Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 1896; doi:10.3390/ijerph16111896

*Article*

# What to do when accumulated exposure affects health but only its duration was measured? A case of linear regression.

**Igor Burstyn [1]*, Francesco Barone-Adesi [2], Frank de Vocht [3] and Paul Gustafson [4]**

[1] Department of Environmental and Occupational Health, Dornsife School of Public Health, Drexel University; ib68@drexel.edu
[2] Department of Pharmaceutical Sciences, University of Eastern Piedmont, Novara; francesco.baroneadesi@uniupo.it
[3] Population Health Sciences, Bristol Medical School, University of Bristol; frank.devocht@bristol.ac.uk
[4] Department of Statistics, The University of British Columbia; gustaf@stat.ubc.ca
* Correspondence: ib68@drexel.edu; Tel.: 1-267-359-6062

**Abstract:** *Background*: We considered a problem of inference in epidemiology when cumulative exposure is the true dose metric for disease, but investigators are only able to measure its duration on each subject. *Methods*: We undertook theoretical analysis of the problem in the context of a continuous response caused by cumulative exposure, when duration and intensity of exposure follow log-normal distributions, such that analysis by linear regression is natural. We present a Bayesian method to adjust duration-only analysis to incorporate partial knowledge about the relationship between duration and intensity of exposure and illustrate this method in the context of association of smoking and lung function. *Results*: We derive equations that (a) describe under what circumstances bias arises when duration of exposure is used as a proxy of cumulative exposure, (b) quantify the degree of such bias and loss of precision, and (c) describe how knowledge about relationship of duration and intensity of exposure can be used to recover an estimate of the effect of cumulative exposure when only duration was observed on every subject. *Conclusions*: Under our assumptions, when duration and intensity of exposure are either independent or positively correlated, we can be more confident in qualitatively interpreting the direction of effects that arise from use of duration of exposure per se. To make reliable inference about the magnitude of effect of cumulative exposure on the outcome, we can use external information on the relationship between duration and intensity of exposure even if intensity of exposure is not available at the individual level.

**Keywords:** Measurement Error, Dose-Metric, Bayes, Cumulative Exposure

## 1. Introduction

We considered a problem of inference in epidemiology when cumulative exposure is the true dose metric for disease, but investigators are only able to measure its duration on each subject. We nest most of our presentation within the context of occupational and environmental epidemiology, while recognizing that the issue also arises in other sub-disciplines of epidemiology. This problem was first highlighted by Johnson who observed that an association with duration can indicate a causal relationship with cumulative exposure when intensity of exposure is independent of its duration, also highlighting that when duration and intensity are inversely associated, a trend with duration can be observed that is in the wrong direction[1]. We are not aware of systematic investigations of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2019 doi:10.20944/preprints201904.0306.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 1896; doi:10.3390/ijerph16111896



correlation structure between duration and intensity of occupational exposures in the context of this problem. However, there is an example of negative correlation between the two, e.g. if new hires are assigned to "dirtier" jobs that then leads them to change employment to avoid such exposure [2]. There are also reports of a positive correlations when such feedback is either unlikely [3], or when selection out of the workforce due to high exposures may not be strong [4]. There are settings where duration and intensity of exposure appear to be unrelated within a subject (e.g. for exposures emitted intermittently) [5], and between subjects (e.g. after selection on the basis of vulnerability to exposure, as has been shown to exist in bakers) [6]. Thus, specifics of the workplace, health condition, and selection of the study sample may all influence the correlation of duration and intensity. This raises concerns about both false positive and negative findings that could result from procedures that use duration as proxy for cumulative exposure. De Vocht et al.,[4] when intensity and duration had correlation of 0.3, observed stronger association with cumulative exposure compared to duration alone. Similarly, McDonald et al.,[7] reported that cumulative exposure to silica, but not duration alone, was associated with lung cancer, implying that if only duration was the available, then the likely causal association would have been missed. Another case in point is the lack of association of cancer mortality with trichloroethylene that may be due to absence of information on exposure intensity [8]. This is suspected, because a finding of an association of trichloroethylene with non-Hodgkin lymphoma was based on cumulative exposure, but was not observed for either duration or intensity alone [9]. Conversely, when an association is reported with duration of exposure and information on intensity is not available, there is a concern that error in exposure due to use of duration as a proxy for cumulative exposure may have created a false positive finding [10,11].

The reason why sometimes duration of exposure is available, but intensity is not, relates to cost associated with assessments of intensity of (workplace) exposure. Duration of exposure is typically derived from employment records or self-reports of occupational histories, which are the minimal requirements in occupational epidemiology. Estimating intensity of exposure requires an additional effort that assigns intensity of exposure to occupational histories and involves estimation processes based on either expert judgments or a typically limited collection of workplace measurements. At best, in most retrospective epidemiological studies researchers have information on the (historic) distribution of exposure intensity, but not individual values. In occupational epidemiology, this led to development of practice and theory of job-exposure matrices [12,13] and group-based exposure assessment [14-16]. However, such approaches raise the question of how to proceed with the analysis of health impact of accumulated exposure, when duration is assessed with a high degree of accuracy, while exposure intensity is subject to various modeling assumptions and is known, at best, in terms of its mean and variance. The naive practice in the field has been to compute cumulative exposure indices as if duration and intensity are of equal accuracy, using some form of best guess of intensity, or to resort to analysis by duration of exposure only. The improvement on this practice may lie in framing it in the context of missing data or measurement error problem.

We considered the problem from the theoretical perspective by exploring the expected behavior of the effect estimates. The focus of our work is not on false positive or false negative occurrences (as would arise from hypothesis-testing) but rather on a more pragmatic path of reasoning in epidemiology that deals with bias and precision of effect estimates as measure of their usefulness [17-19]. For the sake of clarity in describing the key features of the problem, we limit our analysis to the theoretically more tractable situation of continuously measured health outcome suspected be related to logarithm of cumulative exposure (e.g. relationship of noise to blood pressure [20] or hearing loss [21]), where analysis by linear regression could apply. Such constraints are most directly applicable to cross-sectional studies with continuous exposure and outcome measures (or any design where time-course of exposure is either not collected, or not relevant to the hypothesis). Thus, we do not address here the problem of time-varying variables. However, working out the details of this relatively simple case is a useful first step towards tackling the problem in more complex study designs, and in other disease models applicable to estimation of effects of exposure on binary and survival-time outcomes. We consider the realistic situation where duration and intensity of exposure may not be independent. Next, using synthetic data motivated by cross-sectional study of Kennedy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2019 doi:10.20944/preprints201904.0306.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 1896; doi:10.3390/ijerph16111896



et al.,[22] we outline and illustrate a Bayesian method aimed at recovering an estimate of cumulative exposure on the outcome, when only duration is assessed for every subject and some information on exposure intensity is available, i.e. is disjointed at the individual (sample) level from duration, following an approach reminiscent of Gustafson and Burstyn [23]. Finally, we illustrate our methodology using data from two waves of the National Health and Nutrition Examination Survey (NHANES) that can be used to assess association of smoking and lung function. Note that we do not aim to add to the underlying etiological questions, but that this is merely added as a practical example of the proposed methodology.

## 2. Theoretical analysis of impact on estimate of effect of cumulative exposure

For continuously measured health outcome $Y_i$ on the $i^{th}$ of n persons, the outcome model is assumed to be

$$Y_i=\beta_0+\beta_1 \mathit{log}\ C_i+e_i, \tag{1}$$

where $C_i$ is the cumulative exposure, $e_i$ is the error term distributed as $N(0, \sigma^2)$, and $\sigma^2$, $\beta_0$ and $\beta_1$ are the parameters. The cumulative exposure of the $i^{th}$ person is defined as the product of duration of exposure ($D_i$) and intensity ($I_i$), such that the outcome models can be re-written as: $(Y|D, I)\sim N(\beta_0+\beta_1(\mathit{log}\ D_i+\log\ I_i), \sigma^2)$. There is theoretical and empirical evidence that many occupational exposures are well-described by the lognormal distribution[24,25] and emerging evidence that age up to an event, such as either development of illness or selection into an epidemiologic study, can follow the lognormal distribution [25,26]. Consequently, we focus on situation where ($\mathit{log}\ I_i$, $\mathit{log}\ D_i$) follows a bivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with means $\mu_I$ and $\mu_D$, variances $\sigma_I^2$ and $\sigma_D^2$, respectively, and a correlation $\varrho$. This assumption is not necessary to linear regression in general, so we are considering a special case where such an assumption is defensible. Mathematical details pertinent to the rest of this section are in Appendix A, while the R [27] code need to reproduce Figures 1-3 is provided in Supplemental Material 1.

