PRECEDENTIAL

UNITED STATES COURT OF APPEALS
FOR THE THIRD CIRCUIT

No. 16-2247

IN RE: ZOLOFT (SERTRALINE HYDROCHLORIDE)
PRODUCTS LIABILITY LITIGATION

Jennifer Adams, et al, Plaintiffs appealing dismissal
by order entered April 5, 2016,

Appellants

On Appeal from the United States District Court
for the Eastern District of Pennsylvania
(D. C. Civil Action No. 2-12-md-02342)
District Judge: Honorable Cynthia M. Rufe

Argued on January 25, 2017

Before: CHAGARES, RESTREPO and ROTH, Circuit Judges

(Opinion filed: June 2, 2017)

David C. Frederick [Argued]
Derek T. Ho
Kellogg Huber Hansen Todd Evans & Figel
1615 M Street, N.W.
Suite 400
Washington, DC 20036

Dianne M. Nast
NastLaw
1101 Market Street
Suite 2801
Philadelphia, PA 19107

Mark P. Robinson, Jr.
Robinson Calcagnie Robinson Shapiro Davis
19 Corporate Plaza Drive
Newport Beach, CA 92660
Counsel for Appellants

Sheila L. Birnbaum
Mark S. Cheffo [Argued]
Quinn Emanuel Urquhart & Sullivan
51 Madison Avenue
22nd Floor
New York, NY 10010

Robert C. Heim
Judy L. Leone
Dechert
2929 Arch Street
18th Floor, Cira Centre
Philadelphia, PA 19104
Counsel for Appellees

Cory L. Andrews
Washington Legal Foundation
2009 Massachusetts Avenue, N.W.
Washington, DC 20036
Counsel for Amicus Washington Legal Foundation

Brian D. Boone
Alston & Bird
101 South Tryon Street
Suite 4000
Charlotte, NC 28280

David R. Venderbush
Alston & Bird
90 Park Avenue
15th Floor
New York, NY 10016
Counsel of Amicus Chamber of Commerce of the United States

Joe G. Hollingsworth
Hollingsworth
1350 I Street, N.W.
Washington, DC 20005
Counsel for Amicus American Tort Reform Association and Pharmaceutical Research and Manufacturers of America

---

O P I N I O N

ROTH, Circuit Judge:

This case involves allegations that the anti-depressant drug Zoloft, manufactured by Pfizer, causes cardiac birth defects when taken during early pregnancy. In support of their position, plaintiffs, through a Plaintiffs' Steering Committee (PSC), depended upon the testimony of Dr. Nicholas Jewell, Ph.D. Dr. Jewell used the "Bradford Hill" criteria[1] to analyze existing literature on the causal connection between Zoloft and birth defects. The District Court excluded this testimony and granted summary judgment to defendants. The PSC now appeals these orders, alleging that 1) the District Court erroneously held that an expert opinion on general causation must be supported by replicated observational studies reporting a statistically significant association between the drug and the adverse effect, and 2) it was an abuse of discretion to exclude Dr. Jewell's testimony. Because we find that the District Court did not establish such a legal standard and did not abuse its discretion in excluding Dr. Jewell's testimony, we will affirm the District Court's orders.

I.

This case arises from multi-district litigation involving 315 product liability claims against Pfizer, alleging that Zoloft, a selective serotonin reuptake inhibitor (SSRI), causes cardiac birth defects. The PSC introduced a number of experts in order to establish causation. The testimony of each of these experts was excluded in whole or in part. In particular, the court excluded all of the testimony of Dr. Anick Bérard (an epidemiologist), which relied on the "novel

[1] *See* Section II.B *infra*.

technique of drawing conclusions by examining 'trends' (often statistically non-significant) across selected studies."[2] The PSC filed a motion for partial reconsideration of the decision to exclude the testimony of Dr. Bérard, which the District Court denied. The PSC then moved to admit Dr. Jewell (a statistician) as a general causation witness. Pfizer filed a motion to exclude Dr. Jewell, and the District Court conducted a *Daubert*[3] hearing.

The District Court considered Dr. Jewell's application of various methodologies, reviewing his expert report, rebuttal reports, party briefs, and oral testimony. The District Court first examined how Dr. Jewell applied the traditional methodology of analyzing replicated, significant results. While Dr. Jewell discussed many groupings of cardiac birth defects, he focused on the significant findings for all cardiac defects and septal defects. Dr. Jewell presented two studies reporting a significant association between Zoloft and all cardiac defects (Kornum (2010)[4] and Jimenez-Solem (2012)[5]). He also presented five studies reporting a

---

[2] *In re Zoloft (Sertraline Hydrochloride) Prods. Liab. Litig. (Zoloft I)*, 26 F. Supp. 3d 449, 465 (E.D. Pa. 2014). Since Dr. Jewell seems to provide similar testimony, we take into account the District Court's rationale in excluding Dr. Bérard.

[3] *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993).

