Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2009; all rights reserved. Advance Access publication 17 April 2009

International Journal of Epidemiology 2010;**39**:89–94
doi:10.1093/ije/dyp174

How to assess the external validity of therapeutic trials: a conceptual approach

O M Dekkers,[1,2]* E von Elm,[3,4] A Algra,[1,5,6] J A Romijn[2] and J P Vandenbroucke[1]

Accepted 3 March 2009

Background External validity of study results is an important issue from a clinical point of view. From a methodological point of view, however, the concept of external validity is more complex than it seems to be at first glance.

Methods Methodological review to address the concept of external validity.

Results External validity refers to the question whether results are generalizable to persons other than the population in the original study. The only formal way to establish the external validity would be to repeat the study for that specific target population. We propose a three-way approach for assessing the external validity for specified target populations. (i) The study population might not be representative for the eligibility criteria that were intended. It should be addressed whether the study population differs from the intended source population with respect to characteristics that influence outcome. (ii) The target population will, by definition, differ from the study population with respect to geographical, temporal and ethnical conditions. Pondering external validity means asking the question whether these differences may influence study results. (iii) It should be assessed whether the study's conclusions can be generalized to target populations that do not meet all the eligibility criteria.

Conclusion Judging the external validity of study results cannot be done by applying given eligibility criteria to a single target population. Rather, it is a complex reflection in which prior knowledge, statistical considerations, biological plausibility and eligibility criteria all have place.

Keywords Clinical trial, external validity

Introduction

In clinical trials the effect of therapeutic interventions is estimated in persons who enrolled in a trial: the *study population*. Internal validity refers to the question whether the study results are valid for the original study population. External validity concerns the generalizability of study results to persons other than the

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.
[2] Department of Endocrinology and Metabolic Diseases, Leiden University Medical Center, Leiden, The Netherlands.
[3] Institute of Social & Preventive Medicine (ISPM), University of Bern, Bern, Switzerland.
[4] Department of Medical Biometry and Medical Informatics, University Medical Centre, Freiburg, Germany.
[5] Julius Center for Health Sciences and Patient Care, University Medical Center Utrecht, Utrecht, The Netherlands.
[6] Rudolf Magnus Institute, Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands.
* Corresponding author. Department of Clinical Epidemiology C7-99, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands.
E-mail: o.m.dekkers@lumc.nl

original study population.[1] We call the population of patients to whom the results should be generalizable the *target population*.

Internal validity is a prerequisite for the external validity. Study results that deviate from the true effect due to systematic error lack the basis for generalizability. Next, however, from a clinician's point of view the generalizability of study results is of paramount importance. According to the CONSORT statement external validity should be addressed in reporting randomized clinical trials (RCTs).[1] It is surprising how often external validity is neglected in methodological considerations about research.[2,3] Two reasons might account for this neglect: first, most clinical trails are directed towards the assessment of treatment effects in an ideal setting, and not towards the question whether the intervention has a positive effect when applied in routine clinical practice; and secondly, the concept of external validity is more complex than it seems to be from the deceptively simple description that it concerns generalizability of results to persons who were not included in the original study.

Clinical trials: internal validity, external validity and applicability

Trial results are obtained in persons who are included in the study as a representative fraction of an underlying source population. The internal validity of study results refers to the question whether the results suffer from systematic error. A study can be judged to be internally valid when the effect estimate differs from the true effect that would be obtained in the total source population only because of random error.

The terms generalizability, external validity, applicability, transferability and extrapolation are all used with overlapping meanings. We will not dwell on potential differences in meaning, but we will make a distinction between 'external validity'and 'applicability'. *External validity* will be used to denote the question whether the study results are valid for patients other than those in the original study population in a treatment setting that is in all respects equal to the treatment setting of the original study. External validity therefore involves *patient and disease characteristics*. Study results can be generalized to a specific target population if, and only if, the results are externally valid for this specific target population.

In contrast, we refer to *applicability* as the question whether study results are valid for patients to whom results are generalizable but who are in a different treatment setting than the original study population. Consequently, applicability involves characteristics of the *treatment setting*. For instance, if a neurosurgical intervention works well within a hospital with an experienced neurosurgeon, the results might in principle be generalizable to a similar population of patients. However, if a specific hospital lacks experienced neurosurgeons the applicability of the study findings might be limited. The proposed distinction between external validity and applicability may not always be clear-cut. For example, treatment effects of antibiotics depend on patient characteristics as well as resistance patterns within a specific treatment setting. However, the fact that this conceptual distinction is not clear-cut in all respects is no compelling argument against its use. Similarly, the distinction between selection bias and confounding by indication is widely used, although the boundaries are not clear-cut for all situations.