## 3. Naïve analysis

The relationships above in eq. (1) imply that $(Y|D)\sim N(\alpha_0+\alpha_1 \mathit{log}(D), \lambda^2)$, where expressions for ($\alpha_0$, $\alpha_1$, $\lambda^2$) in terms of the original parameters are given in Appendix A. When the investigators have no information about intensity of exposure and naively regresses outcome on $\mathit{log}(D)$ to estimate $\beta_1$ with $\hat{\alpha}_1$, we show that they incur bias

$$\alpha_1-\beta_1=\varrho k\beta_1, \tag{2}$$

where $k=\sigma_I/\sigma_D$. (In such an analysis, when the model in eq. (1) is assumed to be true, any interpretation of $\hat{\alpha}_1$ must be a reflection of the true causal association mediated by non-zero intensity of exposure.) Outside of some uncommon settings (particular combinations of parameter values paired with a very small sample size), this estimator has a root-mean-squared-error (RMSE) greater than that obtained in the complete-data case by the regressing outcome on log(C) exposure to obtain $\hat{\beta}_1$ (estimate of slope with complete data). In the special case where $\varrho=0$, bias is not incurred but variance of the estimator is inflated: $Var(\hat{\alpha}_1)=n^{-1}(\sigma^2+\beta_1^2\sigma^2_I)/\sigma^2_D > Var(\hat{\beta}_1)= n^{-1}\sigma^2/(\sigma^2_D+\sigma^2_I)$ (general expressions for estimator variances are in Appendix A). This is the same as Berkson-type error when log(D) is used as a surrogate of log(C) with error term $\log(I)\sim N(\mu_I, \sigma^2_I)$ [28]. When $\varrho k<-1$, the naïve analysis will estimate a target (tend to yield an estimate) that is in the opposite direction from the true effect (Figure 1). In other words, this situation can only occur when (a) intensity and duration are inversely related with sufficiently high correlation and (b) intensity is more variable than duration to a large enough degree to produce $\varrho k<-1$, leading to the case highlighted by Johnson [1]. Clearly, in such circumstances, as well as when bias is expected to be substantial, there is a motivation to either collect data on exposure intensity, or use knowledge about the joint distribution of intensity and duration to account for it in data analysis. Furthermore, when the RMSE of a naïve analysis is much

 doi:10.20944/preprints201904.0306.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 1896; doi:10.3390/ijerph16111896



worse than that obtainable with cumulative exposure, either further data collection, or adjustment are motivated, such as when duration and intensity are noticeably correlated (e.g. Figure 2). We develop intuition as to whether the adjustment can achieve worthwhile improvements in the next section; it is important to consider this because, where possible, the resources involved in additional statistical analyses and validation studies are less than the cost of full-scale assessment of intensity of exposure.

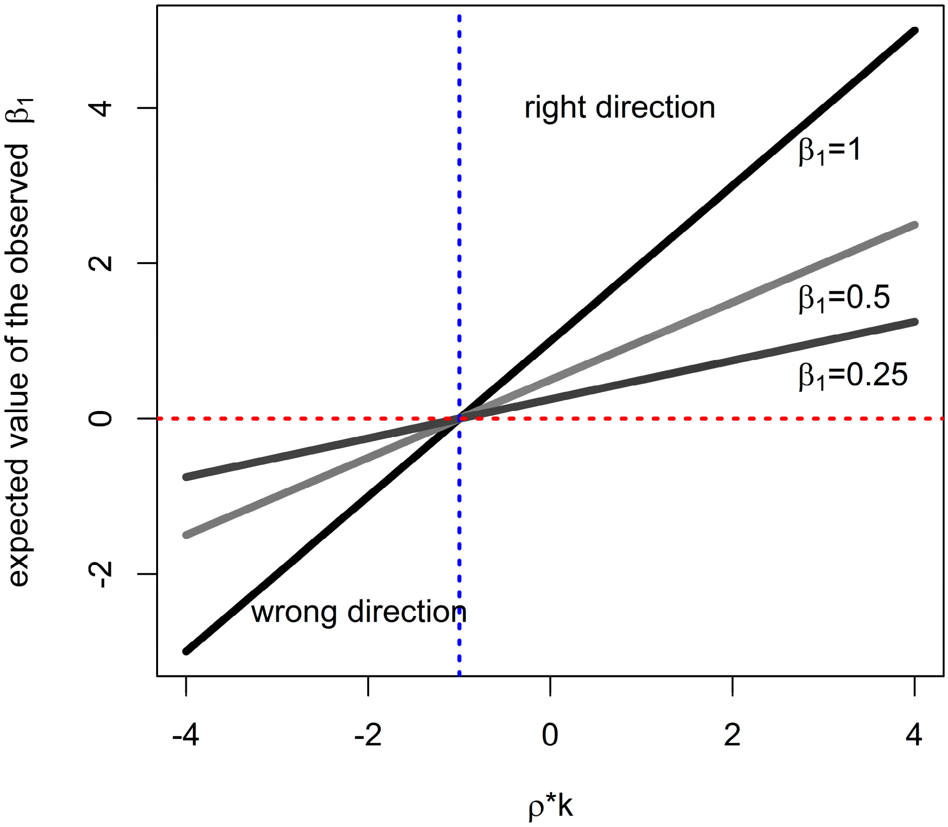


**Figure 1.** The expected direction of the apparent association with duration of exposure, as a function of correlation of intensity and duration ($\varrho$), ratio of variances of intensity and duration (k), and strength of causal effect ($\beta_1$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2019 doi:10.20944/preprints201904.0306.v1

Peer-reviewed version available at *Int. J. Environ. Res. Public Health* **2019**, *16*, 1896; doi:10.3390/ijerph16111896



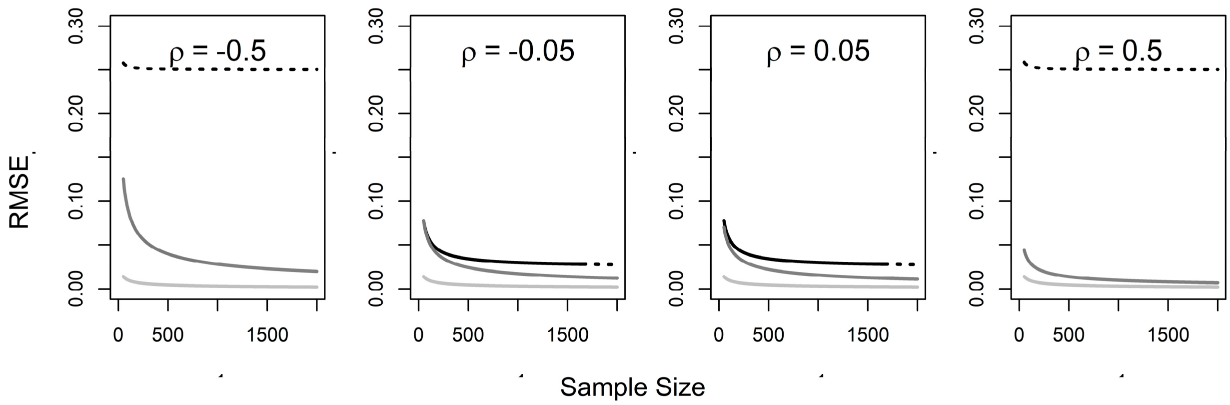


**Figure 2.** The root mean squared error (RMSE) as function of sample size in analysis (n) with duration of exposure (black), duration of exposure adjusted for distribution of intensity (grey), and cumulative exposure (light grey); dotted lines indicate that 95% confidence internal coverage is less than 50%. NB: correlation of intensity and duration varies by panel (ϱ), ratio of variances of intensity and duration (k=1), and strength of causal effect ($\beta_1$=0.5).

