Drug-Induced Birth Defects
Exploring the Intersection of Regulation, Medicine, Science, and Law
An Educational Module

Prepared by

Nathan A. Schachtman
Lecturer in Law
Columbia Law School

For

Committee on Preparing the Next Generation of Policy Makers for Science-Based Decisions
Committee on Science, Technology, and Law

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

June 2016

Contents

Introduction: Birth Defects in Children Whose Mothers Used Medications in Pregnancy

Welcome to this short course on a topic of medical causation for policy makers. A popular television series from the early 1960s started with the solemn recitation and inscription of "man, woman, birth, death, infinity"— ♂, ♀ , ✳, †, ∞ (*Ben Casey*). This teaching module focuses on birth, and in particular with a pathology of birth—birth defects and their causes.

Over the next several days, we will explore a specific controversial causal claim that the use of antidepressants by pregnant women causes birth defects in their children. The issue whether maternal use of antidepressants causes birth defects is itself a topical issue of public health in current medical journals. The issue is scientifically complex, emotionally charged, and personally relevant to us all.

Our reason for looking at this issue, however, will go beyond this particular controversy, important as it is. The broader goal of the course is to develop and refine our general understanding and appreciation of the conceptual and scientific bases for reaching sound conclusions of medical causation, upon the available evidence. Issues of medical causation are commonplace in policy making, and sound policy requires being informed by sound causal conclusions. Even when causal conclusions cannot be reached on available evidence, policy makers will need to understand the nature and extent of the unavoidable uncertainty in making policy.

In this short course, we will focus on a claim of biological causation between a claimed cause, medication use, and a claimed effect, birth defects. The use of medications among women of child-bearing age raises serious, difficult, and perhaps intractable policy questions. These questions are fundamentally scientific and epistemic. How, when, and to what level of certainty can we determine whether maternal use of a medicine causes a particular birth defect, or several birth defects? As you work through the evidence, you should consider these issues in the context of the widespread use of antidepressants and whether this medication use causes or contributes to the population burden of birth defects.

The policy questions raised by widespread antidepressant use turn on evaluating evidence, of various strengths, which may weigh for or against a causal relationship between a mother's use of medication in pregnancy and her child's being born with a serious birth defect.

The module thus highlights a real, complex fact pattern that involves the interactions among the consuming public, the medical and scientific communities, the pharmaceutical industry, the regulatory agencies, and the judicial system. The issues arise as a result of concerns about the teratogenicity of a class of medications, the serotonin selective reuptake inhibitors (SSRIs), which are commonly indicated for depression and anxiety, and for which indications, they are currently first-line therapies The module will explore the difficulties created by uncertainty over the full safety profile of medications when they are initially marketed, the uncertainty in regulating them after marketing, and the uncertainty in adjudicating consumers' claims of harms. The module should help policy makers evaluate the scientific, regulatory, and legal issues that arise in the context of the pharmaceutical industry, as well as understand more generally the controversial and ambiguous concepts in assessing risks.

This course introduces some of the basic issues in medical causation, allowing students to understand debated controversial scientific policy issues. There are many components of causal assessment, including:

a. Animal toxicologic research and its relevance or lack thereof to the human situation.
b. Epidemiologic research, study design, and threats to study validity.
c. Basics of statistical and causal inference.
d. Integration of study results, both qualitatively and quantitatively, to support conclusions and decisions.
e. Assessment of risk and causation in the scientific, medical, regulatory, and legal communities.

This short segment cannot address all these components in depth. The module does focus on epidemiology, which is especially important because of the expectation that children of mothers will have birth defects, and heart birth defects in particular, regardless of whether the mothers have used any medication or have had any other suspected teratogenic exposures.

The module may also be used to impart a basic understanding of how scientific evidence is adduced, challenged, and received in legislatures, agencies, and in courtrooms, and the competence of legislators, regulators, and judges and juries to assess such evidence.

The module represents an abridgement of a much larger set of material, but it can be sampled at various levels of detail and complexity. To explore the implications of the problem will require participants to develop some familiarity with epidemiologic, statistical, and causal concepts, evidence synthesis, regulatory action, and perhaps with some basics of tort litigation and civil procedure. The module should be taught in at least five classes, with lectures, preparatory and follow-up readings, and discussions.

For the first session, the instructor will lead discussion of the students' base-line understandings of the initial reading materials, and will enrich the subject with lecture on the historical development of our understanding of causality from deterministic (with universal, necessary, and certain laws) to stochastic processes. This is a shift that certainly affects our contemporary understanding of many biological (and especially genetic) processes, but also physics, as in the Copenhagen interpretation of quantum mechanics.

ASSIGNMENT

The following two popular media articles help frame some of the issues we will be discussing in this short course. Please read both articles before our first session. They are both nontechnical introductions to medical causation, public health, depression, pregnancy risks, and risk-benefit analyses under conditions of uncertainty.

- Andrew Solomon, "The Secret Sadness of Pregnancy With Depression: Pregnant women often fear taking the antidepressants they rely on. But not treating their mental illness can be just as dangerous," *New York Times* (May 28, 2015), available at http://www.nytimes.com/2015/05/31/magazine/the-secret-sadness-of-pregnancy-with-depression.html.

- John P. A. Ioannidis, “An Epidemic of False Claims: Competition and conflicts of interest distort too many medical findings,” *Scientific American* (May 17, 2011), available at http://www.scientificamerican.com/article/an-epidemic-of-false-claims/.

CLASS ACTIVITY: DISCUSSION OF KEY CONCEPTS FROM THE SOLOMON AND IOANNIDIS ARTICLES

What do we mean when we say that a mother’s use of a medication causes her child to be born with a deformity?

[Instructors’ note: See discussion of causation, causal laws, below.]

[Are there meaningful differences between and among the use of various terms, such as *link*, *association*, *risk*, *risk factor*, *increased risk*, and *cause*?]

[See discussion below about links, risks, risk factors, and so on. In discussing the readings for this first segment, note the loose language of causality in Solomon’s article. The use of links is particularly typical of journalists’ approaches to scientific issues in that it avoids the more precise meanings of *association*, and *causation*.]

How do we decide whether the medication causes birth defects?

[This discussion should allow for a preliminary exploration of types and hierarchy of evidence, such as case reports, case series, analytical epidemiologic studies that involve control groups, and randomized clinical trials. See below for some additional ideas and suggestions.]

Who decides whether the medication causes birth defects?

[See notes below on stakeholders and on the multiple perspectives on risk. The instructor should help the students to identify the various stakeholders and to elaborate upon their differing competencies, understanding, and perspectives on potential and actual risk, the determination of causation, and causal attribution in a given case. These different stakeholders include patients’ (or consumers’), physicians and health care providers, insurers and prescription drug-plan managers, pharmaceutical companies, regulatory agencies, and the legal system.]

Do the beneficial effects of the medication affect our judgment about whether the medication causes a harm such as birth defects?

Are there varying degrees of certainty, accuracy, and validity required by different audiences for making causal claims?

What are the practical difficulties and barriers to arriving at consensus judgments about the causality of birth defects?

What are the ethical requirements and implications of describing risks accurately, for journalists, clinicians, scientists, regulators, and others?

[Note that Solomon quotes Dr. Urato as likening SSRIs to thalidomide. The controversial, uncertain risk ratios for cardiac birth defects and SSRIs are on the order of 50% increases in the background rate. The risk ratios for phocomelia among offspring of mothers who used thalidomide in early pregnancy are increased over a thousand fold.[1]]

What are the ethical obligations of scientists, journalists, publishers, legislators, and lawyers to refrain from overwrought, unsupportable analogies of this sort?

What is the physician's or other health care provider's ethical responsibility to discuss, at a minimum, birth control with a woman, of child-bearing years, who suffers from chronic depression?

If the young woman must take prescription medicine, of uncertain teratogenicity, who should bear the (known or potential) risk for an untoward pregnancy outcome?

Who are the stakeholders in the scientific controversy surrounding the ability of SSRI medications to causes birth defects?

[See notes below about the multiple perspectives of (1) the patients (or consumers), (2) the physicians or health care prescriber, (3) the pharmaceutical companies, (4) the regulatory agencies, (5) the tort bar.]

INSTRUCTORS' LECTURE NOTES

The targeted skills and understanding of medical causation issues obviously arise in the regulation, prescription, and reimbursement of medicines, vaccines, and medical devices. Increasingly, legislatures have waded into the practice of medicine to mandate that physicians and health care providers include disclosures of legislated "facts" in their informed consents, such as the risk of suicide or of breast cancer from abortion. Medical causation issues arise in the context of insurance coverage, disease-screening protocols, and resource allocation. Causation issues lie behind many occupational and environmental disease controversies. Although medical causation issues involving birth defects can be especially difficult, and emotional because of the harms involved, the general approach to understanding medical causation certainly has a much wider applicability to many issues.

Individual human survival depends upon learning mechanistic concepts of causality. Touching a burning stove invariably causes pain. Gravity causes objects to fall, and to fall on our heads if we are below them. Ultimately, we infer causal connections all around us. A billiard ball hitting another billiard ball at constant angle and speed, for instance, results in the struck ball invariably moving off in the same direction, and speed.

[1] Michael B. Bracken, Risk, Chance, and Causation Investigating the Origins and Treatment of Disease at 243 (2015) (relative risk about 1,500).

Biological causation, on the other hand, is more difficult to describe as invariable scientific "laws." Lung cancer occurs among people who never smoked tobacco, and many long-term smokers never develop lung cancer. Still, scientists observed that the incidence (the rate of new cases each year) at which lung cancer occurs in a population increases about two decades after the prevalence of smoking increases. Today, there is no doubt that smoking causes lung cancer, but the model of causation involved is obviously very different from the invariable experience with mechanistic models of causation, such as how we understand the movement of billiard balls. For tobacco-caused lung cancer, a person's smoking tobacco is neither necessary nor sufficient for that person to suffer the consequence. Nonetheless, on a population basis, there is no question that increasing the prevalence of smoking, of sufficient quantity and duration, will result in increased rates of lung cancer in the population.

The question whether SSRIs cause birth defects invokes a concept of biological causation in which the claimed causality is neither necessary nor sufficient. Many children with birth defects will be born to mothers who did not use the medication in question. Birth defects have been documented in human infants for thousands of years, and so, clearly, modern pharmaceutical interventions are not necessary for their causation. Not all mothers who use a particular medication, even one known to cause birth defects, will have children with a birth defect. And so, equally clearly, the use of even a known bad-acting medication is not sufficient to produce the birth defect outcome.

Biological causal processes that change rates of disease or other health outcomes present difficult policy problems in public health. First, our way of learning about causal connections from childhood has the potential to deceive us. We see a bad outcome following a suspect exposure, and we are tempted to infer causation simply because one followed the other. In biological processes, there is, however, often a baseline rate at which the bad outcome occurs, which means that the temporal relationship between some event, say, medication use, and bad outcome, say, a child born with a birth defect, is merely a coincidence. Indeed, given the baseline rate, there is, or should be, an expectation that there will be bad outcomes, regardless of medication use, or any other putative cause. The fallacy of inferring causation based upon temporal relationship is so prevalent that it has long been described by a Latin expression: *Post hoc, ergo propter hoc*. After this, therefore because of this.

Perhaps the biggest policy issue is how to resolve uncertain, emotional, or even specious claims about medication use, in the face of chance, systematic bias, and confounding, in an evidentiary dataset that is never quite as big as we would like.

Belief, Knowledge, and Certainty. This is an area in which the instructor can add some background understanding of exactly what we mean by "epistemic," in the tradition of western philosophy. Although there are some detractors, and claimed paradoxes, there is widespread agreement that "knowledge" consists of true, justified belief. So there may be instances of believing something that happens to be true, as a matter of dumb luck, but it is hard to envision this as a sound basis for policy, although blind squirrels sometimes find their nuts. And more important, something that is false has no claim on us as knowledge.

As noted above, the notion of causation may not involve an invariable outcome from maternal use, and the use of the medication may be neither necessary nor sufficient for producing the outcome. The causation of birth defects is complicated. Both mother and father contribute their genetic material to a fertilized egg, and both potentially contribute genetic causes of a manifested birth defect. This new organism may contain an inherited defect, or a defect may arise spontaneously. Furthermore, the parents' environment may change (cause mutations) in their genetic material. Some genetic causes of birth defects have been identified, but many remain unknown.

The environment of the fertilized egg, embryo, and fetus provide additional opportunities for interference with development, and the causation of structural birth defects. The mother's diet, lifestyle, overall health, exposure to viruses and bacteria, and exposure to chemicals and medications, all are considerations in that they may possibly affect the environment of the unborn.

INSTRUCTORS' NOTE

There are, of course, limits to how well we can quantify all maternal and paternal exposures, with respect to known causes of birth defects, such as folate levels in the maternal diet, or exposure to rubella and other viruses. As we will see, in the specific epidemiologic studies of SSRIs and birth defects, there is no effort to quantify or characterize paternal exposures, although these exposures may have potential effects as well.

Risk and Causation

The language of biological causation can be surprisingly imprecise. From one scientific perspective, risk is simply causation ex ante; that is, some causal factor (exposure or condition) that is present before the effect is observed. Because of the nature of biological causation, in many instances, a risk may exist without the effect's ever later occurring. Additionally, the outcome may occur after exposure to a risk, but the outcome occurs as a result of another independent risk or process.

There is another common use of the word *risk*. Less compelling evidence, or perhaps even no evidence at all, may lead us to think of exposures or conditions as risks in the sense that we do not know whether they are causally related to the outcome in question. Physicians and others certainly use risks this way, and they even suggest that they can manage to weigh the unknowable harm against a known benefit in order to reach a clinical decision about the appropriate therapy or surgery in their patients.

Communications about risk might be much more precise if we could dictate the meaning of a term such as risk by semantic fiat. There are other, competing uses of risk that come closer to meaning a possible or a potential risk. Sometimes, the term risk is qualified with one of these adjectives, but often authors expect their readers to understand that they are not making a causal claim, and sometimes authors actually are exploiting the ambiguity.

Risk Assessment

Risk assessment is a technical enterprise that attempts to put an upper boundary on the magnitude of human risk, and many risk assessments are published for exposures or conditions that have not been shown to be causes of the human diseases that are being “assessed.” Again, the use of “risk” here is technical in that there may be no risk at all, or even benefit, but there is a need to make some worst-case assumptions to prioritize regulatory or legislative activity.

Terminological Confusion

There are other examples of the scientific parlance’s being systematically misleading. Statisticians and epidemiologists refer to the measurement of an increased incidence or prevalence of disease in a sample of a population as the “effect size,” or as an “increased risk,” even though the study may be quite incapable of establishing that there is a (known) causal relationship or true risk at all.

Epidemiologists sometimes refer to exposures or conditions that are implicated as “increasing the risk” of an outcome as risk factors, but again their language does not necessarily connote causation. Statisticians and epidemiologists refer to the magnitude of a risk ratio, or risk difference, as an effect size, even though causality may still be an open question.

Scientists and journalists sometimes refer to suggestive findings between an exposure, to medication or other exposure, and an outcome, as links. This term seems to be favored as a way of avoiding clear language of causality, which has a more definite meaning. Nonetheless, a report of finding a link between a medicine and a birth defect sounds quite definitive to a lay audience, which raises ethical and professional questions about the degree of linguistic precision that scientists and journalists use to describe study results. As we work through the evidence on a problem involving birth defects, we will consider the ambiguities and technicalities of the language of biological and medical science.

Multiple Perspectives on Risk

There are many ways to think about risk, including risk assessment, perception, management, and communication. In this course, we will work through a set of actual data that has implications for all these modes of thinking about risk. The evidence inevitably involves multiple perspectives from which scientific evidence, reasoning, and conclusions are viewed:

Patients’ (or Consumers’) Perspective

There are patients who have a clinical need or desire for a medication. In the case of psychotropic medications such as antidepressants, the patients may be highly dysfunctional because of depression or anxiety. These patients may represent a risk of harm to themselves or others. There is, of course, the perspective of the unintended consumer, the “conceptus,” either an embryo or a fetus that receives a dose of the medication by virtue of its mother’s ingestion. There are risks and benefits for the developing embryo, in utero, that follow from the mother’s ability to cope with her pregnancy, and remain compliant with prenatal care. After birth, the perspective of the mother and of the infant may shift radically to a need to understand the cause or causes of life-threatening or devastating malformations.

Physician or Health Care Prescriber's Perspective

The prescribing health care provider may want to manage a disease or disorder that threatens both the mother and the viability of the unborn child. The provider will have professional responsibilities to obtain informed consent from the patient, and to make a reasoned judgment about the appropriateness of the medication, and its risk-benefit balance for the patient. In the face of uncertainty about the causation of adverse outcomes, the informed consent process may be an impossibility.

Professional medical societies may have mixed motives in addressing birth defects issues. One the one hand, professional societies will want to give up-to-date, authoritative advice to their members about appropriate therapy. On the other, they may wish to reduce the potential liability of their members.

The balancing of risk against benefits can be extremely difficult and personal. In some instances, even a medication known to cause serious birth defects may be indicated for a woman who has a serious disease. Consider the woman with epilepsy, who wants to complete a pregnancy. Antiepileptic medications may be known to cause some specific birth defects at a rate of 1/100, which is several fold higher than would be expected in the unexposed population of newborns. As high and concerning as is the known increased rate among children of mothers who take the antiepileptic medication, the disease for which the medication is given (known as the indication), the epilepsy, may lead to fetal death or deformity in 1/10 cases. The decision to support a pregnancy and prescribe a medication known to cause birth defects involves difficult ethical issues for physicians, patients, and third-party payers.

Pharmaceutical Companies' Perspective

The sponsors and manufacturers of medications have important legal and reputational interests in properly labeling their products. While these companies are primarily responsible for innovating and marketing medicines (or devices or biologicals) that improve health, the scientific appreciation of the efficacy of medications includes the realization that there will be "off-target" effects. Because of the limitations of premarketing testing, many off-target effects will likely be unknown. The uncertainty about off-target effects is particularly acute for women who will become pregnant while using the medications because it touches upon their own well-being and that of their offspring, and because there is generally no sufficient basis for forming judgments about risk and causation when a new medicine comes to market. Of course, after new medicines come to market, the pharmaceutical companies have obligations of "pharmacovigilance," which include working with the appropriate regulatory agencies on a range of activities, from monitoring adverse event reports (AERs), to assessing "signals" of associations in AERs, to establishing pregnancy registries of birth defects, and to sponsoring and reporting observational studies of birth defects among offspring of women who have used the medicine in pregnancy, to changing the labeling of the medicine. In some situations, the sponsor may act preemptively, without agency (Food and Drug Administration [FDA]) prior approval of a new warning, by making the change in labeling, pursuant to the FDA's so-called changes being effected (CBE) regulation,[2] for "newly discovered" risks, and hoping that the agency agrees.[3]

Regulatory Agencies' Perspective

In the first instance, the agency must guide, supervise, and evaluate the testing of a proposed drug in animals and then in humans. One of the first regulatory agency steps is to grant an investigational new drug (IND) application.[4] The data from animal studies and later from human clinical trials of an IND become the basis for a new drug application (NDA).[5]

In the initial approval process, the agency must determine whether a new drug is safe and effective for its proposed use(s), and whether the benefits of the drug outweigh the risks.[6] The determination of safety is always made in the context of safety as labeled, and thus the agency must determine whether the sponsor's proposed labeling (in the form of a proposed package insert) is appropriately designed and worded. The agency's approval of a drug's proposed labeling includes review and approval of the package insert's stated indication that specifies the conditions or diseases for which the drug may be used, and for which patients. The agency may require a statement of counterindications for patients or situations for which the medicine should not be used. Furthermore, the agency must also address the appropriateness of the use of the drug in specific populations, such as pregnant and lactating women.

The regulatory agency's responsibilities for assessing risk do not end with the initial approval of a new drug. The agency must work with the sponsor to evaluate potential signals generated by adverse event reporting, and to decide on whether further testing or changes in

[2] 21 C.F.R. § 314.70(c).

[3] 71 Fed. Reg. 3922, 3934 (FDA Jan. 24, 2006) ("[T]he determination whether labeling revisions are necessary is, in the end, squarely and solely FDA's."). The FDA promulgated Section 314.70 in 1982, and it has not changed materially since. In making this rule, the FDA characterized CBE labeling changes as intended for facilitating procedures for "newly discovered" drug risks. 47 Fed. Reg. 46622, 46623 (FDA Oct. 19, 1982) ("[S]ome information, although still the subject of a supplement, would no longer require agency preclearance. These supplements would describe changes placed into effect to correct concerns about *newly discovered risks* from the use of the drug.) (emphasis added). In April 2004, the FDA published in a final rule on changes to approved NDAs, including what kinds of changes to warnings could be made using the CBE procedures. See FDA, Supplements and Other Changes to an Approved Application, 69 *Fed. Reg.* 18728 (April 8, 2004) (final rule). Major label changes cannot be made by a sponsor without prior approval of the agency. See FDA, Guidance for Industry Changes to an Approved NDA or ANDA (April 2004) (Revision 1), available at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm077097.pdf, last visited on March 1, 2015.

[4] See FDA description of the Investigational New Drug (IND) application and process at http://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/investigationalnewdrugindapplication/default.htm, last visited March 28, 2015.

[5] For a description of the FDA's regulatory framework and guidances for sponsors on navigating the NDA process, see

http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/NewDrugApplicationNDA/, lasted visited on March 28, 2015.

6 See FDA Regulations, 21 U.S.C. § 355(d), (e) (requiring drug sponsor to show adequate testing, labeling, safety, and efficacy).

labeling are necessary.[7] The agency's decision that warnings need to be added or intensified do not necessarily turn on its determination that the drug has been shown to cause some adverse event. The agency has some discretion to impose warnings on a precautionary basis in the absence of evidence that would permit a conclusion that the drug causes the adverse outcome.[8]

Tort Bar's Perspective

A manufacturer's failure to warn or failure to warn adequately may be actionable under state law for compensatory and punitive damages for injuries sustained by consumers of the product. In the context of pharmaceutical products, the state law's mandate to warn adequately typically creates a duty to warn the prescribing physician of the alleged causal relationship between the drug and some untoward event, such as a birth defect.[9] The state law duty creates several layers of problems, not the least of which may be that the state law duty may conflict directly or indirectly with the warning label approved by the FDA, pursuant to federal law. Under the United States Constitution, federal law trumps state law (by a legal doctrine descriptively known as "preemption"), but the plaintiffs' bar may argue that drug sponsors can avoid the FDA's supervision by voluntarily changing warning labels under the so-called CBE provisions, discussed above. With respect to medical causation between drug and alleged harm to the plaintiffs, the plaintiffs' bar will have an interest in diluting standards for causal assessment, whereas the defense bar, on behalf of the pharmaceutical industry, will have an interest in heightening the same standards. As we have seen, the FDA may require warnings in the absence of demonstrated causation between drug and adverse event. When the evidence is uncertain, but the FDA's perception of the medication's utility is strong, the agency may disfavor adding warnings that will dissuade clinicians from prescribing the medication. On the other hand, uncertain evidence combined with a sense that the medication has minor utility, or relatively low efficacy, or the availability of a competing medication, may lead the FDA to impose strong warning language or even to withdraw the medication for market. The opportunities for judicial and jury confusion are rife.

ASSIGNMENT: READING FOR SECOND SESSION

Dana L. Shuey, Thomas W. Sadler, and Jean M. Lauder, "Serotonin as a Regulator of Craniofacial Morphogenesis: Site Specific Malformations Following Exposure to Serotonin Uptake Inhibitors," 46 *Teratology* 367 (1992) [Shuey 1992].

[7] See, e.g., Selena Ready, FDA Pharmacovigilance Overview (Feb. 11, 2014), available at http://fdaguidance.net/download/1234/, lasted visited March 28, 2015.

[8] See, e.g., 21 C.F.R. § 201.57(e) (requiring warnings in labeling "as there is reasonable evidence of an association of a serious hazard with a drug; a causal relationship need not have been proved."); 21 C.F.R. § 803.3 (adverse event reports address events possibly related to the drug or the device); 21 C.F.R. § 803.16 (adverse event report is not an admission of causation).

[9] The nature of the duty to the prescribing physician thus denies recovery for an alleged failure to warn when the physician, by virtue of training, skill, and knowledge, has an independent understanding of the deleterious effect of the prescribed drug at the time of prescribing.

FDA, Public Health Advisory: Paroxetine (Dec. 8, 2005)[10]

FDA, Guidance for Industry Establishing Pregnancy Exposure Registries (2002) [FDA Pregnancy Registries Guidance][11]

FDA, Reviewer Guidance Evaluating the Risks of Drug Exposure in Human Pregnancies (2005) [FDA Pregnancy Guidance][12]

FDA, Adverse Event Reporting System (FAERS) (formerly AERS) (Dec. 7, 2013)[13]

Medwatch Form[14]

FDA, Guidance for Industry Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment (2005) [Guidance for Pharmacovigilance][15]

RESOURCE MATERIALS

Michael D. Green, D. Michal Freedman, and Leon Gordis, "Reference Guide on Epidemiology," in *Reference Manual on Scientific Evidence* 549 (3d ed., 2011)

David H. Kaye and David A. Freedman, "Reference Guide on Statistics," in *Reference Manual on Scientific Evidence* 209 (3d ed., 2011)

[10] This public health advisory on paroxetine is available at http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm051731.htm, last visited March 1, 2015.

[11] This FDA guidance document is available at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071639.pdf, lasted visited on March 1, 2015.

[12] This FDA guidance document is available at http://www.fda.gov/downloads/Drugs/.../Guidances/ucm071645.pdf, lasted visited on March 1, 2015.

[13] This FDA webpage is available at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm, lasted visited on March 1, 2015.

[14] Various versions of the current Medwatch Form are available at http://www.fda.gov/Safety/MedWatch/HowToReport/DownloadForms/ucm2007307.htm, last visited March 15, 2015.

[15] This FDA guidance document is available at http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126834.pdf, lasted visited on March 1, 2015.

Assessing Causation of Birth Defects from a New Drug

CLASS ACTIVITY: DISCUSSION OF KEY CONCEPTS FROM READINGS OF FDA MATERIALS

What did the Food and Drug Administration (FDA) warn against in its 2005 Public Health Advisory on Paxil/paroxetine?

[Note the specificity of end point or outcome, and the care in not asserting that the agency had reached a conclusion of causation.]

Did the FDA announce that it, or anyone, had concluded that paroxetine causes heart (cardiac) birth defects?

Did the FDA suggest that its concern involved all possible kinds of birth defects?

What kind of evidence did the FDA cite in support of its Public Health Advisory?

From your reading of the FDA guidances, what kinds of evidence might be used to address the question whether a medication causes a birth defect?

- Clinical trials (available at time of first licensing?).
- Observational epidemiologic studies (analytical versus descriptive).
- Anecdotal evidence (case reports, adverse event reports, case series).
- Animal (toxicologic) studies (whole animal, in vitro).
- Analogies to similar medications (structural or indication similarity).

Did the FDA's Public Health Advisory state that all serotonin selective reuptake inhibitors (SSRIs) were implicated by its review of the available evidence?

As you probably know, there are many different kinds of antidepressant medications, and several in the class of antidepressants known as SSRIs:

- Paxil (paroxetine)
- Prozac (fluoxetine)
- Zoloft (sertraline)
- Lexapro (escitalopram)
- Citalopram (citalopram)

REPRODUCTIVE TOXICOLOGY – ANIMAL STUDIES

Class Activity: Discussion of In Vitro Studies – Shuey Article

What kind of study is reported in the paper by Shuey and colleagues?

