SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2019, Vol. 62 212205:1–212205:23
https://doi.org/10.1007/s11432-019-9866-3

Data-driven group decision making for diagnosis of thyroid nodule

Chao FU[1,2*], Wenjun CHANG[1,2], Weiyong LIU[3] & Shanlin YANG[1,2]

[1]*School of Management, Hefei University of Technology, Hefei* 230009*, China;*
[2]*Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei* 230009*, China;*
[3]*Department of Ultrasound, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei* 230001*, China*

Received 2 January 2019/Revised 26 February 2019/Accepted 29 March 2019/Published online 20 September 2019

Abstract Emerging information technologies' integration into various fields has enhanced the development of these fields. Large volumes of data have been accumulated in this process. The accumulated data offer opportunities and challenges for people facing practical problems. On the one hand, it is essential to depend on a group's capabilities rather than an individual's capabilities to handle practical problems because the individual may lack sufficient expertise and experience to use data. In this situation, the practical problems can be considered as group decision making (GDM) problems. On the other hand, the accumulated data can help generate quality solutions to GDM problems. To obtain such solutions under the assumption that the accumulated data regarding a specific decision problem are available, this paper proposes a data-driven GDM method. In the method, decision makers' weights are learned from historical overall assessments and the corresponding gold standards, while criterion weights are learned from historical overall assessments and the corresponding decision matrices. The learned expert weights and criterion weights are used to produce the aggregated assessments, from which alternatives are compared or the overall conclusion is made. In a tertiary hospital located in Hefei, Anhui Province, China, the proposed method is applied to aid radiologists in diagnosing thyroid nodules.

Keywords data-driven group decision making, interval number, learning of expert weights, learning of criterion weights, diagnosis of thyroid nodule

Citation Fu C, Chang W J, Liu W Y, et al. Data-driven group decision making for diagnosis of thyroid nodule. Sci China Inf Sci, 2019, 62(11): 212205, https://doi.org/10.1007/s11432-019-9866-3

1 Introduction

In an era of the Internet and big data, almost every event is accompanied by a large amount of information and knowledge. In this situation, when a problem associated with an event is analyzed, it is usually found that generating a rational or acceptable result from a single decision maker is difficult owing to lack of sufficient expertise and experience. Thus, it is more feasible and reasonable to depend on group expertise and experience than to depend on individual expertise and experience to achieve such a goal. In real life, different group decision making (GDM) methods have been developed to handle various problems. For example, to perform supplier evaluation and categorization, Galo et al. [1] proposed a GDM method based on ELECTRE TRI; to evaluate and select green suppliers, Qin et al. [2] developed a multi-criteria group decision making (MCGDM) method in the context of interval type-2 fuzzy sets; to select hotel location, Cheng [3] proposed an MCGDM method based on interval-valued intuitionistic fuzzy sets; to

* Corresponding author (email: wls_fuchao@163.com, chaofu@hfut.edu.cn)

construct the house of quality in the process of quality function deployment, Chen et al. [4] designed a fuzzy GDM method; to select the best one from alternative marine fuels, Ren and Liang [5] proposed a fuzzy MCGDM method by combining logarithmic least squares with TOPSIS; to select a way to control the safety of ships not under control, Wu et al. [6] developed a hybrid GDM method based on TOPSIS; to evaluate and compare product prototypes in a socio-economic theme-based new product development, Lu et al. [7] proposed a fuzzy MCGDM method; and to help a car company to select a suitable supplier to purchase automobile parts, Li et al. [8] proposed a GDM method with heterogeneous information.

These studies show that different GDM methods have been developed to satisfy various application requirements. In general, one method cannot satisfy all possible requirements. In past studies, researchers focused on GDM's different perspectives, such as the aggregation of individual preference information [9–11], the convergence of group consensus [12–15], the consistency of preference relations [16–21], large scale GDM [22–24], GDM with various types of preference information [25–27], the construction of value functions of preference information [28–30], the determination of criterion weights [31–34], and the determination of decision makers' weights [35–37]. The aggregation of individual preference information is a necessary step in GDM that is usually associated with the expression of preference information. The convergence of group consensus, which can be accelerated by a feedback mechanism [16,38] is a prerequisite for generating a group-satisfactory solution in theory. In the convergence process, group analysis and discussion (GAD) helps to reach the expected group consensus, however, decision makers are encouraged to update their preference information independently [12]. When GAD and the updating of decision makers' preference information are difficult to conduct because of the limitations of time, space, and other factors, reaching the predefined group consensus becomes nearly impossible. The consistency is an inherent issue when objects (alternatives) are compared in pairs by decision makers. Large scale GDM aims to handle the situation where more than 20 decision makers are involved. It is not intended for the GDM process in which only several (e.g., 3–5) decision makers are involved [22]. Various types of preference information facilitate decision makers to flexibly provide assessments on alternatives. Their applicability depends on the characteristics of decision problems and decision-makers' choices between different formats of preference information. The purpose of constructing value functions of preference information is to facilitate a direct comparison among the preference information of different alternatives. It is particularly useful when the expression of preference information is so complex that such a direct comparison is difficult for all cases. Regarding criterion weights and decision makers' weights, due to the fact that they are a prerequisite to generating a GDM solution, they must be determined in appropriate ways in the GDM process. These analyses show that different from other requirements of GDM, criterion weights and decision-makers' weights are closely related to generating a GDM solution instead of only a group-satisfactory solution, and are not related to the characteristics of decision problems.

Owing to the significance of criterion weights and decision-makers' weights in GDM, their determination is a crucial issue in GDM's studies. In relevant studies, criterion weights and decision-makers' weights are determined by the subjective judgments of facilitator and/or decision makers, decision matrices, or a specific goal of GDM such as group consensus (see Section 2 for the detailed analysis). Little attention has been paid to determining criterion weights and decision-makers' weights from the data collected in very similar decision situations. This may be caused by the difficulty in data collection and the availability of decisions in similar situations. In practice, there are problems in which decision makers make judgments in the same decision framework. For example, radiologists in a hospital provide overall diagnoses of thyroid nodules by considering the perspectives of margin, contour, echogenicity, calcification, and vascularity. When historical data regarding such problems are accumulated and available, to determine the criterion weights and decision-makers' weights from historical data becomes a new and significant issue, and this facilitates the generation of a solution that is satisfactory to decision makers.

For decision-makers' limited capabilities to recognize all information and data related to decision problems under consideration, decision makers usually select to evaluate alternatives in an uncertain way. For example, the following ways are used to characterize decision-makers' uncertain preference information in different contexts: linguistic distributions [34], fuzzy preference relations [16], triangular fuzzy numbers [13], belief distributions [12], interval-valued intuitionistic fuzzy sets [22], and interval type-

2 fuzzy sets [23]. The selection of the expression of uncertain preference information is determined by the requirements of decision problems. Therefore, there is no uniform expression of uncertain preference information for all decision problems. In some contexts, as a simple and useful expression of uncertain preference information, interval numbers are applied in GDM [39–42], such as the evaluation and selection of suppliers, the evaluation of transnational corporations for a strategic alliance, and the diagnosis of a thyroid nodule. This shows that GDM with interval numbers is meaningful and useful in practical applications.

In this paper, a data-driven method is proposed to learn criterion weights and decision-makers' weights from historical data for MCGDM problems with interval numbers. The modeling of MCGDM problems by using interval numbers is presented first. With the assumption that historical decision data of decision-makers can be collectable under the same decision framework, the learning of expert weights from historical overall assessments of alternatives and the corresponding gold standards is discussed. Learning criterion weights from historical decision matrices and overall assessments is demonstrated from a theoretical perspective when historical decision matrices corresponding to historical overall assessments are available. Based on the elicitation of expert weights and criterion weights from historical decision data, the process of the proposed method is presented. In the process, the aggregated assessments of all alternatives, which are derived from a decision matrix, the learned expert weights, and the learned criterion weights are applied to compare these alternatives or make an overall conclusion (see Section 4). The proposed method is applied to aid radiologists, who are serving on a tertiary hospital located in Hefei, Anhui Province, China in diagnosing thyroid nodules. The effect of auxiliary diagnoses is examined using the historical examination reports and the relevant pathologic findings as gold standards in the period from 2011 to 2018.

The rest of this paper is organized as follows. In Section 2, existing studies regarding the determination of decision-makers' weights and criterion weights are reviewed to show the importance of learning them from historical decision data. The operations and distance of interval numbers are presented in Section 3. Section 4 discusses the proposed method. Furthermore, the proposed method is applied to aid radiologists in diagnosing thyroid nodules based on historical examination reports and pathologic findings in Section 5. Finally, Section 6 presents the conclusion of the paper.

