VALUE IN HEALTH 20 (2017) 992–998

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

event (i.e., discharge), and X is used as the time to event. The tests are then performed the same way as described earlier.

All simulations and statistical tests were performed using R software and the *MASS*, *survival*, and *quantreg* libraries (R Foundation for Statistical Computing, Vienna, Austria) [28–30]. The statistical tests and their corresponding R codes are given in Table 2. Hereafter, the tests are referred to by the abbreviations given in Table 2 (e.g., *Student* for Student *t* test).

Table 2 – List of the statistical tests and the corresponding R code*,†.

Statistical method (and abbreviation)	R code to get the *P* value
Statistical tests	
Kruskal-Wallis test (KruskalWallis)	obj <-kruskal.test(x=X, g=G) pval<-obj$p.value
Student *t* test (Student)	obj <- t.test(X1,X2, var.equal=FALSE) pval<-as.numeric(obj$p.value)
Student *t* test with log transformation (StudentLog)	X1 <- log(X1) X2 <- log(X2) # then the same code as Student
Student *t* test on the ranks (StudentRanks)	X1 <-rank(X)[G==0] X2 <-rank(X)[G==1] # then the same code as Student
Wilcoxon or Mann-Whitney test (Wilcoxon)	obj <-wilcox.test(X1,X2) pval<-obj$p.value
Methods based on regression models	
Gamma regression with a log link (GammaReg)	fam <-Gamma(link="log") obj <-glm(X~G, family=fam) pval<-summary(obj)$coefficients[2,4]
Linear regression with a log link (LinearReg)	fam <-gaussian(link="log") obj <-glm(X~G, family=fam) pval<-summary(obj)$coefficients[2,4]
Median regression (MedianReg)	library(quantreg) obj <-rq(formula=X~G, tau=0.5) pval<-summary(obj,se="ker")$coefficients[2,4]
Poisson regression (PoissonReg)	fam<-poisson obj <-glm(total ~ groups, family=fam) pval<-summary(obj)$coefficients[2,4]
Survival methods	
Cox proportional hazard model (CoxPH)	library(survival) obj <-coxph(formula=Surv(X)~G) pval<-summary(obj)$coefficients[5];
Log rank test (LogRank)	library(survival) obj <-survdiff(formula=Surv(X)~G) pval<-1-pchisq(obj$chisq, df=1)
Weibull survival (Weibull)	library(survival) obj <-survreg(formula=Surv(X)~G, dist="weibull") pval <-1-pchisq(2*diff(obj$loglik), sum(obj$df)-obj$idf)

LOS, length of stay.

* An example of implementation is available in Appendix B in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

† In the code, X1 and X2 are two samples of LOS. X results from pooling X1 and X2, then G is the binary variable corresponding to the groups. pval is the *P* value computed for each test.

Estimation of the Type I Error

When a *P* value of less than 0.05 is the threshold for statistical significance, the statistical tests' type I error is expected to be 5%. We empirically estimated this risk by simulation. For each fixed sample size n ∈ {10; 15; 20; 25; 30; 40; 50; 60; 80; 100; 130; 160; 190; 220; 260; 300; 350; 400; 450}, two samples were randomly drawn with replacement from the LOS population described earlier. The statistical tests were performed under the null hypothesis, and the *P* value of each statistical test was recorded. This process was iterated 1,000,000 times.

Next, the empirical cumulative distribution function (ECDF) of the *P* values, *F*(*P*), from each test was estimated under the null hypothesis. In theory, each ECDF should be a straight line corresponding to Equation 2, given that Equation 3 is true. To verify this hypothesis, Equation 4 was then used to estimate the empirical type I error for each test with a 5% threshold.

$$\forall X \in [0;1],\ Y = X, \tag{2}$$

$$\forall P_0,\ P_0 = P(P < P_0), \tag{3}$$

$$\text{Empirical type I error} = P(P < 0.05/H_0). \tag{4}$$

Estimation of the Type II Error, Power, and Relative Efficiency

The type II error can be estimated only for a specific alternative hypothesis and should generally be as high as possible—although the value is constrained by the type I error. In the present study, we simulated three simple hypotheses. According to the alternative hypothesis H_{1a}, the first sample was drawn from the original LOS distribution and the second sample was drawn from a distribution that is shifted to the right by 1 day. In the alternative hypothesis H_{1b}, each individual in the second sample had its LOS value increased randomly by 0, 1, or 2 days, with a probability of 1/3 for each possible value. Last, the alternative hypothesis H_{1c} is a real-life hypothesis: the first sample was composed of inpatient stays without surgery, and the second sample was composed of inpatient stays with surgery.

