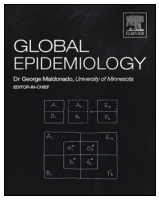
Global Epidemiology 8 (2024) 100154

Contents lists available at ScienceDirect

Global Epidemiology

journal homepage: www.sciencedirect.com/journal/global-epidemiology



Commentary

Adjustment for sparse data bias in odds ratios: Significance to appraisal of risk of diabetes due to occupational trichlorfon insecticide exposure

Igor Burstyn [a,*], David Miller [b]

[a] Department of Environmental and Occupational Health, Dornsife School of Public Health, Drexel University, Philadelphia, PA 19104, USA
[b] US Environmental Protection Agency (Retired), USA

A B S T R A C T

Background: Bias away from the null in odds ratios (OR), aggravated by low power, is a well-known phenomenon in statistics (sparse data bias). Such bias increases in presence of selection of "significant" results on the basis of null hypothesis testing (effect size magnification, ESM).
Objectives: We seek to illustrate these issues and adjust for suspected sparse data bias in the context of a reported more than doubling of the odds of new onset type 2 diabetes in presence of occupational trichlorfon insecticide exposure reported in the Agricultural Health Study.
Methods: We performed ESM analysis on the crude ORs extracted from the contingency table in the published report, which is done by simulating selected OR given a posited true OR. Next, we applied easily accessible methods that adjust for sparse data bias to the extracted contingency tables, including data augmentation, bootstrap, Firth's regression, and Bayesian methods with weakly informative priors.
Results: During the ESM analysis, we observed that there was a reasonable chance that a "statistically significant" OR of around 2.5–2.6 would be observed for true OR of 1.2. Adjustment for sparse data bias revealed that Bayesian methods outperformed alternative approaches in terms of yielding more precise inference, while not making unjustified distributional assumptions about estimates of OR. The OR in the original paper of about 2.5–2.6 was reduced on average to OR of 1.9 to 2.2, with 95% (Bayesian) credible intervals that included the null.
Discussion: It is reasonable to adjust ORs for sparse data bias when the reported association has societal importance, because policy must be informed by the least biased estimates of the effect. We think that such adjustment would lead to a more appropriate evaluation of the extent of evidence on the contribution of occupational exposure to trichlorfon pesticide to risk of new onset diabetes.

Introduction

Bias away from the null in odds ratios (OR) obtained by maximum likelihood (ML) methods in logistic regression is a well-known phenomenon in statistics [1]. However, its significance has not penetrated epidemiologic practice despite valiant efforts that date back at least decades, [2] with accessible overviews produced by Cole et al. [3] and Greenland et al. [2]. Greenland et al. [2] advocate a Bayesian approach via data augmentation within standard statistical software [4,5], but other implementations are now easily accessible [6]. The data augmentation and semi-Bayesian approaches are favored over Firth's regression that implies an unrealistic prior. [7] Bootstrap is also an option, although the recommended quadratic bootstrap is not trivial to implement in standard software [8].

Sparse data bias is another phenomenon which is similarly not often recognized by epidemiologists. It will further exacerbate the bias away from the null in OR obtained by ML methods since regression estimates may not be well-behaved when the number of events is small [2]. Sparse data bias in OR is proportional to the observed effect size estimate and a measure of information contained in the data that is related to sample size. [9] Berkson [1] observed that the ML estimate exhibited more pronounced bias with deviation from prevalence of events away from 50%, implying that it is worse for both "rare" and "common" exposures and that ML methods will further exacerbate sparse data bias. "Rare" and "common" exposures also happen to be conditions where statistical power degrades. It is essential to note that the sparse data bias is largely driven by the combination of covariates (cell of contingency table at its simplest) that contains the least data (low counts). Therefore, a study with a very large number of subjects and no measurement errors or other biases can still be biased by sparse data away from the null (for additional details, see Supplemental Materials **1**).

