Assessing health system interventions: key points when considering the value of randomization

Mike English,[a] Joanna Schellenberg[b] & 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.

Abstracts in عربي, 中文, Français, Русский and Español at the end of each article.

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 studies[4,5] 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 guidelines[9] have helped establish randomization as a defining feature of "a good intervention trial", a concept that extends to cluster randomized designs.

We agree that randomization is an extremely important tool in the researcher's armoury and do not dispute its importance in reducing the effects of various types of bias and confounding, especially when combined with concealment and blinding. These benefits are readily apparent when specific interventions, such as new drugs or vaccines, are tested at the individual level in safety, efficacy or effectiveness studies. Yet despite its undisputed value, randomization may not automatically provide the expected safeguards against confounding and bias, especially in research on what Lilford et al. have termed "targeted" or "generic" service interventions.[4] To help the general reader understand why the normal benefits of randomization are potentially reduced in the study of interventions delivered to components of the health system rather than directly to individuals we offer six points to consider. These points are also intended to illustrate the pitfalls of relying on the results of RCTs alone, without additional approaches to enquiry.

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

[a] KEMRI-Wellcome Trust Research Programme, PO Box 43640, Nairobi, 00100, Kenya.
[b] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, England.
[c] Department of Population Studies, London School of Hygiene and Tropical Medicine, London, England.
Correspondence to Mike English (e-mail: menglish@nairobi.kemri-wellcome.org).
(*Submitted: 20 April 2011 – Revised version received: 3 August 2011 – Accepted: 4 August 2011 – Published online: 6 October 2011*)

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 occur[12] 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 be 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.[11] 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 conditional cash transfers,[14] providing plausible evidence of a causal relationship may be possible. However, when interventions are complex, like the diagnostic

test described above, even if studies are designed in such a way as to demonstrate the link between steps (for example, that training increases use of the diagnostic), one cannot conclude that the results of studies of the individual components of a pathway can simply be combined to indicate an overall effect across the pathway. Thus, RCTs of individual intervention components, even if well conducted, do not necessarily provide information on the effects to be expected when interventions are combined. Conversely, randomized studies of complex interventions can provide evidence of an overall cause and effect relationship but are unable to attribute any specific effect to any single component.

Point 5: contamination

Working within routine health systems may limit our ability to control the spread of an intervention in part or in full.

Ensuring that intervention and control groups receive the correct intervention (which is often no active intervention in the control group) is critical in an RCT. In drug or vaccine trials incorrect treatment should be rare and rather easy to detect and often results in a clear protocol violation. Analyses should usually be conducted on the basis of "intention to treat" and the scale of protocol violation should normally be quantified to facilitate the interpretation of results. In studies of health systems, however, limiting the spread of an intervention, in part or in full, may be much harder, particularly if the studies are of long duration. For example, the knowledge possessed by study participants in the intervention group can spread to the control group through staff transfers, early uptake by training institutions targeting new employees or, perhaps increasingly, through expanding social and professional networks facilitated by widespread dissemination of communications technologies. Thus, the integrity of a control group may be threatened and it may be very difficult to assess, and hence to account for, the extent to which contamination may be undermining the interpretation of the magnitude of an effect or a negative study result.

Point 6: informative heterogeneity

Chance is less likely to explain outcome heterogeneity when units of study and interventions are complex.

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.[15] For these and other reasons researchers may have to consider the relative strengths of alternative designs, as discussed in Victora et al.[5] However, randomization can and often should be used, as illustrated by Zurovac et al.,[16] or it can be problematic, as shown by Basinga et al.[17] 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.[18,19] 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.[20] Therefore, several reasons exist for recommending that RCTs of complex interventions, hereto regarded as high quality evidence, might be downgraded when applying tools such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE).[21] Finally, considering the points we have presented may strengthen the rationale for broader approaches to evaluation, including detailed investigations of pathways to effect.[22,23] ■

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.