[This is a whole embryo culture (WEC) study, done in vitro, literally, in a bottle with embryos removed from a pregnant rat mother (dam).]

What did Shuey conclude?

What is the relevance of this study to the causation of human birth defects?

The key data from Shuey (1992) are presented in Table 1 from the published paper, below:

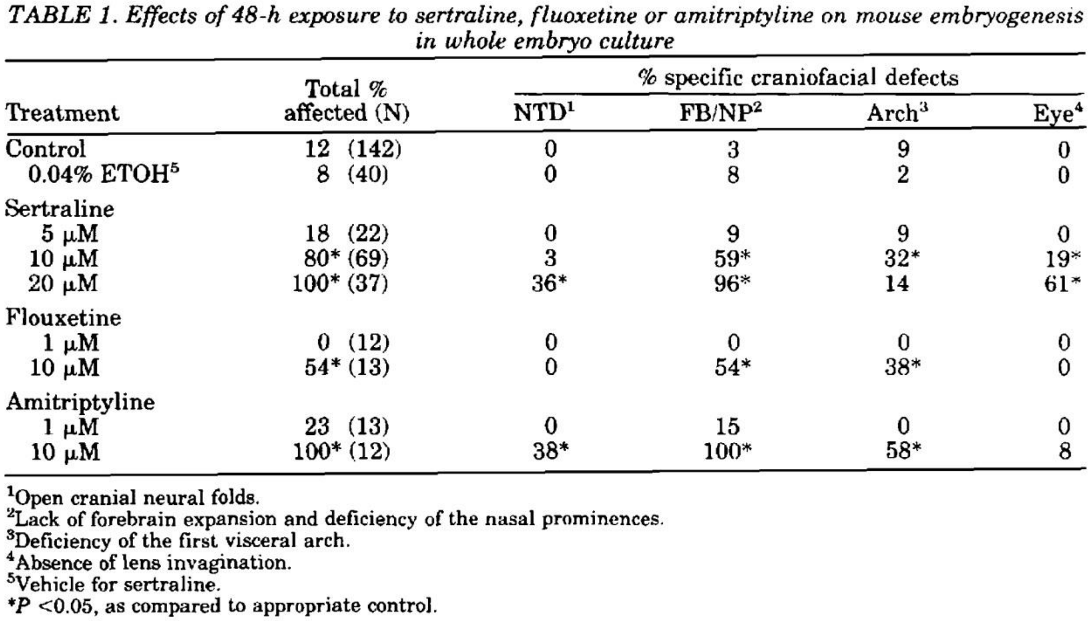


TABLE 1. Effects of 48-h exposure to sertraline, fluoxetine or amitriptyline on mouse embryogenesis in whole embryo culture

Treatment	Total % affected (N)	% specific craniofacial defects			
		NTD[1]	FB/NP[2]	Arch[3]	Eye[4]
Control	12 (142)	0	3	9	0
0.04% ETOH[5]	8 (40)	0	8	2	0
Sertraline					
5 μM	18 (22)	0	9	9	0
10 μM	80* (69)	3	59*	32*	19*
20 μM	100* (37)	36*	96*	14	61*
Flouxetine					
1 μM	0 (12)	0	0	0	0
10 μM	54* (13)	0	54*	38*	0
Amitriptyline					
1 μM	23 (13)	0	15	0	0
10 μM	100* (12)	38*	100*	58*	8

[1] Open cranial neural folds.
[2] Lack of forebrain expansion and deficiency of the nasal prominences.
[3] Deficiency of the first visceral arch.
[4] Absence of lens invagination.
[5] Vehicle for sertraline.
*P <0.05, as compared to appropriate control.

SOURCE: Dana L. Shuey, Thomas W. Sadler, and Jean M. Lauder, "Serotonin as a Regulator of Craniofacial Morphogenesis: Site Specific Malformations Following Exposure to Serotonin Uptake Inhibitors" 46 Teratology 367 (1992) Copyright © 1992 Wiley-Liss, Inc. Reproduced with permission.

What are the exposures to the rat embryos in this WEC experiment?

- Nothing.
- 0.04 percent ethyl alcohol (ETOH).
- Sertraline (Zoloft), an SSRI, at three doses in ethanol solution.
- Fluoxetine (Prozac, another SSRI antidepressant), at two doses.
- Amitriptyline (Elavil - tricyclic antidepressant), at two doses.

What structural abnormalities are the outcomes of the experiment?

f. **NTD** (neural tube defect) or open cranial neural folds.

g. **FB/NP** (lack of forebrain expansion and deficiency of the nasal prominences).

h. **Arch** (deficiency of the first visceral arch) usually associated with facial or cranial abnormality.

i. **Eye** (absence of lens invagination), deficient, dysfunctional eye structure.

The dosages are given in terms of micromolar concentrations. From your reading of the Shuey paper, how would you characterize the exposure levels used for the SSRI medications?

[They were extremely high. The lowest level of sertraline used in the WEC was acknowledged to be twice the lethal dose for a living mother rat. The doses are calibrated in terms of µM (micromolar) concentrations, which are based upon the molecular weight of each of the different chemicals. Elsewhere in the article, Shuey reported that maternal death was observed at serum concentrations of higher than 2µM, in vivo (living animals). All dosing groups of embryos in the WEC are thus at levels that would have killed the mother animals.]

Was there a dose-response relationship between sertraline exposure in the WEC and the structural abnormalities?

[Not in any meaningful sense. Note that for NTD and Eye, there are no differences between control and sertraline at 5 µM. For FB/NP, there is a numerical difference that was not reported to be statistically significant. Recall that the lowest exposure level is already at multiples of a lethal dose for the rat dam, which means that the observed differences between the dose categories cannot be translated into differences between responses of in utero embryos at different levels of pharmacologically relevant doses.

Note that the sertraline had to be placed in the alcohol solution in order to deliver it to the embryo in the culture. Thus, the appropriate comparison is between the sertraline in various concentrations and the dilute alcohol solution.]

What advantages might this kind of animal toxicology study have for investigating the causation of birth defects?

[This kind of study creates an artificial environment, known as whole embryo culture, or WEC, for embryos removed from their mothers, by placing them in a rotating glass jar, containing a nutrient fluid. Although the embryos survived only 2 or 3 days in this artificial environment, the researchers could expose the developing embryos to various levels of drugs, without the interference of placental or other tissue barriers between mother and embryo. Shuey and colleagues were interested in the role of serotonin as a signaling molecule in the early embryological development of rats and mice, and thus they could attempt to see how the addition of the SSRIs changed organ formation at a very early stage in rat embryogenesis (Shuey et al., 1992).]

From your reading of the FDA guidances, what is your sense of the importance or unimportance of this kind of study to the FDA's decision process about drug approval and warning labels?

[In vitro WEC experiments have never been part of the regulatory approval process, but the scientists who conduct such experiments tend to place a great deal of importance on their results. These scientists were intrigued by the potential for an SSRI to inhibit serotonin transport, and thus affect serotonin signaling, not in brain synapses, where the SSRIs have their pharmacologic effects, but in the developing animal embryos, in which serotonin signals developmental movements and structural changes.]

Shuey and colleagues reported their results to the FDA, about the time that sertraline was first approved, but the FDA ultimately regarded these experimental results as too dissimilar to the nature and quantity of in utero exposures, and thus not important to assessing the risk to humans. Neither the sponsor nor the FDA believed that the WEC experiments should be described in the package labeling for the drug when it was approved for marketing.[16] Of course, those determinations back in 1991 and 1992 may not be dispositive of the issue today, either in the regulatory agency or in a courtroom.

The Shuey study was carried out at levels many times higher than any possible pharmacologic dose. Indeed, the lowest sertraline (Zoloft) dose was more than twice a lethal dose in the mother. The authors would argue that the point of the experiment, however, was to determine whether very high levels could invoke a particular mechanism of inducing malformations in the rat embryos.

[By today's standards, the statistical analysis is problematic. The authors present no p-values, measures of effect, or confidence intervals. They do not identify the nature of their statistical analysis or test. Still, the raw percentages are reasonably understandable.]

Shuey (1992) seemed to add little to the FDA's assessment of risk before permitting sertraline onto the market. After the FDA issued its Public Health Advisory for paroxetine and birth defects in 2005, the Shuey paper took on renewed significance, as agency critics used its results to suggest that the FDA should have expanded the scope of its toxicological assessments of the SSRI drugs to include WEC experimental data.

Class Activity: Discussion of Whole Animal Studies – "The Dose Makes The Poison"[17]

The following summary is from an FDA-approved package insert[18] for another SSRI, describing the whole animal studies done with citalopram (Celexa), one of the SSRI medications:

[16] Perhaps a sign of how researchers become invested in their research results, one of the authors of Shuey 1992 has become a regular expert witness for plaintiffs in birth defects litigation against the manufacturers of SSRIs.

[17] Shortened from "[a]ll substances are poisonous—there is none which is not; the dose differentiates a poison from a remedy," originally in Paracelsus' 16th century German, "Alle Ding sind Gift und nichts ohn' Gift; allein die Dosis macht, das ein Ding kein Gift ist." Paracelsus, or Auroleus Phillipus Theophrastus Bombastus von Hohenheim, as he was known to friends, is often regarded as the father of toxicology, who shocked 16th century Europe with his experimental approach and his use of the vernacular rather than Latin for his lectures. See Joseph F. Borzelleca, "Paracelsus: Herald of Modern Toxicology," 53 *Toxicological Sci*. 2 (2000), available at http://toxsci.oxfordjournals.org/content/53/1/2.full.pdf+html, lasted visited March 1, 2015.

"In two rat embryo/fetal development studies, oral administration of racemic citalopram (32, 56, or 112 mg/kg/day) to pregnant animals during the period of organogenesis resulted in decreased embryo/fetal growth and survival and an increased incidence of fetal abnormalities (including cardiovascular and skeletal defects) at the high dose. This dose was also associated with maternal toxicity (clinical signs, decreased body weight gain). The developmental no-effect dose was 56 mg/kg/day. In a rabbit study, no adverse effects on embryo/fetal development were observed at doses of racemic citalopram of up to 16 mg/kg/day. Thus, teratogenic effects of racemic citalopram were observed at a maternally toxic dose in the rat and were not observed in the rabbit."

What are the findings in rat embryos when the mothers are fed citalopram?
What are the findings in rabbit embryos when the rabbit mothers were fed citalopram?
Are humans more like rats or rabbits?
What is the significance of association of certain dosage levels with maternal toxicity?
In the rat studies, what was observed when the mothers were fed 56 mg/kg/day?
What is an adult rat's weight in kg? (laboratory versus New York City subway)?

[About 250 to 300 gm for an adult female laboratory rat. Rats in the New York subway system can reach 1 kg, over 2 pounds.]

The results in whole animal experiments are reviewed in view of the exposure levels in terms of the maximum human recommended dose (MHRD), calculated on a body surface area (mg/m2) basis. The no-effect level of 56 mg/kg/day translates into about 9 times the MHRD. The level at which the embryo abnormalities was observed is twice that, 18 times the MHRD.

So-called whole animal studies of teratogenicity involve exposing pregnant animals to the drug, at various doses until the dosing causes maternal toxicity and death. If the pregnant animal herself is showing overt signs of toxicity, with organ damage and failure, any damage to unborn embryos is at best ambiguous, as it may result from the drug or from the mother's inability to provide a supportive uterine environment, as the mother's individual organ systems are compromised by the deliberate overdose of the medication.

FDA generally requires developmental toxicity testing that consists of several approaches, in two species, usually rats and rabbits. One approach involves multiple groups with increasing doses to the pregnant animals, with allowing the animals to carry their litters to term. Another approach requires removal of embryos by surgery, preterm, at the end of the critical stage of organ formation.

[18] Lexapro (escitalopram oxalate) Tablets/Oral Solution NDA 21-323/NDA 21-365, in section on Pregnancy, available at http://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4065b1-22-tab11C-Lexapro-Tabs-SLR015.pdf.

RANDOMIZED CLINICAL TRIALS – APPROVAL FOR MARKETING

Randomized clinical trials (RCTs) are experiments on human beings, in which investigators randomly assign study participants to an experimental therapy or a placebo or comparator drug. The point of randomization is to remove potential bias or confounding from the study and to ensure a valid comparison, as much as possible, between the two groups. Ideally, the investigators are unaware of (blinded to) which participants receive the experimental therapy and which participants receive the placebo or comparator; and similarly, the participants themselves are blinded to which group they are in.

Although assessing potential teratogenicity (ability to cause birth defects) is an important part of FDA's and the sponsor's initial evaluation of a new drug's risk-benefit profile, pregnant women are almost never included in RCTs. Similarly, because the drug has not yet been marketed, there will be no basis for obtaining data from observational studies on birth defects among children of mothers who used the new drug. As a result, the only data on embryonic and fetal effects for assessing risk will come from animal reproductive toxicology studies.

Ethical Considerations

Women of child-bearing age constitute an important population subgroup that is needed within RCTs. Sponsors of RCTs obviously can provide informed consent only to the extent that they outline the lack of knowledge about the potential for teratogenicity to women recruited for the trial. RCT sponsors cannot require women to use effective birth control, or to terminate pregnancies that begin while women are receiving an experimental therapy. Under the law of many states, women who become pregnant during an RCT cannot provide effective waivers on behalf of their unborn children, despite their mothers' decisions to remain in an RCT after learning of their pregnancies.

The sponsor and the FDA are thus both in the unenviable position of trying to anticipate risks without any experimental clinical trial data or even observational human data in the investigational new drug/ new drug application (IND/NDA) process. Only animal toxicology data are available. Although such data are essential to the approval process, in the context of assessing risk of human birth defects, the animal data are often inadequate and misleading; if extrapolations from nonhuman animals to human animals are relied upon, the inferences will lead to a high rate of false-negative and false-positive conclusions. Here is how the FDA characterizes the problem:

> Animal reproductive toxicology studies are an essential tool for estimating potential risks of exposure to medical products in pregnancy. However, the positive and negative predictive values of such studies for humans are often uncertain (Mitchell, 2000). Animal models can be misleading when screening for specific fetal effects by detecting associations that ultimately turn out to be false positive (e.g., hydrocortisone and clefts in mice) or false negative (e.g., thalidomide and no teratogenesis in rats) (Ward, 2001). The strongest concordance between animal findings and human effects is when there are positive findings from more than

> one species, although even in this case the results cannot always be used to predict specific human effects or incidence in humans (Rogers et al., 1996).
>
> —FDA, *Guidance for Industry Establishing Pregnancy Exposure Registries* at 3 (2002).[19]

ANALOGIES TO, AND EXTRAPOLATIONS FROM, SIMILAR DRUGS – CLASS EFFECT

A pharmacologic class can arise from one or more of the following:

- Same mechanism of action.
- Same physical effect
- Similar chemical structure.

The SSRIs have the same pharmacologic mechanism with respect to inhibiting the serotonin transporter in the human brain. Their chemical structures vary considerably. The sponsors for the SSRIs had to establish safety and efficacy of their individual drug, without relying upon the safety and efficacy studies of other drugs within the class of SSRIs.

The FDA specifically addresses arguments for class effects in teratogenesis:

> Understanding the structure/activity relationships and pharmacological mode of action of a class of therapeutic agents in some circumstances can provide a prediction of the possible safety and efficacy of a new agent. However, such knowledge is generally not predictive of human teratogenesis (Mitchell, 2000).
>
> —FDA, *Reviewer Guidance Evaluating the Risks of Drug Exposure in Human Pregnancies* at 12 (2005), available at http://www.fda.gov/downloads/Drugs/.../Guidances/ucm071645.pdf, lasted visited on March 1, 2015.

The agency pointed to the experience with thalidomide, clearly a major teratogen, as a drug within a chemical and pharmacological class that lacked a "class effect." Like thalidomide, glutethimide (trade name Doriden) is a member of the class of drugs known as glutarimides. Both thalidomide and glutethimide are sedatives; both have similar chemical structures. Unlike thalidomide, however, glutethimide is not teratogenic (id. at 12 [citing Heinonen, 1977]). The FDA thus cautions that:

> While the introduction of a new product from a class of drugs with known human teratogenicity will solicit heightened scrutiny, it cannot be assumed that the product will also be teratogenic.
>
> —Id.

[19] This FDA guidance document is available at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071639.pdf, lasted visited on March 1, 2015.

A well-respected textbook on pharmacoepidemiology (the study of medication effects by using the methods of epidemiology) states the problem of using pharmacologic or structural "class" to infer teratogenicity thus:

> The fallacy of "class action" teratogenesis.
>
> Another clinically important concern specific to teratogenesis is the issue of "class action." It is widely recognized that an understanding of structure/activity relationships shared by members of a given drug class can be helpful in predicting a given class member's efficacy and adversity (indeed, this view is incorporated into regulatory action in the form of class labeling). However, class-based pharmacologic effects cannot be assumed to hold when the adversity at issue is teratogenesis. Given our ignorance about the causes of most birth defects, we cannot know whether it is the chemical structure common to the class that is responsible for teratogenesis or whether the responsible component is that part of the structure that differentiates one class member from another.
>
> For example, thalidomide and glutethimide (Doriden® and other brands) are both glutarimides, and both are sedative/hypnotics. Despite their structural and clinical similarities, thalidomide is clearly a high-risk teratogen and glutethimide is not.[15] Thus, we cannot assume that if one drug is a high-risk teratogen, all other members of its class will share that effect; conversely, we cannot assume that reassurance about the safety of one drug can be extended to other members of that drug's class.
>
> —Allen A. Mitchell,[20] *Studies of Drug-Induced Birth Defects*, chapter 28, in Brian L. Strom, Stephen E. Kimmel, and Sean Hennessy, eds., *Pharmacoepidemiology* 487, 491 (5th ed., 2012). Copyright © 2012 John Wiley & Sons, Ltd. Reproduced with permission.

The Teratology Society is a professional organization that includes several disciplines, epidemiology, toxicology, obstetrics, and others, to study birth defect causation and prevention. The society publishes an important series of journals, *Birth Defects Research*, as well as position papers on issues in birth defects research, prevention, and education. In the society's position paper on assessing causation in litigation, it suggests that regulatory classification of a drug based upon class membership might be reasonable, but that inferring teratogenicity from class membership alone would not.[21]

The Evolution of the Problem in the Postmarketing Phase

Class Activity: Discussion of Key Concepts from Readings of FDA Materials on Pharmacovigilance

When a new drug comes to market, what evidence will be available to show its safety for children exposed in utero?

[20] Dr. Mitchell is director of the Slone Epidemiology Center, at the Slone Epidemiology Center at Boston University, Boston University Schools of Public Health and Medicine.

[21] The Public Affairs Committee of the Teratology Society, "Teratology Society Public Affairs Committee Position Paper -- Causation in Teratology-Related Litigation," 73 *Birth Defects Research (Part A)* 421, 423 (2).

[None, although, as discussed, there will be some nonhuman safety evidence.]

What can the drug sponsor, licensing agencies, and clinical communities do to be vigilant about emerging evidence, trends, or signals of potential harm?

[Case reports, spontaneous adverse event reports, signals of disproportionate reporting of adverse events, pregnancy registries, epidemiologic studies; but not randomized clinical trials.]

What are the limitations to the available methods (in your or in the FDA's views)?

Which of the available methods are most likely to yield accurate answers (that is, not fail to identify a true safety concern, but not falsely identify a safe medication as harmful)?

Let's focus on adverse event reports (AERs) or other case reports of a child, born with a serious cardiac birth defect, to a mother who used an SSRI antidepressant in the first trimester of pregnancy. Are such adverse events unexpected?

What percentage of pregnancies are planned in the United States, on average, each year?

[50%]

How many live births occur in the United States each year?

[4 million]

What is the rate of serious heart birth defects among the live births?

[~ 1%]

What is the rate of exposure to SSRI antidepressants among pregnant women?

[≥ 2%]

What is the expected number, on average, of babies born each year with serious heart defects to women who used SSRI medications in pregnancy?

[(≥ 4 million) X (0.01) X (0.02) ≥ 800 babies each year expected to be born with serious heart defects to mothers who used SSRIs in the first trimester of pregnancy.]

What do those 800 babies born each year represent in terms of pharmacovigilance or safety assessment of SSRIs?

[This number of babies would be expected to be born with serious heart defects even if the mothers' use of SSRI medications contributed nothing to the burden of such birth defects among their children.]

What complications for pharmacovigilance are created by the presence of other risks among the mothers who use SSRI antidepressants in pregnancy?

Can an AER tell us anything about a causal relationship between the mother's use of the drug and the child's birth defect?

Does an AER have to be viewed as evidence of causation between the mother's use of the drug and the child's birth defect by the person filing the report?

What are the likely motivating factors in filing an AER?

[concern, curiosity, conscientiousness, response to publicity or litigation]

Are there circumstances in which the number or the pattern of AER filing might support a conclusion of causality between the mothers' use of a drug and the children's birth defects?

[The thalidomide case is perhaps the exception that helps define the rule: a new, prevalent exposure with another extremely rare outcome, phocomelia, a very specific limb-reduction defect.]

How strong must a relationship be between the exposure and the outcome to give a pattern such as seen with thalidomide in Europe, in the 1950s and 1960s?

[The risk ratio for thalidomide and phocomelia was over 1,000.]

What methods are available that address the limitations of case reports and pregnancy registries, and the unavailability of randomized clinical trials?

What are the key features of (observational) epidemiologic studies?

[Controlled observations require that both exposed and unexposed study participants are included, with and without the outcome of interest; the study is based upon a sample of the total population, or a sample of cases with the congenital abnormality of interest; the "2 x 2" table at the heart of epidemiologic studies.]

INTRODUCTION TO KEY CONCEPTS OF EPIDEMIOLOGY

Birth defects in the form of structural abnormalities of organs and bodily structures have been observed and recorded since ancient Babylonian times. The understanding of the causation of birth defects advanced little until the 20th century. In 1941, an Australian physician, Norman Gregg, published a study that identified maternal rubella as a cause of birth defects. Thalidomide was marketed in Europe (but not in the United States) in 1957. Within a few years, in 1961, physician Widukind Lenz suggested that the emerging European epidemic of birth defects, and in particular, an extremely rare limb defect known as phocomelia (seal-like appendages), were caused by thalidomide. William McBride, an Australian physician who later gained notoriety for his role in the scare over Bendectin and birth defects, made the connection between thalidomide and birth defects about the same time.

Teratology is the discipline that studies birth defects, as well as more broadly, fetal death and miscarriage, their origins and causes. The problem in this learning module will focus more narrowly on structural abnormalities, known variously as congenital anomalies, birth defects, congenital abnormalities, or congenital dysmorphology.

Depression. Chronic depression is widely prevalent in the United States, with a higher concentration among women than among men.[22] Roughly half of all pregnancies are unplanned, which means that many women will be taking medications of various kinds well into the first trimester of their pregnancies. Estimates of the prevalence of women taking SSRI medications range from about 6 to 10%.[23]

One reason for the difficulty in studying whether birth defects are associated with antidepressant medications is that depression itself is associated with several risk factors for various birth defects. The risk factors are usually behavioral, although there may be some risk from depression itself. Women who are depressed are different from women without depression in some relevant respects for their pregnancy outcomes. Women who are depressed tend to be more overweight, to drink alcohol more, to smoke more, to have poorer diets, to be more likely to use street drugs, and to be less compliant with prenatal dietary supplements than nondepressed women. To some extent, researchers can hope to capture some of this information and "adjust" for the differences in depressed and nondepressed women's pregnancy outcomes. Largely, however, researchers realize that they are not ever likely to obtain accurate data on some of the important measures of these lifestyle variables.

The Exposure of Interest – Sertraline (Zoloft)

Serotonin, known chemically as 5-hydroxytryptamine (5HT), is made in the body from an amino acid, tryptophan, which is a normal component of animal and human protein. In the human, most of the body's serotonin is made and found in the gut, where it regulates intestinal movements. Serotonin is also taken up and stored in the platelets of the circulating blood. In the brain, serotonin is made in particular nerve cells (neurons). The serotonin acts as a neurotransmitter to move signals between neurons in certain portions of the brain. The serotonin is released into the space between neurons (the synapse or synaptic space).

Sertraline (trade name Zoloft) is an antidepressant that works by selectively inhibiting serotonin reuptake in synaptic spaces between neurons of the brain. This medication is thus part of a class of antidepressants known as selective serotonin reuptake inhibitors. Fluoxetine (trade name Prozac) was the first SSRI antidepressant marketed in the United States, starting in late 1987. Sertraline followed in 1991, with paroxetine (trade name Paxil) in 1992, citalopram (trade name Celexa) in 1998, and escitalopram (trade name Lexapro) in 2002. All the SSRIs are

[22] The National Institutes of Mental Health webpage for "Depression," has some important background information on the scope of the problem, available at http://www.nimh.nih.gov/health/topics/depression/index.shtml, lasted visited on March 1, 2015.

[23] William O. Cooper, Mary E. Willy, Stephen J. Pont, and Wayne A. Ray, "Increasing use of antidepressants in pregnancy," 196 *Am. J. Obstet. Gynecol*. 544.e1, e1, e3 (2007).

prescription medications. The patents for all these antidepressants have expired, and all are available as generic drugs.

The SSRIs work by binding with the serotonin transporter. Sertraline, like the other SSRIs, is indicated primarily for major depressive disorder, as well as anxiety disorders.

Pharmacovigilance

Pharmacovigilance is the practice of identify and evaluating safety signals from postmarketing information and data. The term *signal* or *safety signal* generally refers to an observation or a concern that more adverse events of a particular kind are being observed for a drug than are expected. A signal triggers a need for additional study and investigation, which may, or may not, support a conclusion that the drug causes the adverse event seen in the signal. Safety signals can be generated from postmarketing adverse event data or other case reports and studies. Signals represent a hypothesis in need of testing, and not conclusions about causation.[24]

Pharmacovigilance may consist of monitoring AERs for unusual or unique patterns of birth defects in the reports. Various “data mining” approaches to databases of AERs exist to test whether a safety “signal” is emerging. One approach, known as disproportionality analysis looks at only AER counts from within a database of all reported AERs. This approach then assesses whether one outcome for a particular drug is becoming disproportionally more common with respect to all AER outcomes for that drug, compared with the proportion of AERs for this particular outcome among all AER outcomes for all other prescription products. Alternatively, the comparison may be made with the proportion for all other drugs in the therapeutic class. A signal of disproportionality results when there is a measure of statistical association that suggests the particular outcome is more commonly reported for the drug of interest than for other drugs in the database. A finding of disproportional reporting is not a measure of causality between the drug and a particular medical event, but a trigger for the use of more rigorous, analytical epidemiologic methods.

Spontaneous Adverse Event Reporting and Other Case Reports

Adverse Event Reports (AERs)

After drug is marketed, physicians or their patients may file an AER (Medwatch form) with the FDA. In other countries, there are similar reporting systems. Even lawyers may, and do, file AERs on behalf of their clients, or direct their clients to file such reports.
An adverse event is defined as:

> Adverse drug experience. Any adverse event associated with the use of a drug in humans, whether or not considered drug related, including the following:

[24] See, generally, FDA, Guidance for Pharmacovigilance.

.

