A way to make sense of their view is to construe it as saying the observed mean is so out of line with what's known that we suspect the assumptions of the test are questionable or invalid. Suppose you have considerable grounds for this suspicion: signs of cherry picking, multiple testing, artificiality of experiments, publication bias, and so forth – as are rife in both examples given in Gelman and Carlin's paper. *You have grounds to question the result* because you *question the reported error probabilities*. Indeed, no values can be inferred if the error probabilities are spurious, the severity is automatically low.

One reasons, if the assumptions are met, and the error probabilities approximately correct, then the statistically significant result *would* indicate $\mu > 150.5$, P-value 0.07, or severity level 0.93. But you happen to know that $\mu \leq 150.5$. Thus, that's grounds to question whether the assumptions are met. You suspect it would fail an audit. In that case put the blame where it belongs.[6]

Recall the (2010) study purporting to show genetic signatures of longevity (Section 4.3). Researchers found the observed differences suspiciously large, and sure enough, once reanalyzed, the data were found to suffer from the confounding of batch effects. When results seem out of whack with what's known, it's grounds to suspect the assumptions. That's how I propose to view Gelman and Carlin's argument; whether they concur is for them to decide.

5.6 Positive Predictive Value: Fine for Luggage

Many alarming articles about questionable statistics rely on alarmingly questionable statistics. Travelers on this cruise are already very familiar with the computations, because they stem from one or another of the "P-values exaggerate evidence" arguments in Sections 4.4, 4.5, and 5.2. They are given yet another new twist, which I will call the diagnostic screening (DS) criticism of significance tests. To understand how the DS criticism tests really took off, we should go back to a paper by John Ioannidis (2005):

> Several methodologists have pointed out that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles should

[6] The point can also be made out by increasing power by dint of sample size. If n = 10,000, $(\sigma/\sqrt{n}) = 0.1$. Test T+ (n = 10,000) rejects H_0 at the 0.025 level if $\overline{X} \geq 150.2$. A 95% confidence interval is [150, 150.4]. With n = 100, the just 0.025 significant result 152 corresponds to the interval [150, 154]. The latter is indicative of a larger discrepancy. Granted, sample size must be large enough for the statistical assumptions to pass an audit.

> be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors or associations.
>
> It can be proven that most claimed research findings are false. (p. 0696)

First, do medical researchers claim to have "conclusive research findings" as soon as a single statistically significant result is spewed out of their statistical industrial complexes? Do they go straight to press? Ioannidis says that they do. Fisher's ghost is screaming. (He is not talking of merely identifying a possibly interesting result for further analysis.) However absurd such behavior sounds 80 years after Fisher exhorted us never to rely on "isolated results," let's suppose Ioannidis is right. But it gets worse. Even the single significant result is very often the result of the cherry picking and multiple testing we are all too familiar with:

> . . . suppose investigators manipulate their design, analysis, and reporting so as to make more relationships cross the $p = 0.05$ threshold . . . Such manipulation could be done, for example, with serendipitous inclusion or exclusion of certain patients or controls, post hoc subgroup analyses, investigation of genetic contrasts that were not originally specified . . . Commercially available 'data mining' packages actually are proud of their ability to yield statistically significant results through data dredging. (ibid., p. 0699)

The DS criticism of tests shows that if

1. you publish upon getting a single P-value < 0.05,
2. you dichotomize tests into "up-down" outputs rather than report discrepancies and magnitudes of effect,
3. you data dredge and cherry pick and/or
4. there is a sufficiently low probability of genuine effects in your field, the notion of probability to be unpacked,

then the probability of true nulls among those rejected as statistically significant – a value we call the false finding rate (FFR)[7] – differs from and can be much greater than the Type I error set by the test.

However one chooses to measure "bad evidence, no test" (BENT) results, nobody is surprised that such bad behavior qualifies. For the severe tester, committing #3 alone is suspect, unless an adjustment to get proper error probabilities is achieved. Even if there's no cherry picking, and your test has a legitimate Type I error probability of 0.05, a critic will hold that the FFR can be much higher than 0.05, if you've randomly selected your null hypothesis from a group with a sufficiently high proportion of true "no

[7] Some call it the false discovery rate, but that was already defined by Benjamini and Hochberg in connection with the problem of multiple comparisons. (see Section 4.6).

effect" nulls. So is the criticism a matter of transposing probabilistic conditionals, only with the twist of trying to use "prevalence" for a prior? It is, but that doesn't suffice to dismiss the criticism. The critics argue that the quantity we should care about is the FFR, or its complement, the positive predictive value (PPV). Should we? Let's look at all this in detail.