[4] JA 1059-67. Jette B. Kornum, *et al.*, *Use of Selective Serotonin-Reuptake Inhibitors During Early Pregnancy and Risk of Congenital Malformations: Updated Analysis*, 2 Clin. Epidemiol. 29 (2010).

[5] JA 1040-51. Espen Jimenez-Solem, *et al.*, *Exposure to Selective Serotonin Reuptake Inhibitors and the Risk of*

significant association between Zoloft and septal defects (Kornum (2010), Jimenez-Solem (2012), Louik (2007),[6] Pedersen (2009),[7] and Bérard (2015)[8]). After excluding two studies from its consideration,[9] the District Court expressed two concerns with the remaining studies: Jimenez-Solem (2012), Kornum (2010), and Pedersen (2009). First, despite the fact that the remaining studies produced consistent results, the District Court did not consider them to be independent replications because they used overlapping Danish

---

Congenital Malformations: A Nationwide Cohort Study, 2 British Med. J. Open 1148 (May 2012).

[6] JA 5622-34. Carol Louik, *et al.*, *First-Trimester Use of Selective Serotonin-Reuptake Inhibitors and the Risk of Birth Defects*, 356 N. Eng. J. Med. 2675 (June 2007).

[7] JA 1030-39. Lars H. Pedersen, *et al.*, *Selective Serotonin Reuptake Inhibitors in Pregnancy and Congenital Malformations: Population Based Cohort Study*, 339 British Med. J. 3569 (Sept. 2009).

[8] JA 5987-99. Anick Bérard, *Sertraline Use During Pregnancy and the Risk of Major Malformations*, 212 Am. J. Obstet. Gynecol. 795 (2015).

[9] The District Court noted that during the trial, a transcription error was found in Louik (2007), which led to a significant result for septal defects being reclassified as insignificant. JA 65. The New England Journal of Medicine (NEJM) required the author to revise his discussion in light of this change. Additionally, multiple people tried to replicate the results in Bérard (2015)—including Dr. Jewell, a member of the PSC's legal team, and Pfizer's experts—and failed. The District Court did not allow Dr. Jewell to rely on Bérard (2015) after Dr. Jewell consequently "expressed a lack of confidence" about its reliability on cross-examination. JA 64-65.

populations. Second, a larger study, Furu (2015),[10] included almost all the data from Jimenez-Solem (2012), Kornum (2010), and Pedersen (2009) and did not replicate the findings of those studies. Dr. Jewell did not explain the reasons why this attempted replication produced different results or why the new study did not contradict his opinion.

The court then examined Dr. Jewell's reliance on insignificant results, noting that it was very similar to Dr. Bérard's methodology. The court noted that Dr. Jewell did not provide any evidence that the epidemiology or teratology[11] communities value statistical significance[12] any

---

[10] JA 4395-4404. Kari Furu, *et al.*, *Selective Serotonin Reuptake Inhibitors and Venlafaxine in Early Pregnancy and Risk of Birth Defects: Population Based Cohort Study and Sibling Design*, 350 British Med. J. 1798 (Mar. 2015). This study was not available to Dr. Jewell when he prepared his report, but the District Court noted that Dr. Jewell testified that he was familiar with it. JA 63, 7297-327.

[11] As the District Court noted, "[t]eratology is the scientific field which deals with the cause and prevention of birth defects. . . . [Where a drug is alleged to be] a teratogen, it is common to put forth experts whose opinions are based on epidemiological evidence." JA 52.

[12] The findings in these studies are often expressed in terms of "odds ratios." Odds ratios are merely "a measure of association." JA 2446. An odds ratio of 1, in the context of these studies, generally means that there is no observed association between taking Zoloft and experiencing a cardiac birth defect. Since these odds ratios are just estimates, a confidence interval is used to show the precision of the estimate. JA 2439-40. If the confidence interval contains the

less than it has traditionally been understood.[13] The court also expressed concern that Dr. Jewell inconsistently applied his "technique" of multiplying p-values[14] and his trend analysis.

The District Court critiqued several other techniques Dr. Jewell used in analyzing the evidence. First, Dr. Jewell rejected meta-analyses on which he had previously relied in a lawsuit against another SSRI, Prozac. The meta-analyses reported insignificant associations with birth defects for Zoloft but not for Prozac. Dr. Jewell rationalized his decision to ignore these meta-analyses because the "heterogeneity"[15] within its Zoloft studies was significant; the District Court

---

odds ratio of 1, the risk of cardiac birth defects while taking Zoloft is not considered "significantly" greater than the risk while not taking Zoloft.

[13] The District Court instead noted that the NEJM's treatment of the Louik (2007) transcription error suggests that the epidemiology and teratology communities still strongly value significance. JA 67.

[14] A "p-value" indicates the likelihood that the difference between the observed and the expected value (based on the null hypothesis) of a parameter occurs purely by chance. JA 2396. In this context, the null hypothesis is that the odds ratio is one; rejecting the null hypothesis suggests there is a significant association between Zoloft and cardiac birth defects.

