BMJ Quality & Safety

# How to attribute causality in quality improvement: lessons from epidemiology.



| | |
|---|---|
| Journal: | *BMJ Quality & Safety* |
| Manuscript ID | bmjqs-2017-006756.R1 |
| Article Type: | Viewpoint |
| Keywords: | Quality improvement, Health services research, Evaluation methodology |
| | |

## BACKGROUND

Quality Improvement and Implementation (QI&I) initiatives face critical challenges in an era of evidence-based, value-driven patient care. Whether frontline staff, large organisations, or government bodies design and run QI&I, there is increasing need to demonstrate impact to justify investment of time and resources in implementing and scaling up an intervention.

Decisions about sustaining, scaling up, and spreading an initiative can be informed by evidence of causation and the estimated attributable effect of an intervention on observed outcomes. Achieving this in healthcare can be challenging, where interventions often are multi-modal and applied in complex systems.[1] Where there is weak evidence of causation, credibility in the effectiveness of the intervention is reduced with a resultant reduced desire to replicate. The greater confidence of a causal relationship between QI&I interventions and observed results, the greater our confidence that improvement will result when the intervention occurs in different settings.

Guidance exists for design, conduct, evaluation, and reporting of QI&I initiatives;[2-4] SQUIRE and StaRI guidelines were developed specifically for reporting QI&I initiatives.[5, 6] However, much of this guidance is targeted at larger formal evaluations, and may require levels of resource or expertise not available to all QI&I initiatives. This paper proposes QI&I initiatives, regardless of scope and resources, can be enhanced by applying epidemiological principles, adapted from those promulgated by Austin Bradford Hill.[7]

## APPLYING BRADFORD HILL CRITERIA AND QI&I METHODS TO STRENGTHEN EVIDENCE

Hill proposed nine "aspects of association" that could be considered before "…*deciding that the most likely interpretation is causation*".[7] His objective was to improve the ability to form scientific judgments about causality. The nine aspects, subsequently referred to as the "Bradford Hill Criteria" (BHC), are considered in the following sections. With roots in causes of disease, the BHC have natural alignment with health care.[8] They can help make sense of causation in complex health care systems and, by extension, interventions to improve those systems. We posit that QI&I methods can be utilised to provide evidence towards meeting the criteria and infer causality. We offer a QI&I-oriented interpretation of the BHC, and match criteria with relevant QI&I methods (Table 1), and in the main text refer to the Michigan's Keystone Project to show the criteria in practice.[9]

**Table 1.** The Bradford Hill Criteria, epidemiological meaning, a translation for Quality Improvement and Implementation (QI&I) in italics, a brief description of QI&I methods that can provide evidence, and advice to practitioners.

