



// NO.26-048 | 10/2026

DISCUSSION PAPER

// KATRIN HUSSINGER AND MAIKEL PELLENS

Quantifying Undetected Severe Misconduct in Medical Research

Quantifying Undetected Severe Misconduct in Medical Research

Katrin Hussinger^{a,b,d*} and Maikel Pellens^{c,b,d}

October 2026

This article is forthcoming in *Research Policy*.

Abstract: Scientific misconduct in the medical and health sciences remains a major concern for practitioners, patients, scientists, and funders. Even as retractions of fraudulent articles have increased sharply in recent years, concerns remain that detected cases represent only the tip of the iceberg and that actual prevalence of misconduct in the scientific system may be much higher. In this study, we estimate the size of the hidden population of scientists associated with articles that have been retracted for severe misconduct by applying capture-recapture methods. Distinguishing first authors and last authors associated with retracted articles, we find a capture rate of approximately 23% and a population of around 4,250 first authors associated with articles that have been retracted due to severe misconduct in the medical and health sciences. Among last authors, we find a capture rate of 34% and a population of around 4,125 scientists. Our results imply that most fraud in the medical and health sciences remains undetected, yet suggest that only a small fraction of the scientific workforce is associated with articles affected by severe scientific misconduct.

Keywords: scientific misconduct; dark number estimation; capture-recapture methods; medical sciences

JEL: I18, O33

a University of Luxembourg (katrin.hussinger@uni.lu)

b Centre for European Economic Research (ZEW), Mannheim, Germany

c K.U. Leuven, Centre for R&D Monitoring (ECOOM), Belgium (maikel.pellens@kuleuven.be)

d K.U. Leuven, Dept. of Managerial Economics, Strategy and Innovation, Belgium

* Corresponding author

1. Introduction

Detected cases of scientific misconduct, as measured by retractions of fraudulent scientific articles, have risen sharply over the past decade. In 2022 alone, some 2,600 articles were retracted in biomedical sciences, more than twice the number in 2018 (“There Is a Worrying Amount of Fraud in Medical Research”, 2023). These cases, however, may represent the tip of the iceberg (Walby, 1990) because the true scale of scientific misconduct is inherently difficult to detect. Many instances of fraud or questionable research practices likely go unnoticed, either because they are never reported, hidden by the individuals involved, or not identified by peer review and editorial processes. As a result, the cases that come to light most likely are only a small fraction of the total misconduct occurring in the scientific community. The hidden nature of scientific misconduct implies that, ultimately, the size of the population of misconducting authors, on which our analysis focuses, remains elusive.

Given the costs that fraudulent science, particularly in the medical and health sciences, imposes on society, it is essential to understand the extent of undetected misconduct. The societal costs of scientific misconduct in the medical and health sciences include direct harm to patients from erroneous treatments (Berggren and Karabag, 2019) and disrupted regulatory oversight (Glenna and Bruce, 2021). More broadly, scientific misconduct erodes trust in science (Hussinger and Pellens, 2019a; Rodrigues and Pizzolato, 2024) and, as a consequence, slows scientific progress in “tainted” fields (Azoulay et al., 2015).

Misconduct also imposes economic costs by funding false science and directing research efforts to build on it (Feng et al., 2020; Stern et al., 2014). When fraudulent or unreliable findings are published, other researchers may build upon false results, wasting time and other resources. In addition, detecting misconduct has a direct cost (Michalek et al., 2010).

Investigating suspected cases of fraud requires substantial resources, including the time and effort of journal editors, peer reviewers, institutional committees, and, at times, legal experts.

These investigations can be lengthy and expensive, diverting funding and attention away from productive research activities. For example, Michalek et al. (2010) estimate that the costs of handling a single case of misconduct, including inquiry, investigation, and subsequent administrative actions, can amount to hundreds of thousands of dollars.

Scientific misconduct also harms scientists' careers. When scientists retract papers because of misconduct, trust in their work declines, which can trigger a "retraction penalty" that reduces publications and citations to prior work (Azoulay et al., 2017; Lu et al., 2013). This penalty extends to co-authors (Mongeon and Larivière, 2016) and prior collaborators (Hussinger and Pellens, 2019a). These impacts are most severe for early-stage scientists (Jin et al., 2019; Memon et al., 2025).

Given these costs, it is important to know the extent of scientific misconduct. Yet estimating it poses a methodological challenge: misconduct is, by definition, concealed, and traditional measurement approaches can capture only part of the phenomenon. In this study, we contribute to understanding the scale of scientific misconduct in the medical and health sciences. We complement existing analyses of misconduct based on retraction databases, survey-based studies (e.g., Fanelli, 2009), and statistical detection methods or statistical forensics (e.g., Bordewijk et al., 2021; Fang et al., 2012), by addressing the "dark figure" problem, that is, the fact that traditional methods cannot estimate misconduct that remains undetected.

We focus only on severe intentional misconduct, defined as the fabrication, falsification, manipulation, or duplication of data and images, rather than on honest mistakes, for both conceptual and methodological reasons. Conceptually, fraud represents intentional deception and is empirically the dominant cause of retractions in major scientific fields (Fang et al.,

2012), whereas honest error follows different correction dynamics.¹ Methodologically, fraud is also particularly prone to under-detection due to stigma, legal risk, and institutional incentives, making it especially suitable for unseen-event estimation (Fang and Casadevall, 2015). Restricting attention to fraud reduces heterogeneity in detection mechanisms, improves interpretability, and aligns the analysis with research-integrity policies that are explicitly designed to deter and sanction intentional misconduct rather than error. Importantly, because policy responses to fraud differ from those addressing error, estimating the magnitude of fraud specifically is more relevant for research integrity governance.

We estimate the size of the hidden population of authors associated with articles affected by severe misconduct in the medical and health sciences using capture-recapture methods (Hook and Regal, 1995). These methods, developed primarily in ecology and epidemiology, infer the size of an otherwise hidden population from repeated observations of its members. In ecology and epidemiology, they are typically used to answer research questions about species richness and biodiversity (i.e., how many individuals of a specific species are present but unobserved) (Hook and Regal, 1995), while typical applications in the social sciences include estimating the size of the homeless population (e.g., Fisher et al., 1994), the population of people who inject drugs (e.g., Mastro et al., 1994), the population of clients of sex workers (Roberts and Brewer, 2006), and the population of criminal offenders (Frank and Carrington, 2007). These questions share the goal of estimating the size of a human population of policy interest that is difficult to measure using conventional methods.

We use capture-recapture methods to estimate the incidence of scientific misconduct. Like all capture-recapture approaches, our design relies on assumptions about independence between

¹ In a detailed analysis of 2,047 retracted biomedical articles, Fang et al. (2012) found that 67.4% of retractions were attributable to misconduct, with fraud and suspected fraud accounting for the largest share and only about one-fifth due to error.

detections and a relatively closed population. In our setting, violations of these assumptions are likely to bias estimates downward, so our results should be interpreted as conservative lower bounds on the true prevalence of scientific misconduct. We use the total number of scientific publications retracted for fraudulent behavior by medical scientists to estimate the total population of fraudulent medical scientists. Our study is based on data on retractions in the medical and health sciences, retrieved from Retraction Watch, a comprehensive database that systematically tracks retractions across scientific disciplines, documents the reasons for retraction, and provides detailed data for each case.² We estimate the population of scientists associated with articles affected by severe misconduct using either the first or last author of the affected article as the focal unit. For first authors, our results imply a capture rate of approximately 23% (973 observed misconducting first authors divided by an estimated population of 4,235 misconducting scientists). For last authors, the capture rate is 34% (1,393 observed last authors compared to an estimated population of 4,123). Both results indicate that most fraud cases in the medical and health sciences remain undetected. At the same time, the estimated population remains a tiny fraction of medical scientists.

With this, we complement previous attempts to quantify the incidence of scientific misconduct, such as those based on problematic images and figures (e.g., Berrío and Kalliokoski, 2025; Bik et al., 2016), statistical forensics (e.g., Bordewijk et al., 2021), surveys (e.g., Fanelli, 2009), or retraction counts (Fang et al., 2012).³ These studies highlight that scientific misconduct, often hidden, may not be marginal and may affect a substantial proportion of the scientific literature. Our study adds to this literature by addressing a specific

² <https://retractionwatch.com/>

³ Berrío and Kalliokoski (2025) find that 1 in 4 or 5 images is problematic, and Bik (2016) reports that 3.8% of screened papers contain problematic figures. Based on a meta-analysis of surveys that attempt to quantify scientific misconduct, Fanelli (2009) concludes that about 2% of reporting scientists commit severe misconduct, while up to 34% admit to questionable research practices. At the same time, Smith (2021) notes that only 0.04% of medical research articles have been retracted.

dimension of the “dark figure” of scientific misconduct. Rather than treating observed authors of retractions due to severe fraud as a direct measure of prevalence, we use repeated observations of authors in retraction databases to estimate the broader population of scientists associated with articles containing severe misconduct.

2. Conceptual basis and alternative approaches

2.1. Conceptual basis

We conceptualize the population of scientists associated with articles prone to retraction for severe misconduct as a hidden population generated by purposeful decisions under incentive constraints (Lacetera and Zirulia, 2011). Researchers choose between honest research and misconduct by weighing the expected career and resource benefits of producing honest and novel results against the expected costs, including detection risk, formal sanctions, reputational damage, and internal moral or professional norms (Fanelli, 2009; Lacetera and Zirulia, 2011). Because only a subset of misconduct is ever detected and formally recorded, observed cases form a non-random sample of the true underlying population (Furmann et al., 2012; Böhning and van der Heijden, 2009). The size of the hidden population therefore depends on two coupled processes: (i) the incidence of misconduct, shaped by incentive structures and competitive pressures, and (ii) the probability of detection, which determines whether misconduct becomes observable. We cannot disentangle these processes because only their combined outcome is observable. Notably, exposure through scientific publications increases the probability of detection.⁴

Our analysis takes a reduced-form perspective, treating author heterogeneity and institutional processes as latent and absorbed into the capture probability. The estimated capture rate

⁴ We address this in our empirical analysis by accounting for the publication stock of the scientists in our covariate-adjusted specifications.

summarizes the combined effects of heterogeneous author visibility, attribution of responsibility through authorship positions, and institutional thresholds for initiating formal investigations (Azoulay et al., 2015). This approach lets us quantify how much which misconduct remains unobserved under prevailing authorship and governance structures. By focusing on the size of the hidden population rather than on covariate effects, the analysis provides a system-level measure of underdetection that is robust to incomplete information about individual environments and incentives and directly relevant for assessing the adequacy of current research integrity regimes.

2.2. Overview of alternative approaches

Our study is situated among several existing approaches for measuring the incidence of scientific misconduct. These can be broadly grouped into analyses of retraction databases, survey-based studies, and statistical detection methods. While all three approaches have merits, they share a limitation: they suffer from the “dark figure” problem and cannot estimate misconduct that remains undetected. We complement these existing approaches by estimating this number.

