scientific reports

Check for updates

OPEN

Identifying critical States of complex diseases by local network Wasserstein distance

Changchun Liu, Pingjun Hou✉ & Lin Feng

Complex diseases often undergo abrupt transitions from pre-disease to disease states, with the pre-disease state is typically unstable but potentially reversible through timely intervention. Detecting these critical transitions is crucial. We propose a model-free method, Local Network Wasserstein Distance (LNWD), for identifying critical transitions/pre-disease states in complex diseases using single sample analysis. LNWD measures statistical perturbations in normal samples caused by diseased samples using the Wasserstein distance, and identifies critical states by observing LNWD score changes. Applied to KIRP, KIRC, LUAD, ESCA (TCGA datasets) and GSE2565, GSE13268 (GEO datasets), the method successfully identified critical states in six disease datasets. This single-sample, local network-based approach provides early warning signals for medical diagnosis and holds great potential for personalized disease diagnosis.

Keywords Complex disease, Critical state, Local network, Wasserstein distance, Single sample, Premorbid state

Research has shown that the progression of complex diseases such as cancer disease[1], diabetes[2], and influenza outbreaks[3] etc. does not follow a steady course, and sudden shifts can occur in the course of disease progression[4]. Overall, the process of complex disease development can be considered as a dynamic change in a nonlinear dynamical system[5] where sudden deterioration corresponds to phase transitions or state transitions at the bifurcation points of the nonlinear dynamical system[6]. Therefore, the development of complex diseases is usually divided into three phases: the relatively normal state, the premorbid state, and the disease state[7].(Fig. 1a) The pre-disease state can be seen as a critical boundary of the normal state, characterized by significantly elevated dynamic instability and high sensitivity to perturbations. At this stage, the system is near a critical threshold, and timely interventions may reverse its trajectory toward disease. Disease states (e.g., advanced fibrosis[8] or metastatic cancer[9]) are relatively stable and often irreversible due to structural or functional damage[10], highlighting the importance of early detection. Therefore, it is of great significance to discover the critical warning signals of disease transformation and identify the critical point of disease development.

However, for many complex diseases, identifying pre-disease states remains a challenging task due to modeling and data limitations. Specifically, the dynamics of biological systems during disease progression usually involve a large number of molecules, clinical data are overly dimensional and contain noisy interferences, and the available sample size is small[11]. Dynamic Network Biomarker (DNB) is an effective method for identifying the tipping points of complex diseases[12]. This method considers the developmental process of complex diseases as a high-dimensional nonlinear dynamical system, based on the theory of critical slowing down[13], and provides early warning signals by identifying a group of key biomolecules (e.g., genes, proteins, etc.), which constitute the DNB group and exhibit significant dynamic features. Specifically, before the system undergoes a critical phase transition, the members of the DNB group display unique changes in their statistical properties: their intermolecular correlations suddenly increase, while their individual standard deviations rise significantly. These dynamic features of synergistic fluctuations are quantitatively captured through the construction of composite statistical metrics, ultimately serving as reliable Early Warning Signals (EWS) that indicate the system is approaching a critical threshold. This provides a crucial window of opportunity for early intervention in complex diseases. Currently, the DNB theory and its extensions have been widely applied, such as identifying the premorbid state of type 1 diabetes by this method[2], detecting early warning signals of cancer immune-detection point blocking reactions[14] and identification of critical points in endocrine resistance processes[15] and studying cell differentiation[16] etc. However, since the DNB theory requires evaluating its three statistical conditions with multiple sample data at each time point, the application of the method is greatly limited for most clinical datasets

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang 471000, China. ✉email: haust_hpj@haust.edu.cn

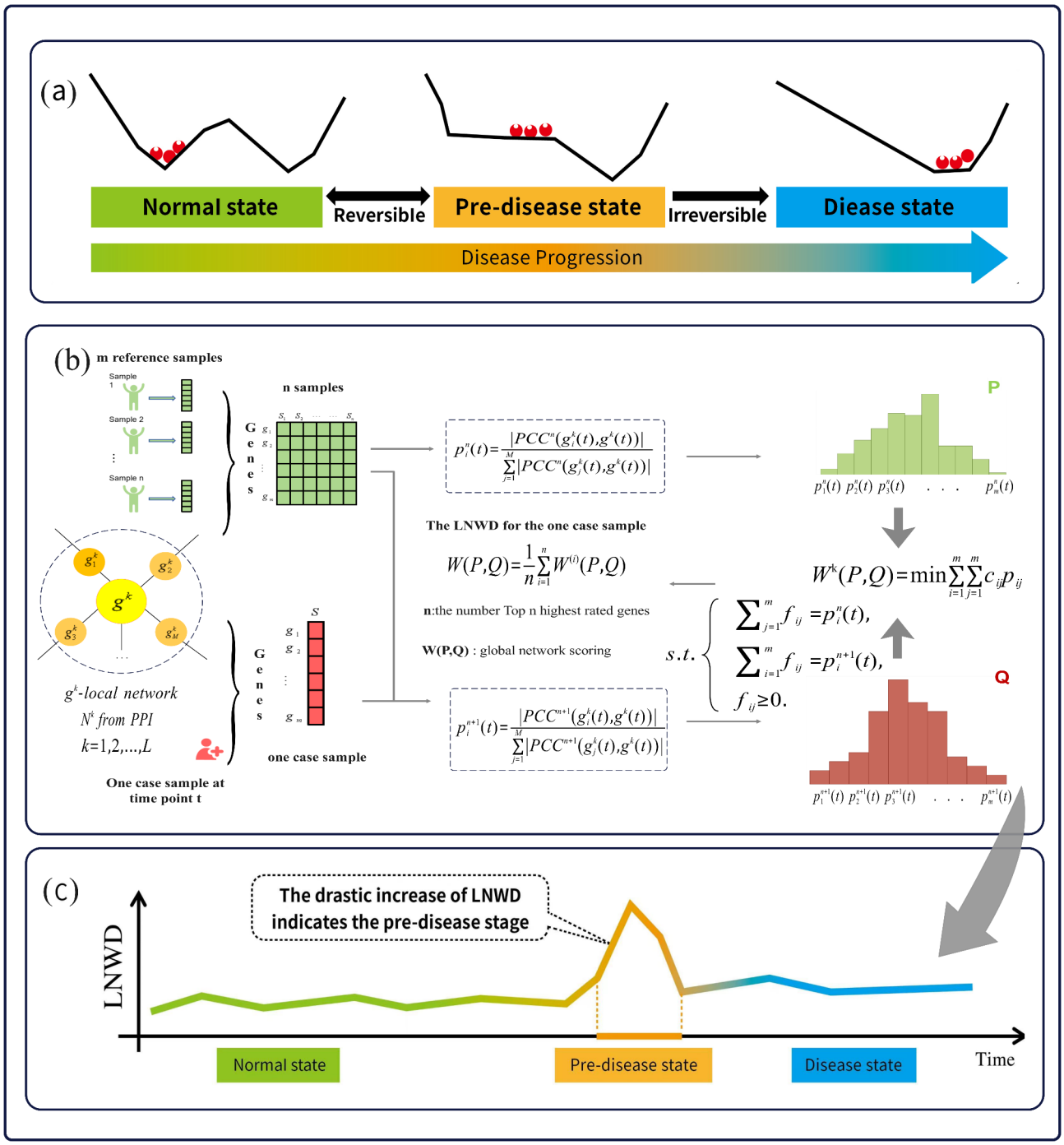

Fig. 1. Identification of critical state of complex disease based on LNWD method. (**a**) Three stages in the development of a complex disease. The vertical axis represents the hypothetical potential function, a theoretical model used to describe the dynamic stability of a biological system. This model visually depicts the critical transition of a system from a healthy state to a diseased state through changes in the shape of the potential energy surface. (**b**) Take a set of normal group samples as reference samples, add a single diseased sample into the normal group samples to form a mixed group, measure the statistical perturbation of the single diseased sample relative to the reference samples by calculating the LNWD scores of the local networks of the normal group and the mixed group, and finally detect the early warning signals of the pre-disease state by calculating the LNWD scores of the global network. (**c**) In the process of complex disease development, when the disease is close to the critical state, the LNWD score changes significantly, indicating that it is about to enter the critical state, and the fluctuation of the LNWD score can effectively identify the pre-disease state.

with small sample sizes. Therefore, when only a single sample is available in a clinical dataset, a new approach is needed to provide early warning signals to identify pre-disease states.