2 Literature review

Two important issues in the GDM process are the determination of criterion weights and the generation of decision-makers' (or experts') weights. In existing studies, much attention has been paid to these two issues.

Regarding criterion weights, the weights can be subjectively determined by a facilitator or experts through different methods, such as point allocation [43], direct rating [44], eigenvector method [45], and the model of goal programming [46]. As analyzed by Fu et al. [47], there are advantages and disadvantages to using these methods to determine criterion weights. More importantly, different methods may result in different criterion weights [48, 49], meaning that the facilitator or experts must select a way to subjectively determine the criterion weights. To avoid subjective judgment biases toward determining criterion weights, the facilitator or experts can select to objectively derive criterion weights from decision matrices by following some principles [47]. One of the principles is that the weight of a criterion is reflected by the amount of discriminating power contained in alternatives' performances on the criterion. Following this principle, some methods have been proposed such as the methods of standard deviation [46], correlation coefficient and standard deviation integrated [50], entropy [31, 33, 49, 51], and deviation maximization [34,52]. Another principle is that criterion weights should be assigned for a specific goal. For example, in Fu and Xu's work [32], weights are assigned to criteria to achieve high solution reliability. Subjective and objective methods may be applicable in some situations. When historical data regarding decision matrices and overall assessments are available, learning criterion weights from historical data may be more reasonable than deriving criterion weights using subjective and objective

methods.

For expert weights, the weights can subjectively be specified by a facilitator based on the background, expertise, and experience of the experts. In particular, expert weights on different criteria can be different [12]. When the subjective assignment of expert weights is not easy for the facilitator, the objective assignment is adopted. Several methods have been developed to determine expert weights. In the work of Shi et al. [37], to update expert weights for the purpose of reaching group consensus, the performance values of expert behaviors in different categories are computed. In the work of Liu et al. [36], to obtain expert weights, the variance of experts' weighted assessments is minimized. In the work of Liu and Li [40], the average deviation between an expert's assessment and others' assessments is measured and used to update expert weights. In the work of Yue [53], to generate expert weights, the projection of each expert's assessment on the mean of all the experts' assessments is used. Additionally, the entropy of experts' decision matrices [54] and the similarity degree between experts' decision matrices [55] are used to assign expert weights. These studies show that the assignment of expert weights is usually lack of support of historical decision data. Although in some contexts, the subjective determination and objective assignment of expert weights are relatively reasonable, historical decision data can facilitate the obtainment of expert weights that are more consistent with experts' decision preferences.

The above analyses show that learning of criterion weights and expert weights is interesting and important for MCGDM and it should be given much attention. In Section 4, it will be discussed in detail.

3 Operations and distance of interval numbers

Let $x = [x^-, x^+] = \{r \mid x^- \leqslant r \leqslant x^+, x^-, x^+ \in \mathbb{R}\}$ denote an interval number. When $0 \leqslant x^- \leqslant x^+$, x is called a positive interval number, which is the focus of this paper. Given two positive interval numbers x and y, their operations are defined as follows.

Definition 1 ([56])**.** Suppose that x and y are two positive interval numbers. Then, their five arithmetic operations are defined as

$$x + y = [x^- + y^-, x^+ + y^+], \tag{1}$$

$$x - y = [x^- - y^+, x^+ - y^-], \tag{2}$$

$$x \cdot y = [x^- \cdot y^-, x^+ \cdot y^+], \tag{3}$$

$$x/y = [x^-/y^+, x^+/y^-], \text{ and} \tag{4}$$

$$\lambda \cdot x = [\lambda \cdot x^-, \lambda \cdot x^+], \ \lambda \geqslant 0. \tag{5}$$

According to Definition 1, given a set of positive interval numbers $\{x_1, \ldots, x_L\}$ and a set of weights $\{\theta_1, \ldots, \theta_L\}$, the weighted sum of interval numbers $x_1, \ldots, x_L$ is represented by

$$x = [x^-, x^+] = \left[\sum\nolimits_{i=1}^{L} \theta_i \cdot x_i^-, \sum\nolimits_{i=1}^{L} \theta_i \cdot x_i^+\right]. \tag{6}$$

The other three arithmetic operations shown in (2)–(4) can be similarly extended using (5). However, this is not useful in the proposed method and thus omitted here.

The distance between positive interval numbers is an important concept in the proposed method. It will be used in learning criterion weights and expert weights. In past studies, the construction of distance between positive interval numbers has attracted much attention. To facilitate the discussion on existing distance measures between positive interval numbers, suppose that $\bar{x} = 0.5 \cdot (x^- + x^+)$, $\bar{y} = 0.5 \cdot (y^- + y^+)$, $l(x) = x^+ - x^-$, and $l(y) = y^+ - y^-$. Then, several distance measures are presented in the following.

(1) Tran and Duckstein's distance measure [57]:

$$D_1^2(x, y) = \int_{-0.5}^{0.5} \int_{-0.5}^{0.5} \left\{[\bar{x} + u \cdot l(x)] - [\bar{y} + v \cdot l(y)]\right\}^2 \mathrm{d}u\mathrm{d}v$$

$$= [\bar{x} - \bar{y}]^2 + \frac{1}{12}[(l(x))^2 + (l(y))^2]. \tag{7}$$

(2) City-Block distance measure [58]:

$$D_2(x, y) = |x^- - y^-| + |x^+ - y^+|. \tag{8}$$

(3) Hausdorff distance measure [59]:

$$D_3(x, y) = \max\{|x^- - y^-|, |x^+ - y^+|\}. \tag{9}$$

(4) Ramos-Guajardo and Grzegorzewski's distance measure [60]:

$$D_4(x, y) = \sqrt{(\bar{x} - \bar{y})^2 + 0.25 \cdot \theta \cdot (l(x) - l(y)^2} \tag{10}$$

with $\theta > 0$.

(5) Wasserstein distance measure [61]:

$$\begin{aligned} D_5^2(x, y) &= \int_0^1 [(\bar{x} + u \cdot l(x)) - (\bar{y} + u \cdot l(y))]^2 \mathrm{d}u \\ &= \int_0^1 [(\bar{x} - \bar{y}) - (0.5 \cdot l(x) - 0.5 \cdot l(y)) \cdot (2u - 1)]^2 \mathrm{d}u \\ &= (\bar{x} - \bar{y})^2 + \frac{1}{12}(l(x) - l(y))^2. \end{aligned} \tag{11}$$

(6) Zhang and Liu's distance measure [56]:

$$D_6(x, y) = \frac{1}{\sqrt{2}}\sqrt{(x^- - y^-)^2 + (x^+ - y^+)^2}. \tag{12}$$

To analyze the above six distance measures, the properties that a distance measure between interval numbers should satisfy are presented.

Definition 2 ([61])**.** Let x, y, and z be three interval numbers. If $d(x, y)$ is regarded as a distance measure between x and y, then it should satisfy:

$$\text{(reflexivity)} \quad d(x, x) = 0, \tag{13}$$

$$\text{(symmetry)} \quad d(x, y) = d(y, x), \text{ and} \tag{14}$$

$$\text{(triangular inequality)} \quad d(x, y) \leqslant d(x, z) + d(z, y). \tag{15}$$

Additionally, $d(x, y) \leqslant 0$ is a basic property that is not presented in Definition 2. Using Definition 2, the six distance measures defined in (7)–(12) are analyzed. $D_1(x, y)$ defined as (7) does not satisfy the reflexivity property shown in (13). Its advantage lies in that it considers the distance between each point in one interval and any point in another interval. However, this advantage is not considered by the other five distance measures defined in (8)–(12). To consider this advantage and the reflexivity property simultaneously, Li et al. [62] proposed another distance measure between two interval numbers.

