the latter may be informative and that the reasons behind it should be explored and not ignored. Based on improved understanding of the value and possible limitations of RCTs on health system interventions, the authors show why we need broader platforms of research to complement RCTs.

Introduction
Researchers are being urged to provide evidence on how to fix health systems in developing countries.1–3 These exhortations recognize that health systems play a vital role in achieving global goals for maternal, neonatal and child survival and for reducing HIV infection, tuberculosis and malaria. The type of research providing the best evidence on the effectiveness of health system interventions is a matter of controversy, with quantitative and qualitative approaches often pitted against each other, although researchers are increasingly aware of the limitations of randomized studies4–5 and of the value of mixed methods approaches.6–8 Despite this, researchers who are better acquainted with individually randomized controlled trials (RCTs) than with other research designs still place undue reliance on randomization, particularly in health services research. Most health-care researchers understand that randomization eliminates or reduces bias and baseline imbalances between the groups being compared, and that the control group provides the comparison for the intervention under study. Clear reporting guidelines9 have helped establish standards behind it should be explored and not ignored. Based on improved understanding of the value and possible limitations of RCTs on health system interventions, the authors show why we need broader platforms of research to complement RCTs.

Point 1: numbers
As we try to examine larger units of health care delivery, fewer units are available for randomization.

RCTs were designed to randomize large numbers of people into receiving either the intervention being tested or a placebo. However, interventions targeting the health system are delivered not to individuals, but to groups, clinics, facilities or even larger units of organization such as districts. The larger the organizational unit, the fewer the units to be randomized, the larger the geographic area spanned by each unit and the greater the number of stakeholders involved, particularly if the study is of long duration. Feasibility then tends to constrain sample size. Unfortunately, if we recruit the sample and intervene at a given organizational level (a clinic, for example), we also need to randomize and to compare the results at that level (cluster). We can measure effects on clinic users, but these observations take place within a cluster, and within a cluster or clinic there are likely to be similarities in how people behave or are treated, thus the observations made within a clinic are not entirely independent but may be influenced to a greater or lesser degree by characteristics of the clinic (a point often overlooked).10 Consequently, it

907
may not be helpful to perform a large number of observations within a clinic (or cluster), as additional within-cluster recruitment typically yields diminishing returns.11 Because the cluster is the unit of analysis, a limited ability to “recruit” units reduces study power considerably, an effect for which it is seldom possible to compensate by increasing the number of within-unit observations.

Point 2: balance

The more complex the system or unit of randomization, the less likely it is that randomization will achieve baseline balance successfully.

The randomization of large numbers of individuals, such as children, to a vaccine trial increases the confidence that the baseline characteristics (for example age, sex, etc.) of the groups being compared will be balanced. Larger systems, such as group practices, clinics or hospitals, are more complex; they involve considerably more baseline characteristics that could confound the observed results. For example, clinics can vary in many respects, of which staff complement, staff skill-mix, the types of leaders or managers, location, or the population served represent just a few. Thus, when larger systems are randomized, imbalance between trial arm characteristics at baseline is more likely to occur12 and there is less confidence that balance has been achieved.

Consider, for instance, the randomization of modestly-sized clinics offering primary care services and staffed by 7 to 10 health workers to an intervention aimed at improving staff practices. The range of factors that could affect the successful adoption of better practices might include: (i) facility characteristics such as location and the availability and constancy of resource supplies; (ii) health-worker characteristics such as skills and experience, team functioning, staff turnover and morale, and (iii) more general factors such as clinic ownership, supervision, workload, and the nature of the population served. If 20 clinics were randomized to two equal groups of 10, how confident could we be that their baseline characteristics were the same? Could we be as confident that we achieved a balance in baseline characteristics as we would if individuals had been randomized in a carefully conducted clinical trial with clear eligibility criteria for participants? Because there are more potentially important factors to balance, we would need to randomize more facilities (clusters) than individuals to attain similar confidence in baseline balance. If the units in the study were even larger, perhaps small hospitals, how many more factors might differ and influence the success of the intervention under study?

There are, of course, appropriate statistical methods that allow adjustment for multiple assessments made within a cluster and for characteristics at the subject and cluster level (and indeed at even higher levels) that could influence the effect under study.13 However, as noted above outcomes could be influenced by many factors, some of which could be difficult to measure, and with relatively limited numbers of clusters (see point 1) adjustment could be only partial.13 Thus, the larger the organizational units under study, the greater the number of factors and interactions influencing outcomes. Hence it is less safe to draw inferences based on the assumption that baseline characteristics are balanced, even after statistical adjustment, especially if the number of units studied is small.

Point 3: bias

Effect sizes may be attenuated as the intervention becomes more complex.