Finally, sparse data bias can itself aggravate issues associated with effect size magnification (ESM). ESM is likewise a known phenomenon in statistics [10,11] and occurs when results are deemed important on the basis of rejection of a null hypothesis or exceeding a pre-determined effect size. In practical terms, ESM means that low-powered studies that find evidence of an effect often provide inflated estimates of the size of that effect that are (consequently) less able to be replicated. These

* Corresponding author at: Nesbitt Hall, Room 614, 3215 Market Street, Philadelphia, PA 19104, USA.
E-mail address: igor.burstyn@drexel.edu (I. Burstyn).

https://doi.org/10.1016/j.gloepi.2024.100154
Received 7 March 2024; Received in revised form 1 July 2024; Accepted 1 July 2024
Available online 8 July 2024

phenomena have been held responsible for the facetious claims in the popular press that "the truth wears off" [12]. Ioannidis [13] summarized that "[*effect size*] inflation is expected when, to claim success (discovery), an association has to pass a certain threshold of statistical significance, and the study that leads to the discovery has suboptimal power to make the discovery at the requested threshold of statistical significance." To avoid being misled by an artificially inflated effect size in early discovery, Ioannidis [13] offers several suggestions including "rational down-adjustment of effect sizes" and "analytical methods that correct for anticipated inflation". However, he does not offer any specific tools for either investigating the extent or magnitude of ESM in published (epidemiological) data or for adjusting study results to account for such effects, choosing to emphasize instead the importance of these considerations in the planning stages of studies. Indeed, unless the whole sequence of model selection is documented, it is difficult to determine when biasing selection decisions occurred. Nonetheless, it is possible to consider the need to shrink effect estimates that are suspected of being positively biased regardless of the exact mechanism by which this occurred.

This issue of bias in OR is significant because ORs of 2 or greater with 95% confidence intervals (CI) that excludes 1 is seen by some courts as important for arguing general causation, [14] i.e., has immediate real-world consequence for some litigants, and has been an element of causal arguments within epidemiology [15]. As another example, the US EPA's Office of Pesticide Programs uses a "statistically significant" OR of 2.0 as the demarcation between classifying as study as having "evidence of a positive association" ($1.30 \leq OR < 2$ and $p < 0.05$) between exposure and health outcome and "evidence of a moderately strong positive association" ($2.0 \leq OR < 3.0$ and $p < 0.05$) between exposure and health outcome (e.g. see statements to this effect on page 11 of [16]). Different authors proposed other rules of thumbs for judging and characterizing effect sizes [17–19]. If the elevation of an OR can be ascribed to such sparse data bias, it is imperative to consider and ideally account for this to guard against ensuing "false positives".

It is important to note that low power can cause sparse data bias even in absence of added selection on the basis of null hypothesis tests and thresholds for effect sizes (i.e., ESM). That is, each can act separately and independently to introduce an artificial inflation of observed effect sizes, with these phenomena, introducing a systematic, insidious, and often unrecognized bias away from the null. In fact, this bias away from the null from sparse data bias and ESM can be sufficient to entirely counter and reverse what some epidemiologists claim as bias to the null under conditions of non-differential misclassification. Apparent low power leading to false positives can also be the consequence of other unmodelled problems in statistical analysis, like exposure misclassification [20].

As a motivating example for this paper, we draw on a report of more than doubling of the odds ($p < 0.05$) of new onset type 2 diabetes attributed to exposure to insecticide trichlorfon by Montgomery et al. [21], which, despite being based on a large cohort (>30,000 subjects), is driven by 7 "highly" exposed cases. Given the likely low power of the study due to the very small number of exposed cases, we suspect that this effect size estimate is inflated relative to the true value and would not replicate in any future similar study or subsequent analysis. Curiously, Montgomery et al. [21] reported a largely null association of diabetes with active form of trichlorfon, dichlorvos: an adjusted OR of 1.26 (95% CI 0.91, 1.73) based on 44 cases in the highest exposure category vs. never-exposed, p-trend 0.15. Juntarawijit and Juntarawijit [22] conducted a case-control study of diabetes using exposure assessment method "derived" from that of Montgomery et al. [21]. Based on 10 ever-exposed cases, they reported for dichlorvos an adjusted OR of 1.03 (95% CI: 0.41, 2.62) [22]. Thus, the initial characterization of the finding by the authors as a "strong" [21] positive result for trichlorfon is not internally consistent and was not replicated in a study with similar number of exposed cases. There appear to be no other reports evaluating the association of either dichlorvos or trichlorfon with type 2 diabetes in humans. We further note that a more than doubling (here, OR of 2.5–2.6) of the OR for diabetes as a result of exposure to trichlorfon would be quite dramatic given that family history is known to play a sizable role in the onset of diabetes and this is observed to have a similar magnitude of effect as reported here. For example, a EU-wide prospective cohort found that a family history of type 2 diabetes was associated with on average doubling to tripling of incidence of the condition [23]. It would be unusual for a typical environmental exposure to have an effect size on diabetes that is of similar magnitude to that of family history and genetics.

