Bioinformatics, 37(23), 2021, 4451–4459
doi: 10.1093/bioinformatics/btab495
Advance Access Publication Date: 13 July 2021
Original Paper

OXFORD

Genetics and population analysis

On the stability of log-rank test under labeling errors

Ben Galili [1,*], **Anat Samohi**[2,*] **and Zohar Yakhini**[1,2,*]

[1]Faculty of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel and [2]Arazi School of Computer Science, Interdisciplinary Center, Herzliya, Israel

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on January 21, 2021; revised on June 25, 2021; editorial decision on June 30, 2021; accepted on July 2, 2021

Abstract

Motivation: Log-rank test is a widely used test that serves to assess the statistical significance of observed differences in survival, when comparing two or more groups. The log-rank test is based on several assumptions that support the validity of the calculations. It is naturally assumed, implicitly, that no errors occur in the labeling of the samples. That is, the mapping between samples and groups is perfectly correct. In this work, we investigate how test results may be affected when considering some errors in the original labeling.

Results: We introduce and define the uncertainty that arises from labeling errors in log-rank test. In order to deal with this uncertainty, we develop a novel algorithm for efficiently calculating a stability interval around the original log-rank *P*-value and prove its correctness. We demonstrate our algorithm on several datasets.

Availability and implementation: We provide a Python implementation, called LoRSI, for calculating the stability interval using our algorithm https://github.com/YakhiniGroup/LoRSI.

Contact: benga9@gmail.com or anatsamohi@gmail.com or zohar.yakhini@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The comparison of different treatments or, more generally, policies or protocols, in terms of survival rates or in terms of success rates is a central aspect of investigating these regimes and of taking related decisions. There are two approaches that are generally taken in analyzing survival data. The first uses a permutational null distribution (Heimann and Neuhaus, 1998; Vandin *et al.*, 2015) and is more appropriate for imbalanced data. The second, more popular approach, uses a conditional null model, based on the hypergeometric distribution. This second approach is also the focus of this article. The log-rank test was introduced by Mantel (1966) and is extensively used since then. It is a standard tool in survival analysis, e.g. Kleinbaum and Klein (2012). In Tourneau *et al.* (2015), reporting on the SHIVA study, the log-rank test was used to determine whether the use of several targeted therapies outside their intended indications will improve progression-free survival in cancer. In Pitt *et al.* (1999), the authors used log-rank test to conclude that the use of Spironolactone is effective to lower the risk of death in patients who suffered from severe heart failure. Galili *et al.* (2021) investigate efficient gene signatures that characterize a breast cancer subtype related to the patient's immune response. The signature is optimized using a survival criterion based on the log-rank test. Levy-Jurgenson *et al.* (2020) report how cancer intratumor heterogeneity can affect patient survival.

When applying the log-rank test to a set of data, we are implicitly assuming that the association of a subject, or, more generally, a sample, to one of the two labels, is not in doubt. In reality, however, this assumption is often compromised. In some cases, the label assignment is, indeed, rather straight forward. This is typically the case in the assignment to treatment arms. In other situations, it may be much less well defined.

This is the case, as a first example, when label assignment is determined by a human judgment, e.g. based on inspection by pathologists, which is often prone to errors. Literature explicitly reports inconsistency in pathology. Jackson *et al.* (2017) report a study that found that the decision of the same pathologist varied when examining the same samples in different times. They showed that two diagnostic calls of the same pathologist, separated by at least 9 months, on the same biopsy, have an agreement rate of 92% (95% CI 88–95%) for invasive breast cancer and even less for other breast cancer types. Jackson *et al.* (2017) also showed that for different pathologists testing the same biopsy the agreement rates dropped by additional 3–10%. Elmore *et al.* (2017) reported similar results. In a different context, any subjective scoring approach, such as the Eastern Cooperative Oncology Group (ECOG) score, as used in Loprinzi *et al.* (1994), depends, based on the same principles, on the individual making the calls.

As a second example consider labeling that follows a machine decision. Three recent studies (Ha *et al.*, 2019; Islam *et al.*, 2020; Jaber *et al.*, 2020) introduce machine-learning models to determine breast cancer subtypes. They all reported around 70% accuracy.

Finally, consider sample labeling, which is based on the results of some molecular measurement assay. Ebbert *et al.* (2011) showed how intrinsic errors in the laboratory process, specifically in gene

expression profiling, affect the final results. They test this on PAM50 results and reported around 5% error in the classification.

In the context of survival analysis, wrong sample labels can lead to dramatically different statistical assessments. Consider the MAINZ cohort, Schmidt *et al.* (2008), that describes survival data for breast cancer patients. As extensively reported, including in Fallahpour *et al.* (2017) and Howlader *et al.* (2018), Luminal A patients have better prognosis than the other types. This can be seen also in the MAINZ cohort, see the left panel in Figure 1. We note a significant difference in the Luminal A prognosis with *P*-value =0.014. Now, what will the effect be, on the resulting *P*-value, of changing one Luminal A label out of the 200 samples (0.5%) in the MAINZ cohort? Figure 2 shows the original data and the data after one label change. Figure 1 shows the Kaplan–Meier graphs, before (left) and after (right) the change, and the corresponding *P*-values. Examining Figure 1 shows a dramatic change in the *P*-value from 0.014 to 0.029, when it is not so simple to notice any change in the plots themselves. Expanding this observation to the actual labeling error, e.g. as reported in Ebbert *et al.* (2011) for breast cancer subtypes, can lead to even more dramatically changes in the *P*-value.

Previous investigations addressed several aspects of uncertainty in survival analysis. Heterogeneity between individuals is not taken into account in the basic form of log-rank test. To address this bias, Hougaard (1995) introduced the concept of frailty models for survival analysis. Under this approach, the null model does not assume that the distribution of time to event is the same for all subjects. In order to overcome the unobserved heterogeneity in the survival data the frailty models use random effect to create different time to event distributions.

Addressing a different issue, it is common to report (and plot) confidence intervals for each of the observed hazard ratios, resulting in a confidence envelope around the survival lines. In Vandin *et al.* (2015), the authors demonstrated that asymptotic approximation, as in log-rank test, can be misleading when the two groups under consideration have very different sizes. They introduced a novel approach to accurately calculate the log-rank *P*-value regardless of the group sizes. Splitting subjects to two groups, in order to determine an association between the split and, potentially, low risk and high risk, is an important task in the context of survival analysis. Standard studies use treatment types, protocols etc. When studying a quantitative potential determinant of survival, we are often interested in splitting according to that quantity. For example, the expression level of some gene or maybe BMI. Trying all possible cut-points (thresholds) is not practical due to multiple testing problems. In an important paper treating this issue, Hothorn and Lausen (2003) developed a method for calculating an upper bound on the log-rank test *P*-value, which efficiently takes into account multiple testing. This approach checks different label assignments, similar to our work, but limits the splits being considered. In addition, it focuses on finding the best split. In this work, we address general labeling and not necessarily such which is driven by a quantitative feature. Our study also considers completely general labeling changes and not ones related to consistent threshold splits.