### 4. Adjusted analysis: the limit of what we can learn when only D is available, but ϱ and k are known

We imagine that the investigator can either conduct an exposure measurement campaign, or access existing measurements that yield insights into the relationship between duration and intensity of exposure. This can be done for a subset of subjects, so long as such sample is deemed representative. If we know ϱ and k (or more generally know $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$), then it is possible to remove bias but not possible to recover all the precision achievable with complete data. We remove the bias via the relationship implied by eq. (2), so the *adjusted* estimator is

$$\hat{\beta}_{1,A} = (1+\rho k)^{-1}\hat{\alpha}_1 , \quad (3)$$

We emphasize that this simple form of adjustment arises because the (Y|D) relationship arising from the presumed (Y|I,D) and (I,D) relationships has a simple form. We could arrive at essentially the same adjusted estimator by explicitly casting the problem as a missing-data imputation problem (I must be imputed for all subjects), or as a measurement error problem (D is a surrogate for C with certain properties). That is, the same likelihood function would underpin the inference, whether this is implicit or explicit in the implementation of the estimation scheme. Of course imputation or latent-variable measurement error approaches could still be applied in more elaborate versions of the problem, when a simple form for Y|D is no longer manifested.

The RMSE of the adjusted estimator shows complex behavior relative to the naïve estimator (Figure 3). It must be noted that the adjusted estimator (and its RMSE) are undefined when ϱk=-1 (denoted by vertical dotted blue line in Figure 3), and the RMSE tends to very large values near this value (see Appendix A). To develop further intuition about this relationship, we focus on special case of $\beta_1$=0 and note that when -2<ϱk<0, the RMSE of the adjusted estimate is worse than that of the naïve one: although there is no bias, precision deteriorates. This arises when the intensity and duration are inversely related. This is illustrated in Figure 3, that compares RMSE of adjusted and naïve estimators for ϱk <0: the red line indicates where RMSE's are equal, such that values above the line indicate a situation where adjusted estimators outperform naïve ones. As the strength of the association with cumulative exposure increases (denoted by solid lines in Figure 3, each associated with different $\beta_1$), the range of ϱk values that result in worse RMSE in adjusted analysis declines. However, it is noteworthy that the degree to which the naïve estimator can outperform the adjusted estimator is small relative to the advantage of the adjustment under most conditions. The exact shape of solid lines in Figure 3 depends on parameters for which the figure is generated, but Figure 3 depicts the expected general pattern of inter-dependence of the ratio of RMSE, $\beta_1$, and ϱk. Furthermore, the relative magnitude of RMSE grows less favorable for the adjusted estimate for small sample size, because the variance contributes disproportionately to the RMSE, and dwarfs the contribution of bias that plagues the naïve estimator. Conversely, for large sample sizes, variances make little contribution to the RMSE whereas bias remains constant, leading to smaller RMSE for the unbiased adjusted estimator relative to the biased naïve estimator.

The gap predicted by theory between the RMSE values under naïve and complete data analyses that can be narrowed by adjustment tends to be greater when duration and intensity are more strongly correlated (positively or negatively) (Figure 2) and intensity is more varied than duration (large k; not illustrated). In Figure 2, the dotted lines indicate that 95% confidence interval coverage is less than 50%. The confidence interval coverage of naïve analyses degrades with increase in sample size and strength of the correlation between duration and intensity, but tends to be recovered in adjusted analyses. These are the circumstances where we can expect to gain by infusing naïve analyses with knowledge about the joint distribution of intensity and duration. However, when duration and intensity are weakly associated, much more accurate estimates can only be obtained by collecting data on intensity for all subjects (the two middle panels of Figure 2), because the RMSE and coverage of naïve and adjusted data analyses are anticipated not to differ substantially; this also tends to occur when duration is more varied than intensity of exposure (small k; not illustrated).

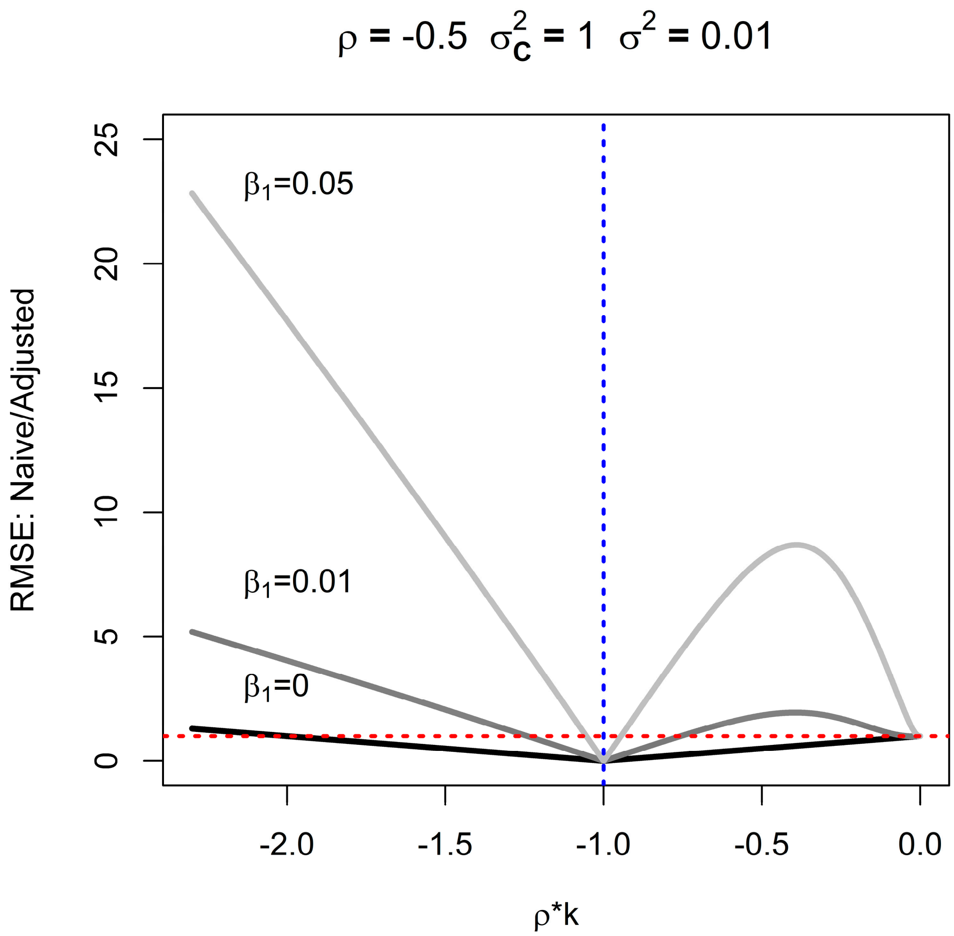


**Figure 3.** Circumstances when infusion of analysis with additional information on exposure intensity is expected to degrade root mean squared error (RMSE), as a function of correlation of intensity and duration ($\varrho$= -0.5), ratio of variances of intensity and duration (k), and strength of causal effect ($\beta_1$) for n=5,000, $\sigma^2$= 0.01, *Var*(log C)=1; red line indicates where RMSE's are equal; blue line indicates where adjusted RMSE is undefined.