ملخص

تقييم تدخلات النظام الصحي- النقاط الرئيسية عند مراعاة قيمة الاختيار العشوائي

إن البحوث ضرورية للمساعدة في التعرف على التدخلات التي تسعى لتحسين قدرات وتشغيل النظم الصحية، حتى تساهم في تحقيق المرامي الصحية العالمية. وتقلل التجارب ذات الشواهد المختارة عشوائياً والجيدة الإعداد من التحيز والالتباس، وتقدم بينات راسخة لتحديد أي التدخلات المقدمة مباشرة للأفراد مأمونة وفعّالة. وعندما تكون هذه البحوث مقبولة أخلاقياً، يمكنها أن تساعد في الحد من التحيز والالتباس عند تقييم التدخلات التي تستهدف جميع النظم الصحية. لكن هناك تحديات إضافية ظهرت عندما ركزت البحوث على التدخلات التي تستهدف وحدات متعددة للمنظمة داخل النظم الصحية. ولذلك لا يستطيع أحد بقناعة تامة أن يفترض قدرة الاختيار العشوائي على الحد من أو القضاء على التحيز أو الالتباس بنفس الدرجة وفي جميع الحالات. وقد صاغ آخرون الجدل الدائر لصالح التصميمات البديلة، إلا أن هذه الورقة العلمية تهدف إلى مساعدة الناس في فهم الخطر الذي تتعرض له القيمة المحتملة للتجارب ذات الشواهد المختارة عشوائياً. وبالتحديد، تقترح الورقة العلمية ست نقاط لتؤخذ عين الاعتبار عند الكشف عن التحديات أمام تصميم أو تقييم البحوث ذات الشواهد المختارة عشوائياً لدراسة تدخلات النظم الصحية، وهي: عدد الوحدات المتاحة للاختيار العشوائي؛ تعقيد الوحدة التنظيمية قيد الدراسة؛ تعقيد التدخل؛ تعقيد المسار بين السبب والتأثير، والتلوث. بالإضافة إلى ذلك، اقترح الباحثون أن تغايرية النتائج قد تكون غنية بالمعلومات، وأنه يجب استكشاف أسبابها وعدم إهمالها. واستناداً إلى وضوح الفهم لقيمة التجارب ذات الشواهد المختارة عشوائياً والقصور المحتمل لها على تدخلات النظم الصحية، أوضح الباحثون الحاجة إلى خططٍ أوسع نطاقاً للتعليق على البحوث ذات الشواهد المختارة عشوائياً.

摘要

卫生系统干预评估——考虑随机化价值时应注意的要点

需要研究能够改善卫生系统能力或功能的干预措施，进而有助于实现全球健康目标。在确定哪些直接作用于个体的干预措施才是安全有效的问题上，组织妥善的随机对照试验（RCTs）在减少偏见和混淆方面提供了最有力的证据。在道德上可行的同时，该等试验还能在针对整个卫生系统的干预措施进行评估时帮助减少偏见和混淆。然而，当研究集中在针对卫生系统内部多个组织机构的干预措施时，额外的挑战也随之出现。因此，我们不能自以为是地认为随机化在每种情况下都能够相同程度地减少或消除偏见和混淆。尽管一些人已经发表了赞成替代设计的论点，然而本文的目的在于帮助人们了解为何随机对照试验的潜在价值可能受到威胁。具体而言，文章建议在探讨设计或评估关于健康系统干预的随机对照试验所蕴含的挑战时需要考虑六个要点：可进行随机化的机构的数量，所研究的组织机构的复杂性，干预的复杂性，因果关系的复杂性以及混淆性。此外，文章作者还建议结果的异质性也可以说明问题，应该探讨而不应忽视背后的原因。通过更好地了解有关卫生系统干预的随机对照试验的价值和可能的局限性，本文作者展示了我们为何需要更加宽广的研究平台从而完善随机对照试验。

Résumé

Évaluation des interventions des systèmes de santé: points-clés lors de l'examen de la valeur de la randomisation

Des recherches sont nécessaires pour permettre d'identifier les interventions qui amélioreront la capacité ou le fonctionnement des systèmes de santé, contribuant ainsi à atteindre les objectifs mondiaux en termes de santé. Des essais contrôlés randomisés (ECR) correctement réalisés, en ce sens qu'ils réduisent les biais et les confusions, offrent la preuve la plus solide pour identifier les interventions sûres et efficaces directement réalisées sur les personnes. Lorsque l'éthique le permet, ils peuvent également permettre de réduire les biais et les confusions lors de l'évaluation d'interventions ciblant les systèmes de santé dans leur intégralité. Toutefois, d'autres questions se posent lorsque la recherche est orientée sur les interventions qui ciblent les nombreuses unités d'organisation présentes dans les systèmes de santé. Ainsi, il est impossible de présumer avec légèreté que la randomisation est en mesure de réduire ou d'éliminer les biais ou les confusions dans la même mesure dans chaque instance. Alors que certains ont des arguments clairs en faveur de conceptions alternatives, cet article est destiné à expliquer pourquoi la valeur potentielle relative aux ECR peut être menacée. Il suggère en particulier six points à considérer lors de l'étude des questions qui apparaissent dans la conception ou l'évaluation des ECR sur les interventions des systèmes de santé: le nombre d'unités disponibles pour la randomisation, la complexité de l'unité d'organisation étudiée, la complexité de l'intervention, la complexité du parcours cause-effet et la contamination. De plus, les auteurs suggèrent que l'hétérogénéité des résultats peut être instructive et que les raisons sous-jacentes doivent être étudiées, et non ignorées. Se basant sur une meilleure compréhension de la valeur et des limitations possibles des ECR sur les interventions des systèmes de santé, les auteurs montrent les raisons pour lesquelles nous avons besoin de plateformes de recherche plus vastes afin de compléter les ECR.

