ORIGINAL RESEARCH & CONTRIBUTIONS

On the Use of Sampling Weights for Retrospective Medical Record Reviews

Ernest Shen, PhD[1]

Perm J 2020;24:18.308

E-pub: 06/10/2020

https://doi.org/10.7812/TPP/18.308

ABSTRACT

Retrospective medical record review is often used to answer the "why" questions that statistical modeling cannot. In addition to its utility as an explanatory tool, it can be used to generate hypotheses using available retrospective data and so is a convenient guide for developing future prospective studies. A recent review of articles that used the retrospective medical record review method listed 10 best practices that ought to be followed. However, an issue that is not listed is the use of sampling weights, which are important when one can only conduct retrospective medical record review for a sample of the target population. Although that review acknowledged the importance of carefully selecting a sampling strategy for such a scenario and indeed had outlined 3 commonly used sampling methods (convenience, simple random, and systematic), the authors say nothing of the use of sampling information at the data analysis stage. This article aims to fill that gap and to demonstrate why the use of sample weights ought to be another best practice to add to the list by reviewing well-known theoretical details and some published data analysis examples.

INTRODUCTION

In the current era of electronic health records and big data analytics, there is still a place for retrospective medical record review (RMRR).[1] RMRR is often used to answer the "why" questions that structured data and statistical modeling usually cannot. RMRR can also be used to capture ill-defined or nondiscrete variables, validate structured data, and generate hypotheses based on qualitative data. It is also used to validate phenotypes and outcomes that have been ascertained via International Classification of Diseases codes using administrative databases or from natural language processing methods.[2] There are explicit RMRR best practices, such as the use of standardized data abstraction forms and assessment of interrater reliability when multiple reviewers are involved.[3] These RMRR guidelines also acknowledge the importance of considering sampling issues a priori as well as conducting a power analysis in the design of a sampling strategy. Unfortunately, no recommendations are available to guide the use of the sampling information at the data analysis stage of a RMRR study. We focus on addressing this gap, demonstrating why the use of sample weights should be added to the list of best practices.

Many studies have performed sampling from well-defined target populations for RMRR; many of these studies report sample statistics and tests to compare subgroups of the population while ignoring the sampling strategy in the analysis.[4-7] Of the commonly used sampling methods, the most obvious choice is for RMRR studies to conduct a simple random sample (SRS) to select records for review. However, summary data are often not presented with measures of sampling variability to report the uncertainty in the sample statistics. Moreover, not applying the sampling weights in even a descriptive analysis can lead to misleading findings, especially when sampling is performed in a stratified manner or clinically relevant strata exist within a population from which an SRS was drawn.

This issue was explicitly acknowledged in the article by Belletti et al,[4] who used RMRR to ascertain adherence to primary care guidelines in treatment of patients with chronic obstructive pulmonary disease. They found that "37% [of participants] had documentation of some level of pulmonologist care," and thus their sample statistics on guideline adherence may have not accurately reflected the rates in primary care settings, and so a sampling design that somehow incorporated other care settings (ie, specialty care visits with a pulmonologist) may have led to more generalizable results. This finding reflects the suggestion of Worster and Haines[8] that data arising from RMRR "are more likely to yield less valid and reliable study results than those based on relatively objective data sources," but appropriate incorporation of sampling weights into one's analysis can allow for valid population estimates and further describe the uncertainty in the sample statistics. The importance of sampling weights has been long recognized in other areas of medical research, such as national survey research, as recommended for the Centers for Disease Control and Prevention's annual National Health and Nutrition Examination Survey (NHANES).[9] Although the goals of national surveys may differ from those of RMRR, we posit that it is similarly important to account for sampling design when analyzing RMRR data as well.