Definition 3 ([62])**.** Given two interval numbers $x = [x^-, x^+]$ and $y = [y^-, y^+]$, suppose that $\bar{x} = 0.5 \cdot (x^- + x^+)$, $\bar{y} = 0.5 \cdot (y^- + y^+)$, $l(x) = x^+ - x^-$, $l(y) = y^+ - y^-$, $d = [d^-, d^+] = x \cap y$, and $l(d) = d^+ - d^-$; then the distance between x and y is defined as

$$d(x, y) = \sqrt{I_D - I_\cap} \tag{16}$$

$$= \sqrt{(\bar{x} - \bar{y})^2 + \frac{1}{12}\big((l(x))^2 + (l(y))^2\big) - \frac{1}{6}(l(d))^2}, \tag{17}$$

where

$$I_D = \int_0^1 \int_0^1 \big((x^- + u \cdot l(x)) - (y^- + v \cdot l(y))\big)^2 \mathrm{d}u\mathrm{d}v \tag{18}$$

and

$$I_{\cap} = \int_0^1 \int_0^1 \left((d^- + u \cdot l(d)) - (d^- + v \cdot l(d))\right)^2 \mathrm{d}u\mathrm{d}v. \tag{19}$$

The distance measure shown in Definition 3 is verified to satisfy the properties of reflexivity, symmetry, and triangular inequality listed in Definition 2 [62]. However, the proof is not very clear. To make it clear, a relevant lemma is presented and a formal proof is offered.

Lemma 1. Given three interval numbers $x = [x^-, x^+],\ y = [y^-, y^+]$, and $z = [z^-, z^+]$, suppose that $l = [l^-, l^+] = x \cap y,\ m = [m^-, m^+] = y \cap z$, and $n = [n^-, n^+] = x \cap z$; then it is satisfied that

$$-(n^+ - n^-)^2 \leqslant (y^+ - y^-)^2 - (l^+ - l^-)^2 - (m^+ - m^-)^2. \tag{20}$$

This lemma is proved in the Appendix. Based on Lemma 1, it can be verified that the distance measure listed in Definition 3 satisfies three properties listed in Definition 2. Thus, this is presented in the following theorem.

Theorem 1. Given three interval numbers $x = [x^-, x^+],\ y = [y^-, y^+]$, and $z = [z^-, z^+]$, the distance measure listed in Definition 3 satisfies the three properties listed in Definition 2.

This theorem is formally proved in the Appendix. Next, the operations of interval numbers listed in (1)–(6) and their distance measure listed in Definition 3 will be used in the proposed method.

4 Proposed method

In this section, MCGDM problems are modeled using positive interval numbers. Based on the operations and distance of positive interval numbers presented in Section 3, criterion weights and expert weights are learned from historical data and used to generate solutions to the modeled problems.

4.1 Modeling of MCGDM problems with positive interval numbers

Suppose that T experts $(t_j,\ j = 1, \ldots, T)$ face a common problem in which each alternative A_l $(l = 1, \ldots, M)$ is evaluated on L criteria $(e_i,\ i = 1, \ldots, L)$ by using positive interval numbers. To analyze the problem, each expert t_j offers an interval-valued decision matrix, which is

$$I^j_{L\times M} = (I^j_{i,l})_{L\times M} = \begin{bmatrix} [I^{j-}_{1,1}, I^{j+}_{1,1}] & \ldots & [I^{j-}_{1,l}, I^{j+}_{1,l}] & \ldots & [I^{j-}_{1,M}, I^{j+}_{1,M}] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [I^{j-}_{i,1}, I^{j+}_{i,1}] & \ldots & [I^{j-}_{i,l}, I^{j+}_{i,l}] & \ldots & [I^{j-}_{i,M}, I^{j+}_{i,M}] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [I^{j-}_{L,1}, I^{j+}_{L,1}] & \ldots & [I^{j-}_{L,l}, I^{j+}_{L,l}] & \ldots & [I^{j-}_{L,M}, I^{j+}_{L,M}] \end{bmatrix}. \tag{21}$$

To facilitate the analysis of the problem, $[I^{j-}_{i,l}, I^{j+}_{i,l}]$ is normalized to be limited to $[0, 1]$, which means that $0 \leqslant I^{j-}_{i,l} \leqslant I^{j+}_{i,l} \leqslant 1$. Assume that the relative weight of expert t_j is denoted by λ_j and the relative weight of criterion e_i is denoted by w_i. To generate a solution to the problem by integrating the opinions of T experts, λ_j and w_i should be determined. In general, $I^j_{i,l}$ is combined using w_i to generate the aggregated assessment I^j_l and the expert t_j is not required to offer I^j_l. However, this is not always the case. For example, when a radiologist diagnoses an inpatient's nodule to be malignant or benign, he or she is required to offer the overall diagnosis by comprehensively considering the observations on different criteria. Based on the expertise and experience of experts or standards commonly accepted in the decision field, the transformation from observations on criteria to assessments can be implemented.

As presented in Section 2, expert weights can be subjectively specified by a facilitator based on the background, expertise, and experience of experts or objectively determined using decision matrices. While criterion weights can be subjectively determined using some methods on the condition that experts can offer preferences as to criteria or objectively generated from decision matrices following some principles,

the generation of expert weights and the determination of criterion weights in existing studies are effective and rational to some extent from a traditional viewpoint. The learning of expert weights and criterion weights from historical data is meaningful when historical decision data are available. The purpose of MCGDM is to find a group solution that reflects the preferences of each expert for each alternative. These preferences are closely related to experts' historical decision data. For this reason, expert weights and criterion weights learned from experts' historical data are more consistent with the preferences of the experts than those generated using traditional methods. In the following, we discuss the learning of expert weights and criterion weights from historical decision data.

4.2 Learning of expert weights

Suppose that the historical decision data of experts can be collectable under the same decision framework. In other words, different alternatives are evaluated on a uniform set of criteria in historical decision processes. With the assumption that historical overall assessments I_k^j $(k = 1, \ldots, K_j)$ with K_j alternatives are obtainable, the learning of expert weights from the overall assessments is discussed.

To learn expert weights, a clear understanding of the weight of an expert is required. In MCGDM, the weight of an expert is positively correlated with his or her capability to make a correct decision. The stronger the capability of an expert, the higher the weight of the expert. This idea is presented in the following assumption.

Assumption 1. In MCGDM, there is a positive correlation between the weight of an expert and the expert's capability to offer correct judgments.

From the idea listed in Assumption 1, an expert's capability to make a correct decision is expected to be measured to determine expert weights. The focus is the indication of a correct decision. The gold standard is a reliably and precisely objective measure to judge the correctness of the thing to be examined and, thus, it can be considered as the indication. In some situations where the gold standards are available, the assessment of an expert that is completely consistent with gold standard can be regarded as a correct decision. For example, a radiologist depends on pathologic findings as gold standards to verify the correctness of the diagnosis of a thyroid nodule. Suppose that the gold standards corresponding to historical overall assessments I_k^j $(k = 1, \ldots, K_j)$ are obtained as $\vec{I}_k^j$ $(k = 1, \ldots, K_j)$. The average similarity between I_k^j and $\vec{I}_k^j$ is calculated as

$$S^j = 1 - \frac{\sum_{k=1}^{K_j} d(I_k^j, \vec{I}_k^j)}{K_j}, \tag{22}$$

where $d(I_k^j, \vec{I}_k^j)$ represents the distance between I_k^j and $\vec{I}_k^j$ that is calculated using Definition 3. For normalized I_k^j and $\vec{I}_k^j$, Eqs. (16) and (17) show that $0 \leqslant d(I_k^j, \vec{I}_k^j) \leqslant 1$, which means that $0 \leqslant S^j \leqslant 1$. According to Assumption 1, the weight of expert t_j is derived from S^j, which is

$$\lambda^j = \frac{S^j}{\sum_{k=1}^{T} S^k}. \tag{23}$$

4.3 Learning of criterion weights

To learn expert weights, historical overall assessments I_k^j $(k = 1, \ldots, K_j)$ with K_j alternatives are assumed to be obtainable. Meanwhile, as presented in Subsection 4.1, historical observations on criteria can be transformed into assessments based on the expertise and experience of experts or standards commonly accepted in the decision field. This means that a historical matrix $I_{L\times K_j}^j$ with K_j alternatives can be assumed to be obtainable. On the condition that historical overall assessments I_k^j $(k = 1, \ldots, K_j)$ and a historical matrix $I_{L\times K_j}^j$ are available, the learning of criterion weights from them is discussed.

Similar to expert weights, to learn criterion weights, a clear understanding of the weight of a criterion is required. In a general case, I_k^j is generated from combining $I_{i\times k}^j$ $(i = 1, \ldots, L)$ using w_i. The larger the w_i, the higher the similarity between $I_{i\times k}^j$ and I_k^j. The converse conclusion may not be always true. To construct a positive correlation between the similarity between $I_{i\times k}^j$ and I_k^j and w_i, assume that the experts do not offer subjective preferences as to criteria.

Assumption 2. For an MCGDM problem, each expert does not offer subjective preferences for the assignment of criterion weights and the assessment on each criterion is involved in the overall assessment.

