Weighted Odds Ratio Estimators

Christiana Drake

Abstract

The odds ratio is a common measure to asses the association between a binary predictor variable and a binary outcome. In epidemiology, the outcome is often disease status and the predictor of interest is a suspected risk factor for the disease. The purpose of a study is an attempt to establish a causal association between exposure and disease. If the object of a study is the estimation of a marginal odds ratio, defined as the ratio of the odds that would be observed in a population if everyone were exposed versus the odds in the same population if no one were exposed, methods such as the Mantel-Haenszel estimator are commonly used. When it is necessary to adjust for many confounders and/or continuous confounders, this approach results in a biased and inconsistent estimator, including matching and stratification by the propensity score. An alternative to matching is inverse probability weighting by the propensity score. The resulting estimator is consistent, provided the propensity score model is correct and adjusts for all confounders.

Introduction

The odds ratio is a common measure to assess the association between a predictor variable and a binary outcome. Often, the predictor is also binary. In epidemiology, the outcome is often disease status and the predictor of interest is a suspected risk factor for the disease. The purpose of a study is to establish an association between exposure and disease, in particular, interest is in establishing that this association is causal. Causality cannot typically be confirmed without a randomized experiment. Nevertheless a suspected causal relationship may be considered to exist in a carefully conducted epidemiology study. A statistical analysis accounting for other factors that may create spurious associations where there are none or hide causal associations that exist is also necessary when trying to establish a causal relationship.

Another issue to consider in establishing causation is the measure to be used. When outcome variables of interest are continuous, responses are often quantified by mean responses and causal effects are measured by differences of means. If the outcome is binary such as presence or absence of disease, disease risk defined as the probability of disease during a period of follow-up may be used as a measure of interest. Risk differences or risk ratios to allow for comparisons between those who are exposed to and those who are free from exposure to a suspected risk factor may be chosen as measures of exposure effects. Generally, risk and risk differences or risk ratios are well understood as quantities of exposure effects. However, one drawback to these measures is dependence on the study design. If the response is continuous it is possible to calculate differences in mean responses between exposed and unexposed based on cross-sectional as well as prospective studies. Similarly, if exposure is measured on a continuous scale it is possible to compare the average exposure among those with and without disease. Risk, however, cannot be established from cross-sectional or retrospective studies. Risk is defined as the probability of developing disease in a population of disease free subjects over a specified time period. It can be calculated as the proportion of subjects who develop disease over a fixed time period.

This requires knowledge of who is at risk for disease at the start of observation and at a minimum, who will have developed disease by the end of observation.

A variety of models and measures are used to assess associations between diseases and exposures as possible causes of disease. Two predominant approaches used are that of regression modeling and using simple measures of association based on summary measures, typically adjusted for what are called confounders or reported separately for what are considered effect modifiers. One approach allows for causal inference only in randomized studies. Neyman[1] discusses agricultural experiments where randomization allows the extrapolation of measurements on randomly selected plots to what would be observed on average per plot if the whole field had been treated in a specific way. By randomly assigning some plots to one treatment and the others to a comparison treatment, it was possible to estimate what would have been observed on average per plot for both treatments. The difference in average yield between these estimates can then be attributed to different treatments. Randomized experiments and their role in establishing causation are extensively discussed in Fisher[2] and Kempthorne[3][4] who discuss the planning of experiments and the mathematical justification for causal inference. Kempthorne[4], in particular discusses the differences in statistical approaches to a planned, randomized experiment vs an observational study. Experiments with random assignment of human subjects are not always feasible. Humans may not consent to randomization. Randomized studies using human subjects are often conducted on a narrowly defined population. It is unethical to study causes that are harmful as is the case in much of epidemiology. Such studies have to rely on observational studies of human populations. Simple effect measures are desirable and require careful planning of observational studies. The design of such studies is discussed extensively in the literature, particularly the need for careful selection of samples and identification of confounders that impact the calculation of effect measures and the drawing of causal conclusions. Cochran[5] presents a detailed discussion of observational studies of human populations. Such

studies typically involve careful selection of samples from populations to be compared that one might argue are reasonably comparable with respect to the outcome of interest in the absence of an exposure effect but differ with regard to exposure.