In this work, we address, for the first time, the uncertainty that arises from general labeling errors and label instability. Given survival data with n samples and an error rate, α, we find the minimum and maximum log-rank *P*-values that can result from changing the labels of at most αn samples. These minimum and maximum

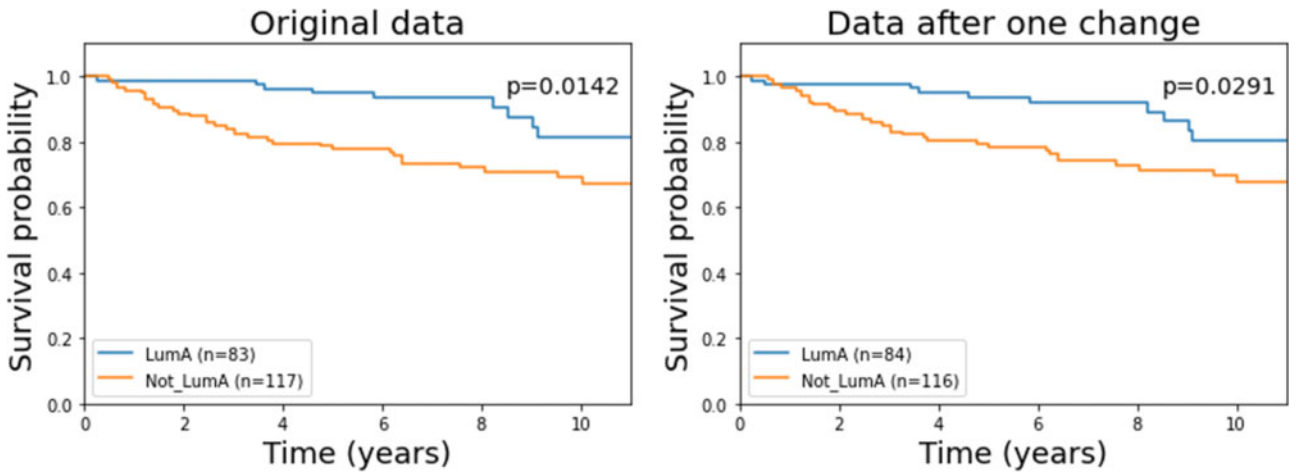

Fig. 1. Kaplan–Meier curve and log-rank *P*-value on the MAINZ cohort—Luminal A versus not Luminal A. On the left the original data and on the right the data after one change

Original data

Time (days)	30	30	90	180	...	6030	6090	6150	7200
Event	0	0	1	1	...	0	0	0	0
Group	LumA	Not	LumA	Not	...	Not	Not	Not	Not

Data after one change

Time (days)	30	30	90	180	...	6030	6090	6150	7200
Event	0	0	1	1	...	0	0	0	0
Group	LumA	Not	LumA	**LumA**	...	Not	Not	Not	Not

Fig. 2. MAINZ cohort data, before and after one change in the labels

P-values, P_L and P_U, define a stability interval $[P_L, P_U]$. To make our analysis less sensitive to extreme cases, we support the use of a confidence level, $1-\delta$, to further narrow our interval. The main contributions of our work are as follows:

- A definition of labeling errors stability intervals for statistical tests.
- A procedure that, given data and a bound on the labeling error, calculates a stability interval for the log-rank test.
- A software implementation of the above procedure.
- https://github.com/YakhiniGroup/LoRSI.
- Applications to several example cases.

2 Materials and methods

2.1 The statistical framework for log-rank stability

2.1.1 Preliminaries

We first set the notation for the log-rank test, in the context of the conditional null distribution (Mantel, 1966), as will be used in the rest of the manuscript.

- Consider time and event data D (survival data) with a partition labeling λ_0 [a binary vector mapping subjects into the groups A (0) & B (1)] over n subjects.
- Let $j = 1, \ldots, J$ be the distinct times of observed events in either group.
- Let $n_{A,j}$, $n_{B,j}$ be the number of subjects 'at risk' (who have not yet had an event nor have been censored) at the time of occurrence of the jth event in the two groups, respectively.
- Let $O_{A,j}$, $O_{B,j}$ be the random variables representing the observed number of events in each group at time j.
- Denote $n_j = n_{A,j} + n_{B,j}$, the number of at-risk subjects at time j.
- Denote $o_j = O_{A,j} + O_{B,j}$, the number of actual events observed at time j.
- Let T be the time of failure of a subject. $P(T = t)$ is the probability distribution function of T. The survival function is defined as $S(t) = 1 - P(T < t) = 1 - F(t)$.
- In log-rank testing, we are working under the null model that assumes that the two groups have identical survival functions, $S_A(t) \equiv S_B(t)$
- We then have

$$O_{A,j} \sim HG(n_j, n_{A,j}, o_j).$$

  Similar for group B (HG stands for Hypergeometric).
- The null model also assumes that the variables $O_{A,j}$ are (collectively) independent.
- The expected value and the variance of $O_{A,j}$ under the null model are:

$$E_{A,j} = \frac{n_{A,j}}{n_j} o_j$$
$$V_{A,j} = \frac{n_{A,j}}{n_j} o_j \left(\frac{n_j - o_j}{n_j}\right)\left(\frac{n_j - n_{A,j}}{n_j - 1}\right).$$

  Similar for group B.
- Putting everything together, for all $j = 1, \ldots, J$, the log-rank statistic compares $O_{A,j}$ to their expected values $E_{A,j}$ under the null model. The statistic is defined as:

$$Z_A = \frac{O - E}{\sqrt{V}},$$

where:

$$O = \sum_{j=1}^{J} O_{A,j} \quad E = \sum_{j=1}^{J} E_{A,j} \quad V = \sum_{j=1}^{J} V_{A,j}.$$

Similar for group B.

If J is sufficiently large and the partition into A and B is reasonably balanced (see e.g. Vandin *et al.*, 2015) then, Z is approximately distributed as $N(0, 1)$. This allows us to compute a P-value for the comparative survival data D, using the value actually observed for O, which we denote $o = o(D)$. This P-value is denoted by $LR(D, \lambda_0)$. By extension $LR(D, \lambda)$ will denote the log-rank P-value that would be obtained for any different partition labeling λ.

2.1.2 Definition of the log-rank stability interval

We now define a log-rank stability interval for given survival data and two parameters $\alpha > 0$ and $\delta \geq 0$.

