THEORY AND METHODS

Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy

Jon Nicholl

J Epidemiol Community Health 2007;**61**:1010–1013. doi: 10.1136/jech.2007.061747

Correspondence to:
Jon Nicholl, Medical Care Research Unit, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK; j.nicholl@sheffield.ac.uk

Accepted 12 March 2007

Observational studies comparing groups or populations to evaluate services or interventions usually require case-mix adjustment to account for imbalances between the groups being compared. Simulation studies have, however, shown that case-mix adjustment can make any bias worse.

One reason this can happen is if the risk factors used in the adjustment are related to the risk in different ways in the groups or populations being compared, and ignoring this commits the ''constant risk fallacy''.

Case-mix adjustment is particularly prone to this problem when the adjustment uses factors that are proxies for the real risk factors.

Interactions between risk factors and groups should always be examined before case-mix adjustment in observational studies.

Assessment of the effects of area-wide service or public health interventions, or the impact of technologies that are evolving over time, often involves non-randomised comparisons between populations in different places or measured at different times.

Although empirical studies comparing randomised and non-randomised controlled trial (NRCT) studies have shown mixed results, some finding that NRCT give biased estimates compared with randomised controlled trials and some that they do not,[1–8] all commentators agree that case-mix adjustment is an essential mark of quality in non-randomised comparisons evaluating interventions.

A recent study by Deeks *et al*.[9] has, however, found that NRCT using before and after or contemporary cohort designs are biased, and furthermore that case-mix adjustment is problematical, always increasing the variability of the estimated effect and rarely eliminating the bias. These findings are perhaps not surprising because in ''one-dimensional'' designs that compare before and after periods, or contemporary populations, there are likely to be some case-mix differences that affect outcome but have not been measured, often because they were not known about, and which have not therefore been taken into account. Ignoring this problem, and assuming that case-mix adjustment leads to unbiased comparisons has been termed ''the case-mix fallacy''.[10]

More surprising was the finding by Deeks *et al*.[9] that case-mix adjustment in one-dimensional NRCT not only failed to eliminate all the bias but sometimes increased it, and this cannot have been caused by failure to adjust for unknown covariates. One possible cause of this problem is that the relationship between the case-mix variable and the outcome differs between the populations or time periods being compared. Ignoring these interaction effects in a case-mix adjustment model commits what might be termed the ''constant risk fallacy''.

THE CONSTANT RISK FALLACY: WHAT IS IT?

Consider, for example, comparing outcomes or health service utilisation between populations or time periods adjusting for different socioeconomic status patterns using car ownership or level of education attainment. Conventional risk adjustment models ignore the fact that car ownership and educational attainment may ''mean'' different things in different populations and at different times. In the United Kingdom, for example, the level of ''risk'' indicated by having a higher education may have changed as the numbers have quadrupled in the past 20 years. Similarly, the level of risk indicated by car ownership may differ between urban and rural populations that are differentially dependent on cars for transport.

Another common example of the constant risk fallacy occurs with age. Known case-mix differences often include different age profiles and nearly all models used for case-mix adjustment include age as a covariate. As age-specific levels of morbidity may differ between populations and also change with time, adjusting for age assuming constant risk will always bias evaluations in favour of those services or interventions introduced into healthier populations or at healthier times. For example, evaluating the benefits of changes to trauma services introduced over the past 20 years using risk adjustment models including a constant age effect will mean that in the later years, if age-specific mortality is lower as a result of improved health, more deaths will be predicted than should have been and the evaluation will conclude that trauma care has improved even if services have not changed.

This type of effect is found in table 1, which shows the ''risk'' of death in seriously injured road traffic accident (RTA) casualties by mode of transport and age, for 1984 and 2004. Has the risk of death changed over time? It seems natural to analyse these data with a simple model comparing risk between years and including age and mode of transport to adjust for their changing distributions.

This model suggests that the odds of dying have actually increased by 35% over these 20 years. There is, however, a statistically significant decrease over time in the odds of dying associated with being elderly, by 24%, suggesting that ignoring this effect may be biasing the estimate of the effect of changing care.

