Documented: Footnotes and Misreported Sources

Abstract

Though footnotes may seem like technicalities to the reader of medical and allied literature, work in these disciplines is by no means independent of textual sources. How often are sources checked? Here I look into a number of medically-related papers, several of them historic, which misreport or otherwise misuse footnoted material. Some of the abuses are inadvertent, some clearly not, but all are nontrivial. In the unique case of D. L. Rosenhan's celebrated—and as we now know, fabricated—study 'On Being Sane in Insane Places', a check of any of several listed sources (or even an ordinarily attentive reading of the text itself) would have suggested strongly that something was not right. In all cases of misrepresented sources examined here, detection of the abuse requires nothing more or less than due diligence—the review of source material with appropriate care. In the absence of such diligence, serious abuses can go undetected for decades. Regardless of the presumption that the humanities are tied to pre-existing texts as the sciences are not, or even that the sciences free us from dependence on the past and its works, the evaluation of published work will require the scrutiny of sources as long as sources are used.

Keywords

Evidence, trial, placebo, source, text

Shadowing the medical literature is a body of information of unknown magnitude consisting of clinical-trial data withheld from publication (Ioannidis, 2014). Necessarily, this dark matter is invisible to readers of the literature. However, certain information seems to possess a degree of invisibility despite being part of the literature itself: footnotes. After all, it is not an article's sources but its findings, in particular its novel findings, that solicit our interest. A long train of footnotes at the end of a medical article strikes us as tiresome, like a parade that lasts so long that the spectators go home.

How many readers have worked through the 78 footnotes attached to Freud's early attempt at fame, 'On Coca' (1884)? The sheer length of the list conveys an impression of erudition, even though the author delved into the literature on cocaine but two weeks before sending his article for publication, and by his own admission became acquainted with some of the literature only at second hand (Byck, 1974: 55). That the notes garble names, titles, dates and places suggests to Freud's freethinking biographer Frederick Crews that he was under the influence of the drug itself when he composed them (Crews, 2017: 65-66). Inserted among references to other investigators in the text of the article are reports of Freud's own experiments with cocaine and pronouncements on the drug's merits and possibilities. In effect, a multitude of names and notes provides a backdrop against which Freud stages his originality, and gives a eulogy to cocaine the aspect of a sober medical judgment.

'On Coca' embodies in a striking form the distinction between findings and sources that structures medical and other literature to this day. While a research paper offers findings, it also embeds them in a context of prior work and may be tied to textual sources as intricately as work in the humanities, regardless of the customary distinction between these fields of inquiry. Such a paper needs to be read with attention to its sources, as the product of a historical discipline—all the more because the very depth and extent of its entanglement with texts heighten the possibility of reporting errors and distortions. But attending to sources means delving into footnotes, and who does that?

To the impatient reader, footnotes in a discipline like medicine are at once a tedious formality and a distraction from the page, and attaching importance to them may seem like a great stir over not much, in the tradition of the private war between Gibbon and a sniping critic

over 383 references in two chapters of his *Decline and Fall of the Roman Empire* (Grafton, 1997: 99f). Here I challenge the prejudice against footnotes not by declamation but by showing the cost of ignoring them. Specifically, I examine a number of papers which inadvertently or otherwise misrepresent sources, in each instance with important issues at stake. Of these papers, one (Beecher's 'The Powerful Placebo') is the acknowledged foundation of a literature and two others (Feighner et al.'s 'Diagnostic Criteria for Use in Psychiatric Research' and Rosenhan's 'On Being Sane in Insane Places') have garnered well over 10,000 citations between them, facts suggesting that a publication can sit squarely in the spotlight for years on end without irregularities in its use of source material attracting attention. In the unique case of Rosenhan's sensational paper, even his critics at the time seem not to have caught on to his distortion of one cited source after another. In that all misrepresentations instanced here should have been apparent to anyone who reviewed the source with ordinary diligence, the papers serve to expose the practices of their own readers.

How common is the misrepresentation of sources? No one can say, but it seems naïve to assume that a general neglect of footnotes by readers encourages the careful use of source material by authors. Somewhere in the back of their mind, authors who misreport a source or cherry-pick a finding or remove data from its original context must sense that no one is watching.

Case 1: 'The Powerful Placebo' (1955)

The foundation of the extensive literature on the placebo effect is usually identified as Henry Beecher's 'The Powerful Placebo', since cited over 2500 times (Beecher, 1955). While other investigations of the placebo effect took place around the same time, Beecher seems in retrospect to have been chosen by history, in that the methodologically demanding trial of drug against placebo for which he argues in 'The Powerful Placebo' later rose to the status of a norm. According to Beecher, the placebo effect must be controlled for because left to itself it can muddle clinical judgment, making a worthless drug appear effective and vice versa. It is to establish the confounding potential of this treacherous variable that he tabulates 15 studies in which subjects treated with placebo improved.

The documentation of 'The Powerful Placebo' is a shambles, with some bibliographical data given in footnotes, some in the table just mentioned, some in both, some in neither. It is an index of a collective neglect of source-checking that almost a half century elapsed before two investigators, Gunver Kienle and Helmut Kiene, discovered that 10 of the 15 tabulated studies are misreported by Beecher. Exposing serious flaws in 'The Powerful Placebo' of which inaccurate reporting is but one, they provide details in a single case of the latter: Beecher's claim that in a trial of treatments for experimental cough, placebo relieved 36% of a group of 22 patients and 43% of another group of 22.

> However, the actual result was, that under none of the placebo administrations could any significant change for the better be demonstrated. Besides, there were no 22 placebo-treated patients (the groups were much smaller), and there were no reports about any 36% or 43% of patients. Thus, Beecher's quotation was wrong (which is

amazing, as Beecher himself had been one of the authors of the original publication). (Kienle and Kiene, 1997: 1316)

Alerted by Kienle and Kiene, I looked into Beecher's sources and soon encountered another misreport. The same Table 2 in 'The Powerful Placebo' lists a 1952 study of cough suppressants in which 37% of patients were relieved by placebo even though the trial had but one subject. The author of the trial considers a study population of one an advantage in that it eliminates the variability of placebo responses (Hillis, 1952: 1231). Beecher purports to derive his calculation of the power of the placebo from a survey of such variations.

Then there is Beecher's gross misrepresentation of a historic trial of a drug for seasickness.

Listed in Table 2 in 'The Powerful Placebo' is a study by Gay and Carliner, with no title or journal given, which supposedly had 33 subjects, 58% of whom were relieved by placebo—the highest response recorded in the table. In actuality there are three overlapping reports of this study, one in *Bulletin of the Johns Hopkins Hospital*, one in *Science*, one in *Transactions of the Association of American Physicians*, all published in 1949 (Gay and Carliner, 1949a; Gay and Carliner, 1949b; Gay, Carliner and Moore 1949), and on reviewing them we discover that Beecher cherry-picked a single finding from a large trial which, all in all, constitutes as complete a triumph of drug over placebo as we are ever likely to witness. (Here I rely mainly on the fullest exposition of the trial, the first.) Contrary to Beecher's representation of the trial as an exhibition of the power of the placebo, the test drug proved to be almost 100% effective

whether used to prevent or treat seasickness, while placebo proved worthless as a preventive and of undetermined, but at best slender, value as a treatment.

One might have expected Beecher to handle the Gay and Carliner trial data with care, given his stated belief in the importance of studying stress and distress outside the laboratory (Beecher, 1955: 1606). Conducted aboard the notably unstable General Ballou in rough seas and involving not 33 but 182 subjects who received placebo under one experimental condition or another (including 59 who received it specifically as a treatment), this trial established Dramamine as both a preventive and treatment of seasickness. Meticulously designed, the study not only paired drug and placebo groups but used subjects as their own control by switching them from one to the other. The crossover phases in particular brought out the categorical difference between drug and placebo, with subjects who fell sick on placebo being relieved by Dramamine in short order, and others who had been relieved by Dramamine falling sick when switched to placebo, in all cases without knowing which treatment was which. In the final tally, 372 of 389 cases of seasickness—that is, 96%—were completely relieved by Dramamine within one hour. In no phase and no report of the trial does placebo perform in a manner remotely comparable; indeed, the difference in the failure rates of drug and placebo is significant at the level of $1/10^6$ (Gay and Carliner, 1949a: 482). While the placebo effect can indeed confuse the evaluation of drugs, here was a trial where placebo could not possibly obscure or be mistaken for the signal of drug with a success rate approaching totality, either before or after the onset of seasickness. By contrast, in another study in Table 2 of 'The Powerful Placebo' drug and placebo performed indistinguishably (Wolf and Pinsky, 1954).