In addition, the value of clinical outcome of the treatment for an individual patient needs to be assessed. Moreover, individual circumstances could be taken into account to assess whether an individual patient should be treated according to the results of a trial or not.[4] For the purpose of this article, we will not go into the details of clinical relevance and individual treatment decisions.

Can external validity be formalized?

How can external validity be conceptualized to facilitate its assessment? One way that seems at first sight appealing is to define it in analogy to internal validity. A study is assessed to be internally valid when the study results differed from the true effect only because of random error. Analogously, a study would be externally valid if the study results differed from the value that would be obtained in the target population only because of random error. Thus, the assessment whether study results are externally valid would mean judging whether the results are valid for a defined target population. However, this approach is problematic.

Because external validity depends on a target population, the first step in the assessment of the external validity is to define this target population. The problem is that in clinical practice different doctors may want to apply the same research evidence to different target populations. For instance, suppose a study exists on the effect of antihypertensive drugs in patients between ages 45 and 74 years, with diastolic dysfunction but without severe co-morbidity. There are several possibilities to define target populations for this specific study. One doctor might strictly want to generalize to persons in the age bracket 45–74 years. However, another may wish to apply the results to all adult hypertensive patients with diastolic dysfunction without severe co-morbidity; referring to hypertensive patients <45 years as well as >74 years. Indeed, if patients >74 years were excluded from a study, should its external validity be restricted to patients below this age limit? Is there any reason to believe that the effects of the therapeutic intervention are not generalizable to 76-year-old patients? Or to those who are 77 years old? Likewise, would

the results not be generalizable to a 40-year old? But where should this extension of generalizability stop? Next, the severity of co-morbidity might be perceived in different ways. Should uncomplicated diabetes mellitus, treated with oral medication, considered to be a severe co-morbidity? And what about diabetes treated with insulin? It becomes clear that there is no single commonly agreed predefined target population for a given study. The question on generalizability must be pondered for various target populations, i.e. different types of patients, and the external validity should be assessed for each.

Of course, a counterargument might be that the target population is strictly defined by the eligibility criteria of the original study, which seems to solve the problem of the multiplicity of the external validity for one single study. Even if we leave aside that these criteria are often incompletely described in original reports,[5] this definition of the target population does not allow for a satisfactory conceptualization of external validity either. A target population that perfectly fits the eligibility criteria will, by definition, still differ from the original study population with respect to geographical, ethnical and temporal conditions. Each of these differences might affect the outcome of interest. For instance, in a study on the effect of clopidrogel in patients with acute myocardial infarction, 45 852 Chinese patients were included.[6] A target population could be defined strictly according to the inclusion criteria, i.e. patients within 24 h after onset of acute myocardial infarction. However, the question whether the results are generalizable to patients with the same condition from other countries or ethnic groups would remain unanswered. This may be relevant, since, for instance, for another drug, the ACE-inhibitor trandolapril, it has been shown that its effect on blood pressure is modified by ethnicity.[7] These considerations show that, in a very strict sense, the external validity of study results cannot be taken for granted even when the target population is defined by the same eligibility criteria of the original study.

In addition, external validity can also be affected because the original study population was not truly representative for the intended eligibility criteria. In a comparison between eligible patients with myocardial infarction who were and were not included in an RCT, important differences were found.[8] The included patients were younger, had less co-morbidity and their mortality was lower compared with the excluded patients. In other studies, patients who gave consent to participate in research differed from non-consenters in characteristics that can determine outcome.[9,10] This highlights that study populations can differ in a meaningful way from all patients who would fulfil the eligibility criteria.

These considerations show that external validity cannot be readily formalized. Although the eligibility criteria of the original report might be used in an attempt to define a target population, this does not guarantee the generalizability of the results to this specific target population that consists of new patients treated in other countries in diverse settings. The only formal way to establish the external validity would be to repeat the study in the specific target population, which would be rather unpractical given the large number of RCTs and an even larger number of potential target populations. Nonetheless, despite the lack of a clear definition and the lack of a formal way of proving that external validity for a specified target population exists, we have to judge external validity as being able to use the results of a particular randomized trial.

