Sample Size Considerations in the Design of Orthopaedic Risk-factor Studies

Running title: Sample Size Considerations

Richard Evans †

† Corresponding Author

Clinical and Translational Science Institute
University of Minnesota

evan0770@umn.edu

##Abstract

Objective

Sample size calculations play a pivotal role in study design, as they influence study interpretability, costs, and the allocation of hospital resources and staff time. In the context of most orthopedic risk-factor studies, either the sample size calculation or the post-hoc power calculation assumes the precise ascertainment of disease status among control subjects, which might not always hold true. Consequently, control groups might consist of a mixture of both unaffected cases and some unidentified affected cases. Negative control groups containing misclassified positive data are denoted as “unlabeled.” Treating unlabeled groups as disease-negative control groups is recognized to introduce misclassification bias. However, scant research has been conducted on the impact of such misclassification on the statistical power of risk association tests. In this study, we elucidate the repercussions of employing unlabeled groups as control groups on the power of risk association tests. Our aim is to demonstrate that even minor misclassification rates within control groups can substantially diminish the power of association tests. Consequently, disregarding the unlabeled aspect of control groups in sample size calculations may lead to underpowered studies. Additionally, we present a range of correction factors to recalibrate sample size calculations to achieve 80% power.

Materials and Methods

This study employed a simulation approach, utilizing study designs from published orthopedic risk-factor studies. The methodology involved adopting these designs and subsequently simulating the data to incorporate predetermined proportions of misclassified affected subjects within the control group. The simulated dataset was then employed to compute the power of a risk-association test. We calculated the statistical power for various study designs and misclassification rates, subsequently comparing these results against a reference model.

Results

Treating unlabeled data as disease-negative consistently resulted in a reduction of statistical power in comparison to the reference power. Moreover, the extent of power loss increased as the misclassification rate escalated. In the context of this study, restoring the statistical power to 80% was attainable by increasing the sample size by a factor ranging from 1.1 to 1.4.

Conclusion

Researchers should exercise caution when calculating sample sizes for risk-factor studies and should incorporate adjustments for estimated misclassification rates.

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

Introduction

Sample size calculations are pivotal in study design, influencing study costs, allocation of hospital resources, and staff time. Underpowered studies, fraught with Type II errors, can be challenging to interpret and pose ethical concerns with minimal prospects of success. [1, 2].

In orthopedic risk factor studies, while the positive disease status of affected subjects is ascertained with perfect sensitivity and specificity, control subjects' disease status is sometimes uncertain. This results in control groups comprising a blend of unaffected and unidentified affected cases. Control groups with misclassified data are termed "unlabeled." Such data scenarios, featuring truly affected cases in the positive group and an unlabeled control group, are referred to as "positive-unlabeled" (PU) data within the data science community.

Although examples of positive-unlabeled data are well-documented in human medicine, their presence in veterinary medicine is less explored. Nevertheless, many veterinary studies fall within the positive-unlabeled framework. For instance, genome-wide association studies on cranial cruciate ligament disease (CCLD) in dogs adopt case-control designs. The affected cases, derived from dogs undergoing knee stabilization surgery, are unequivocally positive for CCLD. Control cases are typically five years old or older, displaying no CCLD history, and clearing an orthopedic veterinary examination by a board-certified surgeon. However, some control dogs might experience future spontaneous ruptures, thus genetically aligning with the CCLD-affected group. Other controls might exhibit sub-diagnostic disease, masking latent CCLD. [3]

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. [4] They said, "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 osteosarcoma 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. [5] The authors noted the positive-unlabeled aspect of their data, "Our study relied on owner-reported diagnoses of endocrinopathic conditions, and this may have introduced misclassification bias."

Abstract Objective Sample size calculations are pivotal in study design, influencing study costs, allocation of hospital resources, and staff time. Underpowered studies, fraught with Type II errors, can be challenging to interpret and pose ethical concerns with minimal prospects of success (Halpern, 2002; Hofmeister, 2007).

In orthopedic risk factor studies, while the positive disease status of affected subjects is ascertained with perfect sensitivity and specificity, control subjects' disease status is sometimes uncertain. This results in control groups comprising a blend of unaffected and unidentified affected cases. Control groups with misclassified data are termed "unlabeled." Such data scenarios, featuring truly affected cases in the positive group and an unlabeled control group, are referred to as "positive-unlabeled" (PU) data within the data science community.