Diagnostic Screening

In scrutinizing a statistical argument, particularly one that has so widely struck a nerve, we attempt the most generous interpretation. Still, if we are not to jumble up our freshly acquired clarity on statistical power, we need to use the proper terms for diagnostic screening, at least one model for it.[8]

We are all plagued by the TSA (Transportation Security Administration) screening in airports, although thankfully they have gotten rid of those whole body scanners in which all "your junk" is revealed to anonymous personnel. The latest test, we are told, would very rarely miss a dangerous item in your carry-on, and rarely trigger the alarm (+) for nothing. Yet most of the alarms are false alarms. That's because the dangerous items are relatively rare. On the other hand, sending positive (+) results for further scrutiny – usually in front of gloved representatives who have pulled you aside as they wave special wands and powder – ensures that, taken together, the false findings are quite rare. On the retest, they will usually discover you'd simply forgotten to remove that knife or box cutter from the last trip. Interestingly, the rarity of dangerous bags – that is, the low prevalence of D's (D for danger) – means we can be comforted in a negative result. So we'd often prefer not to lower the sensitivity, but control false positives relying on the follow-up retest given to any "+" result. (Mayo and Morey 2017.)

Positive Predictive Value (PPV) (1 – FFR). To get the (PPV) we are to apply Bayes' Rule using the given relative frequencies (or prevalences):

$$\text{PPV: } \Pr(\text{D}|+) = \frac{\Pr(+|\text{D})\Pr(\text{D})}{[\Pr(+|\text{D})\Pr(\text{D}) + \Pr(+|{\sim}\text{D})\Pr({\sim}\text{D})]} = \frac{1}{(1+B)}$$

$$B = \frac{\Pr(+|{\sim}\text{D})\Pr({\sim}\text{D})}{\Pr(+|\text{D})\Pr(\text{D})}.$$

The *sensitivity* is the probability that a randomly selected item with D will be identified as "positive" (+):

$$\text{SENS: } \Pr(+|\text{D}).$$

[8] The screening model used here has also been criticized by many even for screening itself. See, for example, Dawid (1976).

The *specificity* is the probability a randomly selected item lacking D will be found negative (−):

SPEC: Pr(−|~D).

The *prevalence* is just the relative frequency of D in some population.

We run the test on the item (be it a person, a piece of luggage, or a hypothesis) and report either + or −. Instead of populations of carry-on bags and luggage, imagine an urn of null hypotheses, 50% of which are true. Randomly selecting a hypothesis, we run a test and output + (statistically significant) or – (non-significant). So our urn represents the proverbial "prior" probability of 50% true nulls.

The criticism turns on the PPV being too low. Even with Pr(D) = 0.5, with Pr(+|~D) = 0.05 and Pr(+|D)= 0.8, we still get a rather high PPV:

$$\mathrm{PPV} = \frac{1}{\left[\frac{1+\Pr(+|\sim \mathrm{D})}{\Pr(+|D)}\right]}.$$

With Pr(D) = 0.5, all we need for a PPV greater than 0.5 is for Pr(+|~D) to be less than Pr(+|D). It suffices that the probability of ringing the alarm when we shouldn't is less than the probability of ringing it when we should. With a prevalence Pr(D) very small, e.g., <Pr(+|~D), we get a PPV < 0.5 even if we assume a maximal sensitivity Pr(+|D) of 1 (Van Belle 2008). In the field of diagnostics, it's scarcely worthless: there is still a boost from the prior prevalence.

Ioannidis rightly points out that many researchers are guilty of cherry picking and selection effects under his "bias" umbrella. The *actual* Pr(+|~D), with bias, is now the probability the "+" was generated by chance plus the probability it was generated by "bias." ~D plays the role of H_0. Even the lowest presumed bias, 0.10, changes a 0.05 into 0.14.

Actual Pr(+|~D):= "alleged" Pr(+| ~D) + Pr(−| ~D)(0.10) = (0.05) + (0.95)(0.10) = 0.14.

