

Abstract

Background: When releasing vital registration data, data authorities are known to suppress parts of the data to protect privacy. Suppressed data lose data entries below a suppression threshold, which damages the accuracy of modeling conducted using this data and often makes modeling impossible. This study aims to develop methods that mathematically compensate for the effects of the omitted data.

Methods: The Annual Vital Statistics Report of 2015 from the Ministry of Health, Labour and Welfare of Japan was used to generate masked datasets (with suppression thresholds at 3, 5, and 10). A test model used was a basic model of mortality risk by five-year age group, treated as a categorical variable. In the scaling and filling method, an optimal scaling value was calculated for each threshold and all observations were scaled down by it and rounded. Masked values were given the value of one. In the imputation method, a multiple imputation using constrained Poisson distributions constructed with the true age-specific mortality rates from the unmasked dataset and with mortality rates from the masked datasets. Both methods were evaluated on the effects on coefficients. Bias (average, minimum, and maximum variances) and inflation of standard error (average % change) were reported.

Results: By masking, the dataset with the suppression threshold at 3 lost about 1% of death counts and over 10% of observations. Amount of omitted death counts and observations jumped up with higher thresholds. In the scaling and filling method, the dataset with the threshold at 3 showed average coefficient bias of 1 to 6% and average inflation of standard error of 7 to 9%. However, average coefficient bias rose to 45 to 58% and average inflation of

standard error was 35 to 28% with the datasets with the threshold at 10. In the imputation method, the bias with true mortality rates declines with the increasing threshold but the standard error grows rapidly. However, the increase in standard error inflation was less than 15%. The estimates imputed using a constrained Poisson distribution with the masked data showed a rapid increase in bias with the threshold, and orders of magnitude larger bias for the higher thresholds. Overall, the imputation method showed better adjustment based on minimal bias in coefficients, especially when the true mortality rates were available. However, when the threshold was 3, the results were comparable between the scaling and filling and the imputation.

Conclusion: It was found that the data suppression damages the accuracy of data by causing bias. But with adjustment methods proposed in this paper, compensation for the omitted data is possible with limited bias. Therefore, cooperation from data authorities such as providing additional information of scaling values or age-specific mortality rates, helps improve the quality of suppressed data analyses.

Keywords: vital registration data, data suppression, data masking, privacy protection, imputation, scaling.