Although examples of positive-unlabeled data are well-documented in human medicine, their presence in veterinary medicine is less explored. Nevertheless, many veterinary studies fall within the positive-unlabeled framework. For instance, genome-wide association studies on cranial cruciate ligament disease (CCLD) in dogs adopt case-control designs. The affected cases, derived from dogs undergoing knee stabilization surgery, are unequivocally positive for CCLD. Control cases are typically five years old or older, displaying no CCLD history, and clearing an orthopedic veterinary examination by a board-certified surgeon. However, some control dogs might experience future spontaneous ruptures, thus genetically aligning with the CCLD-affected group. Other controls might exhibit sub-diagnostic disease, masking latent CCLD.

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 data science literature. [6] 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 and Davis (2020) provides and excellent summary of methods. [6] Sometimes, however, researchers prefer the naive model because the analysis is simpler and they believe 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. [6]

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 decisions. That is, finding P<0.05 when it should. Typically, 80 percent power means that if the group parameters are truly different, then the statistical test has an 80 percent chance of obtaining p<0.05.

During the design phase of risk association studies, researchers often calculate the sample size required for 80% power assuming a zero nondetection rate. In other words, they presume no misclassified affected subjects in the control group. However, if collected data conform to the positive-unlabeled scenario, the naive model becomes inappropriate, and estimated power might be lower than anticipated.

This study examines the impact of positive-unlabeled data on the loss of statistical power under the naive model. For comparison, the reference power is defined when the naive model is accurate, and group sizes are balanced. Results quantify power loss relative to reference power in terms of both percentage and absolute power loss—parallel to relative and absolute risk in epidemiology.

Through simulations, we elucidate how statistical power varies based on different proportions of undetected positives in naive controls and varying imbalances between case and control numbers. Our first objective is to demonstrate that naive analysis of positive-unlabeled data diminishes statistical power in risk-factor studies, even with minor nondetection rates. Our second objective is to provide correction factors to adjust sample sizes upwards, rectifying the power loss outlined in the first objective.

Methods and materials

The Test of Association

This study encompassed a simulation-based approach aimed at evaluating the alterations in the power of a univariate association test across various positive-unlabeled (PU) conditions. While numerous statistical tests gauge association, our analysis specifically focused on one of the most prevalent tests, namely the Fisher exact test. This test is commonly employed to assess the statistical significance of a binary risk factor. Broadly, the test can be extended to evaluate the significance of any risk factor by utilizing predicted values derived from a univariate logistic regression.

Within the realm of risk association studies, assuming all other factors remain constant, the Fisher exact test achieves its peak power in a balanced study design when the naive model holds true (i.e., no undetected positives in the control group). This maximum power is termed the “reference power,” and we present our findings in terms of both the percentage of power loss relative to the reference power and the absolute power loss from the reference power. In essence, we employed the Fisher exact test to illustrate the extent of statistical power reduction resulting from the oversight of the nondetection rate.

The Sample Size and Group Imbalance

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). [7] [8] 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. [6]

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. [7] [8] 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 approach is to simulate data, and then use that data to calculate the p-value of the Fisher exact test. The process is repeated 5000 times for each combination of sample size and nondetection rate, and then the 5000 p-values are compared to 0.05. The proportion of p-values less than 0.05 is the estimated power.

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 135 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.

The simulation algorithm is most easily described using examples, and we begin with calculating power for the Fisher exact test under the reference model, which is 100 cases and 100 correctly labeled (i.e., 100 truly unaffected) controls. That is, there are no affected cases in the control group, so this is not positive-unlabeled data, and the naive model is the correct model. Next we associate a binary risk factor variable, X (e.g., sex), with the cases and controls.The negative controls were 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 $Pr(X = 1) = 0.2 + 0.21$. Using the sex example, that means that having sex=1 predisposes the animals to about double the baseline risk of disease (the baseline is 0.2, and with sex = 1, the risk is double, 0.2+ + 0.21 = 0.41.).