The difficulties posed by small sample sizes and the many factors that could influence and explain the observed effects can feasibly be addressed through good design and statistical analysis. However, the pathway from cause to effect is not as straightforward for many interventions in the health services arena as it is for a new drug for a specific disease, which produces a directly observable effect in its recipient. Health system interventions often rely on individual or group behaviours requiring successful completion of several (or sometimes numerous) process steps along the causal pathway from the intervention to its measured effect. For example, for a new desktop diagnostic test to produce the desired health effects, a consistent supply, user knowledge, correct and appropriately targeted use, appropriate post-test treatment and good patient compliance are all required. Each of these steps is fraught with opportunities for bias and confounding, which are in addition to any imbalance in baseline characteristics; multiple factors can affect and upset the intervention pathway influencing the observed effects. The greater the number of intermediary or contextual conditions potentially influencing the processes that link an intervention to the desired outcomes, the greater the likelihood of reduced effect size and of bias and confounding. It may be possible, and is often desirable, to reduce such effects by limiting variability at each step or component of a more complex intervention by carefully controlling the design and conduct of a study or even by adjusting for process variation in the analysis. However, it is seldom possible to eliminate such effects altogether, and if such careful implementation or process control cannot be achieved under real life conditions (often because of costs), the generalizability and value of the study’s findings may be threatened.

We now have two sets of factors that can influence the observed study results despite randomization. One set of factors increases the possibility of bias when causal pathways between the intervention and its effect are long: the other (covered in Point 2) increases outcome heterogeneity as organizational size and complexity increase. It is obviously possible for these two sets of factors to interact or modify each other. Although many researchers recognize the potential influence of these effects on outcomes, they typically ignore them in their initial estimates of effect and Type I and Type II errors (false positive and negative trial results, respectively).

Point 4: proving cause

As the complexity of interventions or contexts increases, randomization alone will rarely suffice to identify true causal mechanisms.

We often employ the reductive nature of individually randomized experiments to isolate a single input (intervention or therapy), make everything else equal, and observe the effects of this input. For example, we isolate the effect of a new vaccine by comparing the outcomes observed in those receiving and not receiving the vaccine. In such scenarios the link between the intervention (cause) and its effect is clear. Similar demands to demonstrate cause–effect relationships may be made of proposed health service interventions. With some highly specific inputs, such as condition-specific cash transfers,11 providing plausible evidence of a causal relationship may be possible. However, when interventions are complex, like the diagnostic
When designing comparative studies, we acknowledge the problem of random error. We anticipate that our observations could deviate from “the truth” because our samples could, by chance, be not entirely representative and our measurement tools could introduce random error. To increase our confidence that any observed differences between groups are not merely the unfortunate result of factors such as these, we estimate the probability that the magnitude of the observed differences could be explained by chance alone. When this probability is very low, we infer that the difference is real in all likelihood and that it resulted from the intervention – an inference strengthened by a high quality RCT design. However, our attention is usually focused on the difference in group means (or another group level summary term) as we try to account for the noise of within-group heterogeneity. Unfortunately, focusing our attention in this way often results in the intuitive but incorrect assumption that any heterogeneity in our observations is only explained by chance. Although using multi-level modelling approaches makes this intuitive leap less automatic, we still tend to focus on the “average effect”.

We should refrain from conflating heterogeneity due to random effects with heterogeneity due to real effects that we are unable to explain. Consider, for example, the familiar analogy for explaining chance, flipping a coin. By chance, we state, the probability of observing a head or tail is 50%. The critical part of the sentence here is “by chance”. However, if we studied things carefully and could consistently exert a force at just the right place on the coin to provide standard upward and rotatory moments, we would produce a specific and constant number of rotations during the coin’s arching rise and fall. The result would be an entirely predictable outcome of heads or tails. So what explains our view that flipping a coin provides a chance result is simply our inability to standardize conditions in line with well established laws of physics. Returning to our example of introducing a new diagnostic test into clinics, the challenge of standardizing conditions within a health system is soon apparent. We may be able to ensure consistent supplies (in a trial), but not to standardize which staff are present, particularly over time, or staff knowledge, or how staff apply that knowledge in every patient encounter, or how each patient responds. Thus, the more complex a setting and the more complex an intervention, the less likely we are to understand the laws governing action (intervention) and reaction and the less safe it is to dismiss heterogeneity as nothing more than error. In fact, the most informative part of any study will most probably be the attempt to understand such heterogeneity in the hope of uncovering new mechanisms that influence outcomes, an argument familiar to many social scientists.