More generally, we wish to illustrate how summary results as are typically presented in published studies can be used by the reader to (a) assess and evaluate effect size magnification and (b) adjust such results for sparse data bias, assuming that the two phenomena have low power in common. Further, we also compare performance of these adjustment methods for the motivating example of reported associations [21]. We do not purport to develop new methods but, as a secondary aim, consider the wisdom of default priors on odds ratios that have been proposed by others in [2,6,7].

Methods

We extracted 2-by-2 tables from Table 3 of Montgomery et al. [21] for trichlorfon. We considered the original (study authors') classification into "high" exposure group (7 exposed cases) as well as ever exposed group that we created (12 exposed cases). We limited our attention to subjects who had complete information to mimic the main analysis, thereby losing one exposed case and focusing our illustrative analysis on the crude odds ratio for categories of "high" vs. never- and "ever" vs. never-exposure (see Supplemental Materials **1,** Table S1). As described below, we then evaluated, in turn, the extent to which effect size magnification might account for these results and how these results might be adjusted, *post-hoc*, for sparse data bias.

Effect size magnification (ESM) analysis

As a first step in any evaluation of a 2-by-2 epidemiological table, an ESM analysis can be useful to determine the extent to which OR for "discovered" associations may be inflated due to low power. The concepts behind ESM analyses are more thoroughly explained elsewhere [13,24–29]. Briefly, the analysis begins by assuming a true OR for an association and estimates the proportion of exposed individuals among n_0 in a non-diseased group (P_o). The expected proportion of exposed individuals among n_1 cases (P_1) is then estimated based on the assumed true OR and P_o. Exposed non-diseased subjects are simulated using Bin (P_o, n_0) and exposed diseased are independently sampled from Bin(P_1, n_1). For each of these simulated studies, the simulated OR* is then computed via ML method. When looking at only those OR* that pass the threshold of the lower 95% CI exceeding the null (equivalent to one-sided test with $p = 0.025$), their median is compared to the true OR, yielding the ratio termed the magnification factor. Our choice of threshold of selection is motivated by the context of pesticide risk assessment and likely other environmental xenobiotics, where effect estimates suggesting a protective effect (e.g., statistically significant OR < 1) are dismissed as statistical anomalies, while only apparently elevated effect estimates are considered for interpretation as potentially causal. All calculations were implemented in Stata (College Station, TX) using its -emagnification- command and – for comparison purposes – in SAS (Cary, NC) and are shared as Supplemental Materials **1**.

Adjustment of OR for sparse data bias

We assumed both no measurement error in any of the variable (which is unrealistic for exposure at the very least [30,31] but we will tackle this elsewhere) and no important confounding that would make confounder adjusted and unadjusted analysis materially different.

When sparse data bias (e.g., as might be potentially identified in ESM) results in a likely problematic interpretation of a published OR, a variety of methods are available for adjusting the ML estimates of an OR produced by logistic regression. These include bootstrap of ML estimates (we used 10,000 replicates with basic method for CI), profile likelihood for CI, Firth's logistic regression, Bayesian estimates with null-centered priors that shrink the OR, and data augmentation to penalize extreme estimates. For this latter method, we implemented data augmentation as described by Greenland et al. [2] that is equivalent to a prior on OR to fall between 1/40 to 40 with 95% certainty; both ML and profile likelihood CIs were generated.