If Gay and Carliner had treated fully half, instead of only a small number, of the seasick with placebo, with no cross-over, the result would have been shocking. Denied the relief available to their companions by the luck of the draw, scores on placebo would have been reduced to prostration simply to make a point that was already undeniable. Beecher, however, holds up the Ballou experiment as a particularly strong example of 'the powerful placebo', quoting the figure of 58% (that is, 19) of a group of 33 who recovered on placebo, as given in Table II of the text in the *Bulletin of the Johns Hopkins Hospital*. The full account of the episode reads as follows:

> [A] group of 99 men was the control group for compartment 4-E. Within 12 hours after departure from New York, 33 (33.3%) reported to the sick bay. These men were given a placebo, one capsule every five hours and upon retiring. Two days later, they again reported to the sick bay and the following facts were recorded: nineteen men whose complaints had been nausea and dizziness had recovered within 12 hours. The lactose treatment of the 19 men was discontinued, and they had no return of symptoms during the remaining days of the voyage. Fourteen men became progressively worse on the placebo and now complained of excessive nausea, extreme dizziness, and prolonged vomiting. . . . Complete relief followed in all 14 within one-half hour after the first dose of 100 mg. [of Dramamine]. (Gay and Carliner, 1949a: 479-80)

Even this, the best showing of placebo in the study, confirms the telltale pattern of drug speedily remedying placebo's failures. But how strong is the showing, really? What does the

figure of 58% signify, considering that seasickness 'may clear spontaneously after two or three days . . . when the apparatus of equilibrium accustoms itself to the motion of a ship at sea' (Gay and Carliner, 1949a: 483), and the 19 recoveries occurred toward the beginning of the voyage? Gay and Carliner distinctly imply that the 19 subjects simply adjusted to the sea—'gained their sea legs' (Gay, Carliner and Moore, 1949: 200)—inasmuch as none needed any further treatment over the remaining seven days of the voyage (even as the ship rolled more violently), in contrast to the many who relapsed when Dramamine was withdrawn. Just as lactose did not make 14 worse, in all probability it did not make 19 better. Beecher, however, interprets any improvement in a placebo group as evidence of a placebo effect, an error flagged by Kienle and Kiene among others.

Additionally, Beecher ignores the telling finding that those given placebo preventively in the Ballou trial fared no better than those given nothing.

Unless we compare a placebo group with similar others left untreated, we run the risk of inflating the placebo effect by crediting it with the results of spontaneous improvement. As we know, Beecher neglected this precaution. In the Ballou trial, however, he was confronted with 881 subjects deliberately given neither drug nor placebo to prevent seasickness, and this imposing fact he ignored, presumably because it endangered his thesis. In the same tables of the Gay and Carliner study where he found the isolated figure of 58% relieved by placebo he would have seen that 29% of 123 men who received placebo when the ship left New York became seasick, *as did 22% of those who received nothing*. Contrary to the legend of the powerful placebo, a preventive effect of taking placebo is nowhere to be found.

As if this were not enough, Beecher also knew that 19 recoveries in a group of 33 subjects do not tell the whole story. An elegant summary of the Ballou data appears in Table V of the *Bulletin of the Johns Hopkins Hospital* report, as follows:

	Dramamine				Placebo		
	Treated	Failures	% Failures		Treated	Failures	% Failures
Prophylactic Use	134	2	1.4		123	35	28.4
Therapeutic Use	319	8	2.5		***59***	38	64.4

We know that 19 of a group of 33 seasick men treated with placebo improved for one reason or another. However, when a second group suffering from nausea and vomiting, this one consisting of 26, were treated with placebo, only two improved, the other 24 complaining that the capsules they were given 'made them much worse' (Gay and Carliner, 1949a: 483). By pairing the two groups, Gay and Carliner arrive in Table V at the total of 59 subjects, of whom 35.6% were successfully treated with placebo—a figure virtually identical to the one proposed by Beecher as a constant across all placebo trials. Beecher uncouples these two episodes specifically linked in the version of the report he consulted, only one of them superficially suggestive of the power of the placebo. Citing a success while muting a failure, he reverts to an old ploy (Justman, 2017), with the twist that he is looking for an impressive showing of placebo, not medication. Beecher's concealment of the unfavourable finding seems to have escaped comment at the time even though the Gay and Carliner study 'attracted widespread attention' only a few years before 'The Powerful Placebo' (Strickland and Hahn, 1949: 359), and anyone

familiar with it in any version should have recognised that the figure of 58% was suspect and that the study as a whole was anything but a demonstration of the power of the placebo.

From Beecher's listing of a seasickness trial with 33 subjects, no one would imagine that the actual trial population aboard the General Ballou consisted of 1366. In effect, Beecher edits the original with a razor blade, eliminating everything and everyone with the exception of the group of 33, of whom 19 were putatively cured by placebo. And that finding he dramatises by claiming in the text of 'The Powerful Placebo' that Gay and Carliner set a high standard for placebo effectiveness: complete relief within one-half hour (Beecher, 1955: 1605). While Table II in the *Bulletin of the Johns Hopkins Hospital* report does use the figure of one-half hour, other recovery times are given in the surrounding pages; indeed, we have just seen that the same 19 who supply Beecher's figure of 58% 'recovered within 12 hours' on placebo (a figure consistent with spontaneous improvement). Not only, then, does Beecher suppress the negative counterpart of the 58% finding, he slants the interpretation of that finding itself. Why did he go to these lengths? Why did he include the Ballou study at all if placebo performed so poorly against Dramamine?

Beecher claims both that 'all studies that presented adequate data have been included' in Table 2 and that the 15 trials were 'chosen at random' (Beecher, 1955: 1603), a contradiction both halves of which are untrue. Far from including all trials with adequate data, Beecher omitted a number of carefully controlled trials in psychiatry (Shorter, 2011), among them a 1939 double-blinded, cross-over trial of benzedrine for severe depression in which placebo, in contrast to the test drug, proved almost completely worthless (Dub and Lurie, 1939). (Strictly speaking, of course, a trial of drug against placebo provides not just inadequate data but no

data for measuring the placebo effect because it lacks an untreated group.) As for Beecher's claim that the studies in Table 2 were randomly chosen, this is impossible to credit, given that seven of the 15 are his own. With so much of the data in Table 2 bearing his name, perhaps he had to admit a number of other sources lest he appear to be manufacturing his results. The figure of 58% plucked from a table in one of the Gay and Carliner reports offers what looks like strong independent confirmation of the placebo's power.

Speaking of himself in the third person, Beecher once wrote, 'This reviewer holds with Lord Kelvin that "when you can measure what you are speaking about, and express it in numbers, you know something about it . . . [but] when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of Science"' (Beecher, 1958: 253). Seeking to express the placebo effect in numbers in 'The Powerful Placebo', Beecher gathered data from a hand-picked assortment of studies—including a series run on a single patient and a single ambiguous finding from the largest and most definitive study of the lot—and concluded that on average 35% (+/- 2.2%) of subjects will respond to placebo.

At the time of 'The Powerful Placebo', a figure in the 35% range had already taken shape among researchers; according to one of the pioneers of the clinical trial, Harry Gold, a placebo response 'can be demonstrated in about 30 or 40 per cent of all patients with all sorts of disorders' (Gold, 1954: 726). It seems strange that Beecher overlooked Table V in Gay and Carliner's *Johns Hopkins* report showing 35.6% of cases of seasickness in two groups relieved by placebo—strange, that is, until we notice that the same table does away with the 58% figure Beecher prefers and records a 97.5% success rate for Dramamine. Though a placebo effect of

35.6% may seem impressive in the abstract, in the case at hand it includes spontaneous recoveries, and even so runs a woeful second to a highly effective drug. If Gay and Carliner had included 35 men who were given placebo preventively, became severely seasick within 12 hours after sailing from New York, and remained on a lactose regimen—to no avail—for another 36 hours before receiving Dramamine, the overall success rate of placebo treatment in the Ballou trial would have been 22% (21 of 94), near the bottom of Beecher's list.

How is it that the errors that fed into Beecher's constant remained unnoticed until Kienle and Kiene called attention to them? In his concluding plea for studies guarded against bias by means of randomisation, double-blinding, placebo controls and, ironically, statistical validation, Beecher drafted the constitution of the clinical trial as we know it. And yet it is not as if Beecher introduced the methodological principles that go into the making of the randomised clinical trial (RCT). To give but one example, a decade before 'The Powerful Placebo', randomised trials involving 563 subjects—including an untreated group of 303—demonstrated the complete non-efficacy of placebo in preventing seasickness (Tyler, 1946). It is a disturbing thought that a paper making such questionable use of sources as 'The Powerful Placebo' should have been 'enormously influential—to the point of changing the way new medicines are evaluated' (Kirsch, 2010: 107-8). In turn, the ascent of the RCT may have served to protect 'The Powerful Placebo' from the sort of critical scrutiny it eventually received from Kienle and Kiene. Perhaps the article's status as herald of such a prestigious institution contributed to a general reluctance to dig into its sources. History looked kindly on 'The Powerful Placebo'.

But despite Beecher's emphasis on randomisation and double-blinding as indispensable safeguards of drug evaluation, Table 2 in 'The Powerful Placebo' does not limit itself to studies that abide by these methodological rules. A study with one subject cannot very well be randomised. As it happens, by randomisation Beecher does not mean a procedure for choosing subjects blindly or without bias but a procedure for varying the administration of treatments, as in one of his own listed studies wherein placebo was shuffled with morphine so that 'some patients received the placebo first, some second, some third, some fourth and some fifth' (Lasagna, Mosteller, von Felsinger and Beecher, 1954: 771). If Beecher had believed in randomisation of subjects, he would not have discussed in 'The Powerful Placebo' (citing the same study) the value of weeding out placebo responders so that the signal of a test drug can make itself heard. Probably the purest example of randomisation in the accepted sense in Table 2 is the Ballou experiment, with its assignment of troops by the hundreds to drug or placebo or nothing.

