

Approximately 1 in 7 Scientific Papers Are Fake

James Heathers¹²

No funding or COI to declare.

jamesheathers@gmail.com

Acknowledgments: I would like to recognise the efforts of the entire forensic metascience community. On this manuscript in particular, assistance from Anna Abalkina, Rene Aquarius, David Bimler, Nick Brown, Rickard Carlsson, Sholto David, Adam Day, Kim Eggleton, Otto Kallikowski, Gideon Meyerowitz-Katz, Kevin Patrick, Matthias Wjst, Nick Wise.

¹ Linnaeus University, Växjö, Sweden

² Basecamp Kitchen Ltd.

Abstract

‘Fake’ science is either intentionally fabricated - where quantitative elements are invented - or intentionally falsified - where results are dishonestly engineered from real data. A frequently cited figure within metascientific research estimates that ~2% of scientists report faking or plagiarizing at least once. In opposition, this paper argues (1) this estimate is contaminated with procedural and social desirability biases, and (2) the proportion of *faking scientists* is a poor frame for understanding failures of research integrity, and is less important than the proportion of *fake scientific output*. N=12 studies can be identified which estimate fake scientific output, and their estimates are variable, but a preliminary approximation is that 1 in 7 published papers have serious errors commensurate with being untrustworthy. This work is too incomplete to support responsible meta-analysis, and research that could more accurately define this figure does not exist yet. ~1 in 7 papers being fake represents an existential threat to the scientific enterprise. This topic demands immediate recognition on the parts of scientists, scientific institutions, and funding bodies.

Keywords: fraud, fabrication, falsification, forensic metascience

How Much Science Is Fake?

Scientists reserve ultimate distaste for *fabrication* (inventing reported data, summary data, or statistical outcomes) and *falsification* (manipulating any part of the research process sufficiently to actively misrepresent real research). Together with plagiarism, these acts form the majoritarian definition of serious scientific misconduct, typically identified by the initials ‘FFP’. While there are systematic treatments of plagiarism (Citron and Ginsparg 2015), this work focuses on *fabrication* and *falsification* (FF) in isolation.

FF is easier to define or investigate if given access to the full data and meta-data of research work. However, published papers rarely supply these, and they are only likely to become accessible within the context of a formal misconduct investigation. The presence of FF is more difficult to define or detect when critically reading research work in the absence of data. Fabrication is more conceptually straightforward – either data is invented or it is not – but may also cross over with data imputation, typographical errors and clerical mistakes, and other forms of negligence or sloppiness (such as the piecemeal cleaning and reconstruction of biological images, which has both benign and nefarious components, or the loss of archival data which makes provenance indeterminable). Falsification is less straightforward – it might be seen as the point where common ‘questionable research practices’ (QRPs) that involve some manipulation of data (such as managing outliers, post-hoc subgroup analysis, outcome switching, promiscuous dichotomisation, p-hacking, etc.) graduate to outright dishonesty. There is no clear dividing line between falsification and QRPs, but rather a substantial gray zone. Any delineation between research misconduct and poor research practice depends on the extent of the manipulation, local norms, historical context, the admixture of errors, etc. A perpetual problem for determining FF is

that the author's intent may be difficult to ascertain even in a formal misconduct investigation where data and experimental material are being examined by skilled investigators for evidence of manipulation. Repeated cases, where data is fabricated over the scale of a career arc, are more definitive.

A canonical figure within the study of fraud and falsification is 2%, derived from the conclusive statement of a systematic review and meta-analysis conducted by Fanelli (2009). The figure is overwhelmingly the salient cited fact in its 1513 citations³ – this generally appears as some variant of “Previous investigations have shown that about 2% of scientists admitted to have fabricated, falsified or modified data or results at least once.” (Frank et al. 2023)

As a comparison, I took a straw poll of colleagues involved in forensic metascience research into the veracity of data within life and social sciences before the below was prepared. While this is highly unsystematic, it produced a substantially higher figure. Correspondents reliably estimated 1-5% *of all papers* contain fabricated data, and 2-10% contain falsified results. Combined, a rate of ‘fakery’ of 3% to 15%. This has a numerical similarity to the Fanelli (2009) estimate – both are low single-digit percentages – but one is an estimate for ‘a minimum of one incident by one researcher over a scientific lifetime’, the other a non-scientific estimate concerning ‘all published papers’. In other words, there is a strong incongruence between self-reported misconduct vs. the estimated level of misconduct observed. This paper attempts to resolve the discrepancy by examining evidence available in the study of research, not of researchers.