Therefore, many researchers have constructed some single-sample-based methods to recognize pre-disease states based on the DNB theoretical framework, such as constructing a single-sample network to recognize complex disease triggers, which successfully solves the sample size limitation of DNB methods and does not require the use of clustering algorithms or other heuristic algorithms[17]. In addition, the Shannon entropy method has been applied to detect changes in the dynamics of complex systems[11], which is effective in identifying the true dynamics of disease development[18]. However, Shannon entropy is sensitive to small probability events, which sometimes affects the judgment of critical state; meanwhile, some researchers have proposed detecting the critical states of complex diseases using Kullback-Leibler Divergence[10]. However, since two probability distributions that do not overlap, calculating the Kullback-Leibler Divergence is meaningless or produces infinite values[19], the method is not always effective.

In this paper, we propose a model-free method for detecting critical states of complex diseases based on a single sample, namely the Local Network Wasserstein Distance (LNWD) method, which is grounded in nonlinear dynamical systems theory. The necessary conditions for LNWD are rigorously derived from mathematical formulations, ensuring a theoretically solid formulas[6].

The Wasserstein distance (also known as Earth Mover's Distance, subsequently abbreviated as WD) is an effective measure of the difference between distributions, which is well suited for quantifying natural perturbations of probability distributions and is very beneficial for the analysis of complex systems[20]. Specifically, WD quantifies the difference between two distributions by measuring the minimum cost required to transform one distribution into the other, and is therefore also known as the optimal transmission distance[21]. When dealing with the difference between two probability distributions, WD has significant advantages over general divergence, for example, WD is symmetric, measures the overall difference between the two distributions, and produces a metric that is not related to the direction of the transformation of the distributions[22] WD has robustness[23] and small probability events have less impact on it; WD can effectively measure the difference between two probability distributions with little overlap[24], and can solve the problems that Kullback-Leibler Divergence may face. WD has been applied in several fields, such as embedding of image signals[25], privacy preserving computation[26], machine fault diagnosis[27] etc. Therefore, this study aims to construct a single-sample based model-free method to detect the critical state of complex diseases by WD.

Specifically, the LNWD method takes a set of normal group samples as a reference sample, adds a single diseased sample to the normal group samples to form a mixed group, and measures the statistical perturbation of the single diseased sample relative to the reference sample by calculating the LNWD scores of the local networks of both the normal group and the mixed group. Then, the top 10% of the local network LNWD scores are selected, and their average is calculated to obtain the global network LNWD score. The global network LNWD scores are used to detect early warning signals of pre-disease states, thereby identifying pre-disease states during the development of complex diseases.(Fig. 1b). The method was applied to four TCGA datasets: renal papillary cell carcinoma (KIRP), renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), and esophageal carcinoma (ESCA); as well as to two datasets from the GEO database: acute lung injury in mice dataset (GSE2565), and type II diabetes mellitus in adipose tissue in rats dataset (GSE13268). The pre-disease progression state was identified by calculating the LNWD score of the global network for each stage, and the critical state identified by the method was validated by later survival analysis as well as molecular network dynamics change analysis.

Data processing and functional analysis

Four unrelated clinical tumor datasets were downloaded from the TCGA database (https://portal.gdc.cancer.gov/): KIRP, KIRC, LUAD and ESCA; and two datasets from the GEO database (https://www.ncbi.nlm.nih.gov/geo/): GSE2565 and GSE13268.

The four datasets from TCGA contain RNA-seq data for tumor and tumor-adjacent samples, and tumor samples were staged based on clinical information provided by TCGA, and samples without staging information were screened out. For the two datasets from the GEO database, the samples without staging information were also screened out, and the gene probe names in the expression matrix were converted to gene names according to the gene annotation information provided by the platform. For genes with multiple probe localizations, the expression amount was averaged as the expression amount of the gene, and the data corresponding to the probes without gene annotation information were screened out. In this study, differential analysis of the data was performed using the R packages edgeR, limma, and DESeq2 for each period. Differential genes with a logFC value greater than 2 and a P-value less than 0.05 were selected. The differential genes for each period were determined by taking the intersection of the genes identified by the three methods, and the final set of differential genes was obtained by taking the union of the differential genes from all periods.

Molecular interaction networks were constructed in the following steps: first, the protein-protein interaction (PPI) networks of Homo sapiens and mouse (https://cn.string-db.org/) were downloaded. Reciprocal linkages of genes with significance levels below 0.800, as well as isolated points without reciprocal linkages, were filtered out. Second, differential genes were mapped onto the interaction network to provide the interaction relationships between genes. Finally, the molecular interaction network was visualized using Cytoscape (https://cytoscape.org/).

Theoretical background

The LNWD approach is based on DNB theory, which has a solid theoretical foundation. DNB theory models the progression of complex diseases as a nonlinear dynamical system that evolves over time[28] and suggests that sudden disease deterioration can be viewed as a phase transition or state transition occurring at a critical point[29]. The theory further proposes that the stability of the system undergoes significant changes during disease

progression[7], particularly near the critical point, where certain groups of molecules (defined as DNB molecules) exhibit specific dynamic characteristics[10]. These molecules may serve as critical early warning signals in the initial stages of disease.

- The correlation between any pair of members of the DNB group (PCC_{in}) increases rapidly;
- The correlation (PCC_{out}) between a DNB group member and any other non-DNB member decreases rapidly;
- The standard deviation (SD_{in}) or coefficient of variation of any member of the DNB group is increased dramatically.

These three properties are necessary conditions for a state transition to occur at a bifurcation point, and can be approximated as the presence of a group of biomolecules that undergo strong fluctuations and are highly correlated, signaling that a critical transition is imminent. These three properties are used to quantify the critical state as a pre-disease early warning signal and to identify the dominant biomolecules comprising the DNB members. These three properties are the theoretical basis of DNB theory, which is now widely used in the analysis of many disease progressions and biological processes to predict critical states and their drivers in complex systems[30]. From these three properties, it can be seen that the critical transitions of a system are in fact "distributional transitions", i.e., when the system approaches a critical state, the distribution of certain variables will change significantly[10]. Therefore, exploring the distribution of some variables can predict some upcoming state transitions.

WD is widely used to measure the difference between distributions, the basic principle is to quantify the difference between distributions by measuring the minimum cost required to transform one distribution into another. This study aims to identify the critical state of a complex disease by means of a metric constructed by WD. For two distributions P and Q, WD is defined as[31]:

$$W(P,Q) = \inf_{\gamma \sim \prod(P,Q)} E_{x,y\sim\gamma}\left[||x-y||\right], \tag{1}$$

Where$\prod(P,Q)$denotes the set of all possible joint distributions of P and Q, $\|.\|$denotes the norms, for each possible joint distributionγ, x and y denote the sample pairs drawn fromγ,$E_{x,y\sim\gamma}\left[||x-y||\right]$denotes the expectation of the distance between the sample pairs, and the smallest of the expected values from all possible joint distributions is the Wasserstein distance between the two distributions.

The Wasserstein Distance (WD) is robust to the problem of non - overlapping distribution support sets[32]. For example, for two Dirac distributions $P = \delta_x$ and $Q = \delta_{x+\varepsilon}$ that are ε apart, their WD is ε. However, when the support sets of the two distributions P and Q have no overlap, the Kullback - Leibler (KL) divergence will diverge, and the Jensen - Shannon (JS) divergence will saturate to a constant value (log2), and it is impossible to accurately measure the distance between the two distributions. That is, if $\mathrm{supp}\,(P) \cap \mathrm{supp}\,(Q) = \emptyset$, then $D_{KL}\,(P \parallel Q) = +\infty$ and $D_{JS}\,(P \parallel Q) = \log 2$.

