Received: 10 September 2019 | Revised: 30 June 2020 | Accepted: 21 July 2020

DOI: 10.1002/bimj.201900273

Biometrical Journal

RESEARCH PAPER

Exact change point detection with improved power in small-sample binomial sequences

David Ellenberger[1] | **Berthold Lausen**[2] | **Tim Friede**[1]

[1] Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2] Department of Mathematical Sciences, University of Essex, Colchester, UK

Correspondence
David Ellenberger, Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
Email: david.ellenberger@outlook.de

Funding information
The Business and Local Government Data Research Centre, Grant/Award Number: ES/L011859/1

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

Abstract

To detect a change in the probability of a sequence of independent binomial random variables, a variety of asymptotic and exact testing procedures have been proposed. Whenever the sample size or the event rate is small, asymptotic approximations of maximally selected test statistics have been shown to be inaccurate. Although exact methods control the type I error rate, they can be overly conservative due to the discreteness of the test statistics in these situations. We extend approaches by Worsley and Halpern to develop a test that is less discrete to increase the power. Building on ideas from binary segmentation, the proposed test utilizes unused information in the binomial sequences to add a new ordering to test statistics that are of equal value. The exact distributions are derived under side conditions that arise in hypothetical segmentation steps and do not depend on the type of test statistic used (e.g., log likelihood ratio, cumulative sum, or Fisher's exact test). Using the proposed exact segmentation procedure, we construct a change point test and prove that it controls the type-I-error rate at any given nominal level. Furthermore, we prove that the new test is uniformly at least as powerful as Worsley's exact test. In a Monte Carlo simulation study, the gain in power can be remarkable, especially in scenarios with small sample size. Giving a clinical database example about pin site infections and an example assessing publication bias in neuropsychiatric drug research, we demonstrate the wide-ranging applicability of the test.

KEYWORDS
binary segmentation, change point, disorder detection, exact test, Worsley's test

1 | INTRODUCTION

The problem of change point detection (also sometimes referred to as threshold, cutpoint, breakpoint, or "disorder" detection) in binomial sequences has been addressed by many authors over the past decades (see, e.g., Carlstein, 1988; Chen & Gupta, 2011; Halpern, 1999; Hinkley & Hinkley, 1970; Lausen & Schumacher, 1992; Miller & Siegmund, 1982; Pettitt, 1979, 1980; Smith, 1975; Worsley, 1983). A distinction is usually made between the fixed sample change point problems and

sequential (online) schemes that are frequently used in *quality control*. Here, we consider the former, an ordered sequence of ζ independent binomial variables, with m_i being the number of events occurring in n_i subjects at risk. We are interested to test the null hypothesis H_0 of constant event probabilities p_i $(i = 1, \dots, \zeta)$ equal to p against the alternative

$$H_1 : p_i = \begin{cases} p, & (i = 1, \dots, \kappa) \\ p', & (i = \kappa{+}1, \dots, \zeta) \end{cases}, \qquad p \neq p'$$

for some period κ in the range $1, \dots, \zeta$ denoting an unknown change point. Most commonly tests to detect such a change point in the sequence are based on taking the maximum of a test statistic $T^k_{1:\zeta}$ that is designed to find differences for fixed candidate change points $k = 1, \dots, \zeta$. Those statistics include the log likelihood ratio, the cumulative sum and variations thereof (Pettitt, 1980), and the p-value of Fisher's exact test (Halpern, 1999); but also statistics based on Doob's martingale decomposition (Brostrom, 1997) as well as Bayesian statistics may be used (Assareh, Smith, & Mengersen, 2015; Smith, 1975). These maximally selected test statistics $T^{\max}_{1:\zeta} = \max_{k=1,\dots,\zeta}(T^k_{1:\zeta})$ arise not only in change point detection but also in various other applications. A real-world problem might be the precise assessment of the probability of an unfavorable realization in random ordering of (clusters of) binary events. Such an example might be the interest to detect manipulations of fixture lists in sports, which define the order of (weak or strong) opponents to be matched against. Another example may be the assessment of potential context effects in surveys. The context effect relates the order of questions asked to a bias in the overall thinking and answers of survey respondents. Applications of change point models in epidemiology and medicine are common and have led to ongoing methodological developments. Examples include the epidemic wave model (Boulesteix & Strobl, 2007; Siegmund, 1986), the assessment of genetic recombination (Halpern, 1999), dose-response models (Lausen, Lerche, & Schumacher, 2002), calendar time effects in clinical registries (Friede & Henderson, 2003), and clinical trials with adaptive designs (Friede & Henderson, 2009). Maximally selected test statistics are also used as *cutpoint methods* for dichotomization, although these may primarily be used when an underlying change can truly be regarded as abrupt. Otherwise these methods may lack statistical power and alter the effect estimates in comparison to continuous regression methods (see, e.g., Royston, Altman, & Sauerbrei, 2006) as long as the latter are correctly specified. Still, the simplicity of the considered change point model avoids instabilities in the parameter estimation in these scenarios.

Asymptotic distributions of maximally selected test statistics in binary sequences were derived by a number of authors (see Miller & Siegmund, 1982; Pettitt, 1979, 1980). For small sample sizes, however, these approximations have poor performance and exact methods are to be preferred (Friede, Henderson, & Kao, 2006; Halpern, 1982). Exact null and alternative distributions were given by Worsley (1983) for log likelihood ratio and cumulative sum test statistics. Halpern (1999) proposed to use Fisher's exact test and compared the different approaches with regard to their statistical power. Hirotsu (1997) gave exact distributions in case of two-way layouts with interaction effects. While these exact methods are designed for small sample sizes, they often lack size of the test in these scenarios. Due to the discreteness of the test statistic, the significance level of the test cannot be used to the full extent. This adds a *degree of conservativeness* to the test procedure (Ross, Tasoulis, & Adams, 2013). Several approaches have been discussed to overcome not only the implied loss in power, but also the lack of precision as a methodological disadvantage (Zhou, Zou, Zhang, & Wang, 2009). The trivial solution of a randomization of the test statistic is only of theoretical interest to achieve a uniformly most powerful test but cannot be recommended for practical application because of a lack of reproducibility among other reasons. The same applies to approaches that use Monte Carlo simulation techniques to obtain the required probabilities, see for example, Ross et al. (2013). An unconditional version of Worsley's test addressing the problem was suggested by Ellenberger and Friede (2016) to gain power with less discrete test statistics. Unlike Worsley's test, this test does not condition on the observed total number of events. The nuisance parameter p is dealt with by maximizing the p-value over the nuisance parameter. We aim at developing a hypothesis test that uses also information in the sequences left and right of each possible change point. We can thus define an ordering of different sequences that all yield the same Worsley's p-value. The ordering will be based on binary segmentation ideas and is used to create less discrete test statistics. To do this, we develop in Section 5 exact null distributions under certain side conditions. These are used to get exact p-values on both subsequences left and right of the potential change point $\hat{\kappa} = \text{argmax}_k(T_k)$ conditional on $T_{\hat{\kappa}}$. With these conditional p-values, we will define a new test in Section 6 by applying a combination function such that both p-values are merged to a single meaningful p-value. The performance of the proposed test is assessed by Monte Carlo simulations in Section 7 and the test is applied to two motivating examples introduced in Section 2. One example searches for change points in a clinical database of orthopedic surgeries using external fixators in children; the other example is concerned with the assessment of publication bias in neuropsychiatric drug research. We close with a brief discussion of the findings in Section 8.

