Defining and Applying a Method for Establishing Gold Standard Sets of Emergency Room Visit Data

George S. Ghneim[1], D.V.M., M.P.V.M., Ph.D., Shiying Wu[1], Ph.D., M.S., Matt Westlake[1], M.S., Matthew J. Scholer[2], Ph.D., M.D., Debbie A. Travers[2], Ph.D., R.N., and Anna E. Waller[2], Sc.D., Scott F. Wetterhall[1], M.D., M.P.H.

[1]RTI International, Research Triangle Park, NC, [2]Dept of Emergency Medicine, UNC-CH.

OBJECTIVE

The goal of this paper is to describe a methodology used to create a gold standard set of emergency department (ED) data that can subsequently be used to evaluate the sensitivity and specificity of syndrome definitions.

BACKGROUND

One of the greatest challenges that has faced syndromic surveillance and early event detection systems to date has been their validation. Intricately tied into validation is the question of how to evaluate sensitivity and specificity, since it is very difficult to define a gold standard data set that is independent from the testing data set. The basis of our methodology is to randomly select a stratified sample of records, basing sample allocation on estimated prevalence, sensitivity, and specificity. Since only a small percentage of records in an ED database should be positive for any one syndrome, we created a broader less specific version of our standard respiratory syndrome definition which does not require a constitutional term that we could then draw our samples from. The next step was for three clinical experts to independently review the sampled records and assign a syndrome status to each record, according to our written syndrome definitions.

METHODS

There are about 1.1 millions total visits in the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) 2005 static database. Once we removed all injury related visits we were left with 956,015 records. Automated and broad respiratory syndrome queries were applied to the entire data set and every record was given a syndrome designation. The data were divided into four strata according to their query result and the availability of triage notes (more detailed notes on reason for ED visit that could be queried for syndrome classification). To choose the smallest sample size and optimal sample allocation we used the algorithm developed by Chromy (1983). In order to apply this algorithm, we used estimates of the prevalence rates for each of the strata obtained from a pilot study that had been conducted in which 1000 cases were sampled from the 2004 database. The prevalence of each stratum was estimated from this sample. The estimated prevalence rates in the four strata are quite different, implying that the stratified sample can be expected to be much more efficient than a simple random sample. Using estimated variances as inputs to the SAS code developed by Chromy to implement his algorithm, we obtained the minimum sample size and best sample allocation that satisfies the constraints on the variances of the estimated prevalence, sensitivity and specificity. Each stratum was selected independent of the others and delivered to the case reviewers for the clinical review process to begin. Cases were reviewed by two clinical experts, and a third reviewer adjudicated any reviewer disagreements.

RESULTS

We set the bounds for the estimated variances to 0.01%, so that the 95% confidence interval for each estimate would be (point estimate - 2%, point estimate +2%). This resulted in a total sample size of 3,699 records. We allocated our sampling into four categories: 503 records from S1 (broad definition positive with triage notes), 585 records from S2 (broad definition positive without triage notes) 418 from S3 (broad definition negative with triage notes), and 2193 records from S4 (broad definition negative without triage notes).

CONCLUSIONS

There were several challenges with creating a gold standard data set to use in evaluating sensitivity and specificity. The main issue we faced was how to draw a truly representative sample of records that remained unbiased. We eventually chose our sample from a subset of records that we processed with a broader respiratory syndrome definition. We feel that with a broader definition we would capture any possible records that meet the syndrome definition, and later be able to better assess sensitivity. Since the broad definition is different from our final syndrome query, the two processes are therefore independent. We believe that this process will allow the development of practical and useful gold standard datasets for the evaluation of syndromic definition sensitivity and specificity.

REFERENCES

[1] Chromy JR (1987), *Design optimization with multiple objectives*, In Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 194-199.