| Criterion | Contribution of methods use in QI&I and advice to practitioners |
|---|---|
| **1 Strength of association**<br>• What is the size of the effect, i.e. what is the relative risk, or odds ratio?<br>• *What is the size of the effect, i.e. what is the efficacy of the intervention on outcomes of interest?* | Statistical process control (SPC) charts enable identification of special cause of variation that is unlikely to be due to chance alone, thereby providing statistical evidence of effect and its magnitude (measures of relative risk, number needed to treat, and estimates of the magnitude of attributable effect are useful measures of effect size).<br>• State the magnitude of change and its clinical or systems meaning<br>• Use SPC charts, with a clear rule set and control limits, to determine changes in process in advance, to maintain objectivity and avoid fishing for a result |
| **2 Consistency of association**<br>• Are repeated observations from different places, at different times, with differing methods, by different researchers, under different circumstances in agreement?<br>• *Does repeated application of intervention provide similar results in different contexts?* | Existing evidence contributes to programme theory and implementation plan, which can be used to demonstrate consistent impact, e.g., through scaling up.<br>• Keep track of intervention-outcome data as scale-up occurs to increase knowledge about causal consistency between intervention and outcome in different settings<br>• Analyse contextual barriers and enablers; make and note amendments to implementation plan |
| **3 Specificity of association**<br>• Is the outcome unique to the exposure?<br>• *Could anything else have produced the observed result?* | Combination of implementation design (e.g., step wedge), SPC charts, and other analyses of change, can inform the specificity of outcomes in relation to the intervention and planned implementation activities.<br>• Establish a comparison or control group, where possible, to identify secular trends (i.e., explore the counterfactual: what might have occurred without the intervention?)<br>• Ensure that the design and evaluation plan mitigate potential bias and confounding<br>• Explore what alternative mechanisms exist that might obtain the effect |
| **4 Temporality**<br>• Does exposure occur before the outcome?<br>• *Does intervention activity occur before outcome?* | SPC charts determine relationships between timing of intervention and observed special cause, Plan-Do-Study-Act (PDSA) cycles document QI&I activity.<br>• Annotate SPC charts with intervention events; include annotations of relevant external events apart from the planned intervention that could have influenced the outcome; make clear how and when special cause is detected and handled<br>• Ensure sufficient baseline data points to understand variation inherent in system<br>• Specify the predicted time period necessary to implement the intervention before improvement is expected to occur |
| **5 Biological gradient**<br>• As more of the stimulus is added, is the response increased?<br>• *Is more effect observed with more intervention, or higher fidelity of intervention?* | A combination of programme theory, implementation design and plan (e.g. step wedge), SPC charts, and other analyses for change, can examine the extent outcomes improve in relation to the intervention "dose" in planned programme activities.<br>• Demonstrate relationship between dose of interventions and outcome, using SPC charts or other analyses to display effect size<br>• In implementation plan design, consider the activities needed to deliver the "dose" |
| **6 Plausibility**<br>• Does the postulated causal relationship make sense?<br>• *Can the intervention explain the outcome?* | Programme theory and process maps should incorporate the plausibility that the intervention is likely to impact the outcome of interest. The implementation plan should consider the amount of intervention required to obtain a response, and statistical evaluations should reflect the degree of confidence in cause and effect.<br>• Draw on existing theories and models (e.g. behaviour sciences, implementation research) to determine the plausibility of the postulated QI&I initiative<br>• Observe how an intervention works in practice and link to PDSA cycles for testing theories; update programme logic in light of learning. |
| **7 Coherence**<br>• Is the association compatible with existing theory and knowledge?<br>• *As above* | Existing literature that demonstrates evidence for case (using knowledge from across disciplines) builds coherence.<br>• Conduct review of current knowledge (including grey literature and experience), assessing evidence regarding the effectiveness of implementation strategies |
| **8 Analogy**<br>• Does the causal relationship conform to a previously described relationship?<br>• *Are there other interventions in different settings that are similar?* | Learning from other improvers and researchers; for instance a similar intervention (e.g. 'care bundle') in one setting has analogy to another.<br>• Find analogies in existing literature to increase confidence that similar approaches will work elsewhere even if the specific intervention or implementation strategy differs |
| **9 Experiment**<br>• Does controlled manipulation of the exposure variable change the outcome?<br>• *Does modification of the intervention provide difference in outcome?* | Programme theory highlights areas for implementation activity. The implementation design should mitigate confounding and bias where possible. PDSA cycles can be used to experiment, recognising that multiple changes may be required.<br>• Test changes using iterative PDSAs along the theoretical causal pathway to build confidence in cause-effect<br>• Document predictions, what changes were made and why; reflect on accuracy of predictions and determine new information gained; update programme theory |

**1 Strength of Association**

QI&I initiatives aim to achieve meaningful impact from the perspective of the delivery system, providers, and patients. The size of change is important. This can be detected using Statistical Process Control (SPC) charts with appropriate control limits,[10] and measures of effect size (e.g. relative risk). Hill did not consider the p-value to be as important for establishing cause, stating: "*Such tests can, and should, remind us of the effects that the play of chance can create, and they will instruct us in the likely magnitude of those effects. Beyond that they contribute nothing to the 'proof' of our hypothesis*." [7] Significant p-values have little meaning if changes are trivial in a clinical or systems improvement sense. The potential over-emphasis on p-values *per se* has been highlighted.[11] Yet, the more speculative the postulated cause-effect relationship, the more stringent should be the design and evaluation (e.g. mitigation for bias and confounding, consideration of the counterfactual, and determination of effect sizes), and the strength of evidence for the intervention.

