Evidence-based medicine. Part 5: Observational evidence

Tim Holt

Article points

1. Robust cause-and-effect relationships are the basis for intervention, and are also required to justify withdrawal of an otherwise beneficial intervention found to be associated with a harm.
2. Observational designs include studies in which treated and untreated groups are compared without the treatments being allocated randomly.
3. Observational evidence is essential for investigating the way disease affects populations, the patterns and distribution of risk within them, and the emergence of trends in health and disease over time.

Key words

- Case-control studies
- Cohort studies
- Confounding
- Cross-sectional and longitudinal surveys
- Pioglitazone
- Self-selection bias

Tim Holt is Academic Clinical Lecturer at the Department of Primary Care, University of Oxford, Oxford.

Observational evidence offers a huge volume of data to support health care, and can detect harms that randomised controlled trials may not be designed or powered to measure. Many research questions are outside the remit of clinical trials. Observational study designs include population-based longitudinal surveys, cross-sectional surveys, cohort studies, non-randomised controlled studies, case-control studies, case reports and case series. The fifth in a series looking at evidence-based medicine, this article explores observational evidence using the recent controversy surrounding the safety of pioglitazone as an example.

This series has highlighted the importance of the randomised controlled trial (RCT) in evidence-based medicine (EBM). A trial is a form of experiment, and randomisation attempts to isolate the intervention under study from other possible explanations for the result. This approach is the "cleanest" in an inferential sense, but is not the only means of understanding which factors influence outcomes. In practice, trial evidence may be unavailable and unobtainable. Other important questions are outside the remit of clinical trials. What is the natural history of a chronic disease? Which factors are associated with adverse outcomes, or with a more benign prognosis? Which groups are at greatest risk of a particular event, and how can we identify them?

RCTs are usually designed to measure benefits and are often under-powered to detect rarer harms. For drug therapies these must wait for detection through observational methods in the post-marketing phase. A purist might claim that no "cause and effect" inference can reliably be made based on such methods. Without randomisation, confounding variables and other biases may create an illusory causal link. But observational evidence offers a vastly greater volume of data, and the discipline of epidemiology relies (and thrives) on it, as does much of public health. Study designs include population-based longitudinal surveys, cross-sectional surveys, cohort studies, non-randomised controlled studies, case-control studies, case reports and case series.

Association does not equal causation

Statistical associations are often assumed to be causal but this assumption is unsound. The direction of causation may appear self-evident, but should always be questioned. Poor glycaemic control was known to be associated with greater incidence and progression of diabetic retinopathy (Brinchmann-Hansen et al, 1992; EURODIAB Complications Study Group, 1994) before RCTs demonstrated that improving control actually reduces the

risk of the condition. Prior to this, it was plausible that greater practical difficulties in achieving control in people with poor vision might explain the link, and this may indeed contribute to it. In other cases, two strongly associated phenomena may be linked through some other confounding factor. At an overall population level, risk of diabetes is strongly associated with risk of osteoarthritis, but only because both risks increase with age. Robust cause-and-effect relationships are the basis for intervention, and are also required to justify withdrawal of an otherwise beneficial intervention found to be associated with a harm.

Cause and effect: the Hill criteria

Austin Bradford Hill was a medical statistician, whose role in the development of RCT methodology was mentioned earlier in this series (Holt, 2011). Another of his major contributions was the stipulation of the nine "Hill criteria" (*Box 1*) supporting an assumption that a statistical association is likely to be causal (Hill, 1965).

Limitations of the Hill criteria

Most of Hill criteria are insufficient on their own, although number 8 is perhaps the most useful. However, there are times when experimental manipulation is either impossible or unethical; for example, investigating a suspected environmental toxin. A consensus based on the achievement of several criteria may be the best evidence available that a relationship is causal, after confounders have been accounted for. Even this may confuse us. The correlation between the local stork population and the community birth rate both in Berlin and in Lower Saxony from 1970–85 is a well-known example where this association was strong, consistent, specific, and temporal, but we assume spurious largely on the grounds of plausibility and coherence (Höfer et al, 2004). However, EBM has demonstrated the unreliability of conclusions based on plausibility, knowledge of underlying processes, and observational evidence alone. The most infamous example of this concerns coronary heart disease (CHD) risk and hormone replacement therapy (HRT).