How can external validity be assessed?

The assessment of the external validity of clinical trials is a complex reflection in which eligibility criteria, prior knowledge, statistical considerations and biological plausibility are all involved. We propose a three-step approach to assess the external validity of therapeutic research. In addition, the applicability of study results should be addressed. These four aspects of generalizability are summarized in Table 1 and discussed in more detail below.

External validity and eligibility criteria

Eligibility criteria provide an approximate guide for generalizability. Indeed, the study population might not be representative for a population defined by its eligibility criteria, which can be the case when only a fraction of all eligible patients is actually included.[9,10] The proportion of patients actually included may serve as a good indicator for unaccounted selectivity in enrolment. Other aspects of a study, like the use of a run-in period, can limit the generalizability. During a run-in period, patients who do not comply or have side effects on study medication are excluded.[11] In an RCT on the effect of carvedilol in patients with chronic heart failure, during the run-in period, 6% of the patients were excluded due to progressive heart failure or death.[12] The exclusion of these patients may have led to an underestimation of the rate of adverse events and an overestimation of beneficial effects.[11] When patients are included in tertiary care referral centres, disease characteristics may differ from patients treated in non-referral centres or in general practice, which is not accounted for by the eligibility criteria. The first element in the assessment of external validity should address the question whether the study population differs from the source population with respect to characteristics that influence outcome. If so, the eligibility criteria are not a proper reflection of the study population and this might limit the study's external validity.

Table 1 Strategy to assess the external validity and applicability of clinical trials[a]

(i) External validity and eligibility criteria	
Are the eligibility criteria a proper reflection of the study population? If this is not the case, the external validity might be limited	
• Selection of study population	Patients who give consent to participate differ from non-consenters in characteristics that can determine outcome. The proportion of eligible patients that is actually included gives useful information about a potential degree of selectivity.
• Run-in period	The use of a run-in period may exclude non-compliant participants or participants at higher risk for side effects.
• Participating centres	Patients in secondary or tertiary referral centres differ from patients in non-referral centres and general practices.
(ii) Temporal, ethnical, socio-economic and geographical aspects	
Do temporal, ethnical and geographical differences between study population and target populations translate in to a limited generalizability?	
• Temporal aspects	The time elapsed since the original study was performed may translate into important changes in medical practice that influence treatment effects.
• Ethnical aspects	Ethnicity may interact with treatment effect.
• Geographical and socio-economic aspects	Geographical and socio-economic differences between study population and target population may affect treatment effects.
(iii) External validity beyond eligibility criteria	
Can study results be generalized beyond the eligibility criteria?	
• Age	RCTs mostly use strict age criteria. Generalizability beyond age criteria should be based on prior knowledge and biological plausibility.
• Co-morbidities	RCTs often exclude patients with co-morbidity. Generalizability to patients with co-morbidities should only be done with caution, and can only be based on external evidence.
(iv) Applicability of study results	
Do differences in treatment setting translate into possible differences in treatment effects?	
• Treating physicians	Treatment effects can depend on skills of treating physicians.
• Treatment setting	The setting of the treatment, i.e. the use of a study nurse, the frequency of controls and the availability of diagnostic procedures, may influence treatment results.
• Administrative policy	Administrative policies will influence treatment effects, especially for acute diseases that require treatment within a defined time window.

[a]This table includes relevant aspects that should be considered when assessing the external validity and applicability of clinical trials but is not exhaustive.

Temporal, ethnical and geographical aspects

Any target population will, by definition, differ from the study population with respect to temporal, ethnical, socio-economic and geographical conditions. Pondering external validity means asking whether these differences may influence the results and limit generalizability. For example, in studies on antibiotics, temporal and geographical differences can translate into differences in resistance patterns with impact on the effect of antibiotic treatment. Studies performed in the past or in another country might therefore not be generalizable to the intended patient populations. In addition, ethnic differences between study population and target population may influence generalizability. However, it is difficult to make a prior estimation on the possible interaction between ethnicity and treatment effects. Whereas the response to antihypertensives clearly depends on ethnicity,[13] the beneficial effects of aspirin in patients with acute ischaemic stroke were similar in a European and a Chinese population.[14,15] Geographic differences can translate into different treatment effects. For instance, the effect of vitamin D may depend on latitude[16] but also on ethnicity and seasonal effects.[17]

