OPEN ACCESS Freely available online

PLOS ONE

A Novel Method for Fast Change-Point Detection on Simulated Time Series and Electrocardiogram Data

Jin-Peng Qi[1,2]*, Qing Zhang[2], Ying Zhu[3], Jie Qi[1]

1 College of Information Science & Technology, Donghua University, Shanghai, P.R. China, **2** The Australia e-Health Research Centre, CSIRO, Brisbane, QLD, Australia, **3** Hunter New England Health, Royal North Shore Hospital, New South Wales, Australia

Abstract

Although Kolmogorov-Smirnov (KS) statistic is a widely used method, some weaknesses exist in investigating abrupt Change Point (CP) problems, e.g. it is time-consuming and invalid sometimes. To detect abrupt change from time series fast, a novel method is proposed based on Haar Wavelet (HW) and KS statistic (HWKS). First, the two Binary Search Trees (BSTs), termed TcA and TcD, are constructed by multi-level HW from a diagnosed time series; the framework of HWKS method is implemented by introducing a modified KS statistic and two search rules based on the two BSTs; and then fast CP detection is implemented by two HWKS-based algorithms. Second, the performance of HWKS is evaluated by simulated time series dataset. The simulations show that HWKS is faster, more sensitive and efficient than KS, HW, and T methods. Last, HWKS is applied to analyze the electrocardiogram (ECG) time series, the experiment results show that the proposed method can find abrupt change from ECG segment with maximal data fluctuation more quickly and efficiently, and it is very helpful to inspect and diagnose the different state of health from a patient's ECG signal.

Citation: Qi J-P, Zhang Q, Zhu Y, Qi J (2014) A Novel Method for Fast Change-Point Detection on Simulated Time Series and Electrocardiogram Data. PLoS ONE 9(4): e93365. doi:10.1371/journal.pone.0093365

Editor: Daniele Marinazzo, Universiteit Gent, Belgium

Received October 31, 2013; **Accepted** March 5, 2014; **Published** April 1, 2014

Funding: This paper is supported by National Natural Science Foundation of China (No.61104154), and the Fundamental Research Funds for the Central Universities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: qipengkai@dhu.edu.cn

Introduction

Detecting abrupt change from time series, called CP detection, has attracted considerable attention in the fields of data mining and statistics. CP detection [1,2,3,4,5] has been widely studied in many real-world problems, such as atmospheric and financial analysis [1], intrusion detection in computer networks [2], signal segmentation in data stream [3], as well as fault detection in engineering systems [4,5]. A good method of CP detection is by comparing probability distributions of time series samples over past and present intervals [6,7], in which a typical strategy is to trigger an alarm for a CP as two distributions are becoming significantly different. Various methods of change detection follow this statistical framework, including the CUSUM (cumulative sum) [7], the GLR (generalized likelihood ratio) [8,9] and the change finder [4]. Generally, these approaches are limited by relying on pre-specified parametric models such as probability density models, autoregressive models, and state-space models. Therefore, these methods tend to be less flexible in real-world CP detection problems.

In community of statistics, some non-parametric approaches for CP detection have been explored, in which non-parametric density estimation is used for calculating the likelihood ratio [10,11]. However, this kind of estimation is a hard problem [12,13], and may not be promising in practice. As a nonparametric method, the KS statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples [14,15]. Moreover, the KS test is for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution or that the sample is drawn from the reference distribution. The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. Recently, non-parametric KS statistic and its modified version are broadly investigated on several of application fields. For example, the use of the KS statistic for testing hypotheses regarding activation in blood oxygenation level-dependent functional MRI data [16]; modeling the cumulative distribution function of rub-induced AE signals and quantifying the goodness of fit with the KS statistic, to offer a suitable signal feature for diagnosis [17]; abrupt change point (CP) detection from electroencephalography signal (EEG) [18], and gene expression time series database [19].

On the other hand, Wavelet Transform (WT) is another promising approach for CP detection. In the past decade, WT approach has emerged as an important mathematical tool for analyzing time series [20,21,22,23,24]. It has found applications in anomaly detection, time series prediction, image processing, and noise reduction [20,24,25,26]. In particular, wavelets can represent general functions at different scales and positions in a versatile and sophisticated manner, so the data distribution features can be easily extracted from different time or space scales [26,27]. The heart of wavelet analysis is Multi-Resolution Analysis (MRA), by which a signal can be decomposed into sub-signals of different size

resolution levels [7,27,28]. The properties of wavelets such as localization, orthogonality, multi-rate filtering, are essential for analysis of non-stationary and transient signals. WT can represent a general function in terms of simple, fixed building blocks at different scales and positions. These building blocks are generated from a single fixed function called mother wavelet by translation and dilation operations [29,30]. In addition, Haar Wavelet (HW), as a simpler WT, owns some attracting features including fast for implementation and ability to analyze the local feature. HW is a very useful to find discontinuities and high frequency changes in time series, and a potential candidate in modern electrical and computer engineering applications, such as signal and image compression, as well as abnormality detection from time series [7,28].

However, most of these methods above are time-consuming, and not scalable to large-scale datasets due to their time complexity. Moreover, some of them, e.g., KS, is occasionally insensitive even invalid for less significant data fluctuation, especially when abrupt change occurs near two endpoint areas. To detect abrupt change from time series quickly and efficiently, a novel non-parametric method is proposed based on multi-level HW and a modified KS statistic. In this method, we combine superiorities of both KS statistic and HW methods, and try to find an abrupt change in terms of maximal data fluctuation existing between two adjacent segments of a diagnosed time series. This paper is organized as follows. Section II implements the integrated HWKS method in detail. First, the two BSTs termed TcA and TcD, are constructed by means of multi-level HW from a diagnosed time series. Then, the framework of HWKS method is implemented by introducing a modified KS statistic and two search criteria based on TcA and TcD. Last, two HWKS-based algorithms are designed to implement CP detection from the diagnosed time series. Section III evaluates the performance of HWKS by comparing KS, HW, and T methods, via simulated time series and real ECG datasets. Section IV gives conclusion from previous sections.

Method

The flow diagram of the integrated HWKS framework (Fig. 1) is composed of three parts. First, the two BSTs, TcA and TcD are constructed from a diagnosed time series. Second, abrupt CP is detected from root to leaf nodes of TcA in terms of a modified KS statistic and two search rules. Last, the performance of HWKS is evaluated by comparing with KS, HW, and T methods.

A. Construction of TcA and TcD

Like all wavelet transforms, multi-level HW decomposes a discrete signal into two sub-signals with half its length. One sub-signal is a running average or trend; the other sub-signal is a running difference or fluctuation. HW is performed in several stages or levels [28]. It can be described using scalar products with scaling signals and wavelets. The discrete signals are synthesized by beginning with a very low-resolution signal, successively adding on details to create higher resolution versions, and ending with a complete synthesis of the signal at the finest resolution. Generally, by using k-level HW, a discrete time series signal $Z=\{z_1, z_2,\ldots, z_N\}$, can be decomposed into the k^{th}-level trend cA^k, and k level fluctuations, i.e.,$\mathbf{c}D^1$, $cD^2,\ldots, cD^k$, $k=1, 2, .., \log_2 N$. As shown in Fig. 2, the k-level HW is the mapping H_k defined by [8],

$$Z \xrightarrow{H_k} (cA^k|cD^k|cD^{k-1}|...|cD^2|cD^1), \tag{1}$$

The multi-resolution analysis (MRA) is the heart of wavelet analysis [23,29], in terms of MRA, we can conceptualize the process of HW as a projection of time series with size $\mathcal{N}$ to total $\mathcal{N}$ different vectors v_i and w_i, termed as scaling signals and wavelet basis vectors, respectively. The discrete signal Z, average and detail signals are expressible as:

$$Z=A^k+\sum_{i=1}^{k} D^i, 1\le k\le \log_2 N, \tag{2}$$

$$A^k=(Z\cdot V^k)V^k=\sum_{i=1}^{N/2^k}(Z\cdot v_i^k)v_i^k=\sum_{i=1}^{N/2^k}(cA_{k,i})v_i^k \tag{3}$$

$$D^i=(Z\cdot W^i)W^i=\sum_{j=1}^{N/2^i}(cD_{k,j})w_j^i, \tag{4}$$

Thereafter, the following equations can be obtained,

$$\begin{aligned} Z&=A^k+\sum_{i=1}^{k} D^i\\ &=\sum_{i=1}^{N/2^k}(Z\cdot v_i^k)v_i^k+\sum_{i=1}^{k}\sum_{j=1}^{N/2^i}(Z\cdot w_j^i)w_j^i\\ &=\sum_{i=1}^{N/2^k}(cA_{k,i})v_i^k+\sum_{i=1}^{k}\sum_{j=1}^{N/2^i}(cD_{i,j})w_j^i\\ &=cA^k\cdot V^k+\sum_{j=1}^{k} cD^j\cdot W^j, \end{aligned} \tag{5}$$

$$\begin{aligned} A^k&=cA^k\cdot V^k\\ &=\left(cA_{k,1},cA_{k,2},\ldots,cA_{k,N/2^k}\right)\cdot\left(v_1^k,v_2^k,\ldots,v_{N/2^k}^k\right),\\ &=(a_1^k,a_2^k,\ldots,a_{N-1}^k,a_N^k,) \end{aligned} \tag{6}$$

$$\begin{aligned} D^k&=cD^k\cdot W^k\\ &=\left(cD_{k,1},cD_{k,2},\ldots,cD_{k,N/2^k}\right)\cdot\left(w_1^k,w_2^k,\ldots,w_{N/2^k}^k\right),\\ &=(d_1^k,d_2^k,\ldots,d_{N-1}^k,d_N^k,) \end{aligned} \tag{7}$$

where v_i^k is k-level Haar scaling signals, w_j^k is k-level Haar wavelets, $|v_i^k| = |w_j^k| = \mathcal{N}$.