Now, the 200 cases are pairs of binary data, one representing the group and the other representing the risk factor. These simulated data were tested with the Fisher exact test. As mentioned above, this process was repeated 5000 times and the resulting 5000 p-values used to estimate power.

For the second example, we calculate the power for a positive-unlabeled example. Suppose that in a 100-patient control group, 10 percent are in fact undetected positives. So the dataset is 10 affected cases in the control group, 90 unaffected cases in the control group, and 100 affected cases in the positive 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 same 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 p-values were calculated. This process is repeated 5000 times and the was estimated as described in the previous example.

The Correction Factor

For aim 2, the sample-size correction factor estimation, we used the same simulation algorithm and effect size (0.21) but multiplied the group sample sizes by possible correction factors, 1.1, 1.2, and so on, increasing sample size and therefore the power, until it reached the 80%.

Results

Table 1 describes power loss for the three study designs with three different group 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 in the analysis) has 150 cases, 15 of which are actually affected cases. When the nondetection rate is zero, 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.

Columns 5 and 6 are the power loss columns and have negative entries because for this simulation positive-unlabeled data analyzed under the naive model always had lower power, as did unbalanced data. Column 5 is the percent loss from the reference power (82%) and column 6 is the absolute power loss from the reference power. For example, the second row shows a -4.81% relative power reduction when the group sizes are balanced but five percent (0.05) of the control group are actually positive cases.

Rows one, four and seven are correct models (i.e., no positives in the control group). Rows four and seven show a power loss due to sample size imbalance only. So, for this small example, group imbalance sometimes caused more power loss than misclassified data as is seen by comparing row 3 to row four. It is known that for equal overall sample size, group imbalance results in less powerful tests. As an aside, more data is often better than less data, and it is sometimes better to have more unbalanced data than fewer balanced data.

Using Table 1, increasing nondetection rate within a study design decreased power. For example, for the (100, 100) study design, power decreased by more than 10% as the non-detection rate increased (row one to three). For the (50, 150) design, rows seven to nine, power decreased by 12.02% (22.47 - 10.45) from the correct model (row seven), but 22.47% from the reference model.

Finally, note that for this simulation, the absolute power loses are marked, but not extreme. For example, in the (100, 100) design, (rows 1 to 3) the power dropped to 0.73 (0.82 - 0.09) for row 3.

Table 1. Power loss. This table orders sample sizes by relative power loss (%). The first row is the reference power, which had an absolute power of 0.82 (82%). The last two columns represent power loss relative to 0.82, both as a percentage and absolute difference. Note that some inconsistancies in the table are due to rounding. For example, in rows 3 to 5, the absolute power is constant at 0.09, but the relative power changes.

	N positive	N naive	nondetection	Relative power	Absolute power loss
Row	cases	controls	proportion	loss (%)	(from 0.82)
1	100	100	0.00	0.00	0.00
2	100	100	0.05	-4.81	-0.04
3	100	100	0.10	-10.29	-0.09
4	150	50	0.00	-10.77	-0.09
5	150	50	0.05	-14.92	-0.13
6	150	50	0.10	-22.50	-0.20
7	50	150	0.00	-10.45	-0.09
8	50	150	0.05	-15.26	-0.13
9	50	150	0.10	-22.47	-0.20

Table 2 hows how many additional subject are require to regain power when the non-detection rate is 10%. Rows 1 to 5 are for the (100, 100) design, rows 6 to 10 are for the (150, 50) design, and rows 11 to 15 are for the (50, 150) design. Sample size was increased by 10% for each row within a study design. As one might expect, lower positive-unlabeled power needs more subjects to bring the power up to 80%. For the unbalanced designs, the increased sample size also fixes the power loss due to imbalance. For example, in row one, the power is 0.69 for the original (50, 150) design with 10% nondetection rate. Row 14 shows that an additional 65 subjects, or 32.5% more subjects are required to bring the power above 80%. However, for the (100, 100) design, only 10% more subjects are required (rows one and two).

Table 2. Power improvement. Rows 1, 6, and 11, show the power for when there are no false positives. The other rows show improvements in power when there is a 10% nondection rate. The sixth column shows the percent increase in sample size, and the last column is power.

