Statistical methods to compare different definitions of disease with an application to severe sepsis

Linda Peelen[1,3] **, Niels Peek**[1] **and Koos Zwinderman**[2]

[1] Dept. of Medical Informatics , [2] Dept. of Epidemiology and Biostatistics
Academic Medical Center, University of Amsterdam
P.O. box 22700 1100 DE Amsterdam, The Netherlands
Email: {l.m.peelen, n.b.peek, a.h.zwinderman}@amc.uva.nl
[3] NICE (National Intensive Care Evaluation) foundation

Abstract

Diagnostic categories are used to assemble, organize, and communicate knowledge from practitioners and clinical research. Unfortunately, the boundaries between these categories are seldomly clear-cut, and therefore multiple, competing definitions of a given disease often co-exist. It is important to assess whether such competing definitions relate to illnesses that differ in severity: in that case, the definitions cannot be used interchangeably. Differences in severity of illness are usually determined by statistical comparison of clinical outcomes in different patient groups. In this case, such an approach is hampered by the fact that many patients may comply to multiple definitions. This paper presents a statistical approach to comparing multiple definitions of the same disease with respect to binary clinical outcomes. Three methods of comparison are described and evaluated theoretically and empirically. The approach was applied to different definitions of the disease of severe sepsis.

Keywords: statistical testing; disease classification; severe sepsis; clinical trials

1 Introduction

Disease categories are the conceptual *loci* of medical knowledge in medicine. In clinical practice, the establishing of a patient's diagnosis (i.e. the assignment of a disease category) is the starting point for retrieving relevant knowledge that will guide further clinical procedures, and for gaining additional experience with the disease in question. It is therefore essential that the diagnostic categories are suited to assemble, organize, and communicate medical knowledge.

Diagnostic categories are often formed based on commonalities between patients that are medically crucial, such as the underlying etiology or pathophysiology. In many medical fields, however, the understanding of these aspects of the disease is insufficient to serve as a base for categories, and diseases are (completely or in part) defined in terms of clinical signs and symptoms. Another consequence is that for many diseases, different definitions circulate and are used by different doctors and in different hospitals. Similarly, each clinical trial defines its own patient inclusion criteria, although they all claim to study the same disease. Note however that many patients having the disease will satisfy all definitions that are in use.

The observation that different definitions are in use, has consequences both for clinical care and research. Results from clinical studies cannot be extrapolated and generalised straightforwardly, because the population under investigation might differ markedly from the population in clinical practice. Also the development of prognostic models is hindered if the data available to develop the model is based on patients who have been diagnosed with the same disease by different definitions. Note that these differences are only a problem if the patient groups are different *with respect to the medically important characteristics*, such as severity of illness.

To study whether there are medically relevant differences between two definitions, we now propose the following approach. First, we select a medical outcome (e.g. death) that may serve as an approximate measure for severity of illness in the disease in question. Second, cases are sampled and observed at random from the population of patients that satisfy either or both definitions. And third, the two groups of observed cases satisfying either definition are statistically compared with respect to the selected outcome. If a significant difference in outcome between both groups is found, then we conclude that the two definitions relate to illnesses that differ in severity and cannot be used interchangeably. However, because the two groups are likely to be overlapping, standard statistical comparison methods are inapplicable as these assume that the groups to be compared are disjoint.

In this paper we present three statistical methods to analyse the difference in outcomes of overlapping patient groups. We restrict to the situation where severity of illness is measured by a binary outcome. In Section 2, we discuss the three methods and evaluate them using simulations on artificially generated data. Section 3 presents an application in the area of severe sepsis, a typical example of a disease for which different definitions are in use. The methods are applied to the dataset from the Dutch National Intensive Care Evaluation [5] to compare six definitions of this disease that have been used in clinical trials. We summarise the main results and discuss our findings in Section 4.

2 Statistical comparisons of nondisjoint subsets

In this section, we discuss three statistical methods to compare the characteristics of nondisjoint subsets of observations. To simplify the presentation, we will restrict to the case where two subsets are to be compared (see [6] for a generalisation of the methods to the situation where k competing definitions are involved, $k > 2$). Furthermore, we assume that the comparison proceeds on a binomial outcome variable; in our study, the outcome of interest is death versus survival. We want to establish whether the mortality differs significantly between the subsets, using a given significance level α, $0 < \alpha \ll 1$.