## 5. Bayesian analysis when information of exposure duration and intensity is disjointed

### *5.1. Models*

If some information is available about the distribution of intensity of exposure, then we can learn about the effect of cumulative exposure by combining this with analysis by duration of exposure. In this case, information about duration and intensity is disjointed in the spirit of analysis presented by Gustafson and Burstyn [23] who considered a problem of estimating gene-environment interactions when information on prevalence of exposure was only available at the aggregate level, susceptible genotype was known for all subjects, and it was admissible to assume that susceptible genotype and disease were independent in absence of exposure. In other words, assumptions about the joint distribution of the unobserved quantity (exposure) and the observed quantity (genotype), plus an assumption about the disease model, allowed inference on the joint effect of exposure and genotype. The similarity with the current problem lies in the fact that the measure available on all subjects, i.e. duration of exposure, is associated with the outcome only though the interplay with intensity of exposure, and that information on intensity of exposure is only available in the form of knowledge about the joint distribution with duration of exposure. In other words, in both problems, the use of a mis-specified model allows for the inference about the parameter of interest when specific assumptions are justified.

Let us recall that if we know $\rho$ and k, we can correct for the bias arising from the use of duration as proxy for cumulative exposure and obtain the associated estimator variance, as shown earlier in eq. (3). In principle, if we do not know $\rho$ and k but can elucidate informative priors for these parameters, we can sample values from these distributions and incorporate them into eq. (3) to obtain a posterior distribution of $\beta_1$. We use a common default prior for the regression parameters (the g-prior [29], see Hoff [30] for an accessible description). We presume that the investigator uses a scaled beta distribution on [-1, 1] to set the prior on $\rho$, and a log-normal distribution to set the prior on k. As described in Appendix A, posterior computation is straightforward since the posterior distribution can be shown to be a truncated version of a distribution itself composed of standard distributions. Thus, simple Monte Carlo samples can be drawn from the posterior distribution and Markov chain Monte Carlo methods are not required. The general flavor of this analysis is in keeping with probabilistic bias analysis [19], including the need to discard some samples that violate a constraint imposed on $\beta_1$ by the residual variance of naïve analysis ($\lambda^2$); the proportion of samples that violate the constraint grows as $\rho k$ nears -1 (details are in Appendix A).

### *5.2. Synthetic example*

We illustrate this estimation procedure and it properties in a synthetic data inspired by a cross-sectional study of respiratory health of saw-filers by Kennedy et al. [22] In doing so, we simply strive to demonstrate the usefulness of informative priors on $\rho$ and k, not to fully evaluate an existing Bayesian procedure for fitting linear regression. Using linear regression, Kennedy et al. [22] showed a decline in forced expiratory volume in one second (FEV1) in relation to both duration and intensity of exposure (without log-transformation) to cobalt (Co) separately, implying that this association also exists with cumulative exposure. Let us imagine a follow-up study that is about 5 times larger than the original (500 subjects) with similar distributions of duration and intensity of exposure, but without measurements of intensity of exposure to Co due to high cost of obtaining individual measurements. We show how information on the distribution of intensity from the original study can be used to estimate the effect of cumulative exposure in a hypothetical follow-up study. We estimated distributions of duration and intensity from the original paper and set $\beta_0$ and $\beta_1$ to be weaker yet consistent with the original work (see Supplemental Material 2 for details, including R code for implementation of all analyses). The value of k consistent with the original paper is on the order of 2.6, implying that bias in duration-only analysis can be substantial according to eq. (2). We imagined two plausible values of $\rho$: -0.5 (e.g. assuming selection of highly exposed workers out of sample available for study due to their deteriorating health) and +0.5 (e.g. assuming a stable workforce with higher exposures in the past); this leads to $\rho k$ values of about -1.3 and 1.3, respectively. Both situations

 doi:10.20944/preprints201904.0306.v1

Peer-reviewed version available at *Int. J. Environ. Res. Public Health* **2019**, *16*, 1896; doi:10.3390/ijerph16111896



are common in occupational and environmental epidemiology and cannot be discounted *a priori*, but these situations are not meant to be all-encompassing of possible correlations. Having generated synthetic datasets using these parameters, we analyzed them via

1. the naïve approach (duration only),
2. four wide priors on $\varrho$ (two of which admit uncertainty about the sign of the correlation, when the prior mean is one standard deviation below) and k (Priors 1),
3. four narrow priors on $\varrho$ and k (Priors 2),
4. assuming known $\varrho$ and k, and
5. complete data.

The details of implementation in *R* can be found in Supplemental Material 2. In both (2) and (3), priors were set such that prior means were either above or below the true values by one prior standard deviation. As such, they represent guesses of various certainty that were off target, as may be expected when priors are reasonably well calibrated, with the best guesses off-target but not so much as to render them blatantly wrong. The results are illustrated in Figures 4 and 5. When $\varrho=-0.5$ and $\varrho k<-1$ (Figure 4), we note that the naïve analysis results in a reversal of direction of effect estimate, which is remedied when using the more informative priors, i.e. priors in (3). We observe that 95% credible intervals (CrI) exclude true values in naïve analyses, but capture them in analyses that assume known $\varrho$ and k (except in one illustrated case of negative correlation of intensity and duration). When priors are placed on $\varrho$ and k, the inference appears to be sensitive to the choice of priors (with inheritance of more uncertainty with broader priors) but is superior to naïve analysis in that it includes the true value in the 95%CrI's (better coverage). It appears that informative analysis is possible even if there is doubt about the direction of $\varrho$, i.e. priors in (2). Analysis with the narrower priors in (3) tend to yield comparable inference to that obtained with known values of $\varrho$ and k. The analysis is clearly challenging when $\varrho$ <0 and k is large, as even knowing these quantities appears to lead to biased inference in some of our synthetic datasets. We repeated all calculations by switching the variances of duration and intensity, leading to k=1/2.6=0.38. As expected, bias in such situation is reduced and the motivation to adjust may be reduced, even where $\varrho$<0 (Supplemental Material 3).

