The NCI Patterns of Care Studies Small Numbers Reporting Guidelines

Benmei Liu and Michael T. Halpern,

Division of Cancer Control and Population Sciences

National Cancer Institute, Bethesda, MD, USA

Contact Information: Dr. Benmei Liu, liub2@mail.nih.gov

August 22, 2022

Purpose

The National Cancer Institute (NCI) Patterns of Care (POC) studies data have enabled investigators to examine disparities in receipt of cancer therapies, diagnostic tests, and biomarkers by age, race/ethnicity, and residential urbanicity/rurality. They also have been used to show trends in cancer therapy and survival over time for cancer sites with multi-year data[1].

The release of public health statistics from data sources such as POC relies on the degree of confidence that this information is statistically reliable and protects the privacy of individuals in published reports and online query systems. Estimates with small numerators and/or denominators constitute special challenges and need to be handled with caution before release[2-4]. The purpose of this document is to provide specific guidelines and recommendations on how to report estimates involving small numbers for the POC studies. Here we refer to less than 20 individuals in the numerator and/or less than 100 in the denominator (i.e., unweighted sample size) as small numbers.

Reporting guidelines

Small cell counts and small numerator counts

The unweighted counts will be presented only for univariate totals that are not less than five (e.g., number of patients by race/ethnicity, number receiving each systemic therapy, etc.) when presenting national-level data (i.e., data that are not stratified by SEER registry). Counts between one and four should be suppressed and be reported as “<5”. When presenting data from a specific SEER registry, unweighted counts will be presented only for univariate totals that are not less than 11. Counts between one and ten should be suppressed and be reported as “<11”. Suppression is not needed for unweighted counts of zero. Similar suppression methods are recommended by CMS[5].

Estimates with small denominators

Estimates (proportions, rate ratios, means, regression coefficients, etc.) with small denominators or sample sizes, weighted or unweighted, are associated with greater risk of disclosure and less reliability. Aggregating multiple years and geographical locations or collapsing across categories may be used to increase denominator size. In addition, weighted 95% confidence intervals may be presented for the weighted estimates. Statistical reliability of these estimates while nonzero may be examined by calculating the relative standard error (RSE), also known as the coefficient of variation (CV). The RSE is the standard error of an estimate (proportion, ratio, mean, regression coefficient, etc.) divided by the estimate itself, usually expressed as a percentage. That is, $RSE = \frac{standard\ error}{estimate} \times 100$. Estimates with RSE$\geq 25\%$ are subject to high sampling error and should be reported with caution. We recommend that estimates with $25\% \leq RSE < 50\%$ be interpreted with caution and marked in all publications and presentations as potentially uncertain. We recommend suppressing estimates with RSE$\geq 50\%$ due to high instability.

Small proportions

For proportions that are close to 0 or 1 ($0 < p < 0.1$ or $0.9 < p < 1$), the Clopper-Pearson (C-P) interval or Wilson score interval with continuity correction may be reported instead of normal confidence intervals. These intervals produce coverage rate closer to the nominal 0.95 than do the normal confidence intervals for proportions that are close to 0 or 1 [6-8].

Zero (or one) proportions

For proportions equal to zero or one, estimates need to be interpreted with caution because the associated normal standard errors are equal to zero. The "rule of three" strategy[9] may be used to compute the upper bound of the 95% confidence interval for zero proportions or the lower bound of the 95% confidence interval for proportions equal to 1 if associated sample size is >30. That is, for zero proportions report $[0, \frac{3}{n^*}]$ and for proportions equal to 1 report $[1 - \frac{3}{n^*}, 1]$ as the 95% confidence intervals, where n^* is the effective sample size (sample size divided by the design effect due to weighting). The design effect due to weighting can be estimated using $1 + CV^2$, where $CV^2 = \frac{variance}{mean}$ is the relative variance of the survey weights in the sample and can be estimated using SAS PROC UNIVARIATE. For example, assume there are zero events in a domain of interest (e.g., non-Hispanic black) which has a sample size of 35. Further assume the design effect is 2.1, so the effective sample size is 16.67. Following the formula, the confidence interval for the zero proportion is [0, 0.18].

For zero (or one) estimates with sample size≤ 30, we recommend the estimates be interpreted with caution and marked in all publications and presentations as potentially uncertain.

References

1. National Cancer Institute. Patterns of Care Studies. Accessed August 22, 2022. https://healthcaredelivery.cancer.gov/poc/

2. Rhode Island Department of Health. Rhode Island small numbers reporting policy. Accessed August 22, 2022. https://health.ri.gov/publications/policies/SmallNumbersReporting.pdf
3. Wasserman C, Ossiander E. Department of Health Agency Standards for Reporting Data with Small Numbers. Accessed August 22, 2022. https://www.doh.wa.gov/Portals/1/Documents/1500/SmallNumbers.pdf
4. Klein R, Proctor S, Boudreault M, Turczyn K. Healthy People 2010 criteria for data suppression. *Healthy People 2010 Stat Notes*. 2002;Jul(24):1-12.
5. CMS Researc Data Assistance Center. CMS Cell Size Suppression Policy. Updated 05/08/2017. Accessed August 22, 2022. https://resdac.org/articles/cms-cell-size-suppression-policy
6. Tobi H, van den Berg P, de Jong-van den Berg L. Small proportions: what to report for confidence intervals? *Pharmacoepidemiol Drug Saf*. 2005;14(4):239--247. doi:10.1002/pds.1081
7. Clopper C, Pearson ES. "The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26 (4):404–413. doi:10.1093/biomet/26.4.404
8. Newcombe R. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17(8):857-72. doi:10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e
9. Eypasch E, Leferinga R, Kuma CK, Troidl H. Probability of adverse events that have not yet occurred: a statistical reminder. *BMJ*. 1995;311:619.