³ <https://scite.ai/reports/10.1371/journal.pone.0005738> Accurate as of 9th Sept, 2024

Expanding the conclusion of Fanelli (2009)

Fanelli (2009) is a competent and straightforward synthesis of n=18 individual surveys of misconduct that were available at the time of writing. The questions pertinent to faking science asked within the aggregated surveys are reasonably equivalent (e.g. “Have you, at one or more points during your career, faked a scientific result?” “Have you ever falsified research data?” “Have you engaged in [falsifying or "cooking" research data] during the past three years?” “Was there [fabrication or misrepresentation] in the target publication?”) The study concludes in part “A pooled weighted average of 1.97% (N=7, 95% CI: 0.86–4.45) of scientists admitted to have fabricated, falsified or modified data or results at least once”, which is usually cited as 2%. Even scientists unfamiliar with research integrity or forensic metascience methodology may have seen this figure before, or the typical phraseology used to express it – e.g. “the most serious types of misconduct, fabrication and falsification (i.e., data fraud), are relatively rare” (George 2016). The 2% figure also seems to dominate discourse over more recent, higher figures (see, for instance, Tjldink, Verbeke, and Smulders 2014; Necker 2014). However, the figure is not definitive, even with survey-based methods of assessing FF or FFP prevalence.

Fanelli (2009) also contains a realistic discussion of its limitations. Specifically (1) *social desirability bias* (see, for example, Krumpal 2014; scientists have strong social norms that forbid FFP, and may simply not report it when asked, even anonymously, “self-reports systematically underestimate the real frequency of scientific misconduct”), (2) *format* (“Questionnaires that are handed to, and returned directly by respondents might better entrust anonymity than surveys that need to be mailed or emailed.”), and hence *reliability* (“it is likely that, if on average 2% of

scientists admit to have falsified research at least once ... the actual frequencies of misconduct could be higher than this”; Fanelli, 2009).

A point which could not be raised at the time is the age of the aggregated figures. The cited studies are published from an assessment period from 1987 through 2008, with the date parameters changing by subsample. The 2% figure is derived from studies published from 1992 through 2005, and does not include nearly a human generation’s worth of interaction between scientists and access to digital tools and resources. Likewise, it predates many of the complex, systematic frauds of the digital era. The following (Table 1) is a selection of events which took place after the figure above was established.

DATE	EVENT
May 2005	Adobe Photoshop CS2 introduced Spot Healing and Vanishing Point features
July 2005	SciGen (an online ‘nonsense paper’ generator) has first conference submission platformed at WMSCI 2005
December 2006	First PLoS articles published.
January 2008	NIH open access mandate begins
2008	Beall’s List (a list of untrustworthy journals) started
May 2011	Bem publishes seminal work on precognition (i.e. magic)
September 2011	Diederik Stapel confesses to serial data fabrication
March 2012	John Carlisle reveals 168 fabricated RCTs by Yoshitaka Fujii
October 2011	Simmons, Nelson, and Simonsohn publish seminal work on undisclosed analytical flexibility
March 2013	Declan Butler publishes on ‘hijacked’ journals in Nature
October 2013	John Bohannon submits obviously fake paper to ~300 journals; more than half accept it

November 2013	Mara Hvistendahl publishes a full-length expose of pay-to-play publishing in China in Science Magazine
May 2015	John Bohannon reveals ‘chocolate for weight loss’ hoax
2017	Beall’s List removed
June 2020	GPT-3 API released
March 2023	IJERPH (2nd largest journal by volume) loses Impact Factor
July 2024	Hindawi (now Wiley) publishing retracts ~12000 paper mill papers in a single incident