Let S be a representative sample of i.i.d. observations from the population of interest. Let $D_1 \subseteq S$ and $D_2 \subseteq S$ be the subsets of observations that satisfy the two definitions in dataset S. For any subset of observations $S' \subseteq S$, let $Y(S')$ be the associated binary vector of measurements that is used to compare the two definitions. Furthermore, let π_i, $i = 1, 2$, be the parameter of interest in the population that is characterised by definition i, and let $p_i = \sum Y(D_i)/|D_i|$ be its estimate in dataset S. Our null hypothesis is that $\pi_1 = \pi_2$, which is immediately accepted if $p_1 = p_2$, and requires further investigation otherwise.

2.1 Parametric tests

The standard method to compare differences in binomial outcomes variables is to perform Pearson's χ^2 test [1]. Let $n = |D_1 \cup D_2|$ be the total number of observations in both subsets and let $\bar{p} = \sum Y(D_1 \cup D_2)/|D_1 \cup D_2|$ be the mean observed outcome in the union of both subsets. Then,

$$G = \frac{n-1}{n} \cdot \frac{(p_1 - p_2)^2}{\mathrm{SE}_0^2} \tag{1}$$

where

$$\mathrm{SE}_0^2 = \bar{p} \cdot (1 - \bar{p}) \cdot (\frac{1}{|D_1|} + \frac{1}{|D_2|}) \tag{2}$$

has a χ^2 distribution with one degree of freedom if $D_1 \cap D_2 = \varnothing$. If there is overlap between both subsets though, then all observations in the intersection $|D_1 \cap D_2|$ are used in the calculations of both p_1 and p_2, and they are used *twice* in the calculation of the variance SE_0^2. In such cases we cannot trust that G really has a $\chi^2(1)$ distribution.

A simple way to alleviate this problem is to randomly split S into two parts, apply one definition to one part and the other definition to the other and compute p_1 and p_2. Now the estimates are based on independent observations and we are certain that G has the appropriate distribution. However, this method is suboptimal for two reasons. First, the standard error will be much larger because the estimates are based on smaller subsets of observations. Second, an 'unfortunate' random split may obscure the difference between π_1 and π_2, and this further reduces the power of this method. If the differences between both definitions are subtle, or if the dataset is small, this method probably is not appropriate.

2.2 Analytic solutions

A rather different approach avoids the χ^2-test altogether by estimating the variances of the binomial parameters p_1 and p_2 from the data. We therefore distinguish three groups of patients: (a) patients that satisfy both definitions and are thus in $|D_1 \cap D_2|$; (b) patients that satisfy only the first definition and are thus in $|D_1 \setminus D_2|$; and (c) patients that satisfy only the second definition and thus are in $|D_2 \setminus D_1|$. Let π_a, π_b, π_c be the associated population parameters, and let as before p_a, p_b, and p_c be their estimates in dataset S, e.g. $p_a = \sum Y(D_1 \cap D_2)/|D_1 \cap D_2|$. Furthermore, let φ_k be the prevalence of group k, $k \in \{a, b, c\}$ and let f_k be its estimate. For instance, $f_b = |D_1 \setminus D_2|/|S|$.

We now write

$$\begin{aligned} p_1 - p_2 &= p_a \cdot \frac{\varphi_a}{\varphi_a + \varphi_b} + p_b \cdot \frac{\varphi_b}{\varphi_a + \varphi_b} \\ &\quad - \left(p_a \cdot \frac{\varphi_a}{\varphi_a + \varphi_c} + p_c \cdot \frac{\varphi_c}{\varphi_a + \varphi_c}\right) \\ &= p_a \cdot w_a + p_b \cdot w_b + p_c \cdot w_c, \end{aligned} \tag{3}$$

where $w_a = \frac{\varphi_a(\varphi_c - \varphi_b)}{(\varphi_a + \varphi_b)(\varphi_a + \varphi_c)}$, $w_b = \frac{\varphi_b}{\varphi_a + \varphi_b}$, and $w_c = \frac{-\varphi_c}{\varphi_a + \varphi_c}$. If we assume that the φ's are fixed and known, we have that

$$\begin{aligned} \mathrm{var}(p_1 - p_2) &= p_a(1 - p_a) \cdot w_a^2 \\ &\quad + p_b(1 - p_b) \cdot w_b^2 \\ &\quad + p_c(1 - p_c) \cdot w_c^2, \end{aligned} \tag{4}$$

and we can compute a $1 - \alpha$ confidence interval for $p_1 - p_2$ using a normal distribution. If the confidence interval does not contain 0, we reject the null hypothesis.

The advantage of this approach is that it takes into account the covariance of the estimates p_1 and p_2 that is due to overlap in the two definitions. There are however two problems with the approach. First, the empirical distribution of $p_1 - p_2$ may not be normal, especially when π_1 or π_2 is at extremities of the [0,1]-interval. Second, the parameters φ_a, φ_b, φ_c are neither fixed nor known, and have to be estimated from the data and 'plugged' into the above formulae. So, the uncertainty in the estimates of these parameters is not taken into account. The derivation of a correct model is highly complicated because the estimates of φ_a, φ_b, and φ_c are also dependent quantities; this issue not pursued here.