For each fixed sample size n ∈ {10; 15; 20; 25; 30; 40; 50; 60; 80; 100; 130; 160; 190; 220; 260; 300; 350; 400; 450}, two samples were drawn with replacement. For the H_{1a} and H_{1b} hypotheses, both samples were drawn from the original LOS distribution, and the second sample was then transformed (depending on the hypothesis). For the H_{1c} hypothesis, the nationwide LOS data were first separated into inpatient stays without surgery and inpatient stays with surgery, and one sample was drawn from each of the data sets. The statistical tests were performed, and the *P* value of each statistical test was recorded. This process was iterated 100,000 times. Next, the ECDF of the *P* value from each test was estimated. The empirical power was then estimated for each test and for each hypothesis with a 5% threshold, using Equation 5.

$$\text{Power} = 1\text{-Type II error} = P(P < 0.05/H_1). \tag{5}$$

For each hypothesis and for each statistical test, we calculated the efficiency relative to *Student* (because it is by far the most frequently used test in the literature). Here, efficiency was defined as the sample size that is required for a method to achieve a predefined power (0.5 in the present study). Hence, the efficiency of a test t_1 relative to *Student* is the efficiency of *Student* divided by the efficiency of t_1. Accordingly, a test t_1 was considered to be more efficient than *Student* if the relative efficiency was greater than 1, and a test t_2 was considered more efficient than a test t_1 if its relative efficiency was greater than the value for t_1.

If a test did not achieve a power of 0.5 with the largest sample size (300), we recorded the highest power achieved. Because only

19 sample sizes were tested, the relative efficiency at a power of 0.5 was estimated by linear interpolation.

Workflow and Decision Thresholds

The type I error was estimated for the 12 statistical methods and the 19 sample sizes. If one or more of the type I errors exceeded 5.5%, the method in question was excluded from the power analysis. If one or more of the type I errors were between 5.1% and 5.5%, the method was classified as being subject to caution. All other methods were classified as being appropriate for use.

Next, each statistical test's power was estimated for each of the alternative hypotheses H_{1a}, H_{1b}, and H_{1c} and each value of n. The efficiency was estimated relative to *Student*. The tests were then classified into three efficiency groups: low, moderate, and high. The overall workflow is depicted in Appendix Figure C3 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

Ethics

The present study used anonymized secondary data that had been collected in the course of routine care. In line with the French legislation on data set analyses, approval by an investigational review board was neither required nor sought. The work complied with the tenets of the Declaration of Helsinki. The collection and analysis of data was authorized by the French national data protection commission (*Commission Nationale de L'informatique et des Libertés*, Paris, France).

Results

LOS Distribution with Real Hospital Data

The LOS distribution was derived from the total number of inpatient stays in France in 2012. The data distribution was discrete, as shown in Figure 1. The LOS values ranged from 1 day to 1247 days (3.41 years). The mode was 1 day and was associated with a probability of 28.1%. The mean LOS was 5.44 ± 7.80 days. The median LOS was 3 days, and 33.4% of all values were concentrated around the median ± 1 day. The distribution was strongly skewed to the right, with a skewness of 9.26 and a kurtosis of 353.9. Detailed results are provided in Appendix A in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009. A large minority of the patients (26.8%) were hospitalized several times, and 2.8% of patients were hospitalized 5 times or more. Removal of duplicates modified the distribution slightly: the mean LOS was 5.28 ± 7.65 days, the median was still 3 days, 34.7% of the LOS values were between 2 and 4 days, the skewness was 11.7, and the kurtosis was 340.7. These metrics are provided for information purposes only; hereafter, only the raw distribution will be described.