The PPV has now gone down to 0.85. Or consider if you're lucky enough to get a TSA official with 30% bias. Your "alleged" Pr(+| ~D) is again 0.05, but with 30% bias, the actual Pr(+|~D) = 0.05 + (0.95)(0.3) = 0.33. Table 5.1 lists some of the top (better) and bottom (worse) entries from Ioannidis' Table, keeping the notation of diagnostic tests. Some of the PPVs, especially for exploratory research with lots of data dredging, get very low PPVs.

Where do his bias adjustments come from? These are just guesses he puts forward. It would be interesting to see if they correlate with some of the better-

Table 5.1 Selected entries from Ioannidis (2005)

Pr(+|D)	PREV of D	Bias	Practical example	PPV
0.8	50%	0.10	Adequately powered RCT, little bias	0.85
0.95	67%	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.8	9%	0.3	Adequately powered exploratory epidemiological study	0.20
0.2	0.1%	0.8	Discovery-oriented exploratory research with massive testing	.001

known error adjustments, as with multiple testing. If so, maybe Ioannidis' bias assignments can be seen as giving another way to adjust error probabilities. The trouble is, the dredging can be so tortured in many cases that we'd be inclined to dismiss the study rather than give it a PPV number. (Perhaps confidence intervals around the PPV estimate should be given?)

Ioannidis will also adjust the prevalence according to the group that your research falls into, leading Goodman and Greenland (2007) to charge him with punishing the epidemiologist twice: by bias and low prevalence! I'm sympathetic with those who protest that rather than assume guilt (or innocence) by association (with a given field), it's better to see what crime was actually committed or avoided by the study at hand. Even bias violations are open to appeal, and may have been gotten around by other means. (No mention is given of failed statistical assumptions, which can quickly turn to mush the reported error probabilities, and preclude the substantive inference that is the actual output of research. Perhaps this could be added.) Others who mount the DS criticism allege that the problem holds even accepting the small α level and no bias.[9] Their gambit is to sufficiently lower the prevalence of D – which now stands for probability of a "true effect" – so that the PPV is low (e.g., Colquhoun 2014). Colquhoun's example retains Pr(+|~D) = 0.05, Pr(+|D)= 0.8, but shrinks the prevalence Pr(D) of true effects down to 10%. That is, 90% of the nulls in your research universe are true. This yields a PPV of 64%. The Pr (~D|+) is 0.36, much greater than Pr(+|~D) = 0.05.

So the DS criticism appears to go through with these computations. What about exporting the terms from significance tests into FFR or PPV assessments? We haven't said anything about treating ~D as H_0 in the DS criticism.

[9] Even without bias, it's expected that only 50% of statistically significant results will replicate as significantly on the next try, but such a probability is to be expected (Senn 2002). Senn regards such probabilities as irrelevant.

[α/(1 – β)] Again

Although we are keen to get away from coarse dichotomies, in the DS model of tests we are to consider just two possibilities: "no effect" and "real effect." The null hypothesis is treated as H_0: 0 effect ($\mu = 0$), while the alternative H_1: the discrepancy against which the test has power (1 − β). It is assumed the probability for finding any effect, regardless of size, is the same (Ioannidis 2005, p. 0696). Then [α/(1 − β)] is used as the likelihood ratio to compute the posterior of either H_0 or H_1 – a problematic move, as we know.

An example of one of their better tests might have H_1: $\mu = \mu^{.9}$ where $\mu^{.9}$ is the alternative against which the test has 0.9 power. But now the denial of the alternative H_1 does not yield the same null hypothesis used to obtain the Type I error probability of 0.05. Instead it would be high, nearly as high as 0.9. Likewise if the null is chosen to have low α, then its denial won't be one against which the test has high power (it will be close to α). Thus, the identification of "effect" and "no effect" with the hypotheses used to compute the Type I error probability and power are inconsistent with one another. The most plausible way to construe the DS argument is to assume the critics have in mind a test between a point null H_0, or a small interval around it, and a non-exhaustive alternative hypothesis $\mu = \mu_1$ against which there is a specified power such as 0.9. It is known that there are intermediate values of μ, but the inference will just compare two.

The DS critics will give a high PPV to alternatives with high power, which is often taken to be 0.8 or 0.9. We know the computation from Goodman (Section 5.2) that "the truth probability of the null hypothesis drops to 3 percent (= 0.03/(1 + 0.03))." The PPV for $\mu^{.9}$ is 0.97. We haven't escaped Senn's points about the nonsensical and the ludicrous, or making mountains out of molehills. To infer $\mu^{.9}$ based on $\alpha = 0.025$ (one-sided) is to be wrong 90% of the time. We'd expect a more significant result 90% of the time were $\mu^{.9}$ correct. I don't want to repeat what we've seen many times. Even using Goodman's "precise *P*-value" yields a high posterior. A DS critic could say: you compute error probabilities but we compute PPV, and our measure is better. So let's take a look at what the computation might mean.