Based on Assumption 2, a positive correlation between the similarity between $I^j_{i\times k}$ and I^j_k and w_i can be constructed.

Assumption 3. On the condition that Assumption 2 is satisfied in MCGDM, there is a positive correlation between the weight of a criterion and the similarity between the assessment on the criterion and the overall assessment.

To determine criterion weights based on the idea presented in Assumption 3, the similarity between $I^j_{i\times k}$ and I^j_k is calculated as

$$S^j_{i,k} = 1 - d(I^j_{i\times k}, I^j_k), \tag{24}$$

where $d(I^j_{i\times k}, I^j_k)$ represents the distance between $I^j_{i\times k}$ and I^j_k that is calculated using Definition 3. According to Assumption 3, the weight of criterion e_i is derived from $S^j_{i,k}$, which is

$$w^j_{i,k} = \frac{S^j_{i,k}}{\sum_{h=1}^{L} S^j_{h,k}}. \tag{25}$$

When $(w^j_{1,k}, \ldots, w^j_{L,k})$ $(k = 1, \ldots, K_j)$ is obtained, the remaining problem is to generate a representative set of criterion weights $(w^j_1, \ldots, w^j_L)$ that can reflect the preferences of the expert t_j as to the criterion weights, i.e., $(w^j_{1,k}, \ldots, w^j_{L,k})$ $(k = 1, \ldots, K_j)$. For this purpose, an optimization model is constructed.

$$\min \quad f(w^j_1, \ldots, w^j_L) = \sum\nolimits_{k=1}^{K_j} \sum\nolimits_{i=1}^{L} (w^j_{i,k} - w^j_i)^2 \tag{26}$$

$$\text{s.t.} \quad \sum\nolimits_{i=1}^{L} w^j_i = 1, \tag{27}$$

$$0 \leqslant w^j_i \leqslant 1, \ i = 1, \ldots, L. \tag{28}$$

In this model, the objective function is nonlinear and the constraints are linear. As a result, the model can be regarded as a nonlinear programming problem (NLP) [63]. To facilitate the generation of a solution to this optimization model, solving the generalized formulation of an NLP is discussed. The generalized formulation of an NLP is given as follows.

$$\min \quad f(a_1, \ldots, a_N) \tag{29}$$

$$\text{s.t.} \quad g_m(a_1, \ldots, a_N) = b_m, \ m = 1, \ldots, M, \tag{30}$$

where $f(a_1, \ldots, a_N)$ and $g_m(a_1, \ldots, a_N)$ $(m = 1, \ldots, M)$ are real-valued functions with at least one of them being nonlinear.

There are necessary and sufficient conditions to determine whether a vector is an optimal solution to the NLP shown in (29) and (30), which can be seen in Theorems A.1 and A.2 of the Appendix. Meanwhile, Theorem A.3 is presented in the Appendix to help discuss the uniqueness of a solution. With the aid of the three theorems, a unique optimal solution to the optimization model shown in (26)–(28) can be obtained.

Theorem 2. For the optimization model shown in (26)–(28),

$$(\bar{w}^j_1, \ldots, \bar{w}^j_L) = \left(\frac{\sum_{k=1}^{K_j} w^j_{1,k}}{K_j}, \ldots, \frac{\sum_{k=1}^{K_j} w^j_{L,k}}{K_j} \right) \tag{31}$$

is the unique optimal solution.

Theorem 2 is proved in the Appendix. Using the learned expert weights and criterion weights, a group-satisfactory solution can be generated. The process is discussed in the following subsection.

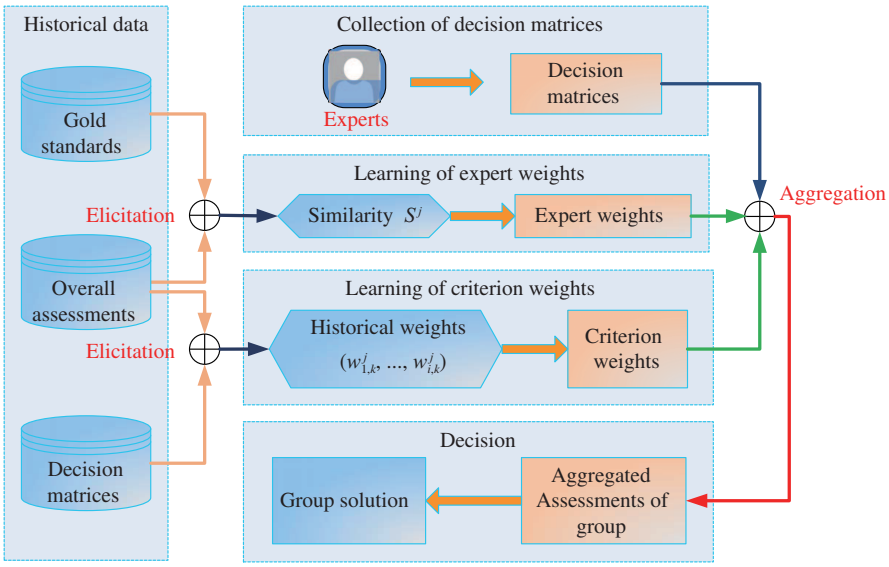

Figure 1 (Color online) MCGDM process for the proposed method.

4.4 MCGDM process

Based on the learning of expert weights and the learning of criterion weights, the process of the proposed method is described in Figure 1.

A facilitator invites T experts to evaluate M alternatives on L criteria. Experts are encouraged to use normalized positive interval numbers to implement the evaluation. For each expert, a historical decision matrix and historical overall assessments with corresponding gold standards are collected. From the historical overall assessments of the expert t_j and the corresponding gold standards, the average similarity between overall assessments and gold standards S^j is obtained using (22). Then using (23), the average similarities of all experts are used to determine the weights of the experts λ^j $(j = 1, \ldots, T)$. For the expert t_j, a historical decision matrix and the historical overall assessments are used to generate K_j sets of historical weights $(w^j_{1,k}, \ldots, w^j_{L,k})$ $(k = 1, \ldots, K_j)$ using (24) and (25). A representative set of criterion weights $(\bar{w}^j_1, \ldots, \bar{w}^j_L)$ that can reflect the preferences of the expert t_j as to criteria is derived from $(w^j_{1,k}, \ldots, w^j_{L,k})$ $(k = 1, \ldots, K_j)$ through the optimization model presented in (26)–(28) and Theorem 2.

When expert weights and criterion weights of each expert are learned from a historical decision matrix and historical overall assessments with corresponding gold standards, the decision matrix of each expert can be aggregated to generate the aggregated group assessment of each alternative. Using the learned expert weights λ^j $(j = 1, \ldots, T)$ and (6), the decision matrix of each expert is first aggregated to generate the group decision matrix, which is

$$I_{L\times M} = (I_{i,l})_{L\times M} = \left(\sum\nolimits_{j=1}^{T} \lambda_j \cdot I^j_{i,l}\right)_{L\times M}. \tag{32}$$

Considering comprehensively the learned expert weights λ^j $(j = 1, \ldots, T)$ and the learned criterion weights of each expert $(\bar{w}^j_1, \ldots, \bar{w}^j_L)$ $(j = 1, \ldots, T)$, a set of criterion weights for the group is obtained as

$$w_i = \sum\nolimits_{j=1}^{T} \lambda_j \cdot w^j_i, \quad i = 1, \ldots, L. \tag{33}$$

Because $\sum_{j=1}^{T} \lambda_j = 1$ and $\sum_{i=1}^{L} w^j_i = 1$, $\sum_{i=1}^{L} w_i = 1$ clearly holds. Using (6) and w_i obtained in (33), $I_{i,l}$ $(i = 1, \ldots, L)$ is aggregated as I_l; i.e.,

$$I_l = \sum\nolimits_{i=1}^{L} w_i \cdot I_{i,l}, \quad l = 1, \ldots, M. \tag{34}$$

The aggregated assessments I_l $(l = 1, \ldots, M)$ can then be used to generate a group solution. From a traditional perspective, to determine the best one or generate a ranking order of the alternatives, which

is considered as a group solution, different alternatives need to be compared using I_l $(l = 1, \ldots, M)$. This is a general purpose of MCGDM. However, this is not the only purpose. Decision makers may attempt to make an overall conclusion using I_l $(l = 1, \ldots, M)$. For example, a group of radiologists work together to provide the diagnosis of a patient's thyroid nodule. Such a situation usually occurs in real life and can be considered as another purpose of MCGDM. The generation of a group solution for the two purposes is discussed in the following.