Резюме

Оценка интервенций в системах здравоохранения: ключевые факторы при определении ценности рандомизации

Чтобы помочь выявлению интервенций, повышающих потенциал функционирования систем здравоохранения и, таким образом, способствующих достижению глобальных целей в области охраны здоровья, необходимы исследования. В той или иной мере уменьшая систематические ошибки и смещения, правильно проведенные рандомизированные контролируемые испытания (РКИ) предоставляют наиболее убедительные данные, позволяющие определить, какие интервенции, применяемые в отношении конкретных индивидов, безопасны и эффективны. Там, где это представляется возможным с этической точки зрения, они также способны уменьшить систематические ошибки и смещения при оценке интервенций, адресно ориентированных на системы здравоохранения в целом. Однако в тех случаях, когда предметом исследования являются несколько организационных ячеек, выявленных внутри систем здравоохранения, возникают дополнительные проблемы. Поэтому не следует наивно полагать, что рандомизация в любом случае одинаково способна уменьшать или устранять систематические ошибки или смещения. В то время как в других работах формулируются аргументы в пользу альтернативных планов исследования, цель нашей статьи – помочь людям понять, почему потенциальная ценность РКИ может быть поставлена под угрозу. В частности, в статье перечисляются шесть факторов, которые следует учитывать при анализе проблем, связанных с разработкой плана или оценкой результатов РКИ, посвященных интервенциям в области систем здравоохранения: (i) число организационных ячеек, доступных для рандомизации; (ii) сложность исследуемой организационной ячейки; (iii) сложность интервенции; (iv) сложность модели «причина – следствие»; (v) статистическое загрязнение и (vi) однородность результатов. Кроме того, авторы указывают, что однородность результатов может быть источником информации, и что ее причины необходимо исследовать, а не игнорировать. Опираясь на углубленное понимание ценности и возможных ограничений РКИ, посвященных интервенциям в отношении систем здравоохранения, авторы объясняют, почему нам необходимы более широкие исследовательские платформы, дополняющие РКИ.

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 los realiza correctamente, los estudios controlados aleatorizados (ECA), siempre que reduzcan el sesgo y la confusión, proporcionan la más sólida evidencia para identificar cuáles intervenciones brindadas directamente a las personas son seguras y eficaces. Cuando es factible desde el punto de vista ético, también pueden ayudar a reducir el sesgo y la confusión cuando se evalúan las intervenciones centradas en sistemas sanitarios completos. No obstante, surgen desafíos adicionales cuando la investigación se enfoca en intervenciones que se centran en múltiples unidades de organización encontradas dentro de los sistemas sanitarios. Por tanto, no se puede suponer con complacencia que la aleatorización puede reducir o eliminar el sesgo y la confusión en el mismo grado en cada caso. Si bien otros autores tienen argumentos expuestos a favor de diseños alternativos, en este documento el objetivo es ayudar a la gente a entender por qué puede verse amenazado el valor potencial de los ECA. Específicamente, propone seis puntos a tener en cuenta al explorar los desafíos del diseño o la evaluación de los ECA en las intervenciones en sistemas sanitarios: el número de las unidades disponibles para aleatorización, la complejidad de la unidad organizativa en estudio, la complejidad de la intervención, la complejidad de la relación de causa y efecto, y la contaminación. Además, los autores sugieren que la heterogeneidad de los resultados puede ser informativa y que deben explorarse y no ignorarse las razones detrás de dicho fenómeno. Basándose en la mayor comprensión del valor y las posibles limitaciones de los ECA en las intervenciones en los sistemas sanitarios, los autores demuestran por qué se necesitan plataformas más amplias de investigación para complementar los ECA.