All these analyses can be implemented in standard statistical software, with the exception of Bayesian methods, which are easy to conduct for 2-by-2 contingency tables in any environment amendable to Monte Carlo sampling (e.g. see section 5.1 of Gustafson [32], where implementation is adopted by assuming no exposure misclassification). In implementing this Bayesian approach, we took advantage of the fact that odds of probability π following a Beta-Prime distribution parameterized by shapes (α,β), which is a conjugate distribution of $\pi \sim \text{Beta}(\alpha,\beta)$ [33]. This allowed us to directly sample from posterior odds from Beta-Prime distributions for cases and controls. We also implemented Bayesian logistic regression with default normal prior on log(OR) suggested by Hamra et al. [6], which presumes that a priori we believe that true OR lies with 95% certainty between 1/10 and 10 (Bayesian-H); it has an advantage of easy implementation in SAS (Cary, NC).

Only the matter of priors requires further explicit elaboration. In what we call Bayesian methods A and B, we induced a prior on the OR by positing identical and independent priors on prevalence of exposure in cases and non-cases via Beta distributions. In the Bayesian-A method, we set prior prevalences to be on average close to that of the prevalence of exposure in cases, and in Bayesian-B method – that of the prevalence of exposure in controls. For example, if we wanted a prior on prevalence centered on 1%, we use Beta($s \times 1$, $s \times 99$), where s is a scaling constant described further below. Then, the posterior of odds becomes Beta-Prime($s \times 1 + a$, $s \times 99 + b$) for observed prevalences of exposure of a/(a + b). This proved to be suitable for Bayesian-A for both high vs. never-exposed and ever vs. never- analyses because the expected exposure prevalence of cases was about 1% in both scenarios. For Bayesian-B, exposure prevalences in controls differed between the two exposure contrasts (0.5% vs 0.3%) and Beta-priors on prevalence were suitably adjusted. For high vs. never-exposed comparison in Bayesian-B, the prior prevalence had 95% certainty interval of about 0.03% to 0.8% with median and mean 0.3%, apparently covering the full range of possibilities in the cohort (note that the apparent prevalence of exposure in the exposed is 0.7%). The priors on OR induced in this manner have means of 1 (null) and the scaling constant s was used to tune the prior on prevalence to induce 95% certainty interval of priors on OR to match that of Hamra et al., [6] (this tuning could have been done for any other target prior precisions).

Most of the sparse data calculations were implemented in R [34] and are shared as Supplemental Materials **2** and **3**, except for Bayesian-H that was implemented in SAS (Cary, NC), see Supplemental Material **4**.

Results

Effect size magnification analysis

From the data reported in the Montgomery et al. [21], crude ORs of 2.62 (95% CI: 1.20, 5.70) and 2.48 (95%CI: 1.37, 4.49) can be calculated for the high vs. ever- and ever vs. never-exposed comparisons, respectively. (The corresponding adjusted OR for the high vs. never-exposed comparison reported in [21] is 2.47, 95%CI 1.10, 5.56).

Table 1 provides evidence that for the high vs. never-exposed comparison for any true OR substantively <3, there is a high potential for the discovered OR of 2.62 to be due to effect size inflation. For example, with a true OR of 1.1 and the observed background exposure prevalence of 0.28%, the expected median "statistically significant" OR will be 2.59, representing an inflation of 136%. At that assumed true odds ratio of 1.1, the observed crude OR of 2.62 is well within the middle 80% of the expected range of 2.3 to 3.4. Thus, such an estimated OR would not at all be unexpected due to ESM if the true OR was only 1.1. It is noteworthy that even when the assumed true OR of 3 for this high vs. ever comparison, power is still only at 75%, below the nominally desired 80%, with the median "selected" OR inflated to 3.3.

Table 1
Simulations for effect sizes passing the chosen threshold of statistical significance (the lower 95% confidence limit above 1) for Montgomery et al. [21]; simulation size 10,000.

True Odds Ratio	Observed odds ratios among "selected" results					
	"High" vs. Never Exposed			Ever vs. Never Exposed		
	$P_0 = 0.28\%$			$P_0 = 0.50\%$		
	Median (p10-p90#)	*Median Fold Inflation*	*Power (%)*	*Median (p10-p90#)*	*Median Fold Inflation*	*Power (%)*
1.0	2.59 (2.31–3.24)	1.59	3.2	2.16 (1.99–2.56)	1.16	2.9
1.1	2.59 (2.31–3.35)	2.36	5.4	2.13 (1.96–2.55)	1.93	5.6
1.2	2.59 (2.31–3.21)	2.16	7.8	2.16 (1.97–2.62)	1.78	8.8
1.5	2.67 (2.31–3.45)	1.78	17	2.24 (1.97–2.83)	1.49	23
2.0	2.85 (2.34–3.86)	1.42	36	2.42 (2.02–3.18)	1.21	53
3.0	3.31 (2.47–4.77)	1.10	75	3.05 (2.24–4.16)	1.02	92