- Again, consider time and event data D with a partition labeling λ_0 (mapping subjects into the groups A & B) over n subjects. Recall that $LR(D, \lambda_0)$ is the log-rank P-value computed for this data.
- Let $0 < \alpha < 1$. Given a different binary labeling λ, we say that λ is an α-modification of λ_0 if the labels have changed in less than a fraction α of the samples.
- Formally, $H(\lambda, \lambda_0) \leq \alpha \cdot n$, where H is the Hamming distance.
- Let $B(\lambda_0, \alpha)$ be the set of all possible labeling partitions λ that are α-modifications of λ_0.
- Let $0 \leq \delta < 1$. We want to compute a tight interval $[p_L, p_U]$ in the following sense: p_U should be the smallest number for which $LR(D, \lambda) \in [p_L, p_U]$ holds for a $1 - \delta$ fraction of $\lambda \in B(\lambda_0, \alpha)$. Note that under this definition, we require tightness on the right hand side, which is, in practice, taking a conservative approach, the more interesting case (see Section 4).
- The interval defined above is the stability interval for the two parameters $\alpha > 0$ and $\delta > 0$ and the input data. We write:

$$SI(D, \lambda_0, \alpha, \delta) = [p_L, p_U]. \quad (1)$$

For example, given data D with $n = 100$ (100 samples), $\delta = 0.05$ and $\alpha = 0.01$, we want to compute an interval SI so that for 95% of the single label changes (1% change) the log-rank P-value will fall in SI.

2.2 Computing stability intervals for log-rank test

In this section, we describe our algorithmic approach and prove its correctness.

2.2.1 Algorithm: LoRSI

We start by some definitions and notations.

Let:

- D– the dataset. Consisting of three vectors of length n:
  a. e: event/censored descriptor. Indicates whether event or censored occurred.
  b. t: time. The time from the beginning of an observation period to an event, censored or end of the study.
  c. l: group. Indicates the group of the subject.
- $d = (l(d), t(d), e(d))$ represents a single instance: an instance $d \in D$ is defined by three quantities—the group label $l(d)$, the time $t(d)$ and the event/censored descriptor $e(d)$.
- F : the set of interest (the Focus set).
- B : the other set (the Background set).
- The set F is typically the one that has a better survival rate. That is: $Z_F < Z_B$. In this article, we also take this approach and

therefore F is the set that has the better survival rate in the input data, D.

We further define the following subsets of F and B:

- EF: the events of the group F, ordered from the earliest to the latest.
- CF: the censored samples of the group F, ordered from the earliest to the latest.
- EB: the events of the group B, ordered from the earliest to the latest.
- CB: the censored samples of the group B, ordered from the earliest to the latest.

Note that $F = EF \cup CF$ and $B = EB \cup CB$.

Now define the prefixes and sufixes of these ordered subsets as follows:

- $EF_L(i) =$ the samples $EF(1), \ldots, EF(i)$
- $EB_L(i) =$ the samples $EB(|EB| - i + 1), \ldots, EB(|EB|)$
- $CB_L(i) =$ the samples $CB(|CB| - i + 1), \ldots, CB(|CB|)$
- $EF_U(i) =$ the samples $EF(|EF| - i + 1), \ldots, EF(|EF|)$
- $EB_U(i) =$ the samples $EB(1), \ldots, EB(i)$
- $CF_U(i) =$ the samples $CF(|CF| - i + 1), \ldots, CF(|CF|)$.

Following standard notation for the set of types of denominator k over a three letter alphabet (Cover, 1999), we denote:

$$T(k,3) = \{(i_1, i_2, i_3) : i_1 + i_2 + i_3 = k\}.$$

Note that:

$$|T(k,3)| = \binom{k+2}{2}.$$

Definition 1. The set of P_U candidates is defined by:

$$C_U = \{(EF_U(i_1) \cup EB_U(i_2) \cup CF_U(i_3)) : (i_1, i_2, i_3) \in T(k,3)\}.$$

We will show that this is the collection of candidate sample sets of size k, amongst which we will identify the set of samples that, if swapped, will lead to the most extreme positive change in the P-value. Note that the size of C_U is the same as that of $T(k,3)$, namely $\binom{k+2}{2}$.

Similarly:

Definition 2. The set of P_L candidates is defined by:

$$C_L = \{(EF_L(i_1) \cup EB_L(i_2) \cup CB_L(i_3)) : (i_1, i_2, i_3) \in T(k,3)\}.$$

Algorithm 1 describes the Log-Rank Stability Interval Algorithm (LoRSI) for finding P_L and P_U, where $\alpha = \frac{k}{n}$. For P_U, the idea of the algorithm is to iterate over the set of all relevant sets of k changes. The size of this collection is relatively small, namely $\binom{k+2}{2}$, due to the monotonicity effect on the z-score in each one of the groups E_F, C_F & E_B as proven below in Section 2.2.2. In each iteration, our procedure calculates the P-value after changing the labels of the current k subjects. Finally, it selects the max P-value among the $\binom{k+2}{2}$ candidates. Note that, if we consider only one label change ($k = 1$), then, the SI (both sides) is determined by only six candidates, three for P_U and three for P_L (see Fig. 3). In the Supplementary Material, we describe the LoRSI algorithm, where $\alpha = \frac{1}{n}$ and for any $\delta > 0$.

Algorithm 1: LoRSI pseoducode.

Log-Rank Stability Interval
input: Dataset—D, $\alpha = \frac{k}{n}$
output: Stability Interval $[p_L, p_U]$
p_L_candidates $= \varnothing$
p_U_candidates $= \varnothing$
//each of these sets will hold all $\binom{k+2}{2}$ relevant P-values
for *current_set_of_changes in* C_U **do**
  //see Definition 1 for C_U
  p = log-rank P-value after swapping the labels of all the k samples in *current_set_of_changes*
  p_U_candidates.append (p)
end
for *current_set_of_changes in* C_L **do**
  //see Definition 2 for C_L
  p = log-rank P-value after swapping the labels of all the k samples in *current_set_of_changes*
  p_L_candidates.append (p)
end
$p_U = \max(p_U\text{\_candidates})$
$p_L = \min(p_L\text{\_candidates})$
return p_L, p_U

2.2.2 Correctness

In this section, we prove the correctness of Algorithm 1. This, in essence, is the content of Theorem 1, stated at the end of this section. We start with some definitions and notations.