In addition, we can represent HW with k-level approximation and detail coefficient vectors by the following matrices, namely McA and McD:

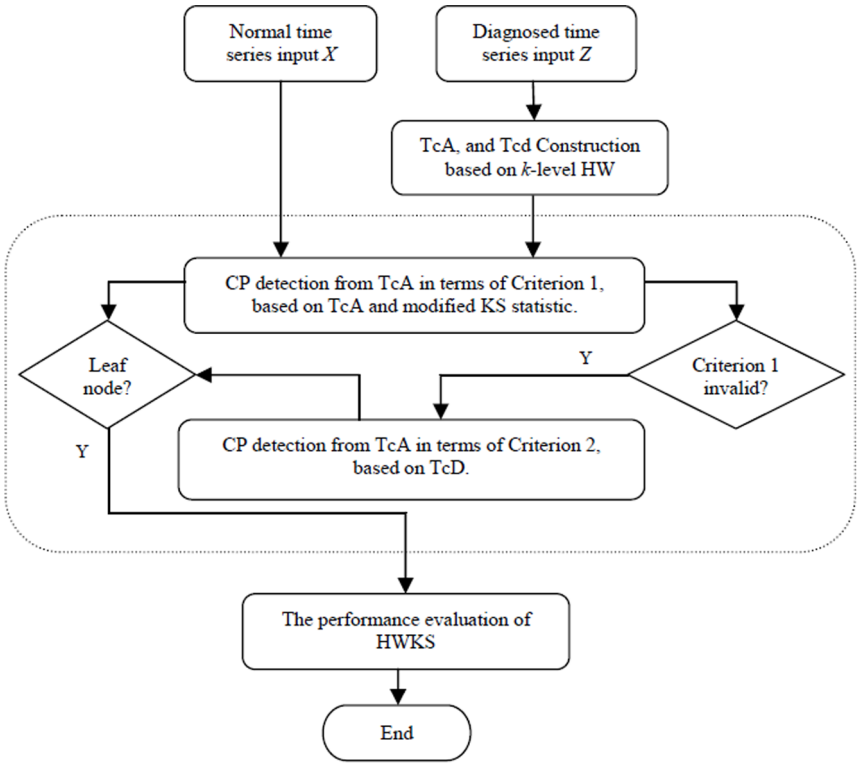

Figure 1. The integrated scheme of HWKS method for fast CP detection, which includes three parts: construction of two BSTs, namely TcA and TcD; CP detection of HWKS in terms of two search criteria; and evaluation of HWKS method.
doi:10.1371/journal.pone.0093365.g001

$$\mathrm{McA}=\begin{bmatrix} \mathrm{cA}_{1,1} & \ldots & \mathrm{cA}_{1,N/2} \\ \ldots & \mathrm{cA}_{k,j} & 0 \\ \mathrm{cA}_{m,1} & 0 & 0 \end{bmatrix},$$
$$\mathrm{McD}=\begin{bmatrix} \mathrm{cD}_{1,1} & \ldots & \mathrm{cD}_{1,N/2} \\ \ldots & \mathrm{cD}_{k,j} & 0 \\ \mathrm{cD}_{m,1} & 0 & 0 \end{bmatrix}, \quad (8)$$

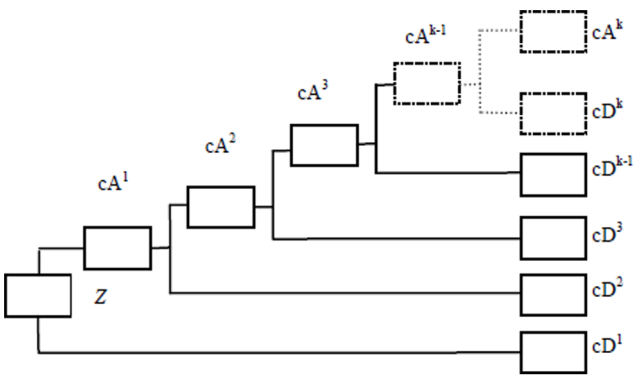

Figure 2. The diagram of multi-level HW for time-series signal *Z*, it is composed of *k*-level cA and cD vectors, i.e., the average and difference coefficients vectors.
doi:10.1371/journal.pone.0093365.g002

where $0 \le k \le m = \log_2 N$, $1 \le j \le N/2^k$. Suppose the size of a diagnosed sample Z is divisible k times by 2, we can further denote the j^{th} element in cA^{k} and corresponding averaged signal in A^{k}, as well as the j^{th} element in cD^{k} and corresponding detail signal in D^{k} as follows:

$$\mathrm{cA}_{\mathrm{k,j}} = \frac{1}{(\sqrt{2})\hat{}k}\left(\sum_{i=a}^{b} z_i\right), \quad (9)$$

$$\mathrm{a}_{\mathrm{k},a} = \ldots = \mathrm{a}_{\mathrm{k},b} = \frac{1}{(\sqrt{2})}\mathrm{cA}_{\mathrm{k,j}}, \quad (10)$$

$$\mathrm{cD}_{\mathrm{k,j}} = \frac{1}{(\sqrt{2})\hat{}k}\left(\sum_{L=a}^{c} z_L - \sum_{R=c+1}^{b} z_R\right), \quad (11)$$

$$\mathrm{d}_{\mathrm{k,a}} = \ldots = \mathrm{d}_{\mathrm{k,b}} = \frac{1}{(\sqrt{2})}\mathrm{cD}_{\mathrm{k,j}}, \quad (12)$$

where $2 \le k \le \log_2 N$, and $2^k(j-1)+1 \le i \le j * 2^k$; $a = 2^k(j-1)+1$, $c = 2^k(j-1)+2^{(k-1)}$, and $b = 2^k * j$. Therefore, a diagnosed Z can be decomposed into cA and cD matrices by means of k-level HW. Thereafter, as shown in Fig. 3, TcA and TcD are built in terms of McA and McD, as well as original elements in $Z = \{z_1, z_2, \ldots, z_N\}$. *In TcA, and TcD, non-leaf nodes in different level are constructed from McA, and McD, respectively; and then leaf nodes are derived*

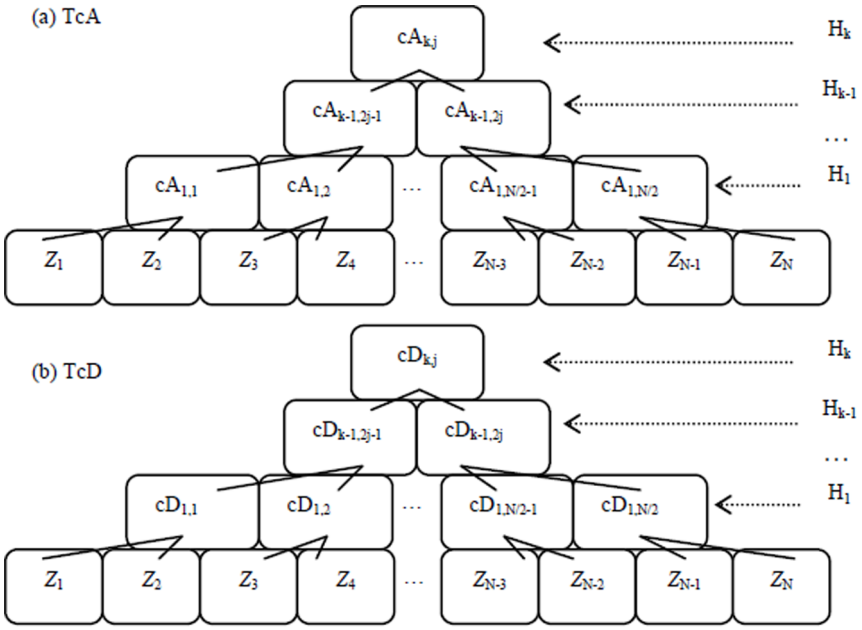

Figure 3. The diagrams of TcA and TcD, derived from a diagnosed time-series Z by means of k-level HW.
doi:10.1371/journal.pone.0093365.g003

directly from the elements in Z. The constructions of TcA and TcD are implemented by Algorithm 1 in detail.

Algorithm 1.

Input: $Z=\{z_i : 1 \le i \le \mathcal{N}\}$,a diagnosed time series

Output: two BSTs, TcA and TcD

Initiate the level of HW, k, ($1 \le k \le \log2(\mathcal{N})$);

Declare two matrices, McA and McD;

For $i=1$ to k do

$[cA_i, cD_i]$ = Call HW(Z, i);

$McA(i) = cA_i$; $McD(i) = cD_i$;

end

Construct TcA and TcD from McA and McD, as well as Z.

Output TcA and TcD

B. HWKS method based on a modified KS statistic

KS test is one of the most useful and general non-parametric methods for comparing two samples, because it is sensitive to those differences in both location and shape of the empirical cumulative distribution functions (e.c.d.f) of two samples [31,32]. Suppose $Y=\{y_1,...,y_N\}$ is a time series, we observe,

$$Y=f(i/N)+X, \quad i=1,...,N,$$

where $X=\{x_i\}_{i=1,...,N}$ are discrete and centred i.i.d. random variables, and f is a noisy signal with unknown distribution. Thereafter, we can deal X as a normal time series with distribution function $F_m(x)$, and Y as an abnormal time series with distribution function $G_n(x)$. Then, we can assemble a diagnosed time series $\mathcal{Z}$ and define it as below:

$$Z=\{X,Y\}=\{Z_1,Z_2\}=\{z_1,\ldots,z_c,z_{c+1},\ldots,z_n\}, \quad (13)$$

To detect an abrupt CP from Z, a modified KS statistic is defined to evaluate the distribution distance between X and Z [15,33,34]:

$$D_{mn}(x)\underline{\Delta}\left(\frac{mn}{m+n}\right)^{1/2}\sup_{x\in\mathrm{R}}|G_n(x)-F_m(x)|, \quad (14)$$

if a change point c occurs in Z, there exists a value z_c satisfies $F_m(z_c)\neq G_n(z_c)$, and $\mathrm{D_{mn}}(z_c)>\delta$, $z_c\in[z_1, z_n]$, $\delta\in R$.

As hypothesized $F_m(x)$ and $G_n(x)$ are not available, but instead, the e.c.d.f of $F_m(x)$ and $G_n(x)$ can be derived from two time series X and Z. Then, $F_m(x)$ and $G_n(x)$ are defined by,

$$F_m(x)=P_m(X\le x)=\frac{1}{m}\sum_{i=1}^{m}I(x_i\le x), \quad (15)$$

$$G_n(x)=P_n(Z\le x)=\frac{1}{n}\sum_{j=1}^{n}I(z_j\le x), \quad (16)$$

where $F_m(x)$ and $G_n(x)$ count the proportion of the sample points below level x. For any fixed point $x\in\mathrm{R}$, the law of large numbers implies that

$$F_m(x)=\frac{1}{m}\sum_{i=1}^{c}I(x_i\le x)\rightarrow \mathrm{E}\ I(x_i\le x)=F(x), \quad (17)$$

$$G_n(x)=\frac{1}{n}\sum_{j=1}^{n}I(z_j\le x)\rightarrow \mathrm{E}\ I(z_j\le x)=G(x), \quad (18)$$

where $F(x)$ and $G(x)$ are true underlying distribution of two time series X and $\mathcal{Z}$, i.e. the proportion of the sample in the set $(-\infty,x]$ approximates the probability of this set. It is easy to say that this approximation holds uniformly over all $x\in\mathrm{R}:\sup_{x\in\mathrm{R}}|G_n(x)-F_m(x)|\rightarrow 0$, *i.e.*, the largest difference between $F_m(x)$ and $G_n(x)$ goes to 0 in probability. The key observation in KS test is that the

distribution of this supremum does not depend on the 'unknown' distribution P in diagnosed Z, if P is continuous distribution. In addition, we have,

$$\begin{aligned} &\mathrm{P}((\frac{mn}{m+n})^{1/2}\sup_{x\in\mathrm{R}}|G_n(x)-F_m(x)|\le\delta)\to \\ &H(\delta)=1-2\sum(-1)^{i-1}e^{-2i^2\sigma}, \end{aligned} \tag{19}$$

where $H(\sigma)$ is the c.d.f. of KS distribution [35].