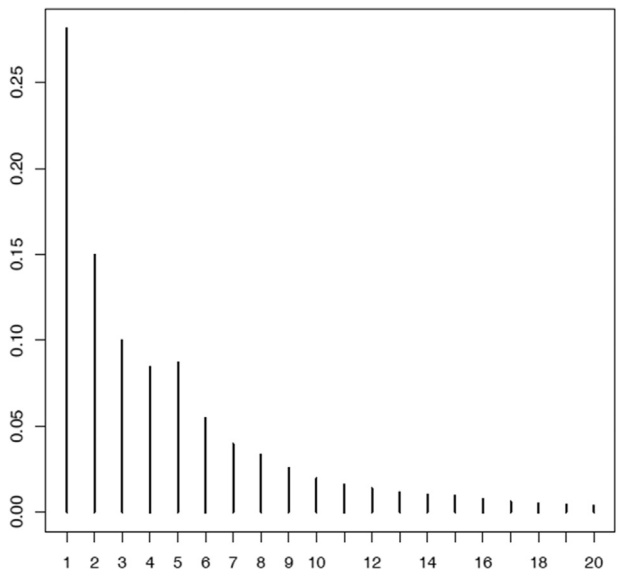

Fig. 1 – Estimated probability density (*y*-axis) function of the length of stay (*x*-axis, truncated to 20 days).

Estimation of the Type I Error

Under the null hypothesis, the ECDF of the *P* values from each of the 12 statistical tests was estimated for each sample size. The 228 ECDF curves (12 tests × 19 sample sizes) were used to estimate the type I error with *P* less than 0.05. Three groups of tests were identified (Fig. 2). The *KruskallWallis*, *LinearReg*, *Student*, *StudentLog*, *StudentRanks*, and *Wilcoxon* tests were found to be appropriate; it is particularly noteworthy that the type I errors for *Student* and *LinearReg* were well less than 5%. Caution should be exercised with regard to *CoxPH*; this method was associated with a moderately high type I error for sample sizes smaller than 80. The *GammaReg*, *LogRank*, *MedianReg*, *PoissonReg*, and *Weibull* tests were excluded because of very high type I errors, and so should not be used for LOS comparisons.

Detailed methods and results are provided in Appendix D in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

Estimation of the Type II Error

The type II error and the power were estimated for the three alternative hypotheses, the 19 sample sizes, and the seven tests with an acceptable type I error. The 399 resulting ECDF curves were used to estimate the power with *P* less than 0.05.

Under the H_{1a} hypothesis, the power was high for *KruskallWallis*, *StudentLog*, *StudentRanks*, and *Wilcoxon* (Fig. 3), moderate for *CoxPH*, and low for *Student* and *LinearReg*. The results under the H_{1b} hypothesis were very similar. Under the H_{1c} hypothesis, the power was low for every test and increased linearly with the sample size. Table 3 presents the relative efficiency with respect to *Student*. In the high-power group, the relative efficiencies ranged from 7.9 to 8.5 under the H_{1a} hypothesis and were around 6.3 under the H_{1b} hypothesis. Within this group of tests, *StudentLog* was the most efficient. Under the H_{1c} hypothesis, the highest relative efficiency was observed for *CoxPH* (1.172). Additional methods and results are provided in Appendix C in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

The Tests' Assumptions

When we looked at whether the statistical tests' underlying assumptions were violated, we noticed that the LOS, its logarithm, and its ranks did not follow a normal distribution. Consequently (given the central limit theorem), *Student*, *StudentLog*, and *StudentRanks* were not valid for an n value of less than 30 but were valid for an n value of 30 or more. The assumptions were always met for *KruskalWallis* and *Wilcoxon*, despite a high proportion of tied data (e.g., ties accounted for 37% of the sample for n = 10, 50% for n = 30, and 68% for n = 300). When we analyzed the residuals of 76 random experiments (4 per sample size), we observed that the residuals of the regressions were never normally distributed for *LinearReg*, *MedianReg*, *GammaReg*, and *PoissonReg*. We next examined survival methods. The proportional hazard assumption of the *CoxPH* was violated in only 4 of the 76 random experiments. When we drew quantile-quantile plots of the event times, we observed that they fitted a *Weibull*