2.3 Bootstrap sampling

Bootstrap sampling [2; 3] is a computationally intensive technique to estimate the uncertainty in statistical estimates. In our case, we generate a sequence of artificial datasets $S_1^*, S_2^*, \ldots$ (called *bootstrap samples*) by drawing with replacement from dataset S; each of the artificial datasets has the same size as the original dataset S. From each bootstrap sample we estimate the relevant parameters; let $p_{i,1}^*, p_{i,2}^*, p_{i,3}^*, \ldots$ be the sequence of estimated parameters associated with definition i ($i = 1, 2$). Each $p_{i,j}^*$ can be regarded as the realisation of a random variable P_i^*. It can be shown that $E(P_i^*) = p_i$, and that $P_i^* - p_i$ has the same distributional properties as $P_i - \pi_i$ [2]. The estimated variance of P_i^* can therefore be used as a measure of uncertainty in our estimate of parameter π_i from dataset S. We use bootstrap sampling to investigate the properties of the statistic $(P_1 - P_2)^*$. To this end, we compute the difference $p_{1,j}^* - p_{2,j}^*$ from each bootstrap sample S_j^*, and then count whether a $1 - \alpha$ fraction of the computed differences has the same sign.

The bootstrap sampling approach is attractive because it can estimate the variance of p_1–p_2 even though the dependence of these estimates was not explicitly modelled. Also note that this approach avoids making assumptions on the distribution of the bootstrap variables P_1^* and P_2^*. However, the theoretical properties of bootstrap distributions might not be obtained in practice if the sample size is small. So, for small sizes of dataset S, the results of bootstrap sampling tests should be regarded with caution.

2.4 Experimental validation

The characteristics of the methods have been studied by means of application onto a series of artificial datasets. We applied the following method to generate the datasets: φ_a, φ_b and φ_c were randomly drawn from a uniform distribution on the [0,1]-interval with the restriction that $\varphi_a + \varphi_b + \varphi_c = 1$. The parameters π_a, π_b, and π_c were also drawn from a uniform distribution on the [0,1]-interval, with the restriction that in 50% of the cases $\pi_a = \pi_b = \pi_c$ (to simulate the situation where there is no difference in expected outcome); in the other cases the π's were generally close to each other (average difference between π_1 and π_2 was 0.04, see [6] for details). Based on these parameters a sample of 1000 observations was generated. To determine whether the associated sample estimates p_1 and p_2 were significantly different, we applied the approaches described above. This procedure was implemented in S-plus and repeated for 2000 times.

The results of this experiment are presented in Table 1 by the sensitivity and specificity of the methods. The table shows that all methods except the analytical method are highly specific, and will therefore rarely report an unwarranted significant difference. The sensitivity values suggest that the bootstrap method is more accurate than the methods based on χ^2 testing; especially the results with performing this test after splitting the dataset are disappointing. The results of the analytic solution are unsatisfactory: both its sensitivity and its specificity are low. It seems that the theoretical deficiencies of this method come with a high penalty in practice.

Method	Sensitivity	Specificity
Pearson's χ^2		
normal	0.35	0.96
after split	0.25	0.94
Analytic	0.07	0.52
Bootstrap	0.46	0.92

Table 1: Comparison of the methods based on the application on artificial data ($n = 2000, \alpha = 0.05$)

3 Application to severe sepsis in the ICU

The methods described above have been applied in the area of severe sepsis, which is a typical example of a disease with different definitions. We have selected six prominent sepsis trials from the last decade and derived their in- and exclusion criteria from publications and study protocols (see [7] for the trials included). These criteria have been applied to the dataset of the NICE (National Intensive Care Evaluation) foundation [5], which currently contains 71,929 records describing the first 24 hours and outcome of ICU stay of part of the Dutch ICU population. In this manner we created 6 subsets of patients that would have been eligible for these trials, one subset for each trial.

Our study focuses on the differences between these subsets with regard to ICU mortality as a proxy for severity of illness. Parsons χ^2 test for multiple proportions yielded a p-value $\ll 0.001$ ($\chi^2 \simeq 64.87$ with 5 degrees of freedom), suggesting that the subsets do not have an equal ICU mortality. We have used the methods described in Section 2 to study the differences in detail by comparing the individual subsets with all other individual subsets. To compensate for the fact that multiple comparisons are being made, the level of significance, $\alpha = 0.05$, has been adjusted in a conservative manner (see [6]), resulting in $\alpha' \simeq 0.0034$. Table 2 shows the results of these comparisons.