Since the distributions constructed in this paper are discrete, and the direct calculation of the WD has a high complexity and a large computational cost, we transform the calculation of the WD into an optimal transport problem between two discrete distributions. Given two discrete distributions $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^{m} b_i \delta_{y_i}$, the definition of the WD is:

$$W(\mu,\nu) = \min \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{i,j} d(x_i, y_j), \tag{2}$$

Among them, δ_{x_i} and δ_{y_i} represent the Dirac delta functions, and $\gamma_{i,j}$ is a non - negative optimization variable, namely the transportation plan, which satisfies $\sum_j \gamma_{i,j} = a_i$ and $\sum_i \gamma_{i,j} = b_i$. $d(x_i, y_j)$ represents the distance between two points, that is, the transportation cost. In this study, the R package (lpSolve) is used to solve this optimization problem.

Critical point identification algorithm based on LNWD method

LNWD uses a one-sample local network Wasserstein distance approach to identify critical states of disease. Specifically, n samples from the normal group are used as the reference group (as background samples representing healthy or relatively healthy individuals), and single diseased sample at a certain point in time are added to the reference samples to form a mixed group, and the critical state of a complex disease is identified by measuring the statistical perturbation that single diseased sample bring to the reference samples, the specific algorithmic process is as follows (Fig. 1b-c) :

[Step1] Firstly, we downloaded the TCGA data (or GEO data) and performed a differential analysis of the genes using TCGA count data (or GEO gene chip data). The screened differential genes were then mapped to protein-protein interaction (PPI) networks. We used the STRING database, a tool that can localize differential genes to PPI networks based on known protein interaction information. Interaction links between genes with a confidence level higher than 0.800 were selected, and isolated nodes without connections were filtered out to form a global networkN^G.

[Step2] Perform log2 transformation of tpms data from TCGA (or gene chip data from GEO), stage the data according to clinical information, and map the processed data to the generated global networkN^G.

[Step3] Construction of local network central gene probability distribution. For each geneg^k as a central gene, the local network N^k $(k = 1, 2, \cdots, l)$ is extracted from the global network N^G,where$\left\{g_1^k, g_2^k, \cdots, g_m^k\right\}$ is the nearest neighbor of g^k. Then the probability distribution for normal group samples is :

$$p_i^n(t) = \frac{\left|PCC^n\left(g_i^k(t), g^k(t)\right)\right|}{\sum_{j=1}^{m}\left|PCC^n\left(g_j^k(t), g^k(t)\right)\right|}, \tag{3}$$

where the constant m denotes the number of first-order neighbors of the gene g^k and $PCC^n\left(g_i^k(t), g^k(t)\right)$ denotes the Pearson correlation coefficient between the central gene $g^k$and the neighboring gene g_i^kbased on n normal samples of data at time point t. Similarly, for the construction of probability distributions of mixed group samples, we have

$$p_i^{n+1}(t) = \frac{\left|PCC^{n+1}\left(g_i^k(t), g^k(t)\right)\right|}{\sum_{j=1}^{m}\left|PCC^{n+1}\left(g_j^k(t), g^k(t)\right)\right|}. \tag{4}$$

[Step4] Calculation of LNWD scores for localized networks. This study transforms the solution of the Wasserstein distance into the solution of the optimal transportation strategy in the optimization problem[33]. In this study, the linear programming model is constructed by taking $P = \left(p_1^n(t), p_2^n(t), \cdots, p_m^n(t)\right)$as the place of production and $Q = \left(p_1^{n+1}(t), p_2^{n+1}(t), \cdots, p_m^{n+1}(t)\right)$ as the place of sale:

$$\begin{aligned} &\min \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}p_{ij} \\ &s.t. \begin{cases} \sum_{j=1}^{m} p_{ij} = p_i^n(t), \\ \sum_{i=1}^{m} p_{ij} = p_j^{n+1}(t), \\ p_{ij} \geqslant 0. \end{cases} \end{aligned} \tag{5}$$

wherep_{ij}denotes the amount of goods shipped from the production place $p_i^n(t)$to the sales place $p_j^{n+1}(t)$,$c_{ij} = \left|\overline{g_i^k}(t) - \overline{g_j^k}(t)\right|$denotes the corresponding transportation cost, $\overline{g_i^k}(t)$ and $\overline{g_j^k}(t)$denote the average expression of genesg_i^k and genesg_j^k based on n samples of the normal group, m denotes the number of production and sales places, and the number of production and sales places in the probability distribution constructed in this study is the same, and the minimized transportation cost solved is Wasserstein's distance for the normal and the mixed groups in the first k localized network,

$$W^k(P, Q) = \min \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}p_{ij}. \tag{6}$$

Use this distance as the *k*th local network LNWD score to measure the statistical perturbation introduced by a single diseased sample in that local network relative to the reference sample.

[Step5] Calculation of global network LNWD scores. According to the above algorithm, the local network LNWD scores of the differential genes at each time point are solved, the local network LNWD scores of the differential genes in each period are sorted, and the top n genes with the highest scores are selected, and the scores of these n genes are averaged to finally get the global network LNWD score :

$$W(P, Q) = \frac{1}{n}\sum_{i=1}^{n} W^{(i)}(P, Q) \tag{7}$$

Where $W^{(i)}(P, Q)$denotes the score of the *i*th local network after sorting the local networks from largest to smallest, and$W(P, Q)$ denotes the global network score. In practice, since each sample is ranked in a different order based on gene scores, we select the top 10% of genes based on their scores for each sample within each period. Then, we take the intersection of these selected genes across all samples in a single period, Then take the union of the genes from all periods., and calculate the average of their scores to obtain the global network score.

According to the introduction in the theoretical background section and DNB theory, when a system approaches a critical state, a group of genes will exhibit significant expression changes and a high degree of correlation. Consequently, the probability distribution constructed from the Pearson correlation coefficients between the central gene and its neighboring genes will shift significantly compared to that of the normal

group samples. In summary, as the system approaches a critical state, the difference between the probability distributions derived from the mixed and normal group samples increases notably. The highest point of the LNWD score represents the critical state we have identified.

Results

In this paper, the LNWD method for identifying critical points is described in the Methods section. To verify its validity, the method is applied to four cancer datasets from the TCGA database and two datasets from the GEO database. Furthermore, the identified critical states in these datasets are validated through survival analysis and molecular network dynamic change analysis. The successful identification of these disease-critical states demonstrates the effectiveness of the LNWD method in detecting critical states before the progression of complex diseases.

Identifying critical transition points of cancers

In this study, the method was applied to four cancer datasets in the TCGA database: KIRP, KIRC, LUAD and ESCA, and the samples were staged according to the clinical information, i.e., the datasets of KIRP, KIRC, and ESCA were divided into four stages: I, II, III, and IV.The LUAD dataset was divided into seven stages: IA, IB, IIA, IIB, IIIA, IIIB, IV. The samples in the normal group were used as reference samples, and the LNWD scores of each sample were calculated according to the algorithm in the Sect. 2, and the LNWD scores of all diseased samples in each period were averaged to obtain the dynamics of the LNWD scores during the progression of the disease, and key shifts in disease progression were identified based on the turning points of the LNWD score curves.

In order to verify the critical state identified through the LNWD method, this study compared the survival rates of samples before and after the identified critical state using survival analysis. The survival curves clearly show that the survival time of samples before the state transition is significantly longer than that after the transition. To further validate the identified critical state at the network level, we downloaded the interactions of differentially expressed genes previously mapped onto the PPI network and plotted the dynamic changes of the network. In the network graph, nodes represent LNWD scores, while edges indicate the Pearson correlation coefficients between genes. The identified critical state was validated by observing the changes in the network graph.