Whereas Beecher emphasizes that 'preservation of sound judgment . . . requires the use of the 'double blind' technique, where neither the subject nor the observer is aware of what agent was used or indeed when it was used' (Beecher, 1955: 1606b), the Ballou study was single blinded and its integrity survived intact. But we can go farther. If a few and then more and more seasick troops aboard the ship had simply been given the experimental drug Dramamine (whose safety had already been established) with no blinding, no controls, no placebo, and no intricate methodology, everyone would have been able to see then and there that it relieves seasickness. In this sense, the Ballou study does not really support Beecher's case for the necessity of methodologically exacting trials. While the military importance of

seasickness guaranteed that only the most rigorous study of a drug like Dramamine would be deemed probative (Tyler, 1949), the fact is that not every drug needs placebo to prove itself. Consider that on the previous voyage of the General Ballou along the same route, but from east to west, the vessel was so ravaged by seasickness that 'more than 100 intravenous injections of saline solution were necessary to relieve a number of dehydrated individuals' (Gay and Carliner, 1949a: 484), whereas over the course of the Dramamine study not a single such injection had to be given, despite rough weather.

One of the founders of clinical pharmacology, Beecher's sometimes-collaborator Louis Lasagna, came to believe that elaborate trials are unnecessary in the case of a drug whose effectiveness speaks for itself (Shorter, 2011: 196). So it was with Dramamine. Hours into the study, when not a single subject on a preventive regimen of the drug grew ill even as the corridors to the latrines filled with others too weak to stand, the effectiveness of the drug was already apparent to the naked eye. Beecher's readers, however, know nothing of this unless they investigate. If a trial in which placebo fails to outperform nothing stands as a showcase of the power of the placebo, then any manipulation of trial data is possible.

Case 2: The Feighner Criteria—Homosexuality (1972)

One of the most cited articles ever to appear in the psychiatric literature, 'Diagnostic Criteria for Use in Psychiatric Research' (1972) by Feighner et al.—the so-called Feighner criteria—presented putatively validated markers of 16 mental disorders at a time when conflicting diagnoses posed a serious embarrassment to the profession itself. These checklists

grew into the 'Research Diagnostic Criteria' (1978) by Spitzer et al., the immediate precursor of the symptom-based diagnostic system of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (1980), whose construction was overseen by Spitzer. The Feighner criteria thus represent the original template for the checklists of a diagnostic universe that has since grown to encompass hundreds of entities.

The Feighner document could not have rescued psychiatry from the awkwardness of conflicting diagnoses, if only because one of its own diagnoses was in conflict with itself. While Feighner et al. sought to ground diagnosis in sound evidence in the medical literature, they were in fact unable to offer evidence of any kind in support of one diagnosis: homosexuality. This is not to say that they cite no studies to back up the classification of homosexuality as a mental disorder, but that the cited studies never claim, suggest or show that it constitutes a disorder at all. In this sense the footnotes in question are spurious—a glaring incongruity that escaped notice and comment even during the civil war over the status of homosexuality that broke out in the American Psychiatric Association soon after (but certainly not because of) the publication of the Feighner criteria. If the anomaly of a Feighner category with false documentation had been noted in a timely manner, proponents of the de-listing of homosexuality would have been able to make a case that it satisfies none of the most rigorous tests of a mental disorder then in play: ironically enough, those employed by the Feighner group itself (Justman, 2020).

If you look into the studies cited by Feighner et al. in connection with the diagnosis of homosexuality, you find, in addition to the 1948 Kinsey report, (a) a study of homosexual men in two British prisons reporting 'only minor differences between homosexuals and normals'

(Hemphill, Leitch and Stuart, 1958: 1322); (b) a study of homosexual women that explicitly disregards the issue of pathology and suggests that 'a homosexual woman is able to produce and achieve, despite any psychological and social handicaps that she might have to cope with' (Saghir and Robins, 1969a: 199); and (c) a companion study of homosexual men reporting that 84% of the study population 'had not had disability, psychiatric or social, of any significant degree' (Saghir and Robins, 1969b: 227). The citation of the latter two studies as validation of a mental disorder is all the more bewildering in that the senior member of the Feighner group, Eli Robins, co-authored them. Robins offering these studies as evidence of the pathological nature of homosexuality is like Beecher misquoting his own work.

A still deeper mystery is the Feighner group's reference to the Kinsey report on the sexual practices of American males—a work that famously fails to deplore non-canonical modes of sexual expression—as support for listing homosexuality as a disorder. In both this volume and its companion on female sexuality, Kinsey 'by providing statistical findings of the prevalence of homosexual behavior in American society, . . . explicitly challenged the mental health profession's description of homosexuality as a psychological illness' (Chiang, 2008: 301). A work intended as an attack on the entrenched view of homosexuality as a pathology—a work that in fact moved influential psychologists and psychiatrists to rethink their position on that question—makes a poor choice as authority for the placement of homosexuality on a list of mental disorders. The Kinsey report no more justifies the classification of homosexuality as a disorder than placebo relieved 96% of cases of seasickness aboard the General Ballou**.**

While Feighner et al. could not know that Kinsey didn't merely discuss homosexuality or that he aspired to write a volume on the topic that would emancipate the United States from

sexual repression once and for all, they could not *but* know that *Sexual Behavior in the Human Male* did not judge disapproved forms of sexual behaviour as deviant, because it had been a national sensation for this very reason. In this celebrated and maligned volume cited by themselves, they could have found the following pronouncement on the psychiatric status of homosexuality:

> In view of the data which we now have on the incidence and frequency of the homosexual, in particular on its coexistence with the heterosexual in the lives of a considerable portion of the male population, it is difficult to maintain the view that psychosexual reactions between individuals of the same sex are rare and therefore abnormal or unnatural, or that they constitute within themselves evidence of neuroses or even psychoses. (Kinsey, Pomeroy and Martin, 1948: 659).

What were Feighner et al. thinking of when they cited the source of this statement as grounds for listing homosexuality as a mental disorder?

The Feighner group's citation of highly inapposite sources as justifying the placement of homosexuality on a short list of mental disorders marks a problem they could not solve. Informing the Feighner criteria is the authors' conviction that a valid mental disorder is distinguished by such objective insignia as characteristic presenting features, family history, and a well-marked clinical course. Using these identifiers in the manner of Occam's razor, they cut through the chaos of ill-defined disorders and reduced their numbers, at least provisionally, to 16. However, one of the finalists—homosexuality—has no presenting features or clinical

course and does not run in families, which is to say that it summarily fails the tests that define a disorder *as* a disorder, according to Feighner thinking. And yet Feighner et al. remain committed to the position that homosexuality is a mental disorder, possibly because gay people were 'seen frequently by the psychiatrist and the allied professionals' (Saghir and Robins, 1971: 506). But to judge the nature of homosexuality as such by cases that claim professional attention amounts to a kind of self-publication bias. Ironically, the studies of male and female homosexuality co-authored by Eli Robins which are cited in the Feighner criteria seek to investigate a world that lies outside the purview of psychiatry; they approach this terra incognita in a spirit of inquiry and with no presumption of pathology.

While the Feighner certification of homosexuality as a disorder became moot when the American Psychiatric Association de-listed homosexuality a year later, in the meantime those in favour of keeping it on the books did not cite the Feighner criteria, possibly because they had no wish to join the diagnostic reform movement the document represented, or because they disliked the empiricism of the Feighner group. Those who favoured diagnostic reform and believed that reform began with the removal of homosexuality from the Diagnostic and Statistical Manual had no wish to bring up the anomalous category in the Feighner document, either. In after years, no one seems to have made an issue of an extinct category on a list of allegedly validated disorders. It had become a kind of historical embarrassment.

Case 3: The Feighner Criteria—Depression (1972)

If the Feighner group's certification of homosexuality as a disorder receded into history, another misapplication of a source in the Feighner criteria—in this case probably as a result of carelessness—had profound implications that continue to be felt.

When Feighner et al. set out to codify specific diagnostic criteria for an array of mental disorders, they began with depression (Kendler, Muñoz and Murphy, 2010: 136), and for this they turned to an important 1957 study of the characteristics of manic-depressive disease led by Mandel Cohen, under whom Eli Robins had studied (Cassidy, Flanagan, Spellman and Cohen, 1957). The upshot was a set of diagnostic criteria that distinctly resembles, and serves as both precedent and prototype for, the DSM-III criteria for depression, with one cardinal symptom ('dysphoric mood' in the Feighner document) and five of eight secondary symptoms (virtually reproduced in DSM-III). Inasmuch as the DSM criteria for depression have changed little since 1980 and the disorder became the most common by far of all DSM diagnoses (Horwitz, 2011), the Feighner criteria for depression, based on the 1957 study, cast a historical shadow whose length the authors could hardly have imagined in 1972.