**2 Consistency of Association**

Confidence in the causal nature of a relationship between stimulus (e.g. intervention) and response (e.g. health outcomes) is increased if the association is demonstrated in multiple studies or projects in diverse contexts and conditions. To determine consistency of association, there is a need to account for context, including differentiating the hard "core" from the adaptable "peripheral" components.[12] Programme theory and implementation plans should incorporate study and documentation of contextual factors to facilitate scale-up of interventions

**3 Specificity of Association**

Specificity is established when a presumed cause produces a specific effect, asking if the outcome is unique to the exposure. The assessment of alleged impact of a QI&I intervention on outcomes needs to consider bias, confounding, and trends unrelated to the intervention. There is a premium on having a comparator not exposed to the intervention of interest (e.g. the counterfactual). If this is impractical or unethical, there should be sufficient baseline and follow-up data to support a claim of temporality. As noted by The Cochrane Effective Practice and Organisation of Care (EPOC) Group, the level of confidence regarding causality is dependent on the degree to which QI&I initiatives address bias and confounding in design, evaluation and analysis, and publication.[3] EPOC provides tools, such as a bias checklist, to mitigate these problems.[3] A strong implementation and evaluation plan coupled with examination for specificity of changes using SPC and other analysis, in light of the planned QI&I interventions, would provide evidence for specificity.

**4 Temporality**

Changes in outcome taking place before an intervention starts are not caused by that intervention. QI&I is well prepared in this regard; temporality is demonstrable by SPC charts, with annotations.[9] When demonstrating temporality, a comparison group is desirable to ensure that changes attributed to the intervention are not secular trend. Baseline data are required to detect change in the system performance. Other interventions or events occurring concurrently to the QI&I intervention should be documented, as they confound the observed association. The anticipated lag time for the intervention to "kick in" should be specified and displayed on time-series graphs.

**5 Biological Gradient**

Often referred to as "dose-response", this is the relationship between exposure to stimulus (amount of input) and outcome (degree of change in the outcome of interest). In QI&I this is relevant as one

should measure the "dose" of an intervention, and whether the intended "dose" was delivered to intended recipients with reliability (all individuals receive an intervention) and fidelity (all components of the intervention are received by an individual). We can consider the "dose-response" on individuals, e.g. direct effect of a care bundle on a patient's health, or on organisations, e.g. the resources required to implement an intervention, the number of people treated, the modifications of the intervention between and within settings, and the population and health economy outcomes.

If an intervention is delivered with diminished "dose", the magnitude of improvement would expected to be reduced. Specification of the "dose" is important in an implementation plan, e.g. in designing a post-operative care plan for patients with hip arthroplasty, the number of nurse home visits and the intensity (e.g. length of visit and quality of interaction) of the nurse's activities could be considered "dose", with speed and magnitude of recovery the "response". Determination of "dose" can be complex when the intervention is multi-modal and interactions among the various elements are difficult to estimate. Programme theory and implementation plans should provide logic for expected gradients,[13,14] demonstrable on SPC charts annotated with key changes in implementation practices based on Plan-Do-Study-Act (PDSA) cycles.[10,15]

**6 Plausibility**

Plausibility requires a credible rationale as to why an intervention might have a specific outcome. It does not imply certainty, and if a claim of cause and effect generates incredulity, a more robust design should be considered in the programme theory and implementation plan. The investigation of an initially implausible premise would require strong evidence for an association through an escalation of confirmatory studies, starting with proof-of-concept studies, then increasing the size and scope of studies as confidence in programme theory grows.[13] For example, is it plausible that one educational seminar would improve colon cancer screening in primary care? Or is it more plausible to incorporate computer alerts, a behavioural economics "nudge," patient education, and feedback of screening rates from comparable primary care practices?

Formal approaches to displaying programme theory incorporate assessments of "plausibility" in predicting the attributable effect on desired outcomes of implementing specific activities. Process maps can be used to understand the processes and mechanisms of care requiring alteration to change outcomes, thereby contributing toward plausibility.

**7 Coherence**

Coherence is concerned with the alternative theories we need to reject to find an idea plausible: does the observed effect conform to expectations, and can variations to those expectations be explained rationally? In QI&I practice, we should ask what other potential mechanisms would be rejected before accepting the QI&I intervention has coherence. If a claim exhibits greater congruence with existing knowledge it should be preferred. Programme theory, drawn from existing literature, considers coherence and should highlight any alternative explanations for any observed improvements.

**8 Experiment**

This criterion asks whether deliberate alterations to a system result in changes in outcome. Numerous designs are available (e.g. step-wedge, factorial, cluster randomised control trials, and Bayesian adaptive cluster randomized trials),[16, 17] providing differing levels of confidence in attribution depending on the design's mitigation of bias and confounding. The choice of a specific design may be dictated by practical or ethical considerations.