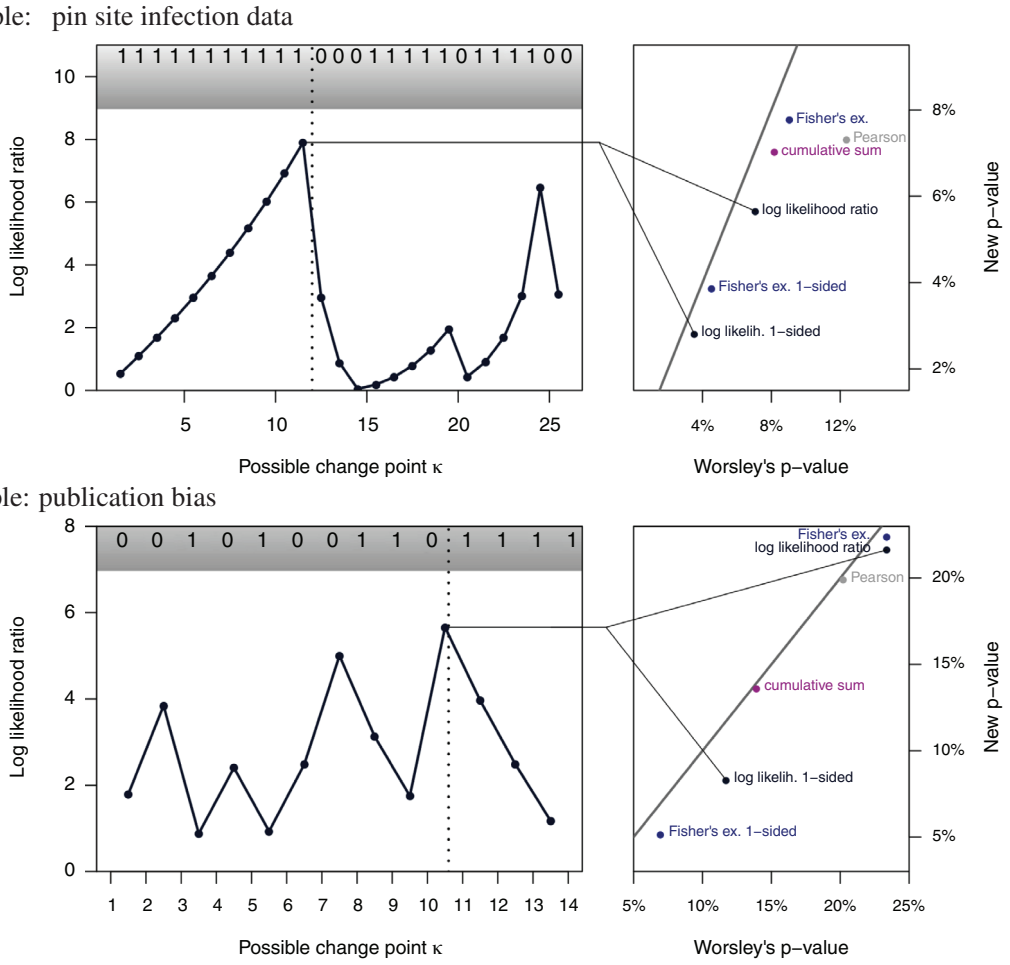

FIGURE 1 Sequence of events (1s and 0s at the top ordered by calendar time) and the according log likelihood ratio statistic (dark area showing the rejection area at a 5% level) for the pin site infection data (a) and the publication bias in neuropsychiatric drug research (b). To the right, the one- and two-sided *p*-values are displayed for both methods and various underlying statistics, indicated by the dots with *x*-axis giving Worsley's *p*-value and the *y*-axis giving the new *p*-value. The extent to which the new *p*-value is less coarse can be deduced from the vertical distance to the diagonal line

2 | MOTIVATING EXAMPLES

2.1 | Pediatric external fixators study

Some orthopedic surgeries require the attachment of external fixators. For this purpose, skeletal pins are cut through the skin. A common side effect of this treatment is the occurrence of infections, so-called pin site infections. The surgeries form a sequence in calendar time with pin site infection being regarded as an event. Analyses of the pin site infection records have previously been carried out in Friede et al. (2006) to investigate the effectiveness of the introduction of a new procedure for pin site care. Measures against pin site infections needed to be taken because those were frequent. The data suggest that the introduction of the new procedure is strongly associated with a decrease in infections. We are now interested whether this association holds for a subgroup of boys who had surgeries with external fixators to treat fractions at their feet. This subgroup of 26 pediatric patients with a total of 20 infections shows to be homogeneous in terms of reason of surgery and other characteristics. Since the covariates age and reason for application were found to be noninformative in previous analyses (Friede et al., 2006), these were not considered here. The binary sequence of pin site infection is shown as 1s and 0s in Panel A of Figure 1. The log likelihood ratio statistic is displayed for all possible change points k.

2.2 | Publication of FDA-approved neuropsychiatric drugs

Publication bias is increasingly recognized as a major problem in scientific publishing. Several authors have addressed the matter and investigated its extent. In clinical trials of neuropsychiatric drugs, Zou et al. (2018) have conducted an extensive literature search and analyzed trends in time. In the past decades, regulatory authorities have taken measures to prevent negative results from not being reported or published. The US Congress passed the FDA Modernization Act (FDAMA) in 1997, which mandated the public registry ClinicalTrials.gov that was established 3 years later. In 2005, the International Committee of Medical Journal Editors (ICMJE) enacted a policy requiring trial registration as a prerequisite for publication in member journals, leading to an increase in the number of registered studies. Zou et al. (2018) point out that FDAMA did not require registration of all studies, and the ICMJE recommendation continued to allow the publication of unregistered studies as compliance was voluntary. In 2007, the US Congress passed the FDA Amendments Act (FDAAA), which applies to all non-phase-I studies with FDA-regulated drugs and requires sponsors and investigators to register all such trials in ClinicalTrials.gov prior to enrolment and report the results to within 30 days post approval. Inappropriately delayed registration and reporting of results, as well as reporting of incorrect results, can be punished by fines and possible loss of funding. The FDAAA applies to trials initiated after September 27, 2007, and to earlier trials in progress as of December 26, 2007. Zou et al. (2018) have studied in detail the registration, results reporting, and publishing of clinical trials supporting FDA approval of neuropsychiatric drugs. They investigated the possible effects of the FDAAA on the publication of negative or unequivocal findings. Regarding this publication bias, the authors found statistically significant effects that the rate of publishing negative findings has increased from the pre-FDAA era to the post-FDAAA era. In the latter, all trials were published, though also some recent trials were found to report inconsistent results in comparison to the FDA approval assessment. In contrast to investigating only this comparison with the date of the change that is allegedly already known, it is also of interest to carry out analyses that consider all possible change points in the chronological order of the drug approvals. We therefore carried out the respective change point analyses on the data. We considered all drugs agents by different pharmaceutical companies that were approved by the FDA but had at least one trial that had either negative or questionable results in the FDA's reports. With data by Zou et al. (2018), we tested the outcome whether all trials of the approved drug were published in a scientific journal as it should be mandatory. These analyses will also investigate in the willingness of drug companies to publish older negative studies on an approved drug, whose retrospective reporting did not became mandatory by the FDAAA. Similarly, further calendar time effects in pivotal studies for the FDA approval of new drugs are the subject of ongoing investigations (Zhang et al., 2020).

3 | MODEL AND WORSLEY'S TEST

We consider the problem of investigating the existence of a change point in a subsequence of interest starting at α and ending at ω with $1 \leq \alpha < \omega \leq \zeta$. Let

$$M_{h:k} = \sum_{i=h}^{k} m_i, \qquad N_{h:k} = \sum_{i=h}^{k} n_i$$

be consecutive sums (h to k) of binomial distributed event numbers $\{m_i \ : \ i = 1, .., \zeta\}$ and the numbers of trials $\{n_i \ : \ i = 1, .., \zeta\}$ also called bin sizes. The indicators $h{:}k$ can be dropped when the whole sequence of interest is referred to in order to simplify the statistical notation. So, $M \triangleq M_{\alpha:\omega}$ and $N \triangleq N_{\alpha:\omega}$ are the total number of events and subjects within the relevant subsequence. Common change point methods for binomial data are based on maximally selected test statistics for 2×2 tables $T(\#\{{}^{\texttt{Events}}_{\texttt{Exposed}}\}, \#\{{}^{\texttt{Total}}_{\texttt{Exposed}}\}, \#\{{}^{\texttt{Events}}_{\texttt{Nonexposed}}\}, \#\{{}^{\texttt{Total}}_{\texttt{Nonexposed}}\})$ for subsequent k. Let

$$T^k_{\alpha:\omega} = T(M_{\alpha:k}, N_{\alpha:k}, M_{k+1:\omega}, N_{k+1:\omega}) \quad | \quad M, \{n_i \ : \ i = 1, .., \zeta\},$$

and $T^{\max}_{\alpha:\omega} = \max_k |T^k_{\alpha:\omega}|$ be the maximum over $k = \alpha, .., \omega$. Conditional on $M(=M_{\alpha:k}+M_{k+1:\omega})$ and all bin sizes fixed, $T^k_{\alpha:\omega}$ is dependent on $M_{\alpha:k}$ only and we may simply write it as a function $T_{\alpha:\omega}(M_{\alpha:k} \mid M)$. Worsley (1983) gives exact distributions of such maximally selected test statistics, which we want to generalize. All inference is made conditional on M with M/N being sufficient for the event probability p to eliminate this nuisance parameter. The only regularity assumption for $T^k_{\alpha:\omega}$ requires some monotonicity in $M_{\alpha:k}$, which is usually fulfilled for sensible choices of $T(\cdot)$. While one-sided

test statistics naturally are monotone, two-sided test statistics have to be monotone for all decreasing $M_{\alpha:k} \leq m_{0k}$ and for all increasing $M_{\alpha:k} \geq m_{0k}$ separately with m_{0k} being the argmin(T) for a given $T(\cdot)$. Usually, $m_{0k}/N_{\alpha:k}$ is close to M/N. With this assumption, it is guaranteed that events $\{T^k_{\alpha:\omega} < t\}$ can be expressed as a set A_k being an interval $\{a_k \leq M_{\alpha:k} \leq b_k\}$ with $a_k = \inf\{M_{\alpha:k} : T^k_{\alpha:\omega} < t\}$ and $b_k = \sup\{M_{\alpha:k} : T^k_{\alpha:\omega} < t\}$. The test statistics given in Worsley (1983) were the log likelihood ratio and the cumulative sum statistic. For fixed k, the log likelihood ratio statistic is

$$L^k_{\alpha:\omega} = 2\{l(N_{\alpha:k}, M_{\alpha:k}) + l(N_{k+1:\omega}, M_{k+1:\omega}) - l(N, M)\}$$

with $l(n, m) = m\log(m) + (n-m)\log(n-m) - n\log(n)$ being the log likelihood function. The statistic was first used by Hinkley and Hinkley (1970). The cumulative sum statistic is

$$Q^k_{\alpha:\omega} = \frac{(M_{\alpha:k} - M \cdot N_{\alpha:k}/N)}{(N^{\frac{1}{2}}\sigma)}$$

with $p_0 = M/N$ and $\sigma = \sqrt{p_0(1-p_0)}$. Q has the same distribution as the Kolmogorov–Smirnov statistic $D_{M,N}$ (Gibbons, 1985; Pettitt, 1979). Also briefly mentioned in Worsley (1983) is the usual Pearson χ^2 statistic (Miller & Siegmund, 1982) that is

$$P^k_{\alpha:\omega} = \frac{\left(Q^k_{\alpha:\omega}\right)^2}{\frac{N_{\alpha:k}}{N} \cdot \frac{N_{k+1:\omega}}{N}}$$