For KIRC, the KIRC dataset yielded 1,848 differentially expressed genes, with sample sizes of 38 for the normal group and 250, 52, 117, and 73 for stages I, II, III, and IV, respectively. A significant shift in LNWD score at stage II signals an impending disease state.(Fig. 2a) According to existing studies, tumors in stages I and II remain localized, and in stage III, the tumor spreads to the major veins or spreads to perirenal tissues, and in stage IV the tumor metastasizes to form metastatic foci at distant sites, this progression pattern supports the validity of the critical state identified in this study[33]. The survival analysis further confirms this finding, (Fig. 2e)the survival curves of the samples before and after the critical point (stage II) were significantly different ($P<0.0001$), and the survival time before the critical point was significantly greater than the survival time after the critical point. The network diagrams for the four phases are shown in Fig. Significant changes occur in the network during period II, indicating the arrival of the critical state, which is consistent with the experimental results described above. The network diagrams for the four phases are shown in (Fig. 3A). Significant changes occur in the network during period II, indicating the arrival of the critical state, which is consistent with the experimental results described above.

For KIRP, the KIRP dataset yielded 1,695 differentially expressed genes, with sample sizes of 20 for the normal group and 162, 20, 46, and 12 for stages I, II, III, and IV, respectively. A sudden increase in the LNWD score at period III suggests an imminent disease state after that period.(Fig. 2b) In a series of studies on staging migration in papillary cell carcinoma of the kidney, it was observed that the 5-year relative survival rate of patients with stage IV was very low, and the tumor had already developed distant metastases, at which point the disease became complex and severe, while survival rates were relatively high in other periods, and timely treatment should be provided before stage III[34]. Observing the survival curves obtained through survival analysis, (Fig. 2f) the survival curves of the samples before and after the critical point (stage III) were significantly different ($P<0.0001$), and the survival time before the critical state was significantly greater than the survival time after the critical state. The network diagrams for the four phases are shown in (Fig. 3B). Significant changes occur in the network during period III, indicating the arrival of the critical state, which is consistent with the experimental results described above.

For ESCA, the ESCA dataset resulted in 283 differential genes, with sample sizes of 7, 15, 75, 56, 9 for normal group samples, stage I, stage II, stage III, and stage IV, respectively. A significant change in the LNWD score at stage III indicates a critical state, (Fig. 2c) after which disease state is imminent. stage IV esophageal cancer tumors have developed distant metastases, and at stage III the tumor has not yet developed distant metastases, and although the disease is already more severe, surgical treatment is usually still possible[35]. Observing the survival curves obtained through survival analysis, (Fig. 2g)the survival curves of the samples before and after the tipping point (stage III) were significantly different ($P=0.00017$), with the survival time before the tipping point being significantly greater than that after the tipping point. The network diagrams for the four phases are shown in (Fig. 3C) Significant changes occur in the network during period III, indicating the arrival of the critical state, which is consistent with the experimental results described above.

For LUAD, the LUAD dataset resulted in 1853 differential genes, with sample sizes of 29, 123, 135, 47, 68, 67, 10, 25 for normal group samples, stage IA, stage IB, stage IIA, stage IIB, stage IIIA, stage IIIB, and stage IV, respectively. A significant mutation in the LNWD score at period IIB indicates that this period is the threshold for identification, (Fig. 2d) after which the disease state is entered.Stages IA-IIB are the early stages in the development of lung adenocarcinoma, and at stage IIIA and beyond, the lung adenocarcinoma enters an

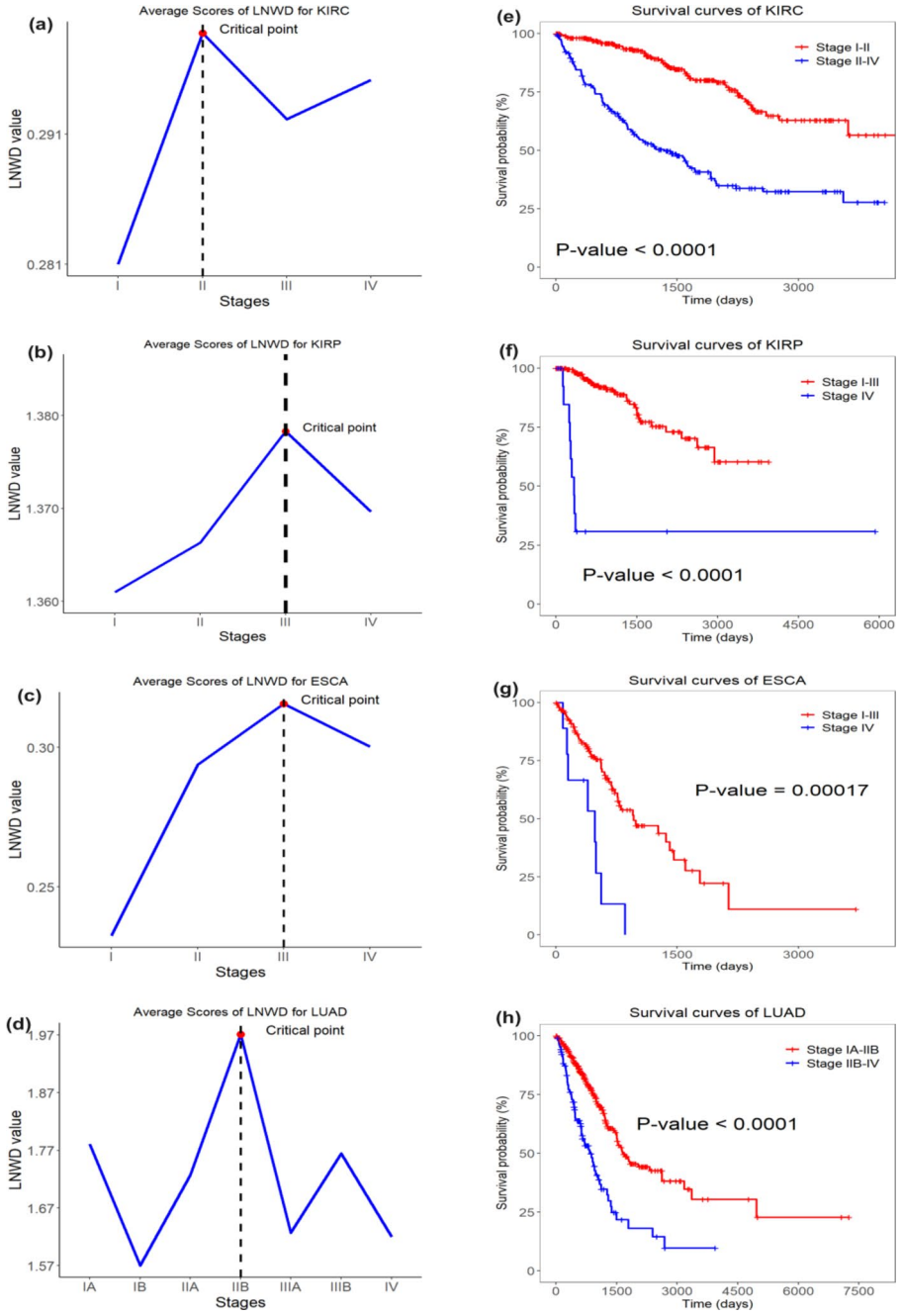

Fig. 2. Identification of Critical States Preceding Tumor Deterioration in Four Cancers Types: (**a**) KIRC; (**b**) KIRP; (**c**) ESCA; (**d**) LUAD. Comparison of survival curves for samples taken before and after the critical state for Four cancers: (**e**) KIRC; (**f**) KIRP; (**g**) ESCA; (**h**) LUAD.

intermediate to advanced stage, often accompanied by metastases to the ipsilateral bronchus or ipsilateral hilar lymph nodes[36]. Observing the survival curves obtained by survival analysis, (Fig. 2h) the survival curves of the samples before and after the critical point (stage IIB) were significantly different ($P < 0.0001$), and the survival time before the critical point was significantly greater than the survival time after the critical point. The network diagrams for the seven phases are shown in (Fig. 3D) Significant changes occur in the network during period IIB, indicating the arrival of the critical state, which is consistent with the experimental results described above.