> An adverse event occurring in the course of the use of a drug product in professional practice. . .
>
> —21 C.F.R. § 314.80 (a) (postmarketing reporting of adverse drug experiences).[25]
> The phrases, "adverse events," "adverse reactions,"[26] and "adverse effects," are sometimes loosely used interchangeably, but they are distinct and different concepts. The phrase "adverse event" carries no suggestion of causality.[27]

Reporting is voluntary for physicians and patients, but mandatory for drug sponsors, at least for serious adverse events. The birth defects that are considered "major malformations," and make up the 3–4% of all live births, are all "serious" adverse events within the FDA's regulatory framework. As a result, drug sponsors must report such adverse events within 15 days of their receiving notice, regardless how the notice was obtained.

The filing of an AER by a sponsor is not an admission of causality; nor is the receipt of the AER by the agency a determination that the drug caused the reported event:

> (k) Disclaimer. A report or information submitted by an applicant under this section (and any release by FDA of that report or information) does not necessarily reflect a conclusion by the applicant or FDA that the report or information constitutes an admission that the drug caused or contributed to an adverse effect. An applicant need not admit, and may deny, that the report or information submitted under this section constitutes an admission.
>
> —21 C.F.R. § 314.80 (k).

Typically, only a small percentage of all adverse events are reported, but the situation for miscarriage and birth defect may be different. The percentage of AERs reported out of all adverse events is subject to the influence of print and electronic news media, social media, other sources of publicity, and litigation.[28]

[25] See also FDA, "Guidance for Industry -- Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products — Content and Format" at 13 (January 2006) (Glossary) ("any untoward medical event associated with the use of a drug in humans, whether or not considered drug-related.").

[26] FDA, "Guidance for Industry -- Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products -- Content and Format" at 13 (January 2006) (Glossary) ("an undesirable effect, reasonably associated with the use of a drug, that may occur as part of the pharmacological action of the drug or may be unpredictable in its occurrence. This definition does not include all adverse events observed during use of a drug, only those for which there is some basis to believe there is a causal relationship between the drug and the occurrence of the adverse event."); see also 21 CFR § 201.57(c)(7)).

[27] See also The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) E2D Guideline on Post-Approval Safety Data Management: Definitions and Standards for Expedited Reporting, quoted in ICH harmonized tripartite guideline, "Post-approval safety data management: definitions and standards for expedited reporting," 336 *The Lancet* 156 (2003).

[28] See Derrick J. Stobaugh, Parakkal Deepak, and Eli D. Ehrenpreis, "Alleged Isotretinoin-Associated Inflammatory Bowel Disease: Disproportionate reporting by attorneys to the Food and Drug Administration Adverse Event Reporting System," 69 *J. Am. Acad. Dermatol*. 398 (2013). ("Attorney-initiated reports inflate the pharmacovigilance signal of isotretinoin-associated IBD [irritable bowel disease] in the FAERS [FDA AER system].")

Role of Case Reports in Assessing Causation between Medication Use and Birth Defects

After a medication is marketed, there will inevitably be spontaneous reporting of adverse events, including birth defects. The March of Dimes estimates that 3 to 4% of all live births have a "major" congenital malformation, and there are about 4 million live births in the United States each year.[29] These rates have been relatively stable over many years, with the result that about 160,000 children are expected to be born with major birth defects each year. Many mothers will have used medications during their pregnancy, especially in the first few weeks after conception, when they may be unaware of their pregnancies. The prevalence of use of antidepressants in particular, as we have seen, is high. Accordingly, there is the expectation that many of children with major birth defects will be born to mothers who coincidentally used any one or more of various prescription or over-the-counter medications. If only 3% of women used an SSRI in their first trimester, we would thus expect about 4,800 children coincidentally born (ignoring multiple births) each year with major birth defects to these mothers.

The FDA acknowledges that case reports may arouse suspicion, but that they "cannot distinguish coincidence from causation and cannot be used to assess teratogenic risk."[30] Most case reports, or AERs, when followed up do not lead to a confirmed finding of causation.[31]

The use of adverse event reports is thus very limited, but there may be times when there is a highly specific outcome that starts to occur shortly after a medication becomes available. Phocomelia, the extreme seal-like, limb-reduction defect, was known before thalidomide was introduced to the European market, but this defect was extraordinarily rare. Not long after thalidomide was marketed, the birth prevalence of phocomelia soared; indeed, the curve of new phocomelia cases (and all defects) follows the curve of increasing use of thalidomide, only lagged by about 1 year.

[29] In 2011, there were 3,953,590 births in the United States. In 2011, there were 3,953,590 births in the United States. See http://www.cdc.gov/nchs/fastats/births.htm.

[30] FDA Pregnancy Guidance at 13.

[31] Id. (citing the Bendectin experience).

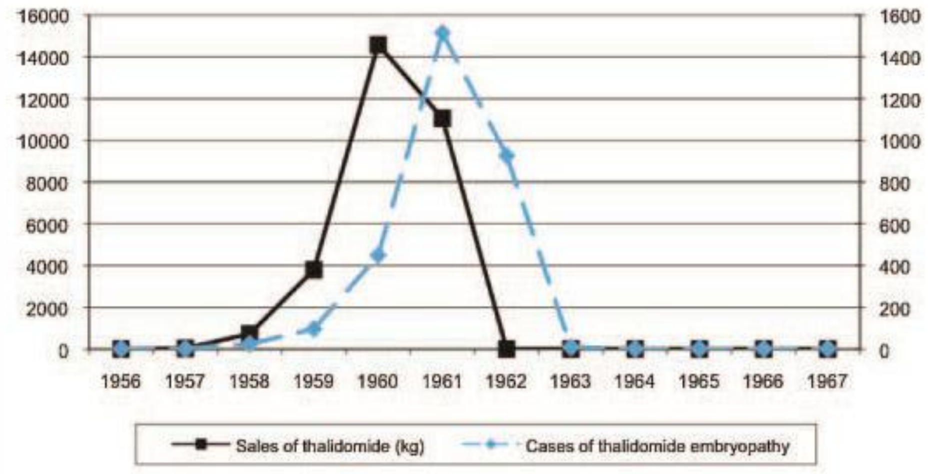


Figure 1. Sales of thalidomide and the appearance of the typical limb malformations in Germany. Drawn from data presented by Lenz [1988].

SOURCE: Sarah Običan and Anthony R. Scialli, "Teratogenic exposures." 157(3) Am J Med Genet Part C Semin Med Genet 150–169 (2011). Copyright © 2011 Wiley-Liss, Inc. Reproduced with permission.

Pregnancy Registries

One method that a drug sponsor may use to search for emerging evidence of teratogenicity after commencing sales is to create a pregnancy registry. Women are recruited voluntarily to participate, or encouraged to participate upon contacting the sponsor after learning that they have become pregnant while taking the drug. In some cases, the FDA may require the sponsor to establish a registry if there is sufficient suspicion of teratogenicity.

Despite their attractiveness and appearance of addressing potential risk seriously, pregnancy registries have limited utility. Such registries tend to be too small to detect any but "major" teratogens, such as thalidomide or isotretinoin (trade name Accutane), which increase risk of birth defects 10- to 1000-fold, and produce birth defects in a relatively high proportion of infants of exposed mothers.[32]

The usefulness of registries is further limited by self-referral bias, created by the proportional over-enrollment of women more likely to have malformed infants.[33] Losses of mothers to follow-up is also a serious problem, with mothers' having healthy babies less motivated to continue to report to the registry.[34]

Although the FDA encourages registries in some instances, the agency acknowledges that they have not ever detected a drug teratogen.[35] In the case of the SSRI antidepressants,

[32] FDA Pregnancy Guidance at 15–16.

[33] Id. at 15.

[34] Id.

[35] Id. at 16.

the paroxetine-birth defect association arose from ongoing studies, the existence of which is considered a reason by the FDA for discontinuing a registry.[36]

Epidemiologic Methods

Epidemiology is the "study of the distribution and determinants of health-related states or events in specified populations and the application of this study to the control of health problems."[37] Apparent and real contradictions in reported epidemiologic associations give rise to popular perception that epidemiology is a failed science. What is lost in this popularization is the many successes of post-World War II epidemiology in identifying causal associations, which today are well accepted. A few examples should help remind students of the power of epidemiologic methods:

- Radiation – various cancers
- Arsenic – lung cancer
- Asbestos – lung cancer
- Tobacco smoking – lung cancer, laryngeal cancer, chronic obstructive pulmonary disease, and so on
- Diethylstilbestrol (DES) – vaginal adenocarcinoma
- Hepatitis (B, C) – liver cancer
- Estrogen – endometrial cancer
- Formaldehyde – nasal sinus cancer
- Ethanol – liver, laryngeal, oral pharyngeal, esophageal, and breast cancer
- Human papilloma virus (HPV) – cervical cancer
- HIV/AIDS – Kaposi sarcoma, non-Hodgkin lymphoma
- Helicobacter pylori – gastric ulcer, stomach cancer

There have been, to be sure, false positives, which tend to linger in the popular imagination. In some instances, case reports or epidemiologic questionable validity have suggested associations between particular exposures and outcomes, for which subsequent, better-designed studies have shown to be unsupportable:

- Abortion – breast cancer
- Ethyl mercury in vaccines – autism
- Coffee – pancreatic cancer, bladder cancer
- Beta-carotene – lung cancer
- Vasectomy – prostate cancer
- DDT – breast cancer
- Silicone – autoimmune connective tissue diseases
- Mercury amalgam dental fillings – various ailments

[36] Pregnancy Registries Guidance at 17.

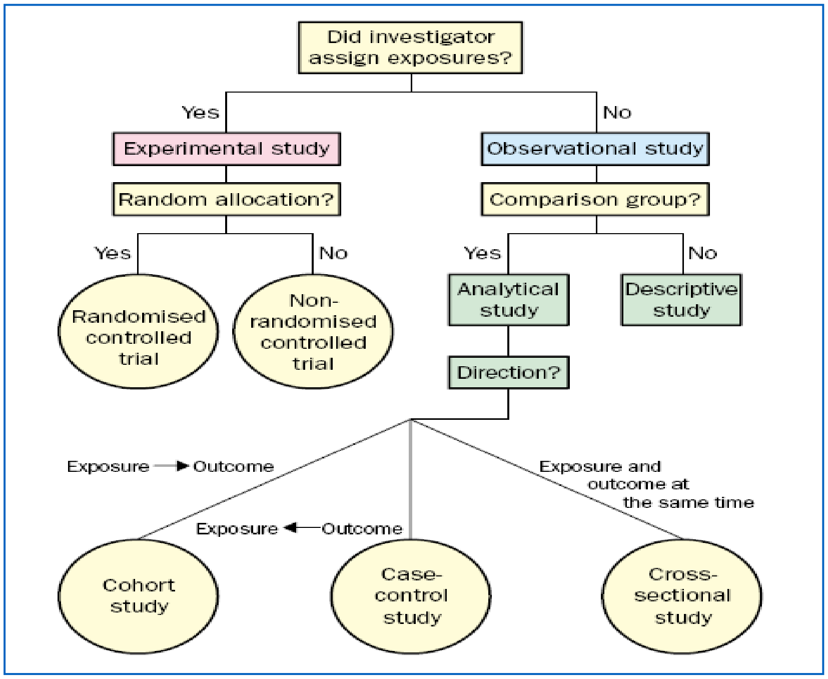
[37] John M. Last, *Dictionary of Epidemiology*, 1988.

- Low exposures to asbestos – lung cancer, asbestosis
- Asbestos – gastrointestinal cancer
- Electromagnetic fields – leukemia, brain cancer

In some cases, significant uncertainty about causality remains, despite many studies, over a long period of time.

Epidemiologic Study Design

The chart below catalogues the general types of epidemiologic studies available to investigate claims of health effects.

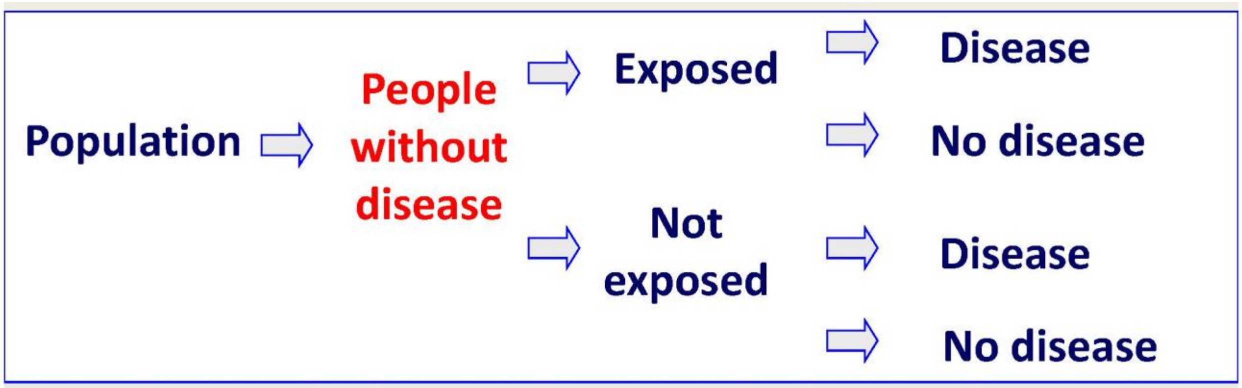


SOURCE: Reprinted from The Lancet, Vol. 359, Issue 9300, David A. Grimes and Kenneth F. Schulz, "An overview of clinical research: the lay of the land," 57, 58, Copyright © 2002, with permission from Elsevier.

Of course, not all studies are created equal. As discussed, the RCT attempts to control bias and confounding through randomization and double blinding, but even the RCT is threatened by noncompliance, loss to follow-up, and other problems. Still, the RCT is considered to be generally a more rigorous test of a hypothesis than any of the observational studies. Cohort and case-control studies are generally regarded as acceptable for causal analysis, whereas cross-sectional, which measure exposure and outcome at the same time, and ecological studies, which do not assess exposure and outcome in identified individuals, are seen as weaker designs not usually suitable for use in causal inference. Weaker studies may be sufficient only to generate a hypothesis, which then in turn would require a stronger study design to test the hypothesis generated for causality.

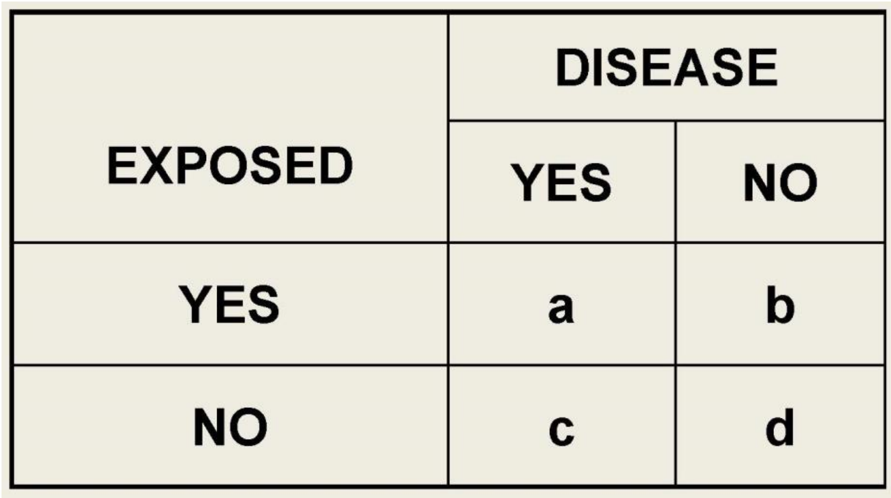
Analytical Epidemiologic Studies

Cohort Studies



SOURCE: Courtesy of the author.

At the heart of the cohort study is a contingency table, or a “2 x 2” table, which shows that the exposed mothers may have children with or without birth defects, and similarly, the unexposed mothers may have children with or without birth defects:



EXPOSED	DISEASE	
	YES	NO
YES	a	b
NO	c	d

SOURCE: Courtesy of the author.

The rate of birth defects in the children of exposed mothers will be numerically a/(a+b).

Similarly, the rate of birth defects in the children of unexposed mothers will be numerically c/(c+d).

These rates are sometimes also called, confusingly, “risks.” If the exposure of mothers to the medication under study does change the rate of birth defects from the (expected) rate of birth defects among children of unexposed mothers, then the rates or the risks would be the same. If the rates were the same, the “risk difference” between the two groups would be zero.

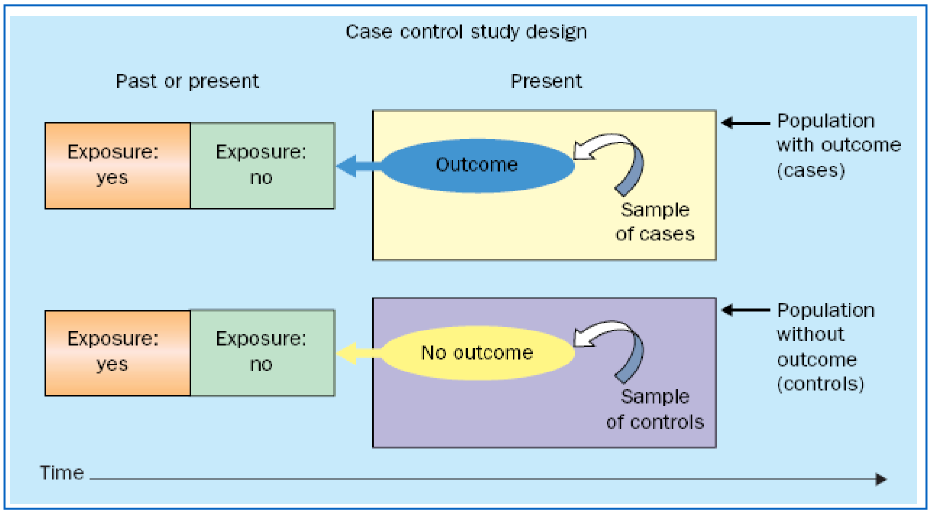
Sometimes, instead of looking at the absolute difference in the rates (the risk difference), epidemiologists look at the ratio of the two rates, for a measure called the “relative risk”:

$$\text{Relative Risk} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

If rates or risks are the same, then the ratio will be equal to one. If the rate of birth defects is higher among the children of mothers who ingested the medication under study, then the ratio will be larger than one. If the rate of birth defects among the children of mothers who ingested the medication under study is less than expected, then the ratio will be smaller than one.

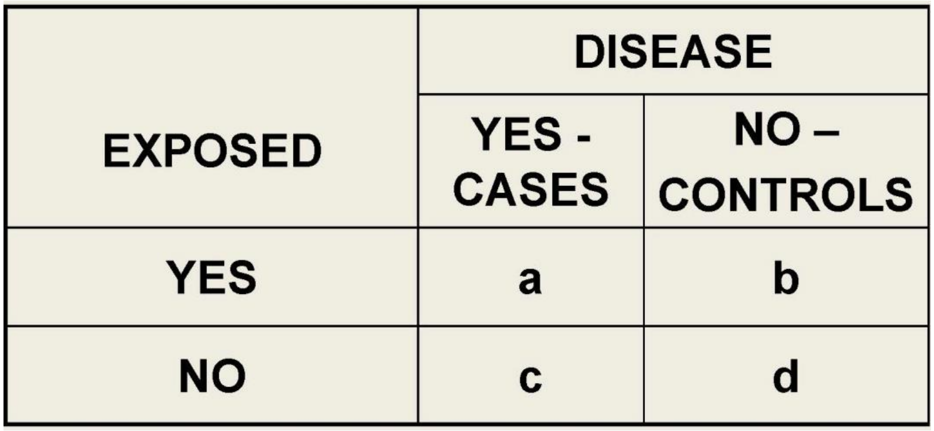
Case-control Studies

The retrospective aspect of a case-control study is shown in the diagram below.



SOURCE: Reprinted from The Lancet, Vol. 359, Issue 9304, Kenneth F. Schulz and David A. Grimes, “Case-control studies: research in reverse,” 431, 432, Copyright © 2002, with permission from Elsevier.

Like the cohort study, the case-control study also yields a 2 x 2 contingency table, from which we can calculate a measure of “risk,” called the odds ratio:



EXPOSED	DISEASE	
	YES - CASES	NO – CONTROLS
YES	a	b
NO	c	d

SOURCE: Courtesy of the author.

$$OddsRatio = \frac{\frac{a}{c}}{\frac{b}{d}} = ad / bc$$

For rare outcomes, the case-control is more efficient in the sense that a cohort study might have to follow a sample of people for many years to see a single rare event, but a case-control study can collect cases of the rare event from a much larger population, and match the cases with appropriate controls.

The case-control study is not, however, without its problems. The selection of controls can create selection biases, and the backwards nature of establishing past exposures leads to recall bias.

The odds ratio is generally larger than the relative risk (more positive if greater than 1.0; more negative if less than 1.0). As *a* becomes small in relation to *b*, and *c* becomes small in relation to *d*, as will be the case for outcomes that are relatively rare in cohort studies, the odds ratio approximates the relative risk:

$$relative\ \ risk = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \approx \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} = odds\, ratio$$

For purposes of birth defects epidemiology, the odds ratio and the relative risk are often treated as interchangeable.

The Paroxetine Signal for Cardiac Birth Defects

In December 2005, the FDA issued a Public Health Advisory on paroxetine (Paxil) and cardiac birth defects.[38] Based upon two unpublished studies (at that time), the FDA changed the pregnancy category for paroxetine from C to D, but the agency did not infer that there was a causal connection between paroxetine and cardiac or any other birth defects. The FDA's advisory did not mention any other SSRI antidepressant to be of concern. The agency's analysis, from its 2005 advisory, is set out below:

> In a study using Swedish national registry data, women who received paroxetine in early pregnancy had an approximately 2-fold increased risk for having an infant with a cardiac defect compared to the entire national registry population (the risk of a cardiac defect was about 2% in paroxetine-exposed infants vs. 1% among all registry infants).
>
> In a separate study using a United States insurance claims database, infants of women who received paroxetine in the first trimester had a 1.5-fold increased risk for cardiac malformations and a 1.8-fold increased risk for congenital malformations overall compared to infants of women who received other antidepressants in the first trimester. The risk of a cardiac defect was about 1.5% in paroxetine-exposed infants vs. 1% among infants exposed to other antidepressants.
>
> Most of the cardiac defects observed in these studies were atrial or ventricular septal defects, conditions in which the wall between the right and left sides of the heart is not completely developed. In general, septal defects are one of the most common type of congenital malformations. They range from those that are symptomatic and may require surgery to those that are asymptomatic and may resolve on their own. It is of note that the data in these studies was limited to first trimester exposures only, and there are not currently data to address whether this or any other risk extends to later periods of pregnancy.

CLASS DISCUSSION TOPIC

Why did the FDA regard a study that used "infants of women who received other antidepressants in the first trimester" as a control group as appropriate for its analysis?

[The study's design with a control group of infants exposed to other antidepressants other than paroxetine accomplished two goals. First, it showed that paroxetine was potentially different from the other antidepressants and guided the FDA in its decision to call for new warnings for paroxetine, but not for other antidepressants. Furthermore, in this area of epidemiology there is a strong concern about "indication bias," or "confounding by indication." If women who are sufficiently depressed or anxious to be taking an antidepressant, then their other lifestyle exposures and behaviors may also be related to the rate of birth defects among their offspring. By using a control group of infants of mothers who have been taking medications with similar

[38] FDA Public Health Advisory: Paroxetine (Dec. 8, 2005), archived at http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm051731.htm.

indications, the results would seem to have less opportunity to be confounded by the underlying condition of depression and its concomitant behaviors and exposures.]

PRACTICE ASSIGNMENT TO BE COMPLETED BEFORE NEXT CLASS

1. In a large national database that collected prescription medication usage and health outcomes (of users as well as their offspring), scientists collect the following information. Out of 100,000 singleton (non-twin) babies born one year, 3,000 infants were exposed in utero to SSRIs as a result of their mothers' use of that medication. There were 1,000 babies born with serious congenital heart defects; with 35 of them to the mothers who had used SSRIs medications in their first trimesters. What was the relative risk for cardiac birth defects? Does this study represent an increased risk? Did the mothers' use of SSRI medications cause cardiac defects in their children? Did the mothers' use of SSRI medications cause all 40 of the observed cardiac defects?
2. A consortium of teaching hospitals in several large U.S. cities, neonatologists identify all children born with serious heart defects. For each child born with a serious heart defect, the researchers match the child to two born without a heart defect, on the same day, or the following day, in the same hospital. At the end of the study period, the researchers had collected 212 infants born with heart defects, and 424 born without heart defects. Of the mothers who gave birth to children with serious heart defects, six took SSRI drugs just before, or in the first 3 months of pregnancy. In the matched group of 424 infants, 10 mothers had used SSRIs during the same time window. What was the odds ratio for SSRI exposure and cardiac birth defects? Does this study represent an increased risk? Did the mothers' use of SSRI medications cause cardiac defects in their children? Did the mothers' use of SSRI medications cause all 16 of the observed cardiac defects?

READINGS FOR NEXT CLASS

Sura Alwan, Jennita Reefhuis, Sonja A. Rasmussen, Richard S. Olney, and Jan M. Friedman, "Use of Selective Serotonin-Reuptake Inhibitors in Pregnancy and the Risk of Birth Defects," 356 *New Engl. J. Med*. 2684 (2007).

Michael D. Green, D. Michal Freedman, and Leon Gordis, "Reference Guide on Epidemiology," in *Reference Manual on Scientific Evidence* 549 (3d ed., 2011).

Evaluating Associations for Causation

CLASS ACTIVITY: REVIEW ASSIGNMENT QUESTIONS

(1) In a large national database that collected prescription medication usage and health outcomes (of users as well as their offspring), scientists collect the following information. Out of 100,000 singleton (non-twin) babies born one year, 3,000 infants were exposed in utero to serotonin selective reuptake inhibitors (SSRIs) as a result of their mothers' use of that medication. There were 1,000 babies born with serious congenital heart defects; with 35 of them to the mothers who had used SSRIs medications in their first trimesters. What was the relative risk for cardiac birth defects? Does this study represent an increased risk? Did the mothers' use of SSRI medications cause cardiac defects in their children? Did the mothers' use of SSRI medications cause all 35 of the observed cardiac defects?

[There is a rate of cardiac malformations in both the exposed and the unexposed groups. The rate in the exposed infants of 3,000, is 35/(35 + 2,965) = 35/3,000. The rate in the unexposed group of 97,000, is 965/(965 + 96,035) = 965/97,000.]

The "relative rate" misleadingly called a relative risk (even though it may not represent a causal relationship) is the ratio of the rate in the exposed to the unexposed:

$$Relative\ Risk = \frac{rate\ in\ exposed\ group}{rate\ in\ unexposed\ group} = \frac{{}^{35}/_{3,000}}{{}^{965}/_{97,000}} = 1.17\ (rounding)$$

So roughly, this represents a 17% increase in the rate of cardiac birth defects among children of women who used SSRIs in their first trimesters.

If the increase were real, then the SSRI use would have been responsible for five "excess cases," which would not have happened if the rate in the unexposed group held for the exposed group as well.

Is the increase real? The relative risk, in this case 1.17, is sometimes referred to as an "increased risk" or a positive association, without any consideration of the validity of the study or the place of chance that may have involved given the sampling in the study. As we will see, discrepancies from an expected value are themselves expected. Discrepancies above or below an expected rate are evaluated for the play of chance. This relative risk, 1.17, is not statistically significant; and the two-sided 95% C.I., 0.84 to 1.64.[39]]

Is the comparison between the rate of heart birth defects in the SSRI-exposed and unexposed groups a fair comparison in terms of other potential differences in potential risk factors?

[39] Using an online calculator at http://www.hutchon.net/confidrr.htm; https://www.medcalc.org/calc/relative_risk.php.