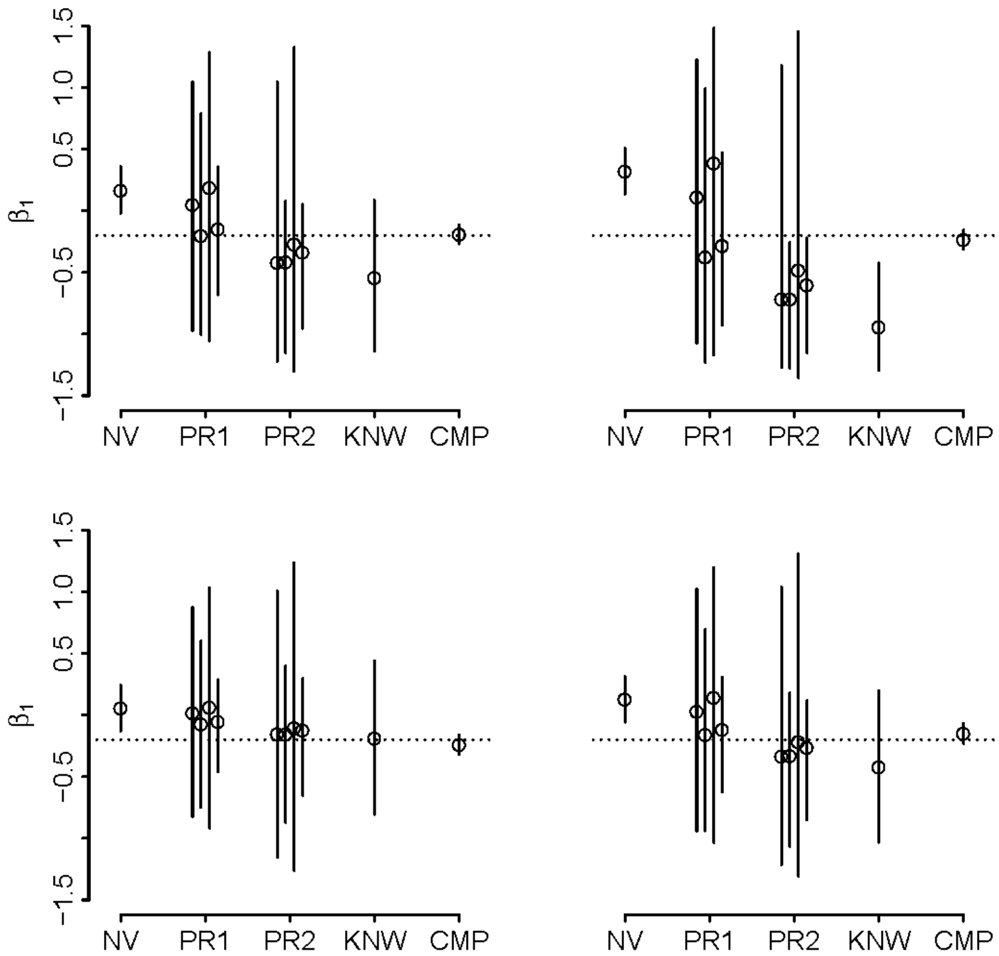


**Figure 4.** Adjusted estimates of $\beta_1$ with different degrees of knowledge about joint distribution of duration and intensity of exposure when $\varrho$ = -0.5 and k=2.6 in four simulations of synthetic example; naïve estimate (NV) is contrasted with adjusted estimates obtained under "well-calibrated" priors on ($\varrho$,k) that are "wide" (PR1), "narrow" (PR2), estimates obtained with $\varrho$ and k known (KNW; the best one can do without complete data), and complete data on intensity and duration (CMP); true value is denoted by dotted line, solid lines represent 95% credible intervals; of see text for details.

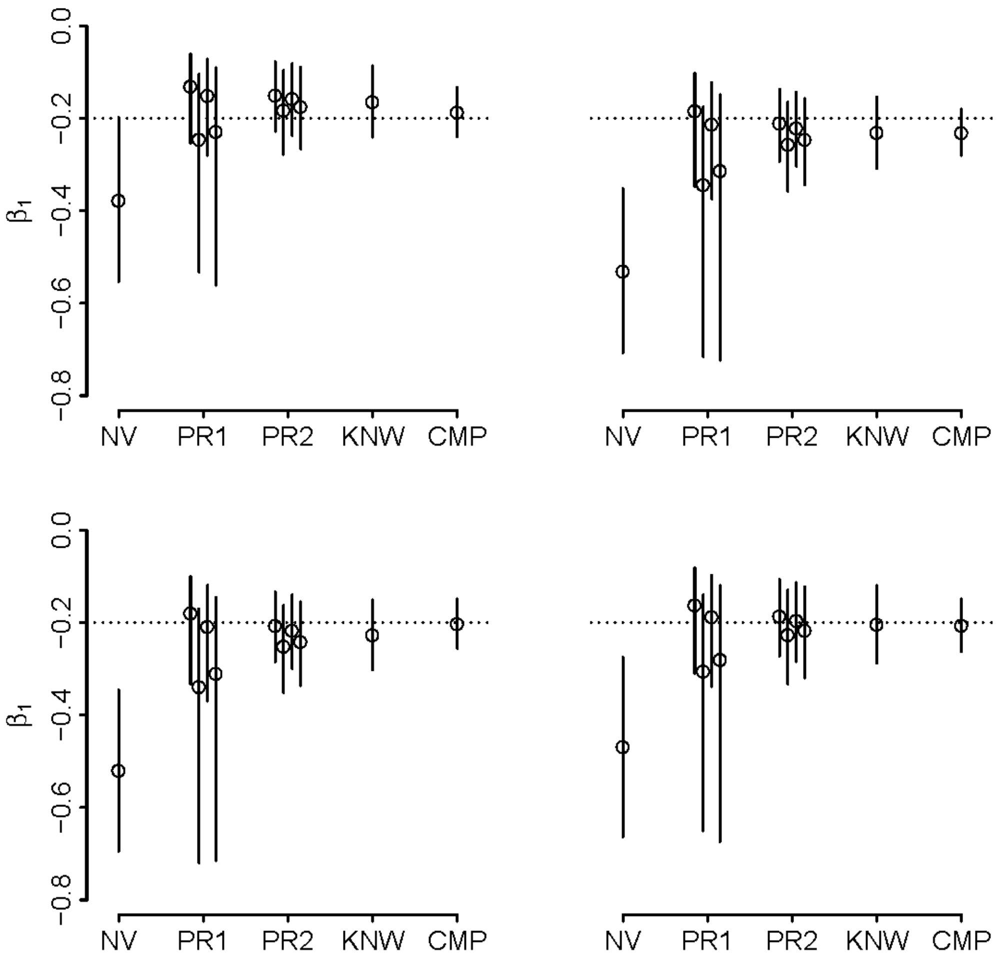


**Figure 5.** Adjusted estimates of β1 with different degrees of knowledge about joint distribution of duration and intensity of exposure when ϱ = +0.5 k=2.6 in four simulations of synthetic example; naïve estimate (NV) is contrasted with adjusted estimates obtained under "well-calibrated" priors on (ϱ,k) that are "wide" (PR1), "narrow" (PR2) and estimates obtained with ϱ and k known (KNW; the best one can do without complete data), and complete data on intensity and duration (CMP); true value is denoted by dotted line, solid lines represent 95% credible intervals; of see text for details.

### *5.3. Real-world application*

To illustrate (the advantages of) our methodology, we use the example of a known association between cumulative exposure to cigarette smoke and forced vital capacity (FVC) in the lungs of male adult smokers (currently smoking and restricted to a cumulative consumption of at least 100 cigarettes in life for this example) using the United States NHANES data. Details of data preparation and all calculations (in *R*) are in Supplemental Material 4. Information on intensity of smoking ("average number of cigarettes per day during past 30 days") and duration ("age at survey" - "age started smoking cigarettes regularly") is available in the 2009-2010 wave of NHANES. We assume that (contrary to the fact) in the subsequent 2010-2012 wave, the decision was made to only collect information on duration of smoking. This would allow us to estimate ϱ (=0.12) and k (=1.2) from 2009-2010 data (595 persons) and use it to derive priors for analysis of the association between duration of smoking and FVC in 2011-2012 data (570 persons), aimed at inferring the association with cumulative exposure (pack-years). The 2011-2012 data is illustrated in Figure S3 in Supplemental Material 4. There is evidence of an inverse linear association of log(FVC) with both log(duration) and log(pack-years) of smoking cigarettes, as expected. We note that ϱk is equal to 0.14, suggesting that the bias due to use of duration as a surrogate of cumulative exposure is expected to be small. We analyze NHANES data using the same priors (except with different numeric values of ϱ and k) as those we employed in the synthetic example with one exception to meaning of a prior previously labeled as "known" is now designated as "fixed" values. To wit, we consider a scenario in which we have the very high confidence that pre-existing data (2009-2010) yielded true values of ϱ and k parameters in the 2011-2012 data and use these fixed values for ϱ and k. However, it should be noted that even if we have a high confidence of in these values, in this case the values of ϱ and k cannot be considered exactly as "known". The outcome of Bayesian analyses is presented in Figure 6. It appears that in this example the existence of the association and its direction could also be inferred from the use of duration of exposure alone, i.e. there is little gain in terms of the qualitative conclusion by incorporating the additional information on intensity in the 2011-2012 wave. The 95% credible intervals of complete data analysis do not overlap with analyses of incomplete data, even when infused with information on how duration and intensity are related (i.e. ϱ and k), except in the case of some wide priors (those among Priors 1). This underscores the challenge of bias-reduction in this specific application, anticipated by theory, due to both small ϱ and large value of k (intensity more varied than duration), and argues for importance of quantifying intensity of exposure at individual level. In this application, our method resulted only in a small improvement in the accuracy of the assessment of the strength of the association.

