Epidemiology: Open Access

Sinaii et al., Epidemiol 2014, 4:3
DOI: 10.4172/2161-1165.1000155

Research Article Open Access

Assessing the Impact of Misclassification when Comparing Prevalence Data: A Novel Sensitivity Analysis Approach

Ninet Sinaii[1,2*], **Sean D Cleary**[3] **and Pamela Stratton**[2]

[1]*Biostatistics and Clinical Epidemiology Service, CC, NIH, Bethesda, MD, USA*

[2]*Program in Reproductive and Adult Endocrinology, Eunice Kennedy Shriver NICHD, NIH, Bethesda, MD, USA*

[3]*Department of Epidemiology and Biostatistics, School of Public Health and Health Services, The George Washington University, Washington, DC, USA*

***Corresponding author:** Ninet Sinaii, Biostatistics and Clinical Epidemiology Service, CC, NIH, Bethesda, MD, USA, Tel: 301-402-9364; Fax: 301-496-0457; E-mail: sinaiin@mail.nih.gov

Received date: January 11, 2014; **Accepted date:** April 25, 2014; **Published date:** April 30, 2014

Abstract

Background: A simple sensitivity analysis technique was developed to assess the impact of misclassification and verify observed prevalence differences between distinct populations.

Methods: The prevalence of self-reported comorbid diseases in 4,331 women with surgically-diagnosed endometriosis was compared to published clinical and population-based prevalence estimates. Disease prevalence misclassification was assessed by assuming over-reporting in the study sample and under-reporting in the general (comparison) population. Over- and under-reporting by 10%, 25%, 50%, 75%, and 90% was used to create a 5×5 table for each disease. The new prevalences represented by each table cell were compared by p-values, prevalence odds ratios, and 95% confidence intervals.

Results: Three misclassification patterns were observed: 1) differences remained significant except at high degrees (>50%) of misclassification; 2) minimal (10%) misclassification negated any observed difference; and 3) with some (25-50%) misclassification, the difference disappeared, and the direction of significance changed at higher levels (>50%).

Conclusions: This sensitivity analysis enabled us to verify observed prevalence differences. This useful, simple approach is for comparing prevalence estimates between distinct populations.

Keywords: Epidemiology; Comorbid diseases; Distinct populations

Introduction

Establishing differences in disease prevalence between populations is a common application of epidemiology. Disease prevalence data may be obtained using surveys, medical record reviews, and surveillance reporting, and thus disease may be over- or underestimated because of unmeasured confounding, misclassification (information bias), and selection bias. While medical researchers strive to collect valid and minimally biased data, missing or limited validation data can be an important obstacle in addressing the effect of misclassification. Analytic techniques may be employed to assess the uncertainty of study results and to correct for potential bias due to misclassification and therefore, are useful in interpreting whether significant differences are real. Sensitivity analysis may be used to quantitatively evaluate the effect of misclassification.

Various sensitivity analyses techniques use basic and matrix algebra to assess and correct for differential, non-differential, or simultaneous misclassification of exposure and disease on epidemiologic measures of association [1-6]. Predictive values are also used to adjust relative risk estimates and to correct for biases resulting from misclassification of outcome status [7,8]. In some instances, computer programs are used to perform more extensive analyses [9]. While these established techniques for conducting a "formal" sensitivity analysis are valuable, there are several important reasons why these comparisons may not be carried out. First, reliable estimates of sensitivity, specificity, and true disease frequency are often required, but may not be available. Second, these methods make assumptions about the data such as misclassification of only the outcome variable, sensitivity and specificity parameters that are the same for each comparison group, or misclassification that is considered in isolation from other forms of bias, such as selection bias or confounding. Third, these methods are not standardized and may be useful only with particular study designs, further hampering their appropriate use. Finally, the methodology is complex, such that most public health professionals or clinicians cannot undertake a sensitivity analysis without formal training in epidemiology or statistics [10].

The prevalence of published self-reported physician-diagnosed autoimmune, chronic fatigue syndrome, and fibromyalgia [11], as well as cancer, and infectious or endocrine diseases [12] in up to 4,331 women with surgically diagnosed endometriosis were compared to prevalences from studies published in the last 30 years. Comparing population prevalences obtained from clinical, population-based, or self-reported studies to those that are solely self-reported may present disparity not only due to differences in study methodology, but also inherent differences in the populations being compared. We assumed that women with endometriosis who self-report a diagnosis may believe they have a disease when they actually do not, therefore biasing prevalence estimates upward. Some diseases were rare and others, like

infectious diseases were more commonly reported, but are less specific and, perhaps, open to interpretation. In both instances, this may lead to overestimation of diagnoses.