Much of the statistical literature is focused on comparisons of means in experimental as well as in observational studies. Two-sample comparisons using t-tests and analysis of variance comparing more than two treatments are well known and used extensively in evaluating the effects of treatments/exposures on outcomes of interest. Blocking is used in experiments and observational studies use matching and subclassification on variables known or suspected of affecting the outcome and having different distributions among exposed and non-exposed. These procedures are attempts to approximate the design and analysis of comparative experiments. If a suspected confounder is continuous, matching can be challenging and exact unexposed matches may not exist for many of the exposed observational units. Cochran[6] suggests that most of the bias due to a continuous confounder can be removed by forming about 5 or 6 strata of similar size. It is assumed that the relationship between outcome and confounder is monotone and additive. Comparative observational studies often compare populations where confounders have different distributions within the treatment groups. To obtain estimates of exposure effects a statistical analysis must account for the effects of these confounders.

Epidemiological studies seeking to investigate or identify risk factors for disease, often have outcomes that are not continuous such as whether or not disease occurs, sometimes it is time to an event or number of occurrences of a disease. In this case comparisons of means are not meaningful. When the outcome is occurrence of disease, we might be interested in the risk of disease, comparing those who are exposed to those who are not exposed. The risk of disease is typically defined as the probability of developing disease within the time horizon of the study. In some studies the outcome of interest is a binary variable $Y = 1$ if disease develops and $Y = 0$ if it does not. We define the risk of disease as $P(Y = 1)$. We might then

define an individuals risk of disease as $P(Y_i = 1 \mid Z_i)$ where $Z_i = 1$ if a subject were to be exposed and $Z_i = 0$ otherwise. A measure of exposure effect could then be defined by the difference R_i or the relative risk RR_i

$$R_i = P(Y_i = 1 \mid Z_i = 1) - P(Y_i = 1 \mid Z_i = 0) \tag{1}$$

$$RR_i = \frac{P(Y_i = 1 \mid Z_i = 1)}{P(Y_i = 1 \mid Z_i = 0)} \tag{2}$$

To estimate risk it is necessary to know the number at risk for disease in a defined population. This typically assumes that a population free from disease is identified and then followed over time to observe who develops disease. For comparison a similar population, also free from disease, is required to be followed for an estimate of risk of disease in an unexposed population. This type of study is feasible only if the disease is not rare and if the time from exposure to disease is much shorter than the duration of the study. When either a disease is rare and/or the time from exposure to development of the disease is long a prospective study may not be feasible. A case-control study samples subjects on the basis of disease status and then establishes exposure. While this design is much weaker for establishing causality, it was the basis for much of the research on establishing causation between smoking and lung cancer as well as between smoking and cardiovascular disease (Doll and Hill[7]).

The Odds Ratio

The odds of disease is defined as the ratio of the probability of the disease occurring divided by the probability of disease not occurring. If we let Y be an indicator variable where $Y_i = 1$ if disease occurs and $Y_i = 0$ if disease does not occur, then the odds are defined as

$$odds_i = \frac{P(Y_i = 1)}{P(Y_i = 0)} \tag{3}$$

An equivalent measure to the risk ratio is given by the odds ratio

$$OR_i = \frac{P(Y_i = 1 \mid Z_i = 1)/P(Y_i = 0 \mid Z_i = 1)}{P(Y_i = 1 \mid Z_i = 0)/P(Y_i = 0 \mid Z_i = 0)} \tag{4}$$

The benefit of the odds ratio over relative risk and risk difference lies in the fact that the odds ratio can be calculated from prospective studies where a cohort of subjects, some exposed, some not exposed, is followed over time to see who develops disease as well as from case-control studies where subjects are selected on the basis of disease status and exposure status is ascertained. The assumption that exposure precedes onset of disease has to be made, however. Another argument for that is that the parameters in logistic regression are log-odds ratios. Though the odds ratio is widely used it is still not always well understood as a measure of association. It should be noted that the relative risk RR_i and OR_i as defined here are unobservable and cannot be estimated directly without assumptions. These are relative risk and odds ratio for an individual if that subject were to be exposed ($Z_i = 1$) or not exposed ($Z_i = 0$). The measures are clearly causal. However, these measures are subject-specific and have to be summarized. There is clearly a need for some statistical measure that allows summary statements regarding association and causal effects. One approach is based on regression models. Typically a mean function is modeled as a function of several factors, one of which is the exposure of interest. For a continuous outcome the effects of several factors are modeled to be additive and we obtain a model of the form