QI&I exploits iterative experiments: PDSA cycles test changes that improvers hypothesise are contributory to achieving the outcome of interest. The cycle of prediction, measurement, analysis, and revision of the intervention based on the analysis is fundamental in QI&I. If processes of care rather than outcomes are measured (because the outcomes are rare or likely to be delayed) there should be strong evidence that targeted processes are linked to those outcomes. Iterative successful PDSAs build confidence that a theorised causal pathway is correct and that improvement in outcomes is attributable to the implemented changes.

**9 Analogy**

Analogy is related to plausibility and coherence, allowing inference to be drawn from related studies and learning from other settings. Hill used as an example: *"In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy?"*.[7]

Improvers could establish analogy by reviewing the literature for related initiatives and plumbing the experience of the QI&I community.

**The Criteria In Practice: Considering Michigan's Keystone Project**

The Michigan Keystone Project sought to reduce catheter based infection. [9] The interventions around disinfection were based on strong evidence, providing **plausibility**. The initial project evaluation shows a **strength of association**: statistically significant effects ($p<0.002$), with relatively large and meaningful sizes (e.g. median 2.7 infections per 1000 days to 0).[9] There was **consistency of effect** with observed reductions in 103 centres. Whilst analyses included time variables, appropriately annotated SPC could have increased confidence in **temporality**, and a comparative arm would have increased confidence in **specificity**. During the initial keystone project the cause and effect relationship of the intervention appeared to have **coherence**, and no alternative explanations were presented for the observed effects. However, in "*Matching Michigan*" and a post hoc evaluation of Keystone, [18,19] it became apparent that the explanation was not fully **coherent**, and that alternative explanations for the success of the intervention existed: the "**dose**" and its delivery were more complex than initially described. This demonstrates how applying the BHC to knowledge gained over time can build or question confidence in causality: in this case, coherence for the programme theory seemed strong at first, with attempts to replicate diminishing that confidence. For the Michigan study, the causality question remains contested as the post hoc theorisation for what constitutes the intervention needs to be applied in practice to determine if it is sufficient for reproducing the desired impact.

**CONCLUSION**

The BHC provide an epidemiological approach to imputing causality in QI&I initiatives. These criteria are compatible with scientific improvement methods and, if properly utilized, QI&I methods can provide evidence towards each criterion. Pragmatic amendments to the BHC are permissible (e.g. combining plausibility and coherence). Refinements to the BHC will be desirable as new scientific advances provide insights into mechanisms by which interventions influence outcomes.[20] Further, the BHC should not be considered "the letter of the law". As Hill stated: "*None of my nine viewpoints* [criteria] *can bring indisputable evidence for or against the cause and effect hypothesis and none can be required as a sine qua non"*.[7] We contend these criteria, in their totality, can build confidence toward causality, and should not be a "checklist" in which every element must be checked for a study to be deemed credible. Yet, lack of temporality would raises a concern, and an implausible intervention would suggest the other criteria need to be addressed with rigour. Apparent conflicts

between criteria should be weighed in reaching a judgment, analogous to how judgment is reached in law courts: is the evidence, in its totality, proof beyond reasonable doubt? For instance, were there weak plausibility and limited specificity, attribution may be tenuous, despite apparent statistical association. Using techniques for grading the quality of evidence, e.g. GRADE, [21] could help this judgment.

A causal relationship can inform a decision to scale-up an intervention: the magnitude of the impact, the number of beneficiaries, the overall cost (time and resources) to the health care delivery system and society, and the policy environment are important considerations.

Improvers could ask if Hill were to examine their QI&I initiative, would he willingly state the results are attributable to the implemented interventions. The BHC can provide a lens through which improvers can gain "casual confidence" in their initiatives. Achieving an "exemplary" causal confidence would require a plausible programme theory specifying a causal pathway to improving measurable processes and outcomes. There would be a sound implementation plan incorporating real-time learning from iterative PDSA tests with bias and confounding addressed during design and evaluation. The timing of improvements in relation to implementation would be clear, with dose-response to interventions. Alternative explanations for the observed effect would be explored. Similar interventions would be successful in other settings. Contextual factors accelerating or impeding the intervention would be presented to enhance the likelihood that replication of core elements of the intervention, with adaptation to local context, would lead to improvement.

