Long abstract: Design Considerations for studies with positive-unlabeled data

2023-06-12
Richard Evans
Clinical and Translational Science Institute
University of Minnesota
Richard Evans' email

Purpose

In risk association studies, the affected subjects in the positive group are often identified 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. Those kinds of control groups are called *unlabeled*, because we are not completely sure about the disease labels (affected or unaffected). The entire dataset is referred to as *postitive-unlabeled* (PU).

Accounting for the unlabeled aspect of the control groups is usually handled during data analysis, and there has been little investigation into PU sample size considerations. (Bekker, Hastie, Gu) In this study, we investigate the effect of PU data on the power of association tests with the intention of improving sample size calculations.

Background

Examples of positive-unlabeled data are well documented in human medicine, but less so in veterinary medicine. Nevertheless, there are some 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 enrolled in the control group, but actually have force-platform-detectable hindlimb lameness. (Waxman) Such a dog should not be in the control group because the lameness might be subclinical CCLD.

There are other plausible 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 control group cases were actually disease positive, but diagnosed outside the study. For 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. These kind of data are called positive-unlabeled data. Treating the unlabeled control group as entirely unaffected is called the *naive model*. The proportion of affected dogs in the unaffected control group is called the *nondetection* rate or *undetected rate*.

The misclassification biases are well documented in the medical literature. The biases due to misclassification can be sometimes 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)

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 in studies. Power is used in planning a study as a measure of the ability of the study to make the correct decision. That is, find $P<0.05$ 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 using the naive model. However, if 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.

Aim

Our aim is to demonstrate that PU data affects statistical power in risk-factor studies, even for small nondetection rates.

Objective

The objective is to report the loss in statistical power for specific, reasonable examples based on the published literature. Loss is reported two ways: as a percentage of the reference power, and as absolute power loss.

Methods

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 (for GWAS, see Pan), 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 risk-factor binary. 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 Wald 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 (N=216) and Baird (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 non-detection proportion 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 detection rates below 10 percent and use the naive model, but change to a PU analysis for rates greater than 10 percent. (Bekker)

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. That study was chosen because it was generally similar to other GWAS studies, and because it has the most extreme imbalance. 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 sample sizes of (50, 150), (150, 50), and (100, 100)

The simulation algorithm is most easily described using two examples, and we begin with calculating power for the χ^2 test using simulated reference data, which has no positives in the negative group and balanced group sizes (100 positive and 100 negative cases). That is, there in no PU data. The binary risk factor variable, X (e.g., sex), 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 $Pr(X=1) = 0.2 + 0.21$. These data and the disease status labels (0=negative, 1=positive) allow for a 2x2 table,

		Risk factor		**Total**
		0	1	
status	0	a	c	100
	1	b	d	100

These simulated data are then treated like pilot data and used for a χ^2 test power calculation with one degree of freedom and $\alpha = 0.05$. That process was repeated 5000 times, each time generating new risk factor data from the two binomial distributions. The average of the 5000 powers give the estimated power for the reference model, which, as mentioned above, is 80 percent by design.

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 positives in the positive group, 10 positives in the control group, and 90 negatives in the control group. As in the previous example, the risk factor is simulated by sampling from binomial distributions. Now, 90 negatives are sampled from the binomial distribution with $Pr(X=1) = 0.2$ and 155 positives from the binomial distribution with $Pr(X=1) = 0.2 + 0.21$. But this time, the 10 mislabeled positives are analyzed as negatives, so as to measure the effect of treating PU data naively.

Findings

Table 1 shows the loss of power due to group size imbalance only. For this table, the naive model is the correct model because the nondetection rate is zero. 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 0.82. As expected, the power drops with imbalance, but it is important to note that the power drop is asymmetric.

Table 1 Power loss due to group size imbalance

Row	N positive cases	N naive controls	nondetection proportion	Relative power loss (%)	Absolute power loss (from 0.82)
1	100	100	0	0	0
2	50	150	0	-8.24	-0.0676
3	150	50	0	-13.4	-0.11

Table 2 shows power loss when PU is modeled naively, for selected nondetection proportions and groups sizes, sorted by descending relative power loss. 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, PU data models treated as 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. Rows three has a higher nondetection rate than rows five and size, but a smaller power loss.

