A Users' Guide to Interpreting Observational Studies of Pediatric Anesthetic Neurotoxicity:

The Lessons of Sir Bradford Hill

Randall P. Flick, M.D. and **David O. Warner, M.D.**
Department of Anesthesiology, College of Medicine, Mayo Clinic, Rochester, Minnesota

David O. Warner: warner.david@mayo.edu

At a meeting of the Royal Society of Medicine in 1965, Sir Bradford Hill proposed to answer the question he himself posed to the assembled audience: "How do we determine what are physical, chemical and psychologic hazards of occupation and in particular those that are rare and not easily recognized?"[1] Although the question Hill proposed was directed at problems of occupational medicine, the article subsequently published from this lecture has become the sentinel guide for the assessment of causation in epidemiologic research. Since the publication of the paper, more than a decade ago, in Science by Ikonomidou,[2] there has been a steady stream of work convincingly demonstrating that anesthetic agents and other drugs that act as *N*-methyl-D-aspartate agonists and *γ*-aminobutyric acid antagonists can produce widespread apoptotic neurodegeneration, with associated cognitive and behavioral decrements in a variety of animal species, including nonhuman primates. Predictably, these studies have prompted a great deal of concern. They have also generated a series of observational studies seeking evidence for similar effects in children, with varying results.[3–8] The study by Block *et al.* in this month's issue of Anesthesiology[9] is yet another examination of the association between exposure to anesthesia in young children and outcome, in this case performance on a test of academic achievement.

The central concern of those who provide anesthesia to children is that of causation: Does anesthetic exposure at a young age cause neurodevelopmental problems? In the hierarchy of study designs, the randomized control trial reigns as the gold standard. Unfortunately, such studies are expensive, time-consuming, and, in this area, may be ethically impossible. An ongoing randomized clinical trial comparing regional and general anesthesia for infants receiving inguinal herniorraphy will be valuable, but in the meantime, the anesthesia community is left to make judgments regarding the potential applicability of animal findings to children based on a growing number of retrospective observational studies. These studies provide insight, but how should they be interpreted? For example, in studies that find an association between anesthesia and subsequent neurodevelopmental problems, is anesthesia merely a marker for another causative factor (*e.g.*, the stress of a surgical procedure, or the underlying condition which makes surgery necessary)?

We here present a "users' guide" of several questions that should be considered when interpreting observational studies of the association between anesthetic exposure and neurodevelopmental or other outcomes. This discussion is by no means exhaustive, but is meant to alert consumers of literature in this area regarding some of the potential strengths and weaknesses of available and to-be-published studies. This guide should not be seen as criticism of Block *et al.* or others authors who have contributed to this body of work. Indeed, as nicely typified by Block *et al.*, most authors spend a great deal of time injecting caution into the discourse and highlighting study limitations.

What is the population receiving anesthesia?

The population of children exposed to anesthesia may be chosen as those anesthetized at a single medical facility (as in the Block study), those covered by a defined payment mechanism (*e.g.*, Medicaid recipients), or those within a given geographical area. The population may be further refined as only those receiving selected types of procedures. Examining children anesthetized at a single center allows for precise definition of the conditions of exposure, but has the potential for referral bias, as these are usually tertiary care centers that typically care for more complicated patients. Geographical population-based studies and large payer-based designs are potentially attractive, but may or may not be not representative of a more general population.

Who is actually included in the analysis?

Not every child in a target population is analyzed, regardless of how the population is defined. Children and families may decline to participate, may be lost to follow-up for a variety of reasons, or may not have received the outcome assessment. This is a potential source of bias, which can be important in studies like Block *et al.*, in which the minority of those contacted agreed to participate. For example, in general those with lower levels of educational attainment are less likely to agree to participate in clinical research, and parental attainment is correlated with child performance. Also, if school achievement tests are the outcome, children with impairment sufficient that they cannot complete the tests will not be included; if anesthesia caused the impairment, this is an obvious source of bias. For geographically based studies, children may migrate from the region. This potential problem is common to many clinical studies, and authors should address this concern, typically by attempting to show that those population members included are similar in defined respects to those not included (as was nicely done by Block *et al.*).

What is the definition of anesthetic exposure?

Although there is clearly a window of vulnerability to anesthesia exposure in animal studies, it is not clear how to translate these age ranges in animals to comparable developmental stages in children. Suggestions from various authors range from the neonatal period up through 4 yr old, and the extant observational studies encompass this range in their definitions of exposures. There is also wide variation between studies in what is known about the exposures. Some (like Block *et al.)* have access to detailed information regarding the anesthetics and procedures, whereas others utilize procedure codes to define exposure, which indicate only that some type of anesthesia was probably administered (*e.g.*, it may not be possible to determine whether general or regional techniques were used for inguinal herniorrhaphy). Clearly, if the definition of exposure differs, or exposure is not well-defined, it can be difficult to compare studies.