CHD risk and oestrogen replacement

The myth of benefit of HRT on CHD risk arose during the 1980s and 1990s. A whole generation of post-menopausal women were advised to take HRT in the expectation of avoiding CHD (among other benefits). This policy was supported by assumed pathophysiological mechanisms and observational evidence. CHD is more common in men, particularly in those under 50 years of age, but after the menopause, when oestrogen levels fall, women start to catch up. This provided a plausible underlying model to understand the lower incidence observed consistently in women taking HRT (Stampfer and Colditz, 1991), explained by a protective effect of oestrogens. However, one large RCT, the Women's Health Initiative Study, demonstrated that in fact, HRT increases, rather than reduces, CHD risk (Rossouw et al, 2002). The lessons learnt over the limitations of observational evidence were summarised in a sobering commentary by David Sackett (2002).

Page points

1. Robust cause-and-effect relationships are the basis for intervention, and are also required to justify withdrawal of an otherwise beneficial intervention found to be associated with a harm.
2. Evidence-based medicine has demonstrated the unreliability of conclusions based on plausibility, knowledge of underlying processes, and observational evidence alone. The most infamous example of this concerns coronary heart disease risk and hormone replacement therapy.

Box 1. The Hill Criteria (Hill, 1965).

1. **Strength of association:** A stronger correlation may support a causal link but this criterion is unreliable on its own.
2. **Consistency:** Studies of different populations using different methods should indicate the same association.
3. **Specificity:** If the suspected causal influence is associated with a single (rather than multiple) outcome this strengthens the evidence that the link is causal.
4. **Temporality:** It is clearly a requirement that the cause precedes the effect.
5. **Biological gradient:** A dose-response relationship supports a causal link.
6. **Plausibility:** We are legitimately influenced by the plausibility of suspected causal mechanisms.
7. **Coherence:** A causal association would fit with existing understanding of the relevant processes.
8. **Experiment (reversibility):** The manipulation of one phenomenon alters the level of the other. Reversing the effect by withdrawing the influence strengthens the evidence for causation.
9. **Analogy (considering alternative explanations):** An attempt to identify a causal relationship and a direction of causation should consider other possible models. The exclusion of confounding effects is the most important safeguard in non-randomised population data.

Page points

1. Observational designs include studies in which treated and untreated groups are compared without the treatments being allocated randomly.
2. A cross-sectional study determines the patterns present at a single point in time, and might report prevalence of a disease, a risk factor or any health parameter and its age-standardised distribution in the study population.
3. A prospective cohort study follows a defined group of people over a period of time to determine the influence of factors measured at baseline on health outcomes.

Data dredging and "fishing expeditions"

A danger in exploring observational data is the risk of detecting associations that, even though statistically significant, are entirely spurious, i.e. occurring through chance. At the traditional 5% level of significance, we expect to find these in about one in 20 such analyses, and repeated testing makes such discoveries almost inevitable. In RCTs, subgroup effects (for example, identifying greater benefit in certain subgroups) that were not pre-specified in the protocol should be treated with similar suspicion (Sleight, 2000). Exploratory "fishing expeditions" (a term used by the National Institute of Health Research [2007] as a warning to grant applicants) are of limited use on their own but may lead on to prospective, hypothesis testing studies including RCTs. An RCT designed to detect pre-specified subgroup effects based on the findings of prior exploratory studies is a perfectly legitimate approach.

Non-randomised controlled studies

Observational designs include studies in which treated and untreated groups are compared without the treatments being allocated randomly. The Freemantle Study is a large, ongoing observational study based in Australia, designed to investigate many different aspects of diabetes care. One of the recently published sub-studies concerned the influence of aspirin on cardiovascular outcomes (Ong et al, 2010). Cardiovascular event rates were compared between patients without prior cardiovascular disease treated with aspirin during routine care with those who were not. The reduced incidence found in those treated supports the use of aspirin in primary prevention, but the authors are careful to point out that this evidence is observational. Those taking aspirin in the study might be different in some ways from the comparator group, ways that might in fact explain the apparent benefit. For instance, any tendency for clinicians to avoid using aspirin in people whose cardiovascular risk is higher (perhaps due to general frailty) might lead to an apparent benefit that is not in fact due to aspirin. If so, "frailty" (however we define it) would be a confounding factor. There might also be a patient preference for taking the drug that is associated in some way with lower risk, as may well have happened in the HRT example above. This is termed "self-selection bias" and may be impossible to identify (Howick, 2011). Only an RCT of aspirin versus placebo will reliably answer the question of whether its benefits outweigh the risks, and the current ASCEND (A Study of Cardiovascular Events in Diabetes) trial is designed to resolve this issue (see www.ctsu.ox.ac.uk/ascend).