The basic idea is that one should already know the appropriate weights on the basis of the sampling design and then apply them as weights in whatever the data analysis might be, for example, calculating a weighted mean for each of several strata in a stratified random sample. A general approach to constructing such weights is the well-known Horvitz-Thompson estimator,[10] which has been extended to the analysis of health survey data in cancer research.[11] This article provides 2 pragmatic examples of how to account for the sampling strategy in the analysis stage, using long-standing methods from the survey sampling literature,[12] and demonstrates the potential consequences of reporting sample statistics without accounting for the sampling weights. Although the

Author Affiliations

[1] Kaiser Permanente Department of Research and Evaluation, Pasadena, CA

Corresponding Author

Ernest Shen, PhD (ernest.shen@kp.org)

Keywords: big data analytics, electronic medical record, retrospective medical record review, sampling methods

methods presented in this article are specific to estimating population frequencies and proportions, analogous methods exist that apply to means, ratios, regression, and differences.[11,12]

PRAGMATIC EXAMPLES AND THE BASIC PRINCIPLE

Example 1

Our first example deals explicitly with sampling weights of clinical encounters for acute sinusitis (AS). RMRR was used to assess the rates of guideline-concordant care for patients treated in 3 different care settings: Emergency Department, Primary Care Department, and Urgent Care.[13] For this study, the investigators used a stratified random sampling approach by which they randomly selected 100 medical records to review from each of the 3 care settings to ascertain whether recommended care had been delivered for specific "AS encounters … which resulted in antibiotics filled, the performance of CT [computed tomography] imaging or both." It was important to estimate the proportions of length of symptoms (LOS) in the patients presenting at the 3 care settings because that was a key variable that indicated a recommendation for antibiotic treatment (ie, LOS ≤ 7 days). The study followed best practice recommendations[3] for RMRR and showed excellent interrater reliability (93.3% agreement) with both raters using the same protocol and data abstraction forms. In addition, the study team knew the sampling weights based on the design and incorporated them when conducting descriptive analyses.

Example 2

A second example involved the assessment of guideline-concordant use of imaging for staging of early-stage breast cancer in patients at low risk for metastasis[14] as recommended by the American Society of Cancer Oncology.[15] The authors also used stratified random sampling, but this time the population stratum corresponded to 3 different types of imaging, with different numbers of medical records randomly sampled from each group of imaging within each study site: CT only and radionuclide bone scan or positron emission tomography. As with the study above, the authors carefully vetted the target population and incorporated the known sampling weights in descriptive analyses.

CALCULATING WEIGHTS FOR A STRATIFIED RANDOM SAMPLE

We demonstrate the basic principle here by using the above AS example, and all calculations can be performed by hand or in any spreadsheet program (we used Microsoft Excel, Microsoft Corp, Redmond, WA), and the following notation illustrates how simple it can be to incorporate sampling weights.

(1) $\hat{p}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$

(2) $\hat{V}\left(\hat{p}_j\right) = \left(\frac{N_j - n_j}{N_j - 1}\right)\left(\frac{\hat{p}_j(1-\hat{p}_j)}{n_j}\right)$

(3) $\hat{p}_{overall} = \frac{1}{N}\sum_{j=1}^{3}\sum_{i=1}^{n_j}\left(\frac{N_j}{n_j} x_{ij}\right)$

(4) $\hat{V}\left(\hat{p}_{overall}\right) = \frac{1}{N^2}\sum_{j=1}^{3} N_j^2 \hat{V}\left(\hat{p}_j\right)$

Starting with equation 1, let n_j denote the sample size for population stratum j and x_{ij} be an indicator of the outcome of interest for person i in stratum j (eg, x_{ij}=1 if LOS ≤ 7 and 0 otherwise). If we let j index the different care settings, p_j denotes the sample proportion of the outcome in care setting j. Variances of the p_j are estimated using equation 2, where N_j and n_j are the population and sample sizes for group j, respectively. The 95% confidence intervals (CIs) can then be computed for each care setting using the normal approximation to the Binomial distribution given by $p_j \mp 1.96*[V(p_j)]^{1/2}$, where 1.96 is the 97.5th percentile of a standard normal distribution (ie, with a mean of 0 and a variance of 1), and the stratum-specific variances are weighted by a ratio of N_j and n_j. The weighted overall sample proportions and variances for the population strata are estimated using equations 3 and 4, respectively. Estimates of population parameters are accented with carets (ie, ^).

Although the same formulas apply to SRS, in either case the weights are determined by the sampling fractions of a properly conducted sampling scheme as discussed elsewhere.[2] Instead of sampling from different population strata $j = 1 \ldots c$, one simply modifies equations 1 to 4 by dropping the index j. Thus, the weights are the same as for a stratified random sample with c = 1, and equations 1 and 3 become redundant because $p_{overall} = p$. Such a strategy would be relevant, for example, had the AS study only been interested in the primary care setting.