References

1. Mexico, 2004: global health needs a new research agenda. *Lancet* 2004;364:1555–6. doi:10.1016/S0140-6736(04)17322-4 PMID:15519610
2. The Bellagio Study Group on Child Survival. Knowledge into action for child survival. *Lancet* 2003;362:323–7. doi:10.1016/S0140-6736(03)13977-3 PMID:12892965
3. First Global Symposium on Health Systems Research [Internet]. Symposium background. More and better health systems research; 2010. http://www.hsr-symposium.org/index.php/symposium-background [accessed 22 September 2011].
4. Lilford RJ, Chilton PJ, Hemming K, Girling AJ, Taylor CA, Barach P. Evaluating policy and service interventions: framework to guide selection and interpretation of study endpoints. *BMJ* 2010;341:c4413. doi:10.1136/bmj.c4413 PMID:20802000
5. Victora CG, Habicth JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health* 2004;94:400–5. doi:10.2105/AJPH.94.3.400 PMID:14998803
6. Medical Research Council. *Developing and evaluating complex interventions*. London: MRC; 2009.
7. Mills A, Hanson K, Palmer N, Lagarde M. What do we mean by rigorous health-systems research? *Lancet* 2008;372:1527–9. doi:10.1016/S0140-6736(08)61633-5 PMID:18984174
8. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. *Health Technol Assess* 1998;2:1–274. PMID:9919458
9. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869. doi:10.1136/bmj.c869 PMID:20332511
10. Campbell MK, Elbourne DR, Altman DG; CONSORT group. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702–8. doi:10.1136/bmj.328.7441.702 PMID:15031246
11. Hayes RJ, Moulton LH. *Cluster randomised trials*. London: CRC Press; 2009.
12. Raab GM, Butcher I. Balance in cluster randomised trials. *Stat Med* 2001;20:351–65. doi:10.1002/1097-0258(20010215)20:3<351::AID-SIM797>3.0.CO;2-C PMID:11180306
13. Ayieko P, Ntoburi S, Wagai J, Opondo C, Opiyo N, Migiro S et al. A multifaceted intervention to implement guidelines and improve admission paediatric care in Kenyan district hospitals: a cluster randomized trial. *PLoS Med* 2011;8:e1001018. doi:10.1371/journal.pmed.1001018 PMID:21483712
14. Morris SS, Flores R, Olinto P, Medina JM. Monetary incentives in primary health care and effects on use and coverage of preventive health care interventions in rural Honduras: cluster randomised trial. *Lancet* 2004;364:2030–7. doi:10.1016/S0140-6736(04)17515-6 PMID:15582060
15. Carlo WA, Goudar SS, Jehan I, Chomba E, Tshefu A, Garces A et al. Newborn-care training and perinatal mortality in developing countries. *N Engl J Med* 2010;362:614–23. doi:10.1056/NEJMsa0806033 PMID:20164485
16. Zurovac D, Sudoi RK, Akhwale WS, Ndiritu M, Hamer DH, Rowe AK et al. The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: a cluster randomised trial. *Lancet* doi:10.1016/S0140-6736(11)60783-6 PMID:21820166
17. Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet* 2011;377:1421–8. doi:10.1016/S0140-6736(11)60177-3 PMID:21515164
18. Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be? *BMJ* 2004;328:1561–3. doi:10.1136/bmj.328.7455.1561 PMID:15217878
19. Ranson MK, Sinha T, Morris SS, Mills AJ. CRTs – cluster randomized trials or "courting real troubles". *Can J Public Health* 2006;97:72–5. PMID:16512334
20. Plsek PE, Greenhalgh T. Complexity science: the challenge of complexity in health care. *BMJ* 2001;323:625–8. doi:10.1136/bmj.323.7313.625 PMID:11557716
21. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6. doi:10.1016/j.jclinepi.2010.07.015 PMID:21208779
22. Mackenzie M, O'Donnell C, Halliday E, Sridharan S, Platt S. Evaluating complex interventions: one size does not fit all. *BMJ* 2010;340:c185. doi:10.1136/bmj.c185 PMID:20123834
23. Litaker D, Tomolo A, Liberatore V, Stange KC, Aron D. Using complexity theory to build interventions that improve health care delivery in primary care. *J Gen Intern Med* 2006;21(Suppl 2):S30–4. doi:10.1007/s11606-006-0272-z PMID:16637958