A primary method for estimating misconduct prevalence is self-reporting through surveys. Fanelli (2009) provides the first meta-analysis of such surveys and concludes, based on pooled weighted averages, that 1.97% of scientists admit to severe research misconduct, i.e., fabrication, falsification, or modification of research data or results. In addition, 33.7% of respondents admit to other questionable research practices. Studies that asked scientists to report colleagues’ behavior found rates of 14.2% for severe misconduct and 72% for questionable research practices. An updated analysis by Xie et al. (2021) estimates severe misconduct at 2.9% and questionable research practices at 12.5%. For observations of colleagues, the estimates are 15.5% and 39.7%, respectively. A recent survey of academic researchers in the Netherlands (Gopalakrishna et al., 2022) reports fabrication at 4.3% and

falsification at 4.2% across all disciplines and academic ranks, with 51% engaging frequently in one or more questionable research practices.

However, survey-based estimates of misconduct incidence tend to underestimate it because of measurement limitations and social desirability bias (Fanelli, 2009; Gopalakrishna et al., 2022; John et al., 2012; Xie et al., 2021). Moreover, Fanelli (2009) argues that it is difficult to assess whether surveys of colleagues' behavior yield better estimates, as respondents might still, in their own interests, under- or over-report misconduct. It is also unclear how a single incident might be reported by many respondents. At the same time, some studies include specialized techniques to address social desirability bias. For example, Gopalakrishna et al. (2022) make use of the randomized response technique for questions on fabrication and falsification. Roberts and St. John (2014) compare results using unmatched count techniques, the crosswise model, and direct questioning across types of misconduct, and find that direct questioning systematically underestimates misconduct.⁵

Another approach is to analyze retraction databases (Hesselmann et al., 2017). These databases provide objective and verifiable evidence of misconduct but capture only visible cases. Fang et al. (2012) examine all retracted biomedical and life-science studies in PubMed up to May 2012, and find that 67% of retractions reflect misconduct and that the proportion of retractions due to fraud has grown tenfold since 1975. While this is informative, estimates of misconduct incidence based on observed retractions have limitations. For instance, retraction rates are influenced by retraction policies, which vary across journals and publishers and

⁵ Randomized response techniques ask respondents to randomize their choice between two questions, without revealing the choice to the researcher. This encourages truthful answers by destroying the link between the respondent and the individual response (Warner, 1965). Unmatched count techniques randomly assign respondents into two groups. The control group receives a list of non-sensitive statements and reports how many apply. The treatment group reports the same, with one sensitive item added to the list. The prevalence of the sensitive behavior can then be estimated as the difference in means between the two groups, without identifying individuals (Raghavarao and Federer, 1979). The crosswise model asks a sensitive question and a question with a known prevalence, and asks only whether the two answers match or differ (Yu et al., 2008).

evolve over time, as well as by the detection of fraudulent science. Therefore, higher retraction rates may reflect increasing capacity for detection and stronger policies rather than more fraud (Fanelli, 2013), and these studies' results are best understood as lower bounds (Stricker and Günther, 2019).

A third family of methods uses statistical or computational tools to analyze published data for anomalies indicative of fabrication or falsification (Bordewijk et al., 2021). Examples include detecting and studying problematic scientific images or figures (Berrío and Kalliokoski, 2025; Bik et al., 2016), examining the distribution of digits against expected distributional properties (Diekmann, 2007; Mosimann et al., 2002), studying mathematically improbable means and response distributions (Brown and Heathers, 2017) or distributions of treatment and control characteristics (Carlisle, 2012, 2017), and examining studies for signs of suspiciously high levels of consistency (Simonsohn, 2013). While useful for detecting scientific misconduct, these methods can only examine anomalies in published data and are best suited to flagging potential misconduct, which must be confirmed through further investigation.

Each of these methods is affected by the “dark figure” problem in detecting undesirable behaviors, the gap between observed cases and the true total. However, they capture different aspects of research misconduct. For instance, survey-based studies, such as Fanelli (2009) and Xie et al. (2021), measure self-reported or colleagues' misconduct, which may include isolated acts, while retraction databases identify articles retracted for misconduct without necessarily establishing individual responsibility. Our study addresses one dimension of this dark figure by using observed cases of severe misconduct that are captured in retraction databases to estimate the broader population of first and last authors associated with articles containing severe misconduct, including those whose associated articles remain undetected and unretracted. Consequently, our estimates should not be interpreted as measures of

individual misconduct prevalence or as directly comparable to survey-based prevalence rates. Rather, these approaches provide complementary evidence on research misconduct, each subject to distinct measurement limitations.

3. Methods and Data

3.1. Methods

We rely on one-source capture-recapture estimators for incomplete population estimation. These estimate hidden populations using only the observed count distribution, without requiring strong assumptions about truthful reporting, replication validity, or classification algorithms. Originating in ecology and later adapted and widely developed in epidemiology, these methods routinely estimate the size of populations that cannot be fully observed, such as wildlife populations or individuals affected by underreported diseases (Seber, 1982; Hook and Regal, 1995). The core insight is that repeated, imperfect observations of the same underlying population can be used to infer the number of unobserved units, even when detection probabilities are heterogeneous.

Researchers have also used capture-recapture methods in the social and medical sciences (Böhning et al., 2018). Recent applications in the medical sciences include estimating total hospital admissions among people experiencing homelessness (Luchenski et al., 2025) and estimating the size of hidden populations of individuals infected with COVID-19 in Thailand that were not captured by official surveillance systems (Böhning et al., 2023).

The application of this framework to scientific misconduct is conceptually analogous: confirmed cases represent only a subset of all misconduct events, detected through selective and imperfect institutional processes (Bohannon, 2013). As in ecological and epidemiological settings, observed cases cannot be assumed to represent the full population, making capture-recapture methods well suited to estimating the size of its hidden component. The capture-

recapture approach allows us to move beyond observed counts and formally quantify under-detection in research integrity systems using well-established statistical tools.

More specifically, we rely on Chao’s (1987) estimator, the generalized Chao estimator (Böhning, 2010; Böhning et al., 2013), and the Zelterman (1988) and modified Zelterman estimators (Böhning, 2010; Böhning and van der Heijden, 2009). Böhning (2010) compares and tests these methods in simulations. They are designed for sparse, heterogeneous, and zero-truncated count data, conditions under which more complex approaches such as Bayesian capture-recapture models, zero-truncated count regression, or multiplier methods become highly sensitive to modeling assumptions or require external information (e.g., Jones et al., 2016). For instance, Bayesian capture-recapture models and multiplier methods require external data to inform priors for population size, detection probabilities, list dependence, and covariate effects (Jones et al., 2016). We do not have this information. Chao-type and Zelterman estimators rely on minimal assumptions, draw primarily on the most reliable portion of the count distribution, and remain stable even when many individuals are observed only once. Their covariate-adjusted extensions incorporate demographic or behavioral factors while preserving the robustness of the underlying estimators. Given these advantages, they represent the most empirically suitable choice for estimating the size of a hidden population in the presence of data sparsity and uncertainty.⁶

Generally, capture-recapture methods in the social sciences rely on registration files.⁷ One-source capture-recapture methods rely on a single register, i.e., a data source. The advantage

⁶ We also used zero-truncated models. They yield smaller point estimates because the distributional assumption constrains the tail of the count distribution and limits the implied number of unseen cases. We do not present the results here, which are available upon request, as the zero-truncated models do not converge for the specifications with the most covariates.

⁷ The method could be applied to other databases, such as self-reported surveys, provided that the data contain repeated or count-type observations that allow for modeling capture–recapture structures. In such cases, the estimators could help adjust for underreporting or missing cases. Triangulating these estimates with results from

of using only one data source is that it does not require the (in our application) unverifiable assumption that the two sources are statistically independent.⁸

Suppose there are n uncovered scientists associated with articles retracted for severe misconduct, organized in a register, i.e., a data source, with each included with frequency f_j , where j denotes the number of retractions by an author associated with articles retracted for severe misconduct, such that f_1 denotes the number of scientists with one retraction of a fraudulent article, f_2 the number of scientists with two retractions, and so on. Since only scientists with uncovered misconduct are included in such a database, frequencies $f_1, f_2, \dots$ are observed, but f_0 , denoting the cases of undetected misconduct, is not. We aim to estimate f_0 to arrive at the total population of scientists associated with articles retracted for severe misconduct, which can be expressed as the sum of the observed number of scientists associated with articles to be retracted for severe misconduct and the estimated number of their undetected counterparts: $N = n + f_0$.

Capture-recapture estimators infer the number of undetected scientists associated with articles to be retracted for severe misconduct f_0 by exploiting information from the observed distribution of counts. Chao (1987)'s estimator provides a lower bound of f_0 by leveraging the counts of scientists with exactly one, f_1 , and exactly two detected retractions, f_2 .⁹ Specifically, Chao (1987) showed that:

independent data sources, such as administrative records or survey-based estimates, would strengthen the findings and provide a more comprehensive assessment of the hidden population size.

⁸ Using two databases as sources, such as Scopus and the Retraction Watch Database, does not work in our case because the registered events should be largely the same, as both databases strive for completeness and as journal retraction records are readily available.

⁹ Chao- and Zelterman-type estimators rely only on the frequencies of one- and two-count observations, making them semiparametric (or locally parametric) estimators. They do not require full specification of the count distribution or the heterogeneity in detection probabilities across units. Instead, they exploit minimal moment conditions or a local likelihood approximation that is sufficient to identify the mass at zero.

The limited reliance on distributional assumptions makes these estimators more robust to misspecification than fully parametric alternatives, which depend on correctly modeling the upper tail of the count distribution and the

$$\hat{f}_0 = \frac{f_1^2}{2f_2}.$$

This ratio compares scientists observed once with those who are observed twice to infer the likely size of the completely undetected population. Intuitively, scientists associated only once with an article retracted for severe misconduct can be thought of as “near misses” that the detection system barely captured, suggesting many similar cases went undetected. By contrast, observing scientists multiple times indicates that detection probabilities are relatively high. Thus, a distribution dominated by one-time detections implies that the detection system captures only a small share of the underlying population associated with articles to be retracted for severe misconduct, and that the true number of these scientists is likely substantially larger than the observed count.

The Chao estimate follows from the Cauchy–Schwarz inequality, which provides a lower bound on the number of unobserved individuals based on the numbers observed once and twice. The Chao estimator allows detection probabilities to vary across individuals without requiring the sources of this heterogeneity to be observed or explicitly modelled. The resulting estimate of the total population size is therefore:

$$\hat{N}_C = n + \frac{f_1^2}{2f_2}.$$

full heterogeneity structure. In settings with strong detection heterogeneity and sparse high-count observations, such as misconduct, focusing on one- and two-count observations reduces sensitivity to extreme cases and model errors. As a result, Chao- and Zelterman-type estimators trade some efficiency for greater reliability and stability, yielding conservative estimates that underestimate the true population size but remain credible.

A simple illustration from wildlife population estimation is as follows: to estimate the number of fish in a lake, researchers rely not on the few fish caught many times, but on how many fish are caught once versus twice across sampling rounds. If many fish are caught only once and very few twice, this suggests many fish were never caught. Translating this logic to the retraction context, many authors with exactly one detected retraction relative to those with two indicates substantial under-detection of misconduct. The estimator therefore uses the frequency of one- and two-time detections to infer the size of the hidden population. At the same time, higher counts add little identifying power and may introduce instability. This reflects that information about unseen cases lies mainly in the lower tail of the detection count distribution, not in its extreme upper tail.