(1) Purpose of comparing alternatives. When the purpose of MCGDM is to compare different alternatives by considering group opinions, I_l and I_m $(l \neq m)$ need to be compared. For this purpose, the mean and standard variance of $I_l = [I_l^-,\ I_l^+]$ denoted by μ_l and σ_l are used to construct the score function of the alternative A_l. Suppose that a random variable V on I_l follows the uniform distribution. Then, μ_l and σ_l are calculated by

$$\mu_l = E(v) = \frac{I_l^- + I_l^+}{2} \tag{35}$$

and

$$\sigma_l = \sqrt{E(V^2) - (E(V))^2} = \sqrt{\int_{I_l^-}^{I_l^+} \frac{v^2}{I_l^+ - I_l^-} \mathrm{d}v - \left(\frac{I_l^- + I_l^+}{2}\right)^2} = \frac{I_l^+ - I_l^-}{\sqrt{12}}. \tag{36}$$

Using μ_l and σ_l defined in (35) and (36), respectively, the score function of alternative A_l is constructed as

$$S(A_l) = \mu_l(1 - \sigma_l) = \frac{I_l^- + I_l^+}{2} \cdot \left(1 - \frac{I_l^+ - I_l^-}{\sqrt{12}}\right). \tag{37}$$

The constructed score function satisfies some properties.

Property 1. Suppose that the overall assessments of two alternatives A_l and A_m are obtained as $I_l = [I_l^-, I_l^+]$ and $I_m = [I_m^-, I_m^+]$. The score functions of A_l and A_m calculated using (37) satisfy:

(1) If $I_l^- = I_m^-$ and $I_l^+ > I_m^+$, then $S(A_l) > S(A_m)$;

(2) If $I_l^+ = I_m^+$ and $I_l^- > I_m^-$, then $S(A_l) > S(A_m)$;

(3) If $\frac{I_l^- + I_l^+}{2} = \frac{I_m^- + I_m^+}{2}$ and $I_l^+ - I_l^- < I_m^+ - I_m^-$, then $S(A_l) > S(A_m)$; and

(4) $0 \leqslant S(A_l),\ S(A_m) \leqslant 1$.

The conclusion presented in Property 1 can be clearly verified using (37) and, thus, the relevant proof is omitted. Using $S(A_l)$ $(l = 1, \ldots, M)$, all alternatives are compared to generate a group solution.

(2) Purpose of making an overall conclusion. In some contexts, such as diagnosis of a thyroid nodule, MCGDM attempts to draw an overall conclusion. The conclusion is usually drawn from the overall assessments of alternatives. According to the imaging reporting and data system (TIRADS) [64–67] in the diagnosis of a thyroid nodule, the cancer risk of the detected thyroid nodule is divided into five grades {3, 4A, 4B, 4C, 5}. Each grade is represented by a normalized positive interval number. When multiple radiologists diagnose a thyroid nodule, a group solution obtained is the group diagnosis of the nodule that is represented by one of the five grades. From a theoretical perspective, suppose that there is a set of grades $\{G_1, \ldots, G_N\}$ with the corresponding normalized positive interval numbers $\{[H_1^-, H_1^+], \ldots, [H_N^-, H_N^+]\}$. Under this assumption, the grade to which each alternative is determined to belong is considered as the group solution. For this purpose, the distance between $I_l = [I_l^-, I_l^+]$ and $[H_n^-, H_n^+]$ $(n = 1, \ldots, N)$, i.e., $d([I_l^-, I_l^+], [H_n^-, H_n^+])$, is calculated using Definition 3. Assume that

$$\hat{n} = \arg\min\{n \mid d([I_l^-, I_l^+], [H_n^-, H_n^+])\}; \tag{38}$$

then alternative a_l is evaluated as grade $G_{\hat{n}}$.

5 Case study

Thyroid nodule is a frequently-occurring disease in the general population, especially in adults. In clinical practice, radiologists can depend on imaging techniques to identify thyroid nodules [65]. Among practical imaging techniques, ultrasonic examination, a technique without radiation is the first imaging

mode to identify and diagnose thyroid nodules [68]. Through ultrasonic examination, a radiologist offers the diagnosis of a detected nodule used to show that the nodule is malignant or benign, from which a clinician determines the treatment of the nodule. Pathologic examination can correctly indicate whether a thyroid nodule is cancerous or benign, and, thus, pathologic findings are considered as gold standards for diagnosing thyroid nodules. A large number of diagnoses and pathologic findings have been accumulated in the clinical practice of a hospital. Accumulated data can help radiologists offer consistent diagnoses. More importantly, to improve diagnostic capability, a radiologist without a wealth of experience can learn experience from historical data, which are collected from radiologists with sufficient experience. Using the proposed method, the improvement of the diagnostic capability of a radiologist without a wealth of experience through historical data collected from radiologists with sufficient experience is examined. Historical examination reports and pathologic findings are collected from a tertiary hospital located in Hefei, Anhui Province, China with the help of the third author. A solution system developed in the Matlab environment is used to support the examination process.

5.1 Diagnosis of thyroid nodules

When a radiologist diagnoses thyroid nodules through ultrasonic examination, he or she observes the features of the nodules. Possible features include margin, contour, size, vascularity, calcification, tallness, halo, echogenicity, and solid component [66,69–72]. In clinical practice, a radiologist usually selects some of these possible features to achieve observations and offer an overall diagnosis of a detected nodule. Different sets of features are selected by radiologists who serve in different hospitals located in different regions. Given a set of features, a radiologist offers the diagnosis of a thyroid nodule by considering the observations on all features comprehensively. When each feature is considered as a criterion, the diagnosis of a thyroid nodule can be regarded as a multi-criteria decision making problem.

When radiologists diagnose thyroid nodules, their overall diagnoses represent the cancer risk of the nodules. Based on clinical practice, the TIRADS has been developed to characterize the cancer risk. There is no international standard of TIRADS although many TIRADSs have been proposed in existing studies. It is mostly accepted that the cancer risk of the detected thyroid nodule is divided into five categories {TIRADS 3, TIRADS 4A, TIRADS 4B, TIRADS 4C, TIRADS 5} with five risk intervals $\{[0\%, 3\%),\ [3\%, 24\%],\ [25\%, 75\%],\ [76\%, 95\%],\ (95\%, 100\%]\}$ [64, 68]. From a discussion with the third author, it is known that clinicians may find difficulties in correctly treating nodules when the categories 4A and 4B are used as the overall diagnoses of the nodules. To address this issue, in the tertiary hospital located in Hefei, Anhui Province, China, where the third author serves, categories 4A and 4B are further divided into two subgrades and three subgrades, respectively. This is found by checking historical examination reports in the period from 2011 to 2018. Table 1 shows the relevant details regarding the TIRADS used in the ultrasonic department of the hospital. Here, FNAB represents fine needle aspiration biopsy. As shown in Table 1, there is a correspondence between TIRADS category and cancer risk. When radiologists use TIRADS categories to offer the overall diagnoses of nodules, they actually aim to describe the cancer risk of the nodules. Based on this fact, interval numbers limited to $[0,\ 1]$ are used to describe the overall diagnoses of nodules in the case study.

Through discussing with the third author and analyzing historical examination reports in the period from 2011 to 2018, five criteria used by radiologists in the ultrasonic department of the hospital are identified. They are margin, contour, echogenicity, calcification, and vascularity and are denoted by e_i $(i = 1, \ldots, 5)$. The observations on the five criteria in the hospital are collected from the historical examination reports, which can be seen in [68]. In clinical practice, the radiologists transform observations on the five criteria into TIRADS categories and then using criterion weights w_i $(i = 1, \ldots, 5)$, they combine the categories to generate the overall TIRADS category, which is the overall diagnosis of a detected thyroid nodule. For each radiologist, the transformation is dependent on his or her expertise and experience. To consider the ultrasonic department as a whole, the common expertise and experience of the department are generated from the third author communicating with representative radiologists. Using the common expertise and experience of the department, the relationship between the observations on the

Table 1 TIRADS categories applied in the hospital

Category	Finding	Cancer risk	Recommendation
TIRADS 3	Probably benign	<3%	Follow-up/FNAB
TIRADS 4	Suspicious	3%–75%	
TIRADS 4A	Low suspicion	3%–24%	
TIRADS 4A-1	Tending towards benign nodule	3%–15%	Follow-up/FNAB
TIRADS 4A-2	Not excluding the possibility of malignant nodule	16%–24%	FNAB
TIRADS 4B	Intermediate suspicion	25%–75%	
TIRADS 4B-1	Not excluding the possibility of benign nodule	25%–40%	FNAB
TIRADS 4B-2	Medium possibility of malignant nodule	41%–65%	FNAB
TIRADS 4B-3	Large possibility of malignant nodule	66%–75%	FNAB
TIRADS 4C	High suspicion	76%–95%	FNAB
TIRADS 5	Suggestive of malignancy	> 95%	FNAB

Table 2 Details about the eight radiologists

Radiologist	Serving period	Diagnostic record
D_1	2013–2018	591
D_2	2011–2018	586
D_3	2012–2018	628
D_4	2015–2018	397
D_5	2013–2017	179
D_6	2011–2016	180
D_7	2017–2018	202
D_8	2018–2018	93

five criteria and the TIRADS categories shown in Table 1 is constructed, as shown in [68]. Furthermore, based on the correspondence between TIRADS categories and cancer risk, as shown in Table 1, the observations on the five criteria are transformed into assessments represented by interval numbers limited to $[0, 1]$.