What is the comparison group (those not exposed)?

Every study must compare children exposed to anesthesia with those who are not. One option (used by Block *et al.)* is to use the rest of the chosen population who do not meet the exposure definition (such as all other children receiving achievement tests, covered by a payer, in the geographical region, and so forth). This has the advantage of usually including very large numbers, but also may "contaminate" the group with children who do not meet study exposure criteria but may nonetheless be at risk (misclassification bias). For example, a child receiving herniorrhaphy at a University of Iowa hospital at 366 days of age would be included in the comparison group in the Block study, as would an infant receiving anesthesia for a procedure not included in the exposure criteria. In addition, all children in Iowa who were not cared for at the University of Iowa would also be included. Given that the

population frequency of procedures in children younger than 1 yr old is relatively low, this may not be a practical concern, and would bias toward finding no differences. Another option is to match exposed children with unexposed children based on some criteria (such as age, sex, burden of illness, or even siblings or twins) to control for characteristics that may be associated with outcomes (genetic, social, environmental). With any of these approaches, the concern is whether the comparison group is truly comparable, or whether there are – usually unmeasured – confounders that may bias comparisons.

What is the outcome?

Three types of outcome measures have been employed in extant studies: group-administered achievement tests, individually administered tests, and diagnostic codes potentially indicative of neurodevelopmental anomalies. In general, group-administered tests of cognitive abilities and achievement (such as used by Block *et al.)* serve as screening tests, whereas individually administered tests are used to make clinical diagnoses (*e.g.*, learning disabilities), because group-administered tests do not allow for in-depth observations of individuals as they complete the test. Individually administered tests provide direct, one-on-one interaction between an examiner who has control of the test environment and can directly observe behavior during testing. They also provide a broad sample of neuropsychological abilities (*i.e.*, sample a broader range of domains that could be potentially affected by any anesthetic-induced neurotoxicity), and thus provide detailed information that cannot be obtained from group-administered test data. The primary advantages of group-administered tests are that they are readily available in relatively large and diverse populations with well-defined population norms. Conversely, individual achievement tests tend only to be available in smaller, less-representative populations. The use of diagnostic codes to ascertain potential impairment allows investigators to take advantage of very large administrative data sets, but has obvious limitations in regards to ascertaining neurodevelopmental anomalies.

Regardless of the outcome chosen, it is not always clear how it should be analyzed. For example, for test scores, are mean values more relevant, or the proportion of children who fall below a threshold? How should the threshold be chosen? The answer may depend on whether all children are affected, or whether there is a subset at particular risk. In addition, multiple exploratory analyses employing a variety of approaches may be useful, but performing multiple analyses increases the risk that spurious associations will be found. As the phenotype of any anesthetic-induced injury remains to be robustly defined (even in the animal studies), it is actually helpful that a wide variety of outcome measures are used in the various human studies, as this can help generate hypothesis for later evaluation in both animals and humans. However, it does make it difficult, or impossible, to compare the results of studies that use very different types of outcomes, much less different particular measures.

How are the data analyzed?

Two major issues need to be addressed: statistical power and confounding. Studies that extensively characterize a smaller number of children (such as those examining children exposed at a single center) pay the price of having more limited power, which may complicate the interpretation of "negative" results. Regarding confounders (*e.g.*, a variable factor that is correlated with both an independent variable – in this case, anesthetic exposure – and dependent variable, or outcome), neurodevelopment is a complex process that may be affected by a multitude of constitutional and environmental factors. Anesthesia is but one of the potential factors. Authors attempt to control for the potential confounding effects of these other factors using the information available through their particular study design, or

matching procedures, multivariate analysis, or both. Studies should be evaluated regarding how the authors attempted to control for these confounders, although it is important to remember that any analysis can control only for what is known and measured; relevant factors may be unknown, difficult or impossible to measure, or unavailable in the dataset analyzed. Perhaps most importantly, the condition that makes anesthesia necessary, or other factors that necessarily accompany the anesthetic exposure (such as the stress of surgery), may themselves be causative, and it is almost impossible for observational studies to make this distinction.

What is the potential clinical relevance of any observed association?