Similarly, Table 1 illustrates the potential for ESM for an ever vs. never-exposed comparison which we created by combining the study's 5 "low"- and 7 "high"- exposed cases. The background exposure prevalence for this comparison was 0.50%. We estimated that the median inflation can vary from a substantial 1.93-fold for a true OR of 1.1 to a trivial 1.02-fold for a true OR of 3.0. This means that the study's reported OR of 2.47 for the association between trichlorfon exposure and incident diabetes [21] could arise with non-ignorable probability from a true OR of as little as 1.1 due to ESM alone, if there were no other sources of error or bias (which almost never can be ruled out with certainty in occupational epidemiology). Only when we assume true OR of 3 for this comparison, did the power exceed the nominally desired 80%, and the median "selected" OR was not importantly inflated.

The above analyses are dependent at least in part on the observed background exposure prevalence among the non-diseased. Hence, it is important to examine impact of uncertainty in the assumption of prevalence of exposure in non-cases. This sensitivity analysis is detailed in Figs. S1 and S2 in Supplementary Materials 1 where it is shown that its impact does not materially affect our conclusions.

Finally, we note our observation that distribution of plausible log(*OR*)s under the null, given the data structure of the ESM analysis we undertook, is multimodal and poorly approximated by Gaussian distribution, signaling that any inference that relies on assumption of asymptotic normality of log(OR) is questionable in the studied circumstances of very low exposure prevalence (see Fig. S3 in Supplementary Materials 1).

Sparse data bias adjustment

Table 2 contains the results of the adjustment for sparse data bias. There is evidence that ML estimate of OR is affected by sparse data bias because OR point estimates are reduced, and the 95% CI shifted to lower range when ML estimates are bootstrapped and when data is augmented to penalize extreme estimates. This is more evident for the sparser "high" vs. never exposed comparison in data augmentation, where point

Table 2
Adjustment for sparse data bias in the association of incident diabetes and trichlorfon insecticide reported in Montgomery et al., [21].

Odds ratio estimation method	Estimated odds ratio (OR) and 95% interval for different exposure contrasts			
	"High" vs. Never Exposed (>10 cumulative days) (cumulative days of use, 7 exposed cases)		Ever vs. Never Exposed (12 exposed cases)	
	OR	95% (CI/CrI)[b]	OR	95% (CI/CrI) [b]
Maximum likelihood (ML) [a]	2.62 [c]	1.20, 5.70	2.48	1.37, 4.49
Bootstrap of ML (basic)	2.62 [d]	0.23, 4.42	2.50	0.83, 3.80
Profile likelihood (PL)	2.62	1.09, 5.31	2.48	1.30, 4.30
Firth's	2.79	1.20, 5.54	2.57	1.36, 4.42
Data augmentation (ML)	2.45	1.13, 5.31	2.38	1.32, 4.31
(PL)	2.45	1.04, 4.96	2.38	1.26, 4.13
Bayesian-A [e] (prior on exposure prevalence akin to cases)	2.65	1.22, 5.07	2.33	1.27, 3.93
Bayesian-B [f] (prior on exposure prevalence akin to controls)	1.94	0.88, 3.70	2.01	1.10, 3.39
Bayesian-H [g] (not informed by exposure prevalence)	2.24	0.99, 4.60	2.26	1.21, 3.99

a: The usual default method as deployed by Montgomery et al. [21]
b: CI = confidence interval, CrI = credible interval.
c: crude OR while the adjusted OR in Montgomery et al. [21] is 2.5, 95%CI 1.1, 5.6. uncertainty is in the first decimal place.
d: median 2.55; inference challenging due to very high skew (−10) and kurtosis (214).
e: prior OR derived from expected exposure prevalence 1%, akin to that among cases.
f: prior OR derived from expected exposure prevalence 0.5% for ever vs never exposed and 0.3% for high vs never exposed, akin to that among controls.
g: the default prior on OR recommended by Hamra et al. [6]