Pathologic findings are gold standards of overall diagnoses of thyroid nodules. They can correctly determine whether nodules are malignant or benign. When the pathologic finding of a nodule shows that the nodule is malignant, the finding is represented by an interval number $[1, 1]$. On the contrary, when a nodule is judged to be benign, the finding is represented by an interval number $[0, 0]$. Because the overall diagnoses of nodules and the corresponding pathologic findings are represented by interval numbers limited to $[0, 1]$, the average similarity between the overall diagnoses and the pathologic findings for a radiologist, which is calculated using (22), can be used to characterize the diagnostic capability of the radiologist. Therefore, different radiologists in the department have different diagnostic capabilities. For a radiologist without a wealth of experience, an important issue is to help improve his or her diagnostic capability through accumulated data collected from radiologists with sufficient experience. To address this issue, the proposed method is applied to comprehensively consider the advice of radiologists with a wealth of experience, which is reflected by their historical data, to generate diagnostic recommendations for a radiologist without sufficient experience.

5.2 Generation and examination of diagnostic recommendations

To apply the proposed method to generate diagnostic recommendations for a radiologist who is of insufficient experience, the historical examination reports of eight radiologists in the department and the corresponding pathologic findings in the period from 2011 to 2018 are collected. Suppose that the eight radiologists are denoted by D_k $(k = 1, \ldots, 8)$. Table 2 presents the periods the eight radiologists served on the department and the numbers of their diagnostic records.

All the records shown in Table 2 are collected from diagnosing inpatients. Using (22), the diagnostic capabilities of the eight radiologists are obtained from the overall diagnoses of nodules and

Table 3 Weights of the five criteria for the eight radiologists

Radiologist	Learned criterion weight
t_1	$w_i^1\ (i=1,\ldots,5)=(0.1893,0.2386,0.1798,0.2119,0.1804)$
t_2	$w_i^2\ (i=1,\ldots,5)=(0.1901,0.2412,0.1778,0.2164,0.1745)$
t_3	$w_i^3\ (i=1,\ldots,5)=(0.2005,0.2352,0.188,0.1938,0.1825)$
$\widetilde{t}_1$	$\widetilde{w}_i^1\ (i=1,\ldots,5)=(0.2002,0.2287,0.1745,0.2115,0.1851)$
$\widetilde{t}_2$	$\widetilde{w}_i^2\ (i=1,\ldots,5)=(0.1928,0.2238,0.1852,0.216,0.1822)$
$\widetilde{t}_3$	$\widetilde{w}_i^3\ (i=1,\ldots,5)=(0.1981,0.2205,0.2126,0.1911,0.1778)$
$\widetilde{t}_4$	$\widetilde{w}_i^4\ (i=1,\ldots,5)=(0.2027,0.2314,0.1668,0.2291,0.1699)$
$\widetilde{t}_5$	$\widetilde{w}_i^5\ (i=1,\ldots,5)=(0.1811,0.251,0.1713,0.2276,0.169)$

the corresponding pathologic findings associated with the records of the eight radiologists, which are $(0.7843, 0.7958, 0.7505, 0.7732, 0.7317, 0.7451, 0.7445, 0.7025)$. The radiologists D_1, D_2 and D_4 are clearly the top three radiologists. They are selected as a group to generate diagnostic recommendations for the other five radiologists and are denoted by t_j $(j = 1, 2, 3)$. The remaining radiologists are denoted by $\widetilde{t}_k$ $(k = 1,\ldots,5) = \{D_3, D_5, D_6, D_7, D_8\}$. Their diagnostic capabilities are denoted by $\widetilde{C}_k^{R_0}$ $(k = 1,\ldots,5) = (0.7505, 0.7317, 0.7451, 0.7445, 0.7025)$. According to (23), the weights of the three radiologists in the group are obtained as λ^j $(j = 1, 2, 3) = (0.3332, 0.3382, 0.3286)$.

As demonstrated in Subsection 5.1, using the common expertise and experience of the department, the assessments on the five criteria can be derived from the observations presented in the records of the eight radiologists. Owing to the correspondence between TIRADS categories and cancer risk presented in Table 1, they are represented by interval numbers limited to $[0, 1]$. From the assessments on the five criteria and the overall assessments provided by each of the eight radiologists in their records, the weights of the five criteria for each of the eight radiologists, i.e., w_i^j $(j = 1, 2, 3,\ i = 1,\ldots,5)$ and $\widetilde{w}_i^k$ $(k = 1,\ldots,5,\ i = 1,\ldots,5)$ are learned according to Subsection 4.3, as presented in Table 3. Using the weights of radiologists t_j $(j = 1, 2, 3)$, i.e., λ^j, their learned criterion weights presented in Table 3, and (33), a set of criterion weights for the group is generated, which is w_i $(i = 1,\ldots,5) = (0.1932, 0.2384, 0.1818, 0.2075, 0.1791)$. By combining the assessments on the five criteria associated with the records of the radiologists $\widetilde{t}_k$ $(k = 1,\ldots,5)$ using w_i $(i = 1,\ldots,5)$ and (34), the aggregated assessments associated with the records are obtained. Based on the aggregated assessments and the correspondence between TIRADS categories and cancer risk, as shown in Table 1, the recommended TIRADS categories for the nodules in the records are generated using (38). From the cancer risk corresponding to the recommended categories and the pathologic findings corresponding to the nodules, the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1,\ldots,5)$ based on w_i $(i = 1,\ldots,5)$ are calculated using (22), which are $\widetilde{C}_k^{G_w}$ $(k = 1,\ldots,5) = (0.6857, 0.6737, 0.7522, 0.5841, 0.5245)$. It can be observed that $\widetilde{C}_k^{G_w} < \widetilde{C}_k^{R_0}$ $(k = 1, 2, 4, 5)$.

According to a discussion with the third author, it is found that the diagnostic capability of a radiologist is not only dependent on the weights of the five criteria, but also associated with the distributions of overall diagnoses provided by the radiologist on the TIRADS categories T_c $(c = 1,\ldots,8) = \{$TIRADS 3, TIRADS 4A-1, TIRADS 4A-2, TIRADS 4B-1, TIRADS 4B-2, TIRADS 4B-3, TIRADS 4C, TIRADS 5$\}$. Based on the group's criterion weights, there is lack of consideration of the distributions of overall diagnoses, which results in the lower diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, 2, 4, 5)$. This reflects the difference between the diagnosis of thyroid nodules and traditional GDM problems. Based on the expertise and experience of the third author, TIRADS categories T_c $(c = 4,\ldots,8) = \{$TIRADS 4B-1, TIRADS 4B-2, TIRADS 4B-3, TIRADS 4C, TIRADS 5$\}$ indicate malignant nodules, while TIRADS categories T_c $(c = 1, 2, 3) = \{$TIRADS 3, TIRADS 4A-1, TIRADS 4A-2$\}$ indicate benign nodules. The higher the degree to which a radiologist is sure about a malignant nodule, the larger the possibility that the radiologist prefers the TIRADS category with high cancer risk. Conversely, the higher the degree to which a radiologist is sure about a benign nodule, the larger the possibility that the radiologist prefers the TIRADS category with low cancer risk.

