Postitive-Unlabeled Data Considerations in the Design of Orthopaedic Risk-factor Studies

Abstract

Objective

In orthopaedic risk-factor studies, the disease status of control subjects is not always perfectly ascertained. That means control groups may be mixtures of both unaffected cases and some unidentified affected cases. Control groups with misclassified data are called *unlabeled*. Treating unlabeled groups as disease-negative control groups is known to cause misclassification bias, but there has been little research on how the misclassification affects the statistical power of the risk association tests. In this study, we investigated the effect using unlabeled groups as control groups on the power of association tests, with the intention of demonstrating that control groups with even small misclassification rates can reduce the power of association tests.

Materials and methods

This was a simulation study using study designs from published orthopaedic risk-factor studies. The approach was to use their designs, but simulate the data to include misclassified affected subjects in the control group. The simulated data was used like pilot data to calculate the power of the risk-association test. We calculated powers for several study designs and misclassification rates, and compared them to a reference model, which is the case then there are no misclassifications and the model is correct.

Results

Treating unlabeled data as disease-negative only always reduced statistical power compared to the reference power. Power loss was greater with increasing misclassification rate.

Conclusion

Researchers calculating sample sizes for risk-factor studies should include adjustments for misclassification rates.

Keywords: risk-factor, case-control, power, sample size

Introduction

In risk factor studies, the positive disease status of the affected subjects is ascertained with perfect sensitivity and specificity, but sometimes the disease status of control subjects is not perfectly ascertained. That means control groups may be mixtures of both unaffected cases and some unidentified affected cases. Control groups with misclassified data are called *unlabeled*. Data with truly affected cases in the positive group and an unlabeled control group is called *positive-unlabeled* data. Examples of positive-unlabeled data are well documented in human medicine, but less so in veterinary medicine. Nevertheless, many veterinary studies the fall into the PU framework. For example, genome-wide association studies of cranial cruciate ligament disease (CCLD) in dogs use case-control designs. The affected cases are truly positive CCLD cases because they are enrolled from the set of dogs who have undergone knee stabilization surgery. The controls are typically five-years-old or older with no history of CCLD and pass an orthopedic veterinary exam by board-certified surgeons. However some control dogs will have spontaneous rupture in the future, and so genetically belong in the CCLD affected group. Other control dogs may have sub-diagnostic disease. For example, a dog might appear sound on physical exam and be enrolled in the control group, but may actually have force-platform-detectable hindlimb lameness. Such a dog should not be in the control group because the lameness might be subclinical CCLD. (Waxman et al. 2008) There are other examples of PU data in the veterinary literature, typically in risk-factor studies using case-control designs. For example, Arthur et al. (2016) used a case-control design to assess the risk of osteosarcoma following fracture repair. They noted, "There may be additional cases [in the control group] in which implant-related osteosarcoma was diagnosed in the private practice setting without referral. . . ," suggesting that the control group may be unlabeled because some control group cases were actually disease positive, but diagnosed outside the study. In another example, Wylie et al. 2013 studied risk factors for equine laminitis using controls obtained from an owner survey. The authors noted the PU aspect of their data, "Our study relied on owner-reported diagnoses of endocrinopathic conditions, and this may have introduced misclassification bias."

As mentioned above, the affected cases are "labeled" positive, but the control data is "unlabeled," because dogs may be affected or unaffected. Treating the unlabeled control group as entirely unaffected is called the *naive model*. The proportion of affected dogs in the control group is called the *nondetection rate* or *undetected rate*.

Using the naive model when the nondetection rate is positive causes misclassification bias (because there affected cases in the control group), and that bias is well documented in the human medical literature. The biases due to misclassification can be mitigated using models other than the naive model and with the appropriate data analysis, and there are many articles describing methods for analyzing positive-unlabeled

data. Bekker provides and excellent summary of methods. Sometimes, however researchers prefer the naive model because they believe that their small nondetection rates induce misclassification biases that are too small for practical consideration. There is some suggestion that nondetection rates under 10% do have little impact on bias. (Bekker et al., 2020)

But bias in estimates (e.g., bias in regression coefficients) is just one part of the results; the other part is inference (e.g., p-values). Central to inference is the power of statistical tests. Power is used in planning a study as a measure of the ability of the study to make the correct decision. That is, finding $P<0.05$ when it should. Typically, 80 percent power means that if the groups are truly different, then the statistical test has an 80 percent chance of obtain $p<0.05$

During the design of a risk association study, a researcher might calculate the sample size they need assuming the nondetection rate is zero. That is, the there are no misclassified affected subjects in the control group. However, if after collection the data are PU, then the naive model is incorrect and the estimated power may not correct.