Table 1: some events relevant to research integrity and digital publication environment (2005-2024)

The above is a whistle-stop tour of stand-out moments in the confluence of science, digital culture, and research integrity – a substantial seachange in the resources, tools, availability, outlets, and culture. Significantly, all of the above happened after the figure of 2% was collected. In particular, much recent FF is driven by developments in auto-generated text, the rise of fabrication-as-a-service businesses (‘papermills’), and the tools necessary to perform sophisticated digital image manipulation. That being said, there are several other past and present estimates for self-reported FFP rates (Table 2) locatable by analyzing citations of Fanelli (2009). These estimates are similar, but also highly variable – as were the inputs to Fanelli (2009), which were dominated by a single large study (Martinson, 2005) which reported a very low FF rate.

The proportion of faking scientists has limited utility

Let us discount the points raised above, and assume this reporting is complete and precise – that every self-reported answer in these aggregated surveys is accurate, and that 2% of researchers participate in FFP at least once. This leaves us with no estimate of how much

STUDY	TYPE	ESTIMATE	TYPE	SAMPLE SIZE
(Xie, Wang, and Kong 2021)	Meta-analysis	2.9% (2.1–3.8%)	FFP	n=42 papers
(Gopalakrishna et al. 2022)	Survey (RR)	4.3% (2.9–5.7%)	Fab.	n=6813
	Survey (RR)	4.2% (2.8–5.6%)	Fals.	n=6813
(List et al. 2012)	Survey (RR)	4.49% (SE=0.30)	Fals.	n=140
	Survey	4.26% (SE=0.22)	Fals.	n=96
(Kaiser et al. 2021)	Survey	0.2%	Fab.	n=7129
	Survey	0.3%	Fals.	n=7127
	Survey	0.5%	P	n=7181
(Agnoli et al. 2017)	Survey (USA)	0.6% (0–1.3%)	Fals.	n=495
	Survey (Italy)	2.3% (0.3–4.2%)	Fals.	n=220

Table 2: similar survey results of self-reported academic misconduct. Fals. = falsification, Fab. = fabrication, FF = both, P = plagiarism, FFP = all of the above, RR = using the 'random response' method of data collection. 95% CI indicated unless stated otherwise.

scientific *output* is fake, or the consequences of this fakery. How many papers is 'one or more'?

Are these 2% of researchers extremely prolific, or do they only produce sporadic or occasional research items? Are these FFP-affected papers invalid in some very small and insignificant part – do they contain single plagiarized sentences or slightly altered numbers, or are they fake in their

entirety? Do they occur earlier in a research career, when researchers are more likely to perform the data collection and analysis themselves? Are they manipulations of data, summary statistics, or interpretation?

There is another way to view the problem – not on a by-author basis, but on a by-paper basis. An analysis designed to address this question ingests papers, analyzes them, and returns the details and nature of anomalies within them, and therefore the likelihood of dishonesty within the entire sample. We can place this work within the growing research tradition of *forensic metascience*. The benefits of this approach are many: (a) identifying the proportion of fake research published is a better *prima facie* answer to the question of ‘how common is dishonesty in scientific publications’; (b) a sufficiently mature analysis of a large enough number of papers also contains estimates for author dishonesty; (c) there are many forensic metascientific approaches to determine hallmarks of accuracy, and problems identified within any specific domain of analysis increases the urgency of its use (for instance, if image manipulation analysis commonly finds problems that data manipulation analysis does not, that approach is a better target for research interest and expansion); (d) the raw material required to perform this analysis is often publicly available; (e) techniques for analysis are additive, and can be grown, extended, revised, or refined; and (f) there are an increasing number of automated and semi-automated tools available to do the work.