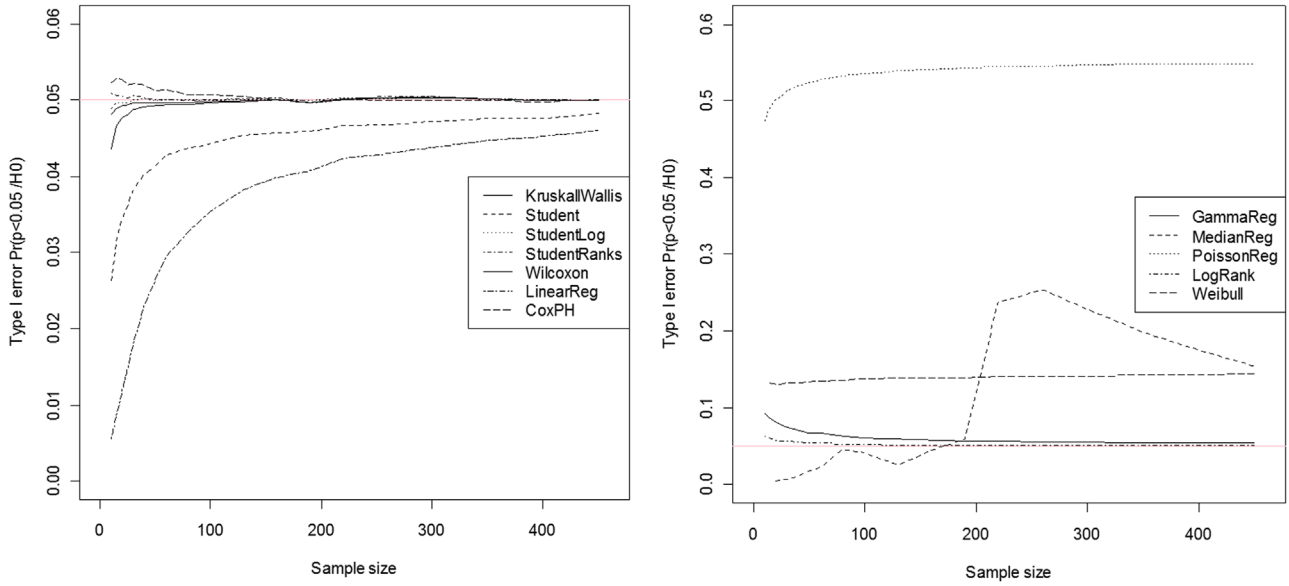

Fig. 2 – Type I error as a function of the sample size (left: seven nonexcluded methods; right: five methods with unacceptable inflation of type I error).

distribution in 73 of the 76 random experiments. Last, the assumptions of *LogRank* were always met, because there was no censoring in our simulations.

Discussion

In the present study, we evaluated 12 statistical tests that are frequently used to compare the LOS of 2 balanced samples. Given the very particular distribution of LOS data in administrative databases (a discrete, highly right-skewed distribution with tied observations and many outliers), the tests' validity cannot be guaranteed and, in practice, is not checked often enough [21]. We had three main findings. First, some methods should not be used because they have a type I error more than 5% if the null hypothesis is rejected at a P value of less than 0.05 (*GammaReg*, *LogRank*, *MedianReg*, *PoissonReg*, and *Weibull*). Second, some of the methods with an acceptable type I error are also similarly and highly efficient, relative to *Student* (*KruskallWallis*, *StudentLog*, *StudentRanks*, and *Wilcoxon*). Third, the differences in the tests' performance under three alternative hypotheses suggest that the relative efficiency depends on the characteristics of the samples being compared.

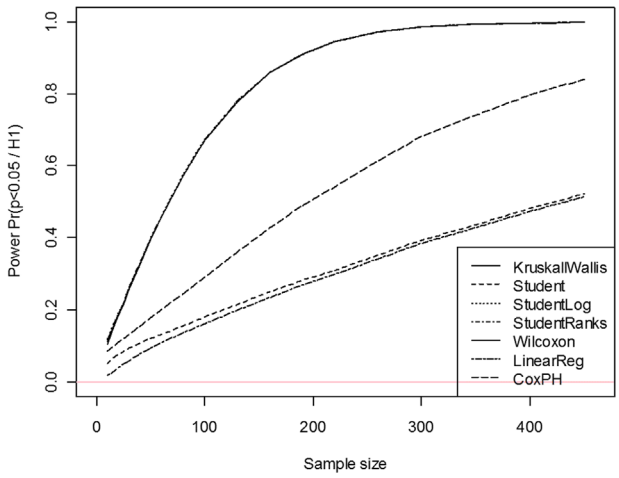

Fig. 3 – Power as a function of the sample size for the seven methods that are not excluded, for the alternative hypothesis H_{1b} (all the LOS values are increased at random by 0, 1, or 2 d). *LOS*, **length of stay.**

We focused on the 12 tests used most frequently in the literature. Some other tests could have been considered, and we did not evaluate approaches such as data truncation. We did not test the effects of outlier removal because a standardized consensus approach is not available, and so the number of solutions would have been infinite [25–27].