External validity beyond eligibility criteria

It is necessary to assess whether the conclusions of a study can be generalized to target populations that do not meet all the eligibility criteria. If studies were only generalizable to target populations that fulfil a study's eligibility criteria, this would encompass only a very small proportion of patients in daily practice. For instance, when the potential eligibility of asthma patients seen in routine clinical practice was examined, only 4–6% of them would have met the eligibility criteria of RCTs in asthma.[18] A similarly low percentage was found in a comparable study in chronic obstructive pulmonary disease (COPD) patients.[19]

About 50 years ago, Doll and Hill assessed the relationship between smoking and lung cancer in 59 600 doctors.[20] There is little doubt that these results are generalizable to athletes and statisticians, although no specific studies were done in these populations. The reason for the generalizability

is 2-fold. First, the relationship between lung cancer and smoking was established in additional studies, in very diverse situations—even if not for all potential combinations: for instance there are no specific studies in white female athletes. Secondly, there is no reason to doubt that the effect of smoking is different in the lungs of doctors compared with others, such as athletes. For reasons that are grounded in prior knowledge and biological plausibility the external validity could be extended to populations that were not eligible in the original study, nor ever studies specifically. In the same way, study results are often generalized beyond their specific age bands (see above).

Applicability of study results

Generalizability does not necessarily mean that study results are applicable.[21] As noted above 'applicability' concerns *treatment settings*. For example, a positive effect of spironolacton in patients with heart failure was shown in an RCT.[22] However, in observational studies the rate of severe hyperkalemia was reported to be much higher than in the original RCT.[23] In the RCT, patients were probably seen more frequently, thereby enhancing the probability of detecting the hyperkalemia at an early stage. This example shows that, although the results might have been perfectly generalizable, their applicability was limited because routine treatment settings differ from the setting in the original study. Also, other aspects of health care settings, such as administrative policies and diagnostic facilities, may influence treatment results in a way that is not determined by disease or patient characteristics.

Conclusion

Unlike internal validity, external validity cannot be easily formalized. The term external validity was called a misnomer,[24] because it suggests objectivity and a clear definition that it cannot satisfy. It is therefore not surprising that there is currently no consensus about how to assess the external validity of study results. Judgment of external validity of study results cannot be done by applying eligibility criteria to arrive at one single target population. Rather, it is a complex reflection in which prior knowledge, statistical considerations, biological plausibility and eligibility criteria all have place. The assessment of external validity is, at best, a well-argued, but fallible, statement about generalizability.

Acknowledgement

We thank Bruce Psaty for constructive comments on a previous draft of this article.

Conflict of interest: None declared.

References

[1] Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;**357:**1191–94.

[2] Glasgow RE, Green LW, Klesges LM *et al*. External validity: we need to do more. *Ann Behav Med* 2006;**31:** 105–8.

[3] Rothwell PM. External validity of randomised controlled trials: 'to whom do the results of this trial apply?' *Lancet* 2005;**365:**82–93.

[4] O'Connell D, Glasziou PP, Hill S. *How to Use the Evidence: Assessment and Application of Scientific Evidence*. NHMRC, 2000. Canberra, Australia.

[5] Shapiro SH, Weijer C, Freedman B. Reporting the study populations of clinical trials. Clear transmission or static on the line? *J Clin Epidemiol* 2000;**53:**973–79.

[6] Chen ZM, Jiang LX, Chen YP *et al*. Addition of clopidogrel to aspirin in 45,852 patients with acute myocardial infarction: randomised placebo-controlled trial. *Lancet* 2005;**366:**1607–21.

[7] Brunner M, Cooper-DeHoff RM, Gong Y *et al*. Factors influencing blood pressure response to trandolapril add-on therapy in patients taking verapamil SR (from the International Verapamil SR/Trandolapril [INVEST] Study). *Am J Cardiol* 2007;**99:**1549–54.

[8] Steg PG, Lopez-Sendon J, Lopez dS *et al*. External validity of clinical trials in acute myocardial infarction. *Arch Intern Med* 2007;**167:**68–73.

[9] Al Shahi R, Vousden C, Warlow C. Bias from requiring explicit consent from all participants in observational research: prospective, population based study. *Br Med J* 2005;**331:**942.