Row	N positive cases	N naive controls	N false controls	N total	percent increase in N	Power
1	100	100	10	200	0.0	0.78
2	110	110	11	220	10.0	0.83
3	120	120	12	240	20.0	0.87
4	130	130	13	260	30.0	0.89
5	140	140	14	280	40.0	0.90
6	150	50	5	200	0.0	0.67
7	165	55	6	220	10.0	0.70
8	180	60	6	240	20.0	0.77
9	195	79	8	274	37.0	0.84
10	210	80	8	290	45.0	0.87
11	50	150	15	200	0.0	0.69
12	55	165	16	220	10.0	0.73
13	60	180	18	240	20.0	0.78
14	70	195	20	265	32.5	0.83
15	80	210	21	290	45.0	0.85

Discussion

This study demonstrated that under specific conditions, there was a modest decline in power, even with relatively small proportions of undetected positives in the control group. This implies that risk-factor studies might exhibit lower-than-expected power, consequently elevating the risk of encountering Type II errors. Importantly, it should be noted that changes in power can influence more than just statistical power itself. For instance, consider Table 1, rows 1 to 3, where the absolute power decreased from 0.82 (row 1) to 0.73 (row 3). To regain the reference power (as per Table 2), approximately 20 additional subjects would be required. In cases where subjects are expensive, the seemingly minor power drop of 0.09 translates into significant cost implications. Conversely, for an exploratory retrospective study, a power drop of 0.09 (equivalent to 9%) might not be considered substantial.

The working examples were derived from Genome-Wide Association Studies (GWAS), yet the simulation outcomes are applicable to any study utilizing univariate association tests, including various risk factor studies. This encompasses a wide spectrum of study types. For instance, it covers the univariate associations between post-operative surgical infections and different surgical conditions (e.g., board-certified surgeon vs. resident, bone plate manufacturer). In such scenarios, the control group may harbor subdiagnostic infections. Similarly, univariate association tests in veterinary surveys also fall within this category.

In this simulation, undetected positive cases in the negative group were randomly sampled from the same population as detected positives in the affected group. Although this is a common assumption, alternative models exist (Greenland & Rothman, 2004). In one such model, undetected positive cases in the control group could constitute a subpopulation of positives defined by another variable. For example, in a GWAS study focusing on cranial cruciate ligament disease (CCLD), the undetected positive cases in the control group might consist of positive dogs with low body condition scores. However, such models were not explored in this research. Our objective was to identify instances illustrating that in certain studies, misclassified data can lead to power loss. When this power loss is compounded with imbalanced data, the resulting loss can be substantial.

References

[1] Scott D Halpern, Jason HT Karlawish, and Jesse A Berlin. The continuing unethical conduct of underpowered clinical trials. *Jama*, 288(3):358–362, 2002.

[2] EH Hofmeister, J King, MR Read, and SC Budsberg. Sample size and statistical power in the small-animal analgesia literature. *Journal of Small Animal Practice*, 48(2):76–79, 2007.

[3] A. S. Waxman, D. A. Robinson, R. B. Evans, D. A. Hulse, J. F. Innes, and Conzemius MG. Relationship between objective and subjective assessment of limb function in normal dogs with an experimentally induced lameness. *Veterinary SurgeryApr;*, 37(3):241–6, 2008.

[4] E. G. Arthur, G. L. Arthur, M. R. Keeler, and Bryan JN. Risk of osteosarcoma in dogs after open fracture fixation. *Veterinary SurgeryJan;*, 45(1):30–5, 2016.

[5] C. E. Wylie, S. N. Collins, K. L. Verheyen, and 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 JournalOct*, 1(198):1, 2013.

[6] J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. *Machine LearningApr;*, 109:719–60, 2020.

[7] A. E. Baird, S. D. Carter, J. F. Innes, W. E. Ollier, and Short AD. Genetic basis of cranial cruciate ligament rupture (cclr) in dogs. *Connective tissue researchAug*, 1(55):4, 2014.

[8] E. Healey, R. J. Murphy, J. J. Hayward, M. Castelhano, A. R. Boyko, K. Hayashi, U. Krotscheck, and Todhunter RJ. Genetic mapping of distal femoral, stifle, and tibial radiographic morphology in dogs with cranial cruciate ligament disease. *PloS oneOct*, 17:14, 2019.