The essential problem with using chronological age to case-mix adjust non-randomised comparisons is that age itself is not the risk factor, but a proxy for the actual risk factors such as serious co-morbidities, which become increasingly common as people age. The relationship between these true risk factors and chronological age can change over time or differ between

Abbreviations: NRCT, non-randomised controlled trial; RTA, road traffic accident

Table 1 Mortality rates in serious road traffic accidents* by year and age group

	1984		2004	
	0–59	60+	0–59	60+
Type of road user	n† (% died)	n† (% died)	n† (% died)	n† (% died)
Pedestrian	14 463 (6.7)	4834 (18.2)	5865 (6.8)	1479 (18.0)
Pedal cyclist	6033 (4.5)	537 (13.4)	2053 (5.0)	221 (14.5)
Two-wheeled motor vehicle user	19 486 (4.8)	420 (7.4)	6367 (8.9)	191 (8.4)
Car occupant	25 165 (6.9)	3668 (12.0)	13 519 (9.9)	2304 (14.4)
Other	2718 (6.6)	535 (7.9)	1187 (9.3)	303 (5.9)
All	67 865 (6.0)	9994 (14.6)	28 991 (8.7)	4498 (14.7)

*Data taken from Road accidents Great Britain 1984[11] and Road casualties Great Britain 2004.[12]
†Number of casualties killed or seriously injured in the year.

populations. This problem is implicit in age, period, cohort models, which recognise that the effect of age is not the same in different periods but also depends on the ''cohort'' to which the individual belongs. It is not just a problem with age or with longitudinal studies, but is a possibility whenever the covariate is not a direct objective measure of risk and is: (a) a proxy for the real risk factors, and the relationship between the proxy variable and the real risk factors differs between populations or over time; or (b) subjectively assessed, and the subjective assessment changes over time or different assessors are used in different populations; or (c) a ''label'', such as a diagnosis, subject to changes over time or variations between places in the policy or practice of labelling different groups.

The constant risk fallacy can also arise when the risk factor used in the case-mix adjustment is a direct and objective measure if it is measured in different ways in different populations. For example, blood pressure is a direct, objective measure of the risk of stroke, and an observational study comparing stroke incidence in two populations to assess the effectiveness of their different stroke services might want to adjust the comparison for differences in blood pressure distribution. Blood pressure measurements, however, depend on when and where they are measured. If these factors differ systematically between populations then measurements may ''mean'' different things in the two populations and ignoring this commits the constant risk fallacy.

THE CONSTANT RISK FALLACY: DOES IT MATTER?

Whether the constant risk fallacy matters depends on how often it occurs in case-mix adjustment and what the effect of ignoring it has on the bias.

We do not know how common the problem is because uncovering it requires testing whether the association between the outcome and each of the variables used in the case-mix adjustment differs between the populations being compared, and this is rarely done or reported. Nevertheless, opportunity for the fallacy to arise is common. It can be present whenever the covariate is not an objective, directly and consistently assessed measure of risk, and commonly used case-mix adjustment factors such as diagnostic category, age, and clinician-assessed disease severity open up the possibility of committing the constant risk fallacy.

With regard to the consequences, the risk associated with a risk factor may be sufficiently similar in the populations being compared for the interaction not to matter. There may, however, be a substantial interaction in populations that are very different like those in table 1, which are 20 years apart. In these cases ignoring the interaction in case-mix adjustment models may lead to an increase in the bias of the estimate of the treatment effect.

This is illustrated in the hypothetical example shown in box 1. In this example the observed odds ratio (OR) of having the outcome with the intervention (OR = 0.66) is clearly biased for the true odds ratio (OR < 0.5). Conventional case-mix adjustment increases the estimated odds ratio from 0.66 to 0.73 and has made the bias worse.

THE CONSTANT RISK FALLACY: WHAT SHOULD WE DO ABOUT IT?

The first step in risk adjustment in observational studies comparing populations should be to examine the ''interaction'' between risk factor and population, that is whether the relationship between risk factor and outcome differs between populations. Although tests for interactions between characteristics defining subgroups and outcomes in trials are well known to lack power compared with tests of main effects,[13] observational studies tend to be relatively large, and sometimes very large when they are based on routine data, so that this will usually be feasible.

If no evidence of an interaction is detected, then case-mix adjustment using conventional models may be appropriate. When an interaction is detected what should be done?

Model the interactions

One approach could be to let the effect of the risk factor differ between the populations by fitting a model including the interaction between the intervention term and the risk factor,

Table 2 Hypothetical results from an observational study

Risk factor value,	Population A, without intervention			Population B, with intervention			
x	r	n	Risk estimate*	r	n	Risk estimate*	Odds ratio
1	10	100	0.1	10	200	0.05	0.47
2	60	300	0.2	45	300	0.15	0.71
3	180	600	0.3	125	500	0.25	0.77
All levels	250	1000	0.25	180	1000	0.18	0.66

*Note that the risks in population A if they had had the intervention would be half the observed estimates, and in population B if they had not had the intervention would be twice the observed estimate.