Identifying critical transitions in mouse acute lung injury

Application of the LNWD method to microarray data from the mouse carbonyl chloride inhalation exposure-induced acute lung injury dataset (GSE2565) from the phosgene-induced acute lung injury experiment in mice[37], which resulted in 121 differential genes, with a sample size of 6 for the normal group, the experimental group in the other time periods, and the control group corresponding to the same time period. In this experiment,

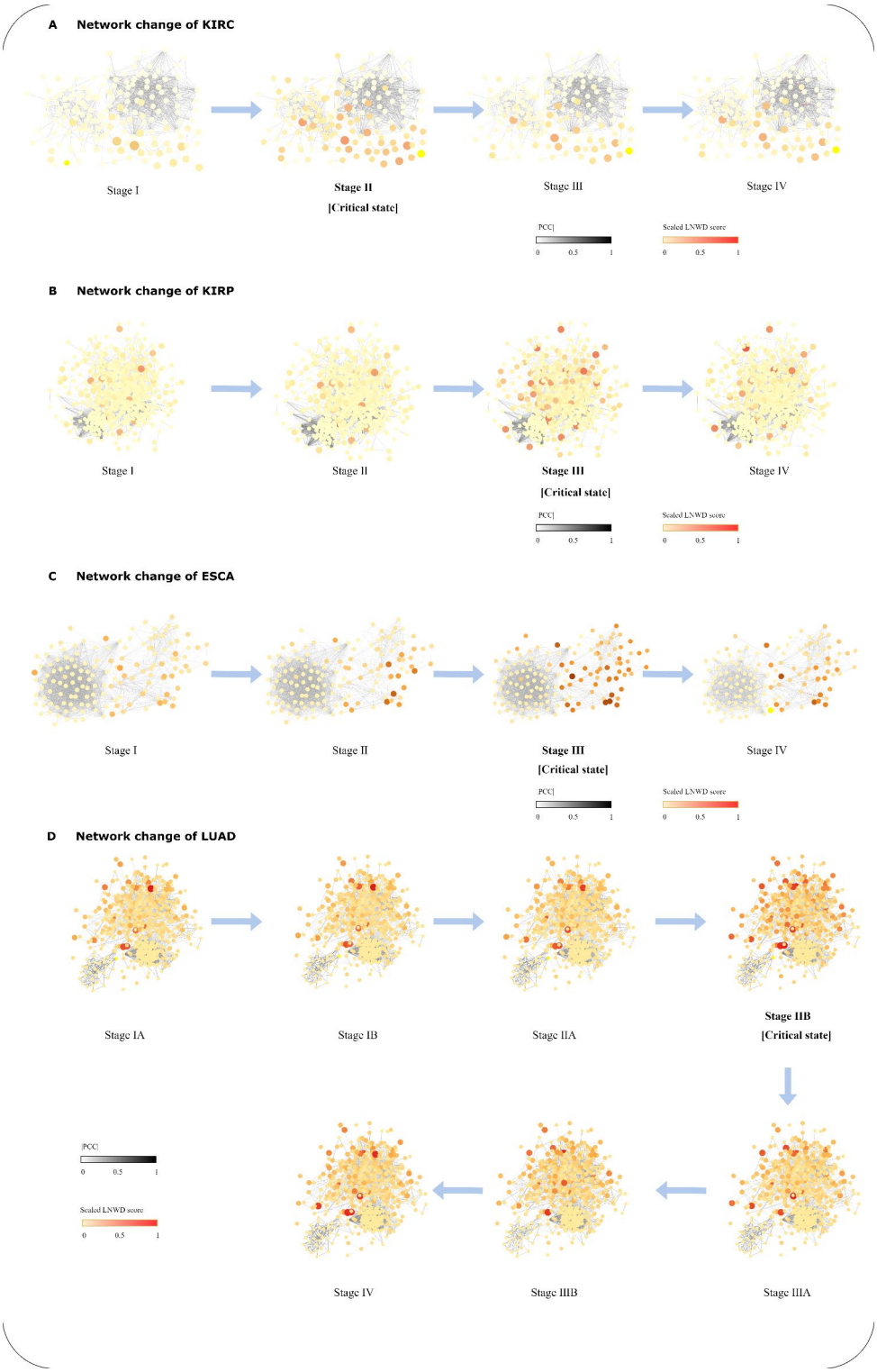

Fig. 3. This figure represents the dynamic evolution of LNWD scores for differential genes and the correlation coefficients between differential genes on the PPI network during the progression of the four diseases: KIRC, KIRP, ESCA, and LUAD.

CD-1 male mice were exposed to phosgene and air for 72 h. Air-exposed mice were used as control group, and lung tissues of phosgene-exposed mice and air-exposed mice were collected at nine time points, 0, 0.5, 1, 4, 8, 12, 24, 48, and 72 h after the exposure, with six case-group samples and six control-group samples at each sampling point[38]. The study showed that the mice were exposed to light and air at eight hours. The study showed that prominent physiological effects occurred in mice by eight hours, with elevated BALF protein levels and

increased pulmonary edema, which led to decreased survival, with mice surviving approximately 50–60% after 12 h of exposure to phosgene, and mice surviving approximately 60–70% after 24 h of exposure to phosgene.

The LNWD method was applied to the high-throughput gene expression data of this experiment, and the results showed that the LNWD scores increased dramatically at 4 ~ 8 h after exposure, (Fig. 4a) and the LNWD scores peaked at the eight-hour sampling point, indicating that it was a pre-disease state before the eight-hour period, and it would be in the disease state after the eight-hour period, and such a result was consistent with the actual experimental results. In the dynamic change network diagram consisting of mouse differential genes, (Fig. 4b) the network structure changed significantly at the eighth hour, signaling that a critical transition was imminent, consistent with the critical state identified by the LNWD method described above and previous experimental results.