[Probably not; but we gave you no information to tell whether the two groups were as much alike in every other respect other than SSRI usage.]

(2) A consortium of teaching hospitals in several large U.S. cities, neonatologists identify all children born with serious heart defects. For each child born with a serious heart defect, the researchers match the child to two born without a heart defect, on the same day, or the following day, in the same hospital. At the end of the study period, the researchers had collected 212 infants born with heart defects, and 424 born without heart defects. Of the mothers who gave birth to children with serious heart defects, six took SSRI drugs just before, or in the first 3 months of pregnancy. In the matched group of 424 infants, 10 mothers had used SSRIs during the same time window. What was the odds ratio for SSRI exposure and cardiac birth defects? Does this study represent an increased risk? Did the mothers' use of SSRI medications cause cardiac defects in their children? Did the mothers' use of SSRI medications cause all 16 of the observed cardiac defects?

	Cases	**Controls**
Exposed	**6**	**10**
Not Exposed	**206**	**414**
(Total)	**(212)**	**(424)**

[Again, we have a 2 x 2 "contingency" table with columns for cases and for controls, and rows for exposed and for unexposed. The cross-product ratio is the so-called odds ratio, which approximates the relative risk: 1.21. (95% C.I., 0.43 - 3.36).[40]

What does this roughly 21% in exposure rate among the cases over the controls mean?

[We observed six exposed cases when we might have expected only five. Because it is a rare outcome, the odds ratio can be taken as an approximation for the relative risk, and we can say that in the population of exposed infants, the rate of malformations of the heart is 21% higher than in the nonexposed infants.]

LECTURE NOTES

We are going to focus our attention on a specific maternal exposure, to the SSRI sertraline, and to cardiac birth defects. For next class, you will read six epidemiologic studies, with a view to the data on the exposure and outcome of our interest. The key data are typically presented in a table, such as the one below, which provides odds ratios or relative risks, and confidence intervals, and information about what potential confounding variables have been included in the "adjusted" risk ratios to address potential confounding. The cardiac defects in this study have been pooled and the adjusted odds ratio reported in the box created by the intersecting rectangles on the chart. The factors considered in the "adjusted" multiple variable analysis are highlighted below.

[40] Odds ratio calculator at http://vassarstats.net/odds2x2.html.

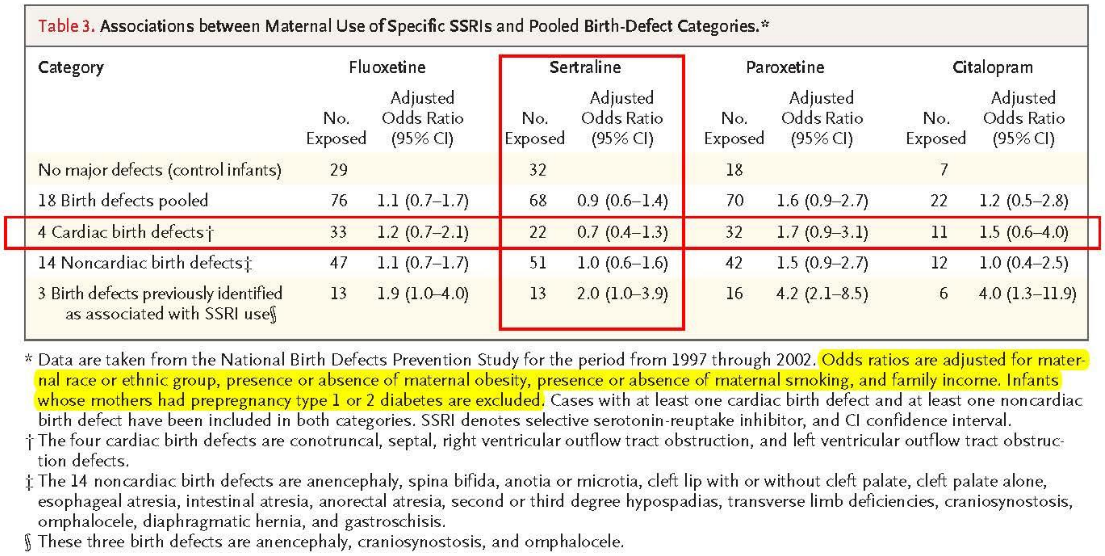


Table 3. **Associations between Maternal Use of Specific SSRIs and Pooled Birth-Defect Categories.***

Category	Fluoxetine		Sertraline		Paroxetine		Citalopram	
	No. Exposed	Adjusted Odds Ratio (95% CI)	No. Exposed	Adjusted Odds Ratio (95% CI)	No. Exposed	Adjusted Odds Ratio (95% CI)	No. Exposed	Adjusted Odds Ratio (95% CI)
No major defects (control infants)	29		32		18		7	
18 Birth defects pooled	76	1.1 (0.7–1.7)	68	0.9 (0.6–1.4)	70	1.6 (0.9–2.7)	22	1.2 (0.5–2.8)
4 Cardiac birth defects†	33	1.2 (0.7–2.1)	22	0.7 (0.4–1.3)	32	1.7 (0.9–3.1)	11	1.5 (0.6–4.0)
14 Noncardiac birth defects‡	47	1.1 (0.7–1.7)	51	1.0 (0.6–1.6)	42	1.5 (0.9–2.7)	12	1.0 (0.4–2.5)
3 Birth defects previously identified as associated with SSRI use§	13	1.9 (1.0–4.0)	13	2.0 (1.0–3.9)	16	4.2 (2.1–8.5)	6	4.0 (1.3–11.9)

* Data are taken from the National Birth Defects Prevention Study for the period from 1997 through 2002. Odds ratios are adjusted for maternal race or ethnic group, presence or absence of maternal obesity, presence or absence of maternal smoking, and family income. Infants whose mothers had prepregnancy type 1 or 2 diabetes are excluded. Cases with at least one cardiac birth defect and at least one noncardiac birth defect have been included in both categories. SSRI denotes selective serotonin-reuptake inhibitor, and CI confidence interval.

† The four cardiac birth defects are conotruncal, septal, right ventricular outflow tract obstruction, and left ventricular outflow tract obstruction defects.

‡ The 14 noncardiac birth defects are anencephaly, spina bifida, anotia or microtia, cleft lip with or without cleft palate, cleft palate alone, esophageal atresia, intestinal atresia, anorectal atresia, second or third degree hypospadias, transverse limb deficiencies, craniosynostosis, omphalocele, diaphragmatic hernia, and gastroschisis.

§ These three birth defects are anencephaly, craniosynostosis, and omphalocele.

SOURCE: From The NEW ENGLAND JOURNAL of MEDICINE, Sura Alwan, Jennita Reefhuis, Sonja A. Rasmussen, Richard S. Olney, and Jan M. Friedman, "Use of Selective Serotonin-Reuptake Inhibitors in Pregnancy and the Risk of Birth Defects," Volume No. 356, Page No. 2684, 2691. Copyright © (2007) Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

As we dig into actual data, there are some important methodological considerations to consider. The measure of risk ratio greater or less than expected suggests that the exposure variable is related in some way to the outcome variable. This correlation of exposure and outcome is sometimes called an "association." An association may be coincidental, or the result of error in drawing samples from the population that are "biased," or it may be the result of a third variable that is independently associated with both the exposure and outcome variables in a study. Sometimes an association is thus said to be a candidate for a causal association when we can reasonably rule out chance, bias, and confounding.

The relevant concept or model of causation for common birth defects by a medication is usually one in which the use of the medication during a vulnerable window of embryogenesis results in an increased number of birth defects over the background population rate. Birth defects will occur without the medication, and not all children of exposed mothers will have birth defects. The measures of risk, whether the relative risk or risk difference, provide some measure of the strength of any putative causal relationship between exposure and outcome.

Sir Austin Bradford Hill, in a famous address to the Royal Society of Medicine, outlined nine considerations to be considered whether an association was causal. Hill's starting point, however, for those considerations were that:

> [o]ur observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance.

—Austin Bradford Hill, "The Environment and Disease: Association or Causation?" 58 *Proc. Royal Soc'y Med*. 295, 295 (1965).[41] Implicit in "perfectly clear-cut," is that the association is not created by biased sampling or confounding. Implicit in Hill's beyond "the play of chance," is some assurance that the observed association is not a coincidence.

Bias

The validity of an experiment or an observational study is the ability to produce true or correct answers. We would expect that the average or the measured proportion from a random sample to provide an unbiased estimate of the population average or proportion. Biases in the sampling, however, may create threats to the validity of the estimates. Some of the common forms of bias that may affect the validity of epidemiologic studies include:

Referral bias: People known to be exposed to a drug or other exposure and to have the disease are more likely to be referred for inclusion in studies and reported.

Disease ascertainment bias: The disease is more likely to be searched for and diagnosed in subjects known to have been exposed.

Exposure ascertainment bias: The exposure is more likely to be searched for and found in subjects with the disease.

Attrition bias: Subjects with the exposure of interest are less likely to drop out of research and so more likely to be followed up and be analyzed.

Indication bias: The underlying disease for which drug is indicated is causing the complication rather than the drug itself.

Interviewer bias: The interviewer prompts or over-records exposure to the chemical of interest in people with disease compared to healthy respondents.

Recall bias: Control subjects (who do not have the disease under study) are less likely than case subjects to recall exposures.

Publication bias: Researchers are more likely to write up and submit studies for publication (and editors to publish) when associations ($p < 0.05$) have been found between chemical or drug exposures and diseases than when the researchers fail to obtain such results.

These systematic errors in reporting, collecting, analyzing, and reporting data are usually interpreted qualitatively, although recent efforts have attempted to state "good practices" for evaluating biases quantitatively for how much of an observed association may be the result of biases.[42] Biases may create the appearance of an association when there is none, or they may obscure or diminish the appearance of true associations. Bias, in the sense of systematic error, should be distinguished from investigator or study sponsor's conflicts of interest, and cognitive biases, such as hindsight or confirmation bias, which distorts our judgment of evidence, or conclusions drawn from the evidence.

[41] This classic paper is available online at http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1898525&blobtype=pdf, last visited March 1, 2015.

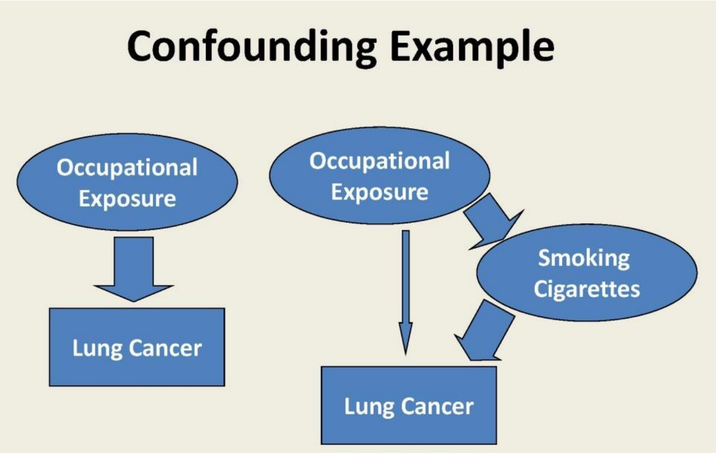
[42] See, e.g., Timothy L. Lash et al., "Good practices for quantitative bias analysis," 43 *Internat'l J. Epidem.* 1969 (2014).

Confounding

Confounding occurs in a study when there is another exposure in the sample that is associated with both the exposure and outcome variables. Sometimes the confounder is called a "lurking" variable because it is present, but its role in producing the outcome is not taken into account. An example may illustrate how a confounder can produce the erroneous appearance of an association.

Consider a study that looks at workers at a particular factory where they have exposures to some chemical. The question is whether this chemical is associated with an increased rate or mortality from lung cancer. The investigators compare the rate of lung cancer of the factory workers with the rate among an unexposed comparison population. Suppose the study finds roughly a doubling of lung cancer rates among the factory workers.

The study of these factory workers may be seriously flawed, however, if the investigators did not take into account the tobacco smoking of the workers as well as their workplace exposures. If an increased rate of tobacco consumption is associated with being a worker in this factory, and if smoking is also (causally) associated with lung cancer, then the workers' smoking may well explain some or all of their apparent increase rate of lung cancer. Smoking is not only a known cause of lung cancer; it is a very strong cause. Modest smoking over a couple of decades can create relative risks of 20-fold or greater. Such a strong confounding variable, if much more prevalent among the workers than among the controls, could easily create the appearance of a doubling of risk even when the chemical exposures in the factory had no propensity to increase the rate of lung cancer.



SOURCE: Courtesy of the author.

Confounding, when not accounted for, may create incorrect conclusions about associations and causations. In the context of birth defects epidemiology, confounding is also a particularly important threat to validity. Rates of birth defects increase with maternal age. These rates are increased by various maternal diseases and disorders, including viral infections, especially rubella, diabetes, and obesity. The rates may be increased for various lifestyle variables including diet, vitamin use, folic acid deficiency, smoking, alcohol consumption, and recreational drug use. Rates may vary among races and ethnic groups. Finally, some variables, such as social class or income may be "proxy" variables for other differences in exposures that lead to differences in birth defect rates. Indeed, depression itself may be a "proxy" for a range of other exposure variables, including increases in obesity, drug use, smoking, alcohol consumption, poor diet, and noncompliance with prenatal care, which may otherwise go unmeasured, or measured inaccurately, in epidemiologic studies.

There are several ways to address potential confounding in an epidemiologic study. Randomization of patients to treatment and placebo groups is, as we discussed, not available in the context of studying birth defects. Studies could restrict participants to those with particular characteristics, or to match cases and controls in a way to distribute confounding variables evenly in the groups.

Perhaps the most common approach is to use what is known as a multivariable analysis. Most of the sertraline birth defect studies use what is known as a multivariable regression model in which other exposure variables, such as age, body mass index, and smoking, are included to try to segregate out the contribution of the sertraline use alone. These additional variables lead to an "adjusted" risk ratio that identifies the contribution of sertraline to the birth defect outcome under study. Even this adjusted risk ratio may be incorrect because of inaccuracies in the data for one or more of the variables in the multivariate analysis. Adjustments cannot be made for important variables when the investigators have no measurements of missing variables. Furthermore, "residual" confounding created by unknown variables lurking in the sample will remain unless the investigators are in the unlikely situation of being able to identify and measure all variables that will contribute to the overall measure of risk.

Ruling Out Chance as a Likely Explanation for Sample Results

In studying an issue such as birth defects, we do not have the opportunity to measure exposures and outcomes in the entire population of women and their offspring. Instead, studies look of samples of mother-child pairs, from which investigators attempt to draw inferences about what is happening in the population itself. Scientists sometimes call the population measure the "parameter."

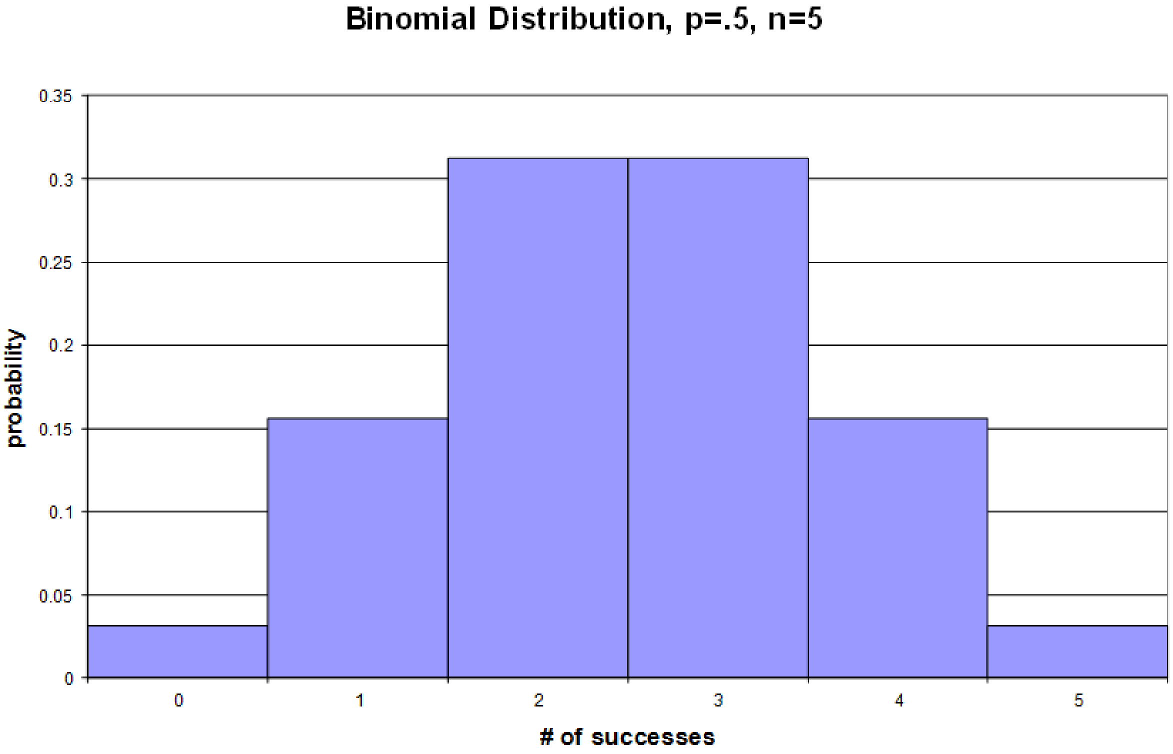
What did Sir Austin Bradford Hill mean by positing the starting point for assessing causality in an association "beyond what we would care to attribute to the play of chance"?[43]

In working with data that is based upon a process that has a fixed success rate, with each event independent of the others, the data will have a predictable expected success rate, and also a predictable, probabilistic variability and distribution. A familiar example can make these ideas clearer.

Toss a coin, with a fixed success rate of obtaining heads, which we might reasonably postulate is 50%. The failure rate is also 50%, since heads and tails are mutually exclusive, and

[43] Austin Bradford Hill, "The Environment and Disease: Association or Causation?" 58 *Proc. Royal Soc'y Med*. 295, 295 (1965).

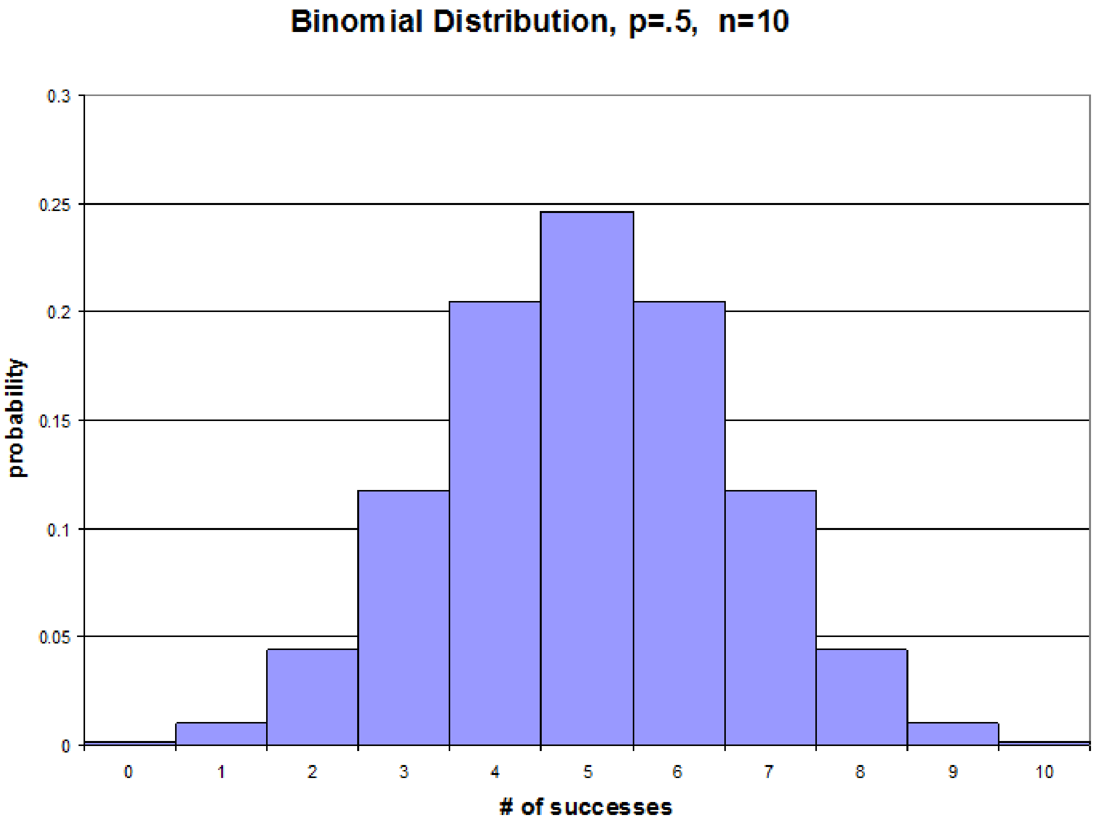
together they make up all the events that can come out of a single toss of a coin. Some interesting, nonobvious things follow from this probability "model," which is called a *binomial distribution*, one of which is that we can calculate the exact probability of obtaining a specific outcome, once we know the number of independent trials, and the fixed rate of success.
If we toss the coin five times (n = 5), assuming that the true probability of observing a heads on any given toss is 50% (p = 0.5), we can graph the exact probability for obtaining each possible outcome (0H/5T; 1H/4T; 2H/3T; 3H/2T; 4H/1T; and 5H/0T:



SOURCE: Courtesy of the author.

This graph shows that obtaining 2H/3T and 3H/2T are equally likely, a little more than 30% of the time, but obtaining a more extreme result, say no heads, or no tails, is much less likely, with each having a probability of less than 5%, respectively.

If we increase the number of trials, or coin tosses, to 10, with the same fixed rate of 50% success for heads, again with each toss result being independent of the other, we see a probability distribution like the one below:



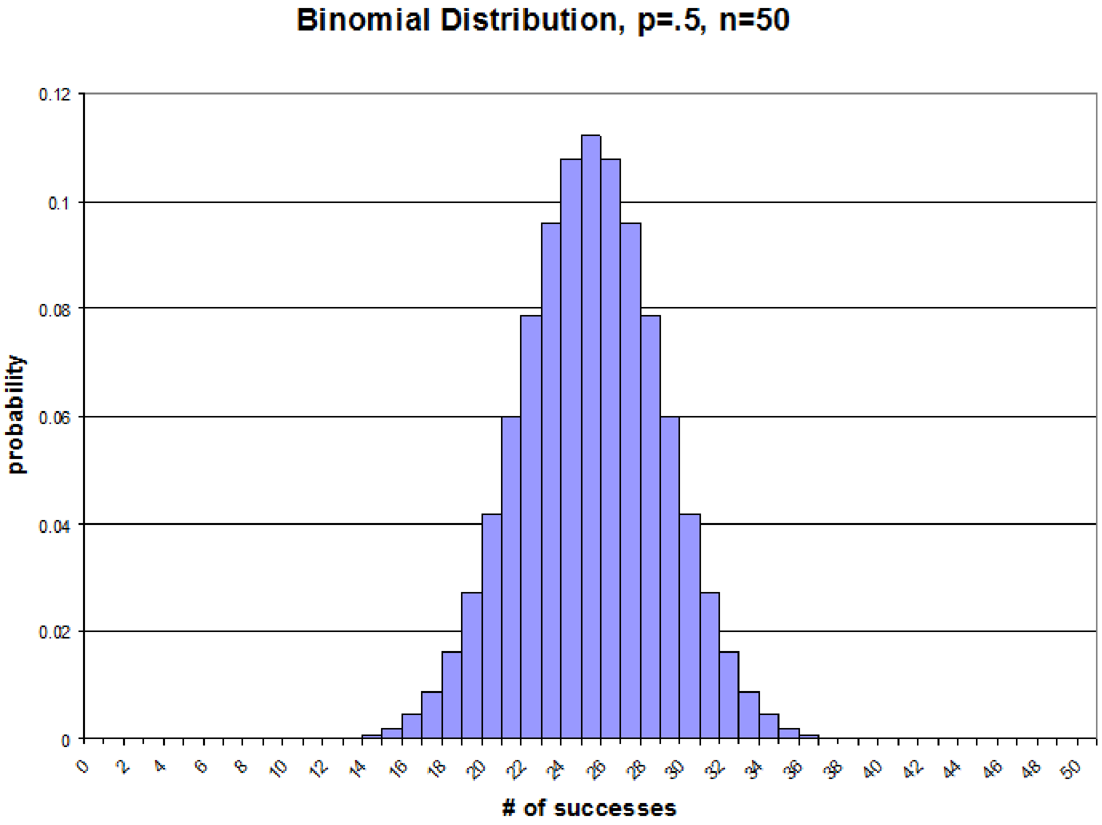
SOURCE: Courtesy of the author.

Again, our expected value is (fixed success rate) x (number of coin tosses) or 0.5 x 10 = 5. Indeed, five heads is the most likely outcome, but the probability of observing either 4H/6T or 6H/4T is almost twice as likely as observing our expected result of 5H/5T. And for a sample this size, 10 coin tosses, we are slightly more than three times as likely to observe a result in which the number of heads and tails are not equal to each other.

If we have randomly selected a sample, our sample average or sample proportion should be our "best estimate" of the population proportion, but our coin tossing examples illustrate how the sample estimate may be "best," and yet still a rather poor estimate of the population value because of random variability. Although we might have an "expected value," we also expect "random error" or "sampling error" to give us a result divergent from the expected value.

Scientists use a probability measure, called the *p-value*, for describing the probability of obtaining sample results as divergent from the expected value, or even more divergent, given the starting assumptions of the probability model and expected value. With our coin tossing examples, we have some experience in believing that the expected value based upon a rate of heads of 50%, but in scientific studies, we may have no basis to assume an expected rate. In studying groups of people (or animals), the assumption is often made that the rate is not different in those "exposed" and "unexposed." This assumption is frequently referred to as the "null hypothesis." If the sample data are sufficiently unlikely under this assumption, then scientists may have to revisit this starting assumption in favor of a belief that the rates are not the same for exposed and unexposed subjects. In other words, if the sample data show a difference sufficiently extreme from the expected, the scientists may reject the starting assumption of the null hypothesis.

Returning to our coin tossing example, we can see that as the number of coin tosses increase, the probability distribution becomes smoother, more bunched around the expected value, and more like the classic bell-shaped curve. The distribution below is based upon 50 tosses or trials, with an expected value of 50% heads.

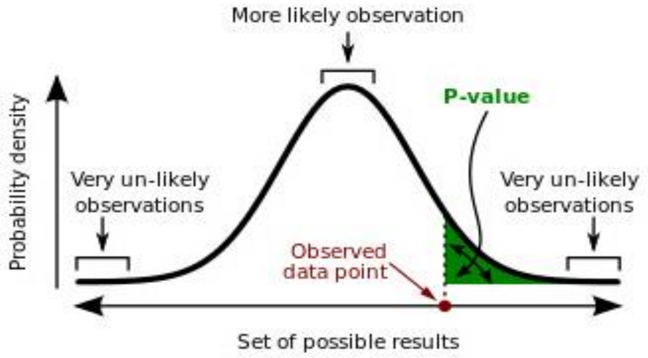


SOURCE: Courtesy of the author.

Because the probability distribution becomes even more “bunched up,” the probability of obtaining the expected value is smaller, but the probability of being off by a little bit, is reduced. Obtaining zero heads or zero tails in 50 tosses is extremely unlikely and does not show up on the graph. In our trial of five coin tosses, we could see that there was about a 3% chance of such an extreme outcome. The bunching of values around the expected value, in our graph of 50 coin tosses, reflects that we have gained a more precise estimate of the population value of the expected value, with the increased sample size.