- Let z_0 be the original Z-statistic obtained from the input labeling.
- Now consider a labeling swap for the instance d. That is, if in λ_0, the instance d is in the group F, then, it is swapped to B and symmetrically otherwise. This swap will affect the value of Z calculated for the new data. Let

$$z_{new}(D,d) = \frac{o_{new} - E_{new}}{\sqrt{V_{new}}},$$

  where o_{new}, E_{new} and V_{new} are obtained for the swapped data as described in the preliminaries.
- We are specifically interested in the resulting change in the observed value of Z, which we denote

$$\Delta z(D,d) = z_{new}(D,d) - z_0.$$

Let $(Y_1, \ldots, Y_n)$ be a set of RVs. We say that $(Y_1, \ldots, Y_n)$ is an independent hypergeometric set (IHS) if:

1. $Y_1, \ldots, Y_n$ are (collectively) independent.
2. $\forall i\ Y_i \sim HG(N_i, B_i, n_i)$.
3. $\forall i\ Var(Y_i) > 0$.

For a single RV X let

$$Z(X) = \frac{X - E(X)}{\sqrt{V(X)}}$$

and then, for a set of independent RVs,

$$Z(X_1, \ldots, X_n) = Z(\sum_{i=1}^{n} X_i).$$

We note that if $(Y_1, \ldots, Y_n)$ is an IHS and if n is sufficiently large then the distribution of $Z(Y_1, \ldots, Y_n)$ is approximately standard

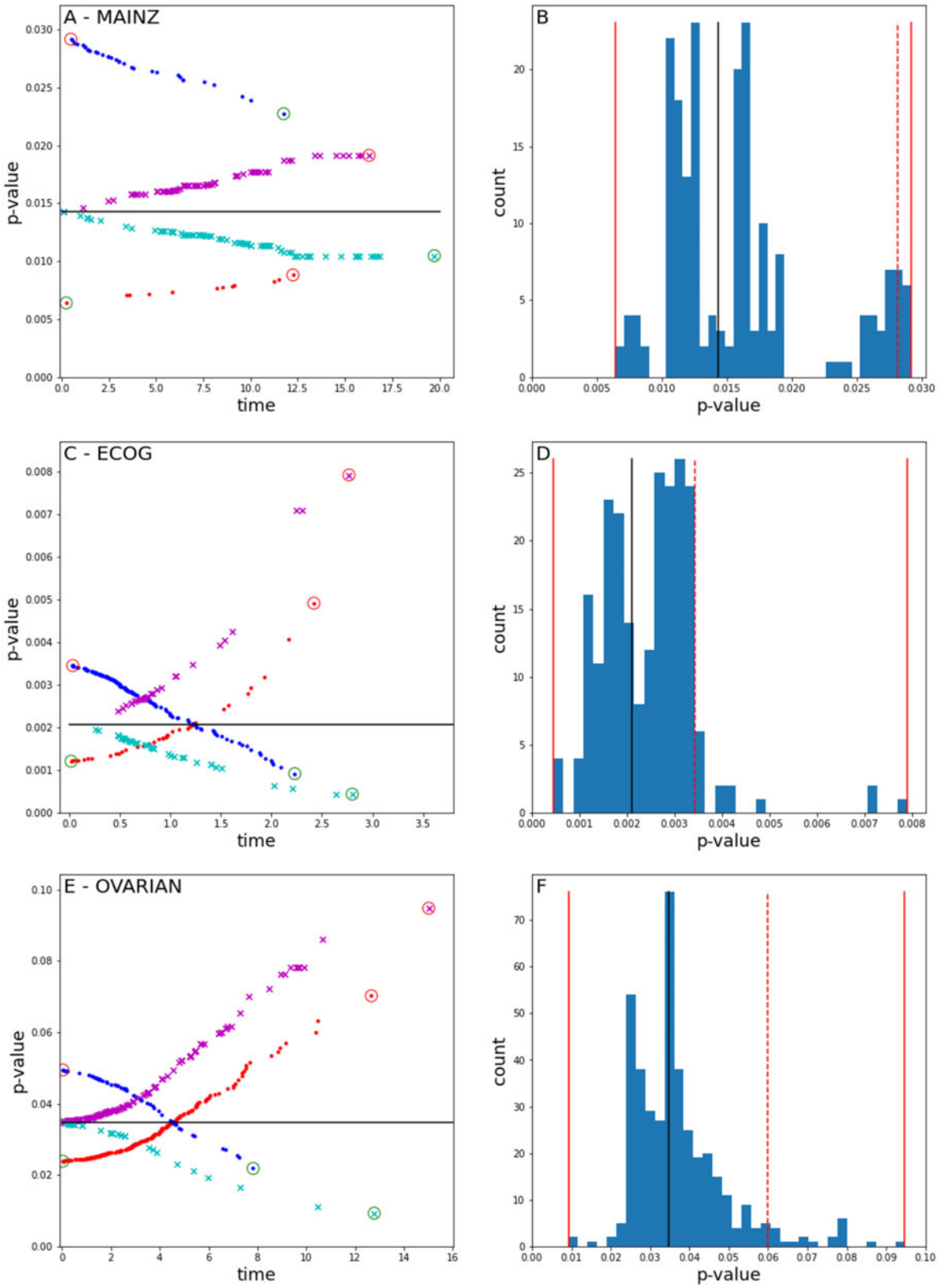

Fig. 3. Log-rank stability analysis for single label changes. The figure represents results for three datasets, as described in the text. Each of the depicted datasets consists of two groups (**F**)—the (actual, as per the original data) group of patients with good prognosis, and (**B**)—the bad prognosis group. Each data point is either an event or a censored point. The combination of the group and the event type leads to four categories of patients. The scatter plots provide a visual representation of the effect, on the *P*-value, that follows from changing the label of a single sample (i.e. $\alpha = \frac{1}{n}$). We can observe the monotonicity of the effect, with a direction depending on the category, as proven in Section 2. For each dataset, we indicate the original (non-swap) *P*-value (solid black lines on the right panel), the numbers P_L and P_U (solid red lines) for δ=0 and the number P_U for $\delta = 0.05$ (dashed red lines). Sample categories, in the scatter plots are represented by shape and color: blue dots—event swap from (B) to (F), red dots—event swap from (F) to (B), cyan Xs—censored sample swap from (B) to (F) and purple Xs—censored sample swap from (F) to (B). The green and red circles represent P_L and P_U candidates, respectively, at $\delta = 0$

normal (Lindeberg, 1922). We also note that, as stated above, the variables $O_{A,j}$, where $1 \leq j \leq J$, constitute an IHS.

For an observation x, derived from an RV X, we further define the Z-transformed value:

$$z(x) = \frac{x - E(X)}{\sqrt{V(X)}}.$$

For a set observation we now define

$$z(X_1 = x_1, \ldots, X_n = x_n) = z(\sum_{i=1}^{n} x_i).$$

For a random variable X and a number $x \in \mathbb{R}$, we use the notation $CDF(X, x)$ to represent the cumulative distribution of X at x. Or, in other words: $CDF(X, x) = P(X \leq x)$.