Because it uses only f_1 and f_2 , i.e., the number of scientists associated with one and two observed retractions, this estimator is robust because it does not require modeling the full count distribution. Therefore, it is less sensitive to extreme observations, and requires no strong distributional assumptions. However, it typically underestimates the true population under substantial detection heterogeneity.

The generalized Chao estimator incorporates covariate information to account for observed heterogeneity (Böhning et al., 2013). If covariates z_i are available for each observed case, the underlying counts are modeled as Poisson with a log-link function:

$$\lambda_i = \exp(\alpha + \beta^T z_i),$$

where λ_i is the Poisson parameter, and α and β are coefficients.

Because only cases with exactly one or two detected retractions are used in the estimation, the truncated Poisson likelihood simplifies to a binomial logistic likelihood with a logit link. In practice, this means the parameters can be estimated by fitting a logistic regression that models the probability of observing a count of 2 versus 1. After fitting the model, the estimated Poisson parameter for each covariate stratum is recovered as:

$$\hat{\lambda}_i = 2 \frac{q_i}{1-q_i},$$

where q_i is the predicted probability of a count of two from the logistic regression. This equation converts the estimated probability of observing a scientist twice rather than once into an estimate of the underlying detection intensity within a given covariate group. As above, this intuitively means that if scientists in a given covariate group are more often observed twice than they are observed once, the detection probabilities are relatively high and that scientists associated with articles to be retracted for severe misconduct in this covariate group are more likely to reappear in the observed data. The equation therefore translates the observed balance between one-time and two-time detections into the Poisson parameter λ_i ,

which summarizes the expected detection rate for that subgroup. A larger λ_i indicates higher detection intensity and, therefore, a smaller hidden population within that covariate stratum.

The estimated zero-counts in each covariate stratum are then:

$$\hat{f}_{0i} = \frac{f_{1i} + f_{2i}}{\hat{\lambda}_i + \hat{\lambda}_i^2 / 2}.$$

This equation estimates the number of undetected scientists associated with articles to be retracted for severe misconduct within a given covariate group by scaling up the observed low-frequency cases using the estimated detection intensity. The numerator represents the number of scientists in that subgroup who were observed once or twice and therefore provide information about the “edge” of the detection system. The denominator reflects the expected probability, under the Poisson model, that a misconducting scientist in this subgroup would be detected once or twice, given the estimated detection rate $\hat{\lambda}_i$. Intuitively, if the estimated detection intensity is low, observing even a small number of scientists once or twice implies that many additional scientists likely remained undetected, leading to a larger estimate of hidden cases.

The total estimated size of the population is:

$$\hat{N}_{GC} = n + \sum_{i=1}^M \hat{f}_{0i},$$

where n is the number of observed scientists in the data and M is the total number of covariate strata included in the model, i.e., the estimated number of scientists who were never detected across the different covariate groups. This estimator reduces bias by adjusting for observed factors influencing detection probabilities and is asymptotically unbiased if all heterogeneity is captured by the covariates.

The original Zelterman estimator (Zelterman, 1988) also uses counts of one and two, but instead of the Cauchy–Schwarz approach, it treats them as arising from a truncated Poisson distribution. The estimate of the Poisson parameter λ is:

$$\hat{\lambda}_Z = \frac{2f_2}{f_1}.$$

This estimates the underlying detection intensity by comparing how often scientists are observed twice relative to once (see above).

Then the probability of observing zero counts is estimated as:

$$P(0) = \exp(-\hat{\lambda}_Z),$$

and the total population size is estimated as:

$$\hat{N}_Z = \frac{n}{1-P(0)}.$$

Here, the observed number of scientists associated with articles retracted for severe misconduct n is treated as only the detected fraction of the total underlying population. $1 - P(0)$ represents the estimated probability that a scientist associated with articles retracted for severe misconduct is observed at least once, while $P(0)$ is the probability of remaining undetected. Dividing the observed population by the estimated detection probability therefore “scales up” the observed data to recover the likely total population size.

Since covariate information is available, estimation is performed by fitting a logistic regression model to the binary outcome indicating whether an observation has a count of 2 (versus 1). This yields predicted probabilities that are transformed back to $\hat{\lambda}_i$ in the same way as in the generalized Chao estimator.

Because the Zelterman estimator overestimates the hidden population size when the Poisson distribution is “contaminated”, i.e., when there is a deviation from the ideal Poisson distribution, mostly due to heterogeneity in the population, we also apply a modified version

of Zelterman's estimator that includes a bias correction (Böhning, 2010; Böhning and van der Heijden, 2009). The modified Zelterman estimator can be written as:

$$\hat{N}_{ZM} = \frac{(f_1 + f_2)}{1 - P(0)} + (n - f_1 - f_2),$$

where the numerator of the first term estimates the size of the subpopulation represented by those scientists who were observed once or twice. $1 - P(0)$ represents the estimated capture probability, the probability that an individual is observed at least once. The remaining scientists $(n - f_1 - f_2)$ are those observed more than twice and are added back in because they do not contribute to the estimation of $\hat{\lambda}$ but are known.

Together, Chao's estimator provides a simple, robust lower bound¹⁰ that relies on minimal assumptions and remains valid under substantial unobserved heterogeneity. Still, its conservative nature can substantially underestimate the true population size when detection probabilities vary widely. The generalized Chao estimator improves on this by incorporating observed covariates to account for detection heterogeneity, reducing bias when relevant factors are measured well. However, it remains sensitive to omitted variables and still depends on information from the lower tail of the count distribution. Zelterman's estimator adopts a local parametric approach that can yield more efficient estimates by leveraging a truncated Poisson approximation. At the same time, it may overestimate the hidden population when heterogeneity violates the Poisson assumption. The modified Zelterman estimator addresses this limitation through bias correction and by separating low-count

¹⁰ The lower-bound property arises because the estimator relies only on the frequencies of individuals observed once and twice and does not fully account for the mass of individuals with extremely low detection probabilities. Under heterogeneity, the observed sample disproportionately represents individuals with relatively high detection probabilities, while those with very low detection probabilities are systematically underrepresented or absent from the observed count distribution. Because these low-detection individuals contribute little information to f_1 and f_2 , the estimator cannot fully recover the lower tail of the detection probability distribution, leading to downward-biased estimates of the hidden population size.

observations from higher-count cases, improving stability at the cost of additional modeling complexity.¹¹

3.2. Critical Assumptions of Capture-Recapture Models

The validity of capture-recapture estimators also hinges on three core assumptions essential to interpreting the results (Bouchard et al., 2019). Full database coverage ensures that all detectable cases have a nonzero probability of appearing in the observed data. If some population segments were systematically excluded, the capture probability would be misspecified, and the hidden population may be underestimated. The independence assumption, that detection events are not systematically correlated, matters because dependence between captures (for instance, detecting one author makes subsequent detections more likely) biases the estimates. The closed population assumption requires that the population of interest remains stable over the observation window. Violations such as authors entering or leaving the sample during the study period confound detection with population dynamics.

3.2.1 Full Coverage of the Register

The first assumption is that the register fully covers the retractions. The Retraction Watch Database is the most comprehensive source and outperforms other sources such as Scopus and CrossRef (Schneider et al., 2023; Sebo and Sebo, 2025). Nevertheless, some retractions may be missing. This implies that some observed misconduct cases are not included in the analysis, which lowers the observed capture rate. The comparative studies cited above suggest that this issue might be small.

¹¹ The variance calculation for these estimators can be found in Böhning & van der Heijden (2009), Böhning (2010), and Böhning et al. (2013). While we performed the analysis in STATA, R code for the Chao and generalized Chao estimators can be found in the supplementary material to Böhning et al. (2013). R also provides the CRAN package for single-source capture-recapture methods, including the estimators used in this paper: <https://cran.r-project.org/web/packages/singleRcapture/index.html>.

3.2.2 Independence

A second assumption is that retraction events are independent, meaning that retractions do not cause behavioral changes among scientists or investigators of scientific misconduct that increase or decrease the likelihood of further retractions. A first potential violation of the independence assumption is that scientists with one detected retraction have a shorter observation period in which another retraction can occur. Following Bouchard et al. (2019), we control for the year of the first retraction in our econometric specification. We further control for the total number of retractions in the previous years, divided by the total publication stock. Like Bouchard et al. (2019), we acknowledge that, while this solution is not perfect, it is the best we can do.¹²

A second point, that the likelihood of detection is independent of previous detections, is more pernicious. High-profile fraud cases have seen the scrutiny of researchers' entire bodies of work. In Diederik Stapel's case, for instance, a committee was formed to determine the extent of the fraudulent research (Levelt et al., 2012). Likewise, over the last decade, scientists have emerged who aim to detect scientific fraud by investigating suspicious research results (Marcus and Oransky, 2018; Shen, 2020). Because of these commendable efforts by the scientific community, scientists with many retractions are likely to violate the assumption of independent discovery. However, while high-profile cases can drive special investigations, they remain rare exceptions in the scientific system, as they typically center on a prominent scientist who has published many articles in top journals. In our sample, the majority of misconduct cases are committed by scientists who perform below average in terms of publication quality in medical science. The average journal impact factor for a fraudulent

¹² In addition, as the corpus of published research grows, the probability of detecting fraudulent research likely decreases, other factors constant. To account for this dynamic, we also control for the cumulative number of articles in biomedical research published up to the current year, as drawn from the Scopus database.

article is 4.2, which is relatively low for a medical journal, where top journals reach 50 or even 100. Hence, the scientists in our sample are unlikely to attract much attention from the scientific community. This is also evidenced by the observation that retracted articles continue to be cited after retraction (Bolland et al., 2022; Budd et al., 1999). Furthermore, most article retractions receive similar media and social media attention to the original publications, whereas retractions of popular articles receive much less attention than the originals (Serghiou et al., 2021). We therefore argue that, rather than leading to media attention or special investigations, detecting such small misconduct cases, even a second case by the same scientist, is more akin to being subject to an uncoordinated pool of individual scientists working in the same field. This strongly suggests that the likelihood of detecting a second or third case by the same scientist is similar to that of the first case. Importantly, our estimators also rely only on cases where a small number of retractions (one or two retractions) are uncovered. We acknowledge, however, that our methodology hinges on the assumption that retractions have little impact on low-profile cases.

Another question is whether scientists adjust their behavior after a fraudulent article has been discovered. While this is expected through channels such as fewer research opportunities due to reputational damage (Azoulay et al., 2017) or because the scientist commits no further misconduct, the researcher's existing articles remain available for detection.

While we cannot formally prove the assumption of independent events, a look at the time until detection shows that it takes, on average, 2.03 years from the publication year to detect the first incidence of fraud for our sample of articles afflicted with serious fraud, while it takes almost the same time, 2.37 years, to detect fraud in other articles by the same scientist. The median is one year for the first and later detections. One interpretation for the relatively stable retraction lag, which suggests that the detection likelihood is unchanged between the first and later retractions, is that for the average misconduct case, detectors more closely

represent an uncoordinated, large pool of scientists, which consists of virtually all scientists working on related topics, and not a focused group of investigators. This also aligns with the notion that scientists face similar incentives to detect the first, second, or later fraudulent articles by the same author.