The drawbacks, also, are many: (a) it is very challenging to find a combination of papers and analysis techniques that can be automated with a low enough error rate to avoid over-detection (and hence raising undue suspicions about honest authors), thus any given estimate requires a very substantial commitment to manual analysis; (b) all techniques are

domain-specific, and not generalisable – they may only be used to analyze specific features of data, and cannot be used if those features are not present; and (c) as a consequence, they provide estimates of fakery which are themselves very context-dependent.

The following estimates are derived from a combination of personal familiarity, stepping up and down all relevant citation chains, and in consultation with the forensic metascientific community. All relevant studies were included, regardless of analysis technique or research area. The list below is comprehensive but not necessarily exhaustive.

Estimates of scientific fakery

Bik, Casadevall, & Fang (2016)

Bik, Casadevall, and Fang (2016) visually inspected 20,621 papers published within the life sciences from a group of 40 journals. Overall, 3.8% of published papers contained problematic figures, with half of those containing features congruent with deliberative editing of the images. The number of papers showing inappropriate image duplications was approximately 1% from 1995 through 2002, then rose quickly to 4%, a figure that was maintained from 2005 through 2014 (as this is the most contemporary figure offered, and was consistent over the final decade of analysis, that figure is used here). Five journals featured image duplication rates over 8% total. As the cohort of data available for analysis finishes in 2014, the last ten years of scientific output are not analyzed. However, over this period, the rate of retracted papers has increased by an approximate order of magnitude (i.e. from ~1000 in 2014, to ~10000 in 2023)⁴. This was (and will likely remain) the largest analysis of its kind.

⁴ <https://www.nature.com/articles/d41586-023-03974-8>

Berrío & Kalliokoski (2024)

Berrío & Kalliokoski (2024) drew a sample of 1,035 studies from the literature on preclinical studies of depression, specifically those describing animal models of chronic stress. n=476 had no analyzable content, and n=588 were amenable to image analysis – of these, n=112 showed anomalies ranging between what were potentially clerical errors to clear hallmarks of fabrication. A reasonable estimate of those which were manipulated is any containing a Class II or Class III error (see Bik, Casadevall, & Fang, 2016), n=49 and n=33 respectively. This estimates an FF rate of 13.9%. This is the most recent exhaustive effort to assign such a figure to a large body of scientific literature.

Further image manipulation work

After the publication of Oksvold (2016) and Bik, Casadevall and Fang (2016), several similar papers in the same tradition were published – all analyze a corpus of papers in the life sciences, specifically check for hallmarks of image manipulation, and use the same system of categorization. They are typically defined by journal or research area, and use a combination of automated and manual detection methods. These are summarized below in Table 3. Where necessary, I have used the same approximation as above (ie. Class 2 and 3 errors are classified as hallmarks of manipulation, Class 1 errors are classified as mistakes).

STUDY	AREA	ESTIMATE	TYPE	SAMPLE
(Oksvold 2016)	Field of oncology	24.2%	Manual	n=120
(Bucci 2018)	Random selection (from PMC)	5.7%	Automated	n=1364
(Bik et al. 2018)	<i>Molecular and Cellular Biology</i>	6.1%	Manual	n=960
		14.5%	Automated	n=83
(Wjst 2021)	<i>American Journal of Respiratory Cell and Molecular Biology</i>	16.2%	Automated + manual	n=37
(David 2023)	<i>Toxicology Reports</i>	10.3%	Manual	n=715
		16.1%	Automated + manual	n=715
(Cho et al. 2024)	Field of rhinology	26.8%	Automated	n=67
		13.4%	Automated + manual	n=67

Table 3: Aggregated FF estimates from image manipulation analysis.