Claim 1. *Given two RVs X, Y where:*

$$X \sim HG(N, B, n), Y \sim HG(N, B, m)$$

and

$$n < m$$

then

$$\forall b \leq n,\ CDF(Y, b) < CDF(X, b).$$

See Supplementary Material for a proof.

Claim 2. *Consider two RVs X, Y where:*

$$X \sim HG(N, B, n), Y \sim HG(N, B, m).$$

Let $T_1, \ldots, T_k$ be k RVs where both $(X, T_1, \ldots, T_k)$ and $(Y, T_1, \ldots, T_k)$ are IHS. Denote $\mu_i = E(T_i)$ and $\sigma_i = \sqrt{V(T_i)}$.

Let:

$$Z_1 = Z(Y, T_1, \ldots, T_k), z_1 = z(Y = a, T_1 = t_1, \ldots, T_k = t_k)$$
$$Z_2 = Z(X, T_1, \ldots, T_k), z_2 = z(X = a, T_1 = t_1, \ldots, T_k = t_k)$$

If

$$n < m$$

then:

1. $CDF(Z_1, z_1) < CDF(Z_2, z_2)$
2. *For sufficiently large values of k* (which *is the interesting case, in the context of log-rank, see comment after the proof), we also have:*

$$z_1 < z_2.$$

See Supplementary Material for a proof.

As noted above, we are interested in working with large values of k in the context of log-rank. Without this assumption, the second part of Claim 2 is not necessarily true. For example, for $k = 1$, consider:

- $T_1 \sim HG(90, 2, 1)$ with observed value $t_1 = 1$
- $X \sim HG(100, 1, 50)$ with observed value $a = 1$
- $Y \sim HG(100, 1, 99)$ with observed value $a = 1$,

which yields: $z_1 = 5.554$, $z_2 = 2.835$.

Claim 3. *Let d_{j_1} and d_{j_2} be censored samples in D from group F. Let $z_1 = z_{new}(D, d_{j1})$ and $z_2 = z_{new}(D, d_{j2})$. If $time(d_{j_1}) < time(d_{j_2})$ then $z_0 < z_1 < z_2$, and therefore:*

$$0 < \Delta z(D, d_{j1}) < \Delta z(D, d_{j2}).$$

Similarly, if the censored samples, d_{j_1} and d_{j_2}, come from group B then if $time(d_{j_1}) < time(d_{j_2})$ then $z_0 > z_1 > z_2$, and therefore:

$$0 > \Delta z(D, d_{j1}) > \Delta z(D, d_{j2}).$$

Proof. We use the fact that the random variables $O_{F,j}$ as defined in the log-rank setup constitute an IHS. Swapping d_{j1} from F to B leads to a change in the at risk numbers $n_{F,j}\ \forall j \leq j_1$. More specifically each one of them is decreased by 1. Nothing changes for the later indices. Assuming that J is sufficiently large, we now iteratively use Claim 2. In every iteration, we decrease $n_{F,j}$ by 1, starting at $j = 1$ and ending at $j = j1$. At every index j let $O_{F,j}$ and $\tilde{O}_{F,j}$ be the hypergeometric variables representing the number of events at time j before and after a hypothetical swap at j, respectively. Claim 2 therefore applies, at every iteration j, with $O_{F,j}$ and $\tilde{O}_{F,j}$ playing the role of Y and X, respectively, and $\tilde{O}_{F,i}$ with $1 \leq i \leq j-1$ and $O_{F,i}$ with $j+1 \leq i \leq J$ playing the role of the T_is. We, thus, get $z_0 < z_1$.

Similarly, since $j_1 < j_2$ the swap of d_{j_2} will affect all at risk numbers above as well as several others $n_{F,j}\ s.t\ j_1 < j \leq j_2$. Continuing the above iterations, we therefore have $z_1 < z_2$.

When swapping away from group B, the effect of the swap will be to increase the at risk numbers, leading to the reverse inequalities. ■

Claim 4. *Consider the RVs X_1, X_2 and Y_1, Y_2 where:*

$$X_1 \sim HG(N, B, l), Y_1 \sim HG(N, B, l)$$
$$X_2 \sim HG(M, C, n), Y_2 \sim HG(M, C, m).$$

Let $T_1, \ldots, T_{k-1}$ be k −1 RVs where both $(X_1, X_2, T_1, \ldots, T_{k-1})$ and $(Y_1, Y_2, T_1, \ldots, T_{k-1})$ are IHS. Denote $\mu_i = E(T_i)$ and $\sigma_i = \sqrt{V(T_i)}$.

Let:

$$z_1 = z(Y_1 = a_1 - 1, Y_2 = a_2, T_1 = t_1, \ldots, T_{k-1} = t_{k-1})$$
$$z_2 = z(X_1 = a_1, X_2 = a_2 - 1, T_1 = t_1, \ldots, T_{k-1} = t_{k-1}).$$

If

$$n < m$$

then

$$z_1 < z_2$$

Proof. First, we write explicitly:

$$Z_1 = \frac{Y_1 - \mu_{Y_1} + Y_2 - \mu_{Y_2} + T_1 - \mu_1 + \cdots + T_{k-1} - \mu_{k-1}}{\sqrt{\sigma_{Y_1}^2 + \sigma_{Y_2}^2 + \sigma_1^2 + \cdots + \sigma_{k-1}^2}}$$
$$Z_2 = \frac{X_1 - \mu_{X_1} + X_2 - \mu_{X_2} + T_1 - \mu_1 + \cdots + T_{k-1} - \mu_{k-1}}{\sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2 + \sigma_1^2 + \cdots + \sigma_{k-1}^2}}$$
$$z_1 = \frac{a_1 - 1 - \mu_{Y_1} + a_2 - \mu_{Y_2} + t_1 - \mu_1 + \cdots + t_{k-1} - \mu_{k-1}}{\sqrt{\sigma_{Y_1}^2 + \sigma_{Y_2}^2 + \sigma_1^2 + \cdots + \sigma_{k-1}^2}}$$
$$z_2 = \frac{a_1 - \mu_{X_1} + a_2 - 1 - \mu_{X_2} + t_1 - \mu_1 + \cdots + t_{k-1} - \mu_{k-1}}{\sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2 + \sigma_1^2 + \cdots + \sigma_{k-1}^2}}.$$

