Sampling Size Calculations for Estimating the Proportion of False Positive and False Negative CLABSIs in the State of Texas

J. Charles Huber Jr, PhD
Associate Professor of Biostatistics

Ryan Hollingsworth
Graduate Research Assistant

Tyler Whittington, MPH
Graduate Research Assistant

Department of Epidemiology and Biostatistics
School of Rural Public Health
Texas A&M Health Science Center

August 31, 2011

Objective

The objective of this project was to calculate the sample size necessary to accurately estimate the proportion of false positive and false negative CLABSIs reported in the State of Texas.

Definitions and Assumptions

All sample size calculations are based on a set of assumptions. Where possible, the assumptions used in these calculations were chosen to represent the "worst case scenario" whereby the largest possible number of samples would be required. Therefore the following estimates represent the upper bound of the sample size required to meet the objective of the project.

The first set of assumptions required the estimation of the number of CLABSIs and potential but non-CLABSIs annually in the State of Texas. The Texas Hospital Discharge data for the years 2005 through 2009 were combined and cleaned. A record was defined as a CLABSI if the field "PRINC_DIAG_CODE" contained an ICD9-CM diagnostic code of 999.31. A record was defined as a potential but non-CLABSI if the patient spent any time in an ICU but did not have diagnostic code of 999.31. The rationale for this second assumption is that patients in an ICU are likely to have been exposed to a central line and thus had the potential for a CLABSI but did not receive a diagnostic code indicating a CLABSI. The largest annual number of records with a CLABSI diagnosis occurred in 2009 which totaled 4000. The largest annual number of potential CLABSIs records that did not have a CLABSI diagnosis was in 2009 which totaled 600,000.

The second assumption was the proportion of false positive and false negative CLABSI diagnoses. After an extensive literature review and multiple interviews with professionals in states with some kind of auditing system for CLABSIs, the best information available came from a December, 2010 paper published the American Journal of Infection Control (Backman, Melchreit & Rodriguez). A trained nurse auditor from the Infectious Disease Section of the Connecticut Department of Public Health reviewed 476 sample of which 27 were reported as CLABSIs to the US Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) database (see Table 2 from Backman, Melchreit & Rodriguez reproduced below. Of the 27 samples that were reported to the NHSN database as CLABSIs, 4 (14.81%) were determined not to meet the criteria for CLABSI by the auditor. Of the 449 samples that were not CLABSIs, 25 (5.57%) were identified as meeting the CLABSI criteria by the auditor. Therefore the proportion of false positives used in the sample size calculations below was 0.1481 and the proportion of false negatives used was 0.0557.

The third assumption which is not logistically or fiscally optimal but again, represents the "worst case scenario" is that the auditing sample will be based on a simple random sample with no use natural clustering or stratification of the CLABSI patients throughout the State.

Table 2. Comparison of central line associated bloodstream infections reported by Connecticut hospitals and the Connecticut Health Department reviewers

	CT hospital reports to the National Healthcare Safety Network		
CT DPH reviewers	**CLABSI**	**No-CLASBI**	**Total**
CLABSI	23	25	48
No-CLABSI	4	424	428
Total	27	449	476

CLABSI, central line-associated bloodstream infections; *CT*, Connecticut; *CT DPH*, Connecticut Health Department.

Methods

The sample size formula below used for the calculations comes from a standard survey sampling textbook by Levy and Lemeshow (1999). This equation was used for two reasons. First, this formula allows the calculation of the sample size based on a 95% confidence interval of a pre-specified size. In this case, the 95% confidence interval was selected to be ±0.03 around the assumed proportion. For example, if the proportion of false positives is assumed to be 15% then the sample size will ensure a 95% confidence interval of 12% to 18%. The second reason for using this equation is that it incorporates a finite population correction factor (FPC) that is advantageous when the sample size exceeds 10% of the size of the target population (van Belle, 2008). Specifically, the equation is:

$$n \geq \frac{z^2 NP\,(1-P\,)}{z^2 P\,(1-P\,)+(N-1)\varepsilon^2 P^2}$$

Where z is the z-score from a standard normal distribution associated with a two-sided alpha level of 0.05 (i.e. 1.96), N is the size of the target population from which the sample is to be drawn, P is the proportion of false positives (or false negatives) assumed, and ε is the "margin of error" (i.e. ±0.03).

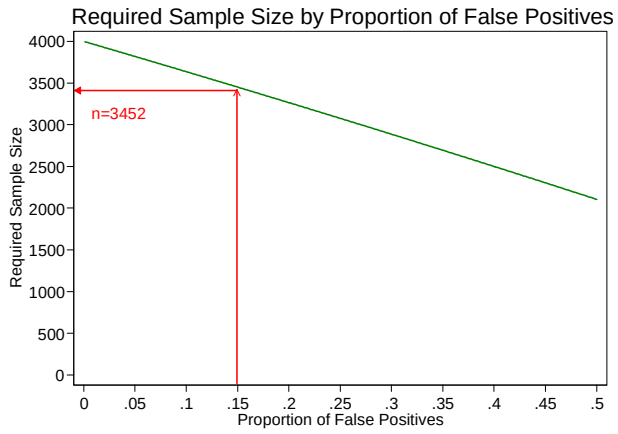
Results

Since the actual proportion of false positives and false negatives in Texas is unknown, programs were written using Stata 12 (see appendix) to calculate the necessary sample sizes for values of false positives ranging from 0.0 to 0.5. The results of these calculations are presented graphically in Figure 1 where the green line indicates the required sample size. The red lines

and number display the required sample size based on the results of the Backman, Melchreit & Rodriguez (2010) paper:

$$n \geq \frac{(1.96)^2(4000)(0.15)(1-0.15)}{(1.96)^2(0.15)(1-0.15)+(4000-1)(0.03)^2(0.15)^2} = 3452$$

