VALUE IN HEALTH ▮ (2017) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests

Emmanuel Chazard, MD, PhD[1,2,*], *Grégoire Ficheur, MD, PhD*[1,2], *Jean-Baptiste Beuscart, MD, PhD*[1,3], *Cristian Preda, PhD*[4,5]

[1]EA2694 *Santé publique: épidémiologie et qualité des soins, Université Lille, Lille, France;* [2]*Public Health Department, CHU Lille, Lille, France;* [3]*Geriatrics Department, CHU Lille, Lille, France;* [4]*Laboratory of Mathematics Paul Painlevé, Université Lille, Lille, France;* [5]*Inria Lille Nord Europe, MODAL, Villeneuve-d'Ascq, France*

ABSTRACT

Background: Although many researchers in the field of health economics and quality of care compare the length of stay (LOS) in two inpatient samples, they often fail to check whether the sample meets the assumptions made by their chosen statistical test. In fact, LOS data show a highly right-skewed, discrete distribution in which most of the observations are tied; this violates the assumptions of most statistical tests. **Objectives:** To estimate the type I and type II errors associated with the application of 12 different statistical tests to a series of LOS samples. **Methods:** The LOS distribution was extracted from an exhaustive French national database of inpatient stays. The type I error was estimated using 19 sample sizes and 1,000,000 simulations per sample. The type II error was estimated in three alternative scenarios. For each test, the type I and type II errors were plotted as a function of the sample size. **Results:** Gamma regression with log link, the log rank test, median regression, Poisson regression, and Weibull survival analysis presented an unacceptably high type I error. In contrast, the Student standard *t* test, linear regression with log link, and the Cox models had an acceptable type I error but low power. **Conclusions:** When comparing the LOS for two balanced inpatient samples, the Student *t* test with logarithmic or rank transformation, the Wilcoxon test, and the Kruskal-Wallis test are the only methods with an acceptable type I error and high power.

Keywords: length of stay, methodology, outcome measurement, statistics.

Introduction

The Inpatient Length of Stay

The *inpatient length of stay* (LOS) is defined as the period during which a patient is confined to a hospital or any other health care establishment [1]. The LOS is often studied by clinical researchers as a guide to the putative benefit of a treatment of interest. A shorter LOS (relative to a reference treatment or standard of care) may indicate clinical benefit, whereas a longer LOS may indicate the greater occurrence of treatment-related adverse events [2,3]. Conversely, the LOS is also an important risk factor for adverse events [4–6]. Furthermore, the LOS is frequently used as a key indicator of operational efficiency and sometimes as a proxy for quality-of-care processes [7]. Health economists also use the LOS to estimate health expenditure because health care establishments mainly have fixed costs (such as salaries). Consequently, more than 2200 articles a year refer to the LOS in their abstract or title (according to the PubMed database; Table 1). Furthermore, researchers use various statistical methods to compare the LOS in two patient samples.

Modeling the LOS

Fitting the LOS distribution using various statistical models has been extensively studied [8–20]. Although the LOS is always considered as the dependent variable (Y), the independent variables (X_i) considered to be statistically significant vary as a function of the selected model [8,9,17,19]. Furthermore, when two methods generate different results after application to real data, researchers are unable to determine which result is true. One can therefore hypothesize that if the goal is to identify statistically significant explanatory variables, the type I and type II errors will differ from one method to the other. The scientific literature, however, does not provide any guidance on choosing the most appropriate method.

Conflicts of interest: The authors declare that there are no conflicts of interest.

* *Address correspondence to:* Emmanuel Chazard, Public Health Department, Faculté de Médecine de Lille, CERIM, Lille Cedex F-59045, France.

E-mail: emmanuel.chazard@univ-lille2.fr.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2017.02.009

Table 1 – Number of articles referenced in PubMed during the last 10 y (January 2006 to December 2015).

Concept	Cited in the title	Cited in the title or the abstract
Length of stay (LOS)*	1,704	22,548
Type I error†	100	2,180
Type II error‡	223	4,925
Type I error† and LOS*	0	4
Type II error‡ and LOS*	0	12

* Keyword: "length of stay."
† Keywords: "type 1 error," "type I error," "first type error," or "alpha risk."
‡ Keywords: "type 2 error," "type II error," "second type error," "beta risk," or "statistical power."