RESULTS

Table 1 gives the raw sample proportions that ignore the sampling weights in the AS study, where the j from the above equations index the 3 care settings. The sampling fractions (N_j/n_j) are 6.7 for Emergency Department, 390.4 for Primary Care Department, and 93.9 for Urgent Care. This means, for example, that each patient selected from primary care represents approximately 391 other Primary Care patients and contributes approximately 4 times more weight (or, alternatively, information) to the population proportion than one seen in urgent care. To obtain the weighted population proportion of LOS of 7 days or less, for example, one divides the sum of the product of the care-specific totals and their weights by the total population size: 74×(601/90) + 37×(32,400/83) + 64×(9114/97) = 20,924/42,115 = 0.49. Repeating this for the other categories gives the weighted proportions (and 95% CIs) of LOS for randomly selected patients with AS treated in different care settings, shown in Table 2. Comparing those proportions between Tables 1 and 2 clearly illustrates that the overall proportion of LOS of 7 days or more *should* be closer to the Primary Care Department proportion, despite the Emergency Department proportion being nearly double, because of the underlying distribution of visits across settings.

We see a similar story in the breast cancer example.[14] In the breast cancer example, sampling fractions correspond to the different sets of imaging types that defined the strata in the sample, which were 15.6 for CT, 13.6 for positron emission tomography or bone scan, and 15.8 for multiple imaging techniques. Similarly, for example, each medical record reviewed from the set of records from all patients who underwent CT represented approximately 16 other patients. Unlike the AS example, however, the patient medical records sampled from each imaging type are all roughly equally weighted and informative. Following the same procedure as in equations 1 to 4

Table 1. Raw counts and unweighted proportions (95% CIs) of LOS for randomly selected patients with acute sinusitis treated in different care settings, determined from medical record review, Kaiser Permanente Southern California, 2012 (N = 100 per setting)

LOS, d	ED (n = 601)	PC (n = 32,400)	UC (n = 9114)	Sample proportion (95% CI)
≤ 7	74 (0.82)	37 (0.45)	64 (0.66)	0.65 (0.59-0.71)
8-13	6 (0.07)	14 (0.17)	9 (0.09)	0.11 (0.07-0.14)
≥ 14	10 (0.11)	32 (0.38)	24 (0.25)	0.25 (0.19-0.30)
Total	90	83	97	270

CI = confidence interval; ED = Emergency Department; LOS = length of symptoms; PC = Primary Care; UC = Urgent Care.

Table 2. Weighted proportions (and 95% CIs) of LOS for randomly selected patients with acute sinusitis treated in different care settings, determined from medical record review, Kaiser Permanente Southern California, 2012 (n = 100 per setting)

	Weighted proportion (95% CI)			
LOS, d	ED	PC	UC	Total
≤ 7	0.82 (0.75-0.89)	0.45 (0.34-0.55)	0.66 (0.56-0.76)	0.49 (0.41-0.58)
8-13	0.07 (0.02-0.12)	0.17 (0.09-0.25)	0.09 (0.03-0.15)	0.15 (0.09-0.226)
≥ 14	0.11 (0.05-0.17)	0.38 (0.28-0.49)	0.25 (0.16-0.34)	0.35 (0.27-0.446)

CI = confidence interval; ED = Emergency Department; LOS = length of symptoms; PC = Primary Care; UC = Urgent Care.

described above, the unweighted proportion of inappropriate imaging works out to 9% vs 16% when weighted.

For the AS example, had the unweighted proportions with LOS of 7 days or less been reported and the overrepresentation of Emergency Department and Urgent Care visits in the sample vs Primary Care been ignored, the population proportion would have been overestimated by roughly 32%. Notably, the 95% CIs for the weighted and unweighted proportions of patients with LOS of 7 days or less do not overlap, providing further evidence of the danger of ignoring the sampling design in the data analysis. In the breast cancer example, the unweighted sample proportion of all imaging performed for surveillance in the cohort appears to underestimate the population proportion by nearly half (9% unweighted vs 16% weighted). In both cases, assuming SRS and then reporting the unweighted sample proportions could be misleading because of ignoring the underlying distribution of the outcome across those population strata.