In terms of the definition of non-leaf nodes in TcA in equations (9), we can denote the j^{th} element z_j of Z and define a new element $z_{k,j}$ used in HWKS method as,

$$z_j=a_{k,j}+\sum_{i=1}^{\mathrm{k}}d_{i,j}\Rightarrow z_j\ge a_{k,j}, \tag{20}$$

$$\begin{aligned} z_{k,j}&=\frac{1}{2^k}(\sum_{i=a}^{b}z_i)=\frac{(\sqrt{2})\hat{}k}{2^k}\mathrm{cA_{k,j}} \\ &=\frac{(\sqrt{2})\hat{}(k+1)}{2^k}\mathrm{a_{k,a}}=\ldots=\frac{(\sqrt{2})\hat{}(k+1)}{2^k}\mathrm{a_{k,b}}\le z_{\mathrm{a}},\ldots,z_{\mathrm{b}}, \end{aligned} \tag{21}$$

where $\mathrm{cA_{k,j}}$ is a non-leaf node of TcA, $z_{k,j}$ is a new element defined in terms of $\mathrm{cA_{k,j}}$, and $1\le k\le\log_2 N$, $1\le j\le N/2^k$; $a=2^k(j-1)+1$, and $b=2^k*j$. Then, a revised KS statistic for HWKS is defined as:

$$\begin{aligned} &D'_{mn}(k,j)\underline{\Delta}(\frac{mn}{m+n})^{1/2}\sup_{z_{k,j}\in R}|G_n(z_{k,j})-F_m(z_{k,j})| \\ &=(\frac{nm}{n+m})^{1/2}\sup_{z_{k,j}\in R}\left|\{\frac{1}{n}(\sum_{j=1}^{n}I(z_j\le z_{k,j}))-\frac{1}{m}\sum_{i=1}^{m}I(x_i\le z_{k,j})\}\right|, \end{aligned} \tag{22}$$

where $D'_{mn}(k,j)$ measures the distribution distance between X and Z at a selected node $\mathrm{cA_{k,j}}$ in TcA, and larger value of $D'_{mn}(k,j)$ means that more significant change occurs in Z.

Thereafter, we define a KS test for X and Z as,

$$H_0: F_m=G_n \text{ vs. } H_0: F_m\ne G_n, \tag{23}$$

if null hypothesis is true then, the distribution of D'_{mn} can be tabulated as it depends only on n. Moreover, if n is large enough then the distribution of D'_{mn} is approximated by KS distribution. On the other hand, suppose $F_m\ne G_n$, since F_m and G_n are the true c.d.f. of X and Z, bylaw of large numbers the e.c.d.f, F_m will converge to G_n, and for large n we will have,

$$\sup_{z_{k,j}\in R}|G_n(z_{k,j})-F_m(z_{k,j})|>\delta, D'_{mn}>(\frac{mn}{m+n})^{1/2}\delta, \tag{24}$$

If H_0 fails then,

$$D'_{mn}>(\frac{mn}{m+n})^{1/2}\delta\to+\infty \text{ as } n\to+\infty. \tag{25}$$

Therefore, to test H_0 we make a detection rule,

$$\delta=\begin{cases} H_0: D'_{mn}\le c \\ H_1: D'_{mn}>c \end{cases}, \tag{26}$$

where c depends on the level of significance α, and can be found by using KS distribution when n is large.

$$\alpha=\mathrm{P}(\delta\ne H_0|H_0)=\mathrm{P}(D'_{mn}\ge c|H_0)\approx 1-H(c), \tag{27}$$

If $\sup_k\sup_j|D'_{mn}(k,j)|\le C(\alpha)$, H_0 is true i.e., no change point occurs. On the other hand, if $\sup_k\sup_j|D'_{mn}(k,j)|>C(\alpha)$, then hypothesis H_1 is true *i.e.*, an abrupt change is detected.

C. HWKS-based CP Detection

To detect abrupt change from diagnosed Z, an optimal path needs to be obtained from root to leaf nodes in TcA accurately and quickly. Therefore, as shown in Fig. 4, two search criteria are introduced in terms of TcA and TcD. The first search criterion is defined based on TcA as follows:

Criterion 1. Suppose the current non-leaf node in TcA we selected is $\mathrm{cA_{k,j}}$, and its left-child and right-child nodes are $\mathrm{cA_{k\text{-}1,2j\text{-}1}}$ and $\mathrm{cA_{k\text{-}1,2j}}$, respectively,

(a) if $(D'_{mn}(k-1,2j-1)>D'_{mn}(k-1,2j)$ and $D'_{mn}(k-1,2j-1)>C(\alpha))$ hold true, then the left-child node $\mathrm{cA_{k\text{-}1,2j\text{-}1}}$ is selected to be involved into the current search path in TcA;

(b) if $(D'_{mn}(k-1,2j-1)<D'_{mn}(k-1,2j)$ and $D'_{mn}(k-1,2j)>C(\alpha))$ hold true, then the right-child node $\mathrm{cA_{k\text{-}1,2j\text{-}1}}$ is selected to be involved into the current search path in TcA.

Proof. In terms of the definitions of $z_{k,j}$ and D'_{mn} in equation (21) and (22), $D'_{mn}(k-1,2j-1)$, and $D'_{mn}(k-1,2j)$ can be written as

$$\begin{aligned} &D'_{mn}(k-1,2j-1)\underline{\Delta}(\frac{nm}{n+m})^{1/2}|G_n(z_{k-1,2j-1})-F_m(z_{k-1,2j-1})| \\ &=(\frac{nm}{n+m})^{1/2}\left|\{\frac{1}{n}(\sum_{a=1}^{n}I(z_a\le z_{k-1,2j-1}))-\frac{1}{m}\sum_{b=1}^{m}I(x_b\le z_{k-1,2j-1})\}\right|, \\ &=(\frac{nm}{n+m})^{1/2}\left|\{\frac{1}{n}(\sum_{a=1}^{n}I(z_a\le\frac{(\sqrt{2})\hat{}(k-1)}{2^k}\mathrm{cA_{k\text{-}1,2j\text{-}1}}))-\frac{1}{m}\sum_{b=1}^{m}I(x_b\le\frac{(\sqrt{2})\hat{}(k-1)}{2^k}\mathrm{cA_{k\text{-}1,2j\text{-}1}})\}\right| \end{aligned} \tag{28}$$

$$\begin{aligned} &D'_{mn}(k-1,2j)\underline{\Delta}(\frac{nm}{n+m})^{1/2}|G_n(z_{k-1,2j})-F_m(z_{k-1,2j})| \\ &=(\frac{nm}{n+m})^{1/2}\left|\{\frac{1}{n}(\sum_{a=1}^{n}I(z_a\le z_{k-1,2j}))-\frac{1}{m}\sum_{b=1}^{m}I(x_b\le z_{k-1,2j})\}\right|, \\ &=(\frac{nm}{n+m})^{1/2}\left|\{\frac{1}{n}(\sum_{a=1}^{n}I(z_a\le\frac{(\sqrt{2})\hat{}(k-1)}{2^k}\mathrm{cA_{k\text{-}1,2j}}))-\frac{1}{m}\sum_{b=1}^{m}I(x_b\le\frac{(\sqrt{2})\hat{}(k-1)}{2^k}\mathrm{cA_{k\text{-}1,2j}})\}\right| \end{aligned} \tag{29}$$

where $z_{k-1,2j-1}$ and $z_{k-1,2j}$ are two diagnosed points in accordance with $\mathrm{cA_{k\text{-}1,2j\text{-}1}}$ and $\mathrm{cA_{k\text{-}1,2j}}$ in TcA. In terms of Criterion 1, if $(D'_{mn}(k-1,2j-1)>D'_{mn}(k-1,2j)$ holds true, as plotted in Fig. 5, it indicates that more significant distribution distance exists in the left sub-tree covered by $\mathrm{cA_{k\text{-}1,2j\text{-}1}}$, than in the right one covered by $\mathrm{cA_{k\text{-}1,2j}}$, and vice versa. That is, abrupt CP occurs in the left segment of Z with more probability than in the right one. On the other hand, If $D'_{mn}(k-1,2j-1)>C(\alpha))$ is satisfied, it means that the distribution distance overtakes a critical value given in an identical data distribution.

Criterion 1 guarantees that if abrupt CP occurs in Z, the left or the right sub-tree with bigger distribution distance is selected to be involved into the current search path, and then the other half part

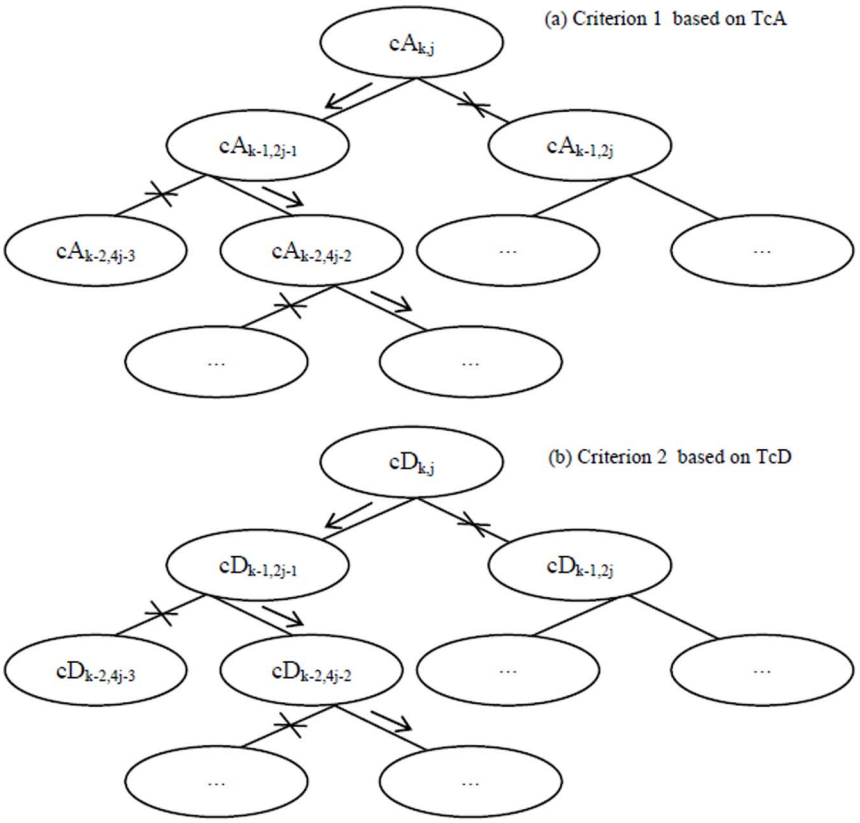

Figure 4. The scheme of two search criteria for CP detection of HWKS. (a) Criterion1 based on TcA, and (b) Criterion 2 based on TcD ensure that an optimal path of abrupt CP can be detected from root to leaf nodes in TcA.
doi:10.1371/journal.pone.0093365.g004

is discarded. Thereafter, an optimal search path can be detected from root to leaf nodes in TcA after about $log_2(n)$ search steps. Unfortunately, if $(D'_{mn}(k-1,2j-1)==D'_{mn}(k-1,2j)$ or $(\max(D'_{mn}(k-1,2j-1),D'_{mn}(k-1,2j))<C(\alpha))$ are true, then Criterion 1 is invalid for CP detection. Therefore, another search criterion is introduced based on TcD, and defined as below.