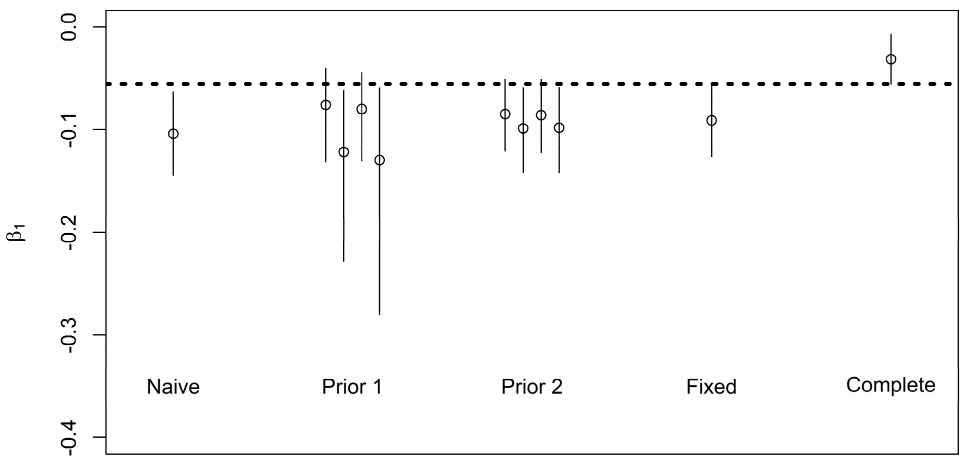


**Figure 6.** Estimated change in log(FVC, ml) among 570 male current smokers in NHANES 2011-2012 under different priors; naïve analysis is the association with log(years of smoking), complete analysis is the association with log(pack-years), see text for description of different priors (Prior 1, Prior 2, Fixed) that use information on correlation of logarithms of duration and pack-years (ϱ) and ratio of standard deviations of logarithms of packs/day and duration (k); circles represent 50th percentile of posterior distributions and line span the 95% credible intervals, dashed line represents lower bound of the 95% credible interval with complete data.

## 6. Discussion

In the context of continuous outcomes amendable to analysis by linear regression, we placed speculations of Johnson[1] about effects of using duration of exposure instead of intensity onto a more solid theoretical foundation and highlighted the importance to bias and precision of the correlation between duration and intensity of exposure, as well as ratio of their variances. Specifically, we stressed the analytical challenges that arise when such correlation is negative, and the intensity is more varied than duration. Lastly, we developed a pragmatic Bayesian approach to the problem.

Our findings are relevant to studies with binary and time-to-event outcomes, although caution is required in drawing analogies. For example, when ϱ=0 and we are reduced to Berkson-type error,

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 1896; doi:10.3390/ijerph16111896



logistic regression will be biased towards the null (unlike linear regression) [31] and the situation with Cox proportional hazard model is nuanced with bias depending on rarity of censoring [32,33]. It is perilous to speculate further, given the complexity we discovered in the case of linear regression. We note that the problem we consider falls within larger domain of scholarship on measurement error problem, [34,35] as well as analytical methods for omitted covariates and latent confounders [36-39], which have advanced solutions for a wider range of models than considered here. It is likely that rapid progress can be made by leveraging such advances where analogy to duration of exposure being a surrogate for cumulative exposure can be defended. At the same time, the mechanics of implementing a Bayesian analysis that we present should be easily adaptable to other study designs and data types, and our approach may inform advances in related statistical problems.

In practice, not only we will be often uncertain about joint distribution of duration and exposure, but also whatever information we have about duration and intensity is typically contaminated by measurement error. This concern is partially addressed when in Bayesian analyses we admit uncertainty about $\varrho$ and k, and may discourage analysis that fixes these quantities as "known". The matter of uncertainty about observed duration of exposure is a more grave concern, as it anchors adjustments that are performed via priors on $\varrho$ and k. We can try to overcome this problem if there is some information about a measurement error model for duration of exposure, such that duration can be modeled as a latent construct, as in established methods for analyses contaminated by measurement error [34]. However, we note that duration of exposure is usually recorded with reasonable accuracy in occupational epidemiology, at least when employment records are used from traditional industrial environments. Thus, in many circumstances, errors in duration of exposure are likely negligible compared to those in its intensity.

Our findings apply only to situations where the disease model is not miss-specified (e.g. the logarithm of cumulative exposure is the correct dose-metric, there are no lags or thresholds, toxicity is not reversible, the effect is linear in the chosen scale). Where this is not the case, extension of our work to a more flexible modeling approach can be contemplated [40,41], but it is equally important to admit that there is a perpetual uncertainty about correct dose-metric in epidemiology, even for well-studied problems. As such, any support for specific dose-metric remains the key element of analysis that must precede consideration of duration of exposure as proxy of true dose-metric [4,42]. Consideration of time-varying measures of duration and cumulative exposure also constitute a natural extension of our work. Where such matters are pivotal, as in analysis of cohort studies, we are willing to speculate that the case of time-varying exposure is not very dissimilar to the one we considered, if viewed from the prism of measurement error problem, in which accumulated exposure up to a given time point or during any discrete time period is approximated by duration of exposure since it start or during a discrete time period.

To circumvent issues involved in the choice of specific functional forms of exposure metrics, such as log(duration) vs. duration *per se*, many analysts conduct analyses using categories of exposure. Although this is certainly a viable approach, there are concerns associated with such methodology that arise from the induction of differential misclassification of exposure [43,44], increase chance in spurious associations [45] and misspecifications of disease models when true risks are expected not to have a threshold. Ideally, different functional forms of exposure metrics yield comparable interpretations of the data, with logarithms of duration and cumulative exposure considered because of theoretical properties that we illustrated and because they tend to counteract undue influence of extreme values.

## 5. Conclusions

When it is reasonable to make assumptions consistent with our work and epidemiologists can be assured that duration and intensity of exposure are either independent or positively correlated, they can be more confident in qualitatively interpreting direction of effects that arise from use of duration of exposure in lieu of true dose metrics when the true dose is captured by cumulative exposure. If they can further substantiate a claim that duration of exposure is more variable than its intensity, they can place more weight on inference about the magnitude of true association with cumulative exposure. However, such analyses are unlikely to be found suitable for quantitative risk assessment. To optimize (or in some cases where individual data on intensity is not available -- make possible) reliable inference about the magnitude of effects of cumulative exposure on the outcome, epidemiologists can use information on the relationship between duration and intensity of exposure even if intensity of exposure is not available at the individual level.

**Supplementary Materials:** The following are available, Supplemental Material 1: R code to generate Figures 1 to 3, Supplemental Material 2: R code to conduct Bayesian analysis with prior on joint distribution of intensity of exposure and its duration and to generate results shown in Figures 4 to 5, Supplemental Material 3: Analysis of synthetic data with value of k inverted compared to that presented in main text; Figures S1 and S2, Supplemental Material 4: Real-world Application, Figure S3 and R-code used to download, select, and analyze NHANES data and to create Figures 6 and S3.

**Author Contributions:** I.B and F.A.-B. conceptualized the project. I.B. and P.G. developed the methodology; I.B. and F.d.V. conducted formal analysis in real-world example. All authors contribute to both original draft preparation and review and editing of the subsequent versions.