Though the Cohen study does have its own footnote in the Feighner paper, it figures simply as one of 11 studies of depression aggregated by the authors. From the text of the Feighner criteria no one could possibly divine the importance of the Cohen study to the framers of the document. However, if we actually review it, we find that while it does feature a checklist, and while this instrument does require one cardinal and six of ten secondary symptoms (almost all of which quite closely predict the corresponding Feighner criteria), by no means does the Cohen group propose a checklist for diagnostic use. The checklist was used in this case to select, among a population of hospitalized patients *already* diagnosed with manic-

depressive disease, those most ill, who would therefore presumably manifest the symptoms of their illness most clearly—the purpose of the study being to distinguish, at least provisionally, the symptoms characteristic of depression. (The group's thinking might have been to identify a degree of impairment so profound that it has virtually a biological character; hence such checklist symptoms as lack of sleep, appetite and sexual interest [Moncrieff, 2009: 126].) The study per se found poor sleep, poor concentration and poor mood more prevalent in the depressed group than controls, which cannot be much of a surprise, given that the study population was chosen on the basis of these symptoms in the first place. Unlike these three symptoms, however, some important clinical features of depression identified and emphasized by Cohen et al., such as a fear of losing one's mind, 'spells of various kinds' (Cassidy, Flanagan, Spellman and Cohen, 1957: 1535), and the attribution of the depression itself to a clearly erroneous cause, did not reappear in the Feighner checklist and eventually sank from view.

Not only did the Cohen group not use a checklist for diagnostic purposes, they recommend against doing so. As they state in the final paragraph of their Conclusion, 'Manic-depressive disease can be diagnosed by the usual medical procedure of using history, examination, and laboratory data' (Cassidy, Flanagan, Spellman and Cohen, 1957: 1545)—nothing less. The group's concept of history-taking was both rich and demanding. To get an idea of the care that went into the eliciting of history in the Cohen study, consider that patients admitted into the study were given a questionnaire with 188 items, an exercise the authors characterize as more systematic than, but not otherwise different in kind from, the usual questioning of patients. The Feighner criteria streamline this labor-intensive method of

diagnosis, thereby making the diagnosis of depression available on a scale even they did not envision.

Constructing an accurate history in cases of depression seems particularly important because the disease can damage the patient's own sense of history. As Cohen et al. stress, depression leads to lost jobs, broken marriages, reckless decisions, all of which can appear to the affected person as 'causes' of his or her condition. In the composite picture of depression that emerges from the Cohen study, such profoundly reversed thinking seems no less characteristic of the disorder than symptoms that lend themselves to a checklist, such as insomnia or poor appetite. But whereas the Cohen group knew enough about the respective histories of the patients in their study to conclude that they traced their illness to an event that could not possibly have caused it (if only for reasons of chronology), the DSM criteria for depression would enable the diagnosis of persons whose sadness really does spring from events—an overreach of psychiatry's writ, as Horwitz and Wakefield emphasise in their trenchant study of the overdiagnosis of depression, *The Loss of Sadness* (Horwitz and Wakefield, 2007).

In removing the Cohen checklist from its original context, Feighner et al. made a serious error. Seeking to reassert or rebuild psychiatry's identity as a medical discipline, they stressed such disease-like traits as the characteristic course of a disorder. Yet depression in response to the adversities of life itself does not follow the same course as depression in the Cohen population, a third of whom had been ill for more than a year, and whose belief that a specific misfortune caused their condition was in most cases a fallacy born of that condition. 'Normal sadness remits when the context changes for the better or as people adapt to their losses'

(Horwitz and Wakefield, 2007: 29]. For many and perhaps most of the Cohen population, this is exactly how depression does *not* behave.

A commemoration of the Feighner criteria published in 2010 by a group, two of whom took part in their composition, gives no sign that Feighner et al. ever recognised the liberties they took with the Cohen study. The construction of a template for the most popular of all diagnostic criteria in the DSM system was an act of inadvertence. How could this have happened? Confronting a detailed, data-rich study that displays a sample checklist (filled out) on its second page by way of illustration, the Feighner authors simply cut and pasted, in the process severing the list from its original context. It's as if the checklist jumped out at them, proposing itself irresistibly as a shortcut through the problems of diagnosis. In turn, the simplification of the disorder under the auspices of DSM-III and its successors (now requiring only two weeks of symptoms) served to abstract it from its human context, enabling the interpretation of appropriate sadness as a mental disorder.

If the triumph of the clinical trial may have inhibited critical examination of 'The Powerful Placebo', the ascendancy of the DSM after the introduction of diagnostic checklists in 1980, and the uptake of the depression diagnosis in particular, would have made a probe of a footnote in the Feighner criteria seem academic.

Case 4: 'On Being Sane in Insane Places' (1973)

Asked about the historical importance of the Feighner criteria, Spitzer replied that if not for Feighner, DSM-III 'would have been delayed and would likely have looked quite different'

(Kendler, Muñoz and Murphy: 141). About another celebrated paper of the 1970's one might say even more. Spitzer's successor Allen Frances once remarked that if not for D. L. Rosenhan's 'On Being Sane in Insane Places' (1973), Spitzer 'could never have done what he did with DSM-III' (Cahalan, 2019: 197), implying that the existential threat to psychiatry posed by Rosenhan gave Spitzer what he needed to rally the profession behind a new diagnostic system.

Published in *Science* and an immediate cause célèbre, Rosenhan's article recounts an experiment in which eight sane individuals, including himself, presented at twelve psychiatric hospitals claiming to have heard voices say improbable things like 'Thud'. Diagnosed as suffering from schizophrenia in every case but one, they were held an average of 19 days despite giving no sign of abnormality, and treated with the utmost contempt. Such was the malign power of the diagnostic label attached to the pseudo-patients (as Rosenhan calls them) that it poisoned the perceptions and judgments of staff and *remained* attached for no other reason than that it was applied in the first place. In Rosenhan's telling, the real insanity lies in the institution of the psychiatric hospital; hence the paper's title. How odd, then, that he would have us believe that 'the administrator and the chief psychologist' of one of these madhouses colluded with him when he got himself admitted (Rosenhan, 1973: 251)—a detail that mirrors the 1952 report of an anthropologist who indeed smuggled himself into a psychiatric hospital with the knowledge of 'two members of the senior staff' (Caudill, Redlich, Gilmore, & Brody, 1952: 315), albeit not with the intention of testing and exposing the hospital.

Recently an investigator who obtained Rosenhan's notes and tracked down every lead concluded that no such experiment as the one reported by him ever took place (Cahalan, 2019). (While Rosenhan did act his way into Haverford State Hospital in Pennsylvania under a false

name, he told the doctor more than a thin story about hearing voices: he claimed that he was 'sensitive to radio signals and could hear what people are thinking' [Cahalan, 2019: 183], and put copper over his ears for his own protection, all of which makes the diagnosis of schizophrenia a lot less casual than it appears in his paper.) By 2019, of course, Rosenhan's study been legendary for so long that its falsehoods had become, for many, conventional wisdom. But the veracity of 'Insane Places' could and should have been challenged at the time. While few besides Spitzer seem to have suspected that Rosenhan's experiment was a fabrication, anyone who read the paper with ordinary care should have been able to tell that something was not right.

While the study protocol supposedly had the pseudos 'cease simulating *any* symptoms of abnormality' upon admission (Rosenhan, 1973: 251), in Rosenhan's account they exhibited no outward symptoms in the first place. They heard voices, but neither their speech nor behaviour was at all disordered. (With great disingenuousness, Rosenhan later emphasized that the study protocol called for the volunteers to simulate one and only one symptom—auditory hallucinations—in order to lessen the psychological demands on them [Rosenhan, 1975: 469].) The symptom of hearing voices was well chosen by Rosenhan, as it accounts almost credibly for the reported fact that each and every pseudo was admitted—notwithstanding that public hospitals at the time actually admitted about 40% of voluntary patients, as documented by the only one of Rosenhan's sources identified as critical of his point of view (Gove, 1970: 877); and all hospitals but one in the Rosenhan study were supposedly public. On the other hand, the normality of the pseudos' speech and behaviour at the hospital door destroys Rosenhan's claim that the institutions failed in every instance to register a clear

and obvious change of demeanour once they were admitted. No change of demeanour occurred. The argument of 'Insane Places' falls to the ground upon an ordinarily careful reading of its own text.

If psychiatric hospitals were as wedded to their own preconceived notions as Rosenhan and other exponents of labeling theory contend, it's a wonder the volunteers in 'Insane Places' were discharged in less than three weeks, on average. Possibly Rosenhan wanted to convey the experience of the many patients in psychiatric hospitals whose stay was brief—a group that actually made up the majority at the time (Gove, 1970: 880). But precisely because their stay was brief, the pseudos in 'Insane Places' simply do not have time to incur the worst of the alleged effects of labeling.