Even for those studies that find "positive" results, is the strength of association clinically relevant? At this point we return to causality, strength of association and Sir Bradford Hill. In his president's address, Sir Hill stressed the importance of strength of association among the factors of greatest importance when determining when an observed association may be considered to be causal. He referred to the classic case of chimney sweeps and scrotal cancer – their risk of cancer was approximately 200 times that of workers not so employed. A more recent example is the association of salicylate and Reye syndrome, in which the relative risk was 26 times greater than the risk observed in any of the observational studies of anesthetic neurotoxicity where most positive studies report hazard ratios of less than 3.[10] For rare events, hazard ratios in this range are often associated with confounding, and should make one suspicious of the association as causal. However, even small hazard ratios may be important in the setting of a common exposure such as anesthesia and relatively common outcomes such as learning disabilities. For example, Flick *et al.* found a hazard ratio of approximately 2 for the association between exposures to multiple anesthetics and the later diagnosis of a learning disability. With approximately 1 in 5 unexposed children experiencing a learning disability, this means that for every six children exposed to multiple anesthetics, one additional child will develop a learning disability. If true, the public health impact of this would be enormous, even though the hazard ratio is not high.

Ultimately, we must heed the lessons of Sir Hill and use great caution when interpreting the observational studies that describe the relationship between anesthetic exposure and learning, behavior, and cognition. Sir Hill described several other factors important to ultimately establish a causative relationship, including consistency (results from study to study that are similar in direction and magnitude), specificity (positive results that are similar from study to study, defining a phenotype), temporality (with exposure reliably preceding the outcome), biologic gradient (a dose-response relationship), plausibility (findings consistent with what is known in animal models or similar human conditions), coherence (an effect fitting an observed pattern in different populations), experiment (the effect being mitigated in a controlled experimental settings), and analogy (similar effects seen with analogous exposures). We have only begun to explore some of these factors. At this point, all that can be concluded is that the available human studies (including that of Block *et al*.) cannot exclude the possibility that the anesthesia-induced neurotoxicity observed in many animal studies may also occur in children – Sir Hill will allow us to go no further. The history of medicine is replete with examples of putative causative relationships detected in observational studies that have proven to be spurious. However, observational studies have also proved seminal in establishing causative relationships of great clinical significance, such as tobacco-related disease and sudden infant death syndrome. Only time and considerable additional effort will determine the factor(s) that are responsible for the associations observed by Block *et al*.

Acknowledgments

References

1. Hill AB. The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine. 1965; 58:295–300. [PubMed: 14283879]

2. Ikonomidou C, Bosch F, Miksa M, Bittigau P, Vöckler J, Dikranian K, Tenkova TI, Stefovska V, Turski L, Olney JW. Blockade of NMDA receptors and apoptotic neurodegeneration in the developing brain. Science. 1999; 283:70– 4. [PubMed: 9872743]

3. Flick RP, Lee K, Hofer RE, Beinborn CW, Hambel EM, Klein MK, Gunn PW, Wilder RT, Katusic SK, Schroeder DR, Warner DO, Sprung J. Neuraxial labor analgesia for vaginal delivery and its effects on childhood learning disabilities. Anesth Analg. 2011; 112:1424–31. [PubMed: 20736436]

4. Flick RP, Katusic SK, Colligan RC, Wilder RT, Voigt RG, Olson MD, Sprung J, Weaver AL, Schroeder DR, Warner DO. Cognitive and behavioral outcomes after early exposure to anesthesia and surgery. Pediatrics. 2011; 128:e1053– 61. [PubMed: 21969289]

5. DiMaggio C, Sun LS, Kakavouli A, Byrne MW, Li G. A retrospective cohort study of the association of anesthesia and hernia repair surgery with behavioral and developmental disorders in young children. J Neurosurg Anesthesiol. 2009; 21:286–91. [PubMed: 19955889]

6. DiMaggio C, Sun LS, Li G. Early childhood exposure to anesthesia and risk of developmental and behavioral disorders in a sibling birth cohort. Anesth Analg. 2011; 113:1143–51. [PubMed: 21415431]

7. Hansen TG, Pedersen JK, Henneberg SW, Pedersen DA, Murray JC, Morton NS, Christensen K. Academic performance in adolescence after inguinal hernia repair in infancy: A nationwide cohort study. Anesthesiology. 2011; 114:1076– 85. [PubMed: 21368654]

8. Bartels M, Althoff RR, Boomsma DI. Anesthesia and cognitive performance in children: No evidence for a causal relationship. Twin Res Hum Genet. 2009; 12:246–53. [PubMed: 19456216]

9. Block RI, Thomas JJ, Bayman EO, Choi JY, Kimble KK, Todd MM. Are anesthesia and surgery during infancy associated with altered academic performance during childhood? Anesthesiology. 2012; 117:494–503. [PubMed: 22801049]

10. Hurwitz ES, Barrett MJ, Bregman D, Gunn WJ, Pinsky P, Schonberger LB, Drage JS, Kaslow RA, Burlington DB, Quinnan GV. Public Health Service study of Reye's syndrome and medications. Report of the main study. JAMA. 1987; 257:1905–11. [PubMed: 3820509]