**Discussion**

When assessing health system interventions, it may occasionally be impossible or unethical to conduct an RCT. For example, for a current study of how to improve practice in a tertiary and university hospital situated in a low-income country, there was no comparable facility to act as a control (ME, personal observation). In a recent large study of the value of training in neonatal care, it was deemed unethical to withhold training to allow for a control group. Thus, for these and other reasons researchers may have to consider the relative strengths of alternative designs, as discussed in Victora et al.1 However, randomization can and often should be used, as illustrated by Zurovac et al., or it can be problematic, as shown by Basinga et al. Yet their increasing familiarity with good practice in RCT leads many researchers to believe that randomization is a reliable, quick fix to prevent, or at least substantially reduce, the possible influence of residual confounding and bias on observed effects.

We do not seek to discount the central importance of randomization, and we have outlined some very good reasons to randomize in interventional research on health systems. Randomization is useful, for example, to prevent investigator-driven selection bias. However, even at the cluster level it is not the simple solution to challenging problems in study design, as is often believed. As units of intervention and study increase in size and complexity, and as interventions and causal pathways become more complex, the protection from bias and confounding that we expect after randomizing the number of units suggested by basic sample size calculations may be considerably less than we imagine. In addition, RCTs, often aimed at addressing narrowly specified
questions and maximizing internal validity, may have limited external validity if our interest lies in applying results in real life settings. Finally, when working with complex units of observation or complex interventions, we may miss valuable insights by assuming that any observed heterogeneity in outcomes, even in an RCT, reflects nothing more than random error.

Providing clear and absolute guidance on what randomization will achieve or on when to use it is, as we have seen, not possible. We therefore suggest thoughtful consideration rather than the automatic assumption that its use will produce an easily interpreted result. We have given here some simple points that may be helpful when considering the value of a randomized design. The same points may prove useful when considering the observed effects of alternative study designs or heterogeneity in the results of studies with the same design but within different health systems. Such points should also be considered when trying to determine the strength of the evidence surrounding an intervention's effectiveness. The more complex an intervention or the organizational units to which an intervention is applied, and the more complex the causal pathways linking the intervention to a given effect, the more complex the task of classifying the strength of the evidence supporting the intervention. Therefore, several reasons exist for recommending that RCTs of complex interventions, heretofore regarded as high quality evidence, might be downgraded when applying tools such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE). Finally, considering the points we have presented may strengthen the rationale for broader approaches to evaluation, including detailed investigations of pathways to effect.

Acknowledgements
This work is published with the permission of the Director of KEMRI, to whom we are grateful. The authors would also like to acknowledge helpful comments from the reviewers and editors on the initially submitted manuscript.

Mike English is also affiliated with the Department of Paediatrics at the University of Oxford, England.

Funding: Funds from a Wellcome Trust Senior Fellowship support Mike English (#076827). Joanna Schellenberg and Jim Todd are employed at the London School of Hygiene and Tropical Medicine. The funders had no role in the writing of this report or in the decision to submit it for publication.

Competing interests: None declared.
Resumen

La evaluación de las intervenciones en sistemas sanitarios: aspectos clave al considerar el valor de la aleatorización

Se necesita realizar investigaciones para facilitar la identificación de intervenciones que mejoren la capacidad o el funcionamiento de los sistemas sanitarios y, por tanto, contribuir a lograr las metas de salud global. Cuando se realiza correctamente, los estudios controlados aleatorizados (ECA) ayudan a reducir el sesgo y la confusión que pueden surgir en la evaluación de intervenciones, brindando la más sólida evidencia para identificar cuáles intervenciones son seguras y eficaces. Cuando es factible desde el punto de vista ético, también se pueden realizar estudios controlados de efecto directo (ECD) para evaluar intervenciones que mejoran la capacidad o el funcionamiento de los sistemas sanitarios, contribuyendo así a la obtención de mejores resultados globales de salud. Además, los autores sugieren que los estudios de caso pueden servir como puntos de partida para explorar nuevas intervenciones que complementen las ECA.

Resumen

La evaluación de las intervenciones en sistemas sanitarios: aspectos clave al considerar el valor de la aleatorización

Se necesita realizar investigaciones para facilitar la identificación de intervenciones que mejoren la capacidad o el funcionamiento de los sistemas sanitarios y, por tanto, contribuir a lograr las metas de salud global. Cuando se realiza correctamente, los estudios controlados aleatorizados (ECA) ayudan a reducir el sesgo y la confusión que pueden surgir en la evaluación de intervenciones, brindando la más sólida evidencia para identificar cuáles intervenciones son seguras y eficaces. Cuando es factible desde el punto de vista ético, también se pueden realizar estudios controlados de efecto directo (ECD) para evaluar intervenciones que mejoran la capacidad o el funcionamiento de los sistemas sanitarios, contribuyendo así a la obtención de mejores resultados globales de salud. Además, los autores sugieren que los estudios de caso pueden servir como puntos de partida para explorar nuevas intervenciones que complementen las ECA.
References