$$E[Y_i \mid \mathbf{X}_i, Z_i] = \beta' \mathbf{X}_i + \delta Z_i \tag{5}$$

In this model the parameter δ represents the exposure effect of the i^{th} individual adjusted for other factors that have an effect on outcome. This exposure effect is causal if we believe $\mathbf{X}$ contains all other factors that are causally associated with Y and possibly associated with the exposure Z. It is a conditional, subject specific effect. However, it is, if the model is to be believed, the same for all subjects. This subject-specific effect is conceptually different from a marginal effect which

compares the potential outcomes in a population under two conditions exposure and non-exposure to the factor of interest. In terms of the above linear additive model it is

$$E_{\mathbf{X}}[E[Y_i \mid \mathbf{X}_i, Z_i = 1]] - E_{\mathbf{X}}[E[Y_i \mid \mathbf{X}_i, Z_i = 0]] = E_{\mathbf{X}}[\beta'\mathbf{X}_i] + \delta - E_{\mathbf{X}}[\beta'\mathbf{X}_i] = \delta \quad (6)$$

Here, expectation is with respect to the distribution of $\mathbf{X}$ in the population of interest. In the linear model with additive effect, marginal and subject-specific effect are the same. This property of equal marginal and subject-specific effect is not retained by nonlinear regression models. In a typical study where a sample of exposed and another sample of unexposed subjects is taken, the marginal effect may be biased since

$$\begin{aligned} &E_{\mathbf{X}}[E[Y_i \mid \mathbf{X}, Z = 1]] - E_{\mathbf{X}}[E[Y \mid \mathbf{X}, Z = 0]] \\ &= \beta'\Big(E_{\mathbf{X}}[\mathbf{X}_i \mid Z = 1] + \delta - E_{\mathbf{X}}[\mathbf{X}_i \mid Z = 0]\Big) + \delta \end{aligned} \quad (7)$$

Here, expectation is over the populations of exposed ($Z = 1$) and unexposed ($Z = 0$). The marginal estimate of δ may be biased if no adjustment is made. In estimating the odds ratio, several issues need to be addressed. First, marginal and conditional odds ratio differ. This is commonly referred to as non-collapsibility (Greenland etal 1999). To further discussion of causal inference and subject-specific vs marginal odds ratio we will use the terminology of the Rubin Causal model (Rubin[8], Holland[9]). Rubin's model postulates two potential outcomes for each subject Y_1, if a subject is exposed, and Y_0 if a subject is not exposed. In any study, one of the outcomes is observed and the other is not. An individual causal effect is defined as the comparison between these two outcomes. In recent years there has been explicit discussion of the differences between marginal and subject-specific measures, their interpretation and what is assessed in relation to the objectives of a study. It is generally recognized that regression based measures using linear, generalized linear models, GAMs or quasi-likelihood estimate effects conditional on the variables included in the model.

The parameter related to exposure measures by how much the response will change for a subject with this particular profile. In the linear model the effect will be the same for all subjects. This is not necessarily true for non-linear models.

Conditional measures may not be useful in epidemiological studies. The question may be by how much a disease experience may change in an exposed population if there were no exposure or similarly, by how much the experience may change in an unexposed population if everyone is exposed. More often we consider a population in which some are exposed and some are not. We might want to study the prevalence or rate of disease in the population if everyone is exposed vs if no one is exposed. Measures of disease association suitable to answering such a question are referred to as marginal measures. The marginal odds ratio is defined as the odds that would be observed in a population if everyone is exposed versus the odds that would be observed if no one is exposed. It is given by the following equation

$$OR_{marg} = \left[\frac{P(Y_1 = 1)}{P(Y_1 = 0)}\right] \Big/ \left[\frac{P(Y_0 = 1)}{P(Y_0 = 0)}\right] = \frac{P(Y_1 = 1)P(Y_0 = 0)}{P(Y_1 = 0)P(Y_0 = 1)} \tag{8}$$