We investigated the effect of PU data on statistical power under the naive model. The reference power is defined as the power when the naive model is correct and the group sizes are balanced. The results are described in terms power loss relative to the reference power, both percent power loss and absolute power loss. For context these two quantities are analogous to relative risk and absolute risk from epidemiology.

Using a simulation, we described how statistical power changes with varying proportions of undetected positives in the naive controls, and varying the imbalance between the numbers of cases and naive controls. Our aim is to demonstrate that the naive analysis of PU data affects statistical power in risk-factor studies, even for small nondetection rates.

Methods and materials

This was a simulation study assessing the changes in power of a univariate association test under different PU conditions. There are many statistical tests for association, but we calculated the power for one of the most common tests, the χ^2 test with one degree of freedom, which is used to test statistical significance of a binary risk factor. For example, if the disease is CCLD, and the risk factor is sex, then the logistic regression coefficient for body condition score would be tested with a χ_2 test. More generally, this test can be used to assess the significance of any risk factor using the predicted values from a univariate logistic regression.

In the context of risk association studies, and all else being equal, the χ^2 test would achieve its maximum power for a balanced study design when the naive model is correct (i.e., the undetected positives rate is zero). We call that maximum power the *reference power* and reported our results as both percent power loss relative to the reference power and as absolute power loss from reference power. In other words, we are using the χ^2 test to show how much statistical power is lost by ignoring the nondetection rate.

The total sample size for the simulation was fixed N=200, which is consistent with Healey et al. 2019 (N=216), and Baird et al. 2014 (N=217). The effect size, 0.21, was chosen because with N=200, the reference power was close to 80 percent, which is value that is commonly used in study design. That way, the reference model is the one with standard power of 80 percent. Note that the sample size and effect size are not a key parameters for the simulation because for any sample size an effect size can be chosen so that power is 80 percent. Also, effect size and sample size are not features of PU data, per se.

The simulation study varied two study design parameters: the nondetection rate and group-size imbalance. The proportion of undetected positives in the control group ranged from 0 (the value for reference power) to 10 percent. We used 10 percent as the upper limit because researchers are generally willing to accept nondetection rates below 10 percent and use the naive model, but change to a PU analysis for rates greater than 10 percent. (Bekker et al. 2020)

We modeled group imbalance using Healey et al. (2019), which used 161 dogs affected with CCLD and 55 unlabeled dogs as controls, and Baird et al. (2014) which used 91 dogs affected with CCLD, and 126 unlabeled dogs as controls, so that imbalance ratios were about 3:1 and 1:3. We only used two imbalance proportions (1:3 and 3:1) and no imbalance (1:1) because the key parameter for this study was the nondetection proportion. That gave simulation sample sizes of (50, 150), (150, 50), and (100, 100)

The simulation algorithm

The overall method is to simulate data, and then use that data like pilot data to estimate power of the χ^2 test. The process is repeated 5000 times for each combination of sample size and nondetection rate, and then the 5000 simulated powers are averaged to give the estimated power for each sample size - nondetection rate combination. Simulating the data works backward from what might be expected. Instead starting with values for a risk factor (e.g., 200 0's and 1's representing sex) and then simulating their disease status, we start with the disease status (e.g., 50 affected cases and 150 controls with 140 unaffected and 15 affected) and then assign binary values for the risk factor. It was done that way to control the nondetection rate and group sizes which needed to be fixed for the power calculation.

The simulation algorithm is most easily described using examples, and we begin with calculating power for the χ^2 test under the reference model, which is 100 cases and 100 correctly labeled (i.e., truly unaffected) controls. That is, there are no affected cases in the control group, so this is not PU data, and the naive model is the correct model. Next we have to associate a binary risk factor variable, X (e.g., sex), with the cases and controls. It is more intuitive to start with the risk factor and then is simulated by sampling 100 negative cases from a binomial distribution with $Pr(X = 1) = 0.2$. That probability means the the baseline risk for the disease in the population is 0.2. It it was chosen arbitrarily, because the baseline risk isn't central to power, the effect size is. As mentioned above, the effect size was 0.21, so the 100 positive cases were sampled from a binomial distribution with $Pr(X = 1) = 0.2 + 0.21$.

Now, the 200 cases are binary pairs of data, one representing the group and the other representing the risk factor. These simulated data are then treated like pilot data and used for a χ^2 test power calculation with one degree of freedom and $\alpha = 0.05$. As mentioned above, this process was repeated 5000 times and the resulting powers average to estimate the reference power.