[10] Yuasa H, Kurita K, Westesson PL. External validity of a randomised clinical trial of temporomandibular disorders: analysis of the patients who refused to participate in research. *Br J Oral Maxillofac Surg* 2003;**41:**129–31.

[11] Pablos-Mendez A, Barr RG, Shea S. Run-in periods in randomized trials: implications for the application of results in clinical practice. *JAMA* 1998;**279:**222–25.

[12] Packer M, Bristow MR, Cohn JN *et al*. The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. U.S. Carvedilol Heart Failure Study Group. *N Engl J Med* 1996;**334:**1349–55.

[13] Materson BJ, Reda DJ, Cushman WC *et al*. Single-drug therapy for hypertension in men. A comparison of six antihypertensive agents with placebo. The Department of Veterans Affairs Cooperative Study Group on Antihypertensive Agents. *N Engl J Med* 1993;**328:**914–21.

[14] CAST (Chinese Acute Stroke Trial) Collaborative Group. CAST: randomised placebo-controlled trial of early aspirin use in 20,000 patients with acute ischaemic stroke. *Lancet* 1997;**349:**1641–49.

[15] International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *Lancet* 1997;**349:**1569–81.

[16] Prince RL, Austin N, Devine A, Dick IM, Bruce D, Zhu K. Effects of ergocalciferol added to calcium on the risk of falls in elderly high-risk women. *Arch Intern Med* 2008;**168:**103–8.

[17] Meier C, Woitge HW, Witte K, Lemmer B, Seibel MJ. Supplementation with oral vitamin D3 and calcium during winter prevents seasonal bone loss: a randomized controlled open-label prospective trial. *J Bone Miner Res* 2004;**19:**1221–30.

[18] Travers J, Marsh S, Williams M *et al*. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax* 2007;**62:**219–23.

[19] Travers J, Marsh S, Caldwell B *et al*. External validity of randomized controlled trials in COPD. *Respir Med* 2007;**101:**1313–20.

[20] Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *Br Med J* 1956;**2:**1071–81.

[21] Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. *Arch Intern Med* 2008;**168:**133–35.

[22] Pitt B, Zannad F, Remme WJ *et al*. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. *N Engl J Med* 1999;**341:**709–17.

[23] Juurlink DN, Mamdani MM, Lee DS *et al*. Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. *N Engl J Med* 2004;**351:**543–51.

[24] Rothman KJ, Greenland S. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins, 1998.

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2010;**39**:94–96
doi:10.1093/ije/dyp305

Commentary: External validity of results of randomized trials: disentangling a complex concept

Peter M Rothwell

University Department of Clinical Neurology, Level 6, West Wing, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK. E-mail: peter.rothwell@clneuro.ox.ac.uk

Accepted 25 August 2009

It is now widely accepted that in most situations, randomized controlled trials (RCTs) and systematic reviews are the most reliable methods of determining the effects of treatment. Yet, the methodology is still relatively new (a few decades old), and so our understanding of trial design, and more especially of how best to make use of results, is less than perfect. RCTs must be internally valid (i.e. their design and conduct must minimize the possibility of bias), and, until recently, guidelines on trial methodology and reporting, such as the CONSORT initiative, concentrated almost completely on issues related to internal validity. However, to be clinically useful, the result of a trial must also be relevant to clinical practice, i.e. be reasonably likely to be replicated when applied to a definable group of patients in a particular clinical setting. The extent to which a result can be extrapolated in this way has been variously termed as 'external validity, applicability or generalizability'.[1]

For some interventions, such as lowering blood pressure in chronic uncontrolled hypertension, the benefits have been shown to be generalizable to the vast majority of patients and settings, but the effects of other interventions will often depend on factors such as the characteristics of the patient, the method of application of the intervention and the setting of treatment. How these factors are taken into account in the design and performance of an RCT and in the reporting of the results can have a major impact on the clinical usefulness of the result. Lack of external validity has always been the most frequent criticism by clinicians of RCTs, systematic reviews and guidelines. Although much more research is required, systematic assessments of the external validity of trials in specific areas of medicine are now beginning to demonstrate the often substantial disparity between the information that is provided by RCTs and the information that is actually required by clinicians.[2,3] This disparity is one explanation for the underuse in routine practice of many treatments that have been shown to be beneficial in trials and are recommended in guidelines.

However, external validity is a 'slippery' concept. It can be defined in broad terms, as above, but is much more difficult to quantify exactly. While the determinants of internal validity are intuitive and