The present work was based on the French national inpatient stay database. The data distribution was exhaustively measured and not estimated. Furthermore, we performed simulations strictly, and the type I and type II errors were estimated in the absence of previous hypotheses. The validity of the statistical methods' assumptions was not checked during simulations because the objective was to evaluate the consequences of this violation, just as many researchers fail to check those assumptions in their peer-reviewed articles [21]. Our results demonstrated that the type I and type II errors obtained are not necessarily related to the validity of the tests' assumptions. For instance, *StudentLog* and *StudentRanks* produced very good results even when the sample size was too small. In contrast, *LogRank*

Table 3 – Efficiency of the nonexcluded tests relative to the Student *t* test (a high value denotes a high power).

Test	Hypothesis H_{1a}	Hypothesis H_{1b}	Hypothesis H_{1c}
CoxPH	2.311	2.164	1.172
KruskallWallis	7.968	6.315	1.005
LinearReg	0.981	0.980	0.971
Student	1	1	1
StudentLog	8.534	6.360	1.040
StudentRanks	7.992	6.331	1.006
Wilcoxon	7.939	6.301	1.005

produced a high type I error and *CoxPH* produced an intermediate power level when used under valid conditions.

The present study had a number of limitations. First, we did not take account of the fact that patients made more than one stay. Second, we chose to use balanced samples ($n_1 = n_2$), and the three alternative hypotheses used to estimate the type II error could not cover all possible configurations. We have focused on mean differences but did not evaluate the effect of differences in variance (heteroskedasticity) or skewness, which might even impair the application of nonparametric tests [24]. A simulation study should be based on a hypothesis that corresponds precisely to the situation to be tested.

Nevertheless, our results also suggest that whichever statistical test is applied, its *P* value should be corrected by using bootstrap techniques [31]. In short, bootstrap techniques enable one to replace the *P* value by an *F*(*P*) value (where *F* is the ECDF of the *P* values under the null hypothesis) for each performance of the test. Even when *P* values are corrected by bootstrapping, differences between statistical tests may nevertheless still exist.

To compare the LOS of two inpatient samples, we recommend using the Student *t* test with logarithmic or rank transformation, the Kruskal-Wallis test, or the Wilcoxon (Mann-Whitney) test. Our results agree with those of a previous study [21] in which the Wilcoxon and Kruskal-Wallis tests had a higher power than the Student *t* test (without transformation) for the analysis of LOS in an emergency department (as a continuous variable). It is often stated that the Wilcoxon and Kruskal-Wallis tests' validity can be questioned for tied data. Nevertheless, our simulation shows that ties are not problematic in this context. The tiny difference between the Wilcoxon and Kruskal-Wallis tests was simply because of the different way in which they handle tied observations, which are prevalent. It is noteworthy that the Student *t* test with logarithmic or rank transformation is just as efficient as the Wilcoxon and Kruskal-Wallis tests, and it may be more accessible for nonstatisticians. Data transformation, however, has some important drawbacks. Transformation can easily be applied if the purpose of the test is solely to compute a *P* value. Nevertheless, such a transformation can be likened to a "black box," which makes it more difficult to interpret the effect size and its confidence interval.

The Student *t* test is by far the most frequently applied test in this context. Our present results showed that this test was very conservative (i.e., with a type I error well less than 5%) but had low power. This observation is very reassuring because in real life, many processes tend to increase the false-discovery rate. Many statistical studies are data-driven (in which tests are performed *because* something is visible in the descriptive analyses); repeated statistical tests are often performed without Bonferroni or Šidák correction, underpublication bias is produced because negative results are not to be submitted or not accepted by journals, and public opinion and decision makers often focus on studies that reject the null hypothesis (even when many other published studies do not).

With regard to methods based on regression models, our results suggest that gamma, median, and Poisson regressions should be avoided because of their type I error, and that linear regression should be avoided because of its low power (the Cox model is preferable). Some researchers have reported the superiority of gamma regression [10,12] and the inferiority of the Cox model [13], although these studies sought to fit the LOS by decreasing the residuals of the prediction and were not designed to evaluate type I and type II errors.