This quantity cannot be estimated directly. The variable Y_1 is the response we observe for a randomly selected subject if the subject experienced the factor and we observe the response Y_0 if the subject did not experience the exposure. We only observe Y_{1i} or Y_{0i} for th eith subject but never both. It should be pointed out that that OR_{marg} is different from the crude odds ratio given by

$$\begin{aligned} OR_{crude} &= \left[\frac{P(Y_1 = 1 \mid Z = 1)}{P(Y_1 = 0 \mid Z = 1)}\right] \Big/ \left[\frac{P(Y_0 = 1 \mid Z = 0)}{P(Y_0 = 0 \mid Z = 0)}\right] \\ &= \frac{P(Y_1 = 1 \mid Z = 1)P(Y_0 = 0 \mid Z = 0)}{P(Y_1 = 0 \mid Z = 1)P(Y_0 = 1 \mid Z = 0)} \end{aligned} \tag{9}$$

Note, the crude odds ratio compares the odds in a population that is exposed to the odds in a population that is not exposed. Such a comparison is for two sub-populations that differ in their background profiles and might have different risk for disease in the absence of an exposure effect and is not causal. A comparison

is causal if the argument can be made, had the cause (the exposure) been different, the effect would have also been different. The marginal odds ratio would be appropriate for a population average causal effect. The marginal odds ratio but it cannot be estimated from observed data without untestable assumptions. Furthermore, there is still a considerable amount of confusion about the effect measured by an odds ratio. Estimating an odds ratio from a logistic model with adjustment for confounders results in a conditional odds ratio. The odds ratio from such a model, if the model accounts for all confounders and risk factors would be the odds ratio for a subject in the population if that subject were to exposed to the factor or not. The marginal odds ratio will be smaller than the subject-specific odds ratio. This is not a bias. It is a reflection of the fact that the marginal odds ratio is an average over conditional effects and is not collapsible (Gail etal[10]). Confounding is different from collapsibility and refers to the fact that exposed and non-exposed populations have different distributions of the confounding variable and this variable is causally related to the outcome of interest (Miettinen and Cook[11], Greenland and Robins[12], Wickramarate and Holford[13]) thus some or possibly all of the observed effect is due to the confounder.

Propensity scores to obtain causal effects

Regression models, in particular logistic regression, are often used to estimate an odds ratio adjusted for confounders. The estimate produced, as stated previously, is a conditional or subject-specific estimate, Gail etal[10]. In a regression model, if one believes all variables related to the outcome have been included, the regression parameter related to the exposure of interest is interpreted as the amount by which an outcome will change if the cause is changed by a specified amount. When the target quantity is a marginal causal effect, the question arises how to obtain an unbiased estimate when some of the quantities required cannot be observed. The Rubin Causal Model[8] provides a framework for a solution. The model defines potential outcomes

(Y_0, Y_1) that would be observed if a subject were to not be exposed $(Z = 0)$ or exposed $(Z = 1)$. In this counterfactual framework each subject has two potential responses with a joint distribution $f(y_0, y_1)$. Treatment assignment (used exchangeably with exposure assignment) is said to be strongly ignorable if

$$f(Y_0, Y_1 \mid z = 1) = f(Y_0, Y_1 \mid Z = 1) \quad \text{and if} \quad 0 < P(Z = 1) < 1 \tag{10}$$

When treatment is strongly ignorable a comparison between an exposed unit and a non-exposed unit is causal. It is unlikely that an observational study of exposure effects would result in groups with strongly ignorable treatment assignment. However, if all confounders are known treatment assignment is strongly ignorable given the values of the confounders. If subjects can be categorized according to these values comparisons among exposed and non-exposed are causal. If the number of confounders is small and they are categorical subclassification on confounders is possible. The typical observational study contains many variables and at least some are continuous. Subclassification by all confounders is usually not possible in such a case since each subject will have a unique set of values for the confounders. Rosenbaum and Rubin[14] showed that adjustment based on the propensity score

