Special Feature

EJA | EVALUATION Journal of Australasia

The best medicine: Lessons from health for policy randomistas

Evaluation Journal of Australasia
2024, Vol. 24(1) 6–13

DOI: 10.1177/1035719X231226437
journals.sagepub.com/home/evj
Sage

Andrew Leigh
Parliament of Australia, Australia

Abstract

The Australian Government has established the Australian Centre for Evaluation, with a mandate to conduct rigorous evaluations, including randomised trials. This approach draws heavily on the transformation of medicine. During the 20th century, medicine evolved into a discipline that is largely driven by evidence from clinical trials. Drawing appropriate lessons from health has the potential to substantially improve the equity and effectiveness of policymaking in Australia.

Keywords

randomised trials, randomised controlled trials, impact evaluation, health policy, health evaluation

Introduction

The creation of the Australian Centre for Evaluation marks a substantial milestone in how the Australian Government regards policy evaluation. The Centre's focus will be on conducting rigorous evaluations, including randomised trials.

The aim of this article is to provide readers with a sense of the philosophy underpinning the Australian Centre for Evaluation, how the Australian Government intends it to operate and collaborate, and our hope for how it will contribute to the evaluation landscape in Australia. To date, randomised trials have been very rare in Australia, constituting a tiny fraction of all evaluations. The Australian Centre for

Corresponding author:
Andrew Leigh, Parliament of Australia, Parliament House, Canberra, ACT 2600, Australia.
Email: andrew.leigh.mp@aph.gov.au

Evaluation aims to expand the prevalence of high-quality impact evaluations across a wide range of policy areas.

The best way to encapsulate the Australian Government's approach to policy evaluation is that the government is taking lessons from health, and in particular how rigorous evaluation has changed medical practices, saving money and lives. As historian David Wootton notes, 'For 2,400 years patients have believed that doctors are doing them good; for 2,300 years they have been wrong' (Wootton, 2006, p. 2). It was only with the rise of germ theory and modern statistics that the medical profession began to make a major contribution to improving human health.

To illustrate this approach, this article considers two areas of health that have been significantly influenced by randomised trials – breast cancer treatment and nutritional studies – before turning to detail the approach of the Australian Centre for Evaluation. It concludes with some thoughts about how new developments in health might influence the use of evidence in policymaking.

The radical mastectomy

In the 1880s, US surgeon William Halsted formed the view that breast cancer was most effectively treated by excising large portions of the patient's tissue. Previous operations, he argued, had been too timid. Observing that patients often relapsed after surgery that removed only the tumour, Halsted advocated removing considerable amounts of surrounding tissue.

Halsted's surgery removed the pectoralis major, the muscle that moves the shoulder and hand. He called it the 'radical mastectomy', drawing on the Latin meaning of radical to mean 'root'. In *The Emperor of All Maladies*, Siddhartha Mukherjee describes how Halsted and his students took the procedure further and further. They began to cut into the chest, through the collarbone and into the neck. Some removed ribs. They sought out the lymph nodes, and claimed they had 'cleaned out' the cancer.

Women who endured these operations were left permanently disfigured, often with gaping holes in their chests. In some cases, they were unable to properly move an arm. In other instances, their shoulders permanently hunched forwards. Recovery could take years. But Halsted was unrepentant, referring to less aggressive surgery as 'mistaken kindness' (Mukherjee, 2010, p. 64).

Halsted persuaded others not through his data, which was shaky, but through the force of his rhetoric and personality. He was supremely self-confident, perhaps fuelled through his cocaine addiction, and belittled his critics for their faint heartedness. Radical mastectomies, he acknowledged, would disfigure patients (Wright, 2018). But these war wounds were the price of winning the battle.

Yet whether a patient survived breast cancer depended not on how much tissue was removed, but whether the cancer had metastasised and spread through her body. If it had not metastasised, a more precise operation to remove the cancer would have been just as effective. If it had metastasised, then a radical mastectomy would still fail to remove it.

In 1967, Bernard Fisher became chair of the National Surgical Adjuvant Breast Project at the University of Pittsburgh School of Medicine. Fisher was struck by the lack of evidence supporting the radical mastectomy. He became interested in the mystery of metastasis and the growing use of clinical trials in medicine. No matter how venerable the clinician, he argued, experience was no substitute for evidence. So Fisher began recruiting for patients to take place in a clinical trial that would test the impact of the radical mastectomy by comparing the surgery against a more moderate alternative, the lumpectomy, that involved removing only the cancerous tissue (Mukherjee, 2010).