Table 2. Power loss due to nondetection rate and group size imbalance.

Row	N positive cases	N naive controls	nondetection proportion	Relative power loss (%)	Absolute power loss (from 0.82)
1	100	100	0	0	0
2	100	100	0.05	-4.3	-0.0353
3	100	100	0.1	-8.8	-0.0722
4	50	150	0.05	-12.8	-0.105
5	150	50	0.05	-17.2	-0.141
6	50	150	0.1	-17.4	-0.143
7	150	50	0.1	-22.3	-0.183

Table 3 combines Tables 1 and 2 to show the combined effects of nondetection rate and group size imbalance. On the whole, increasing nondetection rate increases. 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 2. Power loss using Tables 1 and 2 combined.

Row	N positive cases	N naive controls	nondetection proportion	Relative power loss (%)	Absolute power loss (from 0.82)
1	100	100	0	0	0
2	100	100	0.05	-4.3	-0.0353
3	50	150	0	-8.24	-0.0676
4	100	100	0.1	-8.8	-0.0722
5	50	150	0.05	-12.8	-0.105
6	150	50	0	-13.4	-0.11
7	150	50	0.05	-17.2	-0.141
8	50	150	0.1	-17.4	-0.143
9	150	50	0.1	-22.3	-0.183

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

Arthur EG, Arthur GL, Keeler MR, Bryan JN. Risk of osteosarcoma in dogs after open fracture fixation. Veterinary Surgery. 2016 Jan;45(1):30-5.

Baird AE, Carter SD, Innes JF, Ollier WE, Short AD. Genetic basis of cranial cruciate ligament rupture (CCLR) in dogs. Connective tissue research. 2014 Aug 1;55(4):275-81.

Baker LA, Momen M, McNally R, Berres ME, Binversie EE, Sample SJ, Muir P. Biologically enhanced genome-wide association study provides further evidence for candidate loci and discovers novel loci that influence risk of anterior cruciate ligament rupture in a dog model. Frontiers in Genetics. 2021 Mar 5;12:593515.

Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genetic epidemiology. 2011 Nov;35(7):606-19.

Bekker J, Davis J. Learning from positive and unlabeled data: A survey. Machine Learning. 2020 Apr;109:719-60.

Cook SR, Conzemius MG, McCue ME, Ekenstedt KJ. SNP-based heritability and genetic architecture of cranial cruciate ligament rupture in Labrador Retrievers. Animal genetics. 2020 Oct;51(5):824-8.

Engdahl K, Emanuelson U, Höglund O, Bergström A, Hanson J. The epidemiology of cruciate ligament rupture in an insured Swedish dog population. Scientific Reports. 2021 May 5;11(1):1-1.

Gu W, Swihart RK. Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. Biological conservation. 2004 Apr 1;116(2):195-203.

Healey E, Murphy RJ, Hayward JJ, Castelhano M, Boyko AR, Hayashi K, Krotscheck U, Todhunter RJ. Genetic mapping of distal femoral, stifle, and tibial radiographic morphology in dogs with cranial cruciate ligament disease. PloS one. 2019 Oct 17;14(10):e0223094.

Lydersen S. Balanced or imbalanced samples?. Tidsskrift for Den norske legeforening. 2018 Sep 17.

McManus IC. The power of a procedure for detecting mixture distributions in laterality data. Cortex. 1984 Sep 1;20(3):421-6.

Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. The american journal of human genetics. 2007 Feb 1;80(2):353-60.

Waxman AS, Robinson DA, Evans RB, Hulse DA, Innes JF, Conzemius MG. Relationship between objective and subjective assessment of limb function in normal dogs with an experimentally induced lameness. Veterinary Surgery. 2008 Apr;37(3):241-6.

Wylie CE, Collins SN, Verheyen KL, Newton JR. Risk factors for equine laminitis: A case-control study conducted in veterinary-registered horses and ponies in Great Britain between 2009 and 2011. The Veterinary Journal. 2013 Oct 1;198(1):57-69.

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