DISCUSSION

This work demonstrates how established survey sampling methods can be incorporated into a RMRR at the analysis stage of a study. We argue that this is a necessary step in RMRR studies that use some form of random sampling, which we have highlighted with practical examples. First, a simple method can use estimates of population parameters along with measures of variability for RMRR sample data to appropriately account for sampling design using sample weights. Second, the use of sample weights can offer more valid estimates of population parameters and variances. Third, stratified random sampling can afford certain efficiency advantages over SRS because more information can be obtained per unit sampled when there are important subgroups in the target population. For example, use of SRS may miss patients from clinically meaningful subgroups of the target population, resulting in findings that may not generalize to the target population as in the study by Belletti et al.[4] However, even then one could *still* use the sampling fractions in the random sample in the data analysis assuming one had such information available (eg, in Table 1 of the AS study). In such situations, one can be easily misled by sample statistics that do not appropriately account for sampling design because they may not always reflect quantities of the target population or properly account for underlying population strata.

However, a few important limitations are worth stating. An important caveat is that we are working with the assumption that a binomial random variable (eg, number of patients with AS presenting with LOS ≤ 7 days) can be approximated using the normal distribution, although one could circumvent this assumption by computing exact binomial CIs. A related issue is that the stratum-specific variances depend explicitly on their proportions, and so if one performs SRS and then tries to construct postsampling stratified weights on the basis of underlying population strata, one *must* take care to use the correct variance formulas (ie, equations 2 and 4) to obtain an unbiased estimate of the population parameter. Lastly, the sampling method used may be constrained by the available resources or assume that the cost (in time and effort) of performing medical record reviews is the same for all population strata. In the AS case, this assumption would not be true if reviewing medical records for the Emergency Department setting was more complicated and required more time than for Primary Care Department, and the costs would therefore vary. In that instance, the different costs across the 3 settings could be accounted for in the determination of sample size for each group so as to minimize the cost for a fixed level of variability or to minimize the variability for a fixed cost.[12]

In addition to applying sample weights to the analysis of RMRR data arising from a given random sampling procedure, articles reporting results from such studies should also report complete information on the sampling frame, such as the total number of eligible participants. This information would allow others to assess the degree to which biases or imprecisions in sample statistics or measures of association and lack of generalizability to the target population may be attributable to sampling design. For example, existing methods from the survey sampling literature could allow one to make such assessments, such as the standardization methods used for the NHANES[9] or for health surveys in general.[11] The same could be done to compare the study and target populations by applying sample weights as

we described or by comparing the sampled population to the target population as in 1 study that reported all details of the sampling frame.[16]

This study has demonstrated that relatively straightforward applications of existing survey sampling methods can improve the quality of reporting for RMRR studies by providing more representative estimates of population parameters, along with corresponding estimates of variability. Our study also demonstrates how using such methods can help ensure that underlying population subgroups that may have been overrepresented or underrepresented in a SRS do not bias the parameter estimates, both in one's own study and in the evaluation of others' studies. As discussed earlier, such sampling information can even be used *after* data collection has already been completed to correct for having oversampled or undersampled from some population strata using SRS. It is especially helpful when one is only able to review a limited number of medical records because of resource or time limitations because one can use strategic sampling choices along with the corresponding weights to obtain as much possible information from a limited sample.

CONCLUSION

We recommend that future RMRR studies apply established survey sampling methods in the data analysis stage to improve the quality of their methods and the accuracy of their results. ❖

Acknowledgments

The following people are acknowledged for their many past and ongoing collaborations on such retrospective medical record review studies, the earliest of which provided the impetus for this article. Adam Sharp, MD, MS; and Erin Hahn, PhD, MPH, led the studies and authored the publications from which the example data originated. Corrine Munoz-Plaza, MPH, conducted all the medical record reviews for both example data sources, and Janet Shinn, MS, extracted all relevant structured data from the Kaiser Permanente Southern California electronic medical record system. Michael Gould, MD, contributed to the conception and design of both example studies. All the aforementioned people also provided valuable feedback on an early draft. Finally, thanks to the editors and especially 4 anonymous reviewers whose comments and questions led to a considerably improved article.