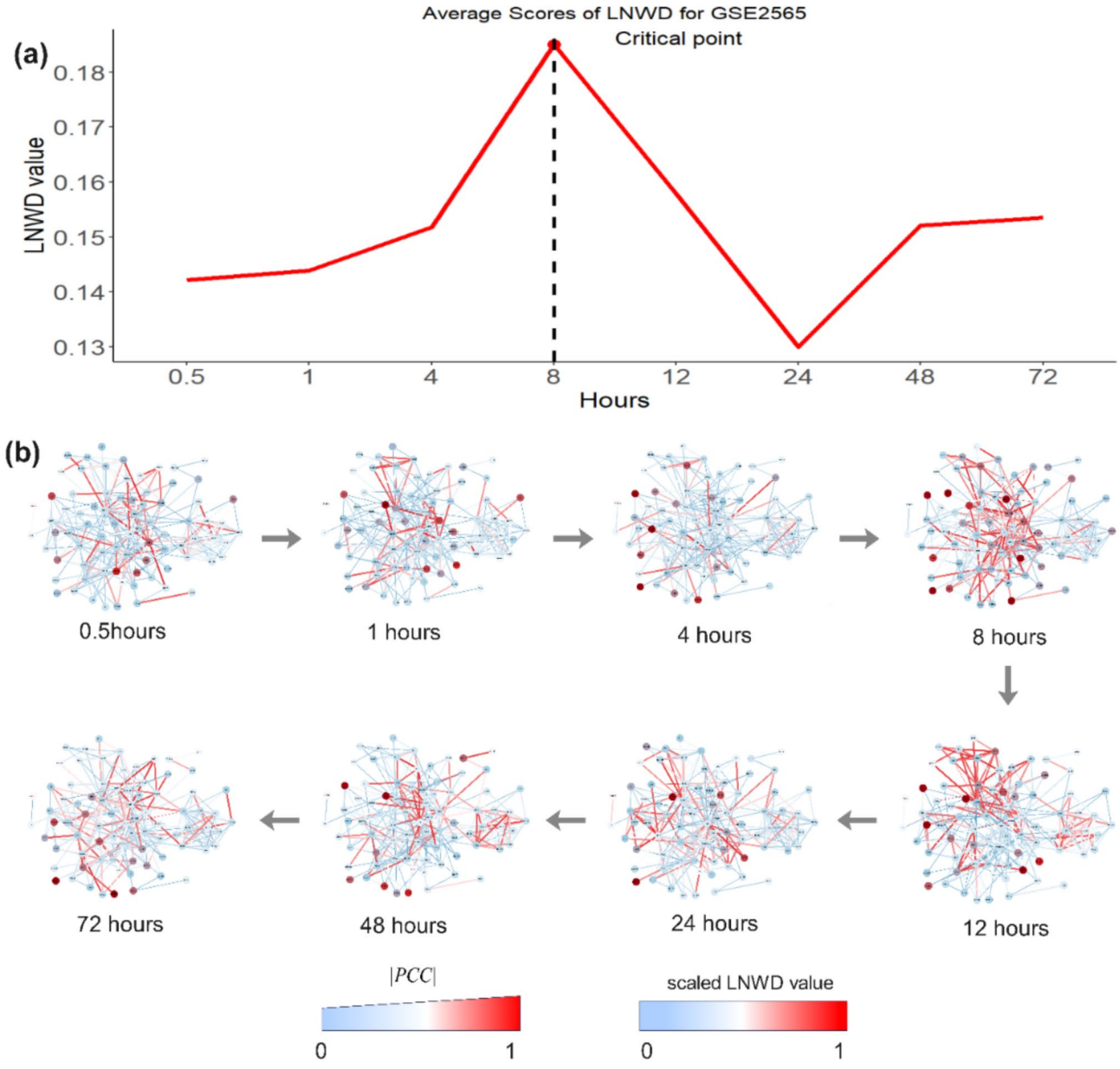

Fig. 4. Application of the LNWD method to the experimental dataset of acute lung injury in mice. (**a**) The red curve represents the significant increase in LNWD score at the eighth hour with increasing exposure time to phosgene, which can be used as an early warning signal for the deterioration of acute lung injury in mice, consistent with the real experimental results. (**b**) The figure represents the dynamic evolution of LNWD scores and correlation coefficients between differential genes on the PPI network during the course of disease development in mice.

Identifying critical transitions in type II diabetes in rats

The LNWD method was applied to microarray data from the type II diabetes dataset (GSE13268) in rat adipose tissue, which resulted in 283 differential genes, with a sample size of 10 for the normal group, the experimental group in the other time periods, and the control group corresponding to the same time period .The case and control groups for this experiment were GK (GotoKakizake) rats on normal and high-fat diets, respectively, and 50 experimental rats were executed at different ages: 4, 8, 12, 16, and 20 weeks, and 50 adipose tissue samples were ultimately collected, with each time point containing five case group samples and five control group samples, respectively[39]. The study showed that after the eighth week the GK rats had further damage to pancreatic β-cells, a sustained decline in insulin secretion, and exacerbation of diabetes mellitus[40].

Applying the LNWD method to the high-throughput gene expression data from this experiment, the results showed that there was a significant mutation in the LNWD scores of rats killed before the eighth week, (Fig. 5a)with the LNWD scores peaking at the eighth week, and the eighth week as a critical transition. In the dynamically changing network map consisting of differential genes in GK rats, (Fig. 5b) the network structure changed significantly at week eight, signaling an upcoming critical transition, consistent with the critical state identified by the LNWD method described above and previous experimental results.

In this study, KL scatter and JS scatter are applied to the above analysis process. Histograms of the mean and error bars representing the relative standard deviation of the scores obtained through the three metrics are plotted to compare the validity of the three methods(Fig. 6). By observing the error bars, it is evident that the relative standard deviation of WD is smaller, indicating that the scores of the samples obtained by this method are closer to the mean. This suggests that our identification of the critical state through the LNWD scores is more accurate, demonstrating the effectiveness of our method.

Discussion

Early intervention during the pre-disease state can effectively prevent disease progression. Therefore, it is important to detect critical warning signals of disease transformation and identify the critical state of disease progression. However, individuals usually have only one sample at a point in time, which means that multi-sample-based DNB methods are not always applicable, and thus it is not easy to recognize the premorbid state of a complex disease before obvious symptoms appear. In this paper, we propose a model-free method for detecting the critical state of complex diseases based on a single sample, i.e., the Local Network Wasserstein Distance (LNWD) method, which has been applied to several real-world datasets and has successfully identified the pre-morbid state of the disease, e.g., the critical stage of renal clear-cell carcinoma is Stage II before the tumor will spread to the main vein or spread to the perirenal tissues; the critical stage of papillary cell renal carcinoma is stage III before distant metastasis occurs; esophageal cancer tumors are stage II before distant metastasis occurs; lung adenocarcinoma is stage IIB before metastasis occurs to the ipsilateral bronchus or ipsilateral hilar lymph nodes; the critical period before elevated BALF protein levels and exacerbation of pulmonary edema in mice is the eighth hour of exposure to phosgene; further damage to pancreatic islet β-cells in the rat, the The critical period before the sustained decline in insulin secretion was the eighth week of high-fat diet in rats. All critical states identified by the TCGA dataset passed the test of survival analysis, i.e., there was a significant difference between the survival rate of the pre-critical state samples and that of the post-critical state, and the survival time of the pre-critical state samples was significantly longer than that of the post-critical state samples, and all the GEO datasets identified by the Critical states all passed the test of dynamic change of molecular network, i.e., the results of molecular network structure changed significantly at the stage of critical transition.

The method proposed in this paper has several advantages. Firstly, the method proposed in this paper is based on a single sample and will not be limited by the sample size. Secondly, the method is a model-free method, unlike traditional classification or machine learning methods, the method does not require a large number of cases or control samples to train the model and does not produce overfitting. Not only that, the Wasserstein distance used in this method is robust, less affected by small probability events, and more robust to noise. The Wasserstein distance can effectively measure the difference between two probability distributions that have almost no overlap. It can effectively avoid the problems that may be faced by the Kullback-Leibler Divergence.

Conclusions

The LNWD method proposed in this paper can effectively detect the critical point or critical state of complex diseases at the single sample level, and is suitable for high-throughput gene expression data of most complex diseases.The computational complexity of the LNWD method is relatively low, easy to implement, and the Wasserstein distance used has a more obvious advantage compared with other metrics. Therefore, this method has great potential for personalized disease diagnosis and preventive medicines.