While arguments suggest the independence assumption is not strongly violated, we also cannot assume it holds unconditionally in our context. The counterarguments center around the fact that, in most countries, a formal investigation is launched quickly following the discovery of misconduct; e.g., for an investigation of an NIH grant suspected of being afflicted by misconduct, an investigation has to be started within the following 30 days,¹³ and additional scrutiny would inevitably be devoted by investigators and by other scientists to the alleged fraudster's prior production.

If the independence assumption were violated, our results would, however, still be informative and would provide a conservative estimate that may underestimate the true extent of scientific misconduct in the medical sciences. More specifically, if misconduct events are positively dependent at the author level, observed retractions will cluster among a subset of scientists who are repeatedly detected. This increases the relative frequency of scientists observed multiple times compared to scientists observed only once. Since our capture-recapture estimators infer the size of the hidden population primarily from the balance between one-time and repeated detections, a higher concentration of repeated detections mechanically implies higher estimated detection probabilities and therefore a smaller estimated hidden population. Intuitively, repeated observations make the detection system appear more effective than it actually is for the broader population, leading the estimators to understate the number of undetected misconducting scientists.

¹³ <https://grants.nih.gov/policy-and-compliance/policy-topics/research-misconduct/handling-process>

3.2.3 Closed Population

A third assumption is that the population is closed. We cannot be sure during which time frame our scientists were involved in research that led to a retraction. Some individuals may have been associated with articles retracted for severe misconduct in the past but have since left academia or stopped publishing. In contrast, others may only begin being associated with articles retracted for severe misconduct after the observation period ends. If the closed-population assumption is violated, we would observe more authors with one or two retractions due to misconduct than would be expected under a closed population. Depending on how the distribution of authors with one or two retractions changes, we might over- or underestimate the true population. This means that the population is effectively dynamic. We have a time window of 28 years, but use shorter time periods for our robustness checks, which to some extent alleviates the concern that many scientists enter and leave the register. Most importantly, retractions are identified retrospectively so that we can be sure that the retractions in our time window are complete. Lastly, closed-population assumption is consistent with the conventions of prior work (Böhning et al., 2013; Bouchard et al., 2019; Chao, 1987). In addition, we present results for split samples, which may better approximate the closed-population assumption, in Section 5.3.2.

4. Data

The underlying data on retractions of fraudulent articles is taken from Retraction Watch, a website that reports retractions of scientific articles since 2010. The data were received in September 2022 and contained 35,527 retracted articles. The Retraction Watch Database¹⁴ publishes reasons for retractions, ranging from data fabrication to missing third-party

¹⁴ <http://retractiondatabase.org/>

approval. For our study, we identified unambiguous cases of severe misconduct defined as fabrication, falsification, manipulation, or duplication of data and images. Retractions based on error, administrative issues, authorship disputes, missing approvals, or other reasons that do not clearly indicate severe misconduct are excluded.¹⁵ This left us with 5,721 cases of retracted publications containing severe misconduct.

We linked the 5,721 fraudulent retractions using their DOI and PubMed number to the Reliance on Science (RoS) database (Sinha et al., 2015).¹⁶ After eliminating DOI duplicates and attempting to retrieve missing DOIs or PubMed numbers, we matched 5,363 articles and retrieved the unique paper identifier from the RoS database.¹⁷ RoS provides a link between the paper identifier and a unique author identifier. Using this link, we retrieved the full publication record of the authors involved in the 5,363 fraudulent retracted articles. After removing unreasonable records, we obtained a sample of 477,764 publications by 14,102 unique authors, according to the RoS author identifiers. We kept only journal publications and excluded other document types, such as working papers, reducing the sample to 456,214 articles. To further focus on relevant articles, we excluded publications from journals without an impact factor (JIF). This left us with 451,819 articles by 14,033 unique authors.

On average, the 5,721 retracted articles had 3.84 authors, with a maximum of 81 authors on a single article. For our study, we follow up only on authors who appear at least once as the first or last author of a fraudulent retracted article. The order of the contributors' names on a publication in the hard sciences typically assigns credit and responsibility (Bhattacharya,

¹⁵ As the unit of classification is the individual retracted article, an article is included only when the Retraction Watch reason satisfies our severe misconduct criteria. Articles involved in collective retractions and paper-mill cases are only included when the recorded retraction reason independently meets our severe misconduct definition.

¹⁶ https://zenodo.org/record/3575146#.Y1FCI_xBw2x

¹⁷ We manually checked 50 articles to validate our match and only found a different match in two cases.

2010; Lapidow and Scudder, 2019; Lissoni et al., 2013). Medical science research is typically conducted in teams, often with a substantial division of labor among coauthors. Within these teams, first and last authors generally occupy particularly central roles. First authors are typically responsible for the execution and coordination of the specific research project. They are therefore likely to be closely involved in, and aware of, the research activities underlying the publication. Consequently, they should also be among the most likely to be aware of scientific misconduct if it occurs (Hussinger and Pellens, 2019b). Last authors, by contrast, are typically senior scientists who lead the laboratory or research team, set its broader research agenda, supervise team members, and secure the funding supporting the research. They are therefore typically prominent researchers with substantial responsibility for organizing and directing the team's research (Bhattacharya, 2010; El-Gilany and Farrag, 2022; Lissoni et al., 2013). Given the distinct but central roles of first and last authors within an author team, both can credibly be considered particularly relevant levels of analysis. We therefore analyze both groups in parallel.

The 1,969 authors who appear at least once as the first author of a fraudulent retracted article have published 13,961 articles in journals with a JIF (or 11,283 articles among the 1,844 last authors). The RoS data were used to calculate publication stocks¹⁸ (P), coauthor stocks (C), the journal impact factor premium (J), which we define as the JIF-weighted publication stock over the publication stock, the total number of publications in the field of medical research according to Scopus (TOTP), and the year of the first retraction (Y) and the total number of retractions in the past year over the publication stock (TOTM), following Bouchard et al.

¹⁸ Stocks are defined as follows: $X_{it}^{stock} = x_{it} + (1 - 0.15) * x_{it-1}^{stock}$, where x stands for the number of publications, impact factor weighted publications, publications afflicted with misconduct, or coauthors. Stock measures account for the past at a depreciated rate, typically set to 15%. The depreciation rate reflects the depreciation of knowledge and coauthor networks, meaning recent publications or coauthor relationships matter more than older ones.

(2019). In addition, we include binary variables to account for the year of retraction. This variable is lagged one year in the regressions.

For the final sample, we exclude retraction years before 1990 and after 2017 because the number of observations per year is negligibly small, suggesting that these years' records are incomplete. As required by the estimation approach, we drop all years in which the authors who appear at least once as the first author on a fraudulent retracted article had no fraudulent articles, so we retain only years with a positive count of fraudulent retracted articles, which is our dependent variable. This leaves us with a sample of 1,172 author-year observations for the authors and 1,647 for last authors.

An inspection of the authors' active fields revealed that most authors (973, or 83%) conduct research in the medical and health sciences. This aligns with Fanelli's (2009) findings, which showed that misconduct was reported more frequently among medical and pharmacological researchers than among researchers in other fields. Because this field drives the results in the total sample, we focus only on the medical and health sciences. Table 1 presents the observed frequencies. Among the first authors, 76% are observed only once as authors of a retracted fraudulent article, 13% twice, and 11% more than twice. Last authors follow a similar distribution: 76% are observed once, 15% twice, and 9% more than twice. A higher number of last authors in the Retraction Watch data is expected because senior researchers typically supervise many more projects and therefore appear in substantially more publications than first authors. Consequently, they have greater exposure to potential retractions and may accumulate retractions arising from different members of their research group, even if their individual propensity for misconduct is not higher. The total number of retractions differs between the first-author and last-author analyses because some first authors, the less prominent author group, could not be identified.

Table 2 presents descriptive statistics. As expected, the largest difference between first and last authors appears in the publication stock, where last authors have a much higher mean.

Table 1. Observed frequencies of misconduct cases.

Frequency	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16+}	Total
First authors	741	128	55	17	9	5	10	1	1	1	3	1	0	0	0	2	974
Last authors	1,063	201	66	19	12	8	6	3	2	4	3	1	1	0	0	4	1,393

Table 2. Descriptive statistics.

	First authors		Last authors	
	mean	s.d.	mean	s.d.
Number of misconduct articles per year (if > 0)	1.53	1.50	1.51	1.56
Publication stock (P)	18.24	22.13	28.38	27.88
Coauthor stock (C)	10.68	7.75	10.26	7.22
Journal Impact Factor premium (J)	3.75	1.91	3.35	1.76
Total publications in the field ('00 000) (TOTP)	44.37	9.26	44.26	10.08
Year of first retraction (centered 2017) (Y)	9.23	5.47	9.84	6.29
Share of retractions in the field (TOTM)	0.15	0.21	0.09	0.19

5. Results

5.1. Main Results

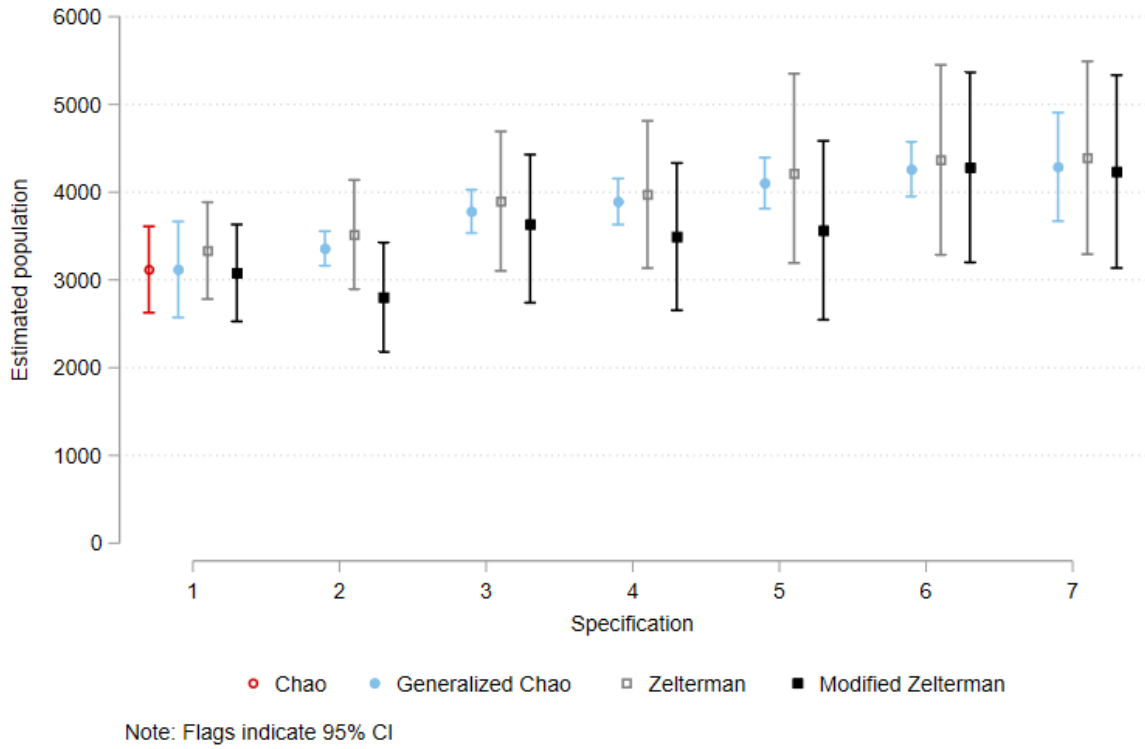
Tables A.1 and A.2 in the Appendix present the logit regression results underlying the Generalized Chao, Zeltermann, and modified Zeltermann population estimators, along with goodness-of-fit indicators for various specifications. Table A.1 presents the results for first authors, while Table A.2 presents the results for last authors. The best-fitting models, in terms of AIC and pseudo- R^2 , for both first and last authors are model 6, which includes misconduct intensity, publication stock, coauthor stock, journal impact premium, year since first retraction, and calendar year effects, and model 7, which also adds publication stock. The specification with total publication stock is not preferred by AIC but yields a slight

improvement in the likelihood ratio test. We focus on the population estimates resulting from these two models.¹⁹