Brown and Heathers (2016)

Brown and Heathers (2016) describes our first published forensic metascientific test; GRIM is a numerical technique designed to evaluate if reported means of granular data are possible given their sample size. We retrieved 260 papers within the social sciences, and n=71 were amenable for GRIM testing (the technique typically only applies to samples or subsamples with n<100). Of these testable articles, half (n=36) contained at least one inconsistent mean, which was not treated as a hallmark of malfeasance, and one in five (n=16) contained multiple inconsistent means, which we deemed ‘substantial’. On requesting the data for some of these, we found a variety of clerical errors which were easily corrected, and one request was based on our

misunderstanding. But one unpublished result which was not included in the initial pre-print and subsequent manuscript is that twelve (12) manuscripts contained both multiple inconsistencies, and the authors refused and/or ignored a request for data, within which three (3) manuscripts contained what we considered definite hallmarks of systematic manipulation. These figures were both sufficiently speculative and controversial at the time of publication to lead us to redact them from the manuscript. However, this puts the percentage of manuscripts with the hallmarks of data manipulation between 3/71 and 12/71, ie. between 4.2% and 16.9%.

Miyakawa (2020)

Miyakawa (2020) describes the author's experience as the editor-in-chief of the journal *Molecular Brain*. Over approximately 3 years, Miyakawa reviewed 181 manuscripts, and for any manuscript that felt 'too beautiful to be true' (n=41), he requested the raw data. Specifically, this was "all the images for entire membranes of western blotting with size markers and for staining, quantified numerical data for each sample used for statistical analyses, etc.)" as well as exact p-values (presumably with the intent of inspecting any that are STALT values; see Heathers and Meyerowitz-Katz (2024)) and any update to corrections for multiple comparisons if necessary. Of those 41 manuscripts, 20 were withdrawn from publication without providing data, 19 were resubmitted with data which was deemed insufficient and rejected, and 1 was published. Of the 40 withdrawn or rejected manuscripts, Miyakawa estimates 26 manuscripts contain fabricated elements. This produces an estimated FF rate of 14.4%.

Carlisle (2021)

Carlisle (2021) analyzed the baseline summary data of RCTs submitted to the journal *Anaesthesia* for ~3 years (02/2017 through 03/2020). The paper deploys a wide variety of forensic metascientific techniques, some of which are identical to traditional forensic accounting techniques, including (a) data re-use from previous publications, (b) incorrectly calculated p-values (c) unlikely omnibus p-values, (d) the GRIM method (see above), (e) trailing digit analysis, (f) strong unexplained randomization failures, (g) unusual deviation from published trial protocols, and more.

Working with both summary statistics and individual patient-level data (which was required by the journal post-2019), the paper concludes 73 out of 526 trials contained false data (13.9%). The ability to analyze patient-level data was extremely strongly associated with the ability to detect false data (OR=10.2, 95% CI 5.3–21.6, $p=2e-16$), raising the detection rate from ~4% to ~29% of submitted trials.

The COPE / STM report on paper mills

‘Paper mill’ papers are fabricated papers prepared by a commercial service that sells them as a service to dishonest researchers. The operation of paper mills has increased significantly in the last 5 years in particular, and paper mill products – typically poorly fabricated work with features such as nonsensical language, meaningless mathematical explanations, inappropriate citations, and other easily detectable features – are increasingly found both before and after publication. A document titled “Paper Mills: Research report from COPE & STM”⁵

⁵ <https://publicationethics.org/files/paper-mills-cope-stm-research-report.pdf>

was published on publicationethics.org in 2022, and does not have identifiable authors. Over 53,000 pre-publication manuscripts from six publishers were analyzed via methods not fully outlined, but presumably including tools similar to the Problematic Paper Screener (Cabanac, Labbé, and Magazinov 2022). As the corpus for analysis is pre-publication manuscripts, the estimate provided is of problems detected before they had a chance to contaminate the formal scientific literature. However, these are also an expression of what that literature will eventually become, as most rejected papers are eventually published, just elsewhere. The percentage of what the authors deem ‘suspect papers’ analyzed before publication ranged from 2% to 46% by journal, and the document describes a right-tailed distribution of paper mill output (as when a journal proves to have inadequate safeguards to prevent paper mill publication, this invites an increased number of submissions). The average percentage of affected articles in each journal analyzed between 2019 and 2021 was 14%.