By contrast, population disease prevalence estimates based on clinical or population-based studies may use more stringent definitions, which might bias prevalence estimates downward. These types of biases may result in conclusions of 1) a difference when one does not exist (Type I error), 2) no difference when there is one (Type II error), or 3) a difference in the opposite direction from the true difference. We therefore considered the degree of underestimation and overestimation of true disease prevalence because even modest amounts of error can profoundly affect results [13].

We developed a novel sensitivity analysis approach to determine the threshold of misclassification that would eliminate the observed differences between the disease prevalence, in two different populations, in this instance, for women with endometriosis and the general female population. This provided us with a visual and quantitative validation of the increased prevalence of comorbid diseases among women with endometriosis. Our method only requires a numerator and denominator for prevalence computation and does not rely on detailed information, assumptions, or complex methodology. Our goal was not to replace formal sensitivity analysis techniques, which should be carried out, when possible, but to offer a simple way to assess the impact of misclassification, and to verify study findings.

Materials and Methods

Prevalence estimates from up to 4,331 female members of the Endometriosis Association (International Headquarters, Milwaukee, Wisconsin) who reported surgical diagnosis of endometriosis and the physician diagnosis of comorbid diseases were compared to the general population [11,12]. Exemptions from Investigational Review Board reviews were granted by the Office of Human Subjects Research at the National Institutes of Health, Bethesda, Maryland, and the Committee on Human Research, The George Washington University, Washington, DC.

Disease prevalence in the general female population for systemic lupus erythematosus, Sjögren's syndrome, rheumatoid arthritis, multiple sclerosis, Hashimoto's thyroiditis/hypothyroidism, Graves' disease/hyperthyroidism, diabetes mellitus, chronic fatigue syndrome, and fibromyalgia were estimated from studies published between 1969 and 2001. Age-specific population estimates of breast cancer, ovarian cancer, non-Hodgkin's lymphoma, and melanoma were obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database. The remaining population prevalence estimates were obtained from published literature or sources such as the Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS). Studies were included if prevalence could be calculated for women by 1) providing the total number in the population for the denominator and 2) prevalence data for women. For some, denominators were determined using U.S. Census Bureau data. The published studies were pooled to derive disease prevalence using standardized medical definitions, patient interviews, clinical and laboratory evaluation, and self-reported surveillance data. Disease prevalences for women with endometriosis compared to the general female population prevalence estimates are presented in Table 1.

Prevalence per 1,000				
Disease	Women with endometriosis	Women in the general population	Prevalence Odds Ratio	95% Confidence Interval
Autoimmune Inflammatory Diseases				
Systemic Lupus Erythematosus	8.42	0.41	20.7c	14.3, 29.9
Multiple Sclerosis	5.16	0.73	7.1c	4.4, 11.3
Rheumatoid Arthritis	18.48	12.48	1.5c	1.2, 1.9
Sjögren's Syndrome	6.25	0.26	23.9c	15.5, 36.5
Cancers				
Melanoma	6.70	1.76	3.81c	2.60, 5.56
Breasta	3.69	6.82	0.54d	0.32, 0.90
Ovary	2.31	0.67	3.43c	1.74, 6.54
Non-Hodgkin's lymphoma	0.46	0.55	0.84	0.14, 3.37
Endocrine Diseases				
Diabetes Mellitus	15.22	13.50	1.1	0.9, 1.5
Hashimoto's thyroiditis/ Hypothyroidism	96.20	14.59	7.2c	6.4, 8.0

Graves' Disease/ Hyperthyroidism	17.12	19.74	0.9	0.7, 1.1
Addison's Diseaseb	2.31	0.09	---	---
Cushing's Syndromeb	0.92	0.00	---	---
Chronic Pain and Fatigue States				
Fibromyalgia	58.97	34.00	1.8c	1.6, 2.1
Chronic Fatigue Syndrome	46.20	0.26	180.5c	147.2, 242.0
Infectious Diseases				
Recurrent upper respiratory Infections	351.65	70.14	7.19c	6.73, 7.68
Candidiasis	376.51	374.88	1.01	0.87, 1.16
Recurrent vaginal infections	292.54	100.00	3.72c	3.48, 3.98
History of mononucleosis	137.61	900.00	0.02c	---
Other Diseases				
Mitral Valve prolapse	184.36	76.19	2.74c	2.32, 3.24
Congenital Birth Defects	27.25	30.00	0.91	0.75, 1.09

Table 1: Prevalence odds ratios of diseases among women with endometriosis and women in the general U.S. population.

a=note lower risk in women with endometriosis; b the prevalence in the general population was extremely low for meaningful POR and 95% CI calculations; c=p<0.001; d=p<0.01

For each disease found to be statistically significantly different in either direction, we propose selecting an appropriate range of under- and overestimation degrees to create an n by n table for comparing the prevalences of each disease (Appendix). In our study, we considered the prevalence from published studies to be underestimated by 10%, 25%, 50%, 75%, and 90% and our study population prevalence overestimated by 10%, 25%, 50%, 75%, and 90%, to create a 5 by 5 table.