Based on the sufficient consideration of such preferences of radiologists and the professional sugges-

Table 4 Distributions of the overall diagnoses on the TIRADS categories for the eight radiologists

Radiologist	Nodule	Distributions of overall diagnoses on T_c
t_1	Malignant	$d^m_{1,c}\ (c=1,\ldots,8)=(10,14,15,25,42,25,36,66)$
t_1	Benign	$d^b_{1,c}\ (c=1,\ldots,8)=(261,29,15,10,24,12,2,5)$
t_2	Malignant	$d^m_{2,c}\ (c=1,\ldots,8)=(10,1,19,1,17,49,26,84)$
t_2	Benign	$d^b_{2,c}\ (c=1,\ldots,8)=(254,12,33,11,20,33,8,8)$
t_3	Malignant	$d^m_{3,c}\ (c=1,\ldots,8)=(5,1,18,0,32,76,17,18)$
t_3	Benign	$d^b_{3,c}\ (c=1,\ldots,8)=(155,18,23,7,13,10,2,2)$
$\widetilde{t}_1$	Malignant	$\widetilde{d}^m_{1,c}\ (c=1,\ldots,8)=(17,1,21,2,36,71,52,42)$
$\widetilde{t}_1$	Benign	$\widetilde{d}^b_{1,c}\ (c=1,\ldots,8)=(206,31,47,26,42,20,9,5)$
$\widetilde{t}_2$	Malignant	$\widetilde{d}^m_{2,c}\ (c=1,\ldots,8)=(9,3,5,0,8,14,8,9)$
$\widetilde{t}_2$	Benign	$\widetilde{d}^b_{2,c}\ (c=1,\ldots,8)=(66,11,11,11,13,9,2,0)$
$\widetilde{t}_3$	Malignant	$\widetilde{d}^m_{3,c}\ (c=1,\ldots,8)=(3,1,3,2,5,9,2,12)$
$\widetilde{t}_3$	Benign	$\widetilde{d}^b_{3,c}\ (c=1,\ldots,8)=(84,2,7,9,20,16,2,3)$
$\widetilde{t}_4$	Malignant	$\widetilde{d}^m_{4,c}\ (c=1,\ldots,8)=(3,5,14,1,14,46,19,14)$
$\widetilde{t}_4$	Benign	$\widetilde{d}^b_{4,c}\ (c=1,\ldots,8)=(60,7,12,1,1,4,1,0)$
$\widetilde{t}_5$	Malignant	$\widetilde{d}^m_{5,c}\ (c=1,\ldots,8)=(3,2,2,2,11,27,10,2)$
$\widetilde{t}_5$	Benign	$\widetilde{d}^b_{5,c}\ (c=1,\ldots,8)=(17,7,4,1,3,2,0,0)$

tions of the third author, following the maximum possibility principle, a way is designed to revise the recommended TIRADS categories derived from the group's criterion weights. For a TIRADS category indicating a malignant nodule, the objective category is selected from the categories whose cancer risk is higher than that of the category under consideration. While for a TIRADS category indicating a benign nodule, the objective category is selected from the categories whose cancer risk is lower than that of the category under consideration. For example, when the recommended TIRADS category is TIRADS 4B-2, the objective category may be one of TIRADS 4B-3, TIRADS 4C, and TIRADS 5. Conversely, when the recommended TIRADS category is TIRADS 4A-2, the objective category may be one of TIRADS 4A-1 and TIRADS 3. To reflect the preferences of a radiologist in the determination of the objective category, the comparison between possible objective categories is dependent on their possibilities of occurring in the overall diagnoses provided by the radiologist. The possibilities are closely associated with the distributions of overall diagnoses on the TIRADS categories T_c $(c=1,\ldots,8)$ for the radiologist. Precisely, the distributions of overall diagnoses represent the numbers of each of the TIRADS categories in the overall diagnoses when nodules are diagnosed by a radiologist as malignant (or benign) ones. By examining the records of radiologists t_j $(j=1,2,3)$ and $\widetilde{t}_k$ $(k=1,\ldots,5)$, the distributions of their overall diagnoses on the TIRADS categories T_c $(c=1,\ldots,8)$ are identified and denoted by $d^m_{j,c}$ $(j=1,2,3,\ c=1,\ldots,8)$ and $d^b_{j,c}$ as well as $\widetilde{d}^m_{k,c}$ $(k=1,\ldots,5,\ c=1,\ldots,8)$ and $\widetilde{d}^b_{k,c}$, as presented in Table 4. According to Table 4 when the recommended TIRADS category for radiologist t_1 is T_5, the objective category is T_8 because the possibility of T_8 occurring in the overall diagnoses is larger than those of T_6 and T_7 for malignant nodules, as shown by $d^m_{1,c}$ $(c=6,7,8)$. This reflects the fact that when a nodule is diagnosed to be malignant and its cancer risk is considered to be at least equal to cancer risk corresponding to T_5, radiologist t_1 prefers T_8 to T_6 and T_7.

Following the maximum possibility principle, both the group's criterion weights and the group's distributions of overall diagnoses on the TIRADS categories T_c $(c=1,\ldots,8)$ are used to generate the recommended TIRADS categories for the nodules in the records provided by radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$. From $d^m_{j,c}$ $(j=1,2,3,\ c=1,\ldots,8)$ and $d^b_{j,c}$, the group's distributions of the overall diagnoses of malignant and benign nodules can be calculated by

$$d^m_c=\left\lfloor\sum\nolimits_{j=1}^{3}\lambda^j\cdot d^m_{j,c}\right\rfloor,\quad c=1,\ldots,8,\ \text{and} \tag{39}$$

$$d^b_c=\left\lfloor\sum\nolimits_{j=1}^{3}\lambda^j\cdot d^b_{j,c}\right\rfloor,\quad c=1,\ldots,8, \tag{40}$$

Table 5 Diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ in different situations

Condition	Diagnostic capabilities of five radiologists
Group's weights and group's distributions	$\widetilde{C}_k^{G_w,G_d}$ $(k = 1, \ldots, 5) = (0.8088, 0.7712, 0.8104, 0.7793, 0.7721)$
Group's weights and radiologists' distributions	$\widetilde{C}_k^{G_w,R_d}$ $(k = 1, \ldots, 5) = (0.7566, 0.7469, 0.8104, 0.6821, 0.6577)$
Radiologists' weights and group's distributions	$\widetilde{C}_k^{R_w,G_d}$ $(k = 1, \ldots, 5) = (0.8012, 0.7712, 0.805, 0.7698, 0.7721)$
Radiologists' weights and radiologists' distributions	$\widetilde{C}_k^{R_w,R_d}$ $(k = 1, \ldots, 5) = (0.7514, 0.7469, 0.805, 0.6745, 0.6577)$
Radiologists' overall diagnoses	$\widetilde{C}_k^{R_0}$ $(k = 1, \ldots, 5) = (0.7505, 0.7317, 0.7451, 0.7445, 0.7025)$

where $\lfloor x \rfloor$ represents the number to which x rounds down. Especially, using (39) and (40), it is obtained that d_c^m $(c = 1, \ldots, 8) = (8, 5, 17, 8, 30, 49, 26, 56)$ and d_c^b $(c = 1, \ldots, 8) = (223, 19, 23, 9, 19, 18, 4, 5)$. Suppose that the recommended TIRADS category of a nodule for radiologist $\widetilde{t}_k$ based on the group's criterion weights w_i $(i = 1, \ldots, 5)$ is $T_{\widetilde{c}}^k$. Then following the maximum possibility principle, $T_{\widetilde{c}}^k$ is revised to be $T_{\hat{c}}^k$ such that

$$\hat{c} = \arg\max\{d_c^m,\ c \in \{\widetilde{c}+1, \ldots, 8\}\}\ (4 \leqslant \widetilde{c} \leqslant 7),\ \text{or} \tag{41}$$

$$\hat{c} = \arg\max\{d_c^b,\ c \in \{1, \ldots, \widetilde{c}-1\}\}\ (2 \leqslant \widetilde{c} < 4),\ \text{or} \tag{42}$$

$$\hat{c} = \widetilde{c}\ (\widetilde{c} = 1, 8). \tag{43}$$

When the recommended TIRADS categories of all nodules in the records of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ are revised using (41)–(43), the diagnostic capabilities of the five radiologists based on the group's criterion weights w_i $(i = 1, \ldots, 5)$ and the distributions of the overall diagnoses $(d_c^m,\ d_c^b)$ can be obtained as $\widetilde{C}_k^{G_w,G_d}$ $(k = 1, \ldots, 5)$ using (22), as shown in the second row of Table 5.