[15] The District Court quoted Dr. Jewell in defining heterogeneity as "the measure of the variation among the effect sizes reported in [various] studies [and] . . . where heterogeneity is significant, the source of variation should be investigated and discussed." JA 70.

accepted this explanation but questioned why Dr. Jewell "fails to statistically calculate the heterogeneity" across other studies instead of relying on trends.[16] Second, Dr. Jewell reanalyzed two studies, Jimenez-Solem (2012) and Huybrechts (2014),[17] both of which had originally concluded that there was no significant effect attributable to Zoloft.[18] The District Court questioned his rationale for conducting, and tactics for implementing, this reanalysis. Finally, Dr. Jewell conducted a meta-analysis with Huybrechts (2014) and Jimenez-Solem (2012). The District Court questioned why he used only those particular studies.[19]

Based on this analysis, the District Court found that Dr. Jewell, tasked with explaining his opinion about Zoloft's effect on birth defects and reconciling contrary studies,

---

[16] JA 72.

[17] JA 4256-67. Krista F. Huybrechts, *et al. Antidepressant Use in Pregnancy and the Risk of Cardiac Defect*, 370 N. Eng. J. Med. 2397 (2014).

[18] Jimenez-Solem (2012) found that both current Zoloft users and SSRI users who "paused" their use during pregnancy had elevated risks of birth defects; this study concluded that the increased risk resulted from a confounding factor. JA 1044, 1047-48. Huybrechts (2014) found the increase in the risk of cardiac birth defects from taking Zoloft to be insignificant. JA 4257-67.

[19] Additionally, the District Court found that Dr. Jewell may have relied on a Periodic Safety Update Report, which contains literature reviews, and email correspondence summarizing a literature review. The District Court excluded this testimony because this is not the type of information statisticians generally rely on. This exclusion is not contested here.

"failed to consistently apply the scientific methods he articulates, has deviated from or downplayed certain well-established principles of his field, and has inconsistently applied methods and standards to the data so as to support his *a priori* opinion."[20] For this reason, on December 2, 2015, the District Court entered an order, excluding Dr. Jewell's testimony, and on April 5, 2016, the court granted Pfizer's motion for summary judgment. The PSC appeals the exclusion of Dr. Jewell and the grant of summary judgment.[21]

[20] JA 82.

[21] The PSC concedes that if the exclusion of Dr. Jewell was proper, it is unable to establish general causation and summary judgment was properly granted. Oral Argument Recording at 13:30-13:59, http://www2.ca3.uscourts.gov/oralargument/audio/16-2247In%20Re%20Zoloft.mp3.

II.[22]

In general, courts serve as gatekeepers for expert witness testimony. "A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if," *inter alia*, "the testimony is the product of reliable principles and methods[] and . . . the expert has reliably applied the principles and methods to the facts of the case."[23] In determining the reliability of novel scientific methodology, courts can consider multiple factors, including the testability of the hypothesis, whether it has been peer reviewed or published, the error rate, whether standards controlling the technique's operation exist, and whether the methodology is

---

[22] The District Court had jurisdiction over this claim under 28 U.S.C. § 1332 and 28 U.S.C. § 1407(a). We have jurisdiction under 28 U.S.C. § 1291. We review questions of law de novo, and questions of fact for clear error. *Ragen Corp. v. Kearney & Trecker Corp.*, 912 F.2d 619, 626 (3d Cir. 1990) (citations omitted). We review the decision to exclude expert testimony for abuse of discretion. *In re Paoli R.R. Yard PCB Litig.* (*In re Paoli*), 35 F.3d 717, 749 (3d Cir. 1994). However, when the exclusion of such evidence results in a summary judgment, we perform a "hard look" analysis to determine if a district court has abused its discretion. *Id.* at 750. An abuse of discretion occurs when a court's decision "rests upon a clearly erroneous finding of fact, an errant conclusion of law or an improper application of law to fact" or "when no reasonable person would adopt the district court's view." *Oddi v. Ford Motor Co.*, 234 F.3d 136, 146 (3d Cir. 2000) (internal quotation marks and citation omitted).

[23] Fed. R. Evid. 702.

generally accepted.[24] Both an expert's methodology and the application of that methodology must be reviewed for reliability.[25] A court should not, however, usurp the role of the fact-finder; instead, an expert should only be excluded if "the flaw is large enough that the expert lacks the 'good grounds' for his or her conclusions."[26]

Central to this case is the question of whether statistical significance is necessary to prove causality. We decline to state a bright-line rule. Instead, we reiterate that plaintiffs ultimately must prove a causal connection between Zoloft and birth defects. A causal connection may exist despite the lack of significant findings, due to issues such as random misclassification or insufficient power.[27] Conversely, a causal connection may not exist despite the presence of significant findings. If a causal connection does not actually exist, significant findings can still occur due to, *inter alia*, inability to control for a confounding effect or detection bias. A standard based on replication of statistically significant

---

[24] *In re Paoli*, 35 F.3d at 742.

