Annals of Epidemiology 23 (2013) 43–48

Contents lists available at SciVerse ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Sparse-data bias accompanying overly fine stratification in an analysis of beryllium exposure and lung cancer risk

Kenneth J. Rothman DrPH *, Paul L. Mosquin PhD

RTI Health Solutions, Research Triangle Institute, 200 Park Offices Drive, Research Triangle Park, NC

ARTICLE INFO

Article history:
Received 24 May 2012
Accepted 14 November 2012
Available online 6 December 2012

Keywords:
Epidemiologic methods
Stratified analysis
Sparse-data bias
Beryllium
Lung cancer

ABSTRACT

Purpose: Beryllium's classification as a carcinogen is based on limited human data that show inconsistent associations with lung cancer. Therefore, a thorough examination of those data is warranted. We reanalyzed data from the largest study of occupational beryllium exposure, conducted by the National Institute of Occupational Safety and Health (NIOSH).
Methods: Data had been analyzed using stratification and standardization. We reviewed the strata in the original analysis, and reanalyzed using fewer strata. We also fit a Poisson regression, and analyzed simulated datasets that generated lung cancer cases randomly without regard to exposure.
Results: The strongest association reported in the NIOSH study, a standardized rate ratio for death from lung cancer of 3.68 for the highest versus lowest category of time since first employment, is affected by sparse-data bias, stemming from stratifying 545 lung cancer cases and their associated person-time into 1792 categories. For time since first employment, the measure of beryllium exposure with the strongest reported association with lung cancer, there were no strata without zeroes in at least one of the two contrasting exposure categories. Reanalysis using fewer strata or with regression models gave substantially smaller effect estimates. Simulations confirmed that the original stratified analysis was upwardly biased. Other metrics used in the NIOSH study found weaker associations and were less affected by sparse-data bias.
Conclusions: The strongest association reported in the NIOSH study seems to be biased as a result of nonoverlap of data across the numerous strata. Simulation results indicate that most of the effect reported in the NIOSH paper for time since first employment is attributable to sparse-data bias.

Introduction

Previously we demonstrated [1] through simulation studies and control of age confounding using stratification that the nested case-control study of beryllium and fatal lung cancer by Sanderson et al [2] was affected by strong confounding by year of birth for lagged measures of average daily exposure. Others have also commented on the age confounding in that study [3,4]. Recently, Schubauer-Berigan et al [5] published an update to the National Institute of Occupational Safety and Health (NIOSH) source cohort study within which the case-control study of Sanderson et al was nested; their update summarizes most of the available data that exist on occupational exposure to beryllium and lung cancer risk. The evidence for a relation between beryllium exposure and lung cancer from their study is mixed. They used several metrics to measure beryllium exposure, with varying lag times, and both internal and external comparisons. The strongest association they reported was for workers with 35 or more years since first employment in the beryllium industry, compared with workers with fewer than 15 years since first employment, for which the standardized rate ratio (SRR) was 3.68. For cumulative exposure, however, the SRR comparing highest quartile with the lowest was only 1.12, based on a 10-year lag, although this value increased to 1.97 after excluding short-term workers. Other measures, such as employment duration (10-year lag) and maximum exposure (unlagged), were also reported, but these associations were smaller.

In this paper, we show that the reported analysis was affected by sparse-data bias, which accounts for most of the reported association between time from first employment in a beryllium plant and lung cancer in the NIOSH cohort study. The NIOSH study included 9199 workers followed from 1940 through 2005 for fatal lung cancer and other endpoints. It encompassed more than 350,000 person-years of follow-up, during which 545 cases of lung cancer were identified. With all these data, it may seem odd that there would be a problem with sparse data. The problem arises because the data were stratified into hundreds of cells, and inferences were drawn from datasets that were populated mostly with zero cell counts.

* Corresponding author. RTI Health Solutions, 200 Park Offices Drive, Research Triangle Park, NC 27709.
E-mail address: KRothman@rti.org (K.J. Rothman).

1047-2797/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.annepidem.2012.11.005

All ratio measures based on counts, such as those reported by Schubauer-Berigan et al [5], are positively biased on the arithmetic scale because errors that exaggerate the ratio are larger than errors that underestimate it. For a single table, a ratio measure that has a nonzero probability of having a zero denominator will consequently have infinite bias, because the mean estimate will be infinity. (For example, consider the ratio of heads to tails in 10 tosses of a fair coin; although the expectation of the proportion heads is 50%, the expectation for the ratio of heads to tails is infinity, because the outcome of 10/0 = infinity is averaged with other outcomes to get the expected value. Even if the outcome of 10 heads were disallowed by recoding it to 9, the ratio of heads to tails would have an expected value above 1.) With stratified analysis, any unconfounded summary measure is essentially a weighted average of stratum-specific estimates, and is subject to the same problem, which can be exaggerated if the numbers within strata are small. Greenland [6] has suggested that such small-sample bias may be more prevalent than commonly realized. Various solutions may be employed to correct for sparse-data problems in stratified data. It may be possible to collapse neighboring strata without introducing substantial residual confounding. A regression model can be employed that avoids stratifying a continuous variable such as age. In addition, various corrections can be applied to mitigate the bias; two possibilities are the use of the Firth correction [7] and the use of data augmentation to implement Bayesian shrinkage for sparse data [8,9].