Scientists can take advantage of the distribution’s becoming approximately a normal curve because the geometry of the normal curve is well known. An important measure of the “spread” of data, or its variability, is the standard deviation, which gives us a measure of the average deviation from the expected value in a sample. (We call this variability the standard error when we are estimating, or making inferences about, the population value from the sample data.) The known geometry of the normal bell-shaped curve tells us that the total probability of deviations at least 1.96 standard errors greater than the expected value, in both directions, will be 5%. Because the bell curve is symmetrical, 2.5% (or half of the 5%) of the data will be 1.96 standard errors below the sample estimate, and 2.5%, above.

The symmetry of the bell curve means that the divergence of the observed sample data from the expected value could be in the same direction as that observed or it could be in the opposite "tail" of the curve. A p-value that is based upon a calculation of the probability of observing a discrepancy at least as large as observed, given our starting assumptions, could be a "two-tailed" or a "one-tailed" p-value. The former will be the probability of the divergence in either direction and in either tail of the curve. The latter, the one-tailed p-value will be the divergence only in the direction seen in the sample data. In the illustration below, the green area represents a one-tailed p-value:[44]



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

SOURCE: Repapetilto and Chen-Pan Liao, "P-value" Wikipedia, online: https://en.wikipedia.org/wiki/P-value, licensed under a CC BBY-SA 3.0 license.

Because the bell curve is symmetrical, a two-tailed p-value will be twice the probability of the one-tail measurement. (Caveat: when we work with asymmetrical probability distributions, the tail probability in each direction may not be the same, and we cannot assume that the two-tail probability will be twice the one-tail probability.)

The p-value has two meanings in statistical practice. The lower the p-value, the stronger is the evidence against the starting assumption that there was no difference between the observed data and the null hypothesis. On this understanding, the p-value is a measure of support for an inference for rejecting the null hypothesis.

The p-value is also taken to provide a long-term frequency of rejecting the assumed null hypothesis, when the assumption is actually true. In other words, in our coin tossing example, the p-value provides the probability of observing results at least as divergent, if we continued

[44] Adapted from the Wikipedia article on p-values, available at http://en.wikipedia.org/wiki/P-value, last visited March 1, 2015.

to sample the population with samples of the same size, over many trials, given our starting assumptions and probability model.

The p-value probability calculation is often misunderstood. The p-value is the probability of observing the evidence obtained, or evidence even more extreme, given that there is no association. The probability is not the probability that there really is an association given the evidence. Notice how the second statement inverts the conditions, and this changes the meaning and the quantification of the probability at issue.

To give another illustrative example of how the conditions and their ordering matter, consider the probability that a person has breast cancer given that she is a woman. This probability is relatively low because the prevalence (women with breast cancer/all women) is small. On the other hand, the probability that a person is a woman given that the person has breast cancer is high, higher than 90%, because most breast cancers occur in women, and not in men.

Multiple Comparisons

The calculation of the p-value, as described above, assumes that we are making only one comparison from the observed data. If we make multiple comparisons from the same data set, comparing several exposure variables with several outcome variables as part of the same study, we increase the likelihood that we will see a false-positive outcome. As a result of the multiple comparisons, we should be less surprised to see a comparison that yields a low p-value. With sufficient number of comparisons within a single study, we would be surprised not to see some associations by chance alone. As we will see when we dig into some of the pertinent studies of SSRIs and birth defects, epidemiologic studies often make multiple comparisons between observed and expected outcomes in the data, and this multiplicity of testing changes our expectations for so-called false positive results. Frequently, the data come from an administrative database or from researchers' collection of cases and controls. Efficient use of the data resources leads to looking for many different associations in the dataset. It is not unusual for hundreds of possible associations to be evaluated from a dataset, with the implication that several of the observed associations, with p-values less than 5%, may be false positives that are artifacts of the multiple comparisons.

Confidence Intervals

In biomedical research, the use of p-values has given way to another measure of random error or variability about a sample proportion or sample mean, from observed sample data. As we saw from our example of tossing a coin 10 times, observing the "expected" value may have had the largest probability for a single result, but the probability of observing results that are off by a bit will be larger yet. Given the probability model of a fixed rate of success, with each outcome independent of the other, we can see that for any given trial of 10 tosses, we are likely to obtain a result that is off by some amount. The standard error provides a measure of this sampling error. The larger the sample, the smaller the standard error, and the more precise the estimate is. Random errors as large as, or even larger than the standard error, are not surprising when sampling from larger populations. When the observed data differ from the expected value by more than two or three standard errors, the original assumption about the expected value becomes suspect.

In estimating an average value or a proportion from a sample, taken from a much larger population, scientists compute a *confidence interval* (sometimes abbreviated "C.I.") around the sample average or sample proportion to show the range of possibilities for the underlying population parameter that is reasonably compatible with the sample data. A confidence interval for the population average or proportion is thus a range of values around the sample average, or the sample proportion, which we would interpret as reasonable estimates of the true population measure, given the sample data. Given the convention of using an alpha or a p-value of 5% or less as a basis for rejecting an assumed null hypothesis, scientists typically use a confidence interval that shows the range of values that would not be rejected as potential population values given the sample data. This confidence interval is described by its coefficient of confidence as (one minus alpha), which works out to 95% confidence when the customary alpha of 5% is used. When the data are known or assumed to be at least approximately "normal" in having a bell-shaped distribution, we know for such a curve, that the sample estimate, plus or minus 1.96 standard errors, will cut off 95% of the data distribution. An advantage of working with confidence intervals is that they focus our attention on the sample estimate (sometimes called the *point estimate*), and the expected variability around that estimate. When the interval is wide, we can tell directly that we have an imprecise estimate from our sample.

One common error in interpreting a reported confidence interval is to opine that there is a 95% probability that the true population value is within that observed interval. The term *confidence* is a technical term, which means that there is a 95% chance that if the sampling and interval construction process were repeated many times, 95% of all the intervals generated would include the true population value[45].

Whether we discuss precision of a sample estimate in terms of p-values or with confidence intervals, we are limiting our focus to one kind of "error," sampling or random error in our sample estimate. These measures do not take into account bias or confounding, or the accuracy of the data, data processing, the underlying probability model, or the appropriateness of the statistical test. For observational studies, such as the epidemiologic studies of birth defects, our total error rates would be considerably higher than what is reflected in the confidence interval.

CLASS ASSIGNMENT

Reading for Next Class

The following six papers are all peer-reviewed epidemiologic studies that address some of the questions raised by SSRI antidepressants and birth defects. Our focus in reviewing and discussing these papers will be the data, analysis, and conclusions with respect to a specific

[45] See *Reference Manual on Scientific Evidence* at 247 (3d ed., 2011).

SSRI, sertraline (Zoloft), and an organ-specific outcome, cardiac birth defects. Note, however, that many of these studies examined several other medications and many other birth defects.

Sura Alwan, Jennita Reefhuis, Sonja A. Rasmussen, Richard S. Olney, and Jan M. Friedman, “Use of Selective Serotonin-Reuptake Inhibitors in Pregnancy and the Risk of Birth Defects,” 356 *New Engl. J. Med*. 2684 (2007).

Carol Louik, Angela E. Lin, Martha M. Werler, Sonia Hernández-Díaz, and Allen A. Mitchell, “First-Trimester Use of Selective Serotonin-Reuptake Inhibitors and the Risk of Birth Defects,” 356 *New Engl. J. Med*. 2675 (2007).

Christina L. Wichman, Katherine M. Moore, Tara R. Lang, Jennifer L. St. Sauver, Robert H. Heise, Jr., and William J. Watson, “Congenital Heart Disease Associated With Selective Serotonin Reuptake Inhibitor Use During Pregnancy,” 84 *Mayo Clin. Proc.* 23 (2009).

Lars Henning Pedersen, Tine Brink Henriksen, Mogens Vestergaard, Jørn Olsen, and Bodil Hammer Bech, “Selective serotonin reuptake inhibitors in pregnancy and congenital malformations: population based cohort study,” 339 *Brit. Med. J*. b3569 (2009).

Margarita Reis and Bengt Källén, “Delivery outcome after maternal use of antidepressant drugs in pregnancy: an update using Swedish data,” 40 *Psychological Med*. 1723 (2010).

Heli Malm, Miia Artama, Mika Gissler, and Annukka Ritvanen, “Selective Serotonin Reuptake Inhibitors and Risk for Major Congenital Anomalies,” 118 *Obstet. & Gynecol*. 111 (2011).

The Human Data on Sertraline and Cardiac Birth Defects

INSTRUCTOR'S NOTE

The module includes these six studies because these are the studies that are included in the Myles meta-analysis, discussed in the next session. There are other relevant studies published before and after the Myles meta-analysis was done, but some restraints had to be placed upon how much the students would read, and drawing the limit at what was included in Myles is a compromise. Even with only six studies, the module requires a fair amount of preparation out of class before the data can be intelligently discussed in class.

The suggested class discussion questions are designed to direct the students to the key information that they need to be able to identify and discuss:

- Study design
- Sample size
- Precision of point estimates
- Lumping and splitting decisions
- Selection of control or comparison groups
- Observed number and rate of birth defects
- Source of information about use of antidepressant
- Adjustments for confounders in the study and differences among studies
- Study limitations
- Apparent class effect
- Study interpretation

Because confounding and bias are such large features of this kind of epidemiology, and because they are often unaccounted for in the provided p-values or confidence intervals, the students' attention should be focused on what covariates are included in the analysis of each study's data.

CLASS ACTIVITY: DISCUSSION OF SELECTED ARTICLES WITH DATA ON SERTRALINE (ZOLOFT) AND HEART BIRTH DEFECTS

The Case-control Studies

(1) Alwan (2007)[46]

Let's begin by returning to the study by Alwan and her coworkers.

[46] Sura Alwan, Jennita Reefhuis, Sonja A. Rasmussen, Richard S. Olney, and Jan M. Friedman, "Use of Selective Serotonin-Reuptake Inhibitors in Pregnancy and the Risk of Birth Defects," 356 *New Engl. J. Med*. 2684 (2007).

What kind of study was done by Alwan?

[Case control study.]

Did the study reach any causal conclusions?

[No; indeed, none of the studies reached a causal conclusion, and any such conclusion in the context of a single study would have been a truly extraordinary occurrence in the world of epidemiology.]

Did Alwan find an increase in the rate (or "risk") of cardiac defects among children born to mothers who used sertraline in the first trimester of pregnancy?

[No; the "point estimate" was actually lower than one, representing a decrease, not an increase, in rate of cardiac birth defects.]

Did Alwan achieve a precise answer to the question whether sertraline was associated with an increase or decrease rate of cardiac birth defects?

[Our best sense of the precision of the relevant point estimate is given by the 95% confidence interval, which ranges from 0.4 to 1.3. A 30% increase in the rate of cardiac defects is still compatible with the study data, but so is a 60% decrease.]

How did Alwan address the issue of the right level of outcome specificity—lumping and splitting?

[Alwan combined several different cardiac defects in her main presentation, but in a data supplement, she broke out the four different cardiac subgroups and presented the adjusted odds ratios and confidence interval for each subgroup.]

For what other exposures, lifestyle factors, and considerations did Alwan adjust by including in her multiple variable model?

[Alwan adjusted for race and ethnicity, obesity, smoking, and family income. The authors did not adjust for maternal diabetes, but excluded such cases from their analyses.]

This paper, published in the *New England Journal of Medicine*, comes from investigators at the National Birth Defects Prevention Study, which has been conducting research on birth defects for a long time. The key data on sertraline and cardiac defects appears in Table 3 of the paper:

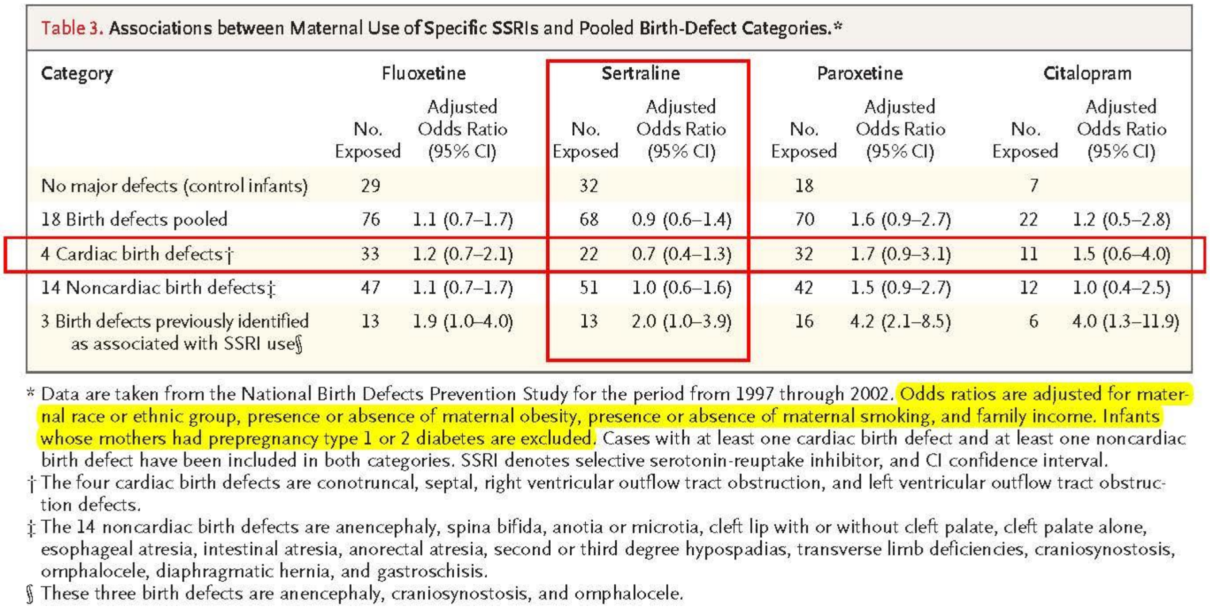


Table 3. Associations between Maternal Use of Specific SSRIs and Pooled Birth-Defect Categories.*

Category	Fluoxetine		Sertraline		Paroxetine		Citalopram	
	No. Exposed	Adjusted Odds Ratio (95% CI)	No. Exposed	Adjusted Odds Ratio (95% CI)	No. Exposed	Adjusted Odds Ratio (95% CI)	No. Exposed	Adjusted Odds Ratio (95% CI)
No major defects (control infants)	29		32		18		7	
18 Birth defects pooled	76	1.1 (0.7–1.7)	68	0.9 (0.6–1.4)	70	1.6 (0.9–2.7)	22	1.2 (0.5–2.8)
4 Cardiac birth defects†	33	1.2 (0.7–2.1)	22	0.7 (0.4–1.3)	32	1.7 (0.9–3.1)	11	1.5 (0.6–4.0)
14 Noncardiac birth defects‡	47	1.1 (0.7–1.7)	51	1.0 (0.6–1.6)	42	1.5 (0.9–2.7)	12	1.0 (0.4–2.5)
3 Birth defects previously identified as associated with SSRI use§	13	1.9 (1.0–4.0)	13	2.0 (1.0–3.9)	16	4.2 (2.1–8.5)	6	4.0 (1.3–11.9)

* Data are taken from the National Birth Defects Prevention Study for the period from 1997 through 2002. Odds ratios are adjusted for maternal race or ethnic group, presence or absence of maternal obesity, presence or absence of maternal smoking, and family income. Infants whose mothers had prepregnancy type 1 or 2 diabetes are excluded. Cases with at least one cardiac birth defect and at least one noncardiac birth defect have been included in both categories. SSRI denotes selective serotonin-reuptake inhibitor, and CI confidence interval.

† The four cardiac birth defects are conotruncal, septal, right ventricular outflow tract obstruction, and left ventricular outflow tract obstruction defects.

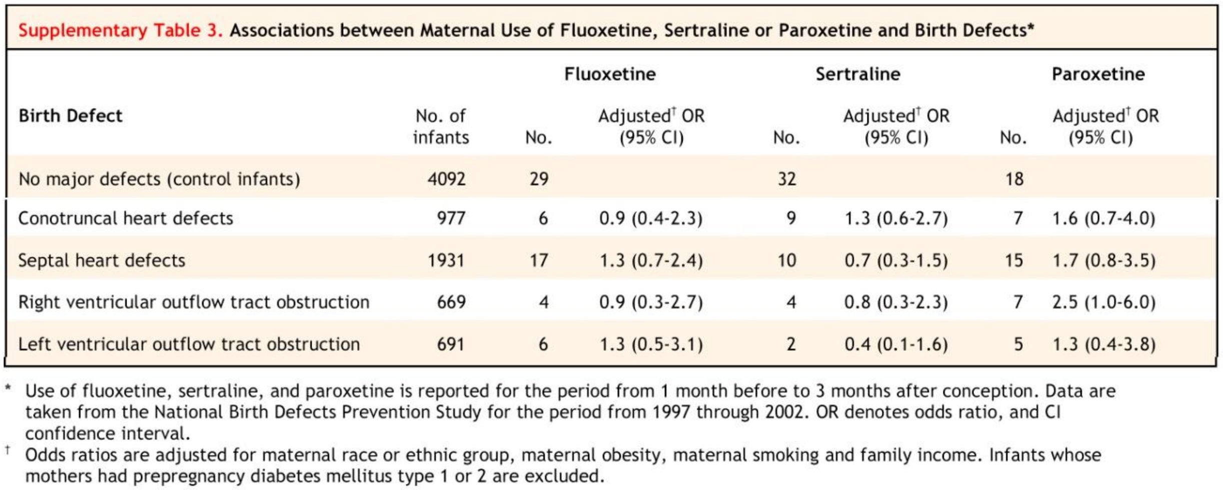
‡ The 14 noncardiac birth defects are anencephaly, spina bifida, anotia or microtia, cleft lip with or without cleft palate, cleft palate alone, esophageal atresia, intestinal atresia, anorectal atresia, second or third degree hypospadias, transverse limb deficiencies, craniosynostosis, omphalocele, diaphragmatic hernia, and gastroschisis.

§ These three birth defects are anencephaly, craniosynostosis, and omphalocele.

SOURCE: From The NEW ENGLAND JOURNAL of MEDICINE, Sura Alwan, Jennita Reefhuis, Sonja A. Rasmussen, Richard S. Olney, and Jan M. Friedman, "Use of Selective Serotonin-Reuptake Inhibitors in Pregnancy and the Risk of Birth Defects," Volume No. 356, Page No. 2684, 2691. Copyright © (2007) Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

Note that this table "lumps" four specific cardiac defects together. In addition to the general points for discussion, noted above, this table can be used to discuss whether the study supports a claim of class effect.

The following chart is adapted from Alwan's data supplement. Note how the specific kinds of heart defects have different point estimates, now with wider confidence intervals, which reflect the smaller size of the subgroups. Although the combined cardiac defect outcomes were actually a little fewer than expected, here we see one subgroup slightly increased, and three subgroups decreased.



Supplementary Table 3. Associations between Maternal Use of Fluoxetine, Sertraline or Paroxetine and Birth Defects*

Birth Defect	No. of infants	Fluoxetine No.	Fluoxetine Adjusted† OR (95% CI)	Sertraline No.	Sertraline Adjusted† OR (95% CI)	Paroxetine No.	Paroxetine Adjusted† OR (95% CI)
No major defects (control infants)	4092	29		32		18	
Conotruncal heart defects	977	6	0.9 (0.4-2.3)	9	1.3 (0.6-2.7)	7	1.6 (0.7-4.0)
Septal heart defects	1931	17	1.3 (0.7-2.4)	10	0.7 (0.3-1.5)	15	1.7 (0.8-3.5)
Right ventricular outflow tract obstruction	669	4	0.9 (0.3-2.7)	4	0.8 (0.3-2.3)	7	2.5 (1.0-6.0)
Left ventricular outflow tract obstruction	691	6	1.3 (0.5-3.1)	2	0.4 (0.1-1.6)	5	1.3 (0.4-3.8)

* Use of fluoxetine, sertraline, and paroxetine is reported for the period from 1 month before to 3 months after conception. Data are taken from the National Birth Defects Prevention Study for the period from 1997 through 2002. OR denotes odds ratio, and CI confidence interval.

† Odds ratios are adjusted for maternal race or ethnic group, maternal obesity, maternal smoking and family income. Infants whose mothers had prepregnancy diabetes mellitus type 1 or 2 are excluded.

SOURCE: From The NEW ENGLAND JOURNAL of MEDICINE, Sura Alwan, Jennita Reefhuis, Sonja A. Rasmussen, Richard S. Olney, and Jan M. Friedman, "Use of Selective Serotonin-Reuptake Inhibitors in Pregnancy and the Risk of Birth Defects," Volume No. 356, Page No. 2684. Copyright © (2007) Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

(2) Louik (2007)[47]

What kind of epidemiologic study was done by Louik and colleagues in their 2007 paper?

[Case control study. This is another case-control study, funded in part by GlaxoSmithKline, the sponsor of Paxil (paroxetine), with the encouragement of the Food and Drug Administration (FDA). The study was published in the same issue of the *New England Journal of Medicine*, as Alwan (2007).]

Was there overlap in the cases or the controls between Alwan and Louik?

[It is hard to be absolutely sure because unique case identifiers are always removed in such studies, but the cases in both studies were drawn from clinical centers that overlapped geographically. One study looked at statewide birth defect registries, and another looked at municipal birth defect registries, for cities within the states whose registries were also used.]

Did Louik and colleagues find any cardiac outcomes that were "statistically significant" at the traditional 5% level?

[Statistical significance with a p-value lower than 5% corresponds to a 95% confidence interval that excludes the null value of no association within the interval. For septal defects, Louik reported a doubling of risk, with an adjusted odds ratio of 2.0, and a 95% confidence interval

[47] Carol Louik, Angela E. Lin, Martha M. Werler, Sonia Hernández-Díaz, and Allen A. Mitchell, "First-Trimester Use of Selective Serotonin-Reuptake Inhibitors and the Risk of Birth Defects," 356 *New Engl. J. Med*. 2675 (2007).

that ran from 1.2 to 4.0. This excludes a risk ratio of 1.0, and would suggest that the result was nominally statistically significant. (There were no adjustments for multiple comparisons.) After the paper was published, one astute observer noted that the confidence interval could not be correctly calculated, and wrote the authors who recalculated the confidence interval. A correction was published with the confidence interval 1.0 to 4.0, and the result was labeled no longer statistically significant because the interval contained 1.0 (although just barely). See http://www.nejm.org/doi/full/10.1056/NEJMoa067407.]

Did Louik and her coauthors interpret any of their findings to be causal?

[No. Although media reports commonly contain language about a new study that "links" this with that, individual studies generally address only apparent associations, and, as we will see, there is a need for a much fuller and comprehensive analysis to address causation.]

What potential confounders did Louik and her colleagues include in their analysis of their data?

[Louik's consideration of confounders was the most extensive, and the only one to include alcohol consumption. In addition to alcohol, they included mother's race or ethnicity, mother's education, year pregnancy began, study center, first-trimester tobacco smoking, family history of birth defect in a first-degree relation, body-mass index before pregnancy, seizures (such as epilepsy), diabetes, hypertension, infertility, and first-trimester use of folic acid supplementation.]

The following table is adapted from Table 2 in the Louik study:

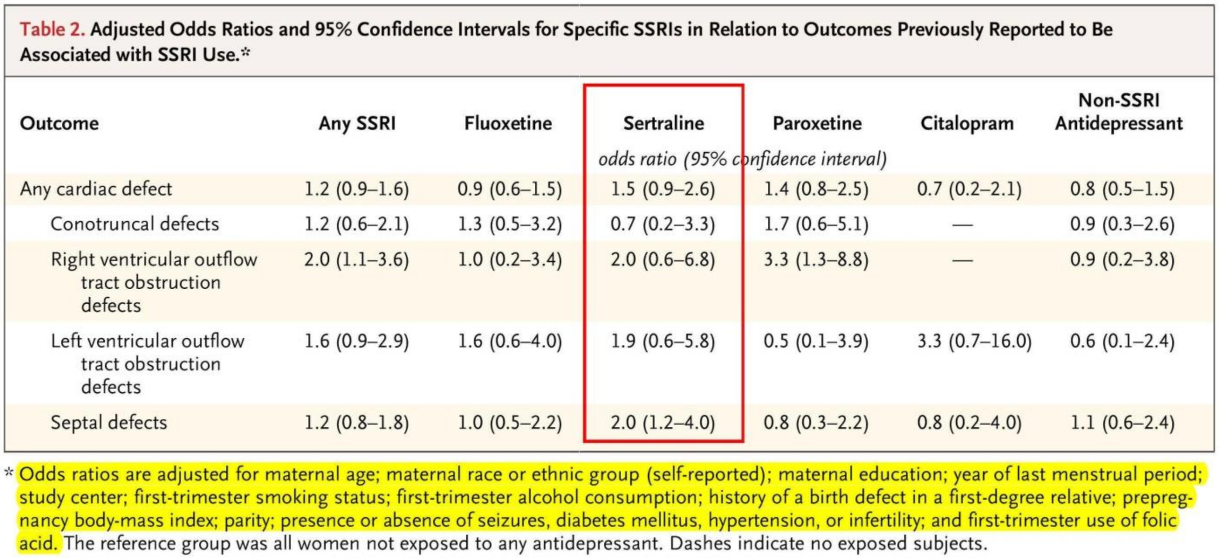


Table 2. Adjusted Odds Ratios and 95% Confidence Intervals for Specific SSRIs in Relation to Outcomes Previously Reported to Be Associated with SSRI Use.*

Outcome	Any SSRI	Fluoxetine	Sertraline	Paroxetine	Citalopram	Non-SSRI Antidepressant
	odds ratio (95% confidence interval)					
Any cardiac defect	1.2 (0.9–1.6)	0.9 (0.6–1.5)	1.5 (0.9–2.6)	1.4 (0.8–2.5)	0.7 (0.2–2.1)	0.8 (0.5–1.5)
Conotruncal defects	1.2 (0.6–2.1)	1.3 (0.5–3.2)	0.7 (0.2–3.3)	1.7 (0.6–5.1)	—	0.9 (0.3–2.6)
Right ventricular outflow tract obstruction defects	2.0 (1.1–3.6)	1.0 (0.2–3.4)	2.0 (0.6–6.8)	3.3 (1.3–8.8)	—	0.9 (0.2–3.8)
Left ventricular outflow tract obstruction defects	1.6 (0.9–2.9)	1.6 (0.6–4.0)	1.9 (0.6–5.8)	0.5 (0.1–3.9)	3.3 (0.7–16.0)	0.6 (0.1–2.4)
Septal defects	1.2 (0.8–1.8)	1.0 (0.5–2.2)	2.0 (1.2–4.0)	0.8 (0.3–2.2)	0.8 (0.2–4.0)	1.1 (0.6–2.4)

* Odds ratios are adjusted for maternal age; maternal race or ethnic group (self-reported); maternal education; year of last menstrual period; study center; first-trimester smoking status; first-trimester alcohol consumption; history of a birth defect in a first-degree relative; prepregnancy body-mass index; parity; presence or absence of seizures, diabetes mellitus, hypertension, or infertility; and first-trimester use of folic acid. The reference group was all women not exposed to any antidepressant. Dashes indicate no exposed subjects.