estimates declined from 2.62 to 2.45 and the lower 95% limit reduced from 1.20 to 1.04 (profile likelihood). Profile likelihood and bootstrap showed similar trends on average, but bootstrap had 95% lower confidence limit below 1, signaling a more skewness with heavier tails than that allowed under the assumptions of normality of log(OR). Specifically, for the ever vs. never-exposed comparison, bootstrap sample had skew = −0.7 and kurtosis = 4; the situation is far worse for the high vs. never-exposed comparison (see footnote to Table 2). The Firth's logistic regression appeared ineffective in adjusting for bias in this case, as it yielded here higher estimates of OR than ML, data augmentation, and bootstrap. All Bayesian methods displayed similar performance, pushing the OR towards lower values, except in the case of Bayesian prior anchored in exposure prevalence among cases for "high" vs. never exposed. The downward adjustment was most pronounced with Bayesian-B method that used prevalence of exposure among controls (median posterior OR 1.94, 95% credible interval (CrI): 0.88 to 3.70) and the default prior of Hamra et al., [6] (median posterior OR 2.24, 95%CrI: 0.99 to 4.60). There is a pronounced skewness without heavy tails in deviation from the normality of posterior of OR in methods that manipulate prior on exposure prevalence (e.g., skew = −0.2 to −0.3, kurtosis = 3 for Bayesian-A and -B). This, together with observation of similar phenomena for bootstrap, signals that the usual assumption of normality of log-OR may not be tenable in this analysis.

It is difficult to be certain which of the adjustments is the most credible. The linear bootstrap that we employed seems to have inherited, unsurprisingly, biases that are a property of ML estimate of the logistic regression [1], as signaled by heavy tails. The recommended quadratic bootstrap [8] is not readily accessible to analysts. Thus, we think that data augmentation, Bayesian-B, and Bayesian-H [6] are the most plausible candidates for being the most credible. They seem to converge on similar conclusions, even though we question whether the Greenland et al.'s suggested default prior on OR of data augmentation [2] (i.e., of prior OR falling between 1/40 to 40 with 95% certainty) is suitable in our application. The Bayesian methods A and B are attractive in that they do not force a normality assumption on the log(OR) and do not rely on logistic regression (unlike bootstrap).

Bayesian-B yields more precise estimates than that with default prior of Hamra et al. [6]: smaller ratios of 97.5 to 2.5 percentiles of posterior ORs. Therefore, we are tempted to favor more precise estimates [35] and to bet that the least biased estimate of OR for "high" vs. never-exposed comparison is centered between 1.9 and 2.2, with the 95% CrI's from about 0.9 to 4 or 5. However, there is uncertainty due to modeling choices. Our conclusion appears to be importantly different from the one in the original manuscript that asserted "the pesticide most strongly associated with diabetes among applicators was the organophosphate insecticide trichlorfon" [21] on the basis of adjusted OR 2.47 (95%CI 1.10, 5.56).

Discussion

Regulatory and other risk assessors should be aware of the bias away from the null when reviewing epidemiological results that suggest a discovered association (e.g., those that cross a pre-determined statistical threshold such at $p < 0.05$ or exceed a pre-determined OR of interest). An ESM analysis can be useful to determine the extent to which the OR may be artificially inflated, which is more likely to occur in underpowered studies. Studies might be underpowered when they are small, prone to measurement error, or they are susceptible to sparse data bias. Gelman and Carlin [26] specifically recommend that such ESM-like calculations (termed "post-hoc design calculations" by them) be done when strong statistically significant evidence for non-null effects have been found because:

> "a [*discovered statistically*] significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect when researchers study small effects using noisy measurements and small sample sizes".

Gelman and Carlin [26] continue, stressing that such calculations may be even more relevant for findings that are found to be statistically significant because the interpretation of such result can change substantially depending on the researchers' prior belief in a plausible size and direction of the true effect. As described earlier, implementation is available in common statistical software to evaluate the potential extent of ESM and determine if sparse-data bias corrections or adjustments might be advisable.