By definition $\mu_{Y_1} = \mu_{X_1}$ and $\sigma_{Y_1} = \sigma_{X_1}$. Let $b_1 = a_1 - 1$. We rearrange the term to get:

$$z_1 = \frac{b_1 - \mu_{Y_1} + a_2 - \mu_{Y_2} + t_1 - \mu_1 + \cdots + t_{k-1} - \mu_{k-1}}{\sqrt{\sigma_{Y_1}^2 + \sigma_{Y_2}^2 + \sigma_1^2 + \cdots + \sigma_{k-1}^2}}$$
$$z_2 = \frac{b_1 - \mu_{Y_1} + a_2 - \mu_{X_2} + t_1 - \mu_1 + \cdots + t_{k-1} - \mu_{k-1}}{\sqrt{\sigma_{Y_1}^2 + \sigma_{X_2}^2 + \sigma_1^2 + \cdots + \sigma_{k-1}^2}}.$$

We now apply Claim 2, with: $b_1 = t_k$, $\mu_{Y_1} = \mu_k$, $\sigma_{Y_1} = \sigma_k$, $a_2 = a$, $\mu_{Y_2} = \mu_Y$, $\sigma_{Y_1} = \sigma_Y$, $\mu_{X_2} = \mu_X$, $\sigma_{X_2} = \sigma_X$ and get $z_1 < z_2$. ■

Claim 5. *Let d_{j_1} and d_{j_2} be events in D from group F.*
Let $z_1 = z_{new}(D, d_{j1})$ and $z_2 = z_{new}(D, d_{j2})$.
If $time(d_{j_1}) < time(d_{j_2})$ then $z_1 < z_2$, and therefore:

$$\Delta z(D, d_{j1}) < \Delta z(D, d_{j2}).$$

Similarly, if the events, d_{j_1} and d_{j_2}, come from group B then if $time(d_{j_1}) < time(d_{j_2})$ then $z_1 > z_2$, and therefore:

$$\Delta z(D, d_{j1}) > \Delta z(D, d_{j2}).$$

Proof. We once again use the fact that the random variables $O_{F,j}$ as defined in the log-rank setup constitute an IHS. Swapping d_{j1} from F to B leads to a change in the at risk numbers $n_{F,j}\, \forall j \leq j_1$. More specifically each one of them is decreased by 1. Similarly, since $time(d_{j_1}) < time(d_{j_2})$ the swap of d_{j_2} will affect all at risk numbers above as well as several others, namely $n_{F,j}\, s.t\, j_1 < j \leq j_2$. In addition, o_{F,j_1} becomes $o_{F,j_1} - 1$ when swapping d_{j1} and o_{F,j_2} becomes $o_{F,j_2} - 1$ when swapping d_{j2}. Therefore, $o_{Fnew} = o_F - 1$ in both swaps.

Now, let Y_1 and Y_2 be O_{F,j_1} and O_{F,j_2} after swapping d_{j1} from F to B, respectively. In addition, let X_1 and X_2 be O_{F,j_1} and O_{F,j_2} after swapping d_{j2} from F to B, respectively. By iteratively using Claim 4 and assuming that J is sufficiently large, we get $z_1 < z_2$.

In the case of swapping away from group B, the effect of the swap will be to increase both the at risk numbers and the observed o_{Fnew}, and therefore we get the reverse inequalities. ■

In summary, we proved that changing one sample will lead to a monotonic effect on the z-score and therefore on the log-rank P-value.

Now consider the case of k swaps. We claim that the k instances that yield the most extreme positive change in the P-value is one of the candidates in C_U. Let λ^* be the labeling that, indeed, yields the largest $LR(D, \lambda)$ within $B(\lambda_0, \frac{k}{n})$. To see why the above claim holds assume, WLOG, that λ^* swaps some instance $EB_U(j)$ but does not swap $EB_U(i)$ for some $i<j$. By the monotonicity proven above (Claim 5), we can swap $EB_U(i)$ instead of $EB_U(j)$ and get a larger Δz. A similar argument holds for an assumed usage, by λ^*, of a non-continuous suffix of EF and CF, respectively. Furthermore, a similar argument holds for the left side of the interval.

We conclude that:

Theorem 1. *For any k (counting label swaps in a data D), $max\{LR(D, \lambda) : \lambda \in B(\lambda_0, \frac{k}{n})\}$ is attained by swapping the labels in one of the sets listed in C_U and therefore determined by a triplet $(i_1, i_2, i_3) \in T(k, 3)$. Similarly, $min\{LR(D, \lambda) : \lambda \in B(\lambda_0, \frac{k}{n})\}$ is attained by a set in C_L and therefore also determined by some (other) triplet $(i_1, i_2, i_3) \in T(k, 3)$.*

3 Results

We now demonstrate the calculation of stability intervals on three different datasets (see Kaplan-Meier curves in Figs 1, 4 and 5). In calculating the interval, we use our efficient LoRSI algorithm, which is considering only a small set of relevant swaps, depending on the value of α, as described above. To provide a more complete information on how labeling errors can affect a given dataset, we also present the full P-value distribution. In order to do this, we calculate the log-rank P-value for all possible swaps according to α. The P-value distributions for $\alpha = \frac{1}{n}$, pertaining to the $n + 1$ swaps (including the non-swap original data) are depicted in Figure 3B, D and F. In addition, we present, in Figure 3A, C and E, for each dataset, the P-value as a function of the time and type of the swapped sample. For the first dataset, we also calculated the interval for 2 and 3 changes (Fig. 6). It should be noted that the calculation of the full distribution introduces a prohibitive time complexity. A more detailed comparison to our efficient approach is given below.

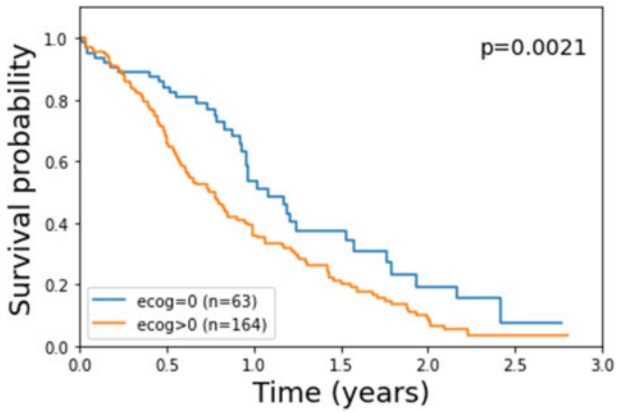

Fig. 4. Kaplan Meier curve and log-rank p-value according to ECOG score of patients with advanced colorectal or lung cancer - ECOG=0 Vs ECOG > 0