SOURCE: From The NEW ENGLAND JOURNAL of MEDICINE, Carol Louik, Sc.D., Angela E. Lin, M.D., Martha M. Werler, Sc.D., Sonia Hernández-Díaz, M.D., Sc.D., and Allen A. Mitchell, M.D., "First-Trimester Use of Selective Serotonin-Reuptake Inhibitors and the Risk of Birth Defects," Volume No. 356, Page No. 2675, 2679. Copyright © (2007) Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

(3) Wichman (2009)[48]

What kind of study was done by Wichman and her colleagues?

[This paper is the published result of a relatively small cohort, coming from the Mayo clinic.]

Do the authors report all cardiac defects together or alone?

[No; Wichman and colleagues reported only one relevant, specific outcome to our question, that is, ventricular septal defects. In all likelihood, given the small size of this cohort, the authors did not find enough of any other kind of cardiac defects to merit reporting another subcategory. Instead, they reported the zero for ventricular septal defects, and lumped all other "congenital heart disease (CHD)" outcomes together as "Other" in the SSRI-exposed and unexposed children. The authors also provided an analysis of "Total" congenital heart disease rates. They did not report risk ratios ventricular septal defects, for "Other," or "Total" CHD, but they did report the numbers in each category (3/808 for the SSRI mothers) and (181/24,406) for the mothers who had not used SSRIs, and they reported a p-value that described the discrepancy for "Other" as 0.29, or a 29% probability that we would have seen a discrepancy this large from expected, or larger, assuming that there was no true difference. The crude risk ratio for ventricular septal defect is zero, and for "Total," the crude risk ratio is (0.4/0.8 = 0.5, with a 95% C.I., 0.1609 to 1.5693.]

What potential or actual confounders are controlled for in Wichman's analysis?

[None.]

[48] Christina L. Wichman, Katherine M. Moore, Tara R. Lang, Jennifer L. St. Sauver, Robert H. Heise, Jr., and William J. Watson, "Congenital Heart Disease Associated With Selective Serotonin Reuptake Inhibitor Use During Pregnancy," 84 *Mayo Clin. Proc.* 23 (2009).

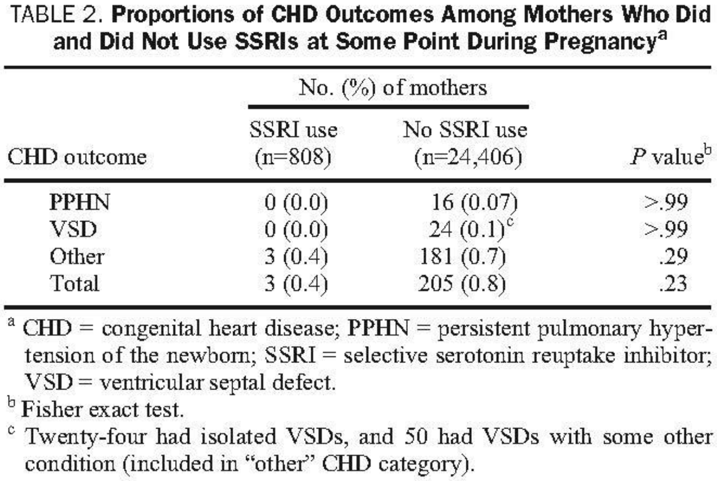


TABLE 2. **Proportions of CHD Outcomes Among Mothers Who Did and Did Not Use SSRIs at Some Point During Pregnancy**[a]

	No. (%) of mothers		
CHD outcome	SSRI use (n=808)	No SSRI use (n=24,406)	*P* value[b]
PPHN	0 (0.0)	16 (0.07)	>.99
VSD	0 (0.0)	24 (0.1)[c]	>.99
Other	3 (0.4)	181 (0.7)	.29
Total	3 (0.4)	205 (0.8)	.23

[a] CHD = congenital heart disease; PPHN = persistent pulmonary hypertension of the newborn; SSRI = selective serotonin reuptake inhibitor; VSD = ventricular septal defect.

[b] Fisher exact test.

[c] Twenty-four had isolated VSDs, and 50 had VSDs with some other condition (included in "other" CHD category).

SOURCE: Reprinted from Mayo Clinic Proceedings, Vol. 84, Issue 1, Christina L. Wichman, DO, Katherine M. Moore, MD, Tara R. Lang, MD, Jennifer L. St. Sauver, PhD, Robert H. Heise, Jr, MD, and William J. Watson, MD, "Congenital Heart Disease Associated With Selective Serotonin Reuptake Inhibitor Use During Pregnancy" p. 23-27, Copyright © (2009), with permission from Mayo Foundation for Medical Education and Research.

(4) Pedersen (2009)[49]

What kind of study was done by Pedersen?

[This is a large cohort study done in Denmark, one of three Scandinavian cohort studies. (There have been updated publications of the Danish cohort since this paper.) Some authors complain that epidemiology is expensive, and time consuming, but this complaint is not always true. In Scandinavia, national health care facilitates the collection of health data, with personal identities concealed, including maternal prescription drug use and infant health outcomes. Because these data are collected on an ongoing basis, the conduct of a study is readily accomplished by designing a program that will sort the available data to compare the rate defects among children of mothers who were prescribed SSRIs and among children whose mothers were not prescribed. Typically, the prescriptions identified are those in the first 3 months of pregnancy (first trimester) because this is the crucial period of organogenesis, in which birth defects arise.]

Did the Danish study achieve a precise estimate of the relative rate of cardiac defects among children born to mothers who had used sertraline?

[The key data on sertraline were included in a supplemental table, available from the publisher's website. From the supplemental data, Table B, we can see that there were only seven cardiac birth defect cases among mothers who used sertraline, which represented a risk ratio of 1.63, and a 95% confidence interval, 0.77 to 3.45. Of these cardiac defects, most of

[49] Lars Henning Pedersen, Tine Brink Henriksen, Mogens Vestergaard, Jørn Olsen, and Bodil Hammer Bech, "Selective serotonin reuptake inhibitors in pregnancy and congenital malformations: population based cohort study," 339 *Brit. Med. J.* b3569 (2009).

them were septal defects (openings in the walls between ventricles or atria). There were five such septal defects in sertraline-exposed infants, for a risk ratio of 2.01 (95% C.I., 0.83 to 4.86). Although not reported, we can see that for the nonseptal defects, the risk ratio had to be lower than 1.63, and the confidence interval had to wider. In any event, the risk ratios are elevated, but the confidence intervals are wide, showing a general lack of precision for these point estimates.]

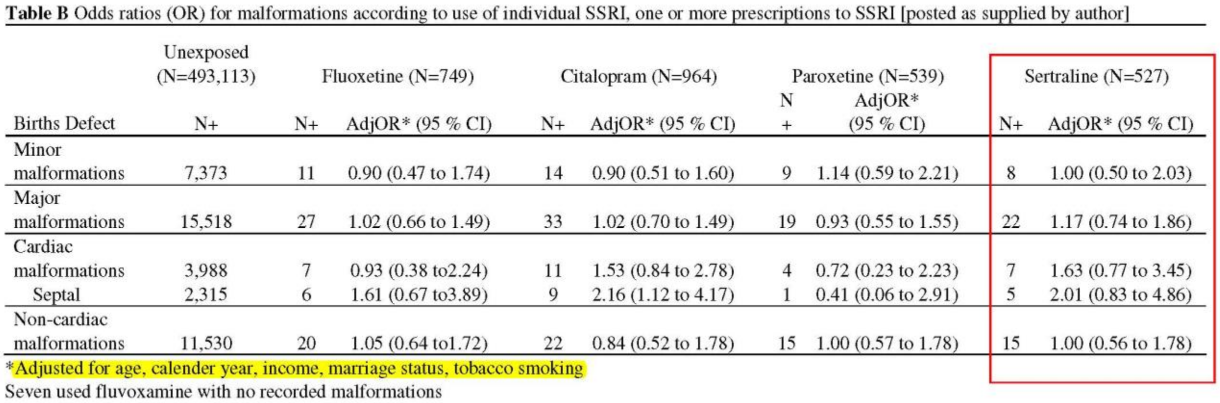


Table B Odds ratios (OR) for malformations according to use of individual SSRI, one or more prescriptions to SSRI [posted as supplied by author]

	Unexposed (N=493,113)	Fluoxetine (N=749)		Citalopram (N=964)		Paroxetine (N=539)		Sertraline (N=527)	
Births Defect	N+	N+	AdjOR* (95 % CI)	N+	AdjOR* (95 % CI)	N +	AdjOR* (95 % CI)	N+	AdjOR* (95 % CI)
Minor malformations	7,373	11	0.90 (0.47 to 1.74)	14	0.90 (0.51 to 1.60)	9	1.14 (0.59 to 2.21)	8	1.00 (0.50 to 2.03)
Major malformations	15,518	27	1.02 (0.66 to 1.49)	33	1.02 (0.70 to 1.49)	19	0.93 (0.55 to 1.55)	22	1.17 (0.74 to 1.86)
Cardiac malformations	3,988	7	0.93 (0.38 to2.24)	11	1.53 (0.84 to 2.78)	4	0.72 (0.23 to 2.23)	7	1.63 (0.77 to 3.45)
Septal	2,315	6	1.61 (0.67 to3.89)	9	2.16 (1.12 to 4.17)	1	0.41 (0.06 to 2.91)	5	2.01 (0.83 to 4.86)
Non-cardiac malformations	11,530	20	1.05 (0.64 to1.72)	22	0.84 (0.52 to 1.78)	15	1.00 (0.57 to 1.78)	15	1.00 (0.56 to 1.78)

*Adjusted for age, calender year, income, marriage status, tobacco smoking

Seven used fluvoxamine with no recorded malformations

SOURCE: Lars Henning Pedersen, Tine Brink Henriksen, Mogens Vestergaard, Jørn Olsen, and Bodil Hammer Bech, "Selective serotonin reuptake inhibitors in pregnancy and congenital malformations: population based cohort study," 339 Brit. Med. J. b3569 (2009).

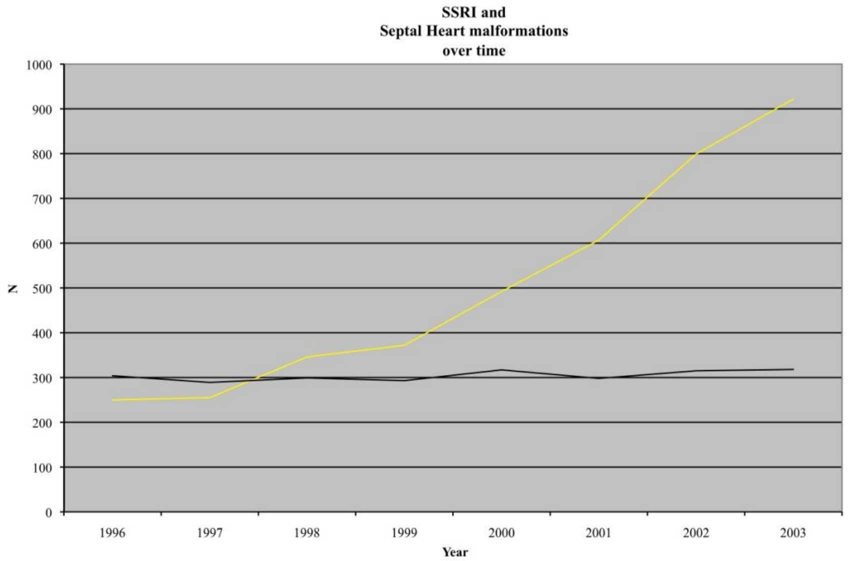
What other factors were considered as potential confounders and adjusted for in the Pedersen analysis?

[Age, calendar year, income, marriage status, and tobacco smoking.]

Why would income be included as a "covariable" in the Pedersen analysis?

[Income is considered a "proxy" for other behaviors, such as diet, alcohol consumption, and attention to prenatal care. It is obviously a very crude, and inexact, way to control for the real issues of concern.]

Also in the data supplement was what is known as a secular trend (or ecologic) analysis for all SSRI use and incidence of septal heart malformations, over time:



SOURCE: Lars Henning Pedersen, Tine Brink Henriksen, Mogens Vestergaard, Jørn Olsen, and Bodil Hammer Bech, "Selective serotonin reuptake inhibitors in pregnancy and congenital malformations: population based cohort study," 339 Brit. Med. J. b3569 (2009).

The yellow line shows the increasing rate of prescriptions of SSRIs between 1996 and 2003. The black line shows the rate of detected septal defects in sample.

Is this time-trend analysis persuasive evidence for or against a possible causal role of SSRIs in producing septal defects?

[Given that the putative risk ratios are not greatly elevated, and that the prevalence of SSRI use is low as a percentage of all children born, a true increase risk could easily be masked by other trends in the heart defect rates over time. The flat line representing a stable rate of annually reported septal defects does, however, weigh against SSRI use's being a serious public health issue for this outcome, in this national sample.]

(5) Reis (2010)[50]

What kind of study was done by Reis?

[This paper reflects data from a cohort study done in the Swedish health registries. An earlier version of these data was used by the FDA in explaining and justifying the Public Health Advisory over paroxetine (Paxil), and a more recent, updated data analysis has since been published in 2013.]

What did Reis and Källén find in terms of heart defects among children of mothers who used sertraline?

[50] Margarita Reis and Bengt Källén, "Delivery outcome after maternal use of antidepressant drugs in pregnancy: an update using Swedish data," 40 *Psychological Med*. 1723 (2010).

[The Swedish authors found a reduced rate of any heart defect, with a risk ratio of 0.74, and a confidence interval of (0.5 to 1.09).]

Were the Swedish results more or less precise than the Danish results?

[Sweden is a more populous country than Denmark, and the larger sample size allows for more precision in the point estimates, as reflected in the narrow confidence interval. Note that the point estimate found in Denmark would have been rejected as a possible null hypothesis by the data collected in Sweden.]

For what other potential confounders did Margarita Reis and Bengt Källén control in their analysis of the Swedish registry cohort?

[Year of birth, mother's age, parity (or number of children for this mother), smoking, and body-mass index.]

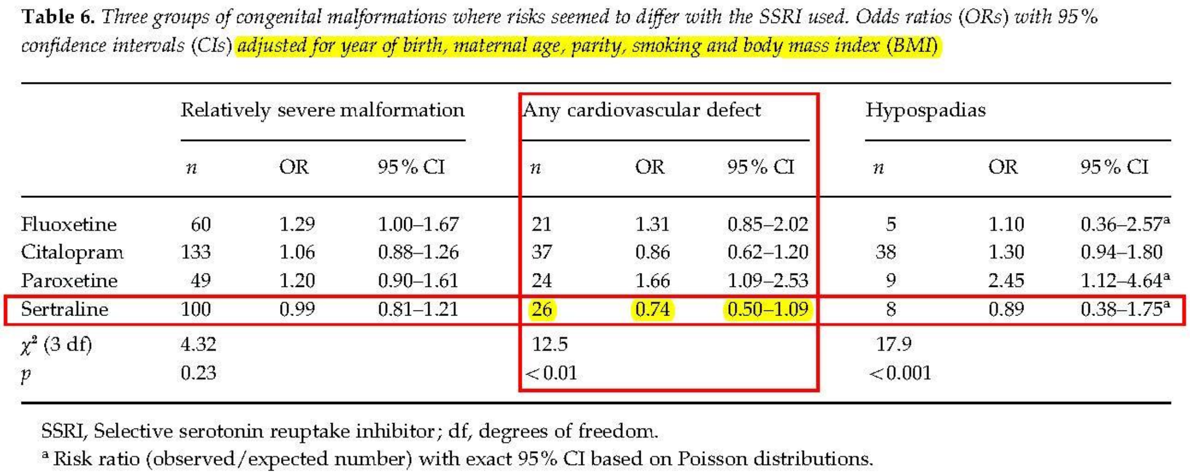


Table 6. *Three groups of congenital malformations where risks seemed to differ with the SSRI used. Odds ratios (ORs) with 95% confidence intervals (CIs) adjusted for year of birth, maternal age, parity, smoking and body mass index (BMI)*

	Relatively severe malformation			Any cardiovascular defect			Hypospadias		
	n	OR	95% CI	*n*	OR	95% CI	*n*	OR	95% CI
Fluoxetine	60	1.29	1.00–1.67	21	1.31	0.85–2.02	5	1.10	0.36–2.57[a]
Citalopram	133	1.06	0.88–1.26	37	0.86	0.62–1.20	38	1.30	0.94–1.80
Paroxetine	49	1.20	0.90–1.61	24	1.66	1.09–2.53	9	2.45	1.12–4.64[a]
Sertraline	100	0.99	0.81–1.21	26	0.74	0.50–1.09	8	0.89	0.38–1.75[a]
χ^2 (3 df)	4.32			12.5			17.9		
p	0.23			<0.01			<0.001		

SSRI, Selective serotonin reuptake inhibitor; df, degrees of freedom.

[a] Risk ratio (observed/expected number) with exact 95% CI based on Poisson distributions.

SOURCE: Margarita Reis & Bengt Källén, "Delivery outcome after maternal use of antidepressant drugs in pregnancy: an update using Swedish data," 40 Psychological Med. 1723 (2010). Copyright © (2010) Cambridge University Press.

(6) Malm (2011)[51]

What kind of study was done by Heli Malm?

[This paper is based upon data from the Finnish National Birth Cohort Registry. An earlier version of the data were analyzed and published in 2005, but without SSRI specific outcomes, and only for "all major malformations." In 2011, the principal investigator, Heli Malm, published an update of her cohort data, with more specific analyses of additional specific birth defects.]

[51] Heli Malm, Miia Artama, Mika Gissler, and Annukka Ritvanen, "Selective Serotonin Reuptake Inhibitors and Risk for Major Congenital Anomalies," 118 *Obstet. & Gynecol*. 111 (2011).

Was Malm a lumper or a splitter?

[Table 3 presents the data for sertraline and cardiac defects, for both any major cardiovascular defect (potentially broader than just cardiac), and for six specific subgroups of cardiovascular birth defects. So Malm lumped and split.]

Was Malm able to achieve a precise point estimate for the risk ratio for cardiac birth defects among children born to mothers who had used sertraline in first trimester?

[When she lumped, she was able to achieve a fair amount of precision in her point estimate for sertraline and any major cardiovascular anomaly: adjusted odds ratio of 0.65 (95% C.I., 0.34 to 1.25). This precision dissipated as she delved into subgroups of cardiovascular defects.

Most of the subgroups were below or near to risk ratios of 1.0, but for transposition of the great arteries, there was a finding of statistically nonsignificant increase in the odds ratio of 2.55, with a very wide 95% confidence interval of 0.35 to 18.62. This was obviously based upon very few cases—one to be exact! Similarly, the point estimate for conotruncal defects was 1.27 (95% C.I., 0.16 to 9.15), which was very broad, and not surprising given that there was also only one exposed case in the analysis.]

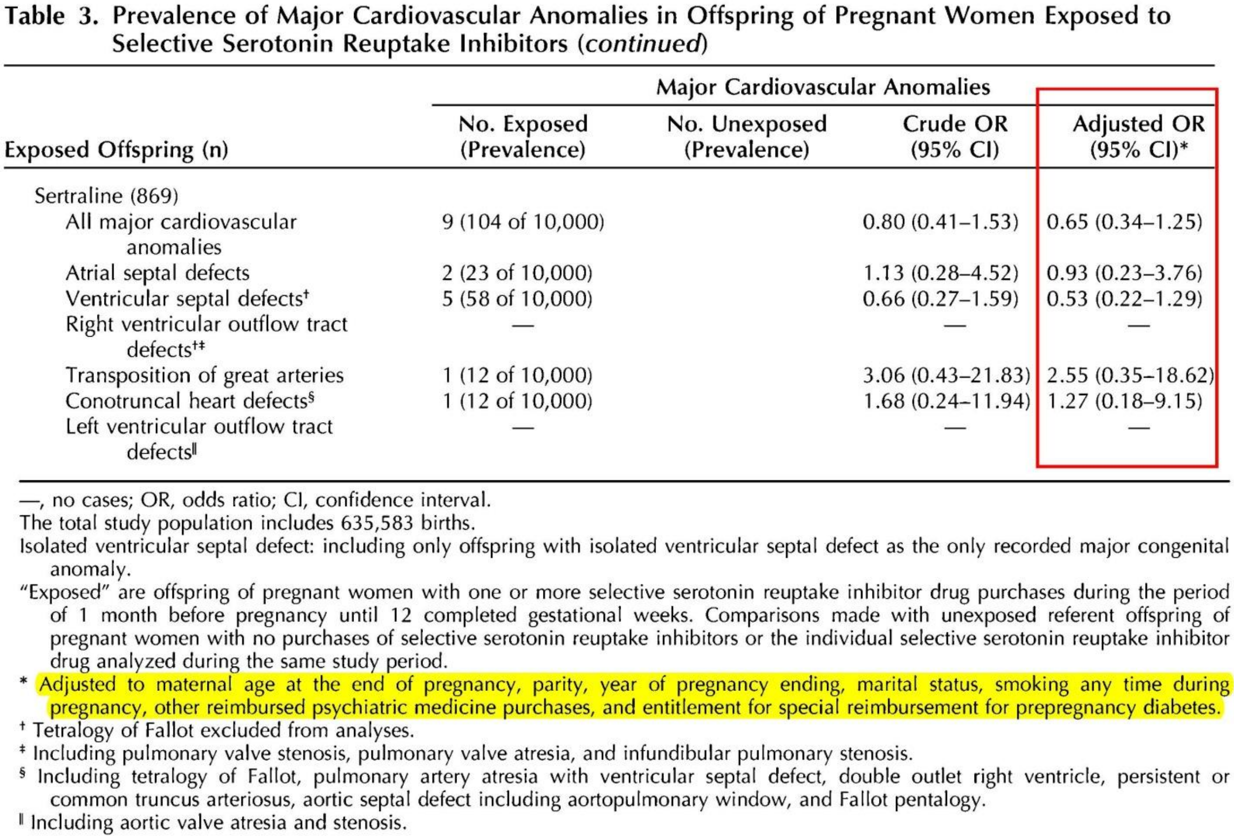


Table 3. Prevalence of Major Cardiovascular Anomalies in Offspring of Pregnant Women Exposed to Selective Serotonin Reuptake Inhibitors (*continued*)

	Major Cardiovascular Anomalies			
Exposed Offspring (n)	No. Exposed (Prevalence)	No. Unexposed (Prevalence)	Crude OR (95% CI)	Adjusted OR (95% CI)*
Sertraline (869)				
All major cardiovascular anomalies	9 (104 of 10,000)		0.80 (0.41–1.53)	0.65 (0.34–1.25)
Atrial septal defects	2 (23 of 10,000)		1.13 (0.28–4.52)	0.93 (0.23–3.76)
Ventricular septal defects[†]	5 (58 of 10,000)		0.66 (0.27–1.59)	0.53 (0.22–1.29)
Right ventricular outflow tract defects[†‡]	—		—	—
Transposition of great arteries	1 (12 of 10,000)		3.06 (0.43–21.83)	2.55 (0.35–18.62)
Conotruncal heart defects[§]	1 (12 of 10,000)		1.68 (0.24–11.94)	1.27 (0.18–9.15)
Left ventricular outflow tract defects[‖]	—		—	—

—, no cases; OR, odds ratio; CI, confidence interval.

The total study population includes 635,583 births.

Isolated ventricular septal defect: including only offspring with isolated ventricular septal defect as the only recorded major congenital anomaly.

"Exposed" are offspring of pregnant women with one or more selective serotonin reuptake inhibitor drug purchases during the period of 1 month before pregnancy until 12 completed gestational weeks. Comparisons made with unexposed referent offspring of pregnant women with no purchases of selective serotonin reuptake inhibitors or the individual selective serotonin reuptake inhibitor drug analyzed during the same study period.

* Adjusted to maternal age at the end of pregnancy, parity, year of pregnancy ending, marital status, smoking any time during pregnancy, other reimbursed psychiatric medicine purchases, and entitlement for special reimbursement for prepregnancy diabetes.

[†] Tetralogy of Fallot excluded from analyses.

[‡] Including pulmonary valve stenosis, pulmonary valve atresia, and infundibular pulmonary stenosis.

[§] Including tetralogy of Fallot, pulmonary artery atresia with ventricular septal defect, double outlet right ventricle, persistent or common truncus arteriosus, aortic septal defect including aortopulmonary window, and Fallot pentalogy.

[‖] Including aortic valve atresia and stenosis.

SOURCE: Heli Malm, Miia Artama, Mika Gissler, and Annukka Ritvanen, “Selective Serotonin Reuptake Inhibitors and Risk for Major Congenital Anomalies,” 118 Obstet. & Gynecol. 111 (2011). Copyright (2011) © 2011 The American College of Obstetricians and Gynecologists. Reproduced with permission.

For what other exposures and factors did Malm adjust in her adjusted odds ratios?

[Mother’s age, parity, year of pregnancy, smoking, other psychotropic drugs, and diabetes insofar as mothers were prescribed medications for diabetes before pregnancy.]

Was alcohol consumption controlled for in Malm’s or in any study?

[Only one of the six studies controlled for alcohol consumption, and truthful, accurate alcohol consumption histories are difficult if not impossible to obtain. Although alcohol was not included among the covariates in Malm’s adjusted model, there was striking, indirect evidence that the Finnish women taking SSRIs were different in their use of alcohol from women who did not take SSRIs. Among the offspring exposed to any SSRI, eight children had fetal alcohol spectrum disorders (prevalence of 11.5 of 10,000). There were 75 children, of women who did not take SSRIs, with such alcohol-related disorders (prevalence 1.2 of 10,000). The difference between the groups represented an odds ratio of 9.6, (95% C.I., 4.6 –20.0). Malm (2011), at 115.]

Does Malm draw any causal conclusions in her paper?

[No.]

The FDA’s revision of the paroxetine pregnancy category from C to D, focused attention on the entire class of SSRIs. The revision no doubt helped fuel litigation against manufacturers of Paxil, and the generic paroxetine, and this litigation ultimately grew to include the manufacturers of all the SSRI antidepressants. The FDA appears to have considered, but declined to conclude, that the observed signal with respect to paroxetine applied to the entire SSRI class. The agency did, however, request the original sponsor of Paxil, GlaxoSmithKline, to sponsor research that included all the SSRIs and evaluated the potential for a class effect.

We considered six studies, two case-control and four cohort studies. They vary in quality, size, design, biases, and the extent that confounders are identified and analyzed. The studies illustrate the difficulty in defining the proper outcome variable in birth defects litigation. If we consider only “major malformations,” as an outcome, we might miss an increased risk for a relatively rare malformation. Suppose the baseline risk of major malformations is 3%, and a hypothetical drug exposure increases the risk of spina bifida (which involves the failure of the spinal cord to close properly from the embryo’s neural tube) from a baseline risk of 1/1,000 to 2/1,000. The overall increased risk of major malformations, including spina bifida, would now be 3% + 0.1% for what should be a predicted overall risk of major malformations of 3.1%. Few studies, however, would be able to discriminate reliably 3% from 3.1%, or what would be a risk ratio of 1.034.

It is similarly unclear how to segregate out the results of specific birth defects. Even within an organ, there are likely to be serious heterogeneity of birth defects, in terms of the gestational and developmental timing of when such defects are formed, and whether they are likely all to result from a specific environmental, infectious, or medicinal exposure.