**Funding:** This research received no external funding.

**Acknowledgments:** The authors are thankful to James Leon Beau Burstyn for allowing lead author enough hours of sleep to complete the revisions, negative correlation of intensity and duration of crying, and for encouraging common sense approach to all complex problems.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

### Theory

Recall that we start with (Y|D, I)~$N(\beta_0 + \beta_1(\log D + \log I), \sigma^2)$ and ($\log$ I, $\log$ D) ~ $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$\boldsymbol{\mu}=(\mu_I,\mu_D)'$ and

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_I^2 & \rho\sigma_I\sigma_D \\ \rho\sigma_I\sigma_D & \sigma_D^2\end{pmatrix}.$$

With complete data on (Y,I,D), we simply estimate $\beta_1$ from regression of Y on $\log$ C, with estimator variance given as $nVar(\hat{\beta}_1)=\sigma^2/Var(\log C)$. For the sake of comparison with later expressions, using $k=\sigma_I/\sigma_D$, this can be re-expressed as

$$nVar(\hat{\beta}_1)=\frac{\sigma^2}{\{(1+\rho k)^2+(1-\rho^2)k^2\}o_D^2}, \qquad \text{(A.1)}$$

To consider the situation without intensity data, note that Y|D ~ $N(\alpha_0+\alpha_1 \log D, \lambda^2)$, where $\alpha_0=\beta_0+\beta_1(\mu_I-\varrho k\mu_D)$, $\alpha_1=(1+\varrho k)\beta_1$, and $\lambda^2=\sigma^2+\ \beta_1^2\ \sigma_I^2\ (1-\varrho^2)$. Thus the naïve estimator can be viewed as $\hat{\alpha}_1$ obtained from regressing Y on $\log$D, which targets $\alpha_1$ rather than $\beta_1$. The bias incurred is then $\varrho k\beta_1$, while the estimator variance is

$$nVar(\hat{\alpha}_1)=\frac{\sigma^2+\beta_1^2(1-\rho^2)k^2\sigma_D^2}{o_D^2}$$

If ($\varrho$, k) are known then the adjusted estimator $\hat{\beta}_{1,A}=(1+\rho k)^{-1}\hat{\alpha}_1$ unbiasedly estimates $\beta_1$. The estimator variance is $Var(\hat{\beta}_{1,A})=(1+\rho k)^{-2}Var(\hat{\alpha}_1)$, which in fact can be written as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2019 doi:10.20944/preprints201904.0306.v1

Peer-reviewed version available at Int. J. Environ. Res. Public Health 2019, 16, 1896; doi:10.3390/ijerph16111896



$$nVar(\hat{\beta}_{1,A}) = \frac{\sigma^2 + \beta_1^2(1-\rho^2)k^2\sigma_D^2}{(1+\rho k)^2 o_D^2}, \qquad \text{(A.2)}$$

Comparing both numerators and denominators in (A.1) and (A.2) respectively, we directly see the reduced efficiency of adjusting without intensity data compared to having such data.

A nuance concerning the adjustment is that the form of $\lambda^2$ induces a constraint in the parameters governing (Y|D) and (D), namely that $\beta_1^2 < \lambda^2 / \{\sigma_I^2(1-\rho^2)\}$ (to see this more clearly, consider that that $\beta_1^2 = \lambda^2 / \{\sigma_I^2(1-\rho^2)\}$ would imply the impossible condition of $\sigma^2=0$). This is relevant to the special case that the known ($\rho$, k) values satisfy $\rho k = -1$. Clearly $\hat{\beta}_{1,A}$ does not exist in this case, and indeed $\beta_1$ is not a point identified by (Y,D) data. However, $\beta_1$ would be interval-identified, in that all quantities in the upper-bound for $\beta_1^2$ are either known, or estimable.

A further consequence of the form of $\lambda^2$ is that in the case that ($\rho$, k) are unknown and described by prior distributions, we must *a priori* rule out parameter values that violate the constraint. Expressed purely in the Y|D and D parameterization, the inequality takes the form

$$\alpha_I^2 < (1+\rho k)^2 \lambda^2 / \{k^2 o_D^2 (1-\rho^2)\}$$

Thus, we use a prior distribution of the form

$$f(\alpha, \lambda^2, o_D^2, \rho, k) \propto g_1(\alpha, \lambda^2) g_2(o_D^2) g_3(\rho) g_4(k) I_R\{\alpha, \lambda^2, o_D^2, \rho, k\},$$

Here $g_1()$ through $g_4()$ are densities specified for the constituent parameters, while *R* is the subset of the parameter space on which the constraint is satisfied. Thus, we are using truncation to obtain a prior distribution that respects the structure of the problem.

As a generic prior for regression parameters, we take $g_1()$ to be the g-prior with default hyper-parameters g=n, $\upsilon_0=1$, $\sigma_0=1$ (as parameterized, for instance, in Hoff PD. Linear regression *A first course in Bayesian statistical methods.*, New York: Springer-Verlag 2009;149-170). Similarly, $g_2()$ is specified as inverse gamma with shape and scale parameters both set to 0.5. As a convenient form for the investigator to specify prior information about $\rho$, $g_3()$ is specified as the scaled-beta distribution on [-1,1], which can be simply parameterized via mean and standard deviation. Further, given the definition of *k* as a ratio of variances, we take $g_4()$ to be a log-normal distribution.

The posterior distribution arising from this prior is tractable in the sense that without enforcing the constraint, the joint posterior is characterized by independent conjugate posterior distributions for ($\alpha$, $\lambda^2$) and $\sigma_D^2$ along with the independent prior distributions for $\rho$ and k (since neither $\rho$ nor *k* appears in the likelihood function). Consequently, independent Monte Carlo draws from the joint posterior without the constraint are easily taken. The constraint can then be enforced simply by discarding those sampled ($\alpha$, $\lambda^2$, $\sigma_D^2$, $\rho$, k) draws that violate it. Markov Chain Monte Carlo methods are not required.

For some datasets and prior specifications, very few, if any posterior draws are discarded. In other cases, however, the discarded proportion can be substantial. Unsurprisingly given the discussion above concerning $\hat{\beta}_{1,A}$, a prior putting some mass for ($\rho$,*k*) near $\rho k = -1$ tends to result in a higher proportion discarded.

Note that by setting $g_3()$ and $g_4()$ to be point mass priors, we obtain a Bayesian version of the known ($\rho$, $k$) adjustment procedure. In doing so, if the dataset is such that there is little to no posterior truncation, then the resulting posterior mean and standard deviation of $\beta_1$ will closely approximate $\hat{\beta}_{1,A}$ and $SE[\hat{\beta}_{1,A}]$, as arises from Bayesian linear regression with a default prior. However, for datasets leading to considerable truncation, this approximate equivalence is no longer guaranteed. In particular, the Bayesian version should be more trustworthy when $\rho k$ is close to -1, with the possibility of achieving more precision than stated in (A.2).

**References**


1. Johnson, E.S. Duration of exposure as a surrogate for dose in the examination of dose response relations. *Br J Ind Med* **1986**, *43*, 427-429.
2. Blair, A.; Thomas, K.; Coble, J.; Sandler, D.P.; Hines, C.J.; Lynch, C.F.; Knott, C.; Purdue, M.P.; Zahm, S.H.; Alavanja, M.C., et al. Impact of pesticide exposure misclassification on estimates of relative risks in the Agricultural Health Study. *Occup Environ Med* **2011**, *68*, 537-541, doi:10.1136/oem.2010.059469.
3. Westberg, H.B.; Hardell, L.O.; Malmqvist, N.; Ohlson, C.G.; Axelson, O. On the use of different measures of exposure-experiences from a case-control study on testicular cancer and PVC exposure. *J Occup Environ Hyg* **2005**, *2*, 351-356, doi:10.1080/15459620590969046.
4. de Vocht, F.; Burstyn, I.; Sanguanchaiyakrit, N. Rethinking cumulative exposure in epidemiology, again. *J Expo.Sci.Environ.Epidemiol.* **2015**, *25*, 467, doi:jes201458 [pii];10.1038/jes.2014.58 [doi].
5. Preller, L.; Burstyn, I.; De, P.N.; Kromhout, H. Characteristics of peaks of inhalation exposure to organic solvents. *Ann.Occup.Hyg.* **2004**, *48*, 643-652, doi:10.1093/annhyg/meh045 [doi];meh045 [pii].
6. Nieuwenhuijsen, M.J.; Lowson, D.; Venables, K.M.; Newman-Taylor, A.J. Correlation between different measures of exposure in a cohort of bakery workers and flour millers. *Annals of Occupational Hygiene* **1995**, *39*, 291-298.
7. McDonald, J.C.; McDonald, A.D.; Hughes, J.M.; Rando, R.J.; Weill, H. Mortality from lung and kidney disease in a cohort of North American industrial sand workers: an update. *Ann Occup Hyg* **2005**, *49*, 367-373, doi:10.1093/annhyg/mei001.
8. Lipworth, L.; Sonderman, J.S.; Mumma, M.T.; Tarone, R.E.; Marano, D.E.; Boice, J.D., Jr.; McLaughlin, J.K. Cancer mortality among aircraft manufacturing workers: an extended follow-up. *J Occup Environ Med* **2011**, *53*, 992-1007, doi:10.1097/JOM.0b013e31822e0940.
9. Purdue, M.P.; Bakke, B.; Stewart, P.; De Roos, A.J.; Schenk, M.; Lynch, C.F.; Bernstein, L.; Morton, L.M.; Cerhan, J.R.; Severson, R.K., et al. A case-control study of occupational exposure to trichloroethylene and non-Hodgkin lymphoma. *Environ Health Perspect* **2011**, *119*, 232-238, doi:10.1289/ehp.1002106.
10. Burstyn, I.; Yang, Y.; Schnatter, A.R. Effects of non-differential exposure misclassification on false conclusions in hypothesis-generating studies. *Int.J Environ.Res.Public Health* **2014**, *11*, 10951-10966, doi:ijerph111010951 [pii];10.3390/ijerph111010951 [doi].
11. Loken, E.; Gelman, A. Measurement error and the replication crisis. *Science* **2017**, *355*, 584-585, doi:10.1126/science.aal3618.
12. Hoar, S. Job exposure matrix methodology. *J Toxicol Clin Toxicol* **1983**, *-84;21(1-2)*, 9-26.