[25] *Id.* at 745 ("However, after *Daubert* [*v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993)], we no longer think that the distinction between a methodology and its application is viable.").

[26] *In re TMI Litig.*, 193 F.3d 613, 665 (3d Cir. 1999), *amended*, 199 F.3d 158 (3d Cir. 2000) (internal quotation marks and citation omitted).

[27] Power is "the chance that a statistical test will declare an effect when there is an effect to be declared. This chance depends on the size of the effect and the size of the sample. Discerning subtle differences requires large samples; small samples may fail to detect substantial differences." JA 2409.

findings obscures the essential issue: a causal connection. Given this, the requisite proof necessary to establish causation will vary greatly case by case. This is not to suggest, however, that statistical significance is irrelevant. Despite the problems with treating statistical significance as a magic criterion, it remains an important metric to distinguish between results supporting a true association and those resulting from mere chance. Discussions of statistical significance should thus not understate or overstate its importance.

With this in mind, we proceed to the issues at hand. The PSC raises two issues on appeal: 1) whether the District Court erroneously concluded that reliability requires replicated, statistically significant findings, and 2) whether Dr. Jewell's testimony was properly excluded.

A.

The PSC argues that the District Court erroneously held that replicated, statistically significant findings are necessary to satisfy reliability. This argument seems to have been originally raised in the motion for reconsideration of Dr. Bérard's exclusion. Explaining its decision to exclude Dr. Bérard, the District Court cited a previous case, *Wade-Greaux v. Whitehall Labs, Inc.*, for the proposition that the teratology community generally requires replicated, significant epidemiological results before inferring causality.[28] The PSC

[28] *Zoloft I*, 26 F. Supp.3d at 454 n.13 (citing *Wade-Greaux v. Whitehall Labs., Inc.*, 874 F. Supp. 1441, 1453 (D.V.I. 1994) *aff'd*, 46 F.3d 1120 (3d Cir. 1994), *for text, see* No. 94-7199, 1994 WL 16973481 (3d Cir. Dec. 15, 1994)).

claims that in so doing, the District Court was asserting a legal standard that required replicated, significant findings for reliability.[29] Pfizer contends that the District Court merely made a factual finding about what the teratology community generally accepts.

Upon review, it is clear that the District Court was not creating a legal standard, but merely making a factual finding. The PSC argues that the District Court must have created a legal standard because it did not cite any sources other than *Wade-Greaux* to support its assertion that the teratology community generally requires replicated, significant epidemiological findings. However, in its initial exclusion of Dr. Bérard, the District Court noted that it looked to the standards adopted by "other epidemiologists, even the very researchers [Dr. Bérard] cites in her report."[30] Similarly, in

---

[29] Relatedly, the PSC claims that the District Court made a legal standard that "it was not reliable for Dr. Jewell to invoke studies observing non-statistically significant positive associations." However, the language cited does not support this conclusion: The District Court merely asserts that "experts may use congruent but non-significant data to bolster inferences drawn from replicated, statistically significant data. However, in this case . . . three of the studies Dr. Jewell relies upon to show replication use overlapping data . . . [and] have not been replicated by later, well-powered studies which attempt to control for various confounding factors and biases." JA 67-68.

[30] *Zoloft I*, 26 F. Supp. 3d at 456 ("There exists a well-established methodology used by scientists in her field of epidemiology, and Dr. Bérard herself has utilized it in her published, peer-reviewed work. The 'evolution' in thinking

its order denying general reconsideration of Dr. Bérard's exclusion, the District Court clarified that it "made this factual finding after review of the published literature relied upon by Dr. Bérard and other experts, as well as its review of the reports and testimony of both parties"[31] and merely used this factual finding as part of its FRE 702 analysis.[32] While the District Court does cite *Wade-Greaux*,[33] it uses it merely to show "that other courts have made similar findings regarding the prevailing standards for scientists in Dr. Bérard's field."[34]

---

about the importance of statistical significance Dr. Bérard refers to does not appear to have been adopted by other epidemiologists, even the very researchers she cites in her report.").

[31] *In re Zoloft (Sertraline Hydrocloride) Prod. Liab. Litig.* (*Zoloft II*), No. 12-2342, 2015 WL 314149, at *2 (E.D. Pa. Jan. 23, 2015); *see, e.g.*, JA 3962, 3971-72.

[32] While general acceptance by the scientific community is no longer dispositive in the Rule 702 analysis, it remains a factor that a court may consider. *Daubert*, 509 U.S. at 594 ("[A] known technique which has been able to attract only minimal support within the community may properly be viewed with skepticism.") (internal quotation marks and internal citation omitted).

[33] *Wade-Greaux*, 874 F. Supp. at 1453 (noting that "[a]bsent consistent, repeated human epidemiological studies showing a statistically significant increased risk of particular birth defects associated with exposure to a specific agent, the community of teratologists does not conclude that the agent is a human teratogen.").