There are further complications in assessing studies and their outcomes. The timing of assessment can create significant disparities in the rate of birth defects. Many birth defects, even serious or major birth defects, are not discovered until later in children's lives. Some birth defects, denominated "major," when found, resolve spontaneously in the young child. Many septal defects, openings in the walls between the atria or ventricles of the newborn's heart, close without surgery. Differential ascertainment (looking more closely, perhaps with more sensitive instrumentation) in newborns of mothers who are anxious or depressed could readily find more such septal defects.

Other birth defects are really a consequence of preterm birth. One such example is patent ductus arteriosus (PDA). This problem affects the newborn, and results when the connection between the aorta and the pulmonary artery (ductus arteriosus), which is open in utero is not closed off soon after birth. The open or PDA allows oxygen-rich blood from the aorta to mix with oxygen-poor blood in the pulmonary artery, straining the heart and increasing blood pressure in the pulmonary arteries. If women taking antidepressants have other exposures, such as smoking, alcohol, or drugs, which lead to more preterm deliveries, then their infants may well have a higher detected rate of PDA. The inclusion of PDA in some but not other studies complicates the comparison of studies and the interpretation of risk ratios for cardiac defects.

READINGS FOR NEXT MEETING

Austin Bradford Hill*, "*The Environment and Disease: Association or Causation*,"* 58 *Proc. Roy. Soc'y Med.* 295 (1965), available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/pdf/procrsmed00196-0010.pdf.

Anthony K. Akobeng, "Understanding systematic reviews and meta-analysis," 90 *Arch. Diseases Childhood* 845 (2005), available at http://adc.bmj.com/content/90/8/845.full. Note that although Akobeng discusses systematic reviews and meta-analyses in the context of "trials" (usually thought of as human experiments), his discussion is relevant to systematic reviews and meta-analyses of observational studies, such as are involved in the epidemiology of birth defects and maternal use of medications.

Nicholas Myles, Hannah Newall, Harvey Ward, and Matthew "Large, Systematic meta-analysis of individual selective serotonin reuptake inhibitor medications and congenital malformations," 47 *Australian & New Zealand J. Psychiatry* 1002 (2013).

Causation – Synthesis, Systematic Reviews and Meta-Analysis

CLASS ACTIVITY: DISCUSSION OF CAUSATION

What was the occasion and context of Sir Austin Bradford Hill's paper?

[This was Hill's presidential speech to the Royal Society of Physicians; it was not a scholarly paper. By way of background, Hill, along with Sir Richard Doll, had just seen his causal claim that tobacco smoking causes lung cancer accepted the previous year by the United States Surgeon General. Not only the tobacco industry criticized their work and conclusions; no less than the great statistician and geneticist, Sir Ronald A. Fisher, the epidemiologist, Joseph Berkson, and the National Cancer Institute scientist, and well-regarded authority on cancer causation, Wilhelm Hueper, had criticized the causal claim. By 1965, the doubts had been quelled by observational epidemiology, without much in the way of experimental support (in humans or in animals). In many ways, Hill's work had shown the power of epidemiology at a time when epidemiology as a formal discipline was still relatively new and frequently dismissed as unable to resolve questions of human causation.]

Before we get to Hill's nine viewpoints or factors or criteria, what did Hill describe as the precondition for the application of these factors?

[Note the first page, before the nine factors are described, where Hill sets up the problem: "Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance."]

With respect to the epidemiology of sertraline and cardiac birth defects, is there an association between the two variables?

[In some studies, but not in others.]

Is the association clear-cut?

[In other words, is the association free from threats to validity from systematic bias and confounding? For those studies that report "associations," do the authors declare the associations clear-cut? What do you think?]

Is the association "beyond what we would care to attribute to the play of chance"?

[There are scattered nominally statistically significant results, but often in the context of multiple comparisons.]

Does this last question require a subjective or an objective assessment?

[Hint: Sir Austin was the author of one of the first important textbooks on the use of statistics in medical research. The 5% cutoff for attained significance probability is customary, and not a law of nature. As a customary cutoff, it is generally prespecified, in advance of conducting a study, and is not subjective in the sense that it varies from researcher to researcher.]

Do the confidence intervals in specific studies address whether the overall pattern of data across all studies suggest associations beyond that what we could care to attribute to the play of chance?

[No; they are statistical analyses based only upon the sample results in the immediate study.]

Taking the six epidemiologic studies of sertraline and cardiac birth defects, do we get beyond the precondition of a clear-cut association "beyond what we care to attribute to the play of chance?"

[Arguably, no; but a meta-analysis may give a more comprehensive answer.]

Sir Austin's Nine Viewpoints or Considerations for Assessing Causality

Regardless of where you come out on the preconditions for the Hill factors, let's look at those nine factors and examine how they would apply to the issue at hand.[52]

Are the considerations equally important?
Are they all necessary?
Is any necessary?

[Hill suggested none was necessary, but clearly temporality is required unless you are an Italian neutrino.]

Was Hill's ordering of the considerations important or revealing about his own thinking?

Strength

Magnitude of association. How does the strength of association between sertraline and birth defects, to the extent it exists at all, compare with Hill's example of a strong association between chimney-sweep work and scrotal cancer?

[Generally, two orders of magnitude. Relative risks of 1.5 or 2, or so versus relative risk of 200 for scrotal cancer mortality among chimney sweeps.]

[52] Austin Bradford Hill*, "*The Environment and Disease: Association or Causation*,"* 58 *Proc. Roy. Soc'y Med.* 295 (1965).

Why does Hill think that weaker associations may not be causal?

[The example he gives is illustrative. Smoking causes a 20-fold increase in lung cancer, but the increased rate of death due to “coronary thrombosis” (heart attack) is only about doubled among smokers. Without foreclosing a causal interpretation of smoking and heart attacks (which is now widely accepted), Hill notes that there may have been unmeasured confounding variables—“features of life that may go hand-in-hand with smoking,” such as lack of exercise or poor diet, which in turn will ultimately explain the apparent, smaller association between smoking and heart attack.]

Consistency

In other words, has the association been corroborated by other studies, done by different researchers, in different places with different study subjects, under different circumstances. How do the studies of sertraline and heart birth defects shape up in terms of a consistent outcome?

[These studies seem discordant.]

Hill notes that with many studies being done, there will be “many an environmental association thrown up.” He further notes that some of these may not satisfy the “customary tests of statistical significance” and “appear to be unlikely to be due to chance.” Is the failure of a study to satisfy the customary tests of statistical significance the end of the matter for Hill?

[No; Hill acknowledges as we must that the resolution of the chance-as-explanation issue may sometimes be resolved only by repeated studies. We will talk about meta-analysis later, and meta-analysis was not widely used at the time of Hill’s speech; nor were confidence intervals. The insight here is that an individual study may be inconclusive, and only repetition of studies will answer the research question to anyone’s satisfaction.]

Suppose, for some logistical reason, a study could not be done in a large sample, and so it was repeatedly done in small samples of the total population, with remarkably consistent results. In six studies, the relative risks were all 1.5 (or so), and in each case, the 95% confidence interval was (0.8 to 3.0). There is no suggestion of bias or confounding, only the problem with random or sampling error from the small sample size. How would you evaluate the consistency of the result in the face of the lack of ability to rule out chance at the individual study level?

[Putting aside the difficult assumption of no bias or confounding in these studies, the hypothetical presents remarkable concordance among study outcomes that should be persuasive under the assumptions. With a relative risk of 1.0 representing no association, and with only chance involved in producing a result above or below this “expected” value, we could calculate the probability of seeing six studies all with relative risks greater than 1.0. This is akin to flipping a coin and obtaining six heads. Assuming that heads and tails were equally likely, the probability of obtaining exactly six heads in a row is equal to (0.5^6 = 1/64 = 0.015625 = 1.6%, rounded.]

Specificity

The exposure associated with a very specific disease as opposed to a wide range of diseases. How does Hill explain specificity as helpful to resolving the causality of an observed association?

[Hill offers a couple of considerations. One is the narrowness of the disease or disorder that is under scrutiny. Another is the limitation on exposures that give rise to the outcome. Clearly, Hill did not think that there had to be an exclusive one-to-one relationship between the exposure variable and the outcome variable. He had been involved in research on arsenic and lung cancer, and he was well aware of Sir Richard Doll's 1955 research on asbestosis and lung cancer.[53] (Interestingly, when Doll was later questioned why he had not included smoking histories of the workers in his 1955 paper on asbestos factory workers, he noted simply that it had not occurred to him. His smoking and his asbestos research were under way at the same time, and he did not try to test both hypotheses in both lines of work. Implicit in Hill's discussion of specificity is that the more unusual the exposure, especially in environmental or occupational circumstances, and the more specific the outcome, the less likely that some other concurrent or intervening exposure may be responsible for the specific outcome. Today, many researchers downplay the need to show specificity, but they are not dismissing the need to identify and measure the exposure circumstance; rather they are simply acknowledging that many different exposures may lead to the same outcome.]

Temporality

Exposure before the outcome of interest. In the sertraline studies, what was the needed temporality?

[The in utero exposure had to have preceded the live birth with the cardiac defect. A refinement of the issue is, however, that the exposure had to take place during a gestational time window in which the structural feature of the heart was forming.]

Could there ever be any doubt about whether the exposure came before the outcome of interest?

[Although it seems obvious, in some study designs, it is possible to confuse which comes first. Sometimes, the circumstances are given names such as *protopathic* or *temporal-precedence bias*. One notorious instance was a study of a cough-cold remedy, phenylpropanolamine (PPA),

[53] Richard Doll, "Mortality from lung cancer in asbestos workers," 12 *Brit. J. Indus. Med*. 481 (1955).

and hemorrhagic stroke.[54] Because of the body's elimination of the PPA fairly quickly, the researchers used a case-control study design and counted an "exposed case" to be a stroke that occurred within three days of a PPA-containing drug. The problem turned out that people suffering from a stroke often present with a sentinel headache that sends them to the medicine cabinet, to take a PPA-containing drug. The headache quickly (within 3 days) evolves into the clinical manifestation of a stroke and leads to diagnosis. The headache that led the person to take the drug was, in some cases, a stroke that was already in progress, and some cases labeled "exposed" were really nonexposed.]

Biological Gradient (exposure- or dose-response)

What might be a dose-response in birth defects studies?

[The rate of an observed specific birth defect in children might increase between across of their mothers as those mothers' doses of medication increased. Alternatively, the severity of the birth defect might have increased in children of mothers, depending upon their dosages of the medication under study.]

Did any of the studies evaluate for dose response?

[No.]

Plausibility (an explanation for the association in terms of a credible mechanism)

What is Hill's assessment of the need for "plausibility" in a candidate for causal association?

[Hill downplays the need for a plausible explanation or mechanism. Indeed, at the time of his triumphant speech, there was no explanation of how smoking tobacco actually caused cancer. In litigation of lung cancer cases, tobacco companies often emphasized that the "cause of cancer is unknown," and downplayed the significance of the available epidemiology on the rationale that the populations studies, even with their impressive effect sizes and statistical precision, still fell short of demonstrating causation in the absence of mechanistic demonstration. Part of the significance of Hill's speech is its recognition that epidemiology had supplied an answer to the question of lung cancer causation without a demonstration of mechanism.]

From your reading of the studies, and the Solomon essay, is the claim that sertraline causes heart birth defects plausible in the sense used by Hill in his describing his factors?

[54] Walter N. Kernan et al., "Phenylpropanolamine and the Risk of Hemorrhagic Stroke," 343 *New Engl. J. Med*. 1826 (2000). The extent of the bias did not become apparent until after the parties to litigation sought underlying data and material, which in turn revealed questionable classifications of cases as exposed.

[Superficially, yes. SSRIs modify cellular uptake of serotonin with the consequence that larger amounts of extracellular serotonin can accumulate outside cells. In frogs and chicks, serotonin acts as a cell-signaling molecule that directs embryogenesis, and so the suggestion that it may do so in humans gains some plausibility. This suggestion leaves a lot of crucial details to be worked out, such as what levels of SSRIs reach the tissues supposedly involved at a vulnerable gestational stage, and at what stage of human embryogenesis do cellular serotonin receptors become available on cell membranes of tissues claimed to be susceptible to some change in serotonin levels.]

Coherence (association fits with natural history of the disease)

How did Hill argue for the coherence of smoking and lung cancer?

[Hill noted that coherence of the claim turned on the general agreement with what we know to be the natural history of the disease or outcome under study. In the case of lung cancer, there had been a dramatic rise in the rate of new lung cancer each year (incidence), which tracked the dramatic increase in tobacco smoking. Although the pattern was not quite as dramatic as the thalidomide-phocomelia pattern we discussed, the tobacco–lung cancer pattern lent further support to the claimed causality of the association.]

Is the claim that sertraline causes cardiac birth defects coherent based upon what you have reviewed to date?

[Most birth defects have no known cause, and therefore, much of birth defects epidemiology is a “black box,” the contents of which remain hidden from us. The thalidomide-phocomelia association was adjudged causal by the early 1960s, but the mechanism was only recently established.[55] Unlike the dramatic tracking of phocomelia incidence to the rate of new prescriptions for thalidomide, the rate of cardiac defects has been relatively stable over a long time. Recall the time-trend analysis done by the Danish researchers, who showed that the rate of septal heart defects was relatively unchanged over the entire time that SSRIs were introduced and became first-line therapies for depression, including in the population of child-bearing women.]

Experimental Evidence (animal or human controlled experiments supporting association)

In addition to experiments with humans, such as randomized clinical trials, and with animals, in which researchers assign or determine who gets exposed to what, what other kinds of experiments does Hill discuss?

[55] Christina Therapontosa, Lynda Erskine, Erin R. Gardner, William D. Figg, and Neil Vargesson, “Thalidomide induces limb defects by preventing angiogenic outgrowth during early limb formation,” 106 *Proc. Nat’l Acad. Sci*. 8573 (2009).

[Since Hill was concerned with environmental and occupational exposures, he was not particularly interested in clinical trial experiments, which would be unethical and impracticable in humans. He does, however, talk about "natural experiments," such as what we might see if a company had two factories with similar workforces, but used asbestos in product formulation in one factory, and fiberglass in another. Another example of a "natural experiment" might occur when an ingredient in a manufacturing process was discontinued for whatever reason, with the subsequent observation of a dramatic change in disease rate in the workforce population.

Analogy (similar outcomes from similar exposures/mechanisms)

How does Hill suggest that analogy functions in supporting a case for causality?

[Hill suggests that analogy operates at a fairly high level to help support claims of plausibility and mechanism. He gives an example from birth defects epidemiology, which is of interest. Hill's suggested that because thalidomide and rubella had already been determined to cause birth defects, we are thus more open to consider other drugs or infectious agents as potential causes.]

Is the Hill consideration of analogy satisfied for the claim that sertraline causes cardiac defects?

[This factor is relatively easily satisfied. When the causality of rubella and thalidomide were under consideration, there was substantial doubt about the ability of maternal exposures and diseases to affect embryologic and fetal development. The evidence for the causality of rubella virus and thalidomide changed the way medicine viewed relationship of the developing embryo to the maternal and the external environment.]

LECTURERS' NOTES

Since Hill's article, some of these considerations have fallen by the wayside as unimportant. Specificity, for instance, connotes that the exposure in question has a unique relationship with the associated outcome. There are some signature diseases, such as manganism (a movement disorder caused only by excessive manganese) or asbestosis (a scarring of the lungs caused only by asbestos), which can be caused only by a single, specific exposure, but they are exceptional in that relative risks, by definition, would be infinite.[56]

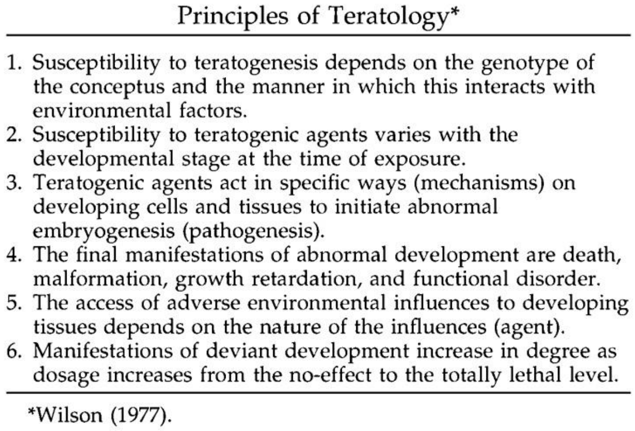
Some of the Hill viewpoints are trivially satisfied such as analogy or plausibility, for which anyone with a keen imagination, and without the necessity of much evidence, can satisfy. Hill modestly suggest that none of his viewpoints was necessary, but one of them,

[56] This ignores the substantial diagnostic difficulties in assessing whether a case is actually manganism, or asbestosis, or some other putative signature disease. The clinical criteria for such diseases may have less than perfect specificity, thus guaranteeing that on a large population basis, there will be many false positive diagnoses.

temporality, would seem to be required unless you were traveling faster than the speed of light.

Hill's viewpoints should be seen as helping to structure the discussion of, and the arguments for or against, causal conclusions. These viewpoints are certainly the only expression of criteria or factors to be considered and used in supporting or negating causal conclusions. In the field of teratology, several well-known and highly regarded scientists have advanced their formulations of the relevant consideration. Below are set out Wilson's principles, stated by one of the founders of the discipline of teratology:

WILSON'S PRINCIPLES OF TERATOLOGY



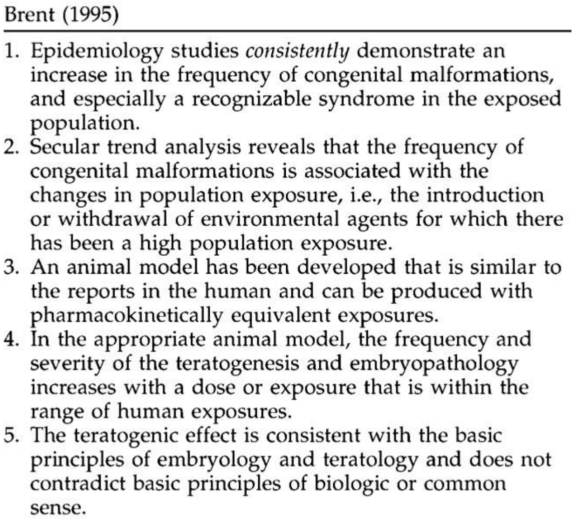
Principles of Teratology*

1. Susceptibility to teratogenesis depends on the genotype of the conceptus and the manner in which this interacts with environmental factors.
2. Susceptibility to teratogenic agents varies with the developmental stage at the time of exposure.
3. Teratogenic agents act in specific ways (mechanisms) on developing cells and tissues to initiate abnormal embryogenesis (pathogenesis).
4. The final manifestations of abnormal development are death, malformation, growth retardation, and functional disorder.
5. The access of adverse environmental influences to developing tissues depends on the nature of the influences (agent).
6. Manifestations of deviant development increase in degree as dosage increases from the no-effect to the totally lethal level.

*Wilson (1977).

SOURCE: The Public Affairs Committee of the Teratology Society, "Causation in Teratology-Related Litigation," 73 Birth Defects Research (Part A) 421, 422 (2005), citing James G. Wilson, "Current status of teratology. General principles and mechanisms derived from animal studies," in James G. Wilson & F. Clarke Fraser, eds., 1 Handbook of Teratology - General Principles and Etiology 47–74 (N.Y. 1977). Copyright © (2005) Wiley-Liss, Inc. Reprinted with permission.

Later versions of relevant considerations were put forward by the noteworthy teratologist, Robert Brent:

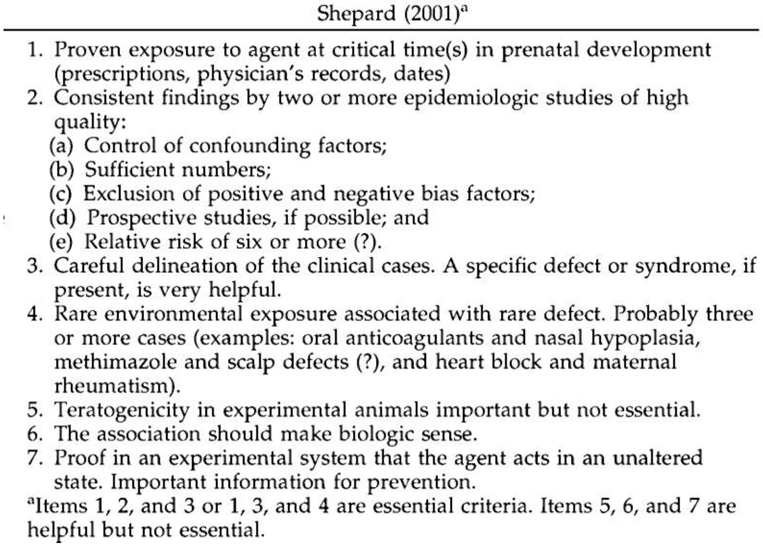


Brent (1995)

1. Epidemiology studies *consistently* demonstrate an increase in the frequency of congenital malformations, and especially a recognizable syndrome in the exposed population.
2. Secular trend analysis reveals that the frequency of congenital malformations is associated with the changes in population exposure, i.e., the introduction or withdrawal of environmental agents for which there has been a high population exposure.
3. An animal model has been developed that is similar to the reports in the human and can be produced with pharmacokinetically equivalent exposures.
4. In the appropriate animal model, the frequency and severity of the teratogenesis and embryopathology increases with a dose or exposure that is within the range of human exposures.
5. The teratogenic effect is consistent with the basic principles of embryology and teratology and does not contradict basic principles of biologic or common sense.

SOURCE: Reprinted from Reproductive Toxicology, Vol. 9, Issue 4, Robert Brent, “Bendectin: Review of the medical literature of a comprehensively studied human nonteratogen and the most prevalent tortogen-litigen” 337-349, Copyright © (1995), with permission from Elsevier.

And criteria from teratologist Thomas Shepard:



Shepard (2001)[a]

1. Proven exposure to agent at critical time(s) in prenatal development (prescriptions, physician’s records, dates)
2. Consistent findings by two or more epidemiologic studies of high quality:
   (a) Control of confounding factors;
   (b) Sufficient numbers;
   (c) Exclusion of positive and negative bias factors;
   (d) Prospective studies, if possible; and
   (e) Relative risk of six or more (?).
3. Careful delineation of the clinical cases. A specific defect or syndrome, if present, is very helpful.
4. Rare environmental exposure associated with rare defect. Probably three or more cases (examples: oral anticoagulants and nasal hypoplasia, methimazole and scalp defects (?), and heart block and maternal rheumatism).
5. Teratogenicity in experimental animals important but not essential.
6. The association should make biologic sense.
7. Proof in an experimental system that the agent acts in an unaltered state. Important information for prevention.

[a]Items 1, 2, and 3 or 1, 3, and 4 are essential criteria. Items 5, 6, and 7 are helpful but not essential.

SOURCE: Shepard, Thomas H., M.D.. Catalog of Teratogenic Agents. Tenth Edition. p. xxiv, Table 1. © 1973, 1976, 1980, 1983, 1986, 1989, 1992, 1995, 1998, 2001 The Johns Hopkins University Press. Reprinted with permission of Johns Hopkins University Press.

Follow-up assignment: Do these other criteria sets for evaluating causality add anything to what is explicitly or implicitly present in Hill’s 1965 paper?

CLASS ACTIVITY: DISCUSSION OF SYSTEMATIC REVIEWS AND META-ANALYSIS

From your reading of Akobeng's introduction to systematic reviews and meta-analysis, how do systematic reviews and meta-analyses differ from the kind of journalistic coverage of SSRIs and birth defects, seen in Solomon's essay?

Have you ever seen a television news show that presented information beyond the latest study mentioned in some university's press release?

In your experience, do news media distinguish between and among "links," "increased risks," "risk factors," and "causes"?

What are the key features of a "systematic review," which are different from your experience of how biomedical causation issues are typically reported and discussed in the media?

[Note that "eligibility" criteria, that is criteria for inclusion and exclusion in the review, are formulated before the search of the literature. In narrative reviews, and popular media, studies are often taken on the allure of their headlines and findings, with little or no attention to the appropriateness or rigor of their methodology.]

What are the key goals of a systematic review?

[Attention to study validity and rigor; comprehensiveness; external validity of studies (applicability to entire population or to other samples beyond the sample studied).]

Does a systematic review inevitably involve a quantitative meta-analysis?

[No; the review may find no qualifying studies, or the studies may be so different by design, or by findings, that combining them would give a "summary estimate," which would be misleading or simply false.]

Is any of the terminology used by Akobeng to describe meta-analysis misleading?

[Akobeng uses the traditional statistical term *effect size* to describe the magnitude of association, whether found in an individual study or found as a summary estimate of effect size from a meta-analysis. The terminology presupposes that the association is causal by referring to an effect.]

LECTURER'S NOTES FOR SYSTEMATIC REVIEW AND META-ANALYSIS

Traditionally, expert opinion was based upon the authority and prestige of the speaker. Courtroom procedures permitted expert witnesses, once qualified as "experts," to say pretty much what they wanted. Juries were free to accept all, some, or none of expert opinion, and appellate courts had little power, and less interest, to revisit the scientific issues behind expert opinion testimony. In the medical literature, narrative review articles and textbook chapters predominated. Consensus statements from learned societies were usually the opinions of GOBSAT ("good ol' boys sitting at the table").

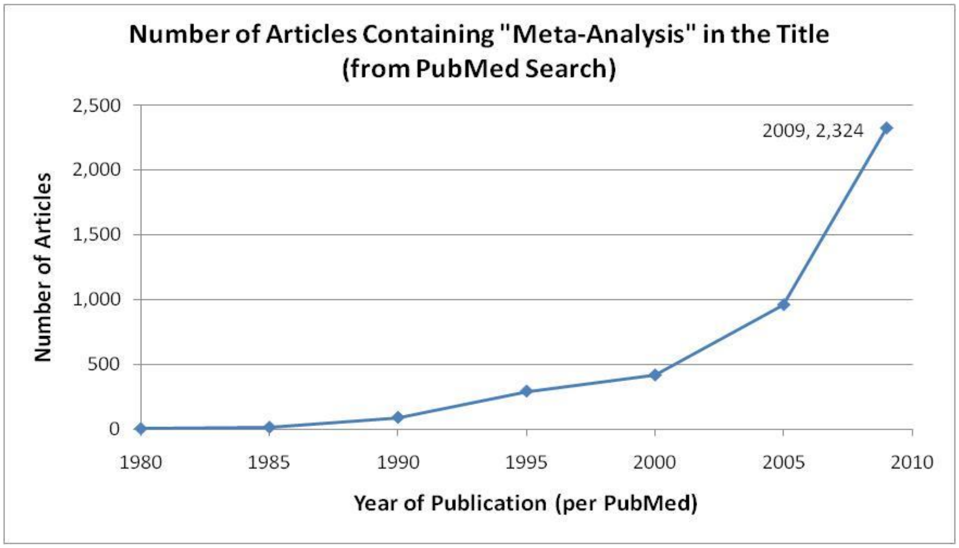
The last couple of decades have seen the emergence of an insistence upon sound methodology and comprehensive assessment of the available evidence. In the courts, judicial decisions and legislation have required trial judges to serve as gatekeepers with respect to the methodological soundness of opinion testimony offered by expert witnesses. In medicine, opinion has given way to systematic reviews, the hallmarks of which include prespecification of appropriate evidence by inclusionary and exclusionary criteria, with a comprehensive search for, and inclusion of, all pertinent evidence.[57]

The systematic review may, in appropriate instances, lead to a meta-analysis, which uses a quantitative method for systematically combining study outcomes to reach a conclusion ("summary estimate of association") about the research issue involved.