$$e(x) = P(z = 1 \mid x) \tag{11}$$

is sufficient for strongly ignorable treatment assignment. They suggest subclassification or matching by the propensity score as well as regression adjustment. While the propensity score is continuous, subclassification based on quintiles of $e(x)$ can reduce bias by as much as 90% (Cochran[6], Rosenbaum and Rubin[15]). Therefore, it is common practice to use subclassification or matching (Rosenbaum and Rubin[16], Dehejia and Wahba[17], Foster[18], Rosenbaum[19]) by the propensity score when the

estimate of the causal effect is defined by

$$\Delta = E[Y_1] - E[Y_0] \tag{12}$$

Another approach is introduced by Robins etal[20] who suggest that the unobserved counterfactuals are treated as missing data and observed values should be assigned to represent the missing counterfactuals similar to the way in which surveys treat sampled units. The Horvitz-Thompson estimator uses as weights the inverse of the selection probability to estimate a population total. A unit u_i sampled with probability p_i represents $w_i = p_i^{-1}$ units like it. In a simple random sample where $p_i = n/N$ with n the sample size and N the population size, each sampled unit represents N/n units in the population. Thus the value $w_i \cdot Y_i$ represents the total value of N/n units. Weighting results in unbiased estimates of totals. Rosenbaum[19] suggested propensity weighting of the observed outcomes. Lunceford and Davidian[21] compare propensity weighting to stratification by the propensity score. They also provide a theoretical framework for calculating standard errors of weighted estimates and show their consistency as long as the propensity score is correctly specified.

Weighted odds ratios

As shown before, the marginal odds ratio is a different measure than the conditional, subject-specific odds ratio. For this reason adjustment for confounders by stratification on confounders may not result in consistent estimates. The Mantel-Haenszel estimator[22] of the odds ratio is commonly used. It can be shown to be consistent under asymptotics where the number of strata remains fixed while the number of observations goes to $n \to \infty$ as well as under sparse asymptotics where the number of strata $k \to \infty$ as n increases as is the case when observations are matched. This property has been applied to propensity matching as well. However, it can be shown (Loux etal[23]) that stratification and matching by the propensity score typically re-

sult in biased estimates of the marginal odds ratio. Austin[24] first investigated several methods of estimating a marginal odds ratio through a series of Monte Carlo studies and concluded that propensity score matching did not result in the same reduction in bias as the estimation of differences of means or risk differences. In a letter to the editor in response to the study by Austin[24] Forbes and Shortreed[25] demonstrate that an IPW (inverse probability weighted) estimator they constructed was nearly unbiased.

Stampf etal[26] proposed another approach to estimating a marginal odds ratio. The authors use a logistic regression model to estimate a conditional probability of disease given exposure $P(Y = 1 \mid X, Z)$ where $Z = 1$ if exposed and $Z = 0$ if not exposed. The covariates X are considered random variables and expectation is taken with respect to X to obtain the marginal probabilities $P(Y_1 = 1)$ and $P(Y_0 = 1)$ where $Z = 1$ is substituted into the logistic model to obtain $P(Y_1 = 1)$ and similarly for $P(Y_0 = 1)$. They also propose a stratified estimator of the probability of disease, with strata defined by the propensity score and then calculate a weighted average of the stratum specific probabilities of disease when $Z = 1$ and $Z = 0$. In both cases the marginal probabilities are used to calculate a marginal odds ratio.

Loux etal[23] calculate marginal probabilities $P(Y_1 = 1)$ and $P(Y_0 = 1)$ by weighting by the estimated propensity score. The propensity score is usually not known and must be estimated. Often, a logistic regression is used Rosenbaum and Rubin[14]. The primary purpose in many of the models using the propensity score has been on stratification and the balance achieved with respect to confounders. Drake[27] has shown that the estimated propensity score is less sensitive to misspecification of the functional form of the propensity score than the maximum likelihood estimates of regression parameters such as the conditional odds ratio. However, balance can only be achieved for known confounders. Loux etal[23] proposed two versions of the weighted estimator

$$\hat{p}_{1,IPW} = \left(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}_i} \tag{13}$$

and similarly

$$\hat{p}_{0,IPW} = \left(\sum_{i=1}^{n} \frac{1 - Z_i}{1 - \hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{(1-)Z_i)Y_i}{\hat{1} - e_i} \tag{14}$$