In six out of the fifteen comparisons, the methods yielded the same result (marked in bold). From this unanimity we conclude that the PROWESS and CORTICUS definitions and the Kybersept and Annane definitions do not lead to significantly different ICU mortality. The ICU mortality obtained by using the Norasept definition differs significantly from all others, except the Annane trial. In five comparisons (marked by an asterisk), the verdict of the 'χ^2 after split' method ('not significant') deviated from the others. Taking into account the low sensitivity of this method we found with the artificial data, we believe

Definition	**ICU**			**Compared to**		
[7]	mortality (%)	Lenercept	PROWESS	CORTICUS	Kybersept	Annane
Lenercept	23.6	-	-	-	-	-
PROWESS	25.5	NS †	-	-	-	-
CORTICUS	26.4	NS †	**NS**	-	-	-
Kybersept	29.3	S*	NS †	NS †	-	-
Annane	30.6	S*	S*	S*	**NS**	-
Norasept	39.1	**S**	**S**	**S**	**S**	S*

Table 2: Differences in ICU mortality in patient groups based on definitions of prominent sepsis trials, S denotes significant, NS denotes non-significant. Results in bold indicate where all methods agreed. * Only 'χ^2 after split' votes NS. † Only the analytic solution votes S.

that in these situations indeed a significant difference does exist. In four comparisons (marked by a dagger) only the analytic solution reported a significant difference. However, the artificial experiments have shown this method to be unreliable, which makes us believe that in these situations the ICU mortalities do not differ significantly. Note that the definitions have been ordered by the ICU mortality in the subset. The verdicts of the methods are in line with the expectations based on the ICU mortality in the different subsets.

4 Discussion and conclusions

This paper presents methods to compare overlapping patient groups which arise from different definitions of a single disease. The methods have been applied to artificially generated data and to real, clinical data pertaining to the disease of severe sepsis. We have shown that significant differences in ICU mortality exist between patient groups based on different definitions for severe sepsis.

Note that the differences we have assessed are significant from a *statistical* point of view. From a *clinical* point of view, a small difference in mean outcome may be uninteresting. In such cases another null hypothesis (e.g. $H_0 : |\pi_1 - \pi_2| < 0.05$) would be more appropriate; the properties of the methods presented here for such alternative testing hypotheses will have to be investigated in the future. Furthermore, clinicians and medical researchers may prefer clustered comparisons of definitions over pairwise comparisons. For example, the cluster of patients that fulfill the criteria of trials with a relatively low ICU mortality (i.e. the Lenercept, the CORTICUS or the PROWESS trial) may be compared to the Kybersept and/or Annane trial, which have a higher ICU mortality.

The results obtained in the artificial experiments deserve further investigation. Future experiments have to show the possible influence of unfortunate sampling on the results (which would explain why the analytic solution does not perform too bad with the real data). Furthermore, the methods can be refined and combined in various, more sophisticated ways. For example, under appropriate distributional assumptions a parametric bootstrap method may outperform the nonparametric bootstrap method that was employed here. Perhaps the limitations of the analytic solution may be overcome by combining it with bootstrapping.

Another limitation of the work presented is that we have restricted ourselves to the comparison of binary outcome variables. However, severity of illness is often quantified in terms of non-binary measures; examples are the time until a specific event occurs (e.g. survival time) and the patient's score on a heuristic, discrete scale (e.g. APACHE-score [4]). We intend to extend the methods currently described to such measures in the near future.

Acknowledgements

The authors wish to thank the NICE foundation for the use of the ICU data; Gert-Jan Scheffer, Nicolette de Keizer, Evert de Jonge and Rob Bosman for their contribution to the discussions about the application in severe sepsis. This study has been supported by a grant of Eli Lilly.

References

[1] D.G. Altman. *Practical Statistics for Medical Research.* Chapman & Hall, London, 1996.

[2] A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[3] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1982.

[4] W.A. Knaus, E.A. Draper, D.P. Wagner, et al. Apache II: A severity of disease classification system. *Crit Care Med*, 13:818–829, 1985.

[5] NICE foundation: http://www.stichting-nice.nl.

[6] L. Peelen and N. Peek. Methods to compare different definitions of disease. Technical Report TR 2003-02, Department of Medical Informatics, University of Amsterdam, October 2003.

[7] L.M. Peelen, N.F. de Keizer, N.B. Peek, E. de Jonge, R.J. Bosman, and G.J. Scheffer. Different definitions in sepsis trials yield significantly different study populations. *Intens Care Med*, 29(9), September 2003. *To appear.*