[34] *Zoloft II*, 2015 WL 314149, at *2.

Second, the course of the proceedings make clear that the replication of significant results was not dispositive in establishing whether the testimony of either Dr. Bérard or Dr. Jewell was reliable. In fact, the District Court expressly rejected Pfizer's argument that the existence of a statistically significant, replicated result is a threshold issue before an expert can conduct the Bradford-Hill analysis.[35] In doing so, the District Court was clear that it was not requiring a threshold showing of statistical significance. Similarly, the District Court did not end its inquiry after analyzing whether there were replicated, significant results. Instead, the District Court examined other techniques of general trend analysis, reanalysis of other studies, and meta-analysis. Even though it ultimately rejected the application of these techniques as unreliable, it did not categorically reject alternative techniques, suggesting that it did not make a legal standard requiring replicated, significant results.

For these reasons, we find that the District Court did not require replication of significant results to establish reliability. Instead, it merely made a factual finding that teratologists generally require replication of significant results, and this factual finding did not prevent it from considering other evidence of reliability.[36]

---

[35] *Id.* ("In so doing, the Court rejected Pfizer's argument that the Court could exclude Dr. Bérard's opinion without even reaching her Bradford–Hill analysis, because the Bradford–Hill criteria should only be applied after an association is well established"); *see also Zoloft I*, 26 F. Supp. 3d at 462.

[36] The PSC also argues that the District Court did not discuss one study providing a significant, positive association between Zoloft and birth defects, Wemakor (2015). The PSC

B.

The second issue on appeal is whether it was an abuse of discretion for the District Court to exclude Dr. Jewell's testimony. Dr. Jewell utilized a combination of two methods: the "weight of the evidence" analysis and the Bradford Hill criteria. The "weight of the evidence" analysis involves a series of logical steps used to "infer[] to the best explanation[.]"[37] The Bradford Hill criteria are metrics that epidemiologists use to distinguish a causal connection from a mere association. These metrics include strength of the association, consistency, specificity, temporality, coherence, biological gradient, plausibility, experimental evidence, and analogy.[38] In his expert report, Dr. Jewell seems to utilize numerous "techniques" in implementing the weight of the evidence methodology. Dr. Jewell discusses whether the

---

claims this is "reversible error because it inaccurately depicted Dr. Jewell's opinion as unsupported by replicated, non-overlapping data." Pfizer argues that the District Court did not have to mention each study and that Wemakor is unreliable, as the authors themselves admit that their findings are "compatible with confounding by depression as indication or other associated factors/exposures." We conclude that this was not an error because it is clear the District Court considered Wemakor in the *Daubert* hearing. Even if the District Court had failed to consider Wemakor, we would find no error because it did not require replicated, statistically significant findings as a legal requirement.

[37] *Milward v. Acuity Specialty Prods. Grp., Inc.*, 639 F.3d 11, 17 (1st Cir. 2011) (internal quotation marks and citation omitted).

[38] JA 5652-56.

conclusions drawn from these techniques satisfy the Bradford Hill criteria and support the existence of a causal connection.[39]

Pfizer does not seem to contest the reliability of the Bradford Hill criteria or weight of the evidence analysis generally; the dispute centers on whether the specific methodology implemented by Dr. Jewell is reliable. Flexible methodologies, such as the "weight of the evidence," can be implemented in multiple ways; despite the fact that the methodology is generally reliable, each application is distinct and should be analyzed for reliability. In *In re Paoli R.R. Yard PCB Litigation*, this Circuit noted that while differential diagnosis—also a flexible methodology—is generally accepted, "no particular combination of techniques chosen by a doctor to assess an individual patient is likely to have been generally accepted."[40] Accordingly, we subjected the expert's specific differential diagnosis process to a *Daubert* inquiry.[41] We noted that "to the extent that a doctor utilizes standard diagnostic techniques in gathering this information, the more likely we are to find that the doctor's methodology is reliable."[42] While we did not require the expert to run specific tests or ascertain full information in order for the differential diagnosis to be reliable, we did require him to explain why his conclusion remained reliable in the face of

[39] Pfizer argues that PSC did not previously use the "weight of the evidence" terminology for the method followed by Dr. Jewell. We assume for the sake of argument that this was the purported methodology all along.
[40] *In re Paoli*, 35 F.3d 717, 758 (3d Cir. 1994).
[41] *Id.*
[42] *Id.*

alternate causes.[43]

This standard, while articulated with respect to differential diagnoses, applies to the weight of the evidence analysis. We have briefly encountered the Bradford Hill criteria/weight of the evidence methodology in *Magistrini v. One Hour Martinizing Dry Cleaning*, a nonprecedential affirmance of the District of New Jersey's exclusion of an expert.[44] The expert followed the weight of the evidence methodology, including epidemiological findings assessed using the Bradford Hill criteria. The District Court acknowledged that although the weight of the evidence methodology was generally reliable, "[t]he particular combination of evidence considered and weighed here has not been subjected to peer review."[45] Similar concerns are arguably present for the Bradford Hill criteria, which are

---

[43] *Id.* at 760 ("[T]he district court abused its discretion in excluding that opinion under Rule 702 unless either (1) Dr. Sherman or DiGregorio engaged in very few standard diagnostic techniques by which doctors normally rule out alternative causes *and* the doctor offered no good explanation as to why his or her conclusion remained reliable, or (2) the defendants pointed to some likely cause of the plaintiff's illness other than the defendants' actions and Dr. Sherman or DiGregorio offered no reasonable explanation as to why he or she still believed that the defendants' actions were a substantial factor in bringing about that illness.").