These semi-parametric estimators are consistent provided the propensity score is correctly specified. Based on the work by Lunceford and Davidian[21], an expression for the variance of the estimator can be derived. It is based on M-estimation in a paper by Stefanski and Boos[28]. The calculations are complicated and it is much easier to use a bootstrap approach to estimating the variance. This was the method used in the example by Loux etal[23]. The propensity score is the probability of being exposed (treated) and some subjects may have very high propensity for exposure or very low propensity. Such extreme probabilities result in potentially very large weights and can lead to a poorly performing estimator. In this case one possibility is to truncate the probability at some pre-specified value. This introduces a small bias as long as the number of truncated weights is small. In the case of a large number of very extreme probabilities and therefore, weights, the data should be carefully inspected and the source of the data should be examined. The two treatment/exposure groups have limited overlap and causal inference is problematic. In a final note, the validity of drawing causal conclusions from the calculated weighted odds ratio as with all other approaches rests on the assumption, not testable, that treatment assignment is strongly ignorable.

References

[1] Neyman, J (1923). Sur les applications de la théoriedes probabilities aux experinecs agricoles: essai des principes. Roczniki Nauk Rolniczki 10 1-51. (In Polish. English translation by D Dabrowska and T Speed, *Statistical Science* **5**, 463-472, (1990).

[2] Fisher RA (1935). *Design of Experiments*. Oliver& Boyd, Edingburgh.

[3] Kempthorne, O (1977). Why randomize. *Journal of Statistical Planning and Inference* **1**, 1-25.

[4] Kempthorne, O (1978). Sampling inference, experimental inference and observation inference. *Sankhya B, 40*, 115-145.

[5] Cochran, WG (1965). The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society Series A* **128**, 134-155.

[6] Cochran WG, (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295-314.

[7] Doll R and Hill AB, (1950). Smoking and carcinoma of the lung. *British Medical Journal* **2(4682)**, 739-748.

[8] Rubin DB, (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* **66(5)**, 688-701.

[9] Holland, PW (1986). Statistics and causal inference. *Journal of the American Statistical Association* (with Discussion and Reply) **81**, 945-970.

[10] Gail, MH, Wieand, S, and Piantadosi, S (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431-444.

[11] Miettinen, O and Cook, EF. Confounding: Essence and detection. *American Journal of Epidemiology*, **114(4)**:593-603.

[12] Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Science*, **14(1)**:29-46 .

[13] Wickramaratne, PJ and Holford, TR (1987). Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. *Biometrics* **43**, 751?765.

[14] Rosenbaum, PR and Rubin, DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

[15] Rosenbaum, PR and Rubin, DB (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516-524.

[16] Rosenbaum, PR and Rubin, DB (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39**, 33-38.

[17] Dehejia, RH and Wahba S (2002). Propensity score-matching methods for non-experimental causal studies. *The Review of Economics and Statistics* **84**, 151-161.

[18] Foster EM (2003). Propensity score matching: an illustrative analysis of dose response. *Medical Care* **41**, 1183-1192.

[19] Rosenbaum PR. Propensity Score. In *Encyclopedia of Biostatistics* Armitahe P, Colton T (eds), vol 5. Wiley: New York, 1998:3551-3555.

[20] Robins JM, Rotnitzky A and Zhao Lue Ping (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-865.

[21] Lunceford JK and Davidian M (2004): Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study *Statistics in Medicine* **23**, 2937-2960.

[22] Mantel N and Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*. **22**, 719-748.

[23] Loux TM, Drake C and Smith-Gagen J (2017). *Statistical Methods in Medical Research*. **26**, 155-175.

[24] Austin P (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*. bf 26, 3078-3094.

[25] Forbes A and Shortreed S (2008). Letters to the Editor: Inverse probability weighted estimation of the marginal odds ratio: Correspondence regarding "The performance of different propensity score methods for estimating marginal odds ratios". it Statistics in Medicine. **27**, 5556-5559.

[26] Stampf S, Graf E, Schmoor and Schumacher M (2010). Estimators and confidence intervals for the marginal odds ratio using logistic regression and propensity score stratification. *Statistics in Medicine*. **29**, 760-769.

[27] Drake CM (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* **49**, 1231-1236.

[28] Stefanski LA and Boos DD (2002). The calculus of M-estimation. *The American Statistician*. **56**, 29-38.