According to Rosenhan, the ultimate harm of diagnostic labels such as those supposedly applied to the pseudos is that the patient *over time* comes to internalize them and even live them out. As he says, 'Eventually, the patient himself accepts the diagnosis, with all of its surplus meanings and expectations, and behaves accordingly' (Rosenhan, 1973: 254). Not once in 'Insane Places' is this ominous process borne out. Not the pseudo-patient held for 52 days, not one of the seven or eight discharged with a diagnosis of 'schizophrenia in remission' (as if the disease were dangling over their heads, ready to strike at any time), not even the long-term patients among whom the pseudos were housed are ever shown acting out the diagnosis they have been labeled with. Again the text of 'Insane Places' clearly fails to support its own polemic.

On what, then, does Rosenhan rest the audacious claim that a diagnosis once imposed becomes the patient's fate? On one of the cardinal works of the anti-psychiatry movement, Thomas Scheff's *Being Mentally Ill*.

The sentence 'Eventually, the patient himself accepts the diagnosis, with all of its surplus meanings and expectations, and behaves accordingly' is footnoted, with the following reference:

T. J. Scheff, *Being Mentally Ill: A Sociological Theory* (Aldine, Chicago, 1966)

Note the subtitle. Rosenhan certifies what he presents as a statement of fact by referring to a theory. But a theory can't certify a claim of fact, still less such an expansive claim as that a psychiatric label will inevitably realise itself at the patient's expense. Evidently Rosenhan didn't really see the need to verify the theory of labeling, since as we now know he never actually conducted the experiment recounted in the pages of *Science*.

The reader who advances beyond the subtitle of Scheff's study will find that after discussing the power of psychiatric labels to shape behaviour, colour the patient's self-conception, and lock in deviations from the norm, the author progressively qualifies his claims until little is left—certainly not enough to serve as a foundation for Rosenhan's principle of inevitability. Thus, on p. 101 of the edition of *Being Mentally Ill* cited by Rosenhan, in the Conclusion of the study's central chapter, Scheff concedes that 'many of the hypotheses suggested are largely unverified', and on p. 152 he summarises the state of the evidence as

follows: 'There is some evidence that too hasty exposure to psychiatric treatment may convince the patient that he is "sick," prolonging what might have been an otherwise transitory episode'. One qualifier is piled on another: 'some'; 'too hasty'; 'may'; 'might have been'. Far from supporting the law that the label inevitably crushes the patient, Scheff's evidence is too weak to support anything.

If it's possible to dilute even further a claim already so attenuated, Scheff does just that in his monograph's Conclusion. Writing of himself in the third person, he encapsulates his case for the labeling theory as follows:

> Acknowledging that the evidence was far from complete, both in amount and quality, the author concluded that the existing state of evidence favored this sociological theory, perhaps only slightly . . . Obviously the author is predisposed to accept the theory, and may not have been sufficiently impartial in his selection and evaluation of the evidence. Other investigators, more objective than the author, might review the state of evidence and come to a contrary conclusion. (Scheff, 1966: 197-98)

As if conscience-stricken, Scheff concludes the case for his theory with a sort of apology for the theory itself. Reviewing *Being Mentally Ill* in 1968, Rosenhan noted that it propounds 'a theory in terms of nine testable propositions', neglected to note the state of the evidence, and praised the author for 'preferring the modest to the flamboyant statement' (Rosenhan, 1968: 361). In 'Insane Places' Rosenhan prefers the flamboyant to the modest, conceals Scheff's qualifiers and disclaimers, cites his theory as if it had the status of a law, and reports his own study as a

confirmation of the theory even though it bears out not at all the baleful effect of diagnosis on patients, no matter whether they are held indefinitely in an 'insane' institution or have an ill omen like 'schizophrenia in remission' hanging over them. Scheff squared the circle in 1974 by holding up the Rosenhan study as strong corroboration of his theory of the social origin of mental illness (Scheff, 1974).

Scenting fraud in 'Insane Places', Spitzer answered the allegation of wholesale misdiagnosis with a study of his own, refusing to conceal the identity of the hospitals he put to the test and implicitly challenging Rosenhan—in a footnote—to do the same (Spitzer, 1975: 444). Once Spitzer came into possession of medical records proving the dishonesty of Rosenhan's anonymized account of his own admission into Haverford State Hospital, he could have exposed Rosenhan at any time. For reasons of his own he kept the secret. But just as anyone who reads 'Insane Places' with ordinary care can see that it does not confirm its own claim that the unnamed hospitals failed to notice an obvious change in the pseudo-patients' behaviour, so any reader who follows up on Rosenhan's citation of Scheff catches on to his opportunistic use of the latter's text. The evidence in this instance is not confidential but in the public domain.

Nor does Rosenhan confine his unscrupulous handling of source material to Scheff. He abuses sources right and left—so many that it's a wonder that none of the cited authors seem to have cried foul. The practice begins in the first footnote, which The first footnote in 'Insane Places' attaches to the statement, 'More generally, there are a great deal of conflicting data on the reliability, utility, and meaning of such terms as "sanity", "insanity", "mental illness", and "schizophrenia"'. The note itself refers the reader to two comprehensive reviews, one

published in 1967, one in 1971, neither of which happens to mention sanity or insanity (Zubin, 1967; L. Phillips and Draguns, 1971).

In defense of the proposition that the mentally ill are 'society's lepers' (Rosenhan, 1973: 254], Rosenhan cites a 1970 article by Sarbin and Mancuso which, as it happens, strongly implies that the allegedly intolerant general public would see no reason to hospitalize someone who behaves as innocuously as the pseudo-patients in 'Insane Places'. 'The survey data have shown repeatedly that only persons who exhibit the most exaggerated deviations will be regarded as mentally ill, and even when this is done, the general public only infrequently makes the recommendation that such persons be hospitalized' (Sarbin and Mancuso, 1970: 169). So much for the notion that society demands the sequestration of the mentally ill in the psychiatric equivalent of a leper colony. Another of Rosenhan's cited sources disputes the cliché of public abhorrence of the mentally ill, reporting that 50% of respondents 'could imagine themselves falling in love with someone who had been mentally ill', 85% agreed that 'people who have some kinds of mental illness can be taken care of at home' and so on (Crocetti and Lemkau, 1965). Predicated as they are on the impermanence of mental illness ('someone who had been mentally ill'), several of these responses belie Rosenhan's contention that the public believes mental illness 'endures forever'. If Rosenhan had planted clues that his study was a hoax, he could hardly have given better ones than some of his footnotes.

In connection with the issue of the arbitrariness of psychiatric categories, Rosenhan cites, among other sources, an article by Derek Phillips that has nothing to do with that topic and does not even mention psychiatric diagnosis (D. Phillips, 1963). After making the provocative claim that diagnosed psychiatric patients are condemned to live out their label,

Rosenhan returns to the theme of arbitrariness, referring to an article by Zigler and Leslie Phillips said to demonstrate that 'there is an enormous overlap in the symptoms presented by patients who have been variously diagnosed' (Rosenhan, 1973: 254): a point which, in turn, has nothing to do with the matter of noxious diagnoses. Rosenhan does not note that the article in question ('Psychiatric Diagnosis: A Critique') censures the excesses of labeling theory, in particular the indiscriminate rejection of psychiatric categories (Zigler and L. Phillips, 1961b). In direct opposition to Rosenhan, the authors *defend* the principle of psychiatric classification. They do not consider the diagnosis of schizophrenia (for example) a meaningless but highly prejudicial tag, as in 'Insane Places'; on the contrary, they observe that by carefully delineating symptoms one can distinguish 'those schizophrenics with good prognosis' from 'those with poor prognosis' (Zigler and L. Phillips, 1961b: 615). It is hard to know why Rosenhan mentions Zigler and Phillips other than to project an appearance of scholarship or perhaps make a tactical retreat from a sensational claim about the power of labels for which he can offer no support even in a piece of fiction.

However, by mentioning Zigler and Phillips as if they somehow corroborated his polemic, Rosenhan opens himself to the charge of using sources dishonestly. It is not just that he obscures the authors' thoughtful analysis of psychiatric classification. The fact is that the same authors, in the same year, in the same journal offer evidence against the inflammatory claim that labels themselves dictate the outcome of cases. They do so by demonstrating a correlation between clinical outcome and maturity or social competence (Zigler and L. Phillips, 1961a). Rosenhan's non-mention of the latter article, even as he misleadingly cites its

companion, revealed his dishonesty as an investigator decades before he was exposed as an academic trickster.

Following the publication of 'Insane Places', Rosenhan went on to test the limits of audacity by toying with an identified source. In a retrospective comment on the controversy he incited, he noted that he and his confederates were not the first to study a psychiatric hospital covertly from within. In what now looks like a private joke, he wrote,

> More than two decades earlier, Caudill (1958; Caudill, Redlich, Gilmore, & Brody, 1952), had spent considerable time in a psychiatric hospital simulating a florid pattern of symptomology throughout. He was consumed with guilt over deceiving his colleagues and his report of his experiences was an excruciating warning to subsequent scientific generations that such elaborate deceptions can have enormous personal consequences. (Rosenhan, 1975: 469)

As brief as it is, this statement abounds with misrepresentations and concealments. In the investigation reported in the 1952 article Caudill, an anthropologist, spent two months undercover in a psychiatric hospital with the permission of two administrators (Caudill, Redlich, Gilmore and Brody, 1952: 315), unlike all of Rosenhan's supposed confederates. (While Rosenhan could have answered his critics by pointing out that he himself had authorization for impersonating a patient, something evidently deterred him from repeating this lie.) Caudill did not simulate his way through the door. His intent was not to expose the insanity of the institution but to learn about the patients' behaviour with one another, especially their group

dynamics. According to the circumstantial account given in the 1952 article, inside the hospital Caudill exhibited no florid symptoms, unless playing bridge falls into that category. Dissatisfied with the results of this covert exercise, Caudill undertook a lengthier study in 1952-53 in propria persona, not as a pseudo-patient. While he did come to feel that the price of conducting undercover research was 'too high' (Caudill, 1958: xiv), he did not mean that anyone following in his footsteps should take care to fake only a single symptom (as Rosenhan preposterously suggests), still less that someone brave enough to shoulder an enormous burden of guilt can reveal the inner workings of a psychiatric hospital as no one else can.