for testing the equality of p and p'. P is equivalent to a two-sample version of the Anderson–Darling statistic (Darling, 1957; Halpern, 1999). Exact two-sided statistical inference with P yields the exact same results as when using the z-statistic

$$Z^k_{\alpha:\omega} = \frac{\frac{M_{\alpha:k}}{N_{\alpha:k}} - \frac{M_{k+1:\omega}}{N_{k+1:\omega}}}{\sigma\sqrt{\frac{1}{N_{\alpha:k}} + \frac{1}{N_{k+1:\omega}}}},$$

which is by some authors referred to as *z-pooled* (Mehrotra, Chan, & Berger, 2003). Since $Z^2 = P$ holds, it is sufficient to use the Pearson statistic when two-sided testing is carried out. Another statistic introduced by Halpern (1999) is the p-value of Fisher's exact test, that is,

$$\mathcal{F}^k_{\alpha:\omega} = \sum_{j \in J} \mathrm{P}(X_1 = j \mid M) = \sum_{j \in J} \frac{\binom{N_{\alpha:k}}{j}\binom{N_{k+1:\omega}}{M-j}}{\binom{N}{M}}$$

with random $X_1 = \#^{\texttt{Events}}_{\texttt{Exposed}}\}$ and $J = \{j : \rho(j) \leq \rho(M_{\alpha:k})\}$ with given $\rho(u) = \mathrm{P}(X_1 = u \mid M)$ such that a two-sided p-value is obtained. Similarly, Blaker's test is obtained with $\rho(u) = \min(\mathrm{P}(X_1 \leq u \mid M),\ \mathrm{P}(X_1 \geq u \mid M))$ (Fay, 2010). The statistics $L, Q, P, \mathcal{F}$ are the ones most prominently used in the literature, despite many more statistics for testing a change in probability are available, for example, variations of *z-pooled* with separate variance estimation, variations of the two-sided Fisher's exact test (Fay, 2010), rank statistics (Hothorn & Lausen, 2003; Hothorn & Zeileis, 2008; Lausen & Schumacher, 1992), martingale statistics (Brostrom, 1997), or Bayesian approaches (Smith, 1975).

Let $F_{\alpha:\omega}(M)$ be the distribution function of any of the above-mentioned $T^{\max}_{\alpha:\omega}$ under H_0 conditional on the number of events M in the (sub)sequence. Worsley (1983) gives an exact iterative procedure to calculate $F_{\alpha:\omega}(M)$ as follows. For each candidate change point k, the probability $\mathrm{P}(T^k_{\alpha:\omega} < t)$ can be expressed in terms of the events A_k of the form $\{a_k \leq M_k \leq b_k\}$ as defined above. The events $\{T^{\max}_{\alpha:\omega} < t\}$ are thus equivalent to $\cap^{\omega}_{i=\alpha} A_i$. Given all A_i and $M = m$ fixed, let

$$F_{\alpha:k}(v) = \mathrm{P}\left(\max_{i=1,..,k}\{T^i_{\alpha:\omega}\} < t \ \mid\ M_{\alpha:k} = v\right) = \mathrm{P}\left(\cap^k_{i=\alpha} A_i \ \mid\ M_{\alpha:k} = v\right)$$

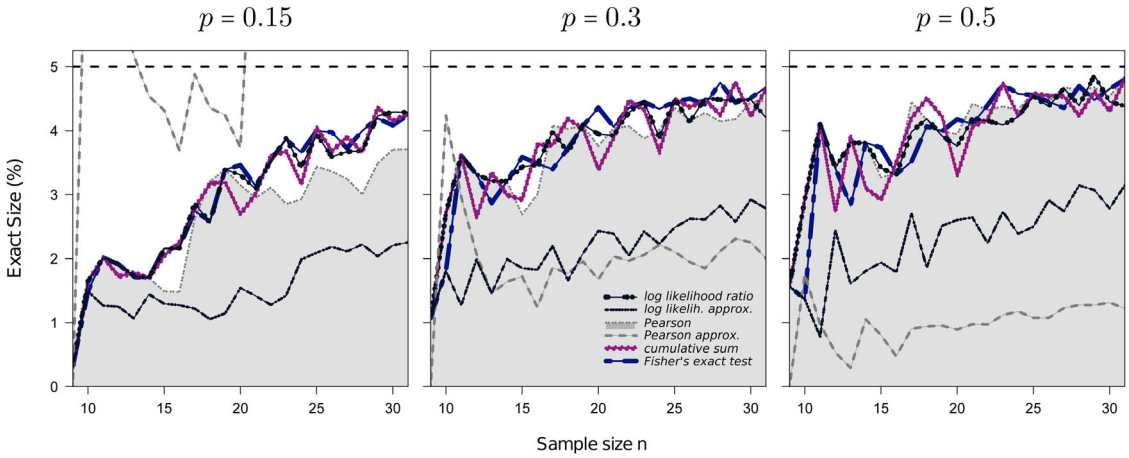

FIGURE 2 Exact size of Worsley's test and the asymptotic Brownian bridge approximation for various test statistics. The event probability is constant $p = .15$ (left), $p = .3$ (center), and $p = .5$ (right)

so that $F_\alpha(v) = 1$ if $a_\alpha \le v \le b_\alpha$ and $F_{\alpha:\omega}(m) = \mathrm{P}(\cap_{i=\alpha}^{\omega} A_i)$. If $a_{k+1} \le v \le b_{k+1}$, that probability is calculated stepwise according to

$$F_{\alpha:k+1}(v) = \sum_{u=a_k}^{b_k} \mathrm{P}(\cap A_i \mid M_k = u, M_{k+1} = v) \cdot \mathrm{P}(M_k = u \mid M_{k+1} = v)$$

$$= \sum_{u=a_k}^{b_k} F_{\alpha:k+1}(u) \cdot h_k(u, v),$$

where $h_k(u, v)$ is the probability function of the hypergeometric distribution, defined for $0 \le u \le N_{\alpha:k}$ and $0 \le v - u \le n_{k+1}$ $(= N_{k+1:k+1})$ as

$$h_k(u, v) = \binom{N_{\alpha:k}}{u}\binom{n_{k+1}}{v-u} / \binom{N_{\alpha:k+1}}{v}$$

and probability zero otherwise. This result can now be used iteratively for $k = 1, .., c-2$ to find $F_{\alpha:k}(v)$ for $a_k \le v \le b_k$, and finally, produces the p-value $1 - F_{\alpha:\omega}(m) = \mathrm{P}(T^{\max}_{\alpha:\omega} \ge t)$.

Besides the exact approach, several approximating distributions exist. In case of large sample sizes, asymptotic procedures may be preferred since exact methods may be time-consuming in their calculation. We will use the Brownian bridge approximation for maximally selected χ^2 statistics by Miller and Siegmund (1982) as a comparator. The approximating probability calculates as follows:

$$\mathrm{P}(T > t) \cong 4\frac{\phi(\sqrt{t})}{\sqrt{t}} + \phi(\sqrt{t}) \cdot \left(\sqrt{t} - 1/\sqrt{t}\right) \cdot \log\left(\frac{N_{\omega:\omega} \cdot N_{\alpha:\omega-1}}{N_{\alpha:\alpha} \cdot N_{\alpha+1:\omega}}\right).$$

The term $N_{\alpha:\alpha}$ hereby represents the first bin only. Since we have binomial data, the first bin should contain all the n_i where we are not interested in seeking for a change point (or maximizing our test statistics). In applications, this is often used to shift the change point detection mildly away from the tails of the sequence. Regarding the Brownian bridge approximation, this is, however, a crucial parameter. In Bernoulli sequences, we restricted our search to the central 90% of the sequence as, for example, in Friede et al. (2006).

With the described procedure by Worsley, one can calculate the exact size of a change point test using any statistic. For all M and all $M_{\alpha:k}$, we can calculate the probability of $T^k_{\alpha:k}$ exceeding a certain threshold, that is, the exact 95% or 99% quantile conditional on M or quantile from an (asymptotic) approximation. With M being drawn from a binomial distribution under H_0, we can assess to what extent the nominal significance level is actually used. For Bernoulli sequences of small sample sizes, we see in Figure 2 a decline in size that relates to a loss in statistical power. The various test statistics

lead to very similar exact sizes, with no clear favorite. The Brownian bridge approximation is mostly conservative for our choices of parameters and their type I error probability is only slowly converging to the nominal level of 5%. Here, it is only applied to the log likelihood ratio statistic L and the Pearson statistic P since those are "maximally selected χ^2" distributed. The approximation was developed by Miller and Siegmund (1982) for the latter.

4 | BINARY SEGMENTATION

Binary segmentation is an iterative procedure to hierarchically split sequences (Scott & Knott, 1974) usually applied in order to detect multiple change points. Initially, the entire data set is searched for one change point. Once a change point is detected, the data are then split into two subsegments: one to the left of the change point and the second one to the right of the detected change point. Subsequent change point detections are then performed on either subsegment, possibly resulting in further splits. This iterative procedure continues until a stopping criterion is met, for example, until significance cannot be achieved at a prespecified level. A plethora of criteria when to split and when to stop has been suggested. The choice has complex implications for consistency. The general trade-off to be made was described by Scott and Knott (1974): that "Choosing an appropriate value for α is difficult. If α is too small, the splitting process will terminate too soon, while if α is too large, the process will go too far and split homogeneous sets of means." Vostrikova (1981) showed consistency of binary segmentation for the number and locations of change points, with rates of convergence of the estimators of locations, under certain technical conditions. For Gaussian processes, Venkatraman (1992) relaxed these conditions on the number and locations of the change points. Furthermore, a simulation study was done to assess the performance of various multiple change point detection methods specifically in small samples, proving real-world applicability. The theory is outlined and discussed for nonnormal cases as well.