Criterion 2. Suppose $D'_{mn}(k-1,2j-1)==D'_{mn}(k-1,2j)$ or $(\max(D'_{mn}(k-1,2j-1),D'_{mn}(k-1,2j))<C(\alpha))$ is satisfied, the non-leaf node $cD_{k,j}$ in TcD is selected, in accordance with the current non-leaf node $cA_{k,j}$ in TcA, with its left-child node $cD_{k-1,2j-1}$ and right-child node $cD_{k-1,2j}$, respectively,

(a) if $|cD_{k-1,2j-1}|>|cD_{k-1,2j}|$ holds true, then the left-child node $cA_{k-1,2j-1}$ is selected to be involved into the current search path in TcA;

(b) if $|cD_{k-1,2j-1}|<|cD_{k-1,2j}|$ holds true, then the right-child node $cA_{k-1,2j}$ is selected to be involved into the current search path in TcA.

Proof. In accordance with the definition of $cD_{k,j}$ in equation (11), $cD_{k-1,2j-1}$ and $cD_{k-1,2j}$ can be written by

$$cD_{k-1,2j-1}=\frac{1}{m}\Big(\sum_{La=L_0}^{L_1} z_{La}-\sum_{Lb=L_1+1}^{L_2} z_{Lb}\Big)=D_L(k-1,2j-1), \quad (30)$$

$$cD_{k-1,2j}=\frac{1}{n}\Big(\sum_{Ra=R_0}^{R_1} z_{Ra}-\sum_{Rb=R_1+1}^{R_2} z_{Rb}\Big)=D_R(k-1,2j), \quad (31)$$

where $L_0=2^k(j-1)+1$, $L_1=2^k(j-1)+2^{(k-2)}$, $L_2=2^{(k-1)}(2j-1)$; $R_0=2^{(k-1)}(2j-1)+1$, $R_1=2^{(k-1)}(2j-1)+2^{(k-2)}$, $R_2=2^{(k-1)}(2j)$, and $m=n=(\sqrt{2})^{(k-1)}$. In terms of equation (30), and (31), D_L , and D_R can reflect data fluctuation of two segments, namely Z_L and Z_R in Z covered by $cD_{k-1,2j-1}$ and $cD_{k-1,2j}$, respectively. None loses of generalization, if $|cD_{k-1,2j-1}|>|cD_{k-1,2j}|$ is true, it means that bigger data fluctuation exists in Z_L than in Z_R. That is, a potential abrupt change exists in Z_L covered by $cD_{k-1,2j-1}$ with more probability than in Z_R covered by $cD_{k-1,2j}$. Therefore, in terms of criterion 2, left-child node $cA_{k-1,2j-1}$ is selected to be involved into current search path in TcA, and vice versa.

In terms of two search criteria above, two HWKS-based algorithms are implemented to detect abrupt change from a diagnosed time series Z. In Algorithm 2, the distribution distance between X and Z is calculated in terms of a selected non-leaf node $cA_{k,j}$ in TcA. In this function, the confidence interval is set by $\alpha=0.05$, and $C(\alpha)=1.3258$. For simplicity, we only output the nodes that the value of D'_{mn} overtakes $C(\alpha)$, otherwise output zero. In Algorithm 3, an optimal path is detected from root to leaf nodes in TcA in terms of two search criteria, and then an estimated CP is obtained from a

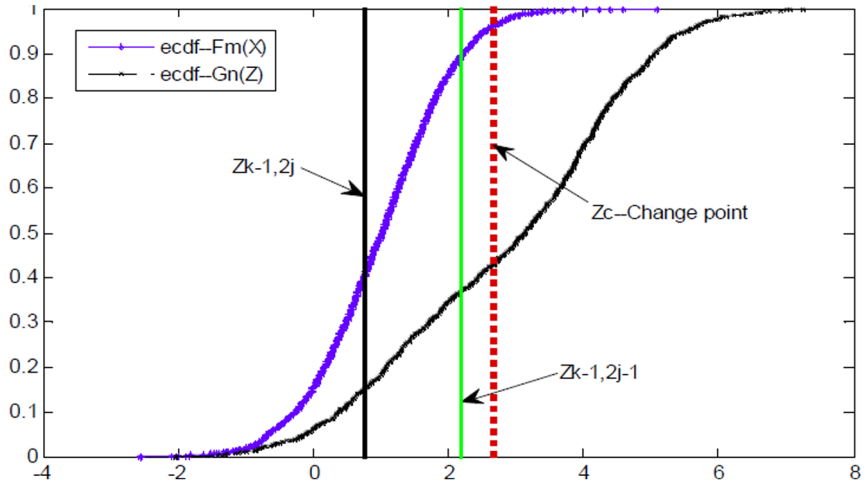

Figure 5. The scheme of search strategy in Criterion 1 in terms of distribution distance between *X* and *Z*. The dotted red line refers to a supposed CP, namely Z_c, the solid black line, and green one stands for the points of $z_{k\text{-}1,2j}$, and $z_{k\text{-}1,2j\text{-}1}$, respectively.
doi:10.1371/journal.pone.0093365.g005

Algorithm.2.
Input: a normal time series input X, and a diagnosed time series $\mathcal{Z}$, as well as current non-leaf node $cA_{k,j}$ selected in TcA.
Output: RS, the distribution distance between X and $\mathcal{Z}$ at the selected non-leaf node $cA_{k,j}$ in TcA.
Set $Ca = 1.3258$;
RS = 0;
Calculate $z_{k,j}$ in terms of $cA_{k,j}$;
Sx = Calculate the e.c.d.f of $z_{k,j}$ in X;
Sz_1 = Calculate the e.c.d.f of $z_{k,j}$ in Z;
Sz_2 = Calculate the e.c.d.f of $z_{k,j}$ in Z;
D1 = abs(Sz_1-Sx); D2 = abs(Sz_2-Sx);
If (D1>D2) &&(D1>*Ca*) Then
{Select D1;
RS = D1;}
elseif (D1<D2) &&(D2>*Ca*) Then
{Select D2;
RS = D2;}
elseif (D1 = = D2) || (max (D1, D2)<*Ca*) Then
RS = 0;
Endif
Output RS($cA_{k,j}$);

Algorithm.3.
Input: X, $\mathcal{Z}$, TcA, and TcD derived from $\mathcal{Z}$.
Output: The estimated abrupt CP from TcA and TcD.
Set $b = 1$; N = length (z); $k = \log_2(N)$;
Set the first node of optimal search path is the root node in TcA:
For $i = 1$ to k do
$a = k\text{-}i+1$;
Call Algorithm 2 to calculate the distribution distance between X and $\mathcal{Z}$ at two non-leaf node $cA_{a\text{-}1,2b\text{-}1}$ and $cA_{a\text{-}1,2b\text{-}1}$, respectively;
Set S1 = RS($cA_{a\text{-}1,2b\text{-}1}$), S2 = RS($cA_{a\text{-}1,2b}$);
If (S1>S2) Then
{The current selected node = $cA_{a\text{-}1,2b\text{-}1}$;
$b = 2b\text{-}1$;}
elseIf ((S1<S2) Then
{The current selected node = $cA_{a\text{-}1,2b}$;
$b = 2b$;}
elseIf ((S1 = = S2) Then
{If($cD_{a\text{-}1,2b\text{-}1}$> $cD_{a\text{-}1,2b}$) Then
{ The current selected node = $cA_{a\text{-}1,2b\text{-}1}$;
$b = 2b\text{-}1$; } endif
If($cD_{a\text{-}1,2b\text{-}1}$< $cD_{a\text{-}1,2b}$) Then
{ The current selected node = $cA_{a\text{-}1,2b}$;
$b = 2b$; } endif }
} Endif
End for
Output b, and z_b

diagnosed Z. The pseudocodes can be found in Algorithm 2 and 3 in detail.

D. Evaluation of HWKS

Many methods have been proposed for CP detection. In this part, the following typical methods are used to evaluate and verify the performance of the proposed HWKS method.

KS statistic [15]. In KS method, firstly, we divide a diagnosed sample data Z into two segments, namely, $Z_m = \{z_1, z_2, \ldots, z_m\}$, and $z_{N\text{-}m} = \{z_{m+1}, z_{m+2}, \ldots, z_{N\text{-}m}\}$. Then, KS statistic for these two segments is defined by,

$$D_{mn}(x) \underline{\Delta} \left(\frac{m(N-m)}{N}\right)^{1/2} \sup_{x \in \mathbf{R}} |F_{N-m}(x) - F_m(x)|$$
$$= \left(\frac{m(N-m)}{N}\right)^{1/2} \sup_{x \in \mathbf{R}} \left| \sum_{i=m+1}^{N-m} I(Z_i < x) - \sum_{j=1}^{m} I(Z_j < x) \right|, \quad (32)$$

where N is the size of the diagnosed sample Z; m is the size of segments $\mathcal{Z}_m$, that is the current diagnosed position in Z.

HW [28,33]. In this method, the fluctuation coefficient vector $cD^1 = \{cD_{1,1}, cD_{1,2}, \ldots, cD_{1,N/2},\}$ is calculated by one-level HW

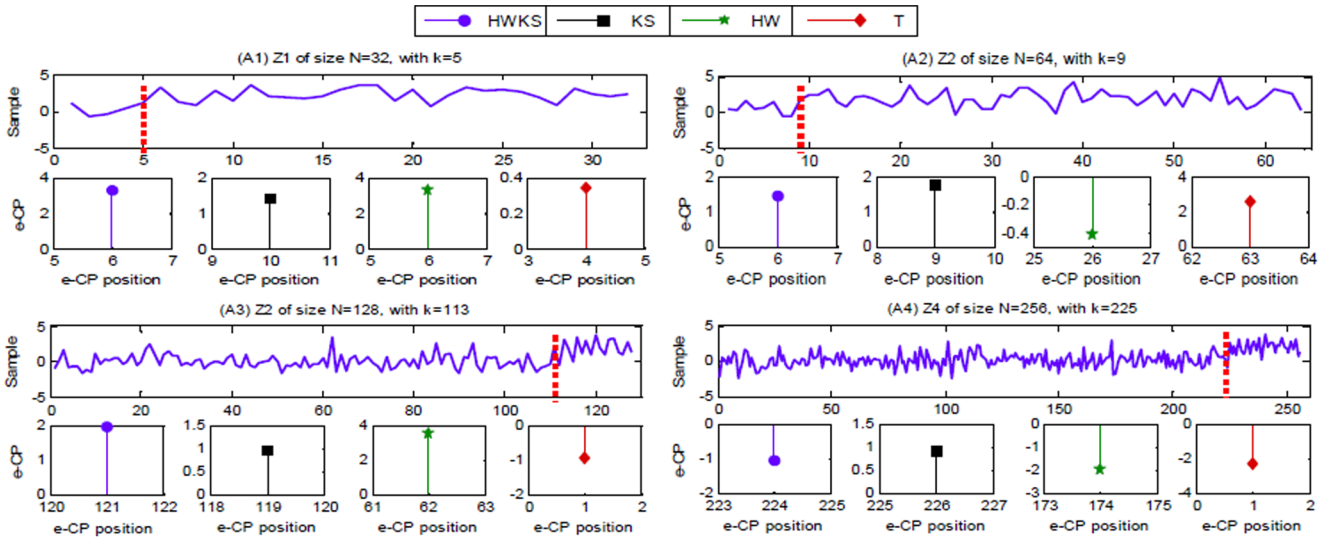

Figure 6. The results of single simulation on single CP test position, with constant variance $v=2$, different sample size N, and CP test position k, by HWKS, KS, HW, and T, respectively. (A1) The estimated CP from diagnosed sample $Z1$, with $N=32$, $k=5$. (A2) The estimated CP from diagnosed sample $Z2$, with $N=64$, $k=9$. (A3) The estimated CP from diagnosed sample $Z3$, with $N=128$, $k=113$. (A4) The estimated CP from diagnosed sample $Z4$, with $N=256$, $k=225$.
doi:10.1371/journal.pone.0093365.g006