[44] *Magistrini v. One Hour Martinizing Dry Cleaning*, 68 F. App'x 356 (3d Cir. 2003).

[45] *Magistrini v. One Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 602 (D.N.J. 2002).

neither an exhaustive nor a necessary list.[46] An expert can theoretically assign the most weight to only a few factors, or draw conclusions about one factor based on a particular combination of evidence. The specific way an expert conducts such an analysis must be reliable; "all of the relevant evidence must be gathered, and the assessment or weighing of that evidence must not be arbitrary, but must itself be based on methods of science."[47] To ensure that the Bradford Hill/weight of the evidence criteria "is truly a methodology, rather than a mere conclusion-oriented selection process . . . there must be a scientific method of weighting that is used and explained."[48] For this reason, the specific techniques by which the weight of the evidence/Bradford Hill methodology is conducted must themselves be reliable according to the principles articulated in *Daubert*.[49]

In short, despite the fact that both the Bradford Hill and the weight of the evidence analyses are generally reliable,

---

[46] *Milward*, 639 F.3d at 17.

[47] *Magistrini*, 180 F. Supp. 2d at 602.

[48] *Id.* at 607.

[49] There has been very little circuit authority regarding the application of the Bradford Hill criteria in the weight of the evidence analysis. The First Circuit has warned against "treat[ing] the separate evidentiary components of [the] analysis atomistically, as though [the] ultimate opinion was *independently* supported by each." *Milward*, 639 F.3d at 23. In contrast, the Tenth Circuit briefly discussed the Bradford Hill criteria, and then separately conducted a *Daubert* analysis for each body of evidence. *Hollander v. Sandoz Pharm. Corp.*, 289 F.3d 1193, 1204-13 (10th Cir. 2002).

the "techniques" used to implement the analysis must be 1) reliable and 2) reliably applied. In discussing the conclusions produced by such techniques in light of the Bradford Hill criteria, an expert must explain 1) how conclusions are drawn for each Bradford Hill criterion and 2) how the criteria are weighed relative to one another. Here, we accept that the Bradford Hill and weight of the evidence analyses are generally reliable. We also assume that the "techniques" used to implement the analysis (here, meta-analysis, trend analysis, and reanalysis) are themselves reliable. However, we find that Dr. Jewell did not 1) reliably apply the "techniques" to the body of evidence or 2) adequately explain how this analysis supports specified Bradford Hill criteria. Because "*any* step that renders the analysis unreliable under the *Daubert factors renders the expert's testimony inadmissible*,"[50] this is sufficient to show that the District Court did not abuse its discretion in excluding Dr. Jewell's testimony.

1.

It was not an abuse of discretion for the District Court to find Dr. Jewell's application of trend analysis, reanalysis, and meta-analysis to the body of evidence to be unreliable. Here, we assume the techniques listed are generally reliable and rest on the fact that they were unreliably applied. As stated in *In re Paoli*, use of standard techniques bolster the inference of reliability;[51] nonstandard techniques need to be well-explained. Additionally, if an expert applies certain techniques to a subset of the body of evidence and other

[50] *In re Paoli*, 35 F.3d at 745.
[51] *Id.* at 758.

techniques to another subset without explanation, this raises an inference of unreliable application of methodology.[52]

First, we find no abuse of discretion in the District Court's determination that Dr. Jewell unreliably analyzed the trend in insignificant results. Dr. Jewell applied this technique by qualitatively discussing the probative value of multiple positive, insignificant results. In justifying this approach, he relied on a quantitative method by which one can calculate the likelihood of seeing multiple positive but insignificant results if there were actually no true effect.[53] However, after alluding to this presumably reliable mathematical calculation technique for analyzing trends in even insignificant results, Dr. Jewell did not actually implement it; instead he qualitatively discussed the general trend in the data. In light of the opportunity to actually conduct such quantitative analysis, his refusal to do so—without explanation—suggests that he did not reliably apply his stated methodology.[54]

Even assuming the reliability of Dr. Jewell's version of

---

[52] *See Magistrini*, 180 F. Supp. 2d at 607 (noting that a scientific method of weighting must be explained to prevent a "conclusion-oriented selection process.").

[53] Dr. Jewell used this as an illustrative example in his report and at the *Daubert* hearing but on appeal PSC identifies this technique as Fisher's combined probability test. Insofar as this is part of a meta-analysis or is sensitive to the same heterogeneity issues articulated by Dr. Jewell, we reiterate our concerns below.