Fisher's randomised trial faced major hurdles. The first was to persuade women to participate in a trial in which randomisation would determine whether the surgeon would remove a lump or their entire breast. The second was to persuade surgeons to refer patients to the trial. Having been trained in radical surgery, many surgeons felt that a lumpectomy was unethical and were hostile to the trial. They flatly refused to refer patients to a trial that might see them receiving anything other than a radical mastectomy (Mukherjee, 2010, pp. 200–201). But Fisher was helped by the burgeoning feminist movement. Feminist activists concerned of the lack of evidence around radical mastectomies wrote letters to newspapers, and asked pointed questions at medical conferences. As activist Cynthia Pearson noted, the 'women's health movement began talking about mastectomy as one of the examples of sexism in medical care' (Otto, 1994).

In the face of opposition in the United States, Fisher's breast cancer surgery trial had to be expanded to Canada to get sufficient sample size. Eventually, it covered 1,665 patients, who were randomised into three groups – radical mastectomy, simple mastectomy and surgery followed by radiation. The results were finally published in 1981 (Fisher et al., 1981). They showed that there were no differences in mortality between the three groups. The women who had undergone radical mastectomies had suffered considerably from the surgery – yet they had not benefited in terms of survival. Fisher's randomised trial changed how surgeons treat breast cancer, but it took a century. Between the 1880s and the 1980s, around half a million women underwent radical mastectomies, an unnecessary surgical treatment (Mukherjee, 2010).

Your choices say a lot about you

Randomised trials are valuable in instances where experts have strong views. In the case of breast cancer treatment, it took a powerful experiment to persuade the experts that the prevailing wisdom had been wrong. An advantage of randomised trials is that they identify a clear counterfactual – what would have happened without the intervention. This can be especially important in instances where people self-select into different treatments.

To see this, suppose that we wished to conduct an experiment on the impact of caffeine on whether people stay awake during a talk by a politician on a Friday afternoon at the end of a stimulating three-day conference.

A randomised trial of this kind might involve a barista producing both regular coffees and decaf coffees. We might pick the coffee blends so that decaf and regular taste as similar as possible. Each time a person walks up to the coffee stand, the barista tosses a coin. Heads, you get a regular coffee. Tails, you get a decaf coffee. The law of large numbers tells us that if we did this experiment with a sufficiently large group, we would end up with roughly half in the heads group, and half in the tails group.

Before taking a sip of the coffee, the heads and tails groups would be similar in every way. With a large number of people, we can reasonably expect that the groups will include a similar number of men and women, a similar number of younger and older people, a similar number of morning larks and night owls. As a result, if we observe differences in alertness between the two groups, then we know that it must be due to the caffeine. We could conclude from this that caffeine keeps people awake, at least for the kinds of people in the room.

What would have happened without randomisation? What if we allowed everyone to ask the barista for regular or decaf, and then tracked the alertness of both groups? How might an observational study turn out differently?

In this case, an observational study would be plagued by selection effects. Those who chose the caffeinated drink might have been the kinds of people who prioritised alertness. Or maybe caffeine consumers were extra tired after a big night. Without a credible counterfactual, the observational data would not have told us the true effect of caffeine on performance. We would have learned a lot about the kinds of people who chose caffeinated drinks, but very little about the true impact of caffeine.

The problem with observational studies isn't just an academic curio. In medicine, researchers using observational data had long observed that moderate alcohol drinkers tended to be healthier than non-drinkers or heavy drinkers. This led many doctors to advise their patients that a drink a day might be good for your health.

Yet the latest meta-analyses, published in the *Journal of the American Medical Association*, now conclude that this was a selection effect (Zhao et al., 2023). In some studies, the population of non-drinkers included former alcoholics who have gone sober. Compared with non-drinkers, light drinkers are healthier on many dimensions, including weight, exercise and diet. Studies that use random differences in genetic predisposition to alcohol find no evidence that light drinking is good for your health (Biddinger et al., 2022). A daily alcoholic beverage isn't the worst thing you can do, but it's not extending your life.

The problem extends to just about every study you've ever read that compares outcomes for people who choose to consume one kind of food or beverage with those who make different consumption choices. Health writers Peter Attia and Bill Gifford point out that 'our food choices and eating habits are unfathomably complex', so observational studies are almost always 'hopelessly confounded' (Attia & Gifford, 2023, p. 300).

A better approach is that adopted by the US National Institutes of Health, which is conducting randomised nutrition studies. These require volunteers to live in a dormitory-style setting, where their diets are randomly changed from week to week.

Nutritional randomised trials are costlier than nutritional epidemiology, but they have one big advantage: we can believe the results. They inform us about causal impacts, not mere correlations.