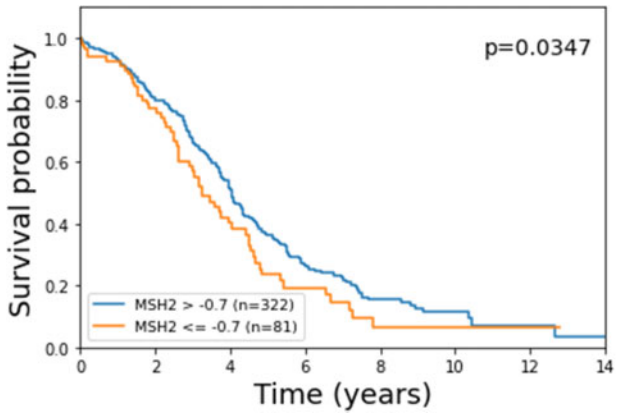

Fig. 5. Kaplan Meier curve and log-rank p-value according to expression of the gene MSH2 in ovarian cancer patients - MSH2 expression >-0.7 Vs MSH2 expression ≤-0.7

The first dataset is the MAINZ cohort (Schmidt *et al.*, 2008). We divided the data according to the subtype—Luminal A versus not Luminal A. It is well known that the breast cancer subtype Luminal A has better prognosis than the rest of the subgroups (Fallahpour *et al.*, 2017; Howlader *et al.*, 2018). As expected and as stated in the introduction, a log-rank test demonstrates this difference with P-value = 0.014 (see the left panel in Fig. 1). The stability interval calculated given $\alpha = \frac{1}{n}$ & $\delta = 0$ is [0.006, 0.029], see Figure 3A and B. This interval represents a 57% and 107% decrease/increase from the original P-value, respectively. We note that $\alpha = \frac{1}{n}$ in this dataset is only 0.5%. Using $\delta = 0.05$, the effect is still dramatic: a 101% increase to the inferred maximum P-value (SI = [0.006, 0.0282]). We further investigate the effect of $\alpha = \frac{2}{n}$ and $\alpha = \frac{3}{n}$. The stability interval calculated for $\alpha = \frac{2}{n}$ is [0.0029, 0.055] and when using $\delta = 0.05$, we got $P_U = 0.034$. The stability interval calculated for $\alpha = \frac{3}{n}$ is [0.0013, 0.095] and when using $\delta = 0.05$ we got $P_U = 0.043$. Here, again, we calculated the full P-value distribution to provide the complete information (see Fig. 6), a time consuming process. In order to find the stability interval, using the full P-value distribution for $\alpha = \frac{k}{n}$, one needs to perform $\sum_{i=1}^{k} \binom{n}{i}$ log-rank calculations. Our LoRSI algorithm needs only $2\binom{k+2}{2}$ such calculations, as described in Section 2. It took 2.5 min to calculate the SI using the full P-value distribution, for $\alpha = \frac{2}{n}$, where LoRSI took 0.5 s. For $\alpha = \frac{3}{n}$, the gap is much larger: almost 3 h to calculate the SI using the full P-value distribution and only 0.75 s for LoRSI.

Furthermore, setting $\alpha = 0.04$, which represents the error rate according to Ebbert *et al.* (2011), we need to investigate $k = 8$ changes, which is totally impractical. LoRSI will take seconds to do the SI calculation.

The second dataset came from a study that was developed to compare descriptive information from a patient-completed questionnaire to that obtained by the patient's physician Loprinzi *et al.* (1994). All the patients suffered from advanced colorectal or lung cancer. We consider the ECOG score calculated by a physician that assess the patients as a label for assessing survival differences. The 0 score represents fully active, able to carry on all pre-disease activities without restriction. Higher ECOG means less ability to perform usual daily activities, where 5 is the highest score. We divided the data according to ECOG $= 0$ and ECOG > 0. The statistical difference in survival between the groups is significant with log-rank P-value $= 0.0021$ (Fig. 4). The stability interval calculated given $\alpha = \frac{1}{n}$ & $\delta = 0$ is [0.0004, 0.0079], see in Figure 3C and D. This interval represents a 81% and 276% decrease/increase from the original P-value, respectively.

The third and last dataset comes from an investigation of the gene expression in ovarian cancer patients, using the TCGA data (Network *et al.*, 2011). The MSH2 gene was shown to be associated with survival in ovarian cancer (Borcherding *et al.*, 2018). We took the relevant part of the TCGA dataset and split the samples into two groups according the optimal cutoff suggested by Borcherding *et al.* (2018). This cutoff is (standardized) MSH2 expression >-0.7 and it yields a log-rank P-value of 0.0347 (Fig. 5). The resulting SI at $\alpha = \frac{1}{n}$ is [0.009275, 0.0947], see in Figure 3E and F. Here, the SI represents a 73% decrease from the original P-value to the minimum P-value and a 273% increase to the maximum P-value. This P_U results from swapping only one single sample (the latest censored sample in C_F), which is 0.24% of the samples in the cohort. Moreover, increasing δ to 0.05 will change P_U to 0.06, still a dramatic effect. To obtain $P_U < 0.05$, we need to set δ to 0.1. The meaning of this result is that 10% of the single sample labeling swaps, applied to a dataset that originally had a significant survival signal, result in a non-significant P-value.

4 Discussion

In this work, we introduce the novel concept of stability interval for log-rank test. This interval represents the possible effects of perturbing the labels from the original survival analysis data. We show that even a small error rate in the labels can lead to dramatically different statistical conclusions. Our calculated stability interval bounds these differences, thus allowing an assessment of the stability of the statistical test, under labeling errors. We focus on the definition of the stability interval for log-rank and develop an algorithm for efficiently calculating the interval for any α.

We present a deterministic approach for addressing the labeling error issue, where we consider all possible label swaps that affect different sample sets representing exactly α fraction of the samples. One can also take a stochastic approach, wherein instances are generated, in which each sample label is swapped with probability α. The number of labels actually swapped will then have a Binom(n, α) distribution. Sampling sufficiently many instances, or analytically characterizing the resulting sample space, will lead to a new way of calculating the stability interval from the resulting P-value distribution. While in the deterministic approach, we (in effect) assume a uniform error distribution, in this stochastic approach we can, theoretically, use any error distribution. This includes, e.g. models that would assign confidence to individual labels, making swaps less or more likely for individual subjects in the cohort. The study of this interesting and potentially useful extension is a topic for future research. Our approach is also extended to address a confidence parameter δ. Specifically, we find the smallest number p_U for which $LR(D, \lambda) \in [p_L, p_U]$ holds for a $1 - \delta$ fraction of possible labeling changes $\lambda \in B(\lambda_0, \alpha)$. This represents a conservative approach to taking δ into account. Namely, one that focuses on the desired significance threshold, as may be determined, by the user, in the study design.