13. Peters, S.; Vermeulen, R.; Portengen, L.; Olsson, A.; Kendzia, B.; Vincent, R.; Savary, B.; Lavoue, J.; Cavallo, D.; Cattaneo, A., et al. SYN-JEM: A Quantitative Job-Exposure Matrix for Five Lung Carcinogens. *Ann Occup Hyg* **2016**, *60*, 795-811, doi:10.1093/annhyg/mew034.
14. Kim, H.M.; Richardson, D.; Loomis, D.; vanTongeren, M.; Burstyn, I. Bias in the estimation of exposure effects with individual- or group-based exposure assessment. *J.Expo.Sci.Environ.Epidemiol.* **2011**, *21*, 212-221, doi:jes200974 [pii];10.1038/jes.2009.74 [doi].
15. Tielemans, E.; Kupper, L.L.; Kromhout, H.; Heederik, D.; Houba, R. Individual-based and group-based occupational exposure assessment: Some equations to evaluate different strategies. *Ann.Occup.Hyg.* **1998**, *42(2)*, 115-119.
16. Xing, L.; Burstyn, I.; Richardson, D.B.; Gustafson, P. A comparison of Bayesian hierarchical modeling with group-based exposure assessment in occupational epidemiology. *Stat.Med.* **2013**, *32*, 3686-3699, doi:10.1002/sim.5791 [doi].
17. Poole, C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology* **2001**, *12*, 291-294.
18. Lash, T.L. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. *Am J Epidemiol* **2017**, *186*, 627-635, doi:10.1093/aje/kwx261.
19. Lash, T.L.; Fox, M.P.; Fink, A.K. *Applying Quantitative Bias Analysis to Epidemiologic Data*; Springer: Dordrecht, Heidelberg, London, New York, 2009.
20. Talbott, E.O.; Gibson, L.B.; Burks, A.; Engberg, R.; McHugh, K.P. Evidence for a dose-response relationship between occupational noise and blood pressure. *Arch Environ Health* **1999**, *54*, 71-78, doi:10.1080/00039899909602239.
21. Seixas, N.S.; Neitzel, R.; Stover, B.; Sheppard, L.; Feeney, P.; Mills, D.; Kujawa, S. 10-Year prospective study of noise exposure and hearing damage among construction workers. *Occup Environ Med* **2012**, *69*, 643-650, doi:10.1136/oemed-2011-100578.
22. Kennedy, S.M.; Chan-Yeung, M.; Marion, S.; Lea, J.; Teschke, K. Maintenance of stellite and tungsten carbide saw tips: respiratory health and exposure-response evaluations. *Occup Environ Med* **1995**, *52*, 185-191.
23. Gustafson, P.; Burstyn, I. Bayesian inference of gene-environment interaction from incomplete data: what happens when information on environment is disjoint from data on gene and disease? *Stat.Med.* **2011**, *30*, 877-889, doi:10.1002/sim.4176 [doi].
24. Koch, A.L. The logarithm in biology. 1. Mechanisms generating the log-normal distribution exactly. *J Theor Biol* **1966**, *12*, 276-290.
25. Limpert, E.; Stahel, W.A.; Abbt, M. Log-normal distributions across the sciences: keys and clues. *BioScience* **2001**, *51*, 341-352.
26. Gualandi, S.; Toscani, G. Human Behavior And Lognormal Distribution. A Kinetic Description. *arXiv* **2018**, *arXiv:1809.01365*.
27. Team, R.D.C. *R: A language and environment for statistical computing. ISBN 3-900051-07-0*; R Foundation for Statistical Computing: Vienna, Austria, 2006.
28. Berkson, J. Are there two regressions? *American Statistical Association Journal* **1950**, *June*, 164-180.
29. Zellner, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision techniques* **1986**.
30. Hoff, P.D. Linear regression. In *A first course in Bayesian statistical methods.*, 1 ed.; Springer-Verlag New York, 2009; 10.1007/978-0-387-92407-6pp. 149-170.

31. Reeves, G.K.; Cox, D.R.; Darby, S.C.; Whitley, E. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Stat.Med* **1998**, *17*, 2157-2177.
32. Prentice, R. Covariate measurement errors and parametric estimation in a failure time regression model. *Biometrika* **1982**, *69*, 331-341.
33. Kim, H.M.; Yasui, Y.; Burstyn, I. Attenuation in risk estimates in logistic and Cox proportional-hazards models due to group-based exposure assessment strategy. *Ann.Occup.Hyg.* **2006**, *50*, 623-635.
34. Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology*; Chapman & Hall/CRC Press: 2004.
35. Carrol, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement error in Nonlinear Models*, 2 ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006.
36. Lin, N.X.; Logan, S.; Henley, W.E. Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics* **2013**, *69*, 850-860, doi:10.1111/biom.12096.
37. Gail, M.H.; Wieand, S.; Piantadosi, S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **1984**, *71*, 431-444, doi:https://doi.org/10.1093/biomet/71.3.431.
38. Lin, D.Y.; Psaty, B.M.; Kronmal, R.A. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **1998**, *54*, 948-963.
39. McCandless, L.C.; Gustafson, P.; Levy, A. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat.Med.* **2007**, *26*, 2331-2347.
40. Seixas, N.S.; Robins, T.G.; Becker, M. A novel approach to the characterization of cumulative exposure for the study of chronic occupational disease. *American Journal of Epidemiology* **1993**, *137(4)*, 463-471.
41. Lubin, J.H.; Caporaso, N.E. Cigarette smoking and lung cancer: modeling total exposure and intensity. *Cancer Epidemiol Biomarkers Prev* **2006**, *15*, 517-523, doi:10.1158/1055-9965.EPI-05-0863.
42. Smith, T.J.; Kriebel, D. *A Biologic Approach to Environmental Assessment and Epidemiology* Oxford University Press: New York, NY, USA, 2010.
43. Wang, D.; Shen, T.; Gustafson, P. Partial Identification arising from Nondifferential Exposure Misclassification: How Informative are Data on the Unlikely, Maybe, and Likely Exposed? *The International Journal of Biostatistics* **2012**, *8*, 1557-4679, doi: https://doi.org/10.1515/1557-4679.1397.
44. Gustafson, P.; Le Nhu, D. Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics* **2002**, *58*, 878-887.
45. Heavner, K.K.; Phillips, C.V.; Burstyn, I.; Hare, W. Dichotomization: 2 x 2 (x2 x 2 x 2...) categories: infinite possibilities. *BMC.Med.Res.Methodol.* **2010**, *10*, 59, doi:1471-2288-10-59 [pii];10.1186/1471-2288-10-59 [doi].