In short, improvers can leverage epidemiology and improvement science to maximize causal confidence when attributing interventions implemented to results observed.

## REFERENCES


1 Ladyman J, Lambert J, Weisner K. 2013. What is a complex system? *European Journal for Philosophy of Science* 3:33. Doi:10.1007/s13194-012-0056-8

2 AcademyHealth. 2016. Evaluating Complex Health Services Interventions: Challenges, Goals, and Proposals for Progress. Available online [accessed 2017-03-01] http://www.academyhealth.org/files/AH_Report_Evaluating_Complex_HSI_June2016_FINAL.pdf

3 Effective Practice and Organisation of Care (EPOC). 2015. *EPOC Resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services*. Available online [accessed 2017-03-01] http://epoc.cochrane.org/epoc-specific-resources-review-authors

4 Chowfla A. 2016. *Salzburg Global Seminar. Better Health Care: How do we learn about improvement?* Available online [accessed 2017-03-01] http://www.salzburgglobal.org/fileadmin/user_upload/Documents/2010-2019/2016/Session_565/SalzburgGlobal_Report_565__online_.pdf

5 Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. *BMJ Quality & Safety* 2016 25:986-992 http://qualitysafety.bmj.com/content/early/2015/09/10/bmjqs-2015-004411

6 Pinnock H, Barwick M, Carpenter CR, Eldridge S, Grandes G, Griffiths CJ, Rycroft-Malone J, Meissner P, Murray E, Patel A, Sheikh A, Taylor SJC. 2017. Standards for Reporting Implementation Studies (StaRI) Statement. *BMJ* 356.i6795. doi.org/10.1136/bmj.i6795

7 Hill AB. 1965. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58: 295-300

8 Howick J, Glasziou P, Aronson JK. 2009. The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? *Journal of the Royal Society of Medicine* 102: 186–194

9 Provonost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, Sexton B, Welsh R, Roth G, Bande J, Kepros J, Goeschel C. 2006. An intervention to decrease catheter related bloodstream infections in the ICU. The New England Journal of Medicine. 355: 2725-2732

10 Benneyan J, Lloyd R, Plsek P. 2003. Statistical process control as a tool for research and healthcare improvement. *Quality & safety in health care* 12(6): 458-464

11 Wasserstain RL, Lazar NA. 2016. The ASA's statement on p-values: Context, Process, and Purpose. *The American Statistician* 70(2) 129-133

12 Denis JL, Hébert Y, Langley A, Lozeau D, Trottier LH. 2002. Explaining diffusion patterns for complex health care innovations. *Health Care Management Review* 27(3): 60-73

13 Rogers P. 2008. Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions. *Evaluation* 14(1):29-48

14 Smith ME, Mourier P. 1999, Implementation: Key to organisational change. *Strategy & Leadership* 27(6):37-41

15 Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed J. 2013. Systematic Review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Quality & Safety* 23: 290-298

16 Campbell MJ, Donner A, Klar N. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Statistics in Medicine* 26(1) 2-19

17 Moen R, Nolan TW, Provost LP. 2012. *Quality Improvement Through Planned Experimentation*. 3rd Edn. McGraw-Hill Education. New York City, NY, USA.

18 Dixon-Woods M, Bosk CL, Aveling EL, Goeschel CA, Pronovost PJ. 2011. Explaining Michigan: Developing an Ex Post Theory of a Quality Improvement Program. *The Milbank Quarterly* 89(2):167-

205. doi:10.1111/j.1468-0009.2011.00625.x.

19 Bion J, Richardson A, hibbert P, Beer J, Abrusci T, McMutcheon M, Cassidy J, Eddleston J, Gunning K, Bellingan G, Patten M, Harrison D. Matching Michigan': a 2-year stepped interventional programme to minimise central venous catheter-blood stream infections in intensive care units in England. BMJ Qual Saf Published Online First: 20 September 2012. doi: 10.1136/bmjqs-2012-001325

20 Fedak KM, Bernal A, Capshaw ZA, Gross S. 2015. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology* 12:14 DOI**:** 10.1186/s12982-015-0037-4

21 Guyatt G, Oxman DA, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, deBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schuënemann HJ. 2011. Grade guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology 64:383-394