The new general population and endometriosis prevalence generated by each cell in the table were then compared by Z-tests, and p-values were reported. To assess the magnitude of the differences and determine the direction of the effect, prevalence odds ratios (POR) and 95% confidence intervals (CI) were calculated. A p-value of less than or equal to 0.05, and a CI excluding 1.0 were considered statistically significant. These results were used to identify the threshold (a "line" connecting the cells), where statistically significant differences between the two groups reversed. The amount of misclassification required for results to change was subjectively defined as the midpoint along the threshold in the table. Generally, a low degree of misclassification was considered to be less than 50%, while misclassification greater than 50% was considered to be high.

Results

Table 2 displays the degree of overestimation in the study population of women with endometriosis and underestimation in the published studies that was necessary to negate the differences between the observed disease prevalence. For most diseases that had significantly different prevalences between the study sample and general population, a high degree (>50% in either direction) of misclassification was needed to eliminate these differences. However, for some diseases, a smaller degree of misclassification nullified the differences in prevalence between populations.

	Overestimation In Study Sample (%) Underestimation	in the General Population (%)
Chronic Fatigue Syndrome	> 90	> 90
Breast Cancerc	>90	> 90
Sjögren's Syndrome	>75	> 90
Systemic Lupus Erythematosus	>75	> 90

Multiple Sclerosis	>50	> 50
Recurrent upper respiratory infections	>50	> 50
Hashimoto's thyroiditis/ hypothyroidism	75	50
Recurrent vaginal infections	50	50
Melanoma	>25	>75
Mitral Valve Prolapse	25	50
Ovarian Cancer	25	50
Fibromyalgia	25	25
Rheumatoid Arthritis	a	a
Diabetes Mellitus	b	b
Graves' Disease/ hyperthyroidism	b	b
Addison's Disease	b	b
Cushing's Syndrome	b	b
Candidiasis	b	b
Congenital Birth Defects	b	b

Table 2: Misclassification threshold for eliminating the observed statistically significant heightened risk of diseases among women with endometriosis.

a=statistically non-significant at first level of misclassification

b=not applicable because statistically non-significant at observed crude level, or comparison with the general population could not be done

c=Breast cancer was observed to be statistically significantly lower in women with endometriosis

Overall, three different patterns were observed in the 5 by 5 tables used for our sensitivity analysis. Figure 1a represents an example of the general pattern for those diseases that reached non-significant levels with increasing degrees of misclassification, or those that never reached non-significant levels. Such a pattern was noted for chronic fatigue syndrome, breast cancer, Sjögren's syndrome, systemic lupus erythematosus, multiple sclerosis, and ovarian cancer, although the threshold varied for each disease. In the second scenario (Figure 1b), non-significant levels were reached with the lowest degree (10%) of misclassification, and the direction of significance was reversed at higher levels (>50%) of misclassification. This pattern was noted for rheumatoid arthritis only. In the third pattern (Figure 1c), statistical significance disappeared at a high level of misclassification, and then at even higher levels (e.g., 90% overestimation and >50% underestimation) the direction of significance was reversed. Recurrent upper respiratory infections, Hashimoto's thyroiditis/hypothyroidism, recurrent vaginal infections, melanoma, mitral valve prolapse, and fibromyalgia displayed such a pattern.

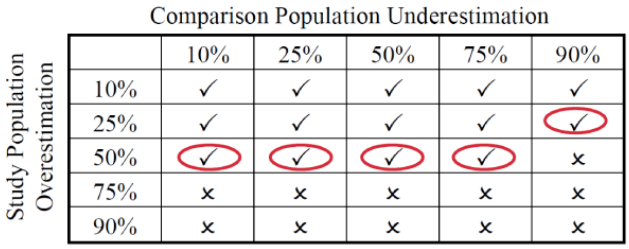

Figure 1a: Example of the first pattern: significant to non-significant or always significant.

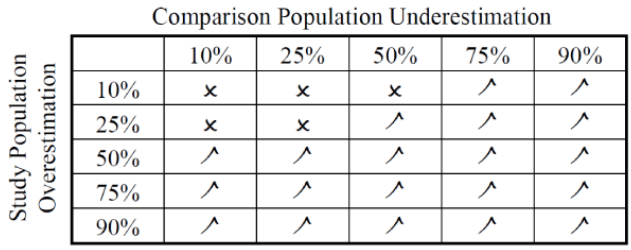

Figure 1b: Example of the second pattern: non-significant to significant in reverse direction.