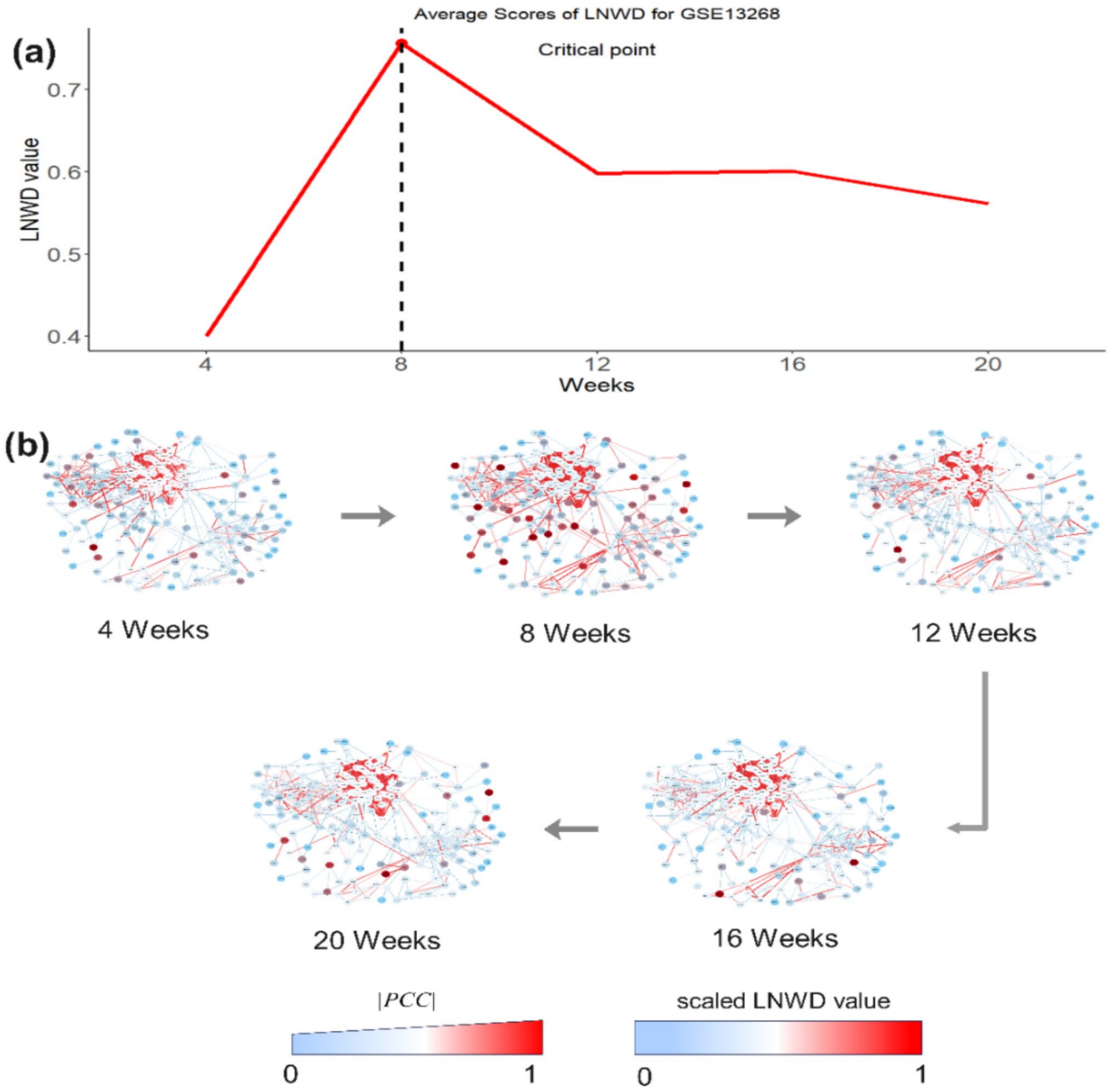

Fig. 5. Application of LNWD method on type II diabetes in rat adipose tissue dataset. (**a**) The red curve represents the significant increase in LNWD score at week 8 over time, which can be used as an early warning signal for the deterioration of type II diabetes in rat adipose tissue, consistent with the real experimental results. (**b**) The figure represents the dynamic evolution of the LNWD scores of rat differential genes and the correlation coefficients between the differential genes on the PPI network during the development of type II diabetes in rat adipose tissue.

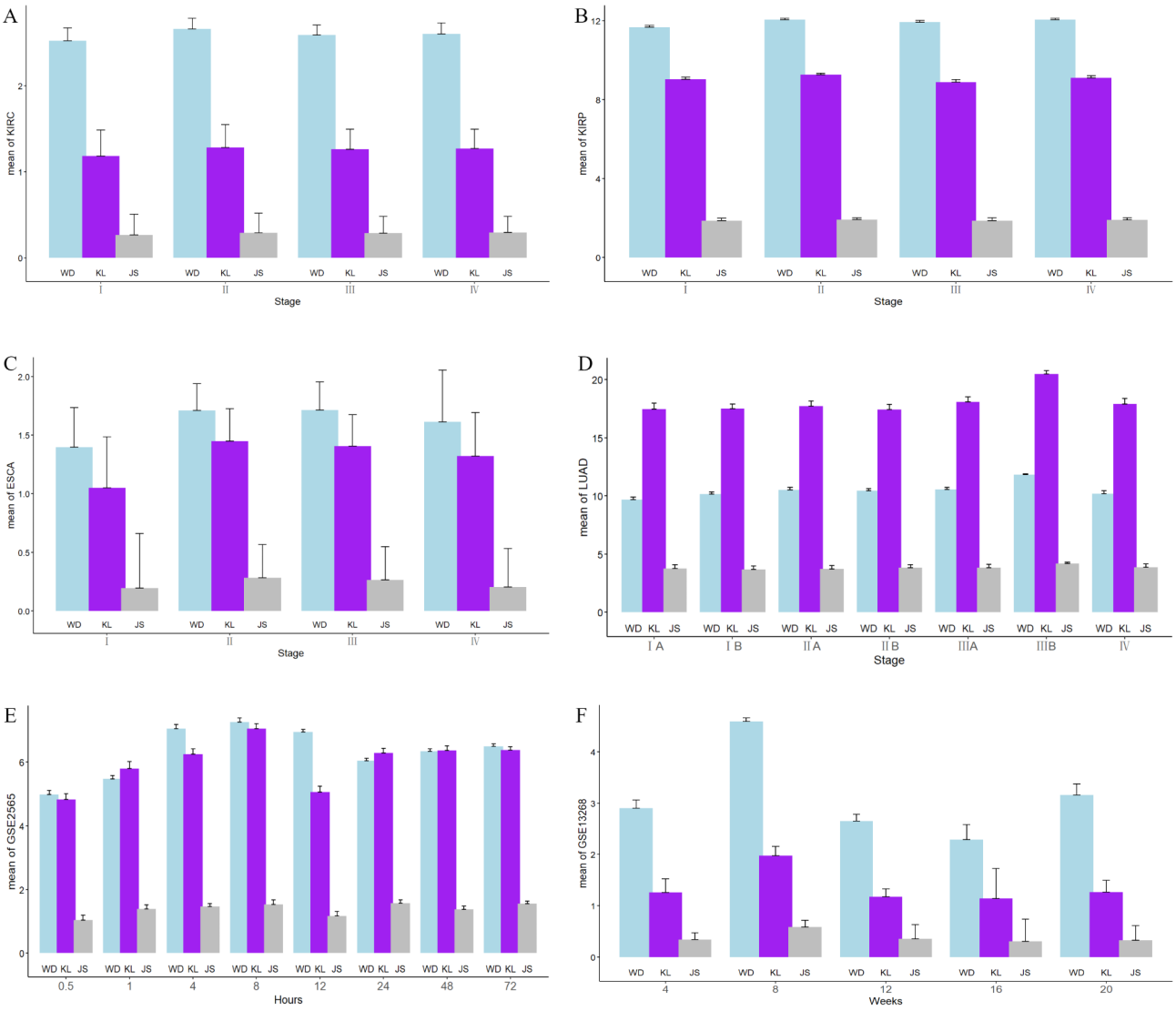

Fig. 6. Histograms of scoring means, constructed using WD, KL scatter, and JS scatter as measures, were combined with relative standard deviation error bars to visualize the characteristics and differences in each set of scoring data.

Data availability

The datasets analysed during the current study are accessible in the TCGA repository and the GEO repository. The web link for the TCGA repository is https://portal.gdc.cancer.gov/, and the web link for the GEO repository is https://www.ncbi.nlm.nih.gov/geo/.

Received: 17 October 2024; Accepted: 14 March 2025
Published online: 20 March 2025

References

1. Chen, P., Liu, R., Chen, L. & Aihara, K. Identifying critical differentiation state of MCF-7 cells for breast cancer by dynamical network biomarkers. *Front. Genet.* **6**. https://doi.org/10.3389/fgene.2015.00252 (2015).
2. Liu, X., Liu, R., Zhao, X. M. & Chen, L. Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med. Genom.* **6**, 1–10. https://doi.org/10.1186/1755-8794-6-s2-s8 (2013).
3. Chen, P., Chen, E., Chen, L., Zhou, X. J. & Liu, R. Detecting early-warning signals of influenza outbreak based on dynamic network marker. *J. Cell. Mol. Med.* **23**, 395–404. https://doi.org/10.1111/jcmm.13943 (2019).
4. Hong, R., Tong, Y., Liu, H., Chen, P. & Liu, R. Edge-based relative entropy as a sensitive indicator of critical transitions in biological systems. *J. Transl. Med.* 22, 333. (2024).
5. Chen, P., Li, Y., Liu, X., Liu, R. & Chen, L. Detecting the tipping points in a three-state model of complex diseases by Temporal differential networks. *J. Translational Med.* **15**, 1–15 (2017).
6. Aihara, K., Liu, R., Koizumi, K., Liu, X. & Chen, L. Dynamical network biomarkers: theory and applications. *Gene* 808, 145997. https://doi.org/10.1016/j.gene.2021.145997 (2022).