Usual descriptions of the method do not consider that the distribution of the maximized statistics in any subsegment is now conditional on all previously found change points. For example, Scott and Knott (1974) write

> This starts with the best split into two groups, based on the between groups sum of squares, and then applies the same procedure separately to each subgroup in turn. The subdivision process is continued until the resulting groups are judged to be homogeneous. [509]

Fryzlewicz (2014) gives an algorithm for "standard" binary segmentation in pseudocode where the same segmentation step is recursively called at each splitting of the data into a left and a right subinterval without any constraints implied through previous steps. To the best of our knowledge, authors so far have not considered any impact of these constraints throughout the repeated/iterative splitting steps, since consistency followed by asymptotic results (Vostrikova, 1981) as well as in simulation studies. Here, we condition on preceding steps and provide exact distributions in all steps throughout the segmentation procedure that considers that subsequences are no longer completely random under H_0 (apart from the initial step). In contexts of developing asymptotic results for multiple change point segmentations of the data, this aspect may be neglectable, and is proven to be asymptotically correct under some regularity constraints. Therefore, failure to consider this aspect during subsequent steps of the procedure will not question the validity nor the usefulness of such a "standard binary segmentation", as we will refer to it in the following. For our purpose, however, we will need to consider rigorous and exact distributions at all steps. We will show that the usage of binary segmentation steps without considering the conditional distribution of the (pseudo) change points found beforehand can lead to conservative but also to liberal results in some scenarios.

5 | EXACT NULL DISTRIBUTIONS UNDER SIDE CONDITIONS

In this section, we extend the iterative procedure by Worsley (1983) for calculating the probability of $T^{\max}_{\alpha:\omega}$ when side conditions are present. Let s be the number of side conditions that are denoted as $C^l,\ l = 1, .., s$ and for intersections, the notation $\bigcap\!\!\!\!+\,_{l=1}^{s} C^l$ is used (to distinguish them from intersections "$\bigcap_{i=\alpha}^{\omega}$" representing possible runs within the sequence). The conditions C^l shall be measurable regarding m_i for $i = 1, .., \zeta$. Similarly to the definition of the A_k in Section 3, we define sets C^l_k of the form $\{c^l_k \le M_{\alpha:k} \le d^l_k\}$ with $c^l_k = \inf\{M_{\alpha:k} : \mathrm{P}(M_{\alpha:k} \mid C^l) > 0\}$ and $d^l_k = \sup\{M_{\alpha:k} : \mathrm{P}(M_{\alpha:k} \mid C^l) > 0\}$. We restrict the C^l_k to be intervals as we did for the A_k. Then intersections between them as $A_k \cap\!\!\!+ C^l_k$ for any $l = 1, .., s$

and $\bigoplus_{l\in\{1,..,s\}} C_k^l$ are thus also intervals. This regularity condition should always be true whenever the A_k and C_k^l arise by meaningful test statistics. This assumption could eventually be dropped but the notation and implementation would be more complex.

Side conditions C^l naturally arise whenever tests were carried out (in a hierarchical fashion) beforehand within the binomial sequence. We can regard C^l as all information received in previous steps to condition on. When steps of a binary segmentation are sequentially done, the conditioning should account for initial change point tests on the sequence $1,..,\zeta$ with a test statistic U_s (possibly different from $T_{1:\zeta}$). The indices l in U_l will also reflect the subsequence $\alpha' : \omega'$ (with $1 \le \alpha' \le \alpha,\ \omega \le \omega' \le \zeta$) of the $s-l+1$th step in the binary segmentation procedure. This fixes the attained maximum value of the test statistic at the observed $U_s^{\max} = \max_{i=1}^{\zeta}\{U_s(M_{1:i}, N_{1:i}, M_{i+1:\zeta}, N_{k+1:\zeta})\}$ at the position of the detected change point. Considering the number of events as random within the subsequence $k = \alpha,..,\omega$, the initial step $l = s$ will impose restrictions of the form

$$C_k^s = \{M_{\alpha:k} \ :\ U_s(M_{\alpha:k}+M_{1:\alpha-1},\ N_{1:k},\ M_{k+1:\omega}+M_{\omega+1:\zeta},\ N_{k+1:\zeta}) < U_s^{\max}\}.$$

We investigate the distribution of $M_{\alpha:k} \mid C^l$ conditional on fixed $M_{\alpha:\omega}$, the number of events in the subsequence, as Worsley's method does. Then all $\{m_i : i = 1,..,\alpha-1,\ \omega+1,..,\zeta\}$ outside of the considered subsequence $(\alpha : \omega)$ are fixed (thus also $M_{1:\alpha-1}$ and $M_{\omega+1:\zeta}$). Hence, U_l is random only in $M_{\alpha:k}$. Denote $U_l^k = U_l(M_{\alpha:k})$.

For the hypothetical binary segmentation procedure, attention must be paid to the decision rule where to split when a maximum is not unique, that is, is attained at multiple possible change points k. A variety of such rulings can be considered from preferring an early or late change point to splitting the sequence directly in multiple subsegments. The ruling we chose picks the change point k that is the most to the left, which corresponds to the earliest change point if the ordering is by time. Any decision rule used will impact the C_k^l since the side conditions might allow the case of equality $U_l(M_{\alpha:k}) = U_l^{\max}$. When the decision rule is to take the left change point in step $s-l+1$, this would forbid the case of equality only on the left subsequence but would allow to attain further maxima on the right subsequence in the following steps. Let $\delta_{\alpha:\omega}^l$ be an indicator function that is 1 when equality is allowed and 0 otherwise. We then define C_k^l in the case of equality as $C_k^l = \{M_{\alpha:k} \ :\ U_l(M_{\alpha:k}) \le U_l^{\max} \ \text{IF}\ \delta_{\alpha:\omega}^l = 1\}$ and in the case of inequality as above. We now want to calculate the probability $\mathrm{P}(T_{\alpha:\omega}^{\max} < t \mid M,\ \bigoplus_{l=1}^{s} C^l)$ under H_0 conditional on the fixed, observed number of successes M of our sequence and conditional on the additional restrictions $\{C^l\}_{l=1}^s$ arising through the hierarchical steps in binary segmentation. Similar to Section 3, we define the probability of a partial maximum $\alpha : k$ not exceeding t as:

$$\begin{aligned} F_{\alpha:k}^{C}(v) &= \mathrm{P}\left(\max_{i=1,..,k}\{T_{\alpha:\omega}^i\} < t \ \middle|\ M_{\alpha:k} = v,\ \bigoplus_{l=1}^{s} \max_{i=1,..,k} \left\{ U_l(M_{\alpha:i}) < U_l^{\max} + \varepsilon \cdot \delta_{\alpha:\omega}^l \,, \forall\, \varepsilon > 0 \right\} \right) \\ &= \mathrm{P}\left(\bigcap_{i=1}^{k} A_i \ \middle|\ M_{\alpha:k} = v,\ \bigoplus_{l=1}^{s} \bigcap_{i=1}^{k} C_i^l \right). \end{aligned}$$

For the implementation of an exact procedure, we need the following theorem.

Theorem 5.1. *Under the null hypothesis H_0 and for any possible change point $k = \alpha, \dots, \omega-1$*

$$F_{\alpha:k+1}^{C}(v) = \frac{\sum_{u=\max(a_k, c_k^1,..,c_k^s)}^{\min(b_k, d_k^1,..,d_k^s)} G_{\alpha:k}(u) \cdot h_k(u,v)}{\sum_{u=\max(c_k^1,..,c_k^s)}^{\min(d_k^1,..,d_k^s)} G'_{\alpha:k}(u) \cdot h_k(u,v)}$$

holds with

$$G_{\alpha:k}(v) = P\left(\bigcap_{i=1}^{k} \bigoplus_{l=1}^{s} \{A_i \cap C_i^l\} \ \middle|\ M_{\alpha:k} = v \right),$$

$$G'_{\alpha:k}(v) = P\left(\bigcap_{i=1}^{k} \bigoplus_{l=1}^{s} C_i^l \ \middle|\ M_{\alpha:k} = v \right)$$

and $\max(a_{k+1}, c^1_{k+1}, .., c^s_{k+1}) \leq \upsilon \leq \min(b_{k+1}, d^1_{k+1}, .., d^s_{k+1})$ *as well as the hypergeometric probability function* $h_k(u, \upsilon)$ *as defined in Section 3.*

Proof of Theorem 5.1. Initially, $F^C_{\alpha:\alpha}(\upsilon) = 1$ for $\max(a_1, c^1_1, .., c^s_1) \leq \upsilon \leq \min(b_1, d^1_1, .., d^s_1)$ and iteratively for $k+1 \leq \omega$, we can write

$$F^C_{\alpha:k+1}(\upsilon) = \mathrm{P}\left(\bigcap_{i=1}^{k+1} A_i \mid M_{\alpha:k+1} = \upsilon, \bigoplus_{l=1}^{s}\bigcap_{i=1}^{k+1} C^l_i\right)$$

$$= \frac{\mathrm{P}\left(\bigcap_{i=1}^{k+1} A_i \cap \bigoplus_{l=1}^{s}\bigcap_{i=1}^{k+1} C^l_i \mid M_{\alpha:k+1} = \upsilon\right)}{\mathrm{P}\left(\bigoplus_{l=1}^{s}\bigcap_{i=1}^{k+1} C^l_i \mid M_{\alpha:k+1} = \upsilon\right)}.$$