Summary

These values are too disparate to meta-analyze responsibly, and support only the briefest form of numerical summary: $n=12$ papers return $n=16$ individual estimates; these have a median of 13.95%, and 9 out of 16 of these estimates are between 13.4% and 16.9%. Given this, a rough approximation is that for any given corpus of papers, 1 in 7 (i.e. 14.3%) contain errors consistent with faking in at least one identifiable element.

Discussion

The figure of 1/7 is probably higher than many expect. One community of people who are not surprised are data sleuths. The estimation that sparked this document was repeated at a later date (Nick Brown, pers.comm), where n=29 data sleuths were asked "What percentage of papers do you think are fake?" without disambiguating the word 'fake'. Perhaps unsurprisingly, as some of these participants were authors in the cited literature above, the median was 15% (the mean was higher – 23.6% (SD=23.7) – driven by some very high estimates). Likewise, those with formal research integrity roles are likely unsurprised – in a normative month, IOP Publishing immediately rejects 7% of submitted manuscripts for ethical issues before review (Kim Eggleton, pers.comm).

For other scientific communities, the question remains of how to reconcile this evidence with the estimate that 2% of scientists self-report their FFP at least once. Even if the present estimate is wildly inflated and we use the single most conservative figure available (i.e. 4%), this is certainly more individual papers than would be supported by '2% of scientists commit FFP *once*'. As expected, as the social desirability of FFP is extremely low, it is likely under-reported. Presumably, it is somewhat psychologically naive to expect dishonest people to honestly report their dishonesty in an environment that cherishes honesty.

However, *should* we reconcile the evidence at this point? The accumulation of papers collected here is, frankly, haphazard. It does not represent a mature body of literature. The papers use different methods of analyzing figures, data, or other features of scientific publications. They do not distinguish well between papers that have small problematic elements which are fake, or fake in their entirety. They analyze both small and large corpora of papers, which are in different

areas of study and in journals of different scientific quality – and this greatly changes base rates; for instance, a recent incident saw the publisher Hindawi (now Wiley) retract ~12,000 papers in a single incident, which is 667x the all-time number of retractions from Nature Publishing Group⁶. They analyze both recent and past publications, and pre- and post-publication manuscripts. They report automated analysis as detecting both more and less manipulation than manual analysis. They are generally focused on specific paper types, with specific problems, within specific research areas of the life and biomedical sciences. And while they return empirical estimates on the trustworthiness of ~70000 individual papers, they are not free of judgment or subjectivity, as there is often a lack of clarity on the question of whether or not paper authors made an inadvertent mistake or committed a malfeasance.

Finally, as this is a controversial area, it is likely there are more estimates that were never published. At least one (Wjst, pers.comm) conducted a retrospective 20-year analysis of papers using a combination of manual and automated tools, and found the presence of image anomalies in around 15%. As a consequence, it would be prudent to immediately reproduce the result presented here as a formal systematic review. It is possible further figures are available after an exhaustive search, and also that pre-registered analytical assumptions would modify the estimations presented.

However, if these figures are in any way accurate, then they constitute the single biggest unsolved problem within modern science, particularly because the above figures represent *lower bounds*. The strong majority of FF estimates included here are subsequent analyzing the fairly obvious hallmarks of manipulation capable of being detected *without* access to study materials

⁶ <http://retractiondatabase.org/RetractionSearch.aspx#?pub%3dNature%2bPublishing%2bGroup>

(ethical application, study materials, reagents, raw data, etc.) – if those additional details were available, the presumptive rates of FF would be higher (e.g. Carlisle, 2021). These details are sometimes available, and have in the past led to the identification of specific features of manipulation, especially at the data level. The false positive rate (FPR) of detecting fake science is almost certainly quite low, as data which are persistently impossible are unlikely to be honest mistakes, so are pixel-identical or deceptively edited images. However, the false negative rate (FNR) is unknown, but it is very likely higher than the FPR as all of the above methods are best alerted to obvious and inexpert fraud – a skilled faker could almost certainly produce less obviously problematic research, and may be able to evade detection entirely under any level of scrutiny. In short, we can say with confidence the $FNR > FPR$, and that the true figures are higher than those listed. Likewise, if this is the rate of *fake* papers, then the presumably higher number of papers containing questionable research practices (which are far more commonly admitted to) is presumably higher still.