In the typical illustrations it's the prevalence that causes the low PPV. But what is it? Colquhoun (2014) identifies Pr(D) with "the proportion of experiments we do over a lifetime in which there is a real effect" (p. 9). Ioannidis (2005) identifies it with "the number of 'true relationships' . . . among those tested in the field" (p. 0696). What's the relevant *reference class* for the prevalence Pr(D)? We scarcely have a list of all hypotheses to be tested in a field, much less do we know the proportion that are "true." With continuous parameters, it could be claimed there are infinitely many hypotheses;

individuating true ones could be done in multiple ways. Even limiting the considerations to discrete claims (effect/no effect), will quickly land us in quicksand. Classifying by study type makes sense, but any umbrella will house studies from different fields with different proportions of true claims.

One might aver that the PPV calculation is merely a heuristic to show the difference between α and FFR, or between $(1 - \alpha)$ and the PPV. It should always be kept in mind that even when a critic has performed a simulation, it is a simulation that assumes ingredients. If aspects of the calculation fail, then of what value is the heuristic? Furthermore, it is clear that the PPV calculation is intended to assess the results of actual tests. Even if we agreed on a reference class, say the proportion of true effects over your lifetime of testing is θ, this probability θ wouldn't be the probability that a selected effect is "true." It would not be a *frequentist* prior probability for the randomly selected hypothesis. We now turn to this.

Probabilistic Instantiation Fallacy. Suppose we did manage to do an experiment involving a random selection from an urn of null hypotheses, $100\theta\%$ assumed to be true. The outcome may be $X = 1$ or 0 according to whether the hypothesis we've selected is true. Even allowing it's known that the probability of $X = 1$ is 0.5, it does not follow that a specific hypothesis we might choose (say, your blood pressure drug is effective) has a frequentist probability of 0.5 of being true – any more than a particular 0.95 confidence interval estimate has a probability of 0.95. The issue, in this form, often arises in "base rate" criticisms (Mayo 1997a, 1997b, 2005b, 2010c, Spanos 2010b).

Is the PPV computation *relevant* to the very thing that working scientists want to assess: strength of the *evidence* for effects or their degree of corroboration?

Crud Factor. It is supposed in many fields of social and biological science that nearly everything is related to everything: "all nulls are false." Meehl dubbed this the crud factor. Meehl describes how he and David Lykken conducted a study of the crud factor in psychology in 1966. They used a University of Minnesota student questionnaire sent to 57,000 high school seniors, including family facts, attitudes toward school, leisure activities, educational plans, etc. Cross-tabulating variables including parents' occupation, education, siblings, birth order, family attitudes, sex, religious preferences, 22 leisure time activities, MCAT scores, etc., all 105 cross-tabulations were statistically significant at incredibly small levels.

> These relationships are not, I repeat, Type I errors. They are facts about the world, and with $N = 57{,}000$ they are pretty stable. Some are theoretically easy to explain, others more difficult, others completely baffling. The 'easy' ones have multiple explanations, sometimes competing, usually not. Drawing theories from a pot and associating them whimsically

> with variable pairs would yield an impressive batch of H_0-refuting 'confirmations.' (Meehl 1990, p. 206)

He estimates the crud factor correlation at around 0.3 or 0.4.

So let's apply Ioannidis' analysis to two cases. In the first case, we've randomly selected a hypothesis from a social science urn with high crud factor. Even if I searched and cherry picked, perhaps looking for ones that correlate well with a theory I have in mind, statistical significance at the 0.05 level would still result in a fairly high prevalence of true claims (D's) among those found statistically significant. Since the test they passed lacked stringency, I wouldn't be able to demonstrate a genuine reproducible effect – in the manner that is understood in science. So nothing has been demonstrated about replicability or knowledge of real effects by dint of a high PPV.

You might say high prevalence could never happen with things like correlating genes and disease. But how can we count up the hypotheses? Should they include molecular biology, proteomics, stem cells, etc. Do we know what hypotheses will be conjectured next year? Why not combine fields for estimating prevalence? With a little effort, one could claim to have as high a prevalence as desired.