Reordering the terms gives

$$= \frac{\mathrm{P}\left(\bigcap_{i=1}^{k+1} \bigoplus_{l=1}^{s}\{A_i \cap C^l_i\} \mid M_{\alpha:k+1} = \upsilon\right)}{\mathrm{P}\left(\bigcap_{i=1}^{k+1} \bigoplus_{l=1}^{s} C^l_i \mid M_{\alpha:k+1} = \upsilon\right)} = \frac{G_{\alpha:k+1}(\upsilon)}{G'_{\alpha:k+1}(\upsilon)}.$$

For the numerator $G_{\alpha:k+1}(\upsilon)$ (defined similar as $F_{\alpha:k+1}(\upsilon)$ in Section 3) we use, the distribution $G_{\alpha:k}(\upsilon)$ for all values $M_{\alpha:k} = u$ can attain

$$G_{\alpha:k+1}(\upsilon) = \sum_{u=\max(a_k, c^1_k, .., c^s_k)}^{\min(b_k, d^1_k, .., d^s_k)} \mathrm{P}\left(\bigcap_{i=1}^{k} \bigoplus_{l=1}^{s}\{A_i \cap C^l_i\} \mid M_{\alpha:k}{=}u, M_{\alpha:k+1}{=}\upsilon\right) \cdot \mathrm{P}(M_{\alpha:k}{=}u \mid M_{\alpha:k+1}{=}\upsilon).$$

Conditional on $M_{\alpha:k+1}$ and $M, M_{\alpha:\alpha}, .., M_{\alpha:k}$ are independent of $M_{\alpha:k+1}$ and hence

$$\mathrm{P}\left(\bigcap_{i=1}^{k} \bigoplus_{l=1}^{s}\{A_i \cap C^l_i\} \mid M_{\alpha:k}{=}u, M_{\alpha:k+1}{=}\upsilon\right) = \mathrm{P}\left(\bigcap_{i=1}^{k} \bigoplus_{l=1}^{s}\{A_i \cap C^l_i\} \mid M_{\alpha:k}{=}u\right) = G_{\alpha:k}(\upsilon)$$

and $\mathrm{P}(M_{\alpha:k}{=}u \mid M_{\alpha:k+1}{=}\upsilon)$ following a hypergeometric distribution $h_k(u, \upsilon)$, it follows:

$$G_{\alpha:k+1}(\upsilon) = \sum_{u=\max(a_k, c^1_k, .., c^s_k)}^{\min(b_k, d^1_k, .., d^s_k)} G_{\alpha:k}(u) \cdot h_k(u, \upsilon).$$

The same derivations for the denominator $G'_{\alpha:k+1}(\upsilon)$ complete the proof. □

With Theorem 5.1, we can use the iterative procedure as given in Section 3 to calculate $G_{\alpha:\omega}(M)$ and $G'_{\alpha:\omega}(M)$, and thus, $\mathrm{P}(T^{\max}_{\alpha:\omega} < t \mid M, \bigoplus_{l=1}^{s} C^l) = 1 - F^C_{\alpha:\omega}(M)$. The rigorous derivation of the conditional distributions allows the realization of exact distributions of the steps in a (hypothetical) binary segmentation procedure. Since binary segmentation is usually referred to as a multiple change point detection method, we use the term *exact binary segmentation steps*, since we do not discuss definitions of a stopping criteria, which would define such a procedure, see, for example, Vostrikova (1981), Venkatraman (1992), and Fryzlewicz (2014). Conversely, we referred to binary segmentation using unconditional distributions as *standard binary segmentation steps*, since this procedure is reliant on asymptotic results. Although multiple constraints arise through change points found beforehand, the method stays a one-dimensional optimization problem in the search for further possible change points. With the number of side conditions only increasing by the depth, going through the exact binary segmentation steps, it stays a *greedy* procedure, that is solvable in polynomial time, whereas approaches that rely on all possible 2^N combinations of the input sequence are only solvable via simulation techniques, as suggested for example by Ross et al. (2013).

6 | PROPOSED TEST UTILIZING AN ORDERING OF SEQUENCES

Let z be an actual instance of a sequence of binomial variables we want to investigate a (single) change point test or some other maximally selected statistic on. The sequence is defined by its bin sizes and events $\{n_i\}(z)$, $\{m_i\}(z)$ and the attained maximum of the test statistic T is $t^{\max}(z)$ whose distribution we derived is conditional on $M(z)$. We have

$$p_W(z) = \mathrm{P}(T^{\max}_{\alpha:\omega} \geq t^{\max}(z) \mid M(z)),$$
$$p^-_W(z) = \mathrm{P}(T^{\max}_{\alpha:\omega} > t^{\max}(z) \mid M(z)),$$
$$p_W - p^-_W = \mathrm{P}(T^{\max}_{\alpha:\omega} = t^{\max}(z) \mid M(z)).$$

A randomized p-value would be achieved with a uniform variable $Y \sim U[0,1]$ on the unit interval by

$$p_R(z) = p^-_W + (p_W - p^-_W) \cdot Y.$$

Randomization yields full size and forms uniformly most powerful test statistics. When testing a change point, we still have unused information in the sequences. The sequences form a natural order regarding the likelihood of further separability in a hypothetical binary segmentation procedure. Fully conditional on the initial change point test, we can use results in Section 5 to determine a p-value of further change points as a "secondary" dimension. The sequence is split at the initial change point $\hat{\kappa}$ into a *left* subsequence z_{left} from $\alpha_{\text{left}} = 1$ to $\omega_{\text{left}} = \hat{\kappa}$ and a *right* subsequence z_{right} from $\alpha_{\text{right}} = \hat{\kappa}+1$ to $\omega_{\text{right}} = \zeta$. Theorem 5.1 is used to determine p-values conditional on the initial estimated change point $\hat{\kappa}$:

$$p_{\text{left}} = \mathrm{P}(T^{\max}_{\alpha_{\text{left}}:\omega_{\text{left}}} \geq t^{\max}(z_{\text{left}}) \mid M(z_{\text{left}}),\ T^{\max}_{1:\zeta} < t^{\max}(z))$$
$$\text{and} \quad p_{\text{right}} = \mathrm{P}(T^{\max}_{\alpha_{\text{right}}:\omega_{\text{right}}} \geq t^{\max}(z_{\text{right}}) \mid M(z_{\text{right}}),\ T^{\max}_{1:\zeta} \leq t^{\max}(z)).$$

To pool p_{left} and p_{right}, a combination function $C(,)$ for p-values will be considered. This approach is used in adaptive clinical trials (Brannath, Posch, & Bauer, 2002) but also in meta-analysis. Fisher's weighted product test (1932) is one possibility to combine p_{left} and p_{right} with the function

$$C(p_{\text{left}}, p_{\text{right}}) = p^w_{\text{left}} \cdot p_{\text{right}}, \quad w > 0.$$

Another popular approach is the inverse normal method (Lehmacher & Wassmer, 1999)

$$C(p_{\text{left}}, p_{\text{right}}) = 1 - \Phi\big(w_1 \cdot \Phi^{-1}(1-p_{\text{left}}) + w_2 \cdot \Phi^{-1}(1-p_{\text{right}})\big)$$

also called Stouffer's method (1949) with possible weights $0 \leq w_i < 1$ and $w_1^2 + w_2^2 = 1$. Many other combination functions have been proposed, some also specifically for discrete p-values. Kincaid (1962) compares methods to pool discrete p-values. Still, these methods are not easily adopted to the setting considered here, since they need full derivations of the exact discrete distributions (which depend on the sequence). The calculations required would be of exponential order and are therefore not considered further. If $C(p_{\text{left}}, p_{\text{right}})$ is not already a valid pooled p-value, it is defined as

$$c(p_{\text{left}}, p_{\text{right}}) = \int_0^1 \int_0^1 \mathbf{1}_{\{C(x,y) \leq C(p_{\text{left}}, p_{\text{right}})\}} dx\, dy.$$

It is guaranteed that under H_0, $c(,)$ is stochastically smaller than a uniform distribution $\mathrm{P}(c(p_{\text{left}}, p_{\text{right}}) \leq x \mid C) \leq F^{\,\mathrm{U}[0,1]}(x)$ since $p_{\text{left}}(z_{\text{left}})$ and $p_{\text{right}}(z_{\text{right}})$ are conditionally independent.

For our purposes, we will use Fisher's product test in the following, since it is better suited for p-values that are not continuous. For instance, p-values need to be truly smaller than 1 for the inverse normal method. Besides the numerical instability, the strong impact of p-values close to 1 will lead to less homogeneous pooling. Furthermore, we only pool the left and the right p-values if both subsequences are informative. This way subsequences with no events or the maximal number of events are not considered further by getting weight zero. If both subsequences are uninformative, the pooled

p-value will be set to 1. With the known distribution for the Fisher's product test, we obtain

$$c_F(p_{\text{left}}, p_{\text{right}}) = F^{-1}_{\chi^2}(-2\,(\log(p_{\text{left}}) + \log(p_{\text{right}}))$$

and a new p-value

$$p_N(z) = p_W^- + (p_W - p_W^-) \cdot c_F(p_{\text{left}}, p_{\text{right}})$$

since $c_F(p_{\text{left}}, p_{\text{right}})$ is determined conditional on the initial change point test including p_W and thus independent. Thus, by construction, the test keeps prespecified significance levels and is at least as powerful as Worsley's exact test since the new test is less discrete and $p_N \leq p_W$ holds.