Figure 1 shows the population estimates for first authors. When covariates are added to models 2 to 7, the estimates from the generalized Chao, Zelterman, and modified Zelterman estimators increase, but the former yields smaller confidence intervals than the latter two. The Zelterman estimator consistently yields the largest estimates, but these do not increase disproportionately relative to the other two estimators when covariates are added. Following Böhning (2010), the generalized Chao and modified Zelterman estimators are considered the most accurate. In models 6 and 7, the Generalized Chao and Modified Zelterman population is estimated around 4,250 first authors (Generalized Chao: model 6: 4,261 [3,950-4,572], model 7: 4,289 [3,670-4,908]; Modified Zelterman: 4,282 [3,199-5,364] and 4,235 [3,136-5,334]). Compared to the 973 observed first authors associated with articles retracted for severe misconduct, this implies a capture rate of around 23% (973 observed first authors associated with articles retracted for severe misconduct divided by an estimated population of scientists associated with articles to be retracted for severe misconduct of 4,235 in the model 7 Modified Zelterman estimate).

¹⁹ In terms of coefficient estimates, the likelihood of involvement in two misconduct publications rather than in only one increases with the overall publication stock, the stock of misconduct publications, and the stock of coauthors. For first authors, the likelihood decreases as the JIF-weighted publication stock increases. For last authors, this effect does not persist, but their likelihood increases the more recent the first retraction year is.

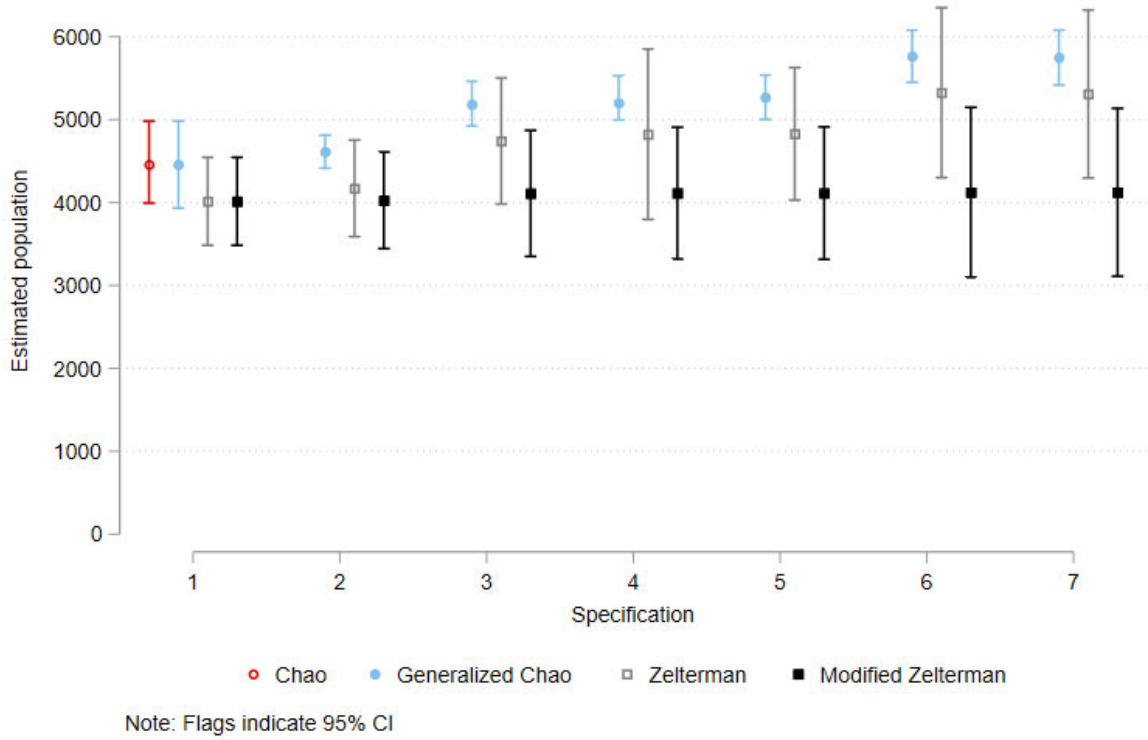
Figure 1: $\hat{N}$ and 95% confidence interval for the population of first authors associated with articles to be retracted for severe misconduct obtained from Chao, Generalized Chao, and Modified Zelterman estimators.



This figure reports the estimated population size $\hat{N}$ for each model specification, obtained using the Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. See Table A.3 in the Appendix for the complete results.

Figure 2 shows the population estimates for the last authors. Compared with the first-author results, the Modified Zelterman estimates are lower than the Zelterman estimates. This suggests that, for the last authors, the Zelterman and Generalized Chao estimators may be upwardly biased due to heterogeneous capture probabilities. Hence, we focus on the Modified Zelterman estimates as a conservative result. In models 6 and 7, the population is estimated at around 4,123 last authors (Model 6: 4,123 [3,098-5,148], Model 7: 4124 [3,111-5,136]). Compared with the 1,393 observed last authors, this implies a higher capture rate of 34%. One potential explanation for this higher capture rate is that, because of their more prominent roles, last authors are more visible and subject to greater scrutiny when misconduct is revealed.

Figure 2: $\hat{N}$ and 95% confidence interval for the population of last authors associated with articles to be retracted for severe misconduct obtained from Chao, Generalized Chao, and Modified Zelterman estimators.



This figure reports the estimated population size $\hat{N}$ for each model specification, obtained using the Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. See Table A.4 in the Appendix for the complete results.

5.2. Robustness Checks: Independence Assumption

As discussed above, we run robustness checks to test how sensitive our results are to potential violations of the independence assumption. Therefore, we run additional analyses that split the samples into more and less prominent scientists. In principle, more prominent scientists are more likely to be caught after being found associated with articles retracted for severe misconduct, yielding a higher recapture rate. For brevity, and since comparison between the first authors and last authors samples would not be possible since different threshold values would be needed for both samples, we focus on first authors in these analyses, on estimates

from the Modified Zelterman estimator, and on capture rates. The results are presented in Table 3.²⁰

Table 3: capture rates as implied by split-sample analyses according to JIF and time of first publication.

JIF	≤ 4	> 4
Capture rate	29%	36%
First publication	After 2004	Up to 2004
Capture rate	24%	42%

This table reports estimated capture rates resulting from Modified Zelterman population estimates (specification 6). See Tables A.5-A.8 in the Online Appendix for the complete results.

We consider two proxies for prominence. First, we split the sample into two groups based on JIF premiums, defined as the JIF-weighted publication stock divided by the publication stock. We use a threshold of 4, which yields two sizeable subsamples: 383 high-JIF and 591 low-JIF authors. We find that for authors with a high JIF premium have a higher capture rate (36%) than those with a low premium (29%). Second, we split the sample by career age, measured as the time elapsed since first publication. We use 2004, the sample mean, as a threshold. This yields 450 older versus 524 younger scientists. Here, too, the capture rate is higher for older scientists, i.e., those who are, on average, more prominent, at 42%, compared with 24% for younger scientists.

This sample split also allows us, to some extent, to assess whether potential violations of the independence assumption are more likely to arise in the subgroup of more prominent scientists. While these authors are arguably more concerning from an impact perspective given the wider diffusion and influence of their research, their higher capture rates suggest they are less likely to go undetected in our setting. This mitigates, to some extent, concerns

²⁰ See Tables A.5 through A.8 in the Online Appendix for the population estimates.

that dependence across detection channels among high-profile scientists would substantially overestimate misconduct within this subgroup.

5.3. Further Analyses

5.3.1. Natural Sciences

To better quantify scientific misconduct in medical science, we compare these results with those in the natural sciences. We select the natural sciences because the Retraction Watch database identifies this field as having the second-highest number of observed cases after the medical sciences: 102 first authors and 102 last authors. The results are presented in Tables A.9 and A.10 in the Online Appendix. For this smaller sample, the AIC criterion recommends the first specification without covariates for both the first- and the last-author samples.²¹ In specification 1, the Modified Zelterman estimates of the populations of first authors associated with articles to be retracted for severe misconduct, and that of last authors, are the same: 281 (95% confidence interval: 159-404), because the distributions are almost identical. This amounts to a capture rate of 36% for first as well as for last authors (102 observed scientists associated with articles retracted for severe misconduct divided by 281 estimated scientists associated with articles to be retracted for severe misconduct).

This means we find a higher capture rate in the natural sciences than in the medical sciences.

This may be driven by differences in the incidence of misconduct or in detectability.

Arguments for the latter build on the fact that natural-science research often relies on highly structured data, strong theoretical constraints, and standardized representations, making anomalies, internal inconsistencies, and manipulations easier to identify through replication, statistical checks, or forensic screening. Experiments may also be more readily reproducible and, in some cases, cheaper in the natural sciences, increasing the likelihood that problematic

²¹ We omit the logit regression results for brevity.

results are independently scrutinized. Medical research, in contrast, typically involves heterogeneous populations, noisy outcomes, flexible modeling choices, and restricted access to primary data due to ethical and privacy constraints, all of which reduce the visibility of misconduct and make it harder to distinguish fabrication or falsification from legitimate variability. However, we cannot exclude impacts of other institutional differences between fields, for example, in ethical standards.²² Our reduced-form approach does not allow us to shed light on the origin of these differences. Still, it highlights that the observed gap in capture rates should not be interpreted as a straightforward difference in the true prevalence of misconduct across fields. Rather, it likely reflects differences in detection environments and underlying behavior.

5.3.2. Evolution over Time

In a second analysis, we examine how scientific misconduct in the medical sciences has evolved over time. We split our sample according to the year of the scientist's first retraction, distinguishing pre-2006, 2006-2011, and post-2011.²³ For brevity, we focus on first authors and present the results of the Modified Zelterman estimates of specification 6 in Table 4. The estimated population of scientists associated with articles to be retracted for severe misconduct increases significantly: from 668 with a first retraction before 2006, to 892 with a first retraction from 2006 to 2011, and to 1,866 with a first retraction from 2011 onward. This implies an increase of approximately 180% between the first and last periods, far above the

²² This possibility was suggested by a reviewer.