Box 1 Hypothetical example

An intervention is being evaluated that is not used in population A but is used in population B, by comparing outcomes r in n = 1000 individuals in each population. Individuals also have an additional risk factor for r that has three values x = 1, 2, 3, but the distribution of these values differs markedly between the populations.

Suppose the effect of the intervention used in population B halves the risk of the outcome r so that the relative risk with the intervention equals 0.5, and this effect is the same whatever the level of the risk factor x. The risk factor, however, has different effects in the two populations. In population A, x = 2 doubles the risk compared with level 1, and x = 3 increases the risk by three times. In population B, level 2 increases the risk by three times, and level 3 by five times. Then the results of our evaluation might look like those shown in table 2.

We could analyse these data using logistic regression models in two different ways:

- We could ignore the risk factor, making no case-mix adjustment, and just fit a term for the population to represent the intervention. This gives an estimate of the odds ratio for the effect of intervention of 0.66.
- We could use a conventional risk adjustment model by also fitting a term for the value of the risk factor x, treating x as a covariate linearly related to the logit of the risk in the same way in each population. This gives an estimate for the odds ratio of 0.73.

As the odds ratio is always less than the relative risk when the latter is less than one, and the data were generated using a relative risk of 0.5, the correct answer for the odds ratio is less than 0.5. Fitting the risk factor assuming that it had the same effect on the risk in both populations thus made the bias in the estimated effect worse.

and then estimating the main intervention effect using this model.

This approach has two problems. First, it assumes that all of this ''interaction'' is caused by differences in the effect of the risk factor in the different populations and that none of the interaction is caused by the ''treatment'' having a different effect in individuals with different levels of the risk factor.

For example, fitting the age by period interaction for the RTA data in table 1 assumes that the interaction arises because the risk of death in serious accidents associated with age has changed between periods. It is possible, however, that other changes, for example changing trauma care, have benefited some age groups more than others, so that the ''treatment'' effect (the effect of changing trauma care) differs between age subgroups. This can also be seen in the table of results for the hypothetical example in box 1. If we did not know how the data had been generated, we could not know whether the different odds ratios in the subgroups are caused by differences between the effects of the treatment in the subgroups or differences between the effects of the risk factor between the populations. The assumption that there are no subgroup treatment effects is, of course, made in nearly all randomised trials, and however unreasonable, would not be exceptional.

The second problem with fitting interactions is that, with different risk factor relationships in the two populations, we can no longer uniquely estimate the treatment effect. Even if we are prepared to assume the treatment effect is constant, at which level of the risk factor do we estimate it? In the RTA example, is the effect of the treatment (changing trauma care, say) estimated as the difference in risks between 1984 and 2004 for younger road users or older road users?

Box 2 Examples of risk factors used for case-mix adjustment

- Directly measured, objective measures of risk, such as blood pressure, Injury Severity Score, tumour size
- Directly measured, but subjectively assessed measures of risk, such as clinician-assessed severity, disease stage, Glasgow Coma Score
- Indirectly measured, but objective, 'proxy' measures of risk, such as age, ethnicity, educational attainment, socioeconomic group
- Indirectly assessed classifications of types of patient, such as diagnostic category made using local definitions

Covariate selection

A second approach results from recognising that different risk factors that might be used in case-mix adjustment are more or less prone to the constant risk fallacy. We can identify several types of risk factor, such as biomarkers like blood pressure or respiratory function, which could reasonably be assumed to have a constant relationship to the risk if they are measured in the same way, but others such as age or socioeconomic status that are clearly proxies for the actual risk factors and are therefore prone to the fallacy (see box 2).

We could thus create a taxonomy or hierarchy of these factors and restrict case-mix adjustment to use only those covariates that are least prone to the constant risk fallacy. For example, if we have measured a biomarker we should use this (when it is consistently measured and related to the outcome) but only use measures of socioeconomic status in contemporary cohort designs and not in before-and-after studies over a period of time when their association with risk is likely to have changed.

Other solutions

A third type of solution may be to use different designs. For example, controlled before-and-after studies may be free of the bias caused by the constant risk fallacy because any change from before to after in the risk associated with the covariate may occur equally in the intervention and control populations.

Other solutions might be available that are specific to types of risk factor or types of design. For example, in the case of age, replacing chronological age with age-specific life expectancy calculated separately for each population or period being compared, which is a measure of health-related age, might avoid the constant risk fallacy.