The described test operates in depth 1, but can be easily extended by applying the same approach recursively to p_{left} and p_{right}. This hierarchical procedure resembles binary segmentation and forms some sort of "segmentation p-value" rather than a single change point p-value. The new test uses an exchange of information through different depths to make the test at a given depth more precise. Also when further segmentation is not of any interest, the approach is natural since it favors sequences with a sharp change in the empirical frequency of events. In order to focus on such local sharp changes, it may also be advisable to change the underlying test statistic from two-sided to one-sided, such that the subsequence on the side of the change point with a low frequency is searched for an increasing frequency the further this subsequence goes away from the change point. Conversely, the subsequence on the side with a high frequency is searched with the one-sided test in the opposite direction of a decreasing frequency. We refer to this procedure in the following as *swapped one-sided alternatives*.

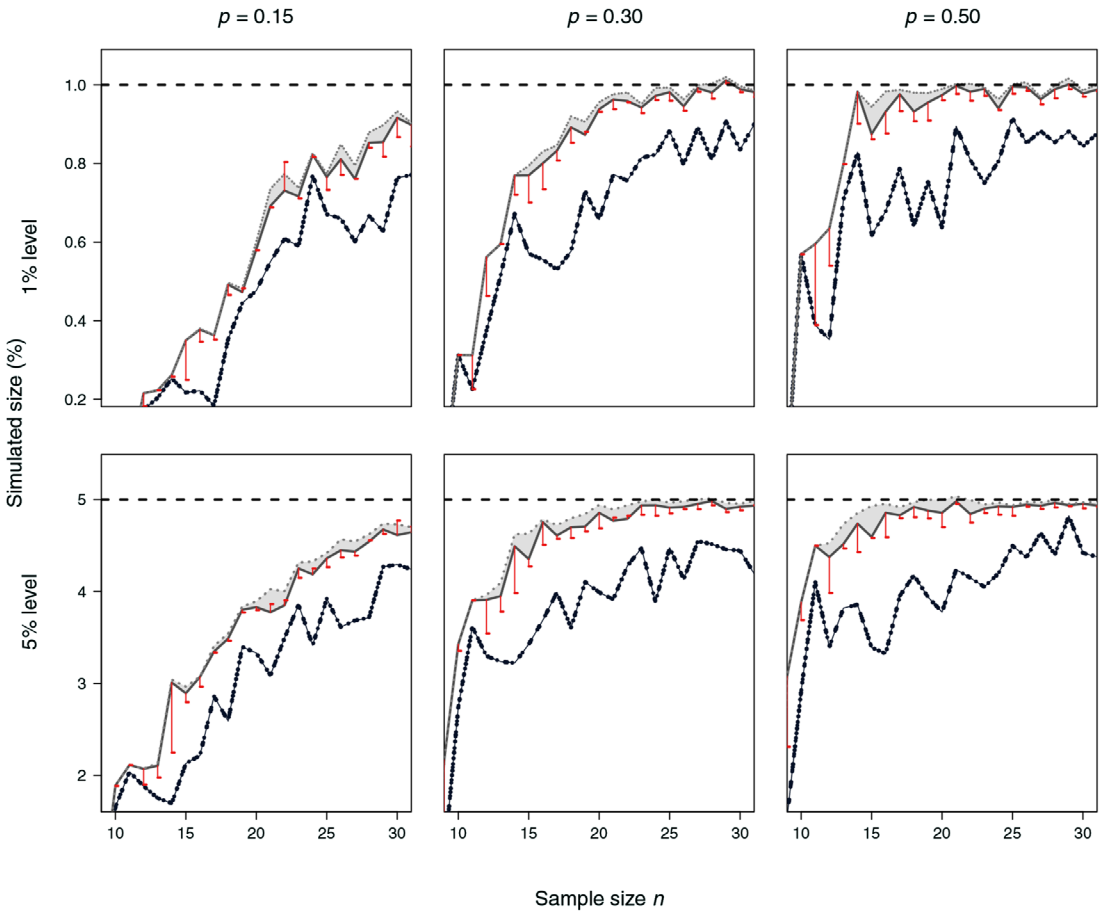

FIGURE 3 Simulated sizes of the *log likelihood ratio* (black) test statistics for Worsley's test (depth 0). The new test (depth 1) is displayed as solid gray line. Dotted lines with shaded area show the gain through deeper segmentation (depth 3+). The whiskers (red) show the difference induced by the *standard approach* (depth 1). The nominal level varies 1% [top] and 5% [bottom]; the number of simulation runs is $n_{\text{sim}} = 500{,}000$ per parameter combination

In our definition of the test, we do not provide specific *stopping criteria* regarding the segmentation other than the prespecified depth, or more precisely, the forced stop whenever the resulting subsequence consists exclusively of events or nonevents. We point out that the stopping criteria defined this way were chosen with the intent to get a less discrete test. However, it is neither a valid nor a sensible criterion for detecting multiple change points. The latter is a separate setting for which we refer to, for example, Scott and Knott (1974) and Vostrikova (1981), in which only a rigorous stopping criteria will be able to obtain a binary segmentation procedure in the original sense. A variety of other methods for (direct) detection of multiple change points exist, many of them proven to provide better results than binary segmentation procedures in certain multiple change point applications (see, e.g., Frick, Munk, & Sieling, 2014; Zou, Yin, Feng, & Wang, 2014). We use simply the idea of a segmentation procedure (without a stopping criteria) to obtain less discrete test statistics. The derived exact conditional distributions, however, can be used to evaluate any given segmentation procedure (with well-defined splitting and stopping criteria) that is based on Worsley's test.

7 | NUMERICAL STUDIES

In the following, we will explore the properties of the proposed methods by means of a simulation study and by application of motivating examples introduced in Section 2.

7.1 | Simulation studies

First, we look at Monte-Carlo simulations to compare the new test with Worsley's test and approaches based on *standard binary segmentation* steps. To investigate to which extent the use of the new ordering can lead to a gain in size, a simulation

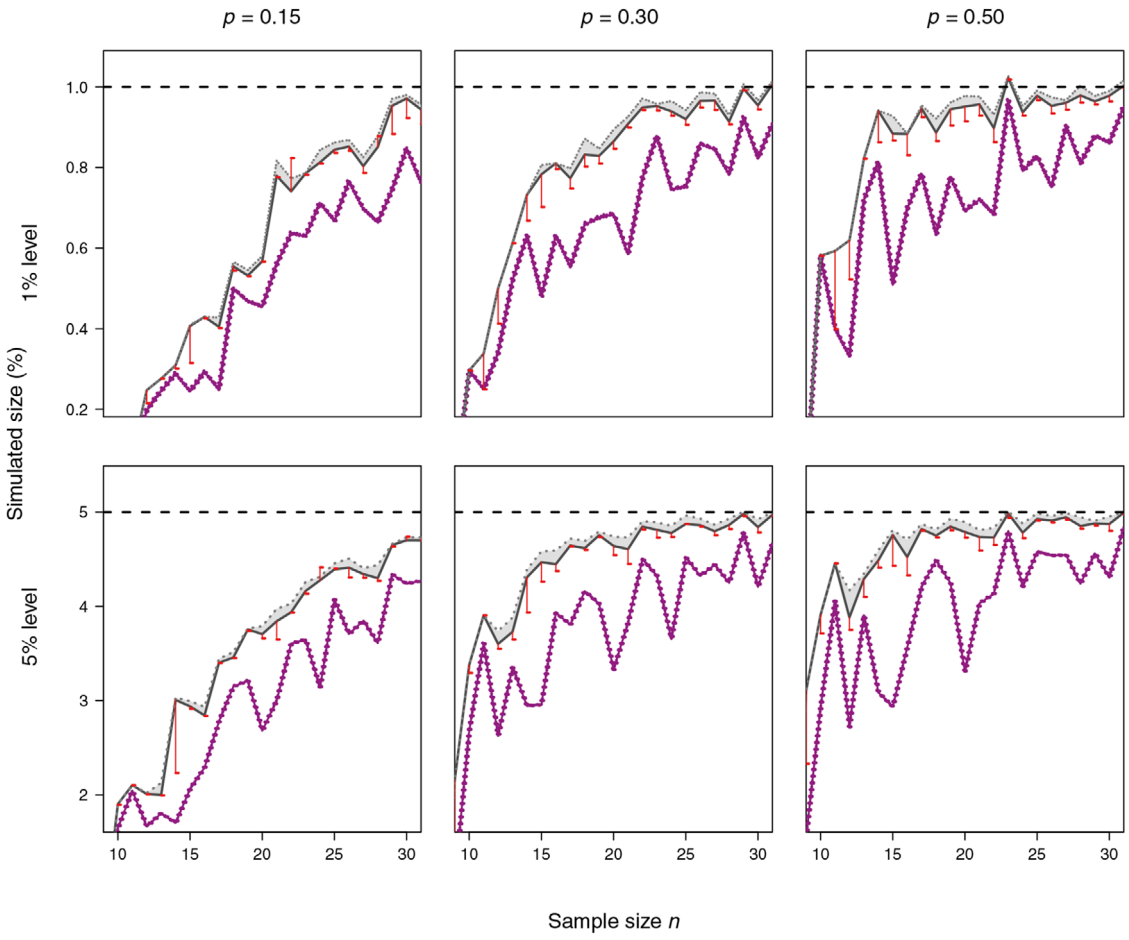

FIGURE 4 Simulated sizes of the *cumulative sum* (purple) as underlying test statistics for Worsley's test, as in Figure 3