Cross-sectional and longitudinal surveys

A huge volume of data is generated through routine health and social care, as well as through survey data gathered prospectively. A cross-sectional study determines the patterns present at a single point in time, and might report prevalence of a disease, a risk factor or any health parameter and its age-standardised distribution in the study population. For example, a study published in 2007 reported the distribution of quality of diabetes care markers among different ethnic groups in Wandsworth (Gray et al, 2007). Such surveys can be repeated at a later date to identify longitudinal trends. Longitudinal surveys can also report incidence rates including standardised mortality ratios.

Cohort studies

A prospective cohort study follows a defined group of people over a period of time to determine the influence of factors measured at baseline on health outcomes. Regression techniques are used to determine the relative importance of the baseline variables, and to identify significant interactions between these predictors. Martínez-González et al (2008) used this method to investigate the association between adherence to a Mediterranean diet and risk of subsequent diabetes development, an area difficult to study using an RCT. Their analysis was adjusted for several potential confounders, including age, sex, years of

university education, total energy intake, BMI, physical activity, smoking status, history of hypertension and family history of diabetes. Adherence to the diet at baseline was associated with a significantly reduced risk of developing diabetes, and there was evidence of a gradient, i.e. the higher the adherence, the lower the risk.

Case-control studies

RCTs are usually underpowered to detect harms, which may be very rare but still important. It may be unrealistic to design and conduct an RCT on the necessary scale to determine whether a rare harm is more likely in the intervention group compared with the control group. It might even be unethical to do so. The harm might take much longer to develop than the usual timescale of a trial. For such problems, a case-control design is useful. This looks at a group of people who actually have the condition of interest (the suspected "harm"). These are the cases, and they are matched by a group of controls who do not have the condition. Matching involves selection of individuals that are otherwise similar in a range of respects that might be relevant confounders. These typically include age, sex, drug treatments, other important comorbidities and socioeconomic status, but could include anything considered relevant.

The object is to determine whether or not the trigger of interest is significantly more common in cases than controls, accounting for (and therefore removing the influence of) these confounders. An example of this technique is a study demonstrating that hepatocellular carcinoma is three times more common in people with diabetes than in those without, accounting for other possible risk factors (Davila et al, 2005). This method is a very efficient means of studying a rare condition. However, while it can confirm a suspected statistical association, it cannot prove the link is causal nor (despite the title of the article by Davila et al [2005]) the direction of causation.

Case reports and case series

Case reports describe interesting or unexpected findings in one or a few individuals. The next stage is to confirm the initial finding, look for other such cases, and if possible, identify underlying mechanisms. A recently reported improvement in psoriasis in a person with type 2 diabetes treated for just a few days with a glucagon-like peptide-1 (GLP-1) receptor agonist was followed-up by studying two other cases followed over 6 weeks, with measurement of immunological parameters considered likely to be mediators (Hogan et al, 2011). This research might ultimately lead on to an RCT of GLP-1 receptor agonists in the treatment of psoriasis, but is at a very early stage. In deciding whether the effect is real, knowing the underlying mechanism is not essential, but an understanding of it may help to design the trial, for instance by selecting the most appropriate patient groups to include.

Serendipity and innovation

Serendipity is an important source of innovation in medicine, and surprises often arise from observational data. In 1962 a laboratory study of seizures in rodents demonstrated anticonvulsant effects in a number of experimental substances tested, and it became apparent that the organic solvent used in all of them was the common denominator: valproic acid (Henry, 2003). After 80 years on the laboratory shelf, this compound became a standard antiepileptic therapy, marketed during the 1970s following confirmatory RCTs, and through unexpected benefits observed in these trials and in case studies, its antimigraine and mood-stabilising effects were also detected and later confirmed.

This story provided a happy ending for thousands of people living with epilepsy and struggling with the notorious side-effects of previous treatment options. However, in studying new drug therapies, cases of unexpected benefit are unfortunately outnumbered by cases of unexpected harm.