CONCLUSION

The cause of the constant risk fallacy is, of course, just a specific type of unmeasured case-mix difference between populations. If we had measured the actual morbidity or ''health'' of the patients in the trauma example, rather than the proxy (age), then we could adjust for this unmeasured covariate. It is also more than this, however, because the constant risk fallacy occurs when the unmeasured risk factor interacts with the measured proxy so that the proxy is related to the outcome in one way in one population and in another way in different populations. When this happens conventional case-mix adjustment can increase the bias in estimated intervention effects.

What is already known

It is known that observational studies comparing groups or populations to evaluate services or interventions usually require case-mix adjustment to account for imbalances between the groups being compared. Simulation studies have, however, shown that case-mix adjustment can make any bias worse.

What this study adds

This paper shows that one reason this can happen is if the risk factors used in the adjustment are related to the risk in different ways in the groups or populations being compared, and ignoring this commits the ''constant risk fallacy''. Interactions between risk factors and groups should always be examined before case-mix adjustment in observational studies.

Policy implications

Interactions between risk factors and groups should always be examined before case-mix adjustment in observational studies.

Many other problems with case-mix adjustment have been highlighted,[14] particularly what Lilford *et al*.[10] have termed the case-mix fallacy, that is the erroneous assumption that case-mix adjustment removes all the variability in the outcome as a result of case-mix differences and therefore leads to unbiased comparisons. The constant risk fallacy, however, points to a more shocking problem: that case-mix adjustment may make the biases worse. The particular concern of Lilford and colleagues[10] was that comparisons of institutional performance based on risk adjustment were unfair and falsely stigmatising because they have only partly adjusted for the case-mix. The problems may, however, be worse than this because the models used for risk adjustment to compare institutional performance typically employ case-mix adjustment factors that may vary in definition, classification and measurement between institutions and are therefore particularly prone to the constant risk fallacy. Even the risk associated with age can vary dramatically between geographically close regions,[15] and comparisons of outcomes between the institutions serving such populations, adjusting for age differences in the demographics of their catchment areas, will always bias the comparison in favour of the institution serving the population with the lower age-specific morbidity and mortality.

More research is needed to understand the scope of the problem and to find the best ways of overcoming it. Nevertheless, some guidelines are clear. Interactions between risk factors and the populations or groups being compared should always be examined before using the factors for case-mix adjustment in observational studies. When interactions do not occur, then the covariates can safely be used in conventional risk adjustment models. When interactions are detected, however, this means either that the risk associated with the risk factor differs between populations, or that the intervention effect differs between subgroups defined by the risk factor. Observational studies are then faced with serious problems because we cannot determine which of the explanations is right. If it is the former we cannot adjust for any observed difference in the case-mix between the populations without running the risk of making the bias in the estimated treatment effect worse. Therefore, when interactions are detected we may simply have to recognise that one-dimensional observational study designs that compare populations or time periods are flawed.

Funding: The Medical Care Research Unit receives funding from the Department of Health.

The views expressed here are those of the author and not necessarily those of the Department of Health.

Competing interests: None.

REFERENCES

1 **Britton A**, McKee M, Black N, *et al*. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;**2**(13).
2 **MacLehose RR**, Reeves BC, Harvey IM, *et al*. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;**4**(34).
3 **Sacks H**, Chalmers TC, Smith H Jr. Randomised versus historical controls for clinical trials. *Am J Med* 1982;**72**:233–40.
4 **Kunz R**, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;**317**:1185–90.
5 **Benson K**, Hartz A. A comparison of observational studies and randomised, controlled trials. *N Engl J Med* 2000;**342**:1878–86.
6 **Concato J**, Shah N, Horwitz RI. Randomised, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887–92.
7 **Ioannidis JPA**, Haidich A, Pappa M, *et al*. comparison of evidence of treatment effects in randomised and non randomised studies. *JAMA* 2001;**286**:821–30.
8 **Lipsey MW**, Wilson DB. The efficacy of psychological educational, and behavioural treatments: confirmation from meta-analysis. *Am Psychol* 1993;**48**:1181–209.
9 **Deeks JJ**, Dinnes J, D'Amico R, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**(27).
10 **Lilford R**, Mohammed AM, Spiegelhalter D, *et al*. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;**363**:1147–54.
11 **Department of Transport**. *Road accidents Great Britain 1984*. London: HMSO, November, 1985.
12 **Department of Transport**. *Road casualties Great Britain 2004*. London: TSO, November, 2005.
13 **Brooks ST**, Whitley E, Peters TJ, *et al*. Sub-group analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;**5**(33).
14 **Iezzoni LI**. The risks of risk adjustment. *JAMA* 1997;**278**:1600–7.
15 **Nicholl JP**. Population intervention. *Lancet* 1989;**i**:718.