[54] JA 69 ("[T]he Court finds Dr. Jewell's failure to apply the methodology he outlined to the studies he reviewed problematic.").

trend analysis, Dr. Jewell identified trends and interpreted insignificant results differently based on the outcome of the study. The District Court concluded that Dr. Jewell "selectively emphasize[d] observed consistency . . . only when the consistent studies support his opinion."[55] Dr. Jewell emphasized the insignificance of results reporting odds ratios below 1 but not the insignificance of those reporting odds ratios above 1. He also paid attention to the upper bounds of the confidence intervals associated with odds ratios below 1, but not to the lower bounds.

Second, we interpret the District Court's discussion of heterogeneity as raising the concern that Dr. Jewell selectively used meta-analyses. He did this in two ways: First, without explanation, Dr. Jewell performed a meta-analysis on two studies but not on any of the other studies. The District Court questioned why Dr. Jewell did not conduct a meta-analysis on the remaining studies instead of using the qualitative general trend analysis. While Dr. Jewell was not required to do specific tests, the lack of explanation made his inconsistent application of meta-analysis to certain studies unreliable.[56] Second, when he did perform a meta-analysis, Dr. Jewell only included two studies utilizing "exposed" and "paused" groups even though each had a different definition

[55] JA 69.

[56] Dr. Jewell admitted that he did not "attempt to do a meta-analysis where [he] defined an a priori – an a priori inclusion/exclusion set of criteria, generated a return set of studies, assessed heterogeneity and then considered whether by further adjustment or accommodation, [he] could come up with a meaningful set of statistics." He cryptically claimed that he "determined you couldn't." JA 4898.

of "paused," without an adequate explanation for why these studies can be lumped together. He also inexplicably excluded another study (Kornum (2010)) utilizing similar methodology. Again, while there may have been legitimate reasons for these inconsistencies, the fact that he did not give an adequate explanation for doing so makes his testimony unreliable.

Finally, Dr. Jewell reanalyzed two studies to control for confounding by indication. The need for conducting this reanalysis on Huybrechts (2014) was unclear. Dr. Jewell said that he wanted to control for indication by comparing the outcomes for "paused" Zoloft users to "exposed" Zoloft users; however, the study already controlled for indication. If Dr. Jewell wanted to correct for misclassification, the original study already controlled for that as well through extensive sensitivity analyses.[57] Given that the study originally concluded that Zoloft was not associated with a statistically significant increase in the likelihood of birth defects, this reanalysis seems conclusion-driven.

Ultimately, the fact that Dr. Jewell applied these techniques inconsistently, without explanation, to different subsets of the body of evidence raises real issues of reliability. Conclusions drawn from such unreliable application are themselves questionable.

[57] It is true that these sensitivity analyses had less power because they involved looking at a subset of the population, making them less likely to find a significant difference; however, we could not find that Dr. Jewell has raised this point as a reason for reanalysis.

2.

Using the techniques discussed above, Dr. Jewell went on to evaluate the Bradford Hill criteria. While Dr. Jewell did discuss the applicable Bradford Hill criteria and how he weighed the factors together, he did not explain how he drew conclusions for certain criteria, namely the strength of association and consistency.

Dr. Jewell concluded that the strength of association weighs in favor of causality. In doing so, he focused on studies reporting odds ratios between two and three (Colvin (2011),[58] Jimenez-Solem (2012), Malm (2011),[59] Pedersen (2009), and Louik (2007)). He rationalized that such a large association is unlikely to be associated with confounding alone.[60] He later bolstered this argument by estimating the percent of the effect generally attributable to confounding by indication. He estimated this percent by observing the percent decrease in odds ratios after controlling for indication over a few studies. When pressed by counsel at the *Daubert* hearing, Dr. Jewell admitted that this was not a scientifically

---

[58] JA 6011-28. Lyn Colvin, *et al.*, *Dispensing Patterns and Pregnancy Outcomes for Women Dispensed Selective Serotonin Reuptake Inhibitors in Pregnancy*, 91 Birth Defects Res. A Clin. Mol. Teratol. 142 (2011).

[59] JA 7697-7707. Heki Malm, *et al.*, *Selective Serotonin Reuptake Inhibitors and Risk for Major Congenital Anomalies*, 118 Obstetrics & Gynecology 111 (2011).

[60] Dr. Jewell also notes that the link between depression and cardiac defects being missing undercuts the confounding by indication argument. JA 7468-69.

rigorous adjustment.[61] Such reliance on ad hoc adjustments supports the District Court's decision to exclude Dr. Jewell's testimony.