Pioglitazone and bladder cancer risk

A controversy has arisen over the safety of pioglitazone, which appears to be associated

Page points

1. Randomised controlled trials (RCTs) are usually underpowered to detect harms, which may be very rare but still important. It may be unrealistic to design and conduct an RCT on the necessary scale to determine whether a rare harm is more likely in the intervention group compared with the control group.
2. An example of the case-control technique is a study demonstrating that hepatocellular carcinoma is three times more common in people with diabetes than in those without, accounting for other possible risk factors.
3. Serendipity is an important source of innovation in medicine, and surprises often arise from observational data.

Page points

1. The thiazolidinedione drug class improves insulin sensitivity as its primary mechanism, and has been found (along with sulphonylureas) to be among the most effective oral antidiabetes agents.
2. The data available for pioglitazone are largely observational, and in addition to the French and USA population-based studies, include a case control study based on the General Practice Research Database.
3. Relying on observational evidence leaves the investigation vulnerable to confounding factors and other biases. For instance, people prescribed pioglitazone in clinical practice may, on average, be more centrally obese than otherwise similar people given other drugs, such as sulphonylureas.

with an increased risk of bladder cancer. This link has been suspected and monitored for some years, but came to general notice following a retrospective French cohort study reported in June 2011 (available in French at: http://bit.ly/oRhbNW). This suggested an overall adjusted hazard ratio of 1.22 for users of pioglitazone (95% confidence interval 1.05, 1.43), and led to the withdrawal of the drug in France. A separate epidemiological study in the USA suggested no overall risk, but an increased risk in those with longer duration of use and cumulative dose (Lewis et al, 2011).

Pioglitazone is the only remaining thiazolidinedione in use following the withdrawal of rosiglitazone. This drug class improves insulin sensitivity as its primary mechanism, and has been found (along with sulphonylureas) to be among the most effective oral antidiabetes agents (Sherifali et al, 2010). Pioglitazone is particularly effective in centrally obese individuals in whom insulin resistance is usually a factor. It is not difficult to find examples among our patients of people who, without this drug would either need insulin, or a newer agent whose long-term safety has not been firmly established. So this is a difficult situation both for clinicians and people with diabetes.

Rosiglitazone was found to be associated with increased risk of myocardial infarction (MI) in the post-marketing phase, and after prolonged discussion it was withdrawn from general use. The same discussions concluded that pioglitazone does not carry the same MI risk (although both cause fluid retention and can precipitate or exacerbate heart failure in susceptible individuals). Despite the gravity of the rosiglitazone issue, the drug was not withdrawn without an exhaustive examination of the most reliable evidence available – meta-analysed RCT data (Nissen and Wolski, 2010). This provided confidence, based on studies of randomised populations, that the association was not only real but also causal.

The data available for pioglitazone are largely observational, and in addition to the French and USA population-based studies, include a case-control study based on the General Practice Research Database. There is also a (not yet published) meta-analysis of RCT evidence, but numbers are small, with only 19 cases of bladder cancer in 12 506 pioglitazone users (0.15%) compared with seven cases in 10 212 non-users (0.07%) reported by the European Medicines Agency (2011). The current recommendation in the UK is that the drug should only be used after weighing up risks and benefits in the individual. This of course is important for prescribing any drug, but for this particular problem, the scale of the risk is unclear.

This exposes the difficulties in establishing "cause and effect" relationships in this situation. Relying on observational evidence leaves the investigation vulnerable to confounding factors and other biases. For instance, people prescribed pioglitazone in clinical practice may, on average, be more centrally obese than otherwise similar people given other drugs, such as sulphonylureas. Obesity itself raises the risk of bladder as well as other cancers (Wolk et al, 2010), so unless this confounder is adequately controlled for, some of the association might be due to a tendency to prescribe pioglitazone in those already at higher risk. No adjustment for obesity has so far been reported in either the French or the USA studies cited.

A sound verdict on this issue should ideally be based on evidence from randomised populations if more of such data become available. This would make the risk estimable, to allow a balanced, person-centred decision to be made for each individual. In the meantime, the conclusion is that the association is significant and likely to be causal, but mild, associated with use of the drug for more than a year, and related to cumulative dosage. This risk may be outweighed by benefit, but this decision needs to be made on an individual basis, taking into account the clinical response to the drug, which benefits some individuals much more than others (Kenny, 2011).

Conclusion

Observational evidence has a central role in medicine, particularly in the disciplines of epidemiology and public health. It often involves much larger volumes of data than those from randomised trials, which are more difficult and at times impossible to collect.