7. Liu, R., Chen, P. & Chen, L. Single-sample landscape entropy reveals the imminent phase transition during disease progression. *Bioinformatics* 36, 1522-32. https://doi.org/10.1093/bioinformatics/btz758 (2020).
8. Friedman, S. L. Mechanisms of hepatic fibrogenesis. *Gastroenterology* 134,1655-. https://doi.org/10.1053/j.gastro.2008.03.003 (2008).
9. Mehlen, P. & Puisieux, A. Metastasis: a question of life or death. *Nat. Rev. Cancer* 6, 449 – 58. https://doi.org/10.1038/nrc1886 (2006).
10. Zhong, J., Liu, R. & Chen, P. Identifying critical state of complex diseases by single-sample Kullback–Leibler divergence. *BMC Genom.* **21**, 1–15 (2020).
11. Liu, J., Ding, D., Zhong, J. & Liu, R. Identifying the critical states and dynamic network biomarkers of cancers based on network entropy. J. Transl. Med. 20, 254. (2022).
12. Liu, R. et al. Identifying critical transitions and their leading biomolecular networks in complex diseases. Sci. Rep. 2, 813. (2012).
13. Dakos, V. et al. Slowing down as an early warning signal for abrupt climate change. *Proc. Natl. Acad. Sci.* **105**, 14308–14312. https://doi.org/10.1073/pnas.0802430105 (2008).
14. Lesterhuis, W. J. et al. Dynamic versus static biomarkers in cancer immune checkpoint Blockade: unravelling complexity. *Nat. Rev. Drug Discovery* 16,264 – 72 (2017).
15. Liu, R. et al. Hunt for the tipping point during endocrine resistance process in breast cancer by dynamic network biomarkers. *J. Mol. Cell Biol.* 11,649 – 64. https://doi.org/10.1093/jmcb/mjy059 (2019).
16. Richard, A. et al. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biol.* 14e1002585. https://doi.org/10.1371/journal.pbio.1002585 (2016).
17. Liu, X. et al. Detection for disease tipping points by landscape dynamic network biomarkers. *Natl. Sci. Rev.* 6,775 – 85. https://doi.org/10.1093/nsr/nwy162 (2019).
18. Liu, Z. P. & Gao, R. Detecting pathway biomarkers of diabetic progression with differential entropy. *J. Biomed. Inform.* 82,143 – 53. https://doi.org/10.1016/j.jbi.2018.05.006 (2018).
19. Van Erven, T. & Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory*. https://doi.org/10.1109/TIT.2014.2320500 (2014). 60,3797 – 820.
20. Panaretos, V. M. & Zemel, Y. Statistical aspects of Wasserstein distances. Annual review of statistics and its application. 6,405 – 31. https://doi.org/10.1146/annurev-statistics-030718-104938 (2019).
21. Anderes, E., Borgwardt, S. & Miller, J. Discrete Wasserstein barycenters: optimal transport for discrete data. *Math. Methods Oper. Res.* **84**, 389–409 (2016).
22. Kiesel, R., Rühlicke, R., Stahl, G. & Zheng, J. The Wasserstein metric and robustness in risk management. *Risks* **4**. https://doi.org/10.3390/risks4030032 (2016).
23. Piccoli, B. & Rossi, F. On properties of the generalized Wasserstein distance. *Arch. Ration. Mech. Anal.* **222**, 1339–1365 (2016).
24. Kolouri, S., Rohde, G. K. & Hoffmann, H. Sliced wasserstein distance for learning gaussian mixture models. Proc. IEEE Conference on Computer Vision and Pattern Recognition.3427-36. (2018).
25. Tong, A. et al. Embedding signals on graphs with unbalanced diffusion earth mover's distance. ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).5647-51. (2022).
26. Blanco-Justicia, A. & Domingo-Ferrer, J. Privacy-Preserving computation of the Earth mover's distance. *Int. Conf. Inform. Secur.* **409**, 23 (2020).
27. Zhu, Z., Wang, L., Peng, G. & Li, S. WDA: an improved Wasserstein distance-based transfer learning fault diagnosis method. *Sensors* **21**. https://doi.org/10.3390/s21134394 (2021).
28. Liu, R., Zhong, J., Yu, X., Li, Y. & Chen, P. Identifying critical state of complex diseases by single-sample-based hidden Markov model. *Front. Genet.* https://doi.org/10.3389/fgene.2019.00285 (2019).
29. Chen, L., Liu, R., Liu, Z. P., Li, M. & Aihara, K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* **2**, 342 (2012).
30. Liu, X., Wang, Y., Ji, H., Aihara, K. & Chen, L. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res.* **44**, e164–e. https://doi.org/10.1093/nar/gkw772 (2016).
31. Duan, X., Li, B., Guo, D., Zhang, Z. & Ma, Y. A coverless steganography method based on generative adversarial network. EURASIP Journal on Image and Video Processing. 2020, 1–10. (2020).
32. Shen, J., Qu, Y., Zhang, W. & Yu, Y. Wasserstein distance guided representation learning for domain adaptation. Proc. AAAI conference on artificial intelligence. https://doi.org/10.1609/aaai.v32i1.11784 (2018).
33. Ruttenberg, B. E. & Singh, A. K. Indexing the earth mover's distance using normal distributions. Preprint at arXiv:1111.7168. (2011).
34. Gu, T. & Zhao, X. Integrating multi-platform genomic datasets for kidney renal clear cell carcinoma subtyping using stacked denoising autoencoders. *Sci. Rep.* **9**, 16668 (2019).
35. Kane, C. J., Mallin, K., Ritchey, J., Cooperberg, M. R. & Carroll, P. R. Renal cell cancer stage migration: analysis of the National cancer data base. *Cancer* **113**, 78–83. https://doi.org/10.1002/cncr.23518 (2008).
36. Wu, S. G. et al. Sites of metastasis and overall survival in esophageal cancer: a population-based study. *Cancer Manage. Res.* 781–788. (2017).
37. Robinson, L. A., Ruckdeschel, J. C., Wagner Jr, H. & Stevens, C. W. Treatment of non-small cell lung cancer-stage IIIA: ACCP evidence-based clinical practice guidelines. *Chest*. https://doi.org/10.1378/chest.07-1379 (2007). 132,243S-65S.
38. Sciuto, A. M. et al. Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation. *Chem. Res. Toxicol.* **18**. https://doi.org/10.1021/tx050126f (2005).
39. Almon, R. R. et al. Gene expression analysis of hepatic roles in cause and development of diabetes in Goto-Kakizaki rats. *J. Endocrinol.* https://doi.org/10.1677/JOE-08-0404 (2009). 200,331 – 46.
40. Mauvais-Jarvis, F. et al. A model to explore the interaction between muscle insulin resistance and beta-cell dysfunction in the development of type 2 diabetes. *Diabetes*. https://doi.org/10.2337/diabetes.49.12.2126 (2000).

Author contributions

Chang Chun Liu was responsible for writing the first draft and data processing.Ping Jun Hou was in charge of the method design of the entire article.Lin Feng was responsible for visualization.All authors participated in the revision of the article.

Funding

This research was funded by Natural Science Foundation of Henan Province (Grant Number: 242300420244).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94521-0.

Correspondence and requests for materials should be addressed to P.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.