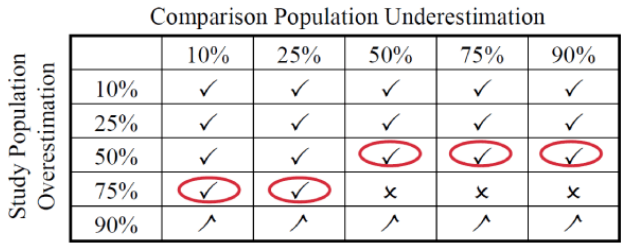

Figure 1c: Example of the third pattern: significant to non-significant to significant in reverse direction

Discussion

This novel sensitivity analysis was a technique to help assess the impact of misclassification in disease prevalence, assuming it existed, and verified the observed significant differences between populations. The resulting 5 by 5 tables pictorially illustrated the analysis results and aided in the determination of the misclassification threshold where statistical difference between the two groups disappeared.

We observed three patterns. In the first pattern, the difference became non-significant with high degrees (>50% in either direction) of misclassification, suggesting the observed difference was truly significant. Thus, higher thresholds provided a greater likelihood that the observed differences were valid and real. In the second pattern, the difference disappeared with the first degree (10% in either direction) of misclassification, resulting in the failure to verify the observed association, and suggesting that there is no association. This occurred for only one disease, rheumatoid arthritis, in which the magnitude of the difference was weak at the observed level. The third pattern, in which differences disappeared and the direction of significance was reversed at higher degrees of under- and overestimation (>50% in either direction), leads to the opposite conclusion and suggests no observed difference.

The interpretation of any of these patterns, the last pattern in particular, depend on the assumed degree of misclassification based on the study design, methodology, source of data, and other differences in the populations that were compared. Furthermore, prevalences from published studies need not always be considered to be underestimated or should population prevalences always be considered to be overestimated. These should be adjusted to what is believed to be true, depending on the diseases in question as well as the sources from which prevalences are being compared.

There is an increasing need for epidemiologic and biostatistical methodology, or "how to" papers, that can be easily applied by public-sector epidemiologists, other public health practitioners, and clinicians [10,14]. Most methods employ complex methodology and require detailed data, making their application by medical researchers impractical. The method presented here is simple, yet powerful in allowing investigators to judge their conclusions of any observed differences against how likely they are to be true. In the absence of the necessary information for conducting a formal analysis, we developed this new sensitivity analysis approach. Its advantage lies in its ease of use by any public health professional, and provides substantial power for validating findings.

In conclusion, we developed a novel, practical sensitivity analysis approach to verify findings by determining the degree of misclassification necessary to negate the difference between population prevalence estimates. The tables for each disease pictorially illustrated three different patterns, which helped to interpret the observed differences and sensitivity analysis results. The sensitivity analysis presented here is a useful alternative to a formal correction method for comparing population prevalence estimates between different populations and could be added to routine study methodology.

Financial Support

The research for this study was supported by the Intramural Program of the Clinical Center and National Institutes of Health and Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, Bethesda.

References

1. Barron BA (1977) The effects of misclassification on the estimation of relative risk. Biometrics 33: 414-418.
2. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH (1977) Bias due to misclassification in the estimation of relative risk. Am J Epidemiol 105: 488-495.
3. Greenland S (1982) The effect of misclassification in matched-pair case-control studies. Am J Epidemiol 116: 402-406.
4. Greenland S, Kleinbaum DG (1983) Correcting for misclassification in two-way tables and matched-pair studies. Int J Epidemiol 12: 93-97.
5. Kleinbaum D, Kupper LL, Morgenstern H (1982) Epidemiologic research: principles and quantitative methods. Van Nostrand Reinhold, New York.
6. Brenner H, Savitz DA, Gefeller O (1993) The effects of joint misclassification of exposure and disease on epidemiologic measures of association. J Clin Epidemiol 46: 1195-1202.
7. Green MS (1983) Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. Am J Epidemiol 117: 98-105.
8. Brenner H, Gefeller O (1993) Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. Am J Epidemiol 138: 1007-1015.
9. Lash TL, Fink AK (2003) Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology 14: 451-458.
10. Rowe AK, Kleinbaum DG, Koplan JP (2004) Practical methods for public health practitioners. Am J Prev Med 26: 252-253.
11. Sinaii N, Cleary SD, Ballweg ML, Nieman LK, Stratton P (2002) High rates of autoimmune and endocrine disorders, fibromyalgia, chronic fatigue syndrome and atopic diseases among women with endometriosis: a survey analysis. Hum Reprod 17: 2715-2724.
12. Gemmill JA, Stratton P, Cleary SD, Ballweg ML, Sinaii N (2010) Cancers, infections, and endocrine diseases in women with endometriosis. Fertil Steril 94: 1627-1631.
13. Greenland S (1998) Basic methods for sensitivity analysis and external adjustment. In: Modern Epidemiology (edn), Lippincott Williams & Wilkins, Philadelphia.
14. Jacobson DL, Gange SJ, Rose NR, Graham NM (1997) Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol 84: 223-243.