Indeed, a clever experiment with mice has shown how problematic nutritional epidemiology can be. The study, led by Keisuke Ejima (Ejima et al., 2016), starts off with a randomised experiment on calorie restriction and longevity. By randomly varying the amount of calories given to different groups of mice, the researchers show that mice that are fed a calorie-restricted diet tend to live longer. This result replicated a well-established finding that calorie restriction boosts longevity.

Next, the researchers looked within those mice that had been allowed to eat as much as they wanted. Within that group, what was the association between calorie consumption and longevity? Now, the result flipped. Those mice that ate more calories lived longer – perhaps because they were doing more exercise or had faster metabolisms.

In case you're in any doubt that the observational study wasn't providing any insights, the researchers paired the free-eating mice with another group. The diet given to these rodents was the amount of food that their paired mouse had chosen to eat the previous day. Among the paired mice, the positive association between calories and longevity disappeared.

The observational study produced exactly the wrong result. As Peter Attia and Bill Gifford's work has observed, the complexity of what we choose to eat can confound any studies about the true effect of food and health.

The Australian Centre for Evaluation

The Australian Centre for Evaluation will not solely conduct randomised trials. But randomised trials will be an important component of the work of the Centre, which is why I have focused on them in this article. That said, there will be instances in which randomisation is considered unfeasible or impractical. This may include existing universal programs, where it would be difficult to induce differences in participation rates. Programs that involve multiple interventions, or are highly tailored to local conditions, may not be suitable for randomised evaluation. In other instances, researchers may defer to community preferences for alternative evaluation methodologies. In certain cases, the question may be one of implementation fidelity rather than impact, in which case a process evaluation may be most appropriate.

The Australian Centre for Evaluation will be located in the Australian Treasury, and will partner with other government agencies to conduct rigorous evaluations. The Centre receives funding of about $2 million per year and employs around 14 staff. Its work will be conducted within a careful ethical framework, ensuring that the Centre is as rigorous about issues of ethics as about issues of causality.

The Australian Centre for Evaluation will not conduct all the evaluations of government programs and policies. Given its size, that would be impossible. At present, the volume of external evaluations is over $50 million a year (Australian Evaluation

Society, 2023), and many agencies have their own in-house evaluation teams. The Australian Centre for Evaluation will partner on a modest number of flagship evaluations, and work to build capacity across the public service on rigorous evaluation.

An important part of the process will be to ensure that evaluations are not unnecessarily expensive. Over recent years, the response rate in government surveys has fallen, while the quality of administrative data has risen. In this environment, researchers are looking for opportunities to see how evaluations can make better use of data that is already held by the government, while maintaining strict privacy protections. Randomised trials need not take decades and cost tens of millions of dollars. Low-cost randomised trials can produce rapid insights at a modest cost (for more examples, see Leigh, 2018).

The Australian Centre for Evaluation will be characterised by its openness. The choice of name is deliberate: this will be a centre for high-quality evaluation nationally. This means that the Centre will be open to engaging as appropriate with states and territories, with non-profits and philanthropic foundations, and with evaluation experts in the private sector and academia. Policymakers and researchers are all on the evaluation journey together, and engagement will help to build the nation's evaluation capacity.

Conclusion

I began by focusing on how randomised trials have helped transform medicine. While randomised trials of pharmaceuticals have been commonplace since the 1950s, randomised surgical trials are still relatively rare. Likewise, the advent of randomised nutritional trials is still in its infancy. In these fields and others, high-quality evidence is helping to displace misplaced dogma, based on low-quality observational evidence.

Over coming years, developments in medicine may have applications in policy. Following Phase 1 safety trials, health researchers typically conduct two phases of clinical trials: Phase 2 trials, on a relatively small population, and Phase 3 trials, on a larger population. The notion of replicating the evaluation before going to market is highly relevant in social science.

Over the past decade, the 'replication crisis' in social science has seen several high-profile findings debunked. A 2015 study led by psychologist Brian Nosek looked at a hundred studies published in top journals and found that they could replicate only one in three (Open Science Collaboration, 2015). Over recent years, a number of top researchers, including food researcher Brian Wansink, behavioural scientist Francesca Gino and psychologist Dan Ariely, have been accused of fabricating data for their studies. Ironically, a study of dishonesty among car dealers appears to have been based on fake data. Single studies with surprising findings that are published in top journals have had too much impact on how we view the world (Ioannidis, 2005). Researchers and policymakers need to do a better job of building in replication in social science – just as Phase 2 and Phase 3 trials do in the clinical world.