Comparing the LOS Distributions in Two Independent Samples

Researchers often want to compare the mean LOS of two independent samples. This can be done with three families of methods [13,21]: 1) bivariate statistical tests (i.e., parametric tests such as the Student t test or nonparametric tests such as the Wilcoxon test); 2) regression models (such as gamma regression [22]), in which the LOS is the dependent variable Y and the samples are labeled by a binary independent variable X; and 3) survival analyses (e.g., the log rank test or the Cox models) in which the discharge is the observed outcome and the LOS is the time to the outcome. Although there are no censored values (because the patient is always discharged), survival analyses can be used as "less parametric" alternatives to traditional statistical tests [9,13]. We will now provide a more precise description of 12 of these methods.

Type I and Type II Errors in the Comparison of Two Samples

In the present case, the type I error (the alpha risk) corresponds to the probability with which a test will detect a significant difference in the mean LOS between two samples—even though the samples have been drawn from the same population. In most studies, the null hypothesis is rejected when the *P* value is less than 0.05, which leads researchers to assume that the type I error is 5%. In the field of medical research, increasing the type I error beyond 5% is considered to be unacceptable, because it could generate erroneous knowledge and prompt physicians to make inappropriate diagnostic and therapeutic decisions [23]. The type II error (the beta risk) corresponds to the probability with which a statistical test will not detect a difference in the mean LOS between two samples, even though the samples have been drawn from populations whose means were different. The power is defined as 1 − type II error.

Type I and type II errors have been extensively studied in the literature on variables not related to the LOS (Table 1). The literature contains many general assertions about type I and type II errors in statistical tests [21]. These assertions are not always evidence-based and cannot be generalized without considering the distribution of the variable under investigation. For example, Skovlund and Fenstad [24] developed an algorithm for determining the best way of comparing two samples. The appropriate choice depended on the equality of variance, the sample imbalance, and (most importantly) the skewness.

The Distribution of LOS Data

LOS data have a very particular distribution: a highly right-skewed, discrete, positive distribution with many tied observations, with values concentrated around the median [8,14,20,25,26]. In some health care establishments, the LOS distribution might be multimodal and depend on how care is organized. Last, LOS samples may contain a few outliers with extremely high values. Expert opinion suggests that these outliers should be excluded from analysis [25–27]; unfortunately, automated approaches have yet to be developed.

A Lack of Specific Research on Comparing LOS

Although mean LOS values are often compared using statistical tests, there is a lack of knowledge in this field. In particular, the type I and type II errors have not been empirically evaluated for this specific distribution (see Table 1). To our knowledge, all the studies in this field to date have assumed that the type I error is always controlled: researchers have assumed that under the null hypothesis, there is a 5% probability that the *P* value is less than 0.05. We believe that this assumption should be questioned. Furthermore, most investigators do not check the validity of the assumptions of the chosen statistical tests [21]. Consequently, it is important to empirically check the type I and type II errors even when statistical tests are inappropriately used (e.g., a parametric test applied to a small sample or a rank test applied to tied observations).

Hence, the objective of the present study was to empirically evaluate statistical tests that are frequently used to compare the mean LOS of two independent samples, with regard to the type I error under the null hypothesis and the type II error under three realistic, alternative hypotheses. We evaluated the tests even when their assumptions were not met so as to determine and consider the consequences of inappropriate application.

Methods

Estimation of the LOS Distribution Function

We first queried the French nationwide hospital discharge database programme de médicalisation des systèmes d'information (PMSI) to obtain the LOS empirical distribution function. This database is based on compulsory, standardized discharge reports on all patients admitted to nonprofit acute-care hospitals in France and is used to calculate a significant proportion of a hospital's public funding. Each discharge report describes the patient's administrative and demographic data, diagnoses, and medical procedures. The database query included all inpatient stays for 2012 and excluded outpatients and iterative ambulatory treatments (dialysis etc.). The total number of included stays was 9,895,673. For each inpatient stay, the LOS is defined as an integral number of calendar days (see Equation 1).

$$\text{LOS} = \text{Discharge\_date} - \text{Admission\_date} + 1. \quad (1)$$

The probability density function of LOS was estimated from the result of the database query. Univariate statistics were calculated.

The Statistical Tests

This section introduces the statistical tests used to assess the difference between the LOS X of two independent samples of patients. Nevertheless, actual examples of use will be given later in the article. Here, X is the random variable representing the LOS, μ is the mean of X, and G is the binary variable "group name," with possible values of $\{G_1; G_2\}$.

Bivariate methods are used to test whether $\mu_{X_{G1}}$ and $\mu_{X_{G2}}$ are different. For each test, the *P* value is considered to be the main output. Methods based on regression models are used to test whether X can be explained by G. In that case, a coefficient β_G is calculated for G and is tested against the null hypothesis $\beta_G = 0$. The *P* value of this test is again considered to be the main output. For survival-based methods, all subjects are assumed to have an