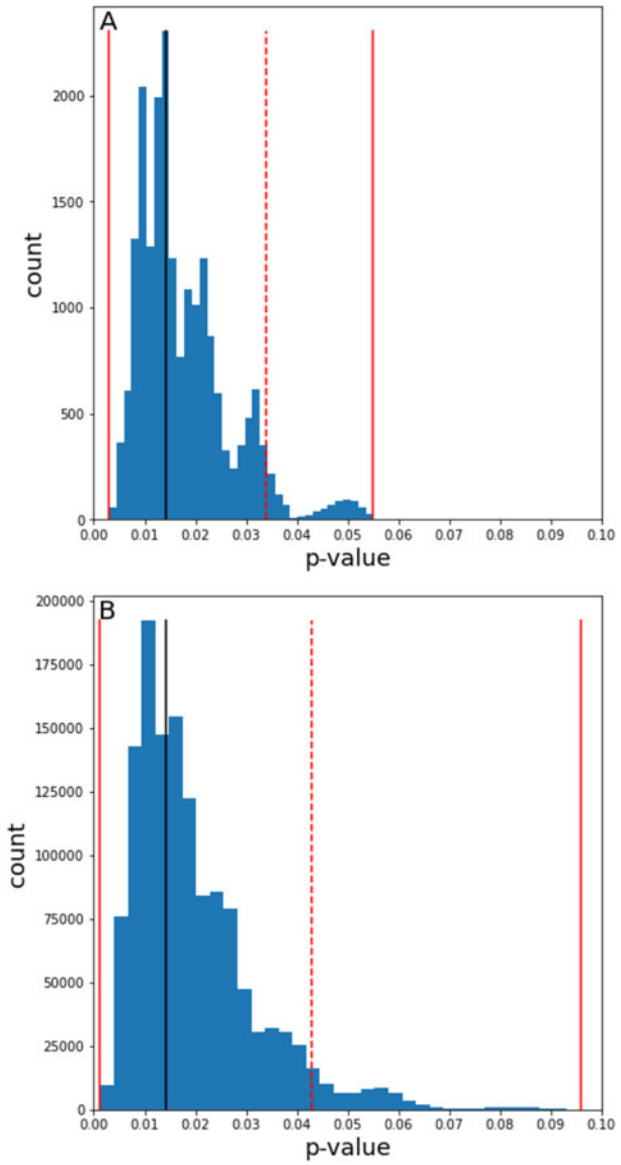

Fig. 6. p-value distribution on the MAINZ cohort (Luminal A versus not Luminal A), when $\alpha = \frac{2}{n}$, panel A, and when $\alpha = \frac{3}{n}$, panel B. The black line is the original p-value. The solid red lines are the P_L and P_U when $\delta = 0$. The dashed red line represents the P_U when $\delta = 0.05$

We note that in the proof of our algorithmic approach, we distinguish between working with the CDFs of sums of hypergeometric distributions and working with their standardized versions. Our result, pertaining to how the ends of the log-rank SI can be obtained by calculating the results of $2\binom{k+2}{2}$ sets of k swapped, holds for large Js as it requires a normal approximation. If differences in survival are directly assessed against the underlying sum of hypergeometric variables null model, then some of our results hold for any J.

We investigated the advantage of using our efficient LoRSI approach as compared to calculating the stability interval by generating the full P-value distribution. While LoRSI performs $\Theta(k^2)$ log-rank calculations to address k changes, the exhaustive approach takes $\Theta\binom{n}{k}$ such calculations. This complexity gap leads to seconds versus hours difference for small values of k and to LoRSI being the only practical approach in higher values.

We provide a Python implementation of the LoRSI algorithm. Current work focuses on the development of more efficient and user

friendly implementations of the methods described herein as well as on visualization tools. All will be made available through future releases. We hope that such efforts will make statistical stability analysis more accessible and useful for the community.

Acknowledgements

We thank the Technion Computer Science Department and the School of Computer Science at IDC Herzliya, for their support of the project. We thank the Yakhini Research Group for important discussions and input. We thank Xavier Tekpli for important discussion and insights.

Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under RESCUER, GA No. [847912].

Conflict of Interest: none declared.

References

Borcherding,N. *et al.* (2018) TRGAted: a web tool for survival analysis using protein data in the cancer genome atlas. *F1000Res.*, **7**, 1235.

Cover,T.M. (1999) *Elements of Information Theory*. John Wiley & Sons, NJ, USA.

Ebbert,M.T. *et al.* (2011) Characterization of uncertainty in the classification of multivariate assays: application to pam50 centroid-based genomic predictors for breast cancer treatment plans. *J. Clin. Bioinform.*, **1**, 37.

Elmore,J.G. *et al.* (2017) Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ*, **357**, j2813.

Fallahpour,S. *et al.* (2017) Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. *CMAJ Open*, **5**, E734–E739.

Galili,B. *et al.* (2021) Efficient gene expression signature for a breast cancer immuno-subtype. *PLoS One*, **16**, e0245215.

Ha,R. *et al.* (2019) Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm. *J. Digit. Imaging*, **32**, 276–282.

Heimann,G. and Neuhaus,G. (1998) Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. *Biometrics*, **54**, 168–184.

Hothorn,T. and Lausen,B. (2003) On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal.*, **43**, 121–137.

Hougaard,P. (1995) Frailty models for survival data. *Lifetime Data Anal.*, **1**, 255–273.

Howlader,N. *et al.* (2018) Differences in breast cancer survival by molecular subtypes in the united states. *Cancer Epidemiol. Biomarkers Prev.*, **27**, 619–626.

Islam,M.M. *et al.* (2020) An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Comput. Struct. Biotechnol. J.*, **18**, 2185–2199.

Jaber,M.I. *et al.* (2020) A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Res.*, **22**, 12.

Jackson,S.L. *et al.* (2017) Diagnostic reproducibility: what happens when the same pathologist interprets the same breast biopsy specimen at two points in time? *Ann. Surg. Oncol.*, **24**, 1234–1241.

Kleinbaum,D.G. and Klein,M. (2012) *Survival Analysis*. Springer, New York.

Levy-Jurgenson,A. *et al.* (2020) Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Sci. Rep.*, **10**, 18802.

Lindeberg,J.W. (1922) Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Math. Z.*, **15**, 211–225.

Loprinzi,C.L. *et al.* (1994) Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *J. Clin. Oncol.*, **12**, 601–607.

Mantel,N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, **50**, 163–170.

Network,C.G.A.R. *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609.

Pitt,B. *et al.* (1999) The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *N. Engl. J. Med.*, **341**, 709–717.

Schmidt,M. *et al.* (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.*, **68**, 5405–5413.

Tourneau,C.L. *et al.*; SHIVA investigators. (2015) Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol.*, **16**, 1324–1334.

Vandin,F. *et al.* (2015) Accurate computation of survival statistics in genome-wide studies. *PLoS Comput. Biol.*, **11**, e1004071.