from a diagnosed time series Z, and then an estimated CP can be found by comparing the values of elements in vector cD^1. The elements in cD^1 is defined as,

$$cD_{1,j}=\frac{1}{\sqrt{2}}\Big(\sum_{i=2(j-1)+1}^{j*2}(-1)^{i+1}z_i\Big), \tag{33}$$

where $1\le j\le N/2$. To find an abrupt change, a critical value δ is given in terms of an identical data distribution. If $\max_{1\le j\le N/2}|cD_{1,j}|>\sqrt{2}\cdot\delta$ holds true, then an abrupt CP occurs in a diagnosed sample Z.

T-statistic[36]. T, also known as Welch'st-test, is used only when two population variances are assumed different (the two sample sizes may or may not be equal) and hence must be estimated separately. A diagnosed sample Z is divided into $Z_m=\{z_1, z_2,\ldots, z_m\}$ and $Z_{N\text{-}m}=\{z_{m+1}, z_{m+2},\ldots, z_{N\text{-}m}\}$. Then, T statistic is calculated as,

$$t=\frac{\bar{Z}_m-\bar{Z}_{N-m}}{s_{\bar{Z}_m-\bar{Z}_{N-m}}}, s_{\bar{Z}_m-\bar{Z}_{N-m}}=\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{N-m}}, \tag{34}$$

where $\bar{Z}_m$, and $\bar{Z}_{N-m}$ are the sample mean of two segments in diagnosed Z, S_* is unbiased estimator of standard deviation, and m, n is the size of two segments in Z.

Results and Discussion

First, we evaluate the performance of HWKS on the simulated time series datasets, the sensitivity, efficiency, and accuracy of HWKS is analyzed by comparing KS, HW, and T methods. Then, we apply HWKS, and other three methods, to distinguish normal and abnormal ECG segments from the assembled ECG time series samples, and diagnose the different states of health from a patient's abnormal ECG time series segments.

Table 1. The summarized results of single simulation with single CP test position.

M \ Z	Size, N	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	Averaged
	CP, k	2	3	5	9	113	225	449	897	
HWKS	*Err*	*0*	*0*	*1*	*−3*	*8*	*−1*	*17*	*−10*	*5*
	Acc	*1.0*	*1.0*	*.97*	*.95*	*.94*	*.99*	*.97*	*.99*	*.97*
KS	Err	−1	−2	5	0	6	1	−4	−9	3.5
	Acc	.88	.88	.84	1.0	.95	.99	.99	.99	.94
HW	Err	0	1	1	17	−51	−51	−93	−503	89
	Acc	1.0	.94	.97	.73	.60	.80	.82	.51	.79
T	Err	−1	−2	−1	54	−112	−224	62	126	72
	Acc	.88	.88	.97	.16	.13	.13	.88	.88	.61

doi:10.1371/journal.pone.0093365.t001

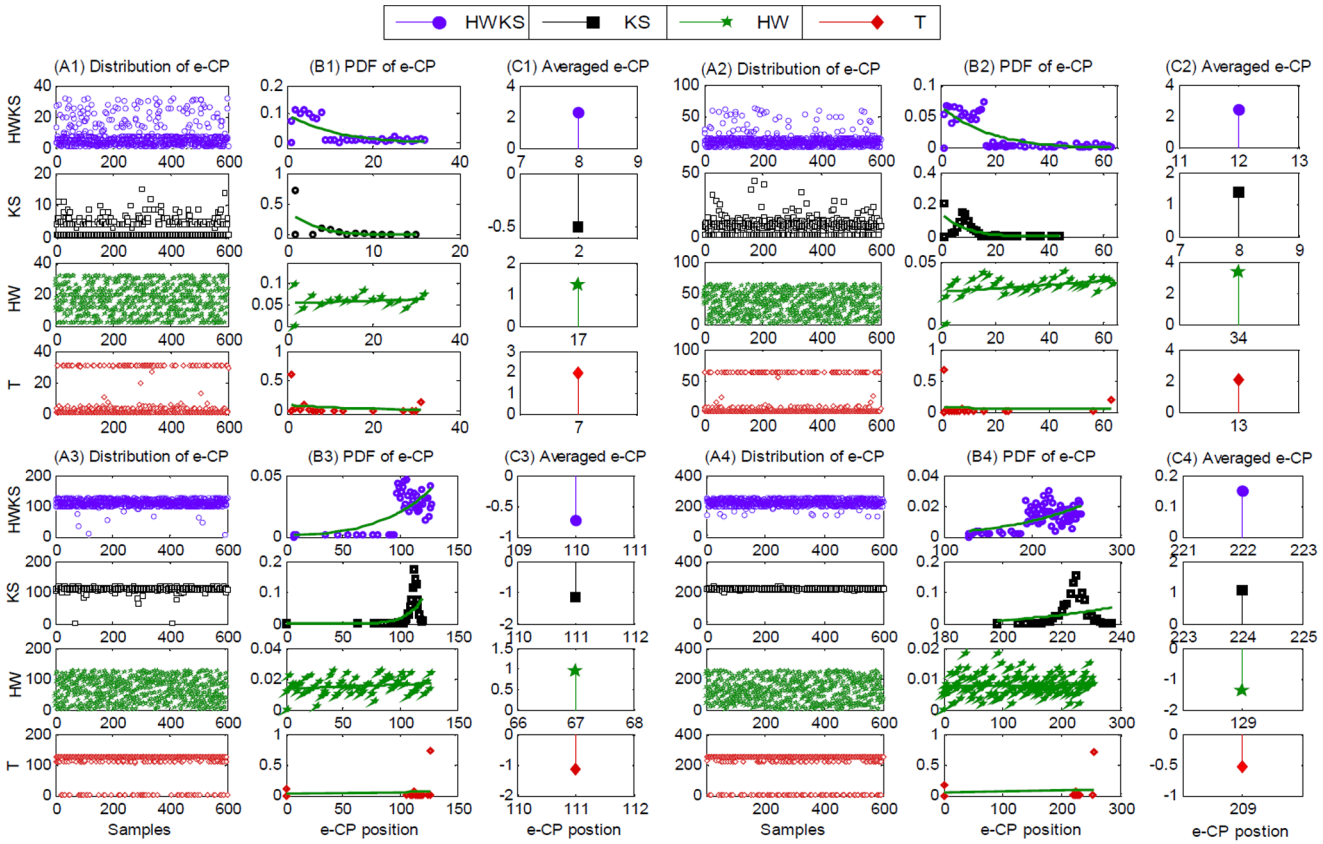

Figure 7. The results of multiple 600 simulations on single CP test position, with $v=2$, different N, and k, by HWKS, KS, HW, and T, respectively. For samples with $N=32$, $k=5$; $N=64$, $k=9$; $N=128$, $k=113$; and $N=256$, $k=225$, (A1)–(A4) the distribution of e-CP, (B1)–(B4) the PDF of e-CP, and (C1)–(C4) the averaged e-CP, by HWKS, KS, HW, and T, respectively.
doi:10.1371/journal.pone.0093365.g007

Table 2. The summarized results of multiple 600 simulation on single CP test position.

M \ Z	Size, N / CP, k	2^3 / 2	2^4 / 3	2^5 / 5	2^6 / 9	2^7 / 113	2^8 / 225	2^9 / 449	2^{10} / 897	Averaged
HWKS	***Tim***	*.17*	*.23*	*.27*	*.33*	*.41*	*.47*	*.60*	*.75*	*.40*
	Hit	*.39*	*.15*	*.10*	*.06*	*.03*	*.03*	*.01*	*.01*	*.10*
	Err	*2*	*2*	*3*	*3*	*−3*	*−3*	*−5*	*3*	*3*
	Acc	*.75*	*.88*	*.91*	*.95*	*.98*	*.99*	*.99*	*.99*	*.93*
KS	Tim	.09	.15	.36	.68	1.6	3.4	9.05	23.8	4.9
	Hit	.0	0	.08	.14	.14	.13	.10	.09	.09
	Err	−1	−2	−3	−1	−2	−1	−1	1	1.5
	Acc	.88	.88	.91	.98	.98	.99	.99	.99	.95
HW	Tim	.05	.06	.26	.70	1.15	4.1	8.7	29.7	5.6
	Hit	0	0	0	0	0	0	0	0	0
	Err	1	4	12	25	−46	−95	−186	−380	93.6
	Acc	.88	.75	.63	.61	.64	.63	.64	.63	.67
T	Tim	.6	1.1	2.2	4.5	9.2	17.7	36.7	75.2	18.4
	Hit	.02	.017	.02	.01	.015	.01	.01	.01	.014
	Err	0	1	2	4	−2	−16	−27	−46	12.3
	Acc	1.0	.94	.94	.94	.98	.94	.95	.96	.95