We examined the results of Schubauer-Berigan et al in several ways. After replicating their results, we inspected their stratified data, a step that reveals the sparse-data problem. Because the sparse-data problem arises from a combination of fine stratification of the data coupled with non-overlapping exposure distributions, and the purpose of the stratification is to control confounding, we then reanalyzed their data to explore the amount of confounding as well as the magnitude of the sparse-data bias. Finally, we conducted simulations using the actual cohort experience with respect to beryllium exposure, but randomly simulating lung cancer deaths, which enabled us to see the extent to which the finely stratified analysis biased the results.

Methods

NIOSH kindly supplied a copy of the dataset used for this analysis. To verify the data, we first attempted to replicate the results reported in the NIOSH paper. In the NIOSH paper, two analytic approaches were used, both based on stratification to control confounding. One involved external comparison with U.S. mortality data, calculating standardized mortality ratios by exposure level for the cohort. The other was an internal comparison across approximate exposure quartiles in the data, using standardization to summarize the results across strata. The standard used to weight the stratum-specific results was the distribution of person-time in the entire cohort across categories of the stratification variables. There were three stratification variables used: Age, calendar year, and race. Both age and calendar year were categorized into 5-year intervals. For age, there were 16 categories ranging from a low of 10 to 14 years, which had very little person-time, to a high of 85 or older. For calendar year, there were 14 categories, starting with 1940 to 1944 and going to 2005 to 2009. There were two categories of race. The data were further divided by exposure level into approximate quartiles. Several exposure metrics were used; these included employment duration, time since first employment, cumulative beryllium exposure, and maximum beryllium exposure. Most of our analyses focused on time since first employment, the measure that had the largest SRR for lung cancer death (3.68) reported by Schubauer-Berigan et al. Following the approach of Schubauer-Berigan et al., we classified person-time into four approximately equally sized categories of time (in years) since first exposure, which were bounded as follows from lowest to highest: [0,15) [15,25) [25,35) [35,80). All analyses conducted by NIOSH used publically available cohort analysis software, the Life Table Analysis System (LTAS.NET) [10–12]. We used LTAS.NET for verification but also wrote our own software as a check on LTAS.NET. LTAS.NET uses standard stratification methods to control confounding, coupled with standardization (“direct standardization”) to summarize effects across strata.

After verifying the integrity of the data and the calculations reported by NIOSH, we inspected the strata to assess the distribution of information across exposure levels and strata. We tried alternative stratification schemes to deal with strong confounding, applying the same statistical methods used in LTAS.NET. We used Mantel-Haenszel methods as an alternative to standardization in some calculations. We also fit a Poisson regression model as an alternative to stratification to control confounding without the sparse-data problems inherent in the stratified analysis. In this model, we included terms for age, age-squared, age-cubed, year, year-squared, year-cubed, and race. Regression modeling can also be affected by bias from sparse data, however, so, in alternative analyses, we fit the Poisson regression using the Firth [7] correction, and we used Greenland’s [8,9] approach of Bayesian shrinkage based on data augmentation. For the data augmentation, for each coefficient we used a weak prior that added two pseudo-records, each with one case, and added an indicator for each pair, corresponding to a prior that offers 95% certainty that the rate ratio (RR) is between 0.026 and 39 [9].

In addition, we conducted a series of simulations of the lung cancer findings, by taking the cohort experience and simulating the occurrence of lung cancer deaths. We obtained cause-, calendar year-, race-, and age-specific population mortality rates used with LTAS.NET. For each cohort member, date and cause of mortality was determined randomly by applying the mortality rates to the corresponding amounts of person-time. Simulations using time since first employment and duration of employment were based on all 9199 cohort members; simulations using cumulative exposure and maximum exposure were based on the subcohort of 5436 workers employed at three plants for which linkage with work history data was possible. The simulation methods are described in more detail in the Appendix. The simulation process guaranteed no association between beryllium exposure and death from lung cancer, so that any departure from a null result in the data analysis reflects bias in the methods applied or the estimator used.

In attempting to verify the NIOSH results, we discovered a small problem in the way that follow-up had been defined in the NIOSH study, a problem that led to a “time-loop” [13] and the exclusion of immortal person-time, inflating the estimated rates. The NIOSH protocol considered workers lost to follow-up if they left employment alive and were not ascertained as a death in subsequent follow-up. Thus, the occurrence of a death determined whether the person-time of these retired workers would be included in the study, introducing a selection bias that inflated the mortality rates. This bias would only affect the rate ratio for beryllium exposure and lung cancer mortality if retirement time is related to exposure. That is the case, however, for time since first employment, because the excluded person-time is concentrated among those with the longest time since first employment, inflating the mortality rate most for those in the highest category of time since first employment. This is a time-loop because whether a worker was actually lost to follow-up at the time of retirement from work depended on a future event, whether the worker was ascertained to have died in the Social Security Administration database or the National Death Index (NDI). Fortunately, because follow-up was lengthy and the NDI is nearly complete, only 123 workers were affected by this