In a Foreword preceding Caudill's admission of what Rosenhan portrays as overwhelming guilt, one of the 1952 co-authors criticises surreptitious study of a psychiatric hospital as ill-judged, unethical and unlikely to yield findings of value (Caudill, 1958: ix). Rosenhan suppresses this word to the wise. A brazen misrepresentation, his reference to Caudill as a predecessor demonstrates just how much an author can get away with unless and until readers check sources.

Case 5: Arts Engagement and Mortality (2019)

Many medical papers entail such a caravan of footnotes that numbers alone seem to favour some sort of reporting error. The following example concerns a recent paper in *BMJ* whose argument is compromised by the inadvertent misreport of information in a single paper among 42 cited.

In the study in question, 'The Art of Life and Death: 14 Year Follow-Up Analyses of the Associations Between Arts Engagement and Mortality in the English Longitudinal Study of Ageing', by Daisy Fancourt and Andrew Steptoe, some form of arts spectatorship (such as visiting museums or exhibitions, or attending theatre, concerts or opera) was associated with a mortality benefit even after possible confounders were taken into account (Fancourt and Steptoe, 2019). It seems implausible that any and all arts events regardless of their content have some element or essence in common beyond their character as social gatherings. The authors, however, find a graded relationship suggestive of a dose response between exposure to the arts and improved survival, such that some exposure yields some benefit, and more frequent exposure more benefit. By stipulation, 'frequent' means in this case 'every few months or more'.

While Fancourt and Steptoe fault two Scandinavian papers with similar findings for devoting 'little attention to the frequency of engagement required for associations with longevity to be seen', source-checking reveals that one of the two papers defines frequency of engagement quite specifically. A quarter of the random sample of Swedes that constituted the study population in that paper attended some form of cultural event 'at least 80 times a year', a figure adopted by the authors as their measure of frequent attendance (Bygren, Konlaan and Johansson, 1996).

This source, too, yielded a gradient of benefits, with occasional attenders of cultural events showing a risk of mortality 24% less than non-attenders, and frequent attenders 57% less. The corresponding figures in the Fancourt and Steptoe study are 14% and 31%. Thus, although engaged Swedes attend cultural events perhaps 20 times more often than their UK

counterparts, they see only about twice the benefit. While it may be possible in principle for a ceiling effect for the benefits of arts attendance to set in at a quite low level, one struggles to take seriously a paper on that topic which incorporates, unbeknownst to the authors, a measure of the key variable so wildly different from its own.

That placebo relieved 58% (19 of 33) of one group on the General Ballou and 8% (2 of 26) of another suggests that the two figures probably don't measure the same thing. There is probably no single thing that visiting museums and attending arts events measures either, beyond mixing with others. In fact, a 2019 review of the literature—with no fewer than 962 references of its own—cited in 'The Art of Life and Death' and co-authored by Fancourt connects improved mortality not specifically to 'arts engagement' at all, but to leisure activities of all sorts, from gardening to pursuing a hobby to dining out. In this case the authors do *not* claim that a mortality benefit remains even after we account for confounders, stating only that 'these associations with mortality appear to be partly explained by socioeconomic factors and partly by reduction in sedentary behaviours' (Fancourt and Finn, 2019: 25-26].

While a bibliography with 962 entries beggars human capacity, it does allow for grazing. Searching studies that identify themselves as randomised, I soon came upon one article supposedly about the health value of 'the arts', which actually concerns 'active video games' as an exercise aid (Staiano et al., 2017). The inclusion of miscellaneous pursuits like this under the rubric of art provokes the question, What counts as art? The 962-reference review begins by defining art as centered on an object 'valued in its own right, rather than merely as a utility' (Fancourt and Finn, 2019: 1). Yet the review proceeds to argue tirelessly that art possesses the highest possible value as a utility—benefiting us in a thousand and one ways, from infancy to

old age—just as 'The Art of Life and Death' suggests that it can enhance and extend life itself. It is entirely in keeping with that suggestion that the editorial accompanying 'The Art of Life and Death' urges a concerted effort to realise the medical potential of the arts, so vital to the health of the community. 'Work must now be done to ensure that the health benefits of these activities are accessible to those who would benefit most' (Gill, Ellis and Clift 2019). But if art is enjoyed for its own sake and not for ulterior reasons, then does art promoted and consumed like some kind of miracle drug or super-vitamin remain art?

Usual Care

Just as it would be foolish to assume that misreported sources cannot remain hidden for long, it would be an error to suppose that originality or whatnot releases investigators from the necessity of borrowing in the first place. Of the articles considered here, the two that seem most original and whose impact proved most stunning—the Feighner criteria and the Rosenhan study—turn out to be less of a departure than they appear. The Feighner group indeed 'introduced for the first time the systematic application of operationalized criteria into psychiatry' (Kendler, Muñoz and Murphy: 140) but grounded the criteria themselves in the existing literature, or sought to. (On the other hand, while they did add their own touches, they did not conduct original studies 'to examine specifically the validity of such key features as the proposed cutoff of five of eight criteria for definite depression' [ibid].) The Rosenhan experiment of course never took place, but if it had, its originality would have consisted in the author's pirating of the Caudill precedent and clever application of labeling theory. As their

footnotes alone indicate, both the Feighner criteria and the Rosenhan study are thoroughly enmeshed in the existing literature.

With borrowing goes the possibility of misreport, a risk heightened when the reporters have a thesis (as with Beecher's 'powerful placebo') and when studies have complex root systems that readers are loath to explore. The misrepresentation of a source or its use out of context can vitiate the conclusions even of papers with great influence, as I have tried to show. The only defense is due diligence: the review of sources with the same sort of care that should be shown to the those of a work of, say, historiography. While a reader confronted with a lengthy bibliography may scarcely know where to begin, a number of the misrepresented sources I have examined call aloud for review. The reported placebo effect of 58% in the Gay and Carliner trial goes well beyond the figure Beecher proposed as a sort of numerical norm and represents the highest such response among all studies tabulated in 'The Powerful Placebo'. The Kinsey reports are so well known for their libertarianism that an article using one of them to the opposite effect opens itself to question. Rosenhan highlights the Caudill paper by placing himself in a tradition it inaugurates (Rosenhan, 1973: 251), whereas the very title of Scheff's work advertises the error of using it to validate an ambitious statement of fact. Fancourt and Steptoe themselves flag the study they misreport.

Certainly the references in medical and psychiatric literature are not exempt from investigation merely because they lie outside the fields most concerned with textual materials. It's not as if the authors reported only findings of such novelty that they render the examination of prior material pointless. And if the reader of, say, a psychiatric article should be prepared to review at least some of its references with the same sort of attention that would or

should be given to works cited in disciplines focused on textual sources, so too the reader of a literary work may find it necessary to check the psychiatric literature. Let the case of a 'nonfiction novel' that quotes extensively from an article in the *American Journal of Psychiatry* illustrate the due diligence required of a reader—but too often neglected in practice—wherever textual sources are in play.

I refer to Truman Capote's *In Cold Blood* (1965), the strangely elegant chronicle of the murder of a family of four in a Kansas farmhouse by a two ex-convicts (only one of whom, it seems, does the killing) drawn by the rumour of a safe holding a small fortune. The lengthy extract from the *AJP* article inserted—with date and title—into the text of *In Cold Blood* in connection with the trial of the two stands out all the more in that its style is entirely foreign to the narrative itself and nothing else like it appears. Few critics seem to have bothered reviewing the original from which it is drawn, even though Capote tells them where to find it. Perhaps readers of *In Cold Blood* skip the pedantic interpolation (which appears in the text just as the murderers' fate is being decided); or perhaps they have the impression that they are in the presence of a reporter so reliable that his reports don't need checking, or an originality of such an order that it reduces the investigation of sources to an exercise in triviality.

In the narrative of *In Cold Blood* we watch the two protagonists prepare for the crime in a deliberate and methodical manner. The victims, too, are bound and then killed deliberately and methodically, with each of those still alive knowing what the others suffered and what awaits them in turn. However, Capote would have us believe that Perry Smith had no intent to kill and knew not what he was doing when he executed at least the first of his victims, because he was in a sort of black-out state. To make his case Capote enters a long discourse on 'Murder

Without Apparent Motive' from the July 1960 issue of the *American Journal of Psychiatry*, as if someone could be said to motiveless who went to the scene of the crime in search of a safe, armed and explicitly resolved to leave no witness to testify against him.