doi:10.1371/journal.pone.0093365.t002

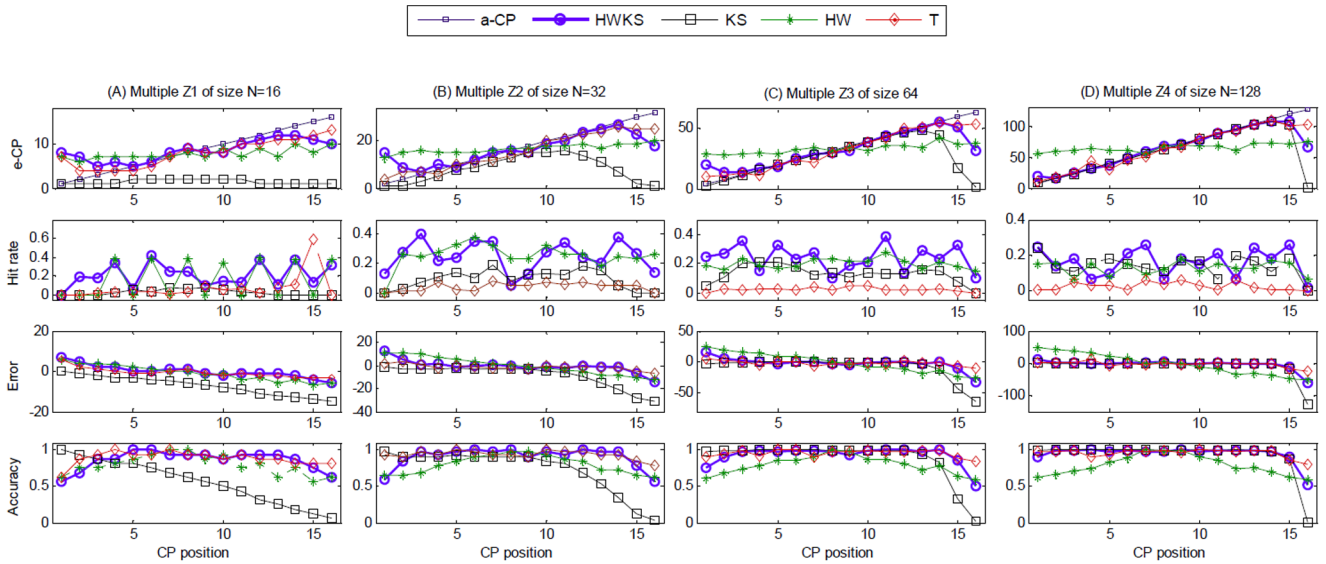

Figure 8. The analysis of e-CP, hit rate, error and accuracy for multiple 100 simulations on different CP test positions, with different *N*, and *k*, by HWKS, KS, HW, and T, respectively. (A) The results of multiple samples *Z*1, with $N = 16$; (B) the results of multiple samples *Z*2, with $N = 32$; (C) the results of multiple samples *Z*3, with $N = 64$; (D) the results of multiple samples *Z*4, with $N = 128$.
doi:10.1371/journal.pone.0093365.g008

A. Analysis on simulated time series

In our simulations, the artificial time series are generated randomly in terms of normal distribution $\mathcal{N}(0,1)$, i.e., $\mathcal{N}$(mean, $u = 0$, standard deviation, $sd = 1$), and then the normally distributed datasets is used for abrupt change detection. Specifically, each diagnosed time series sample of size N is composed of both a normal segment of size k, and an abnormal segment of size N-k, in which a designed abrupt change is contained by adding constant variation v to the normal random numbers of size N-k.

Single simulation on single CP test position. First, CP detection is performed on single time series sample with single CP test position. The simulation results of CP detection are illustrated in Fig. 6, and the summarized analysis of error, and accuracy is shown in Table 1. For the proposed HWKS, it can detect the designed CP from samples of different sizes and CP test positions, with smaller error, and higher accuracy than HW, and T methods. For KS, it has the smallest averaged error in all four methods, but it has lower accuracy than HWKS, especially when sample size N is small. On the contrary, both of HW and T are worse, due to

Table 3. The summary of multiple simulations on different CP test positions.

M \ Z	Size, *N*	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	Averaged
HWKS	***Tim***	*.18*	*.45*	*.59*	*.78*	*.86*	*1.18*	*1.71*	*2.6*	*1.04*
	Hit	*.29*	*.21*	*.25*	*.24*	*.16*	*.10*	*.05*	*.03*	*.17*
	Err	*1.25*	*2.2*	*3.2*	*4.9*	*6.8*	*11.2*	*21.6*	*33.7*	*10.6*
	Acc	*.84*	*.86*	*.90*	*.92*	*.95*	*.96*	*.96*	*.97*	*.92*
KS	Tim	.14	.50	.91	1.8	4.1	9.3	22.9	62.7	12.79
	Hit	.0	.03	.09	.13	.14	.11	.09	.07	.08
	Err	3.5	7.1	8.8	8.1	9.5	16.1	31.0	63.3	18.4
	Acc	.56	.56	.72	.87	.92	.93	.94	.94	.81
HW	Tim	.03	.18	.45	.77	2.2	8.2	28.7	72.7	14.15
	Hit	.08	.08	.04	.02	.02	.03	.02	.03	.04
	Err	1.9	3.3	6.4	13.2	27.3	61.8	123.7	251.9	61.2
	Acc	.76	.79	.79	.79	.78	.76	.76	.75	.77
T	Tim	.78	3.1	6.1	12.1	24.3	49.7	101.7	209.6	50.9
	Hit	.15	.16	.25	.20	.13	.08	.06	.04	.13
	Err	1.38	1.8	2.1	3.1	5.7	12.4	25.3	53.3	13.1
	Acc	.83	.88	.93	.95	.95	.95	.95	.94	.92

doi:10.1371/journal.pone.0093365.t003

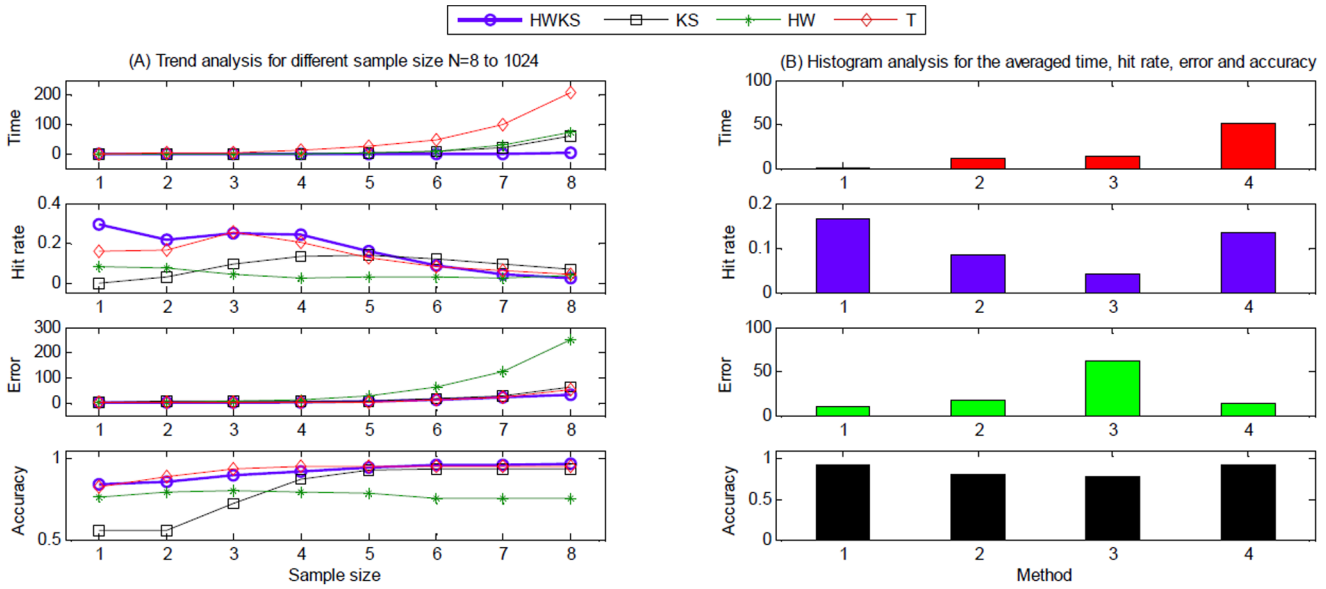

Figure 9. The analysis of computation time, hit rate, error and accuracy on different sample size, for HWKS, KS, HW, and T, respectively. (A) The trend analysis for different sample size from $N=2^3$ to 2^{10}, and (B) the histogram analysis for the averaged computation time, hit rate, error and accuracy. In (B), '1' stands for HWKS, '2' stands for KS, '3' stands for HW, and '4' stands for T.
doi:10.1371/journal.pone.0093365.g009

bigger error and lower accuracy than HWKS and KS. These simulation results indicate that HWKS has the best sensitivity and performance in four methods, especially HWKS is better than KS, due to smaller error and higher accuracy when abrupt change is located near the left or right boundary of samples with smaller size.

Multiple simulations on single CP test position. Second, we test HWKS and other three methods by multiple 600 simulations on single CP test position, with different N, and k. The representative results of CP detection are illustrated in Fig. 7, and the analysis of computation time, hit rate, error, and accuracy is summarized in Table 2. For the proposed HWKS, most of e-CPs are located near the designed CP position, with the shortest computation time, and the highest hit rate in all four methods, as well as smaller error, and higher accuracy than HW and T. For

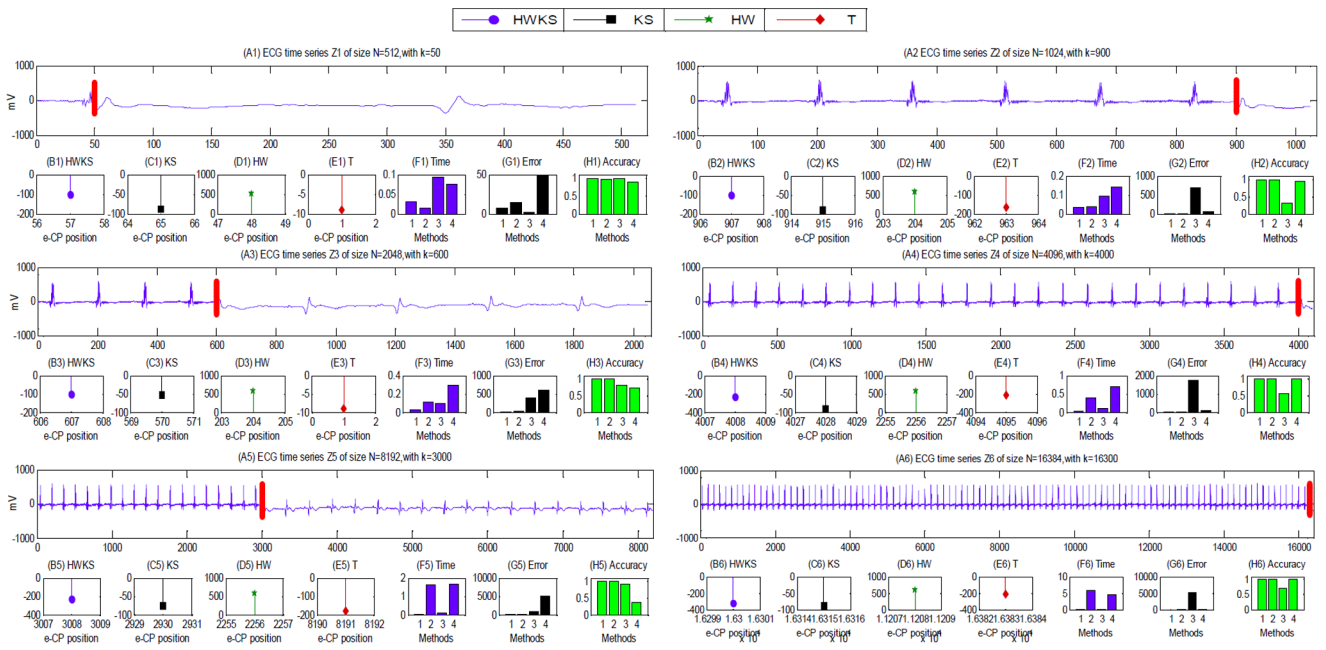

Figure 10. The results of CP detection from assembled ECG time series of size $N=2^k$, $k=9, 10, \ldots, 14$, by HWKS, KS, HW, and T, respectively. (A1)–(A6) the assembled ECG sample $Z1$–$Z6$; (B1)–(B6),(C1)–(C6) ,(D1)–(D6) ,(E1)–(E6) the e-CP detected from $Z1$–$Z6$, by HWKS, KS, HW, and T, respectively; (F1)–(F6) the diagram analysis for the computation time, (G1)–(G6) the error of e-CP, and (H1)–(H6) the accuracy for $Z1$–$Z6$, respectively. In (F)-(H), '1' stands for HWKS, '2' stands for KS, '3' stands for HW, and '4' stands for T.
doi:10.1371/journal.pone.0093365.g010

Table 4. The summary of CP detection from the assembled ECG samples.