But even within isolation, a 1/7 FF rate is essentially a slow-moving local polycrisis. False results waste other scientist's time and money if they are ever chosen for replication or extension. In doing so, they stymie careers and needlessly spend public money, they discourage researchers from continuing their careers, and students from beginning them. They delay pharmacological, surgical, and behavioral treatment of illness. They contaminate meta-analyses, and in doing so, affect the direction of entire fields, or, of more intermediate concern, hurt or kill people if they affect meta-analyses that determine treatment guidelines. They destroy the internal fabric of trust that science relies on, and force the adoption of slower and more substantive open scientific methods. Publicly, they reduce the public profile of science, and threaten the entire

scientific enterprise with a loss of public trust and support. Moreover, they are self-perpetuating – fake science is faster, cheaper, and easier than real science, and if the two traditions compete to see who can produce more results (or produce the same results first), then fake science can quickly engender fake norms.

However, at a university or governmental level, the global financial support for directly detecting, combatting, and publicizing this problem is effectively zero. There are no formal federal or global grant schemes that are available to specifically investigate fake research, and I am not aware of any faculty position anywhere in the world that specifies a research line in scientific error mitigation. There are no dedicated academic journals which publish results, techniques or technological developments in forensic metascience. University Research Integrity Officers frequently complain about the legislation which compels them to investigate ever-increasing numbers of anomalous papers while their roles also include other activities in research integrity, such as training and teaching – essentially, they are hugely under-resourced. The US Office of Research Integrity has a FY 2023 budget of around \$12M, about half the cost of a single Phase 3 drug RCT, and YTD (Sept, 2024) has completed and released 4 misconduct investigations. In contrast, the NIH has a yearly budget of \$47.7B⁷.

However, as stochastic as the estimate here may be, it warrants conducting large-scale investigations into FF, using formal and structured assessment methods that allow us to achieve better formal estimates of the problem. In particular, it seems likely that FF rates change by individual field – in doing so, they may present specific rather than general threats to human health and scientific progress.

⁷ <https://www.nih.gov/about-nih/what-we-do/budget> Figure from 2023.

In conclusion, there is a colossal mismatch between the resources available to investigate and mitigate this problem, and the problem itself. The collective unwillingness to recognize this problem has grown to the point of outrageous wilful ignorance. Priorities must change, or science will start to die.