For the second example, we calculate the power for a PU example. Suppose that in a 100-patient control group, 10 percent are in fact undetected positives. So the dataset is 100 affected cases in the positive group, 10 affected cases in the control group, and 90 unaffected cases in the control group. As in the previous example, the risk factor is simulated by sampling from binomial distributions. Now, 90 controls are sampled from the binomial distribution with $Pr(X = 1) = 0.2$, the 10 affected controls are sampled from the binomial distribution with $Pr(X = 1) = 0.2 + 0.21$ and 100 cases are sampled from the binomial distribution with $Pr(X = 1) = 0.2 + 0.21$. The 10 mislabeled affected cases remain in the control group, so as to measure the effect of treating PU data naively. As before, the simulated data are treated like pilot data and power was calculated. This process is repeated 5000 times and the result averaged.

Results

Table 1 describes power loss for the three sample sizes, (50, 150), (150, 50), and (100, 100), and for three nondetection rates, 0, 0.05 (5%), and 0.1 (10%). To give these parameters context, if the group sizes are (50, 150) and the nondectection rate is 0.1, then the positive (affected) group has 50 cases, and the unlabeled control group (which we are treating naively) has 150 cases, 15 of which are actually affected cases. When the nondetection rate is zero, then the naive model is correct because there no affected cases in the control group. The first row is the reference power, so its loss of power compared to itself is zero. The reference power was calculated in the simulation just like all the other powers, and was estimated to be 0.82.

The first row is the reference power, so there is no power loss for itself. Columns four and five are the power loss columns, and have negative entries because for this simulation because PU data analyzed under the naive models always had lower power, as did unbalanced data. For example, the second row shows a -4.3% relative power reduction when the group sizes are balanced but five percent of the control group is actually positive cases.

There were increasing nondetection rate increases reduces power. However, group size imbalance can affect power loss more than relatively small nondetection rates. For example, consider rows 4, 5, and 6. For those rows, nondetection rate is decreasing, but power loss is increasing due to group imbalance. However, within fixed group sizes (e.g., 100,100, 150,50 or 50,150) increasing nondetection rate alway means increasing power loss.

Table 1. Power loss.

Row	N positive cases	N naive controls	nondetection proportion	Relative power loss (%)	Absolute power loss (from 0.82)
1	100	100	0.00	0.00	0.00
2	100	100	0.05	-4.30	-0.04
3	50	150	0.00	-8.24	-0.07
4	100	100	0.10	-8.80	-0.07
5	50	150	0.05	-12.85	-0.11
6	150	50	0.00	-13.45	-0.11
7	150	50	0.05	-17.22	-0.14
8	50	150	0.10	-17.41	-0.14
9	150	50	0.10	-22.26	-0.18

Discussion

This study showed that under specific conditions there were modest power reductions even for small proportions of undetected positives in the control group. For example, with a 5 percent nondection rate and a balanced study design, the sample size in the context of this simulation would have to be XXXX, which is a zzz percent increase in sample size.

The working example was the association test for a single SNP in a GWAS study, but the simulation results apply to any kind of study with univariate association tests, such as any risk factor study. That is a broad class of studies. Examples include the univariate associations between post-op surgical infections and various surgical conditions (e.g., boarded surgeon vs resident, manufacturer of bone plate). In that case, there may be subdiagnostic infections in the control group. Another example is univariate association tests in veterinary surveys, such as XXX. In that case, there may be subclinical . . .

In this simulaion, the undetected positives in the negative group were randomly sampled from the same population as the detected positives in the affected group. That's a common assumption (ref GU and others) but there are other models. For example, the undetected positives in the control group might be a subpopulation of positives defined by another variable. For example, in a CCLD GWAS study, the undetected positives in the control group might be positive dogs with low BCS only.

it does not apply to the situation where binary "affectedness" changes as the function of a scaled risk factor variable. Using risk factors for CCLD example, BCS may affect spontaneous rupture.

This was a simulation study that considered the effect on statistical power of 10 percent or fewer undetected positives in the control group. The simulation used

Other papers have discussed improved tests and power. (wang, and the pan paper)

Practical implications

However, the undetected-positive rate for CCLD GWAS studies is small, certainly less that 10 percent, and the low rate causes only small biases in odds ratio estimates. However, this study shows that unlabeled data also affect the inferences, and we show that even a few positive cases in the negative controls can affect the power of the study. Fortunately, positive-unlabeled data appears to reduce power, making the results reported in CCLD GWAS studies conservative.

References

[1] T. Wang and Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. *The american journal of human geneticsFeb*, 1(80):2, 2007.

Appendix 1.

Table 1. Sample sizes for four GWAS studies of CCLD. The last columns shows the number of affected cases in the control group assuming 10 percent non-detection rate.

Study	N	Cases	Naive controls	Max No. undetected positives (10%)
Baird et al. (2014)	217	91	126	13
Baker et al. (2021)	397	156	241	24
Cook et al. (2020)	333	190	143	14
Healy et al. (2019)	216	161	55	6