We are not aware of any research that suggests that the application of the methods to adjust for sparse data bias can lead to negating true association, underestimating them. One can be comfortable in this being generally the case because methods that we apply (certainly all the Bayesian ones), at the mechanistic level of the calculations, aim to eliminate extreme values, not to directly penalize inflation of ORs expected in ML estimates of logistic regression. Please note that prior on ORs are all centered on the null; if this was not the case, then there would indeed but push towards higher or lower values in the posterior, but this is not the case in our application and all such applications of Bayesian methods to combat sparse data bias. Thus, even though we are motivated by trying to safeguard against overestimating effects sizes, the methods equally penalize high and low effect size estimates, without any obvious risk of making true associations disappear.

If it is decided that such adjustments are necessary or advisable, it is easy to adjust published ORs for bias arising from sparse data when it is manifested in unadjusted analyses when data is captured in the original publication by a contingency table. This should both enhance interpretation of individual studies and aid their pooling in meta-analyses of studies that are invariably of different size and likely include studies for

which sparse data bias may be important.

If observed confounding is negligible, then little information is lost from working with contingency tables as opposed to record-level data. Extension of our work to account for confounders is possible with a little extra effort as shown by Hamra et al. [6] Of course, the extension to account for other sources of bias, such as measurement errors, selection bias, latent confounding, may be warranted for more reliable inferences; methods to do so should be possible which would also benefit from combining with quantitative bias analysis [36], by adding an adjustment for sparse data bias. For problems that are of considerable importance, full Bayesian analysis that accounts for all sources of bias believed to be important may be recommended, albeit this would require more than application of easily accessible statistical routines.

Analysis that adjusts for bias would still need to be subjectively interpreted in the context of regulatory decisions and care must be taken to not apply the same rules of thumb to bias-adjusted results as one would apply to results where more bias is suspected as being at play. In a sense, the penalty which mentally (perceptually) down-weights the importance of studies for suspected bias should not be kept after quantitative adjustment for such a bias is performed. For example, a bias-adjusted OR of 1.5 in a low-powered study subject to sparse data bias can be interpreted as causal with more confidence compared to crude OR of 2.0 even if they are estimated with about equal precision.

Our experience indicates that Bayesian approach with priors that account for prevalence of exposure does not manifest any obvious advantages: while it yields more precise estimates, it is sensitive to the choice of prior exposure prevalences even within a narrow range. While we are skeptical about default priors on ORs, there is no obvious way to pick exposure prevalence that is used to induce a null-centered prior on OR. Using the observed exposure prevalence as a guide to set such a prior is clearly problematic on the theoretical grounds, because the prior must be elucidated before observing the data. It seems less perilous to set the prior on ORs for binary exposure directly via the default prior of Hamra et al. [6]. And yet if there is prior information on prevalences (as is often that case, setting aside worries about exposure misclassification [37]), one can perhaps learn more from the data (as indicated by narrower credible intervals with Bayesian-B). Furthermore, it is appealing to not make assumption of normality of log(OR) which one is forced to make in approach of Hamra et al. [6] when ESM indicates that such an assumption is dubious. In such a setting, bootstrap and Bayesian methods that postulate an informative prior on prevalence of exposure gain appeal. This remains an interesting avenue for research, even if it proves to be a matter of little practical consequence.

It is reasonable to adjust ORs for sparse data bias when the reported association has societal significance, because policy must be informed by the least biased estimates of the effect. We think that such adjustment would lead to a more appropriate evaluation of the extent of evidence on the contribution of occupational exposure to trichlorfon pesticide to risk of new onset diabetes, which seems to have been originally overstated.

CRediT authorship contribution statement

Igor Burstyn: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David Miller:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gloepi.2024.100154.