Laura King, ELS, performed a primary copy edit.

How to Cite this Article

Shen E. On the use of sampling weights for retrospective medical record reviews. Perm J 2020;24:18.308. DOI: https://doi.org/10.7812/TPP/18.308

References

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf Sci Syst 2014 Feb 7;2(1):3. DOI: https://doi.org/10.1186/2047-2501-2-3 PMID:25825667
2. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 2012 Aug;7(8):1257-62. DOI: https://doi.org/10.1097/JTO.0b013e31825bd9f5 PMID:22627647
3. Vassar M, Holzmann M. The retrospective chart review: Important methodological considerations. J Educ Eval Health Prof 2013 Nov 30;10:12. DOI: https://doi.org/10.3352/jeehp.2013.10.12 PMID:24324853
4. Belletti D, Liu J, Zacker C, Wogen J. Results of the CAPPS: COPD—assessment of practice in primary care study. Curr Med Res Opin 2013 Aug;29(8):957-66. DOI: https://doi.org/10.1185/03007995.2013.803957 PMID:23663130
5. Hodgkiss-Harlow CJ, Eichenfield LF, Dohil MA. Effective monitoring of isotretinoin safety in a pediatric dermatology population: A novel "patient symptom survey" approach. J Am Acad Dermatol 2011 Sep;65(3):517-24. DOI: https://doi.org/10.1016/j.jaad.2010.06.040 PMID:21632153
6. Stracciolini A, Casciano R, Levey Friedman H, Stein CJ, Meehan WP 3rd, Micheli LJ. Pediatric sports injuries: A comparison of males versus females. Am J Sports Med 2014 Apr;42(4):965-72. DOI: https://doi.org/10.1177/0363546514522393 PMID:24567251
7. Vreeman RC, Scanlon ML, Mwangi A, et al. A cross-sectional study of disclosure of HIV status to children and adolescents in western Kenya. PLoS One 2014 Jan 27;9(1):e86616. DOI: https://doi.org/10.1371/journal.pone.0086616 PMID:24475159
8. Worster A, Haines T. Advanced statistics: Understanding medical record review (MRR) studies. Acad Emerg Med 2004 Feb;11(2):187-92. DOI: https://doi.org/10.1111/j.1553-2712.2004.tb01433.x PMID:14759964
9. Curtin LR, Mohadjer LK, Dohrmann SM, et al. National Health and Nutrition Examination Survey: Sample design, 2007-2010. Vital Health Stat 2 2013 Aug;(160):1-23. PMID:25090039
10. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952 Apr;47(260):663-85. DOI: https://doi.org/10.1080/01621459.1952.10483446.
11. Graubard BI, Korn EL. Analyzing health surveys for cancer-related objectives. J Natl Cancer Inst 1999 Jun 16;91(12):1005-16. DOI: https://doi.org/10.1093/jnci/91.12.1005 PMID:10379963
12. Scheaffer RL, Mendenhall W 3rd, Ott L. Elementary survey sampling. Pacific Grove, CA: Duxbury Press; 1996.
13. Sharp AL, Klau MH, Keschner D, et al. Low-value care for acute sinusitis encounters: Who's choosing wisely? Am J Manag Care 2015 Jul;21(7):479-85. PMID:26247738
14. Hahn EE, Tang T, Lee JS, et al. Use of posttreatment imaging and biomarkers in survivors of early-stage breast cancer: Inappropriate surveillance or necessary care? Cancer 2016 Mar 15;122(6):908-16. DOI: https://doi.org/10.1002/cncr.29811 PMID:26650715
15. Schnipper LE, Smith TJ, Raghavan D, et al. American Society of Clinical Oncology identifies five key opportunities to improve care and reduce costs: The top five list for oncology. J Clin Oncol 2012 May 10;30(14):1715-24. DOI: https://doi.org/10.1200/JCO.2012.42.8375 PMID:22493340
16. Turnbull K, Nguyen LN, Jamieson MA, Palerme S. Seasonal trends in adolescent pregnancy conception rates. J Pediatr Adolesc Gynecol 2011 Oct;24(5):291-3. DOI: https://doi.org/10.1016/j.jpag.2011.04.005 PMID:21715196