Like the reader who checks Rosenhan's references, the reviewer of Capote's source is in for a surprise. The first example of 'murder without apparent motive' given in the article in question, but omitted by Capote, reads as follows:

> A 31-year-old chief petty officer in charge of a hospital, while talking casually to the 9-year-old daughter of one of his superior officers, suddenly grabbed the child, choked her, and held her head under water long after she was dead. A discontinuity existed in Thomas' mind as to what happened; he could not remember the beginning of the assault, but "suddenly discovered" himself strangling his young victim. (Satten, Menninger, Rosen and Mayman, 1960: 48)

A case like this has no more to do with the calculated execution of an entire family one by one, as chronicled in the text of *In Cold Blood*, than the Kinsey report has to do with the Feighner criteria. No amount of quoting and no feat of psychoanalytic theorizing can assimilate the consecutive murder of four people in the course of an attempted theft to the sudden and completely gratuitous killing of a child. And yet the fact-checkers of *The New Yorker*, where *In Cold Blood* first appeared, evidently did not object to Capote's importation of an intrusive and highly misleading discourse on unmotivated murder from the psychiatric literature in an obvious effort to palliate the crimes of Perry Smith. Possibly they merely verified the accuracy

of the extract. However, if we ourselves review the article Capote extracted *from*, we come to question not only his judgment in incorporating such incongruous material, at such inordinate length, into his own carefully fashioned text, but his posture as an objective reporter.

Note that the account of the chief petty officer 'suddenly' and for no reason killing a child is on its face irreconcilable with the crime in *In Cold Blood*. And so it is in most of the examples of misrepresented sources I have presented: we need only review an underlying source with usual care—to employ a term common in controlled studies of medical interventions—to see that it has been misapplied.

In the case of Beecher's misrepresentations, which went undetected more than forty years even as his paper was cited ritually in the literature, Kienle and Kiene simply checked the data given in 'The Powerful Placebo' against his listed sources. Following their lead, I reviewed the study listed with the highest placebo effect in Beecher's table of sources, only to find that he (a) picked out a single finding from a large trial in which the test drug totally eclipsed placebo; (b) ignored the strong likelihood that this single finding overstates the placebo effect; (c) ignored the untreated group that fared no worse than a group given placebo preventively; and (d) silently uncoupled his chosen finding from one with which it is paired by the authors of the study in the version he consulted. In effect, Beecher substitutes a cardboard trial for the actual one. Checking Kinsey's report on the sexual practices of American males against a document listing homosexuality as a mental disorder on Kinsey's authority is perhaps more laborious but no less straightforward than comparing Beecher's data with the reports of the seasickness trial itself.

There is much to be said for reading with ordinary care. So read, Rosenhan's famous paper does *not* show hospital staff missing a glaring change in the behaviour of the admitted pseudo-patients, because the pseudo-patients behaved normally to begin with, their one and only symptom being auditory hallucinations. If readers had looked into the cited critique of labeling theory by Gove, they might have realised how unrepresentative the Rosenhan study was, with each and every patient presenting at the door of a public hospital gaining admission. If readers had looked into one of Rosenhan's principal sources of labeling theory—Scheff's *Being Mentally Ill*, which claims in the end only an equivocal preponderance of evidence in favour of the theory itself—they might have marveled that this theory was confirmed unequivocally each and every time Rosenhan tested it: twelve times out of twelve. How flawlessly Rosenhan replicated his own results! Even as social science, like the medical literature, came to confront its replication problem, 'Insane Places', with its improbable success rates, stood decade after decade until at last it was exposed as a sham.

With its sensational account of an investigation like no other, Rosenhan's paper makes footnotes seem like technicalities. Some readers may have had their attention frozen by the drama of his text; some may have assumed the article had already been source-checked (it appeared in *Science*, after all) or that inspecting footnotes may befit a backward-looking discipline but not an advancing one. Each in its own way, the other examples of mishandled textual evidence presented here exploit the same prejudice against digging into underlying sources. Even *In Cold Blood*—a work that by no means pretends to science—counts on a reader's not following up an ostentatious extract of an article from the leading journal of

psychiatry in the United States. If Capote wagered that readers wouldn't bother to investigate this source, he won.

Regardless of the presumption that the humanities are tied to pre-existing texts as the sciences aren't, or even that the sciences free us from dependence on the past and its works, the evaluation of medical and other literature will require the scrutiny of sources as long as it contains footnotes. And the review of sources in a medical paper calls for the same sort of care required by the evaluation of textual evidence wherever else it is found. By the same token, the notion that following up on footnotes is a petty exercise constitutes a formula for carelessness; and as carelessness becomes collective its costs and implications mount up.

A risk of not investigating sources cited in a document like the Feighner criteria (said to rest on validating evidence contained in its sources) is that in time the document may accrue such influence that the original abuse no longer matters. Soon enough after the Feighner criteria for depression were incorporated into DSM-III, their misappropriation of earlier material became a moot point. They had become an enterprise too big to fail. In the case of the Rosenhan study, its exposure as a fabrication almost 50 years after the fact could not undo the damage that had been done over the decades when it was imbibed by generations of Psych 101 students.

A chapter of Susannah Cahalan's exposé of 'Insane Places' chronicles her excavation of its sixth footnote, a particularly mendacious aside in which Rosenhan claims to have omitted the data of one pseudo-patient in the interest of the study's integrity. The title of the chapter is 'The Footnote'. The canonization of sham might have been averted by timely investigation of small print like the reference to Scheff immediately preceding note six, or the spurious citation

of Crocetti and Lemkau in note three or Zigler and Phillips in note 16 or Sarbin and Mancuso in note 19, or the shameless reference to Caudill after 'Insane Places' was published.

A Matter of History

Timely investigation of the sources cited by the first of our authors might have fostered the practice of critical examination of trial data.

Some would say that despite his reporting errors and concealments, Beecher got it right in the end by his advocacy of the placebo-controlled trial. How ironic, then, that in practice the institution of the placebo-controlled trial led to the systematic burial of unwanted findings, as many have pointed out, among them Irving Kirsch. Investigating troves of unpublished trial data, Kirsch found antidepressants to be marginally superior to placebo, by and large. And if we review the devices used by trial sponsors to enhance the showing of these problematic drugs, we discover that some were trialed by Beecher. Thus, long before subjects deemed less likely to respond to the test drug were excluded from antidepressant trials (Kirsch, 2010: 72), Beecher argued in detail in 'The Powerful Placebo' that a clearer signal from a test drug may come through if placebo-responders are culled from the study population. The first of the two tables in 'The Powerful Placebo' illustrates this point. The second, as we know, displays what Beecher takes to be the placebo effect itself in a number of studies.

In an exposé of cherry-picking in *The Emperor's New Drugs*, Kirsch brings up a sponsored paper that reports data on 27 patients given Prozac, even though the study in question actually had 245 (Kirsch, 2010: 41). As we know, Beecher edited the 182 subjects who received placebo

in the Gay and Carliner study to 33, a reduction with the identical effect of making a finding look more important than it otherwise would. Thus he was able to represent the Ballou trial as a demonstration of the power of the placebo, in defiance of the account of the trial in all three versions. (As for the figure of 19 recoveries in this group of 33, it is in all likelihood meaningless.) Kirsch also objects to pooled analyses in which 'drug companies pick and choose which studies they wish to include' (Kirsch, 2010: 42). Beecher's constant of 35% derives from a pooled analysis of a medley of trials selected and interpreted by himself. His assertions notwithstanding, the studies aggregated in 'The Powerful Placebo' were not 'chosen at random' and do not constitute the entirety of available studies 'with adequate data', and obviously could not have satisfied both conditions at once. If readers alerted by this high-flying red flag had investigated Beecher's reporting practices, they and others might have become more critical evaluators of the reporting practices of the trial system on which 'The Powerful Placebo' proved to be 'enormously influential' (Kirsch, 2010: 107).

In opening I noted that while a literature of unreported findings lies buried in archives and other places of concealment, a literature also lies buried, but not exactly invisibly, in footnotes. We are now in a position to see that the second problem anticipates the first.

References

Beecher, H. (1955). 'The Powerful Placebo'. *JAMA* 159 (1955): 1602-06.

Beecher, H. (1958). 'Psychotomimetic Drugs'. *Journal of Chronic Diseases* 8 (1958): 253-85.

Byck, R. (1974). *Cocaine Papers of Sigmund Freud*. New York: New American Library.

Bygren, L., Konlaan, B. and Johansson, S. (1996). 'Attendance at Cultural Events, Reading Books or Periodicals, and Making Music or Singing in a Choir as Determinants for Survival: Swedish Interview Survey of Living Conditions'. *BMJ* 313 (1996); 1577-80.

Cahalan, S. (2019). *The Great Pretender: The Undercover Mission that Changed Our Understanding of Madness*. New York: Grand Central.

Cassidy, W., Flanagan, N., Spellman, M. and Cohen, M. (1957). 'Clinical Observations in Manic-Depressive Disease'. *JAMA* 164 (1957): 1535-46.