Now let's assume we are in one of those low prevalence situations. If I've done my homework and went beyond the one *P*-value before going into print, checked flaws, tested for violated assumptions, then even if I don't yet know the causal explanation, I may have a fairly good warrant for taking the effect as real. Having obeyed Fisher, I am in a good position to demonstrate the reality of the published finding. *Avoiding bias and premature publication is what's doing the work, not prevalence.*

There is a seductive blurring of rates of false positives over an imagined population, PPVs, on the one hand, with an assessment of what we know about reproducing any particular effect, on the other, and fans of the DS model fall into this equivocal talk. In other words, "positive predictive value," in this context, is a misnomer. The number isn't telling us how valuable the statistically significant result is for predicting the truth or reproducibility of *that effect*. Nor is it even assuring lots of the findings in the group will be reproducible over time. We want to look at how well tested the particular hypothesis of interest is. We might assess the prevalence with which hypotheses pass highly stringent tests, if false. Now look what's happened. We have come full circle to evaluating the severity of tests passed. *Prevalence has nothing to do with it.*

I am reminded of the story of Isaac. Not in the Bible, but in a discussion I had with Erich Lehmann in Princeton (when his wife was working at the Educational Testing Services). It coincided with a criticism by Colin Howson (1997a,b) to the effect that low prevalence (or "base rates") negates severity of test. Isaac is a high school student who has passed (+) a battery of tests for D: "college-readiness." It is given that Pr(+|~D) is 0.05, while Pr(+|D) ~1. But because he was randomly selected from Fewready town, where the prevalence of readiness is only 0.001, Pr(D|+) is still very low. Had Isaac been randomly selected from Manyready suburb with high (Pr(D), then Pr(D|+) is high. In fact Isaac, from Fewready town, would have to score quite a bit higher than if he had come from Manyready suburb for the same PPV. There is a real policy question here that officials disagree on. Should we demand higher test scores from students in Fewready town to ensure overall college-readiness amongst those accepted by college admissions boards? Or would that be a kind of reverse affirmative action?

We might go further and imagine Alex from Manyready scored lower than Isaac, maybe even cheated on just one or two questions. Even if their PPVs are equal, I submit that Isaac is in a better position to demonstrate his college readiness.[10]

The Dangers of the Diagnostic Screening Model for Science

What then can we infer is replicable? Claims that have passed with severity. If subsequent tests corroborate the severity assessment of an initial study, then it is replicated. But severity is not the goal of science. Lots of true but trivial claims are not the goal. Science seeks growth of knowledge and understanding. To take the diagnostic-screening model literally, by contrast, would point the other way: keep safe.

> Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high, so that a significant research finding will lead to a post-test probability that would be considered quite definitive. (Ioannidis 2005, p. 0700)

Who would pursue seminal research that challenged the reigning biological paradigm, as did Prusiner, doggedly pursuing, over decades, the cause of mad cow and related diseases, and the discovery of prions? Would Eddington have gone to all the trouble of testing the deflection effect in Brazil? Newton was predicting fine. Replication is just a small step toward getting real effects. Lacking the knowledge of how to bring about an effect, and how to use it to change other known and checkable effects, your PPV may be swell but your science could be at a dead-end. To be clear: its advocates surely don't recommend the "keep safe" consequence, but addressing it is worthwhile to further emphasize the difference between good science and a good scorecard.

[10] Peter Achinstein and I have debated this on and off for years (Achinstein 2010; Mayo 1997a, 2005b, 2010c).

There are contexts in which the screening viewpoint is useful. Beyond diagnostic screening of disease, high-throughput testing of microarray data seeks to control the rates of genes worth following up. Nevertheless, we argue that the PPV does not quantify how well tested, warranted, or plausible a given scientific hypothesis is (including ones about genetic associations where a DS model is apt). I'm afraid the DS model has introduced confusion into the literature, by mixing up the probability of a Type I error (often called the "false positive rate") with the posterior probability given by the FFR: $\Pr(H_0|H_0$ is rejected). Equivocation is encouraged. In frequentist tests, reducing the Type II error probability results in *increasing* the Type I error probability: there is a trade-off. In the DS model, the trade-off disappears: reducing the Type II error rate also reduces the FFR.

Much of Ioannidis' work is replete with sagacious recommendations for better designs. My aim here was the limited one of analyzing the diagnostic screening model of tests. That it's the basis for popular reforms underscores the need for scrutiny.