References

[1] Berkson J. Maximum likelihood and minimum chi2 estimates of the logistic function. J Am Stat Assoc 1955;50(269).
[2] Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. BMJ 2016;352:i1981.
[3] Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. Am J Epidemiol 2014;179(2):252–60.
[4] Sullivan SG, Re Greenland S, Sullivan SG, Greenland S. Bayesian regression in SAS software. Int J Epidemiol 2013;42:308–17. Int J Epidemiol 2014;43(3):974.
[5] Sullivan SG, Greenland S. Bayesian regression in SAS software. Int J Epidemiol 2013;42(1):308–17.
[6] Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems-using weakly informative bayesian priors. Epidemiology 2013;24(2):233–9.
[7] Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Stat Med 2015;34(23):3133–43.
[8] Claeskens G, Aerts M, Molenberghs G. A quadratic bootstrap method and improved estimation in logistic regression. Stat Prob Lett 2003;61:383–94.
[9] Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol 2009;9:56.
[10] Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14(5):365–76.
[11] Reinhart A. Statistics Done Wrong. The woefully complete guide. San Francisco, CA: No Starch Press; 2015.
[12] Lehrer J. The truth wears off: is there something wrong with the scientific method. New York 2010 December 5;2010.
[13] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19(5):640–8.
[14] Carruth RS, Goldstein BD. Relative risk greater than two in proof of causation in toxic tort litigation. Jurimetrics 2001;41(2):195–209.
[15] Doll R. Proof of causality: deduction from epidemiological observation. Perspect Biol Med 2002;45(4):499–515.
[16] Niman AL. Paraquat Dichloride: Tier II Epidemiology Report. In: Programs. UEPAOoP; 2019. p. 11.
[17] Oleckno WA. Essential epidemiology: Principles and applications. 1 ed. Prospect Heights, IL, USA: Waveland Press; 2002.
[18] Monson RR. Occupational Epidemilogy. 2 ed. CRC Press; 1990. p. 312.
[19] Grimes DA, Schulz KF. False alarms and pseudo-epidemics: the limitations of observational epidemiology. Obstet Gynecol 2012;120(4):920–7.
[20] Burstyn I, Yang Y, Schnatter AR. Effects of non-differential exposure misclassification on false conclusions in hypothesis-generating studies. Int J Environ Res Public Health 2014;11(10):10951–66.
[21] Montgomery MP, Kamel F, Saldana TM, Alavanja MC, Sandler DP. Incident diabetes and pesticide exposure among licensed pesticide applicators: agricultural health study, 1993-2003. Am J Epidemiol 2008;167(10):1235–46.
[22] Juntarawijit C, Juntarawijit Y. Association between diabetes and pesticides: a case-control study among Thai farmers. Environ Health Prev Med 2018;23(1):3.
[23] InterAct C, Scott RA, Langenberg C, Sharp SJ, Franks PW, Rolandsson O, et al. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the EPIC-InterAct study. Diabetologia 2013;56(1):60–9.
[24] Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology 2011;22(4):450–6.
[25] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124.
[26] Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci 2014;9(6):641–51.
[27] Lu J, Qiu Y, Deng A. A note on type S/M errors in hypothesis testing. Br J Math Stat Psychol 2019;72(1):1–17.
[28] Yarkoni T. http://wwwtalyarkoniorg/blog/ [Internet] [cited 2019]. Available from: https://talyarkoni.org/blog/2009/11/21/ioannidis-on-effect-size-inflation-with-guest-appearance-by-bozo-the-clown/; 2009.
[29] Miller DJ, Nguyen JT, Bottai M. Emagnification: a tool for estimating effect-size magnification and performing design calculations in epidemiological studies. Stata J 2020;20(3):548–64.
[30] Blair A, Tarone R, Sandler D, Lynch CF, Rowland A, Wintersteen W, et al. Reliability of reporting on life-style and agricultural factors by a sample of participants in the agricultural health study from Iowa. Epidemiology 2002;13(1):94–9.
[31] Lash TL. Bias analysis applied to agricultural health study publications to estimate non-random sources of uncertainty. J Occup Med Toxicol 2007;2:15.
[32] Gustafson P. Measurement error and misclassification in statistics and epidemiology2004. Chapman & Hall/CRC Press; 2004.
[33] Johnson NL, Kotz S, Balakrishnan N. Beta distributions. Continuous univariate distributions. 2. 2 ed. Wiley; 1995. p. 284.
[34] Team RDC. R: A language and environment for statistical computing. ISBN 3-900051-07-02006. Vienna, Austria: R Foundation for Statistical Computing; 2006.
[35] Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2001;12(3):291–4.
[36] Lash TL, Fox MP, Fink AK. Applying quantitative Bias analysis to epidemiologic. Data: Springer; 2021.
[37] MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? Epidemiology 2012;23(1):151–8.