Caudill, W. (1958). *The Psychiatric Hospital as a Small Society*. Cambridge: Harvard University Press.

Caudill, W. Redlich, F., Gilmore, H. and Brody, E. (1952). 'Social Structure and Interaction Processes on a Psychiatric Ward'. *American Journal of Orthopsychiatry* 22 (1952): 314-34.

Chiang, H. H. (2008). 'Effecting Science, Affecting Medicine: Homosexuality, the Kinsey Reports, and the Contested Boundaries of Psychopathology in the United States, 1948-1965'. *Journal of the History of the Behavioral Sciences* 44 (2008): 300-18.

Crews, F. (2017). *Freud: The Making of an Illusion*. New York: Metropolitan.

Crocetti, G. and Lemkau, P. (1965). 'On Rejection of the Mentally Ill'. *American Sociological Review* 30 (1965): 577-78.

Dub, L. and Lurie, L. (1939). 'Use of Benzedrine in the Depressed Phase of the Psychotic State'. *Ohio State Medical Journal* 35 (1939): 39-45.

Fancourt, D. and Finn, S. (2019). 'What is the Evidence on the Role of the Arts in Improving Health and Well-Being? A Scoping Review'. WHO Regional Office for Europe, 2019 (Health Evidence Network synthesis report 67).

Fancourt, D. and Steptoe, A. (2019). 'The Art of Life and Death: 14 Year Follow-Up Analyses of the Associations Between Arts Engagement and Morality in the English Longitudinal Study of Ageing'. *BMJ* 2019;367:l6377 | doi: 10.1136/bmj.l6377.

Feighner, J., Robins, E., Guze, S., Woodruff, R., Winokur, G. and Munoz, R. (1972). 'Diagnostic Criteria for Use in Psychiatric Research'. *Archives of General Psychiatry* 26 (1972): 57-63.

Gay, L. and Carliner, P. (1949a). 'The Prevention and Treatment of Motion Sickness'. *Bulletin of the Johns Hopkins Hospital* 84 (1949): 470-87.

Gay, L. and Carliner, P. (1949b). 'The Prevention and Treatment of Motion Sickness 1. Seasickness'. *Science* 109 (1949): 359.

Gay, L., Carliner, P and Moore, J. (1949). 'The Prevention and Treatment of Motion Sickness'. *Transactions of the Association of American Physicians* 62 (1949): 196-203.

Gill, N., Ellis, V. and Clift, S. (2019). 'Cultural Activities Linked to Lower Mortality'. *BMJ* 2019;367:l6774 | doi: 10.1136/bmj.l6774.

Gold, H. (1954). 'How to Evaluate a New Drug'. *American Journal of Medicine* 17 (1954): 722-27.

Gove, W. (1970). 'Societal Reaction as an Explanation of Mental Illness: An Evaluation'. *American Sociological Review* 35 (1970): 873-84.

Grafton, A. (1997). *The Footnote: A Curious History.* Cambridge, Mass.: Harvard University Press.

Hemphill, R., Leitch, A. and Stuart, J. (1958). 'A Factual Study of Male Homosexuality'. *British Medical Journal,* 7 June 1958: 1317-23.

Hillis, B. R. (1952). 'The Assessment of Cough-Suppressing Drugs'. *Lancet*, 21 June 1952: 1230-35.

Horwitz, A. (2011). 'Creating an Age of Depression: The Social Construction and Consequences of the Major Depression Diagnosis'. *Society and Mental Health* 1 (2011): 41-54.

Horwitz, A. and Wakefield, J. (2007). *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow Into Depressive Disorder*, Oxford: Oxford University Press.

Ioannidis, J. (2014). 'Clinical Trials: What a Waste'. *BMJ* 349 (2014): 349:g7089; doi: 10.1136/bmj.g7089.

Justman, S. (2017). 'James Lind and the Disclosure of Failure'. *Journal of the Royal Society of Physicians of Edinburgh* 47 (2017): 384-87.

Justman, S. (2020). 'Buried in Silence: Homosexuality and the Feighner Criteria'. *Philosophy, Psychiatry, and Psychology* 27 (2020): 283-98.

Kendler, K., Muñoz, R. and Murphy, G. (2010). 'The Development of the Feighner Criteria: A Historical Perspective'. *American Journal of Psychiatry* 167 (2010): 134-42.

Kienle, G. and Kiene, H. (1997). 'The Powerful Placebo Effect: Fact or Fiction?' *Journal of Clinical Epidemiology* 50 (1997): 1311-18.

Kinsey, A., Pomeroy, W. and Martin, C. (1948). *Sexual Behavior in the Human Male.* Philadelphia: W. B. Saunders.

Kirsch, I. (2010). *The Emperor's New Drugs: Exploding the Antidepressant Myth.* New York: Basic, 2010.

Lasagna, L., Mosteller, F., von Felsinger, J. and Beecher, H. (1954). 'A Study of the Placebo Response'. *American Journal of Medicine* 166 (1954): 770-79.

Moncrieff, J. (2009). *The Myth of the Chemical Cure: A Critique of Psychiatric Drug Treatment.* Basingstoke: Palgrave Macmillan.

Phillips, D. (1963). 'Rejection: A Possible Consequence of Seeking Help for Mental Disorders'. *American Sociological Review* 28 (1963): 963-72.

Phillips, L. and Draguns J. (1971). 'Classification of the Behavior Disorders'. *Annual Review of Psychology* 22 (1971): 447-82.

Rosenhan, D. L. (1968). 'Madness: In the Eye of the Beholder'. *Contemporary Psychology: APA Review of Books* 13 (1968): 360-61.

Rosenhan, D. L. (1973). 'On Being Sane in Insane Places'. *Science* 179 (1973): 250-58.

Rosenhan, D. L. (1975). 'The Contextual Nature of Psychiatric Diagnosis'. *Journal of Abnormal Psychology* 84 (1975): 462-74.

Saghir, M. and Robins, E. (1969a.) 'Homosexuality: Sexual Behavior of the Female Homosexual'. *Archives of General Psychiatry* 20 (1969): 192-201.

Saghir, M. and Robins, E. (1969b). 'Homosexuality: Sexual Behavior of the Male Homosexual'. *Archives of General Psychiatry* 20 (1969): 219-29.

Saghir, M. and Robins, E. (1971). 'Male and Female Homosexuality: Natural History'. *Comprehensive Psychiatry* 12 (1971): 503-10.

Sarbin, T. and Mancuso, J. (1970). 'Failure of a Moral Enterprise: Attitudes of the Public Toward Mental Illness'. *Journal of Consulting and Clinical Psychology* 35 (1970): 159-73.

Satten, J., Menninger, K., Rosen, I. and Mayman, M. (1960). 'Murder Without Apparent Motive: A Study in Personality Disorganization'. *American Journal of Psychiatry* 117 (1960): 48-53.

Scheff, T. (1966). *On Being Mentally Ill: A Sociological Theory*. Chicago: Aldine.

Scheff, T. (1974). 'The Labelling Theory of Mental Illness'. *American Sociological Review* 39 (1974): 444-52.

Shorter, E. (2011). 'A Brief History of Placebos and Clinical Trials in Psychiatry'. *Canadian Journal of Psychiatry* 56 (2011): 193-97.

Spitzer, R. (1975). 'On Pseudoscience in Science, Logic in Remission, and Psychiatric Diagnosis: A Critique of Rosenhan's "On Being Sane in Insane Places"'. *Journal of Abnormal Psychology* 84 (1975): 442-52.

Spitzer, R., Endicott, J. and Robins, E. (1978). 'Research Diagnostic Criteria'. *Archives of General Psychiatry* 35 (1978): 773-82.

Staiano, A., Marker, A., Beyl, R., Hsia, D., Katzmarzyk, P. and Newton, R. (2017). 'A Randomized Controlled Trial of Dance Exergaming for Exercise Training in Overweight and Obese Adolescent Girls'. *Pediatric Obesity* 12 (2017): 120-28.

Strickland, Jr., B. and Hahn, G. (1949), 'The Effectiveness of Dramamine in the Prevention of Airsickness'. *Science* 109 (1949): 359-60.

Tyler, D. (1946). 'The Influence of a Placebo, Body Position and Medication on Motion Sickness'. *American Journal of Physiology* 46 (1946): 458-66.

Tyler, D. (1949). 'Dramamine and Motion Sickness'. *Science* 110 (1949): 170.

Wolf, S. and Pinsky, R. (1954). 'Effects of Placebo Administration and Occurrence of Toxic Reactions'. *Journal of the American Medical Association* 155 (1954): 339-41.

Zigler, E. and Phillips, L (1961a). 'Social Competence and Outcome in Psychiatric Disorder'. *Journal of Abnormal and Social Psychology* 63 (1961): 264-71.

Zigler, E. and Phillips, L. (1961b). 'Psychiatric Diagnosis: A Critique'. *Journal of Abnormal and Social Psychology* 63 (1961): 607-18.

Zubin, J. (1967). Classification of the Behavior Disorders'. *Annual Review of Psychology* 18: (1967): 373-406