²³ We chose these periods to roughly equal sample sizes while having meaningfully narrow observation periods. For brevity, we omit the logit regressions for this analysis. See Tables A.11, A.12, and A.13 in the Online Appendix for the full set of population size estimates.

growth rate of the corpus of medical articles.²⁴ In terms of capture rates, the results look more positive, as the capture rate only decreases slightly over time.²⁵

Table 4: Modified Zelterman estimates and capture rates per time of first retraction, first authors.

Period of first retraction	Modified Zelterman Estimate	Capture rate
Before 2006	668 [173–1162]	17%
2006–2011	892 [405–1378]	18%
After 2011	1866 [272–3459]	15%

This table reports estimated capture rates resulting from Modified Zelterman population estimates (specification 6). 95% confidence intervals in square brackets. See Tables A.11, A.12, and A.13 in the Online Appendix for full results.

6. Discussion

We use capture-recapture methods to estimate the population of observed and unobserved scientists associated with articles to be retracted for severe misconduct in medical and health sciences. We estimate the first-author population at approximately 4,250, corresponding to a capture rate of roughly 23%. We estimate a higher capture rate for last authors, at 34%. Both estimates imply that the majority of scientists associated with articles to be retracted for severe misconduct are not detected. With that, we help quantify the size of the iceberg. Based on our estimates, the societal, economic, and personal costs of misconduct are still underestimated, as detected misconduct accounts for only a quarter to a third of total misconduct.

²⁴ For instance, the flow of Medline citations has roughly doubled between 1997 and 2017. Cf: https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

²⁵ Capture rates are lower in the subperiod analyses than in the full-sample analysis. This likely reflects the smaller sample sizes within each subperiod. With fewer observations, the frequencies of individuals observed multiple times become more variable, particularly the number observed exactly twice. Since our estimators rely heavily on the ratio of singletons to repeated observations, smaller samples tend to yield larger estimates of the unobserved population and, consequently, lower estimated capture rates.

Our results paint a more optimistic picture than other estimates of misconduct incidence (e.g., Fanelli, 2009; Fang et al., 2012; Xie et al., 2021) as our results still suggest that a much smaller number of scientists is involved in retracted articles than prior work implies, given that our estimates are conservative. This difference should not be interpreted as evidence that misconduct is unimportant, but rather as indicating that different approaches capture different underlying phenomena. Survey studies may include a broad range of self-reported or observed behaviors, including isolated instances of misconduct, whereas our estimates concern scientists associated with severe cases that would ultimately warrant retraction. Our findings therefore suggest that severe misconduct may be less widespread across the scientific population than some prior prevalence estimates imply, while also being considerably underdetected in the observed retraction record. This points to a more nuanced understanding of the problem: severe misconduct may be concentrated among a small group of scientists yet generate a substantial hidden stock of problematic research.

From a policy perspective, our estimates show that observed retractions represent only a part of the latent population, generated by the joint process of scientific production, involvement in severe misconduct, and detection. Because these components cannot be separately identified, the estimated capture rate should not be interpreted as a direct measure of the effectiveness of existing research-integrity systems. In particular, the results do not establish whether the hidden component primarily reflects a high incidence of misconduct, limited detectability, differences in publication exposure, or a combination of these factors. The main implication is therefore that observed retractions alone provide an incomplete basis for assessing the scale of the problem or evaluating research-integrity institutions. Policy responses should address both margins: prevention-oriented measures that reduce incentives and opportunities for misconduct, and detection-oriented measures such as data and code transparency, auditing, screening, replication, and adequately resourced research-integrity offices. Our estimates cannot

determine the relative effectiveness of these interventions, which would require additional data or research designs capable of separately identifying misconduct incidence and detection.

Our robustness checks allow for more specific recommendations. First, the strong persistence in misconduct detection suggests that initial findings inform future risk, indicating that integrity systems may benefit from greater emphasis on follow-up monitoring after a first confirmed case rather than treating incidents as isolated events. Second, the positive association between repeated detection and publication in high-impact journals points to asymmetric enforcement, implying that oversight mechanisms focused primarily on journal prestige may be misaligned with actual risk patterns. More generally, the evidence supports a shift toward risk-based integrity governance that accounts explicitly for differences in exposure, productivity, and collaboration networks. Third, the comparison between medical and natural sciences further indicates significant differences in observed misconduct rates across fields. This highlights the importance of field-specific integrity infrastructures, particularly in domains characterized by complex data, high analytical flexibility, and restricted access to underlying evidence. Finally, the disproportionate growth in the estimated population of scientists associated with articles to be retracted for severe misconduct relative to publication output over time suggests that integrity mechanisms have not scaled in line with the expansion and increasing complexity of scientific production.

A limitation of this study is that our data identify authors of articles retracted for severe misconduct, but do not establish which coauthors personally engaged in the underlying misconduct. Although authorship position provides a relevant proxy for involvement in the research, neither first nor last authorship conclusively establishes individual responsibility. Consequently, our estimates should be interpreted as the population of authors associated with articles containing severe misconduct, rather than the population of scientists personally responsible for misconduct.

7. Conclusion

This study shifts the focus from prominent individual cases of misconduct to the population of scientists associated with articles to be retracted for severe misconduct by applying capture-recapture methods to retractions. We estimate that retractions capture only a minority of misconduct. However, our estimates should be interpreted as a conservative lower bound; they underestimate the true population size. At the same time, the results suggest that the fraud prevalence, while non-negligible and socially consequential, is lower than some recent aggregate estimates imply, underscoring the importance of estimation approaches grounded in observed detection processes rather than relying solely on self-reports. More broadly, the findings highlight that improving transparency, accountability, and risk-based oversight can substantially reduce the hidden component of misconduct even without changes in underlying behavior. As scientific production continues to expand in scale and complexity, integrity policies that explicitly address detectability, incentives, and institutional structure will be essential for aligning the governance of science with the realities of how misconduct is generated, detected, and corrected.

References

- Azoulay, P., Bonatti, A., and Krieger, J. L. (2017). The career effects of scandal: Evidence from scientific retractions. *Research Policy*, 46(9), 1552–1569. <https://doi.org/10.1016/j.respol.2017.07.003>
- Azoulay, P., Furman, J. L., Krieger, J. L., and Murray, F. (2015). Retractions. *Review of Economics and Statistics*, 97(5), 1118–1136. https://doi.org/10.1162/REST_a_00469
- Berggren, C., and Karabag, S. F. (2019). Scientific misconduct at an elite medical institute: The role of competing institutional logics and fragmented control. *Research Policy*, 48(2), 428–443. <https://doi.org/10.1016/j.respol.2018.03.020>
- Berri , J. P., and Kalliokoski, O. (2025). Fraudulent studies are undermining the reliability of systematic reviews: On the prevalence of problematic images in preclinical depression studies. *FEBS Letters*, 599(11), 1485–1498. <https://doi.org/10.1002/1873-3468.70077>
- Bhattacharya, S. (2010). Authorship issue explained. *Indian Journal of Plastic Surgery: Official Publication of the Association of Plastic Surgeons of India*, 43(2), 233.

- Bik, E. M., Casadevall, A., and Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3), e00809-16. <https://doi.org/10.1128/mBio.00809-16>
- Bohannon, J. (2013). Who's afraid of peer review? *Science*, 342(6154), 60–65. <https://doi.org/10.1126/science.342.6154.60>
- Böhning, D. (2010). Some general comparative points on Chao's and Zelterman's estimators of the population size. *Scandinavian Journal of Statistics*, 37(2), 221–236. <https://doi.org/10.1111/j.1467-9469.2009.00676.x>
- Böhning, D., Bunge, J., and Heijden, P. G. (2018). *Capture-recapture methods for the social and medical sciences*. CRC Press Boca Raton.
- Böhning, D., Lerdsuwansri, R., and Sangnawakij, P. (2023). Modeling COVID-19 contact-tracing using the ratio regression capture-recapture approach. *Biometrics*, 79(4), 3818–3830. <https://doi.org/10.1111/biom.13842>
- Böhning, D., and van der Heijden, P. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *The Annals of Applied Statistics*, 3(2), 595–610.
- Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., and Arnold, M. (2013). A generalization of Chao's estimator for covariate information. *Biometrics*, 69(4), 1033–1042. <https://doi.org/10.1111/biom.12082>
- Bolland, M. J., Grey, A., and Avenell, A. (2022). Citation of retracted publications: A challenging problem. *Accountability in Research*, 29(1), 18–25. <https://doi.org/10.1080/08989621.2021.1886933>
- Bordewijk, E. M., Li, W., van Eekelen, R., Wang, R., Showell, M., Mol, B. W., and van Wely, M. (2021). Methods to assess research misconduct in health-related research: A scoping review. *Journal of Clinical Epidemiology*, 136, 189–202. <https://doi.org/10.1016/j.jclinepi.2021.05.012>
- Bouchard, M., Morselli, C., Macdonald, M., Gallupe, O., Zhang, S., and Farabee, D. (2019). Estimating risks of arrest and criminal populations: Regression adjustments to capture-recapture models. *Crime & Delinquency*, 65(13), 1767–1797. <https://doi.org/10.1177/0011128718807156>
- Brown, N. J. L., and Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4), 363–369. <https://doi.org/10.1177/1948550616673876>
- Budd, J. M., Sievert, M., Schultz, T. R., and Scoville, C. (1999). Effects of article retraction on citation and practice in medicine. *Bulletin of the Medical Library Association*, 87(4), 437.
- Carlisle, J. B. (2012). The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*, 67(5), 521–537. <https://doi.org/10.1111/j.1365-2044.2012.07128.x>
- Carlisle, J. B. (2017). Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*, 72(8), 944–952. <https://doi.org/10.1111/anae.13938>
- Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4), 783–791. <https://doi.org/10.2307/2531532>

- Diekmann, A. (2007). Not the first digit! Using Benford's law to detect fraudulent scientific data. *Journal of Applied Statistics*, 34(3), 321–329.
<https://doi.org/10.1080/02664760601004940>
- El-Gilany, A.-H., and Farrag, N. S. (2022). Authorship in medical research: What should authors know? *Journal of Health Science Research*, 7(2), 1–6.
<https://doi.org/10.7324/jhsr.2022.721>
- Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, 4(5), e5738.
<https://doi.org/10.1371/journal.pone.0005738>
- Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLOS Medicine*, 10(12), e1001563. <https://doi.org/10.1371/journal.pmed.1001563>
- Fang, F. C., and Casadevall, A. (2015). Competitive science: Is competition ruining science? *Infection and Immunity*, 83(4), 1229–1233. <https://doi.org/10.1128/IAI.02939-14>
- Fang, F. C., Steen, R. G., and Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, 109(42), 17028–17033. <https://doi.org/10.1073/pnas.1212247109>
- Feng, L., Yuan, J., and Yang, L. (2020). An observation framework for retracted publications in multiple dimensions. *Scientometrics*, 125(2), 1445–1457.
<https://doi.org/10.1007/s11192-020-03702-3>
- Fisher, N., Turner, S. W., Pugh, R., and Taylor, C. (1994). Estimated numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. *BMJ*, 308(6920), 27–30. <https://doi.org/10.1136/bmj.308.6920.27>
- Frank, O., and Carrington, P. J. (2007). Estimation of offending and co-offending using available data with model support. *The Journal of Mathematical Sociology*, 31(1), 1–46.
<https://doi.org/10.1080/00222500601011423>
- Furman, J. Jensen, K., and Murray, F. (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy* 41(2), 276–290. <https://doi.org/10.1016/j.respol.2011.11.001>
- Glenna, L., and Bruce, A. (2021). Suborning science for profit: Monsanto, glyphosate, and private science research misconduct. *Research Policy*, 50(7), 104290.
<https://doi.org/10.1016/j.respol.2021.104290>
- Gopalakrishna, G., Riet, G. ter, Vink, G., Stoop, I., Wicherts, J. M., and Bouter, L. M. (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. <https://doi.org/10.1371/journal.pone.0263023>
- Hesselmann, F., Graf, V., Schmidt, M., and Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current Sociology. La Sociologie Contemporaine*, 65(6), 814–845. <https://doi.org/10.1177/0011392116663807>
- Hook, E. B., and Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews*, 17(2), 243–264.
<https://doi.org/10.1093/oxfordjournals.epirev.a036192>
- Hussinger, K., and Pellens, M. (2019a). Guilt by association: How scientific misconduct harms prior collaborators. *Research Policy*, 48(2), 516–530.
<https://doi.org/10.1016/j.respol.2018.01.012>