Another feature of the health evidence ecosystem that could be adopted in policy is the notion of the living evidence review. A decade ago, Julian Elliott, Professor of Evidence Synthesis at Cochrane Australia, developed the notion of a living evidence review – a systematic review that is updated in real-time, as new studies are published (Elliott et al., 2021). When COVID-19 hit, Elliott worked with clinicians and researchers from around Australia to produce the National COVID-19 Clinical Evidence Taskforce (archived at clinicalevidence.net.au), a set of living evidence syntheses and recommendations that were updated every week to collate high-quality evidence on everything from how best to treat new coronavirus variants to the impact of masks on the risk of transmission (Global Commission on Evidence, 2022). In policy, the availability of living evidence syntheses would help decision makers identify the most relevant research, and avoid the danger of being swayed by a single low-quality study.

Strengthening the national evidence infrastructure, so that all programs and policies are rigorously assessed for their effectiveness and impact, and evaluation evidence is routinely synthesised and made publicly available, will take time. The Australian Government has taken an important first step in establishing the Australian Centre for Evaluation and looks forward to working with evaluation experts to see high-quality evaluation evidence placed at the heart of policy design and decision-making.

Authors' note

Andrew Leigh is the Assistant Minister for Competition, Charities, Treasury and Employment. This is an edited version of a keynote address delivered to the Australian Evaluation Society's International Evaluation Conference in Brisbane on 29 September 2023.

Acknowledgements

My thanks to editor John Guenther for valuable comments and corrections on an earlier draft.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Andrew Leigh https://orcid.org/0000-0002-5639-0509

References

Attia, P., & Gifford, B. (2023). *Outlive: The science and art of longevity*, Harmony.

Australian Evaluation Society. (2023). *The state of evaluation in Australia: A study of current evaluation volume, drivers, approaches, and trends in Australia*. Australian Evaluation Society.

Biddinger, K., Emdin, C., Haas, M., Wang, M., Hindy, G., Ellinor, P., Kathiresan, S., Khera, A., & Aragam, K. (2022). Association of habitual alcohol intake with risk of cardiovascular disease. *JAMA Network Open*, *5*(3), e223849–e223849. https://doi.org/10.1001/jamanetworkopen.2022.3849

Ejima, K., Li, P., Smith, D. L. Jr., Nagy, T. R., Kadish, I., van Groen, T., Dawson, J. A., Yang, Y., Patki, A., & Allison, D. B. (2016). Observational research rigour alone does not justify causal inference. *European Journal of Clinical Investigation*, *46*(12), 985–993. https://doi.org/10.1111/eci.12681

Elliott, J., Lawrence, R., Minx, J. C., Oladapo, O. T., Ravaud, P., Tendal Jeppesen, B., Thomas, J., Turner, T., Vandvik, P. O., & Grimshaw, J. M. (2021). Decision makers need constantly updated evidence synthesis. *Nature*, *600*(7889), 383–385. https://doi.org/10.1038/d41586-021-03690-1

Fisher, B., Wolmark, N., Redmond, C., Deutsch, M.., & Fisher, E. R. (1981). Findings from NSABP protocol no. b-04: Comparison of radical mastectomy with alternative treatments. II. The clinical and biologic significance of medial-central breast cancers. *Cancer*, *48*(8), 1863–1872. https://doi.org/10.1002/1097-0142(19811015)48:8<1863::AID-CNCR2820480825>3.0.CO;2-U

Global Commission on Evidence to Address Societal Challenges. (2022). *The evidence commission report: A wake-up call and path forward for decisionmakers, evidence intermediaries, and impact-oriented evidence producers*, McMaster Health Forum. https://www.mcmasterforum.org/networks/evidence-commission

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.1004085

Leigh, A. (2018). *Randomistas: How radical researchers changed our world*, Black Inc.

Mukherjee, S. (2010). *The emperor of all maladies: A biography of cancer*, Simon and Schuster.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Otto, M. (1994). Cancer Researcher No Longer The Hero: Bernard Fisher Was Fired Amid Allegations Of Faulty Research. *Philadelphia Inquirer*. July 7, 1994.

Wootton, D. (2006). *Bad medicine: Doctors doing harm since Hippocrates*, Oxford University Press.

Wright, J. R. Jr. (2018). The radicalization of breast cancer surgery: Joseph Colt Bloodgood's role in William Stewart Halsted's legacy. *Bulletin of the History of Medicine*, *92*(1), 141–171. https://doi.org/10.1353/bhm.2018.0006

Zhao, J., Stockwell, T., Naimi, T., Churchill, S., Clay, J., & Sherk, A. (2023). Association between daily alcohol intake and risk of all-cause mortality: A systematic review and meta-analyses. *JAMA Network Open*, *6*(3), e236185–e236185. https://doi.org/10.1001/jamanetworkopen.2023.6185