The mathematical approach to combining evidence across studies has its origins in the work of Karl Pearson and Sir Ronald Fisher, but meta-analysis in the world of biomedical research has gained acceptance only in the late 20th century. In 1952, psychologist Hans Eysenck concluded that psychotherapy was not efficacious, which led another psychologist, Gene Glass, in 1978, after furious debate for decades, to aggregate studies statistically studies in what he called a *meta-analysis*. Eysenck responded by dubbing Glass's approach, "mega-silliness."

Meta-analysis of individual studies did not become common in the biomedical world until the 1990s. A search of the National Library of Medicine's PubMed database suggests that there were only about 10 meta-analyses published in 1980.

[57] See, e.g., Moher et al., "Epidemiology and Reporting Characteristics of Systematic Reviews," 4 *PLoS Medicine* e78 (2007); Liberati et al., "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies that Evaluate Health Care Interventions: Explanation and Elaboration," 151 *Ann. Intern. Med.* W-65 (2009).



SOURCE: Courtesy of the author.

Meta-analysis met with resistance from several prominent writers,[58] but the methodology became increasingly more common in medical publications throughout the 1990s, with dramatic increases after 2000. In addition to the increasing use of meta-analysis, the appearance of various consensus statements about both the methods used and the reporting of findings suggest that meta-analytic techniques have become an important part of the methodological repertoire of modern medical science.[59]

The rise of meta-analysis in biomedical science is mirrored, with some time lag, in the reported judicial decisions in cases involving causal claims in environmental, pharmaceutical, and occupational disease litigation. The earliest proffered meta-analyses met with lawyerly resistance and some judicial skepticism.[60] More recently, meta-analyses have become commonplace in litigation over health claims.[61] In some cases, meta-analyses have been dispositive of mass tort litigations, essentially closing thousands of cases.[62]

[58] See, e.g., Alvan Feinstein, "Meta-analysis: statistical alchemy for the 21st century," 48 *J. Clin. Epidem*. 71 (1995); Samuel Shapiro, "Meta-analysis/Shmeta-analysis," 140 *Am. J. Epidem*. 771 (Nov. 1994) ("[m]eta-analyses begin with scientific studies, usually performed by academics or government agencies, and sometimes incomplete or disputed. The data from these studies are then run through computer models of bewildering complexity, which produces results of implausible precision.").

[59] See, e.g., Blair et al., "Guidelines for application of meta-analysis in environmental epidemiology," 22 *Regulatory Toxicol. & Pharmacol*. 189 (1995); Moher et al., "Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement," 354 *Lancet* 1896 (1999); Stroup et al., "Meta-analysis of observational studies in epidemiology: A proposal for reporting," 283 *J. Am. Med. Ass'n.* 2008 (2000) (MOOSE guidelines).

[60] See, e.g., *In re Paoli R.R. Yard PCB Litigation,* 916 F.2d 829, 856-57 (3d Cir.1990) (''There is some evidence that half the time you shouldn't believe meta-analysis, but that does not mean that meta-analyses are necessarily in error. It means that they are, at times, used in circumstances in which they should not be.'') (Internal quotation marks and citations omitted), cert. denied, 499 U.S. 961 (1991).

[61] Meta-analysis has featured prominently in cases involving asbestos, Bendectin, benzene, silicone gel breast implants, environmental tobacco smoke, fenfluramine, antidepressants (SSRIs, suicide, and birth defects), Baycol,

The goals of meta-analysis include providing a more objective, quantitative summary of the evidence. Combining the evidence enhances precision and power to detect smaller risks, and to answer questions that smaller, single studies cannot. Meta-analysis addresses several of Hill's factors, including strength (risk ratio or risk difference), consistency, and exposure-response, and thus serves as an aid to causal inference. Meta-analysis can also serve as an aid to evaluate bias, and confounding by allowing stratification of studies, and by subgroup analyses. Studies with and without suspected biases or confounders can be analyzed separately and compared to identify whether there are discernible differences.

Meta-analysis, like an individual study, requires a protocol and statistical analysis plan in advance. The designers must specify the criteria for study inclusion and exclusion, data abstraction, statistical methods, analysis of "heterogeneity," planned sensitivity and subgroup analyses, reporting and publishing, and future updating when new studies become available.

The inclusionary criteria typically address study quality, exposure of interest, outcome of interest, and methodological rigor in advance so that inclusion and exclusion does not turn on study outcome. Consideration of heterogeneity has two aspects, statistical and clinical. Statistical heterogeneity arises when the study samples to be included do not seem to come from the same population. Clinical heterogeneity arises from differences in study quality, rigor, and design, as well as differences in exposures, exposure measurements, and outcome definitions and ascertainment. Either kind of heterogeneity may be the basis for not proceeding with quantitative meta-analysis, or it may lead to planned sensitivity analyses to see whether there are differences across studies that relate to their design or other differences.

Study results cannot usually be simply added together.[63] Instead, individual study results are weighted by the inverse of their variability. The larger the study, and the more precise the study's point estimate (and hence the smaller its variance), the more weight is given to the study in the meta-analysis.

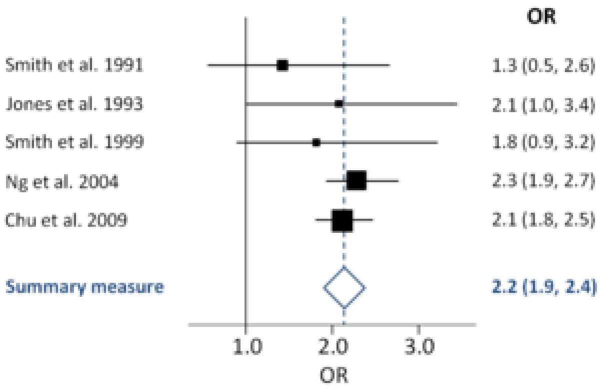
The results of a meta-analysis are usually presented in a summary estimate of risk, and a graphic form, called a forest plot. A generic forest plot is shown, below.[64]

Bextra, Celebrex, Vioxx, hormone therapy (postmenopausal), Fosamax, Avandia, Actos, Neurontin, Seroquel, Zyprexa, gadolinium contrast media, Trasylol, among others.

[62] In the silicone gel cases, meta-analyses were done by the federal court's appointed expert witnesses, who concluded that plaintiffs' claims were not supportable. More recently, a meta-analysis of studies of Parkinson's disease among welders ended hundreds of claims that welders were at risk from Parkinson's disease. See James Mortimer et al., "Associations of welding and manganese exposure with Parkinson's disease," 79 *Neurology* 1174 (2012). In litigation over isotretinoin (Accutane), a party's expert's meta-analysis led to the exclusion of the other side's expert witnesses. *In re Accutane*, No. 271(MCL), 2015 WL 753674, 2015 BL 59277 (N.J. Super. Law Div. Atlantic Cty. Feb. 20, 2015).

[63] Because of Simpson's paradox and other statistical issues.

[64] From the Wikipedia article on "Forest plots," last visited March 1, 2015.

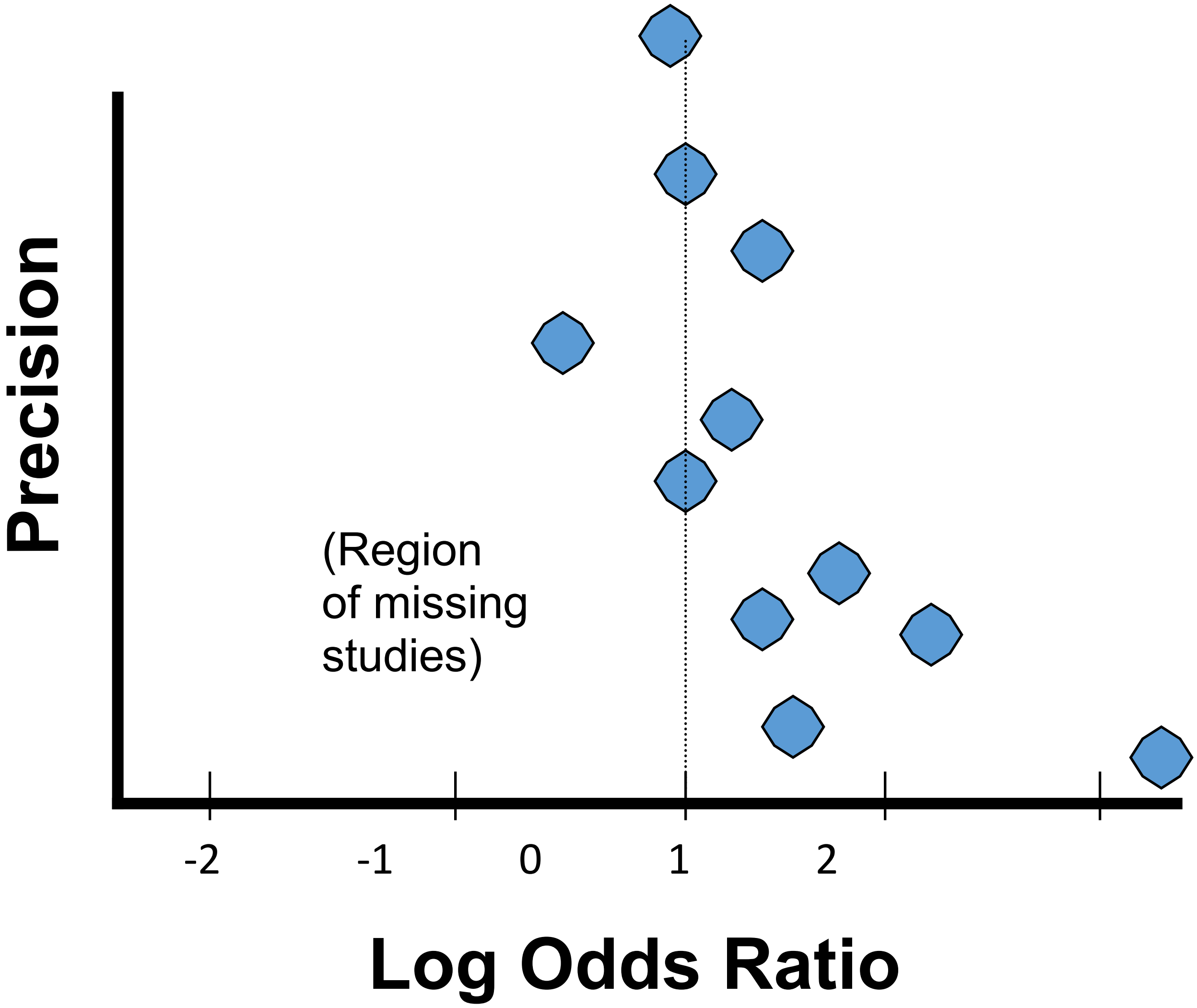


SOURCE: James Grellier, "Forest plot" Wikipedia, online: https://en.wikipedia.org/wiki/Forest_plot, licensed under a CC BY-SA 3.0 license.

Some of the key features of the forest plot include an identification of the included studies, their respective point estimates and confidence intervals, and the summary measure. Horizontal lines through the point estimate mark the boundaries of the 95% confidence interval. The point estimate is sometimes represented by a square, the size of which reflects the weight given to the particular study in the meta-analysis. The summary measure is depicted by a diamond, the width of which reflects the 95% confidence interval around the summary estimate of association.

One of the most worrisome biases that may affect a meta-analysis, beyond the biases in the included studies, is so-called publication bias that results when journals publish studies with splashy findings ($p < 0.05$; confidence intervals exclude the null hypothesis expected value). Conversely, journals may be less inclined to publish "null studies," with risk ratios that are not statistically significant, because they are less news worthy. The result is the publication process curtails the available data and makes some of the data unavailable.

One way to control publication bias is to seek out unpublished papers and dissertations, which may have not been published because of their "null" findings. One way to assess the potential existence of publication bias is to compare published studies, arranged in order by their precision, in a variation of the forest plot, known as the funnel plot. In the forest plot example, the most precise, generally the largest, studies are at the top of the plot, with the smaller, more variable, studies at the bottom. The combined result of shown point estimates may well suggest a summary point estimate greater than 1.0, but the plot's asymmetry suggests that smaller studies with decreased risk ratios are missing, perhaps because they are less "newsworthy" in the pages of the journals:



SOURCE: Marina L. Ritchie, Tamara N. Romanuk, "A Meta-Analysis of Probiotic Efficacy for Gastrointestinal Diseases" PLoS ONE (18 April 2012), online: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0034938, licensed under a CC BY license.

There is a substantial literature, reflecting concern in the medical and scientific community, on how publication bias may be reduced. Some journals have agreed to review study protocols, and if the protocols identify important issues and valid methods, to publish the study's final results, regardless whether the findings are dramatic or otherwise newsworthy.

Meta-analyses are not cure-alls for bias, confounding, or lack of statistical power, but they have added an important set of tools to help identify health effects. By modifying inclusion criteria, by alternately including and excluding questionable studies, and by combining interesting subgroups across studies, meta-analyses can advance and improve our understanding of health effects. Meta-analyses are not "statistical alchemy," but they have their limitations. The GIGO ("garbage in; garbage out") principle ultimately cannot be escaped.

CLASS ACTIVITY: DISCUSSION OF MYLES'S META-ANALYSIS

Did Myles and colleagues state clear objectives and study eligibility criteria?

[The authors purported to do so, but without replicating their search strategy it is impossible to ascertain whether they were faithful to their method.]

Did Myles and colleagues conduct a systematic review that identified each study's design and methodological quality?

[Not really.]

Did Myles and colleagues discuss, explain, and justify their decision to proceed with a meta-analysis of the exposures and outcomes of interest?

[Major Malformations: The combination of study data for "all major malformations" looks rather dubious. Some will arise from the first trimester, and others may involve late pregnancy associations. Furthermore, such a diverse assortment of malformations would seem too broad to be plausibly connected to any one exposure. Including all major malformations certainly increases study power, but defining the outcome at such a high level of generality has the potential to obscure associations in the outcome of interest (heart), and any putative finding for "any major malformation" may or may not actually apply to cardiac defects.

For cardiac malformations, the question is closer. The actual inclusion of defects under the heading "cardiac" was also variable, which raises concerns whether the organ-specific outcome definition will apply to a more specific heart defect, such as "septal defect," or "transposition of the great arteries." Septal defects make up about half of heart defects, but some heart defects addressed in some studies are a small subset of the organ-level category. For our purposes, we will assume, without trying to assess, the propriety of a meta-analysis of all cardiac defects.]

Did Myles's meta-analysis increase the precision of the estimated associations?

[Yes; the confidence intervals were generally narrower in the meta-analysis results than in individual studies.]

Did Myles find a "class effect" for the SSRI medications and any birth defect?

[No; the results suggest that some SSRIs were associated with birth defects, and others were not associated.]

Did Myles and colleagues conduct an analysis of the studies along the lines of Sir Austin Bradford Hill's approach, or any other approach?

[No; although they made some clinical recommendations, they did not make causal judgments.]

Are conclusions, such as made by Myles and colleagues, that maternal paroxetine and fluoxetine use are associated with “significantly increased risk of major malformations and paroxetine in particular with cardiac malformations” intended to state conclusions that these two SSRI drugs cause all major or cardiac malformations?

[The term *increased risk* is generally used to mean that there is a greater than an expected rate of outcome of interest, not likely explained by chance, but without actually asserting a causal relationship. The distinction, however, between an increased risk and a cause that changes (by increasing) a rate of an otherwise expected outcome is often lost in translation to nonepidemiologists.]

With respect to sertraline and cardiac defects, what was the author’s quantitative conclusion?

[The summary point estimate of association is 1.109, with a 95% confidence interval that includes from 0.875 to 1.406.]

How did the authors interpret this result?

[Generally, the authors put forward a clinical recommendation that sertraline (and another SSRI, (es)citalopram), should be used as “first-line SSRI treatments in pregnancy and women of childbearing age.” Although the authors do not purport to show the absence of risk, they interpret the data analyzed to suggest an absence of an association, and relative safety sufficient for their clinical recommendation.]

How do you interpret this result?

What null hypotheses for sertraline and cardiac defects are ruled out by the meta-analysis?

[Any point estimate below 0.875, or above 1.406, would be rejected as a likely true population value based upon the aggregate sample data. But this analysis does not take into account residual bias and confounding in the data.]

What null hypotheses for sertraline and cardiac defects are not ruled out by the meta-analysis?

[Any point estimate greater than 0.875, but below 1.406, would not be rejected as a likely true population value based upon the aggregate sample data. In other words, the aggregate study data are compatible with anything from a 12.5% reduction to a 40% increase in the already existing risk of cardiac birth defects. Again, this analysis does not take into account residual bias and confounding in the data.]

How should the FDA interpret the result?

[The FDA has kept sertraline in the Category C class, which still implicitly urges caution in the prescription of the medication and a clinician's finding of a clear medical need.]

In the crucible of personal injury litigation, how will plaintiffs' expert witnesses and counsel interpret this result in the context of litigation claims?

[They will argue that the risk is actually elevated, although not by much, and that some cases would not have occurred without the exposure.]

Assuming that the upper bound of the sertraline/cardiac defect confidence interval (1.4) is correct, what if anything can reasonably be said by an individual claimant in support of his claim that his birth defect was caused by sertraline?

[Not much; without a biomarker that identifies which cases would not have occurred had there been no maternal drug exposure. If there were a very high relative risk, such as the relative risk of 200 that Sir Austin Bradford Hill cited in connection with scrotal cancer mortality among chimney sweeps, we could be pretty sure that any case would not have occurred had it not been for the exposures involved in chimney sweeping.]

LECTURE NOTES ON MYLES'S META-ANALYSIS

The meta-analysis by Myles and colleagues[65] is typical in how it goes about identifying the research issues, the appropriate evidence, the search for studies, the inclusionary and exclusionary criteria, and the methods to be used to "pool" results. The published paper is actually several meta-analyses of each of the SSRI antidepressants. For our purposes, we focus on the meta-analysis of studies for sertraline. The authors conducted a meta-analysis of sertraline and any "major" malformation and provide a forest plot.[66]

The summary point estimate for any major malformation is very close to 1.0, and it has sufficient precision to rule out even a fairly small increased risk as compatible with the sample data. Looking at only major malformations can obscure rare outcomes however in particular organ systems.

Much of the concern about SSRIs and birth defects comes from the reports of associations between paroxetine and cardiac birth defects. Myles, accordingly, produced meta-analyses for each of the SSRIs and cardiac malformations. A forest plot for sertraline studies and cardiac birth defects is also shown.[67]

[65] Nicholas Myles, Hannah Newall, Harvey Ward, and Matthew Large, "Systematic meta-analysis of individual selective serotonin reuptake inhibitor medications and congenital malformations," 47 Australian & New Zealand J. Psychiatry 1002 (2013).

[66] Ibid, at supplementary figures. See "Figure 1: Forrest [sic] plot of meta-analysis of major malformation."

[67] Ibid. See "Figure 2: Forrest [sic] plot of meta-analysis of cardiac malformation."

OPINIONS, DECISIONS, CONCLUSIONS

After reviewing and synthesizing the available evidence, people have to draw conclusions or make decisions about the casual claims. Conclusions are not necessarily dichotomously in favor or against causation. In the legal arena, courts may reject claims as inadequately supported, with the understanding that future claims may fare differently with different evidence. In litigation over sertraline and birth defects, a federal trial court judge, charged with oversight of the pretrial proceedings in hundreds of cases, excluded plaintiffs' epidemiology expert witness, not because her conclusions did not agree with consensus professional society or regulators' statements, but because her methodology was demonstrably flawed.[68] Plaintiffs' counsel moved for reconsideration, which led the trial court to reaffirm its ruling,[69] but the trial judge gave plaintiffs an opportunity to substitute another expert witness on the representation that he could and would employ a more defensible methodology in reaching the same causal opinion as held by the excluded expert witness.[70] Ultimately, the second, substituted expert witness's methodological approach could not avoid the pitfalls shown by the first witness, and the federal judge excluded the second expert witness as well.[71] In April 2016, the federal trial court granted summary judgment in hundreds of pending cases,[72] and plaintiffs have appealed.

As we have seen, regulators cannot defer the decision about how to regulate a medication. At launch of a new drug, regulators must require sponsors to fashion a warning in the absence of the most important information derived from human experience. Changes in the evidence over time may lead regulators to mandate a change in drug labeling, or they may postpone decisions until the evidence is clearer. In the United States, the FDA has changed the labeling for paroxetine, but not for other SSRI antidepressants.

Science has no "authoritative" checklist or algorithm for parsing the evidence to reach conclusions of causation. Some scientists will advert to Hill's viewpoints, but in the context of teratogenicity, Wilson's principles, and Brent's and Shepard's criteria may compete for attention.

Physicians must make risk-benefit decisions in the absence of information, but have a professional responsibility to stay current with the evidence. In reality, busy practicing physicians rely heavily upon opinions of "thought leaders" in their profession to help them assess evolving evidence. As we saw, the meta-analysis by Myles and colleagues ends in recommending that Zoloft/sertraline be included as a first-line therapy for depression. ("SSRI medications sertraline or citalopram should be considered as first-line SSRI treatments in pregnancy and women of childbearing age"). Students should evaluate the professional society statements as to whether they are up to the standards of "systematic reviews."

[68] *In re Zoloft Prods. Liab. Litig.*, 26 F. Supp. 3d 449 (E.D.Pa.2014)

[69] *In re Zoloft (Sertraline Hydrochloride) Prods. Liab. Litig*., MDL No. 2342; 12-md-2342, 2015 WL 314149 (E.D. Pa. Jan. 23, 2015) (denying plaintiffs' motion for reconsideration).

[70] *In re Zoloft Prods. Liab. Litig.*, No. 12–md–2342, 2015 WL 115486, at * 2 (E.D.Pa. Jan. 7, 2015).

[71] *In re Zoloft Prods. Liab. Litig.*, No. 12–md–2342, 2015 WL 7776911 (E.D.Pa. Dec. 2, 2015).

[72] *In re Zoloft Prod. Liab. Litig*., MDL NO. 2342, 12-MD-2342, 2016 WL 1320799, at *4 (E.D. Pa. April 5, 2016).

The readings should be seen to raise serious questions about the competency of various persons to evaluate the evidence and formulate credible opinions about causation. The print and electronic media often responds to the most recent article published, and the press release from the investigators' university or institution. There is some evidence that these press releases, whether from commercial, academic, or governmental sources are inaccurate and misleading.[73] Prescribing physicians may lack the time to delve into the evidence in a comprehensive fashion. Judges and juries must evaluate evidence in an adversarial cauldron, in which the studies themselves fade into the noise of competing claims and argument.

SUGGESTED FOLLOW-UP READINGS

In re Zoloft Prods. Liab. Litig., 26 F. Supp. 3d 449 (E.D. Pa. 2014).

In re Zoloft (Sertraline Hydrochloride) Prods. Liab. Litig., MDL No. 2342; 12-md-2342, 2015 WL 314149 (E.D. Pa. Jan. 23, 2015) (denying plaintiffs' motion for reconsideration).

In re Zoloft Prod. Liab. Litig., MDL NO. 2342, 12-MD-2342, 2016 WL 1320799, at *4 (E.D. Pa. April 5, 2016) (granting summary judgment in hundreds of cases involving sertraline and cardiac birth defects).

Kimberly A. Yonkers, Katherine L. Wisner, Donna E. Stewart, Tim F. Oberlander, Diana L. Dell, Nada Stotland, Susan Ramin, Linda Chaudron, and Charles Lockwood, "The Management of Depression During Pregnancy: A Report from the American Psychiatric Association and the American College of Obstetricians and Gynecologists," 10 *Focus: The Journal of Lifelong Learning in Psychiatry* 78 (2012).

Marian McDonagh, Annette Matthews, Carrie Phillipi, Jillian Romm, Kim Peterson, Sujata Thakurta, and Jeanne-Marie Guise (for the Pacific Northwest Evidence-based Practice Center Portland, OR), *Antidepressant Treatment of Depression During Pregnancy and the Postpartum Period*, Evidence Report /Technology Assessment No. 216, Agency for Healthcare Research and Quality (2014) (Excerpts).

Etienne Weisskopf, Celine J Fischer, Myriam Bickle Graz, Mathilde Morisod Harari, Jean-François Tolsa, Olivier Claris, Yvan Vial, Chin B Eap, Chantal Csajka, and Alice Panchaud, "Risk-benefit balance assessment of SSRI antidepressant use during pregnancy and lactation based on best available evidence," 14 *Expert Opin. & Drug Safety* 413 (2015).

Krista F. Huybrechts, Kristin Palmsten, Jerry Avorn, Lee S. Cohen, Lewis B. Holmes, Jessica M. Franklin, Helen Mogun, Raisa Levin, Mary Kowal, Soko Setoguchi, and Sonia Hernández-Díaz, "Antidepressant use in pregnancy and the risk of cardiac defects," 370 *New Engl. J. Med*. 2397 (2014) ("The results of this large, population-based cohort study suggested no substantial increase in the risk of cardiac malformations attributable to antidepressant use during the first trimester."), available at http://www.nejm.org/doi/full/10.1056/NEJMoa1312828.

[73] See, e.g., Steven Woloshin et al., Press Releases by Academic Medical Centers: Not So Academic?, *150 Ann. Intern. Med*. 613 (2009).

Assessment

1. Risk Communication

j. Based upon the data discussed in class, and from a "neutral" perspective, draft a narrative warning label to guide prescribing physicians in their use of sertraline for pregnant women and women of child-bearing age.
k. On the assumption that the law required a warning label for the patient, and from the same perspective, draft a narrative warning label for the reasonable woman's use.
l. Identify how various stakeholders (physicians, patients, regulators, industry, and tort lawyers) might challenge your neutral warning label.

2. Understanding and Interpreting New Data in the Context of Prior Studies

A new study is published, which the authors have graciously made available online:

A. Wemakor, K. Casson, E. Garne, M. Bakker, M-C. Addor, L. Arriola, M. Gatt, B. Khoshnood, K. Klungsoyr, V. Nelen, M. O'Mahony, A. Pierini, A. Rissmann, D. Tucker, B. Boyle, L. de Jong-van den Berg, and H. Dolk, "Selective serotonin reuptake inhibitor antidepressant use in first trimester pregnancy and risk of specific congenital anomalies: A European register-based study," *Eur. J. Epidemiol.* (2015). DOI: 10.1007/s10654-015-0065-y (Full text).

This paper may serve as a basis for testing the students' understanding of past concepts, and their developed skill in confronting new data. Key data from Table 2 are presented here:

Congenital Anomalies (CA)	Number	SSRI		Sertaline	
		Number	Adjusted OR (95% CI)	Number	Adjusted OR (95% CI)
Congenital Heart Defect (CHD)	12,876	108	1.41	16	1.51
Severe CHD	2935	28	1.56	6	2.88

What is the study design?
What is the sample size?
Describe the study's findings with respect to sertraline for cardiac birth defects.

m. Does the study report increased risks?
n. Are reported risks strong?
o. Are the measurements of reported risks precise?

Are the study authors justified in using causal language based upon their study? Or based upon their study in conjunction with the prior data?

Do the study authors appropriately interpret their own data?

Are the study results consist with the “class effect” hypothesis for cardiac malformations?

By comparing the prevalence of use of SSRIs antidepressants in this registry-based study with the population-based studies, such as Malm 2011, or Petersen 2009, are there issues with respect to the selection of the control group?

What “factors” are adjusted for in this study’s multivariate analysis? How does the adjustment process compare with the use of co-variates in studies such as Louik 2007 and Alwan 2007?