- Hussinger, K., and Pellens, M. (2019b). Scientific misconduct and accountability in teams. *PLOS ONE*, 14(5), e0215962. <https://doi.org/10.1371/journal.pone.0215962>
- Jin, G. Z., Jones, B., Lu, S. F., and Uzzi, B. (2019). The reverse Matthew effect: consequences of retraction in scientific teams. *The Review of Economics and Statistics*, 101(3), 492–506. https://doi.org/10.1162/rest_a_00780
- John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Jones, H. E., Welton, N. J., Ades, A. E., Pierce, M., Davies, W., Coleman, B., Millar, T., and Hickman, M. (2016). Problem drug use prevalence estimation revisited: Heterogeneity in capture–recapture and the role of external evidence. *Addiction*, 111(3), 438–447. <https://doi.org/10.1111/add.13222>
- King, R. (2012). A review of Bayesian state-space modelling of capture–recapture–recovery data. *Interface Focus*, 2(2), 190–204. <https://doi.org/10.1098/rsfs.2011.0078>
- Lacetera, N., and Zirulia, L. (2011). The economics of scientific misconduct. *The Journal of Law, Economics, & Organization*, 27(3), 568–603. <https://doi.org/10.1093/jleo/ewp031>
- Lapidow, A., and Scudder, P. (2019). Shared first authorship. *Journal of the Medical Library Association: JMLA*, 107(4), 618. <https://doi.org/10.5195/jmla.2019.700>
- Levelt, W. J., Drenth, P. J. D., and Noort, E. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. https://pure.mpg.de/rest/items/item_1569964/component/file_1569966/content
- Lissoni, F., Montobbio, F., and Zirulia, L. (2013). Inventorship and authorship as attribution rights: An enquiry into the economics of scientific credit. *Journal of Economic Behavior & Organization*, 95, 49–69. <https://doi.org/10.1016/j.jebo.2013.08.016>
- Lu, S. F., Jin, G. Z., Uzzi, B., and Jones, B. (2013). The retraction penalty: Evidence from the Web of Science. *Scientific Reports*, 3(1), 3146. <https://doi.org/10.1038/srep03146>
- Luchenski, S. A., Böhning, D., Aldridge, R., Stevenson, F., Tariq, S., and Hayward, A. C. (2025). Estimating the scale of hospital admissions for people experiencing homelessness in England: A population-based multiple systems estimation study using national Hospital Episode Statistics. *BMJ Public Health*, 3(2). <https://bmjpublichealth.bmj.com/content/3/2/e002978>
- Marcus, A., and Oransky, I. (2018). The data thugs. *Science*, 359(6377), 730–732. <https://doi.org/10.1126/science.359.6377.730>
- Mastro, T. D., Kitayaporn, D., Weniger, B. G., Vanichseni, S., Laosunthorn, V., Uneklabh, T., Uneklabh, C., Choopanya, K., and Limpakarnjanarat, K. (1994). Estimating the number of HIV-infected injection drug users in Bangkok: A capture–recapture method. *American Journal of Public Health*, 84(7), 1094–1099. <https://doi.org/10.2105/AJPH.84.7.1094>
- Memon, S. A., Makovi, K., and AlShebli, B. (2025). Characterizing the effect of retractions on publishing careers. *Nature Human Behaviour*, 9(6), 1134–1146. <https://doi.org/10.1038/s41562-025-02154-0>
- Michalek, A. M., Hutson, A. D., Wicher, C. P., and Trump, D. L. (2010). The costs and underappreciated consequences of research misconduct: A case study. *PLoS Medicine*, 7(8), e1000318. <https://doi.org/10.1371/journal.pmed.1000318>

- Mongeon, P., and Larivière, V. (2016). Costly collaborations: The impact of scientific fraud on co-authors' careers. *Journal of the Association for Information Science and Technology*, 67(3), 535–542. <https://doi.org/10.1002/asi.23421>
- Mosimann, J., Dahlberg, J., Davidian, N., and Krueger, J. (2002). Terminal digits and the examination of questioned data. *Accountability in Research*, 9(2), 75–92. <https://doi.org/10.1080/08989620212969>
- Raghavarao, D., and Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1), 40–45. <https://doi.org/10.1111/j.2517-6161.1979.tb01055.x>
- Roberts, D. L., and St. John, F. A. V. (2014). Estimating the prevalence of researcher misconduct: A study of UK academics within biological sciences. *PeerJ*, 2, e562. <https://doi.org/10.7717/peerj.562>
- Rodrigues, R., and Pizzolato, D. (2024). Insights on the socio-economic impacts of research misconduct. *Public Governance, Administration and Finances Law Review*, 9(2), 1–30. <https://doi.org/10.53116/pgaflr.7721>
- Schneider, J., Lee, J., Zheng, H., and Salami, M. O. (2023, May 19). Assessing the agreement in retraction indexing across 4 multidisciplinary sources: Crossref, Retraction Watch, Scopus, and Web of Science. *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*. 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). <https://doi.org/10.55835/6441e5cae04dbe5586d06a5f>
- Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. <https://agris.fao.org/search/en/providers/122621/records/64776160f2e6fe92b3678220>
- Sebo, P., and Sebo, M. (2025). Comparing the performance of Retraction Watch Database, PubMed, and Web of Science in identifying retracted publications in medicine. *Accountability in Research*, 1–25. <https://doi.org/10.1080/08989621.2025.2484555>
- Serghiou, S., Marton, R. M., and Ioannidis, J. P. (2021). Media and social media attention to retracted articles according to Altmetric. *PLoS One*, 16(5), e0248625.
- Shen, H. (2020). Meet this super-spotter of duplicated images in science papers. *Nature*, 581(7807), 132–137. <https://doi.org/10.1038/d41586-020-01363-z>
- Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, 24(10), 1875–1888. <https://doi.org/10.1177/0956797613480366>
- Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (Paul), and Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. *Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion*, 243–246. <https://doi.org/10.1145/2740908.2742839>
- Stern, A. M., Casadevall, A., Steen, R. G., and Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *eLife*, 3, e02956. <https://doi.org/10.7554/eLife.02956>
- Stricker, J., and Günther, A. (2019). Scientific misconduct in psychology: A systematic review of prevalence estimates and new empirical data. *Zeitschrift Für Psychologie*, 227(1), 53–63. <https://doi.org/10.1027/2151-2604/a000356>

- There is a worrying amount of fraud in medical research. (2023). *The Economist*.
<https://www.economist.com/science-and-technology/2023/02/22/there-is-a-worrying-amount-of-fraud-in-medical-research>
- Walby, B. (1990). How much scientific fraud is there? *Interdisciplinary Science Reviews*, 15(2), 108–110. <https://doi.org/10.1179/isr.1990.15.2.108>
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63–69.
<https://doi.org/10.1080/01621459.1965.10480775>
- Xie, Y., Wang, K., and Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, 27(4), 41. <https://doi.org/10.1007/s11948-021-00314-9>
- Yu, J.-W., Tian, G.-L., and Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, 67(3), 251–263.
<https://doi.org/10.1007/s00184-007-0131-x>
- Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18(2), 225–237. [https://doi.org/10.1016/0378-3758\(88\)90007-9](https://doi.org/10.1016/0378-3758(88)90007-9)

Appendix

Table A.1. Logit regression results: first authors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Publication stock		0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)	0.02*** (0.00)
Misconduct stock			2.10*** (0.42)	2.19*** (0.43)	2.31*** (0.44)	2.07*** (0.45)	2.13*** (0.46)
Coauthor stock				0.02 (0.01)	0.03** (0.01)	0.03** (0.01)	0.03** (0.01)
Impact factor stock					-0.14** (0.06)	-0.13** (0.06)	-0.13** (0.06)
First retraction year						-0.04** (0.02)	0.02 (0.04)
Total publication stock							-0.02 (0.02)
Intercept	-1.76*** (0.10)	-2.10*** (0.13)	-2.47*** (0.16)	-2.67*** (0.21)	-2.27*** (0.27)	-2.66*** (0.33)	-1.74 (1.22)
N	869	869	869	869	869	869	869
AIC	728.5	709.1	688.2	687.1	684.9	681.4	682.8
df	1	2	3	4	5	6	7
LR		21.33	44.23	47.38	51.57	57.03	57.67
ll	-363.2	-352.6	-341.1	-339.5	-337.5	-334.7	-334.4
Pseudo-R ²		0.03	0.06	0.07	0.07	0.08	0.08

Logit regression. Dependent takes value 0 when author was identified 1 time as misconducting author, and value 1 when the author was identified 2 times. All stocks apply 15% discount rate. Misconduct stock: stock of fraudulent publications divided by publication stock. Impact factor stock: stock of impact factor weighted publications divided by publication stock. Standard errors in brackets. Stars indicate p-value: ***(**,*): p<0.01(0.05,0.10).

Table A.2. Logit regression results: last authors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Publication stock		0.009*** (0.002)	0.012*** (0.003)	0.011*** (0.003)	0.011*** (0.003)	0.010*** (0.003)	0.009*** (0.003)
Misconduct stock			2.882*** (0.400)	3.013*** (0.408)	3.046*** (0.418)	2.579*** (0.432)	2.482*** (0.450)
Coauthor stock				0.020* (0.011)	0.019* (0.011)	0.027** (0.011)	0.027** (0.011)
Impact factor stock					0.017 (0.048)	0.008 (0.047)	0.006 (0.047)
First retraction year						0.058*** (0.014)	0.073*** (0.024)
Total publication stock							0.011 (0.014)
Intercept	-1.666*** (0.077)	-1.930*** (0.110)	-2.297*** (0.128)	-2.503*** (0.173)	-2.559*** (0.233)	-3.109*** (0.276)	-3.707*** (0.830)
N	1264	1264	1264	1264	1264	1264	1264
AIC	1109.4	1098.9	1050.4	1049.1	1050.9	1035.9	1037.3
df	1	2	3	4	5	6	7
LR		12.44	62.93	66.29	66.42	83.51	84.09
ll	-553.782	-547.461	-522.218	-520.536	-520.472	-511.926	-511.638
Pseudo-R ²		0.01	0.06	0.06	0.06	0.08	0.08

Logit regression. Dependent takes value 0 when author was identified 1 time as misconducting author, and value 1 when the author was identified 2 times. All stocks apply 15% discount rate. Misconduct stock: stock of fraudulent publications divided by publication stock. Impact factor stock: stock of impact factor weighted publications divided by publication stock. Standard errors in brackets. Stars indicate p-value: ***(**,*): p<0.01(0.05,0.10).