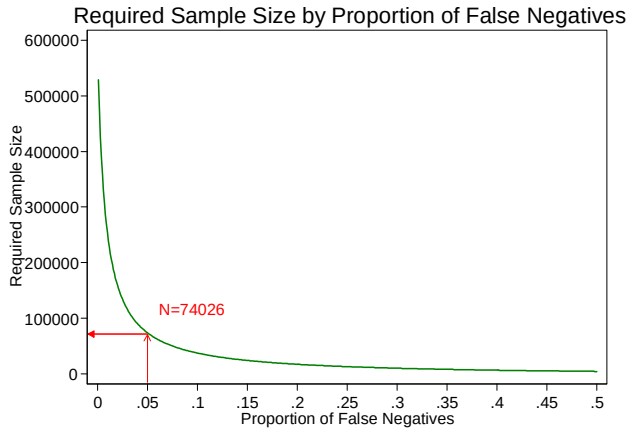
Figure 1



A similar program was written to calculate sample sizes for the range of false negatives and the results are presented in Figure 2. Again, the green line indicates the required sample sizes over the range of false negative values and the red line indicates the required sample size based on the results of the Backman, Melchreit & Rodriguez (2010) paper:

$$n \geq \frac{(1.96)^2(600{,}000)(0.05)(1-0.05)}{(1.96)^2(0.05)(1-0.05)+(600{,}000-1)(0.03)^2(0.05)^2} = 74026$$

Figure 2



Therefore it is estimated that 3452 + 74,026 = 77,478 records would need to be audited to allow the estimation of the proportion of false positive and false negative CLABSIs within ±3% accuracy annually in the State of Texas. If we further assume that a single auditor can review 20 records per day in a standard 250 day work year, the State would need to employ a minimum of 16 full-time auditors to achieve this goal.

References

Backman LA, Melchreit R, & Rodriguez R. (2010) Validation of the surveillance and reporting of central line-associated bloodstream infection data to a state health department. *American Journal of Infection Control*. 38: 832-838

Levy PS & Lemeshow S (1999) *Sampling of populations: Methods and applications*, 3rd Ed. New York: Wiley

Van Belle G. (2008) *Statistical Rules of Thumb*, 2nd Ed. New York: Wiley

Appendix: Stata programs used for the calculations and graphs

```
// False Positives Calculation
//=====================================================================
local p = 0.15
local N = 4000
local e = 0.03
local numerator = 4*`N'*`p'*(1-`p')
local denominator = (`N'-1)*(`e'^2)*(`p'^2) + 4*`p'*(1-`p')
local n = `numerator' / `denominator'
disp "n = `n'"


// False Positives Graph
// ====================================================================
clear
set obs 500
gen p = [_n]/1000
local N = 4000
local e = 0.03
gen numerator = 4*`N'*p*(1-p)
gen denominator = (`N'-1)*(`e'^2)*(p^2) + 4*p*(1-p)
gen n = numerator / denominator
twoway (line n p, lcolor(green) lwidth(medium) lpattern(solid) connect(direct)), /*
       */ title(Required Sample Size by Proportion of False Positives) /*
       */ ytitle(Required Sample Size) ylabel(0(500)4000, angle(horizontal)) /*
       */ xtitle(Proportion of False Positives) xlabel(0.0(0.05)0.5) /*
       */ scheme(s1color)


// False Negatives Calculation
//=====================================================================
local p = 0.05
local N = 600000
local e = 0.03
local numerator = 4*`N'*`p'*(1-`p')
local denominator = (`N'-1)*(`e'^2)*(`p'^2) + 4*`p'*(1-`p')
local n = `numerator' / `denominator'
disp "n = `n'"



// False Negatives Graph
// ====================================================================
clear
set obs 500
gen p = [_n]/1000
local N = 600000
local e = 0.03
gen numerator = 4*`N'*p*(1-p)
gen denominator = (`N'-1)*(`e'^2)*(p^2) + 4*p*(1-p)
gen n = numerator / denominator
twoway (line n p, lcolor(green) lwidth(medium) lpattern(solid) connect(direct)), /*
       */ title(Required Sample Size by Proportion of False Negatives) /*
       */ ytitle(Required Sample Size) ylabel(0(100000)600000, angle(horizontal)) /*
       */ xtitle(Proportion of False Negatives) xlabel(0.0(0.05)0.5) /*
       */ scheme(s1color)
```