M \ Z	Size, *N*	2^9	2^9	2^{10}	2^{10}	2^{11}	2^{11}	2^{12}	2^{12}	2^{13}	2^{13}	2^{14}	2^{14}	Averaged
	CP, *k*	50	500	300	900	600	1400	1600	4000	3000	8100	5000	16000	
Time	*HWKS*	*.032*	*.036*	*.032*	*.033*	*.035*	*.034*	*.034*	*.034*	*.041*	*.036*	*.060*	*.049*	*.038*
	KS	.014	.017	.037	.038	.116	.118	.395	.397	1.59	1.46	5.23	5.58	1.25
	HW	.092	.089	.093	.095	.094	.080	.081	.091	.091	.087	.085	.092	.089
	T	.074	.079	.149	.141	.296	.293	.718	.696	1.65	1.66	4.39	4.38	1.21
Error	*HWKS*	*7*	*0*	*0*	*7*	*7*	*21*	*0*	*8*	*8*	*2*	*12*	*0*	*6*
	KS	15	227	18	15	30	28	18	28	70	6	3	15	39.4
	HW	2	296	96	696	396	1196	1396	1744	744	3762	662	5092	1340.1
	T	49	11	299	63	599	647	2495	95	5191	91	11383	83	1750.5
Accuracy	*HWKS*	*.98*	*1.0*	*1.0*	*.99*	*.99*	*.99*	*1.0*	*.99*	*.99*	*.99*	*.99*	*1.0*	*.99*
	KS	.97	.55	.98	.98	.98	.98	.99	.99	.99	.99	.99	.99	.95
	HW	.99	.42	.90	.32	.80	.41	.66	.57	.90	.54	.95	.68	.68
	T	.90	.97	.70	.93	.70	.68	.39	.97	.36	.98	.30	.99	.74

doi:10.1371/journal.pone.0093365.t004

KS, to some extent, it is better than HWKS and HW for higher accuracy and smaller error; however, it needs much more computation time, and has lower hit rate than HWKS when abrupt change occurs near the left or right boundaries of samples with smaller size N. As for HW, it needs more computation time than HWKS and KS, and has the lowest hit rate and the biggest error in four methods. T is a method with the longest computation time in all four methods, and lower hit rate than HWKS and KS, although it has relatively higher accuracy than HWKS. These simulations show that, the proposed HWKS is a fast and efficient method, due to the shortest computation time and the highest hit rate in all four methods, as well as smaller error and higher accuracy than HW and T. In addition, HWKS has better sensitivity to less significant data fluctuation in sample with small size N, than KS, HW, and T, especially when CP is located near the left or right boundary.

Multiple simulations on different CP test positions. Third, for each diagnosed sample group, multiple 100 simulations on different CP test positions are performed by the proposed HWKS and other three methods. In our simulations, for each N, we select different 16 CP test positions from the different parts of diagnosed samples, *i.e.*, the CP test position is designed by $k = i * 2^{(\log_2 N)-3}$, $N \geq 8$, $i = 1, 2, 3, \ldots, 16$. The selected results

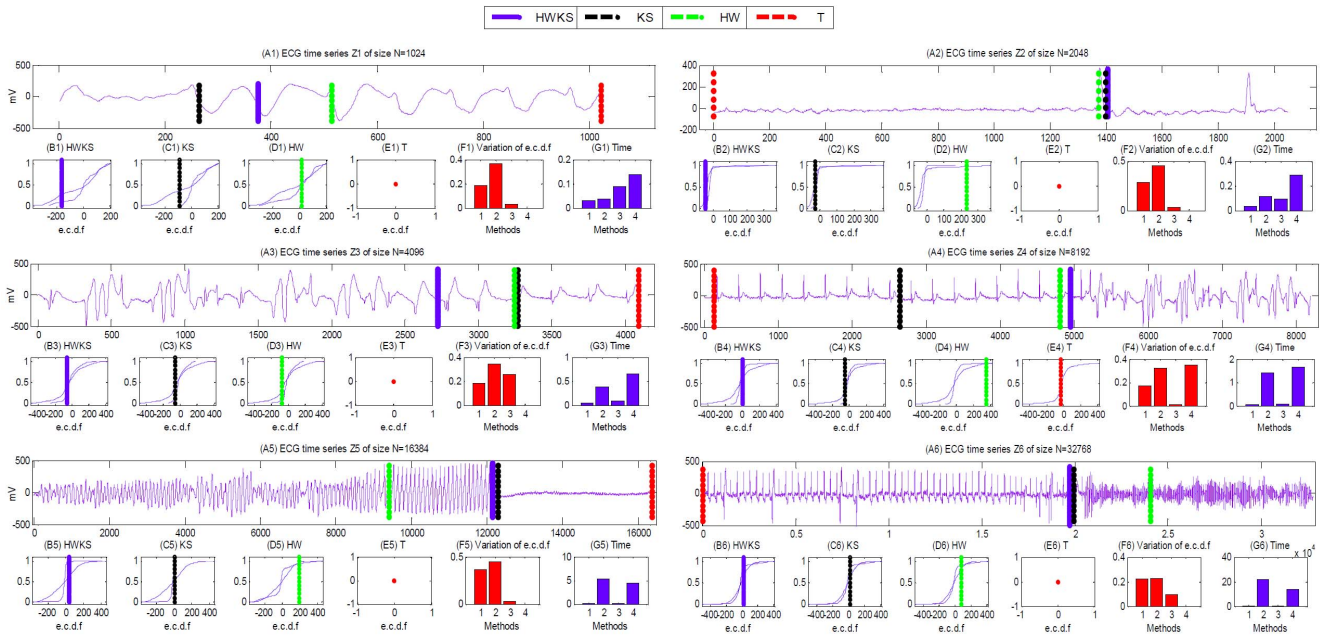

Figure 11. The results of CP detection from abnormal ECG time series of size $N = 2^k$, $k = 10, 11, \ldots, 15$, by HWKS, KS, HW, and T, respectively. (A1)–(A6) the abnormal ECG sample $Z1$–$Z6$; (B1)–(B6), (C1)–(C6), (D1)–(D6), (E1)–(E6) the e.c.d.f derived from two segments of $Z1$–$Z6$, by HWKS, KS, HW, and T, respectively; (F1)–(F6) the diagram analysis of the distance of e.c.d.f, and (G1)–(G6) the computation time of HWKS, KS, HW, and T in $Z1$–$Z6$, respectively. In (F)–(G), '1' stands for HWKS, '2' stands for KS, '3' stands for HW, and '4' stands for T.
doi:10.1371/journal.pone.0093365.g011

Table 5. The summary of CP detection from abnormal ECG samples.

M \ Z	Size, *N*	2_{10}	2_{11}	2_{12}	2_{13}	2_{14}	2_{15}	Averaged
e-CP	*HWKS*	*376*	*1405*	*2722*	*4945*	*12150*	*19711*	*NA*
	KS	264	1399	3268	2646	12296	19930	NA
	HW	514	1374	3244	4810	9416	24066	NA
	T	1023	1	4095	132	16383	1	NA
Computation time	*HWKS*	*.032*	*.034*	*.034*	*.037*	*.049*	*.075*	*.043*
	KS	.036	.114	.376	1.41	5.19	20.82	4.66
	HW	.090	.093	.092	.095	.096	.115	.097
	T	.138	.286	.657	1.65	4.38	13.79	3.49
Variance of e.c.d.f	*HWKS*	*.183*	*.287*	*.183*	*.168*	*.359*	*.223*	*.234*
	KS	.362	.454	.345	.326	.442	.229	.358
	HW	.029	.025	.254	.008	.030	.096	.073
	T	0	0	0	.347	0	0	.057

doi:10.1371/journal.pone.0093365.t005

of simulations on single CP with different test positions are shown in Fig. 8, including e-CP, hit rate, error, and accuracy. In addition, the simulation results for different samples of size from 2^3 to 2^{10} are summarized in Table 3. In terms of computation time, hit rate, error, and accuracy, the trend analysis for different sample size N, as well as the histogram of the averaged analysis for different methods are plotted in Fig. 9.

For the proposed HWKS, it has the best performance for CP detection from samples Z of size N from 2^3 to 2^7, because of the shortest computation time, the highest hit rate, the smallest error, and the highest accuracy in all four methods. Meanwhile, HWKS is better than KS, HW, and T methods for samples Z of size N from 2^8 to 2^{10}, due to the shortest computation time, the smallest error, and the highest accuracy in all four methods, except that hit rate is slightly lower than KS and T. For KS, in general, it has shorter computation time, bigger hit rate, smaller error and higher accuracy than HW. However, KS has the lowest hit rate, the biggest error, and the lowest accuracy as N is below 2^5 in all four methods, which means that KS has worse sensitivity and performance for less significant data fluctuation, especially when N is smaller. For HW, it generally takes shorter computation time than T, whereas, it has the smallest hit rate, the biggest error and lowest accuracy in all four methods. For T, it takes the longest computation time, although bigger hit rate, smaller error and higher accuracy than KS and T. These results show that HWKS has better performance and sensitivity to less significant data fluctuation near the left and right boundary of samples with smaller size N and HWKS is an encouraging method for CP detection on simulated time series, due to shorter computation time, higher hit rate, smaller error, and higher accuracy than KS, HW, and T methods.

B. Analysis on ECG time series

To verify the performance of the proposed method further, we apply HWKS, and KS, HW, and T methods, to detect abrupt change from ECG time series provided by PhysioBank. In ECG experiments, we design the diagnosed ECG samples from different ECG datasets, including the MIT-BIH Normal Sinus Rhythm Database (NSRDB) [37], MIT-BIH Noise Stress Test Database (NSTDB) [38], and MIT-BIH Malignant Ventricular Arrhythmia Database (MVADB) [39,40].