Table A.3. $\hat{N}$ and 95% confidence interval for N obtained from Chao, Generalized Chao, and Modified Zelterman estimators. First authors.

Covariates	Chao	G. Chao	Zelterman	M. Zelterman
<i>null</i>	3119 [2627-3611]	3119 [2572-3666]	3334 [2782-3886]	3080 [2527-3632]
P		3358 [3161-3555]	3516 [2894-4139]	2802 [2179-3425]
P + TOTM		3781 [3534-4028]	3897 [3103-4692]	3635 [2740-4429]
P + TOTM + C		3893 [3631-4155]	3974 [3135-4813]	3493 [2654-4332]
P + TOTM + C + J		4104 [3813-4395]	4213 [3193-5351]	3565 [2546-4585]
P + TOTM + C + J + Y		4261 [3950-4572]	4368 [3285-5451]	4282 [3199-5364]
P + TOTM + C + J + Y + TOTP		4289 [3670-4908]	4392 [3293-5491]	4235 [3136-5334]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. Analysis with first authors. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.4: $\hat{N}$ and 95% confidence interval for N obtained from Chao, Generalized Chao, and Modified Zelterman estimators. Last authors.

Covariates	Chao	G. Chao	Zelterman	M. Zelterman
<i>null</i>	4458 [3933-4983]	4458 [3933-4983]	4014 [3484-4545]	4014 [3484-4545]
P		4613 [4414-4812]	4170 [3588-4753]	4026 [3444-4609]
P + TOTM		5182 [4923-5462]	4742 [3982-5502]	4110 [3350-4870]
P + TOTM + C		5201 [4995-5529]	4822 [3796-5849]	4113 [3319-4908]
P + TOTM + C + J		5269 [5001-5537]	4829 [4030-5628]	4113 [3315-4912]
P + TOTM + C + J + Y		5764 [5450-6077]	5325 [4300-6350]	4123 [3098-5148]
P + TOTM + C + J + Y + TOTP		5748 [5416-6080]	5309 [4296-6321]	4124 [3111-5136]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. Analysis with last authors. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Online Appendix

Table A.5: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators, subsample of authors with JIF premium > 4. First authors.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	1096 [823-1369]	1096 [823-1369]	1044 [791-1297]	1044 [791-1297]
2	P		1187 [1081-1294]	1137 [835-1438]	1065 [746-1366]
3	P + TOTM		1251 [1132-1371]	1201 [855-1547]	1069 [723-1415]
4	P + TOTM + C		1311 [1181-1442]	1261 [867-1655]	1069 [675-1463]
5	P + TOTM + C + J		2298 [1682-2914]	2248 [-369-4864]	1072 [-1545-3689]
6	P + TOTM + C + J+ Y		2302 [1685-2920]	2252 [-905-5409]	1072 [-1585-3729]
7	P+ TOTM + C + J+ Y + TOTP		2313 [1642-2984]	2263 [-415-4940]	1072 [-1605-3750]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTM the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.6: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators, subsample of authors with JIF premium ≤ 4 . First authors.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	2065 [1559-2571]	2065 [1559-2571]	1973 [1521-2424]	1973 [1521-2424]
2	P		2212 [2040-2383]	2120 [1595-2645]	1988 [1463-2513]
3	P + TOTM		2666 [2406-2925]	2575 [1793-3357]	2016 [1234-2798]
4	P + TOTM + C		2662 [2409-2915]	2571 [1792-3350]	2017 [1238-2796]
5	P + TOTM + C + J		2894 [2575-3213]	2804 [1870-3738]	2032 [1097-2966]
6	P + TOTM + C + J+ Y		3150 [2758-3542]	3060 [-2364-8483]	2039 [924-3155]
7	P+ TOTM + C + J+ Y + TOTP		3194 [1818-4569]	3104 [1942-4265]	2039 [877-3200]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTM the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.7: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators, subsample of authors with first publication up to 2003. First authors.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	1120 [902–1339]	1120 [902–1339]	1038 [816–1259]	1038 [816–1259]
2	P		1182 [1090–1274]	1100 [847–1353]	1048 [795–1301]
3	P + TOTM		1267 [1162–1372]	1186 [882–1489]	1056 [753–1360]
4	P + TOTM + C		1343 [1226–1460]	1261 [905–1618]	1059 [703–1416]
5	P + TOTM + C + J		1649 [1418–1880]	1568 [827–2310]	1066 [325–1808]
6	P + TOTM + C + J+ Y		1687 [1430–1944]	1606 [792–2421]	1067 [253–1882]
7	P+ TOTM + C + J+ Y + TOTP		1717 [1433–2001]	1636 [774–2499]	1067 [205–1930]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTM the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTP the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.8: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators, subsample of authors with first publication after 2003. First authors.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	2204 [1649–2760]	2204 [1648–2760]	2143 [1584–2703]	2143 [1584–2703]
2	P		2394 [2199–2588]	2333 [1670–2996]	2161 [1498–2823]
3	P + TOTM		2890 [2630–3150]	2830 [1846–3815]	2179 [1195–3163]
4	P + TOTM + C		2907 [2645–3170]	2847 [1851–3844]	2182 [1185–3178]
5	P + TOTM + C + J		2915 [2652–3178]	2855 [1850–3860]	2182 [1177–3186]
6	P + TOTM + C + J+ Y		3671 [3296–4046]	3611 [1978–5244]	2198 [565–3831]
7	P+ TOTM + C + J+ Y + TOTP		3719 [3209–4229]	3659 [1972–5347]	2199 [511–3886]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTM the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTP the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.9: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators, for natural sciences. First authors.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	289 [168–410]	289 [168–410]	281 [159–404]	281 [159–404]
2	P		299 [250–349]	282 [147–417]	292 [157–427]
3	P + TOTM		318 [263–373]	284 [125–442]	311 [152–469]
4	P + TOTM + C		328 [270–387]	284 [108–460]	321 [145–497]
5	P + TOTM + C + J		338 [275–401]	284 [87–481]	331 [134–528]
6	P + TOTM + C + J+ Y		358 [287–428]	284 [48–521]	350 [114–587]
7	P+ TOTM + C + J+ Y + TOTP		367 [39–695]	284 [29–540]	360 [104–615]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.10: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators, for natural sciences. Last authors.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	289 [266–381]	289 [168–410]	281 [159–404]	281 [159–404]
2	P		299 [250–348]	292 [157–427]	282 [147–417]
3	P + TOTM		318 [261–375]	311 [152–469]	284 [125–442]
4	P + TOTM + C		323 [266–381]	321 [145–497]	284 [108–460]
5	P + TOTM + C + J		338 [275–402]	331 [134–528]	284 [87–481]
6	P + TOTM + C + J+ Y		358 [276–439]	350 [114–587]	284 [48–521]
7	P+ TOTM + C + J+ Y + TOTP		356 [-1025–1759]	360 [104–615]	284 [29–540]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.11: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators. Subsample of first authors with first retraction before 2006.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	720 [560–880]	720 [561–811]	649 [487–811]	649 [487–811]
2	P		737 [671–803]	666 [493–839]	651 [478–824]
3	P + TOTM		822 [740–903]	751 [519–984]	661 [429–893]
4	P + TOTM + C		829 [746–912]	759 [519–998]	661 [422–901]
5	P + TOTM + C + J		991 [848–1133]	921 [466–1375]	668 [214–1122]
6	P + TOTM + C + J+ Y		1013 [856–1169]	942 [448–1437]	668 [173–1162]
7	P+ TOTM + C + J+ Y + TOTP		1019 [835–1203]	949 [444–1454]	668 [162–1173]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.12: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators. Subsample of first authors with first retraction between 2006 up to 2011.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	919 [680–1159]	919 [680–1159]	868 [626–1110]	868 [626–1110]
2	P		1028 [925–1131]	977 [666–1289]	884 [573–1195]
3	P + TOTM		1130 [999–1261]	1080 [686–1473]	890 [497–1283]
4	P + TOTM + C		1222 [1075–1369]	1172 [696–1647]	892 [416–1367]
5	P + TOTM + C + J		1222 [1075–1370]	1172 [696–1647]	892 [416–1367]
6	P + TOTM + C + J+ Y		1232 [1081–1382]	1181 [695–1667]	892 [405–1378]
7	P+ TOTM + C + J+ Y + TOTP		1246 [992–1500]	1195 [701–1690]	894 [399–1389]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.

Table A.13: $\hat{N}$ and 95% confidence interval obtained from Chao, Generalized Chao, and Modified Zelterman estimators. Subsample of first authors with first retraction from 2011 onwards.

Model	Covariates	Chao	G. Chao	Zelterman	M. Zelterman
1	<i>null</i>	1852 [1208–2496]	1852 [1208–2496]	1832 [1185–2478]	1832 [1185–2478]
2	P		2085 [1872–2297]	2064 [1259–2870]	1861 [1055–2666]
3	P + TOTM		2183 [1947–2420]	2163 [1266–3061]	1862 [964–2759]
4	P + TOTM + C		2182 [1947–2418]	2162 [1265–3060]	1862 [964–2759]
5	P + TOTM + C + J		2222 [1978–2466]	2202 [1248–3155]	1862 [909–2815]
6	P + TOTM + C + J+ Y		2751 [2409–3093]	2731 [1137–4325]	1866 [272–3459]
7	P+ TOTM + C + J+ Y + TOTP		2769 [2227–3311]	2749 [1177–4321]	1867 [294–3439]

Table reports population size estimates $\hat{N}$ for model specifications as estimated through Chao, Generalized Chao, Zelterman, and Modified Zelterman estimators. 95% confidence intervals in square brackets. P denotes the publication stock, C the coauthor stock, J the journal impact factor premium, defined as the JIF-weighted publication stock divided by the publication stock, TOTP the total number of publications in medical research according to Scopus, Y the year of the first retraction, and TOTM the number of retractions in the previous year divided by the publication stock. Stock variables are constructed using a 15% annual depreciation rate. All regressions additionally include year-of-retraction fixed effects. TOTM is lagged by one year.



Download ZEW Discussion Papers:

<https://www.zew.de/en/publications/zew-discussion-papers>

or see:

<https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html>

<https://ideas.repec.org/s/zbw/zewdip.html>



IMPRINT

**ZEW – Leibniz-Zentrum für Europäische
Wirtschaftsforschung GmbH Mannheim**

ZEW – Leibniz Centre for European
Economic Research

L 7,1 · 68161 Mannheim · Germany

Phone +49 621 1235-01

info@zew.de · zew.de

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.