References

- Agnoli, Franca, Jelte M. Wicherts, Coosje L. S. Veldkamp, Paolo Albiero, and Roberto Cubelli. 2017. "Questionable Research Practices among Italian Research Psychologists." *PloS One* 12 (3): e0172792.
- Bik, Elisabeth M., Arturo Casadevall, and Ferric C. Fang. 2016. "The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications." *mBio* 7 (3). <https://doi.org/10.1128/mBio.00809-16>.
- Bik, Elisabeth M., Ferric C. Fang, Amy L. Kullas, Roger J. Davis, and Arturo Casadevall. 2018. "Analysis and Correction of Inappropriate Image Duplication: The Experience." *Molecular and Cellular Biology* 38 (20). <https://doi.org/10.1128/MCB.00309-18>.
- Bucci, Enrico M. 2018. "Automatic Detection of Image Manipulations in the Biomedical Literature." *Cell Death & Disease* 9 (3): 400.
- Cabanac, Guillaume, Cyril Labbé, and Alexander Magazinov. 2022. "The 'Problematic Paper Screener' Automatically Selects Suspect Publications for Post-Publication (re)assessment." <https://doi.org/10.48550/ARXIV.2210.04895>.
- Carlisle, J. B. 2021. "False Individual Patient Data and Zombie Randomised Controlled Trials Submitted to Anaesthesia." *Anaesthesia* 76 (4): 472–79.
- Cho, Do-Yeon, Jessica Bishop, Jessica Grayson, and Bradford A. Woodworth. 2024. "Inappropriate Image Duplications in Rhinology Research Publications." *International Forum of Allergy & Rhinology* 14 (1): 119–22.
- Citron, Daniel T., and Paul Ginsparg. 2015. "Patterns of Text Reuse in a Scientific Corpus." *Proceedings of the National Academy of Sciences of the United States of America* 112 (1): 25–30.
- David, Sholto. 2023. "A Quantitative Study of Inappropriate Image Duplication in the Journal *Toxicology Reports*." *bioRxiv*. <https://doi.org/10.1101/2023.09.03.556099>.
- Fanelli, Daniele. 2009. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." *PloS One* 4 (5): e5738.
- Frank, Fabrice, Nans Florens, Gideon Meyerowitz-Katz, Jérôme Barriere, Éric Billy, Véronique Saada, Alexander Samuel, Jacques Robert, and Lonni Besançon. 2023. "Raising Concerns on Questionable Ethics Approvals - a Case Study of 456 Trials from the Institut Hospitalo-Universitaire Méditerranée Infection." *Research Integrity and Peer Review* 8 (1): 9.
- George, Stephen L. 2016. "Research Misconduct and Data Fraud in Clinical Trials: Prevalence and Causal Factors." *International Journal of Clinical Oncology*. <https://doi.org/10.1007/s10147-015-0887-3>.
- Gopalakrishna, Gowri, Gerben Ter Riet, Gerko Vink, Ineke Stoop, Jelte M. Wicherts, and Lex M. Bouter. 2022. "Prevalence of Questionable Research Practices, Research Misconduct and Their Potential Explanatory Factors: A Survey among Academic Researchers in The Netherlands." *PloS One* 17 (2): e0263023.
- Heathers, James and Meyerowitz-Katz, Gideon. 2024. "'Yes, but how much smaller?' A simple observation about p-values in academic error detection". *OSF* <https://doi.org/10.17605/OSF.IO/2SP5B>
- Kaiser, Matthias, Laura Drivdal, Johs Hjellbrekke, Helene Ingierd, and Ole Bjørn Rekdal. 2021. "Questionable Research Practices and Misconduct Among Norwegian Researchers." *Science and Engineering Ethics* 28 (1): 2.

- Krumpal, Ivar. 2014. "Social Desirability Bias and Context in Sensitive Surveys." *Encyclopedia of Quality of Life and Well-Being Research*. https://doi.org/10.1007/978-94-007-0753-5_4086.
- List, John A., Charles D. Bailey, Patricia J. Euzent, and Thomas L. Martin. 2012. *Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior*.
- Miyakawa, Tsuyoshi. 2020. "No Raw Data, No Science: Another Possible Source of the Reproducibility Crisis." *Molecular Brain* 13 (1): 24.
- Necker, Sarah. 2014. "Scientific Misbehavior in Economics." *Research Policy* 43 (10): 1747–59.
- Oksvold, Morten P. 2016. "Incidence of Data Duplications in a Randomly Selected Pool of Life Science Publications." *Science and Engineering Ethics* 22 (2): 487–96.
- Tijdink, Joeri K., Reinout Verbeke, and Yvo M. Smulders. 2014. "Publication Pressure and Scientific Misconduct in Medical Scientists." *Journal of Empirical Research on Human Research Ethics: JERHRE* 9 (5): 64–71.
- Wjst, Matthias. 2021. "Scientific Integrity Is Threatened by Image Duplications." *American Journal of Respiratory Cell and Molecular Biology* 64 (2): 271–72.
- Xie, Yu, Kai Wang, and Yan Kong. 2021. "Prevalence of Research Misconduct and Questionable Research Practices: A Systematic Review and Meta-Analysis." *Science and Engineering Ethics* 27 (4): 41.