Many research questions are outside the remit of clinical trials, but observational evidence should be interpreted with caution. It is less effective at identifying "cause and effect" relationships, where randomised trial evidence is superior. It is nevertheless essential for investigating the way disease affects populations, the patterns and distribution of risk within them, and the emergence of trends in health and disease over time. As a source of innovation, it provides insights at all levels from individual case studies to large population based surveys, and through the generation of hypotheses supports the design of interventional trials underpinning EBM. ■

"Many research questions are outside the remit of clinical trials, but observational evidence should be interpreted with caution. It is less effective at identifying 'cause and effect' relationships, where randomised trial evidence is superior."

Brinchmann-Hansen O, Dahl-Jørgensen K, Sandvik L, Hanssen KF (1992) Blood glucose concentrations and progression of diabetic retinopathy: the seven year results of the Oslo study. *BMJ* **304**: 19–22

Davila JA, Morgan RO, Shaib Y et al (2005) Diabetes increases the risk of hepatocellular carcinoma in the United States: a population based case control study. *Gut* **54**: 533–9

EURODIAB Complications Study Group (1994) Microvascular and acute complications in IDDM patients: the EURODIAB IDDM Complications Study. *Diabetologia* **37**: 278–85

European Medicines Agency (2011) *European Medicines Agency Recommends New Contra-Indications and Warnings for Pioglitazone to Reduce Small Increased Risk of Bladder Cancer.* EMA, London. Available at: http://bit.ly/rdp5El (accessed 12.09.11)

Gray J, Millett C, Saxena S et al (2007) Ethnicity and quality of diabetes care in a health system with universal coverage: population-based cross-sectional survey in primary care. *J Gen Intern Med* **22**: 1317–20

Henry TR (2003) The history of valproate in clinical neuroscience. *Psychopharmacol Bull* **37**(Suppl 2): 5–16

Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* **58**: 295–300

Höfer T, Przyrembel H, Verleger S (2004) New evidence for the theory of the stork. *Paediatr Perinat Epidemiol* **18**: 88–92

Hogan AE, Tobin AM, Ahern T et al (2011) Glucagon-like peptide-1 (GLP-1) and the regulation of human invariant natural killer T cells: lessons from obesity, diabetes and psoriasis. *Diabetologia* 9 Jul [Epub ahead of print]

Holt T (2011) Evidence-based medicine. Part 2: Randomised controlled trials. *Diabetes & Primary Care* **13**: 82–7

Howick J (2011) *The Philosophy of Evidence Based Medicine.* Wiley-Blackwell (BMJ Books), Oxford

Kenny C (2011) Pioglitazone: Balancing risks and benefits. *Diabetes & Primary Care* **13**: 199–200

Lewis JD, Ferrara A, Peng T et al (2011) Risk of bladder cancer among diabetic patients treated with pioglitazone: interim report of a longitudinal cohort study. *Diabetes Care* **34**: 916–22

Martínez-González MA, de la Fuente-Arrillaga C, Nunez-Cordoba JM et al (2008) Adherence to Mediterranean diet and risk of developing diabetes: prospective cohort study. *BMJ* **336**: 1348–51

National Institute for Health Research (2007) Programme Grants for Applied Research. Comments on proposals submitted in the first application round. NIHR, London. Available at: http://bit.ly/ofAjpg (accessed 12.09.11)

Nissen SE, Wolski K (2010) Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. *Arch Intern Med* **170**: 1191–1201

Ong G, Davis TM, Davis WA (2010) Aspirin is associated with reduced cardiovascular and all-cause mortality in type 2 diabetes in a primary prevention setting: the Fremantle Diabetes study. *Diabetes Care* **33**: 317–21

Rossouw JE, Anderson GL, Prentice RL et al (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA* **288**: 321–33

Sackett DL (2002) The arrogance of preventive medicine. *CMAJ* **167**: 363–4

Sleight P (2000) Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them! *Curr Control Trials Cardiovasc Med* **1**: 25–7

Sherifali D, Nerenberg K, Pullenayegum E et al (2010) The effect of oral antidiabetic agents on A1C levels: a systematic review and meta-analysis. *Diabetes Care* **33**: 1859–64

Stampfer MJ, Colditz GA (1991) Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med* **20**: 47–63

Wolk A, Gridley G, Svensson M et al (2001) A prospective study of obesity and cancer risk (Sweden). *Cancer Causes Control* **12**: 13–21