Similarly, while Dr. Jewell found that the causal effect of Zoloft on cardiac birth defects is consistent, it is not clear how he drew this conclusion. As noted above, Dr. Jewell classified insignificant odds ratios above one as supporting a "consistent" causality result, downplaying the possibility that they support *no* association between Zoloft use and cardiac birth defects. While an insignificant result *may* be consistent with a causal effect, Dr. Jewell's discussion is too far-reaching, sometimes understating the importance of statistical significance. For example, Furu (2015)—a study that incorporated almost all the data in Pedersen (2009), Jimenez-Solemn (2012), and Kornum (2010)—included a larger sample but, unlike the former three studies, reported no significant association between Zoloft and cardiac birth defects. Insignificant results can occur merely because a study lacks power to produce a significant result, and, all else being equal, a larger sample size increases the power of a test.[62] Unless there are other significant differences, we

---

[61] JA 7470-71 ("I said, I didn't put that in my report. I put in that if you wanted as a statistician, if somebody came to me now as you're sort of hinting at and said [Colvin] didn't adjust for confounding, well, that could make a big impact, I agree, it could, just if I knew nothing else. . . . [A] statistician knows from doing simulations and computation that we alluded to yesterday how much of an impact could you take -- get from adjusting for confounding even though in this particular population we [aren't] able to do it. It's not a definitive result.")

[62] Insofar as Dr. Jewell finds Furu to be less powerful than the

would expect Furu to be better able to capture a true effect than the preceding three studies. While an insignificant result from a low-powered study does not necessarily undermine a statistically significant result from a higher-powered study, the opposite argument (*i.e.*, that an insignificant finding from a presumably better-powered study is evidence of consistency with significant findings from lower-powered studies) requires further explanation.[63] While there may be a reason that such a result could be consistent with the past significant effects, Dr. Jewell did not meaningfully discuss why this may be.[64] Without adequate explanation, this argument understates the importance of statistical significance. Like the expert in *Magistrini*, Dr. Jewell should have "sufficiently discredit[ed] other studies that found *no association* or a negative association with much more precise confidence intervals, [or] sufficiently explain[ed] why he did not accord weight to those studies."[65] Claiming a consistent result without meaningfully addressing these alternate explanations, as noted in *In re Paoli*, undermines reliability.[66]

---

previous studies based on factors other than sample size, he has not articulated this argument.

[63] For example, Dr. Jewell could have argued that, despite having a larger sample, Furu (2015) was not better powered for other reasons or utilized flawed methodology.

[64] In fact, upon appeal, the PSC argues that Furu (2015) *is* consistent with Dr. Jewell's causal result merely because it reports odds ratios above one (1.05 and 1.13).

[65] *Magistrini*, 180 F. Supp. 2d at 607 (emphasis added).

[66] *In re Paoli*, 35 F.3d at 760 (noting the importance of explaining why a conclusion remains reliable in the face of alternate explanations).

For these reasons, the District Court determined that Dr. Jewell did not consistently assess the evidence supporting each criterion or explain his method for doing so. Thus, it was not an abuse of discretion to find that Dr. Jewell's application of the Bradford Hill criteria was unreliable.

This is not to suggest that all of the District Court's criticisms were necessarily justified. For example, the fact that in his reanalysis Dr. Jewell drew a different conclusion from a study than its authors did is not necessarily a problem. Similarly, his imposition of a different assumption about the "exposed" group in Huybrechts (2014) did not require expert knowledge about psychology; he was merely testing the robustness of the results to Huybrechts' original assumption. Similarly, the District Court credited the claim that overlapping samples did not provide replicated results, despite the fact that Dr. Jewell claimed it provided some informational value.[67] These inquiries are more appropriately left to the jury.

On the whole, however, the District Court did not improperly usurp the jury's role in assessing Dr. Jewell's credibility. There is sufficient reason to find Dr. Jewell's testimony was unreliable. Indeed, "*any* step that renders the analysis unreliable under the *Daubert factors renders the expert's testimony inadmissible*."[68] The fact that Dr. Jewell unreliably applied the techniques underlying the weight of the evidence analysis and the factors of the Bradford Hill analysis satisfies this standard for inadmissibility.

---

[67] JA 7164 (noting that overlapping analysis still "provides a modicum of replication").

[68] *In re Paoli*, 35 F.3d at 745.

III.

This case involves complicated facts, statistical methodology, and competing claims of appropriate standards for assessing causality from observational epidemiological studies. Ultimately, however, the issue is quite clear. As a gatekeeper, courts are supposed to ensure that the testimony given to the jury is reliable and will be more informative than confusing. Dr. Jewell's application of his purported methods does not satisfy this standard. By applying different techniques to subsets of the data and inconsistently discussing statistical significance, Dr. Jewell does not reliably analyze the weight of the evidence. Selecting these conclusions to discuss certain Bradford Hill factors also contributes to the unreliability. While the District Court may have flagged a few issues that are not necessarily indicative of an unreliable application of methods, there is certainly sufficient evidence on the record to suggest that the court did not abuse its discretion in excluding Dr. Jewell as an expert on the basis of the unreliability of his methods. For these reasons, we will affirm the orders of the District Court, excluding the testimony of Dr. Jewell and granting summary judgment in favor of Pfizer.