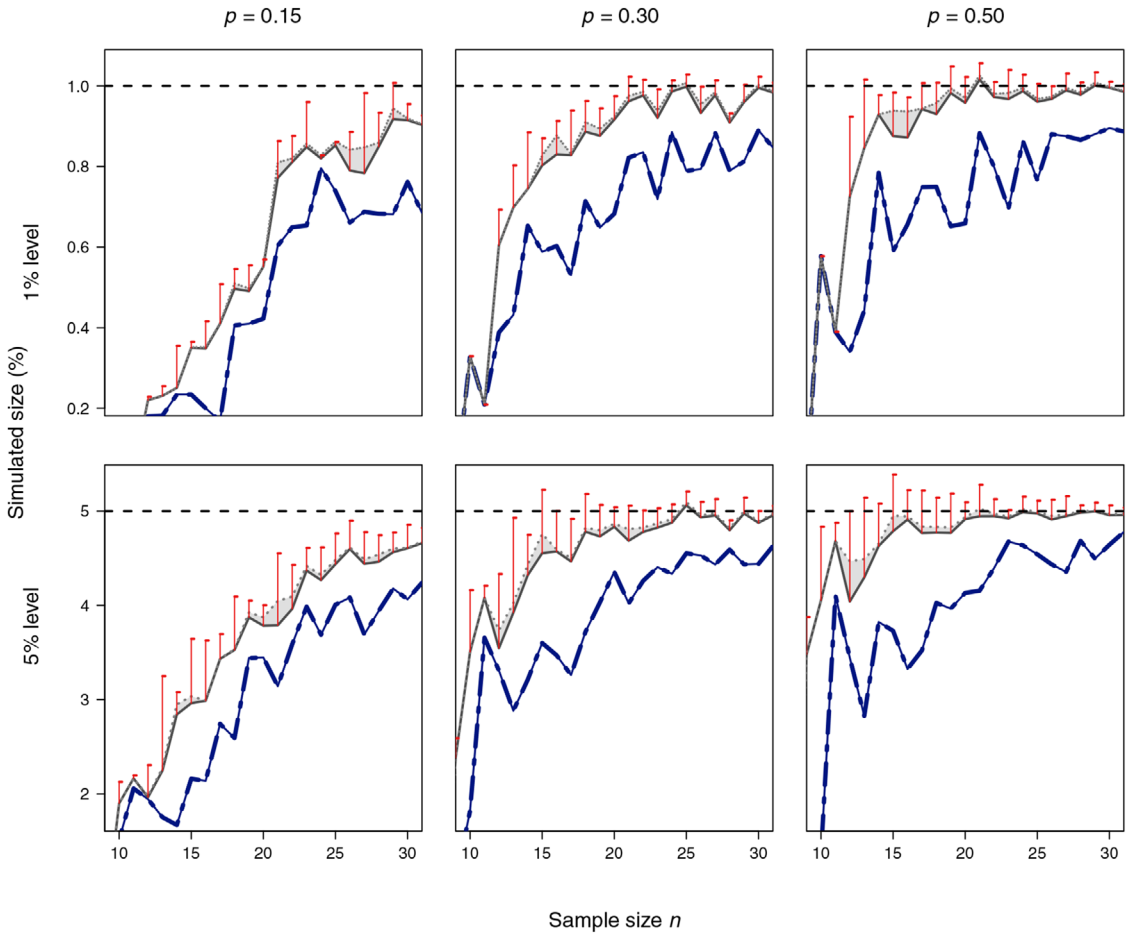

FIGURE 5 Simulated sizes of *Fisher's exact test* (blue) with *swapped one-sided alternatives* as underlying test statistics for Worsley's test, as in Figure 3

study on randomly generated Bernoulli sequences under H_0 was done. Figure 3 shows the results for various depths in the creation of the sequential ordering. A higher depth always leads to a less discrete and therefore smaller p-value and thus a higher size of the test. The log likelihood ratio test and cumulative sum test as underlying test statistics as proposed in Worsley (1983) were used. Also, Fisher's exact test is used with the two-sided version for the initial test but for higher depth swapped one-sided versions as described in Section 6. All test statistics show about equally large size irrespective of the depth, the nominal level (1% or 5%), and the true event probability ($p = .3$ or $p = .5$). The lowest line shows the performance of the original Worsley's test that can be referred to as having depth zero. The new approach leads to a strong increase in the size already with depth 1. Even for sample sizes up to 25, this increase can be above one-fifth of the nominal level. Only when the event probability is very small (or high), the effect diminishes since many randomly generated data sets become rather trivial. When the search depth is further increased, the size slightly improves for depth 2 and very little for depth 3. Search depths beyond three only very occasionally undiscretize a p-values and then to an almost unnoticeable extent. Therefore, the gain in size beyond depth 3 is close to zero. Exact calculations of the size (as displayed in Figure 2) are no longer feasible for depths greater zero, since the number of distinguishable sequences is of exponential order (2^N). When the p-value is achieved by steps of *standard binary segmentation*, the statistical test becomes predominantly liberal in the case of swapped one-sided alternatives, while otherwise it is often conservative (see Figures 3–5).

The power of the test statistics is displayed in Figures 6,7. The gain in power depends strongly on the simulation scenario and the test statistics used. Scenarios with alternatives consisting of one change point only are displayed in the upper tier. In the bottom row, two change points were used and the new method based on segmentation can benefit even more from such an alternative to reject the null hypothesis of no change point. The gain in power can be substantial as shown for the log likelihood ratio statistic that has a large statistical power in the tails of the sequence, as well as for the cumulative sum statistic having a large statistical power in the center of the sequence.

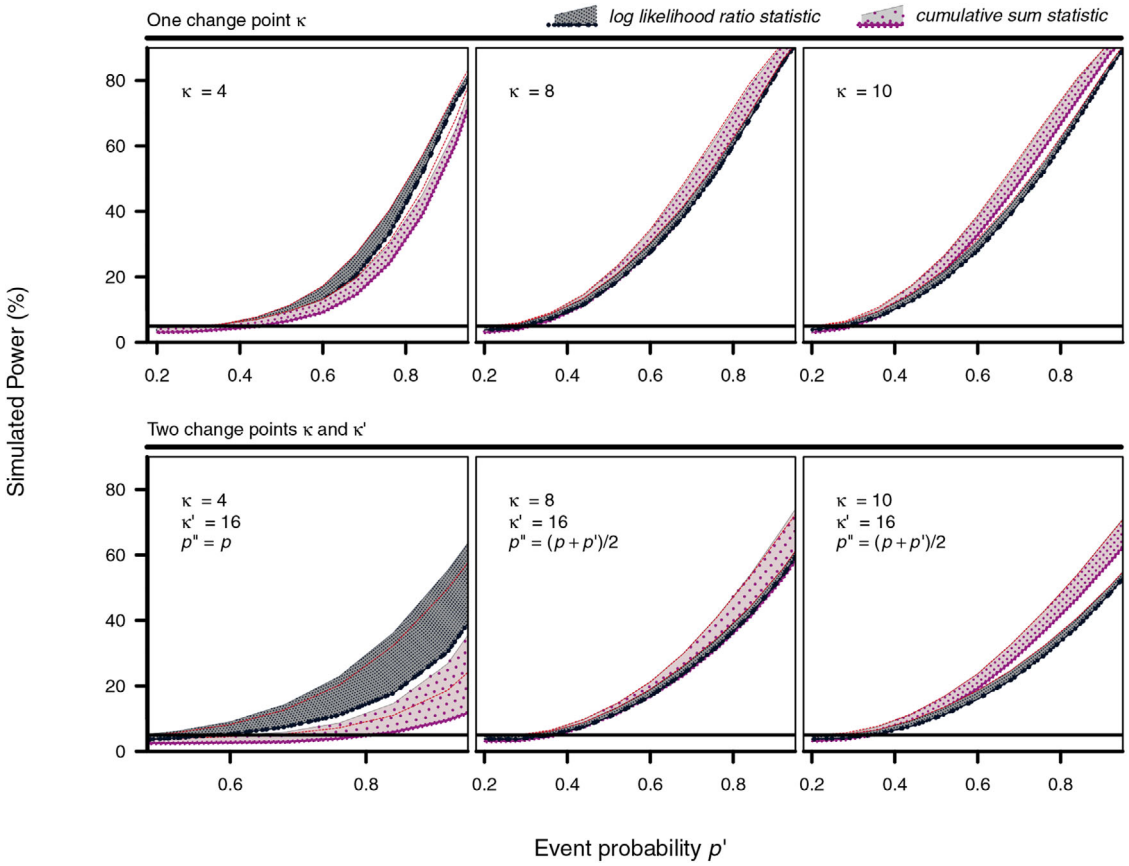

FIGURE 6 Simulated power of log likelihood ratio statistic L and cumulative sum statistic Q comparing depth 0 (solid bottom line) and the gain through depth 3 (area above the line) for different change points κ on a sample of length $N = 20$ with probabilities $p = .2$ fixed and p' varying (x-axis). The bottom tier represents scenarios with a second change point with additional parameters κ' and subsequent probabilities p'' for $\{m_i : i = \kappa', \ldots, 20\}$. The tiny dotted red lines indicate the power of the randomized test version of Worsley's test for comparison. Per parameter combination, $n_{\text{sim}} = 100{,}000$ simulation runs were done