Conclusions

To compare the LOS of two balanced samples of inpatients, we recommend the Student *t* test with logarithmic or rank transformation, the Kruskal-Wallis test, or the Wilcoxon (Mann-Whitney) test. If the LOS distribution differs greatly from that used in the present simulation study, we recommend using bootstrap techniques to recalibrate the chosen statistical method.

Source of financial support: This study did not receive any funding.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2017.02.009 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Medical Subject Headings. Length stay. Available from: http://www.ncbi.nlm.nih.gov/mesh?term=length%20of%20stay. [Accessed May 18, 2016].
[2] Classen DC, Pestotnik SL, Evans RS, et al. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. JAMA 1997;277:301–6.
[3] Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 2003;290:1868–74.
[4] Bates DW, Miller EB, Cullen DJ, et al. Patient risk factors for adverse drug events in hospitalized patients. ADE Prevention Study Group. Arch Intern Med 1999;159:2553–60.
[5] Morimoto T, Sakuma M, Matsui K, et al. Incidence of adverse drug events and medication errors in Japan: the JADE study. J Gen Intern Med 2011;26:148–53.
[6] Hauck K, Zhao X. How dangerous is a day in hospital? A model of adverse events and length of stay for medical inpatients. Med Care 2011;49:1068–75.
[7] Diercks DB, Roe MT, Chen AY, et al. Prolonged emergency department stays of non-ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. Ann Emerg Med 2007;50:489–96.
[8] Austin PC, Rothwell DM, Tu JV. A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. Health Serv Outcomes Res Methodol 2002;3:107–33.
[9] Dudley RA, Harrell FE Jr, Smith LR, et al. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. J Clin Epidemiol 1993;46:261–71.
[10] Marazzi A, Paccaud F, Ruffieux C, Beguin C. Fitting the distributions of length of stay by parametric models. Med Care 1998;36:915–27.
[11] Lee AH, Fung WK, Fu B. Analyzing hospital length of stay: mean or median regression? Med Care 2003;41:681–6.
[12] Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ 2005;24:465–88.
[13] Basu A, Manning WG, Mullahy J. Comparing alternative models: log vs Cox proportional hazard? Health Econ 2004;13:749–65.
[14] Lee AH, Gracey M, Wang K, Yau KKW. A robustified modeling approach to analyze pediatric length of stay. Ann Epidemiol 2005;15:673–7.
[15] Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ 2001;20:461–94.
[16] Samore MH, Shen S, Greene T, et al. A simulation-based evaluation of methods to estimate the impact of an adverse event on hospital length of stay. Med Care 2007;45:S108–15.
[17] Faddy M, Graves N, Pettitt A. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. Value Health 2009;12:309–14.
[18] Singh CH, Ladusingh L. Inpatient length of stay: a finite mixture modeling analysis. Eur J Health Econ 2010;11:119–26.
[19] Ravangard R, Arab M, Rashidian A, et al. Comparison of the results of Cox proportional hazards model and parametric models in the study of length of stay in a tertiary teaching hospital in Tehran, Iran. Acta Med Iran 2011;49:650–8.
[20] Moran JL, Solomon PJ. A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and New Zealand intensive care adult patient data-base, 2008–2009. BMC Med Res Methodol 2012;12:68.
[21] Qualls M, Pallin DJ, Schuur JD. Parametric versus nonparametric statistical tests: the length of stay example. Acad Emerg Med 2010;17:1113–21.

[22] Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. J Health Econ 1999;18:153–71.
[23] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference, part I. Biometrika 1928;20A:175–240.
[24] Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J Clin Epidemiol 2001;54:86–92.
[25] Ruffieux C, Marazzi A, Paccaud F. Exploring models for the length of stay distribution. Soz Praventivmed 1993;38:77–82.
[26] Weissman C. Analyzing intensive care unit length of stay data: problems and possible solutions. Crit Care Med 1997;25:1594–600.
[27] Lee AH, Xiao J, Vemuri SR, Zhao Y. A discordancy test approach to identify outliers of length of hospital stay. Stat Med 1998;17: 2199–206.
[28] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.
[29] Therneau T, Lumley T. Survival: survival analysis, including penalised likelihood. R package version 2.36-5. Available from: http://CRAN.R-project.org/package=survival. [Accessed January 8, 2014].
[30] Koenker R. quantreg: Quantile Regression. 2015.
[31] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press; 1994.