CP detection from assembled ECG samples. First, we select a normal ECG dataset, *16265m* from the NSRDB, and an abnormal ECG dataset, *118e00m* from the NSTDB, and then assemble the diagnosed ECG samples from different segments in the *16265m* and *118e00m*. Specifically, we take the normal ECG segment of size m as X_m, and the abnormal segment of size n as Y_n, respectively, and then assemble the diagnosed ECG sample $Z=\{X_m\ ,\ Y_n\}=\{x_1,\ldots,\ x_m,\ y_1,\ldots,\ y_n\}$. Meanwhile, we design another normal ECG segment from *16265m*, i.e., $X=\{x_1,\ldots,\ x_{m+n}\}$, as normal time series input.

In this ECG experiment, a single CP test position is arranged near the left and right boundary of the assembled ECG sample. For the assembled ECG sample of size from $N=2^9$ to 2^{14} with different CP position k, the results of CP detection from $Z1$–$Z6$ are illustrated in Fig. 10, and then the analysis of computation time, error, and accuracy are summarized in Table 4. Comparing with KS, HW, and T methods, the results show that the proposed HWKS can estimate CP position more quickly, and distinguish the normal and abnormal segments from the assembled ECG samples more efficiently, with smaller error and higher accuracy. For KS, it has smaller error and higher accuracy than HW and T, whereas, it takes much more computation time than HWKS, HW, and T, especially when sample size N gets bigger, meanwhile, KS is less sensitive for less significant statistic fluctuation, with bigger error near the right boundary. For HW, it is inefficient, because of the biggest error and lowest accuracy in all four methods. For T, it is also discouraging for longer computation time, bigger error and lower accuracy than HWKS and KS. Therefore, the proposed HWKS has the best performance in this ECG experiment out of all four methods.

CP detection from abnormal ECG samples. To verify the performance of CP detection further, we apply the proposed HWKS, and KS, HW and T to analyze the abnormal ECG time series directly. In this part, we select the abnormal ECG segment from *118e00m* in the NSTDB, i.e., $Z=\{y_1,\ldots,\ y_n\}$, as a diagnosed ECG sample. Then, we take another normal ECG segment from *16265m* in the NSRDB, i.e., $Z=\{X_n\}=\{x_1,\ldots,\ x_n\}$, as normal input signal. To some extent, the distance of e.c.d.f can partly

reflect the statistical fluctuation. Therefore, we take this variable as an indicator of the data fluctuation between two ECG segments divided by e-CP position. The results of CP detection, including the e-CP position, distance of e.c.d.f, and computation time, are plotted in Fig. 11, and summarized in Table 5.

For abnormal ECG samples $Z1$–$Z6$ with different size N from 2^{10} to 2^{15}, HWKS can detect abrupt change position, and then divide the original ECG sample into two parts, with the shortest computation time out of four methods, and bigger distance of e.c.d.f than HW and T methods. For KS, it can detect CP with the maximal distance of e.c.d.f; however, it needs longer computation time than HWKS, HW and T, especially for ECG sample with big size N. On the other hand, for HW, it is inefficient, due to smaller distance of e.c.d.f than HWKS and KS. T is also inefficient, because of longer computation time than HWKS, HW, the smallest distance of e.c.d.f, and the invalid e-CP position for most of the abnormal ECG samples.

Especially, the results of CP detection from $Z1$–$Z3$ plausibly indicate that, a patient seems recovering from abnormal state of health, after overtaking the critical e-CP position detected by HWKS. On the contrary, the results from $Z4$–$Z6$ suggest that, a patient is encountering a risky situation from the former state of health, after going through the vital e-CP position. These results indicate that HWKS can capture abrupt change position from a diagnosed ECG sample quickly and efficiently, and the detected CP is very useful to find a critical time from ECG time series, where a patient might encounter an important conversion between two different states of health. Therefore, HWKS is an efficient and encouraging method for detecting abrupt change from abnormal ECG time series, and it is very meaningful in inspecting and diagnosing different states of health from diagnosed ECG time series more quickly and efficiently.

Conclusion

In this paper, based on HW and a modified KS statistic, a novel HWKS method is proposed for CP detection from large-scale time series. First, two BSTs are constructed from a diagnosed time series by means of multi-level HW method, the framework of HWKS method is implemented by introducing a revised KS statistic and two search criteria based on TcA and TcD; and then two HWKS-based algorithms are designed to detect an optimal path from TcA in terms of two search criteria. Second, the performance of HWKS is analyzed on simulated time series; the simulations show that HWKS is more sensitive and efficient than KS, HW, and T methods, especially when CP occurs near the left or right boundary with less significant data fluctuation in time series of small size. Last, HWKS is applied to analyze abrupt change on both assembled and abnormal ECG datasets. The results indicate that HWKS can successfully detect abrupt change, and distinguish normal and abnormal ECG segments from assembled ECG samples. In addition, HWKS can estimate an abrupt CP from abnormal ECG segments with different time-scale, and then divide it into two adjacent parts with maximal data fluctuation; therefore, it is very useful to diagnose a patient's different states of health from an abnormal ECG segment more quickly and efficiently. In conclusion, HWKS is a novel and efficient method for fast CP detection; it is a very powerful and promising tool to find useful information from large-scale time series databases.

Acknowledgments

I would like to thank my supervisor Prof. Qing.Zhang in the Australia E-Health Research Centre, CSIRO for his assistance, support and advice on this paper. Meanwhile, I would like thank Prof. Mohan.Karunanithi in the Australia E-Health Research Centre, CSIRO for helps to revise this paper.

Author Contributions

Conceived and designed the experiments: Jin-peng Qi. Performed the experiments: Jin-peng Qi YZ. Analyzed the data: YZ Jie Qi. Contributed reagents/materials/analysis tools: Jie Qi. Wrote the paper: Jin-peng Qi QZ YZ Jie Qi.

References

1. Bolton RJ, Hand DJ (2002) Statistical fraud detection: A review. Statistical Science: 235–249.
2. Ide T, Kashima H (2004) Eigenspace-based anomaly detection in computer systems. Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
3. Kawahara Y, Yairi T, et al. (2007) Change-point detection in time-series data based on subspace identification. Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE.
4. Yamanishi K, Takeuchi J-I, Williams G, Milne P (2004) On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery 8: 275–300.
5. Murad U, Pinkas G (1999) Unsupervised profiling for identifying superimposed fraud. Principles of Data Mining and Knowledge Discovery: Springer: 251–261.
6. Basseville ME, Nikiforov IV (1993) Detection of abrupt changes: theory and application.
7. Darkhovski BS (1994) Nonparametric methods in change-point problems: A general approach and some concrete algorithms. Lecture Notes-Monograph Series: 99–107.
8. Gustafsson F (2000) Adaptive filtering and change detection: Wiley New York.
9. Gustafsson F (1996) The marginalized likelihood ratio test for detecting abrupt changes. Automatic Control, IEEE Transactions on 41: 66–78.
10. Sharifzadeh M, Azmoodeh F, Shahabi C (2005) Change detection in time series data using wavelet footprints. Advances in Spatial and Temporal Databases: Springer:127–144.
11. Brodsky BE, Darkhovsky BS (1993) Nonparametric methods in change point problems: Springer.
12. Huang J, Smola A-J, Gretton A, Borgwardt KM, Schölkopf B, et al. (2007) Correcting sample selection bias by unlabeled data. Advances in neural information processing systems 19: 601–608.
13. Hardle W (2004) Nonparametric and semiparametric models: Springer Verlag.
14. Lin H-D (2007) Automated visual inspection of ripple defects using wavelet characteristic based multivariate statistical approach. Image and Vision Computing 25: 1785–1801.
15. Simard R, L'Ecuyer P (2011) Computing the two-sided Kolmogorov-Smirnov distribution. Journal of Statistical Software 39: 1–18.
16. Aguirre GK, Zarahn E (1998) A critique of the use of the Kolmogorov-Smirnov (KS) statistic for the analysis of BOLD fMRI data. Magnetic Resonance in Medicine 39: 500–505.
17. Hall L, Mba D (2004) Acoustic emissions diagnosis of rotor-stator rubs using the KS statistic. Mechanical Systems and Signal Processing 18: 849–868.
18. Fell J, Kaplan A, Darkhovsky B, Röschke J (2000) EEG analysis with nonlinear deterministic and stochastic methods: a combined strategy. Acta Neurobiologiae Experimentalis 60: 87.
19. Wang Y, Wu C, Ji Z, Wang B, Liang Y (2011) Non-parametric change-point method for differential gene expression detection. PloS one 6: e20060.
20. Alarcon-Aquino V, Barria JA (2001) Anomaly detection in communication networks using wavelets. IEE Proceedings-Communications 148: 355–362.
21. Khalil M, Duchêne J (1999) Detection and classification of multiple events in piecewise stationary signals: comparison between autoregressive and multiscale approaches. Signal processing 75: 239–251.
22. Kobayashi M (2001) Wavelets and their applications in industry. Nonlinear Analysis: Theory, Methods & Applications 47: 1749–1760.
23. Percival DB, Walden AT (2006) Wavelet methods for time series analysis: Cambridge University Press.
24. Salam M, Mohamad D (2008) Segmentation of Malay Syllables in connected digit speech using statistical approach. International Journal of Computer Science and Security 2: 23–33.
25. Qi J, Ding Y, Zhu Y, Wu Y (2011) Kinetic theory approach to modeling of cellular repair mechanisms under genome stress. PloS one 6: e22228.
26. Tseng VS, CH Chen, Chen CH, Hong TP (2006) Segmentation of time series by the clustering and genetic algorithms. Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, IEEE.

27. Alarcon-Aquino V, Barria J (2009) Change detection in time series using the maximal overlap discrete wavelet transform. Latin American applied research 39: 145–152.
28. Walker JS (2002) A primer on wavelets and their scientific applications: CRC press.
29. Percival DB (2008) Analysis of geophysical time series using discrete wavelet transforms: An overview. Nonlinear Time Series Analysis in the Geosciences: Springer:61–79.
30. Yamanishi K, Takeuchi J-i (2002) A unifying framework for detecting outliers and change points from non-stationary time series data. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
31. Brown JR Harvey ME (2008) Rational Arithmetic Mathematical Functions to Evaluate the Two-Sided One sample K-S Cumulative Sample Distribution. Journal of Statistical Software 26: 1–40.
32. Wang J, Tsang WW, Marsaglia G (2003) Evaluating Kolmogorov's distribution. Journal of Statistical Software 8: 1–4.
33. Raimondo M, Tajvidi N (2004) A peaks over threshold model for change-point detection by wavelets. Statistica Sinica 14: 395–412.
34. Stephens MA (1974) EDF statistics for goodness of fit and some comparisons. Journal of the American statistical Association 69: 730–737.
35. Kifer D, Ben-David S, Gehrke J (2004) Detecting change in data streams. Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment:180–191.
36. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron J, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences 100: 8418–8423.
37. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, et al. (2000) Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 101: e215–e220.
38. Moody GB, Muldrow W, Mark RG (1984) A noise stress test for arrhythmia detectors. Computers in Cardiology 11:381–384.
39. Greenwald SD (1986) The development and analysis of a ventricular fibrillation detector, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.
40. Hodges A, Strand AD, Aragaki AK, Kuhn A, Sengstag T, et al. (2006) Regional and cellular gene expression changes in human Huntington's disease brain. Human molecular genetics 15: 965–977.