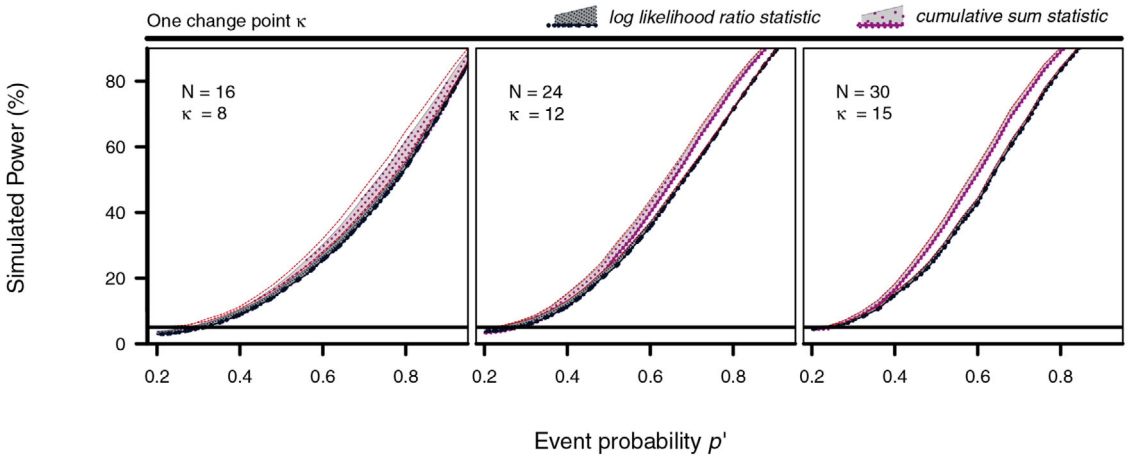

FIGURE 7 Simulated power for different lengths of the sample N and change points κ, as in Figure 6

7.2 | Motivating examples revisited

In the clinical data example about the likelihood of pin site infections in orthopedic surgeries, the estimated change point is located before observation 12 that is before the introduction of the new procedure after observation 17, as displayed in Panel A in Figure 1. Worsley's test gives a p-value of $p_{L0} = .0707$ for the log likelihood ratio statistic and $p_{Q0} = .0819$ for

the cumulative sum statistic. With the new approach of depth 3, respective p-values are undiscretized to $p_{L3} = .0565$ and $p_{Q3} = .0702$. The data indicate no change at the timepoint when the new procedure was introduced, but rather suggest multiple changes in care prior to observation 17. It is plausible that the new procedure in care was in part tested and used beforehand. Also, the preference of hospital based (and community practice) versus ambulant pin site care was changed in this period.

In publication bias example based on data from Zou et al. (2018), we found an estimated change point after observation 10 that was the drug *viibryd* approved on January 21, 2011, and observation 11 that is the drug *aubagio* approved on September 12, 2012 when using the likelihood ratio statistic, as displayed in Panel B in Figure 1. In contrast, the cumulative sum test statistic achieves the maximum after observation 7. We found only a small gain regarding the coarseness of the p-value from $p_{L0} = .234$ to $p_{L3} = .216$ for the two-sided likelihood ratio test statistic. However, if the *one-sided* likelihood ratio test is used for detection of an increasing event probability, the new test achieves a p-value of $p_{L^{1s}3} = .083$, while Worsley's p-value is $p_{L^{1s}0} = .117$. The delay of the change point since the FDAAA in 2007 is not implausible, because the drug approval process usually includes multiple trials, and thus, a delay of over 3 years is likely. Especially, negative findings in the development process will possibly lead to a longer delay.

8 | DISCUSSION

In this paper, we extended the proposal by Worsley by considering a sequential ordering to augment test statistics that compare "before" versus "after" by means to analyze 2×2 contingency tables. The ordering we defined originates from binary segmentation procedures, and to achieve an exact test, we first needed to derive the exact null distributions of the single steps of such procedures. With the "standard" approach not accounting for the conditional distributions, the type I error can be inflated. With the derived exact binary segmentation steps, however, a new test could be defined that is often able to attain a statistical power that is much closer to the randomized version of Worsley's test.

Another promising application of the described exact methods would be the usage in building decision or regression trees with binomial outcomes. When selecting input features, different variables repeatedly compete in being best suited to partition the predictor space into various strata. Here, the p-value can serve as a criterion for selection and the absence of statistical significance subsequently as a possible stopping criterion. In this context, the developed exact methods for binary segmentation steps are promising as they are rigorous. First, current methods do not adjust for any data splits (referred to as internal nodes) that have taken place in advance as *standard binary segmentation* steps do neither. Exact methods would increase validity and objectivity of the procedure. Second, when the explanatory variables are continuously split to create a multitude of strata (so-called tree branches), the sample sizes naturally get small. Increases in power similar to the test developed in Section 6 would be desirable. The simultaneous handling of many covariates in building decision trees is, however, not straightforward but will need some assumptions regarding their dependence structure.

ACKNOWLEDGMENTS

We thank Constance Zou, Joseph Ross, and colleagues for providing us with the data from their publication Zou et al. (2018) and fruitful discussions on the topic. Furthermore, we thank the Reviewers for their suggestions leading to an improved manuscript.

BL acknowledges support from grant number ES/L011859/1, from The Business and Local Government Data Research Centre, funded by the Economic and Social Research Council, UK, to provide researchers and analysts with secure data services.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

ORCID

David Ellenberger https://orcid.org/0000-0002-2274-5025
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES

Assareh, H., Smith, I., & Mengersen, K. (2015). Change point detection in risk adjusted control charts. *Statistical Methods in Medical Research*, *24*, 747–768.

Boulesteix, A.-L., & Strobl, C. (2007). Maximally selected chi-squared statistics and non-monotonic associations: An exact approach based on two cutpoints. *Computational Statistics & Data Analysis*, *51*, 6295–6306.

Brannath, W., Posch, M., & Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, *97*, 236–244.

Brostrom, G. (1997). A martingale approach to the changepoint problem. *Journal of the American Statistical Association*, *92*, 1177–1183.

Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, *16*, 188–197.

Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance*. New York, NY: Springer Science & Business Media.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-Von Mises tests. *The Annals of Mathematical Statistics*, *28*, 823–838.

Ellenberger, D., & Friede, T. (2016). An unconditional test for change point detection in binary sequences with applications to clinical registries. *Methods of Information in Medicine*, *55*, 367–372.

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *R Journal*, *2*, 53–58.

Fisher, R. A. (1932). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Frick, K., Munk, A., & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 495–580.

Friede, T., & Henderson, R. (2003). Intervention effects in observational survival studies with an application in total hip replacements. *Statistics in Medicine*, *22*, 3725–3737.

Friede, T., & Henderson, R. (2009). Exploring changes in treatment effects across design stages in adaptive trials. *Pharmaceutical Statistics*, *8*, 62–72.

Friede, T., Henderson, R., & Kao, C.-F. (2006). A note on testing for intervention effects on binary responses. *Methods of Information in Medicine*, *45*, 435–440.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, *42*, 2243–2281.

Gibbons, J. D. (1985). *Nonparametric Statistical Inference*, 2nd ed. Statistics: Textbooks and Monographs Vol. 65. New York and Basel: Marcel Dekker, Inc.

Halpern, A. L. (1999). Minimally selected p and other tests for a single abrupt changepoint in a binary sequence. *Biometrics*, *55*, 1044–1050.

Halpern, J. (1982). Maximally selected chi square statistics for small samples. *Biometrics*, *38*, 1017–1023.

Hinkley, D. V., & Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika*, *57*, 477–488.

Hirotsu, C. (1997). Two-way change-point model and its application. *Australian & New Zealand Journal of Statistics*, *39*, 205–218.

Hothorn, T., & Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, *43*, 121–137.

Hothorn, T., & Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics*, *64*, 1263–1269.

Kincaid, W. (1962). The combination of tests based on discrete distributions. *Journal of the American Statistical Association*, *57*, 10–19.

Lausen, B., Lerche, R., & Schumacher, M. (2002). Maximally selected rank statistics for dose-response problems. *Biometrical Journal*, *44*, 131–147.

Lausen, B., & Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics*, *48*, 73–85.

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, *55*, 1286–1290.

Mehrotra, D. V., Chan, I. S., & Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*, *59*, 441–450.

Miller, R., & Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics*, *38*, 1011–1016.

Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C*, *28*, 126–135.

Pettitt, A. (1980). A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika*, *67*, 79–84.

Ross, G. J., Tasoulis, D. K., & Adams, N. M. (2013). Sequential monitoring of a bernoulli sequence when the pre-change parameter is unknown. *Computational Statistics*, *28*, 463–479.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*, 127–141.

Scott, A. J., & Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, *30*, 507–512.

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *The Annals of Statistics*, *14*, 361–404.

Smith, A. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, *62*, 407–416.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, Jr., R. M. (1949). *The American Soldier: Adjustment during army life*. Princeton, NJ: Princeton University Press.

Venkatraman, E. (1992). *Consistency results in multiple change-point situations*. Technical Report 24. Stanford, CA: Stanford University.

Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, *259*, 270–274.

Worsley, K. (1983). The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, *70*, 455–464.

Zhang, A. D., Puthumana, J., Downing, N. S., Shah, N. D., Krumholz, H. M., & Ross, J. S. (2020). Assessment of clinical trials supporting US Food and Drug Administration approval of novel therapeutic agents, 1995-2017. *JAMA Network Open*, *3*, e203284.

Zhou, C., Zou, C., Zhang, Y., & Wang, Z. (2009). Nonparametric control chart based on change-point model. *Statistical Papers*, *50*, 13–28.

Zou, C. X., Becker, J. E., Phillips, A. T., Garritano, J. M., Krumholz, H. M., Miller, J. E., & Ross, J. S. (2018). Registration, results reporting, and publication bias of clinical trials supporting fda approval of neuropsychiatric drugs before and after fdaaa: A retrospective cohort study. *Trials*, *19*, 581.

Zou, C., Yin, G., Feng, L., & Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, *42*, 970–1002.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Ellenberger D, Lausen B, Friede T. Exact change point detection with improved power in small-sample binomial sequences. *Biometrical Journal*. 2020;1–17. https://doi.org/10.1002/bimj.201900273