From the comparison between $\widetilde{C}_k^{G_w,G_d}$ $(k = 1, \ldots, 5)$ and $\widetilde{C}_k^{R_0}$, it can be found that the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ based on the group's criterion weights w_i $(i = 1, \ldots, 5)$ and the group's distributions of the overall diagnoses (d_c^m, d_c^b) are beyond their own diagnostic capabilities. The diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ are increased by $\frac{\widetilde{C}_k^{G_w,G_d} - \widetilde{C}_k^{R_0}}{\widetilde{C}_k^{R_0}}$ $(k = 1, \ldots, 5) = (7.77\%, 5.4\%, 8.76\%, 4.67\%, 9.91\%)$. To examine the contributions of w_i $(i = 1, \ldots, 5)$ and (d_c^m, d_c^b) to the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$, three situations are considered. The three situations include: (1) the group's criterion weights and the five radiologists' distributions of overall diagnoses; (2) the five radiologists' criterion weights and the group's distributions of overall diagnoses; and (3) the five radiologists' criterion weights and the radiologists' distributions of overall diagnoses. The diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ in the three situations are obtained as $\widetilde{C}_k^{G_w,R_d}$ $(k = 1, \ldots, 5)$, $\widetilde{C}_k^{R_w,G_d}$, and $\widetilde{C}_k^{R_w,R_d}$, which are also shown in Table 5. It is easy to see that $\widetilde{C}_k^{G_w,G_d} \geqslant \{\widetilde{C}_k^{G_w,R_d}, \widetilde{C}_k^{R_w,G_d}, \widetilde{C}_k^{R_w,R_d}\}$ $(k = 1, \ldots, 5)$. This highlights the contributions of w_i $(i = 1, \ldots, 5)$ and (d_c^m, d_c^b) to the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$. It can also be observed from Table 5 that $\widetilde{C}_k^{R_w,G_d} > \widetilde{C}_k^{R_0}$ $(k = 1, \ldots, 5)$, $\widetilde{C}_k^{G_w,R_d} < \widetilde{C}_k^{R_0}$ $(k = 4, 5)$, and $\widetilde{C}_k^{R_w,R_d} < \widetilde{C}_k^{R_0}$ $(k = 4, 5)$. These indicate that the contribution of (d_c^m, d_c^b) to the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ is larger than that of w_i $(i = 1, \ldots, 5)$. It is worth noting that this finding is closely related to the inherent characteristic of diagnosing thyroid nodules, as indicated in the expertise and experience of the third author.

5.3 Random simulation

As presented in the above process of generating diagnostic recommendations, the criterion weights and the distribution of the overall diagnoses of the group, which is composed of t_j $(j = 1, 2, 3)$, work together to help improve the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$. To further explore the influence of the group's criterion weights and the group's distribution of the overall diagnoses on the recommended diagnostic capabilities of the five radiologists, random simulation experiments are conducted in the following.

From (33), (39) and (40), it can be observed that the weights of radiologists t_j $(j = 1, 2, 3)$ in the group, i.e., λ_j $(j = 1, 2, 3)$, are involved in the determination of the group's criterion weights w_i $(i =$

$1,\ldots,5$) and the determination of the group's distribution of the overall diagnoses (d_c^m, d_c^b). To explore the relationship between the recommended diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$, i.e., $\widetilde{C}_k^{G_w,G_d}$ $(k=1,\ldots,5)$, and λ_j $(j=1,2,3)$, 30 groups of λ_j such that $\sum_{j=1}^{3}\lambda_j=1$ are randomly generated. For each group of λ_j, $\widetilde{C}_k^{G_w,G_d}$ $(k=1,\ldots,5)$ is calculated using the process presented in Subsection 5.2. The movement of $\widetilde{C}_k^{G_w,G_d}$ $(k=1,\ldots,5)$ with random λ_j $(j=1,2,3)$ is plotted in Figure 2. It can be observed from Figures 2(a)–(c) that $\widetilde{C}_{k_1}^{G_w,G_d}>\widetilde{C}_{k_1}^{R_0}$ $(k_1=1,2,3)$ always holds and from Figures 2(d) and (e) that $\widetilde{C}_{k_2}^{G_w,G_d}$ $(k_2=4,5)$ fluctuates around $\widetilde{C}_{k_2}^{R_0}$. These observations indicate that w_i $(i=1,\ldots,5)$ and (d_c^m,d_c^b) derived from different sets of λ_j $(j=1,2,3)$ can always help improve the diagnostic capabilities of radiologists $\widetilde{t}_{k_1}$ $(k_1=1,2,3)$; while w_i $(i=1,\ldots,5)$ and (d_c^m,d_c^b) derived from some specific sets of λ_j $(j=1,2,3)$ can help improve the diagnostic capabilities of radiologists $\widetilde{t}_{k_2}$ $(k_2=4,5)$. It can also be observed from Figure 2 that larger $\widetilde{C}_k^{G_w,G_d}$, which is at least beyond $\widetilde{C}_k^{R_0}$, is generated when λ_1 or λ_2 is the maximum one among λ_j $(j=1,2,3)$. This finding is consistent with the fact that the own diagnostic capabilities of t_1 and t_2 are larger than that of t_3. Consequently, t_1 and t_2 can contribute more to improving the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$ than t_3. Overall, the importance of λ_j $(j=1,2,3)$ with respect to the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$ is reflected by the experimental results shown in Figure 2.

The group's criterion weights w_i $(i=1,\ldots,5)$ and the group's distribution of the overall diagnoses (d_c^m,d_c^b) are used to determine the aggregated diagnoses and further the recommended diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$, i.e., $\widetilde{C}_k^{G_w,G_d}$. To explore the different influence of w_i $(i=1,\ldots,5)$ and (d_c^m,d_c^b) on $\widetilde{C}_k^{G_w,G_d}$ $(k=1,\ldots,5)$, 30 groups of w_i $(i=1,\ldots,5)$ such that $\sum_{i=1}^{5}w_i=1$ and 30 groups of (d_c^m,d_c^b) are randomly generated, respectively. Given the group's random criterion weights and the group's distribution of the overall diagnoses derived from the distributions of t_j $(j=1,2,3)$, suppose that the recommended diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$ are obtained as $\widetilde{C}_k^{\bar{G}_w,G_d}$; while given the group's criterion weights derived from the criterion weights of t_j $(j=1,2,3)$ and the group's random distribution of the overall diagnoses, the recommended diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$ are obtained as $\widetilde{C}_k^{G_w,\bar{G}_d}$. All experimental results are presented and compared in Figure 3. It can be observed from Figure 3 that $\widetilde{C}_k^{\bar{G}_w,G_d}$ $(k=1,\ldots,5)$ fluctuates around $\widetilde{C}_k^{R_0}$ with the group's different random criterion weights, $\widetilde{C}_k^{G_w,\bar{G}_d}$ $(k=1,2,4,5)$ fluctuates around $\widetilde{C}_k^{R_0}$ with the group's different random distributions of the overall diagnoses, and $\widetilde{C}_k^{G_w,\bar{G}_d}$ is always larger than $\widetilde{C}_3^{R_0}$. Meanwhile, from Figures 3(b)–(e), it can also be observed that the group's specific distributions of the overall diagnoses may result in greatly small $\widetilde{C}_k^{\bar{G}_w,G_d}$. These observations indicate that random w_i $(i=1,\ldots,5)$ can cause the larger amplitude of variation in $\widetilde{C}_k^{G_w,G_d}$ than random (d_c^m,d_c^b) and hence, more attention should be paid to w_i $(i=1,\ldots,5)$ in a random environment. It may be a different case when there are some constraints on w_i $(i=1,\ldots,5)$ and (d_c^m,d_c^b) as discussed in Subsection 5.2. Overall, the importance of w_i $(i=1,\ldots,5)$ and (d_c^m,d_c^b) with respect to the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$ is reflected by the experimental results shown in Figure 3.

From the above simulation experiments and analysis, it can be revealed that the group's criterion weights and the group's distribution of the overall diagnoses influence the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$ significantly. On the condition that the constraints on the group's criterion weights and the group's distribution of the overall diagnoses are satisfied, determining appropriately the group's criterion weights and the group's distribution is greatly important for improving the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$.

5.4 Performance comparison between group-recommended diagnoses and overall diagnoses based on binary classification

Subsection 5.2 focuses on using historical data collected from radiologists t_j $(j=1,2,3)$ with sufficient experience to improve the diagnostic capabilities of $\widetilde{t}_k$ $(k=1,\ldots,5)$ without a wealth of experience. When the judgment on whether thyroid nodules are malignant (positive) or benign (negative) becomes the focus, the diagnosis of nodules is regarded as a binary classification problem. To address this problem,

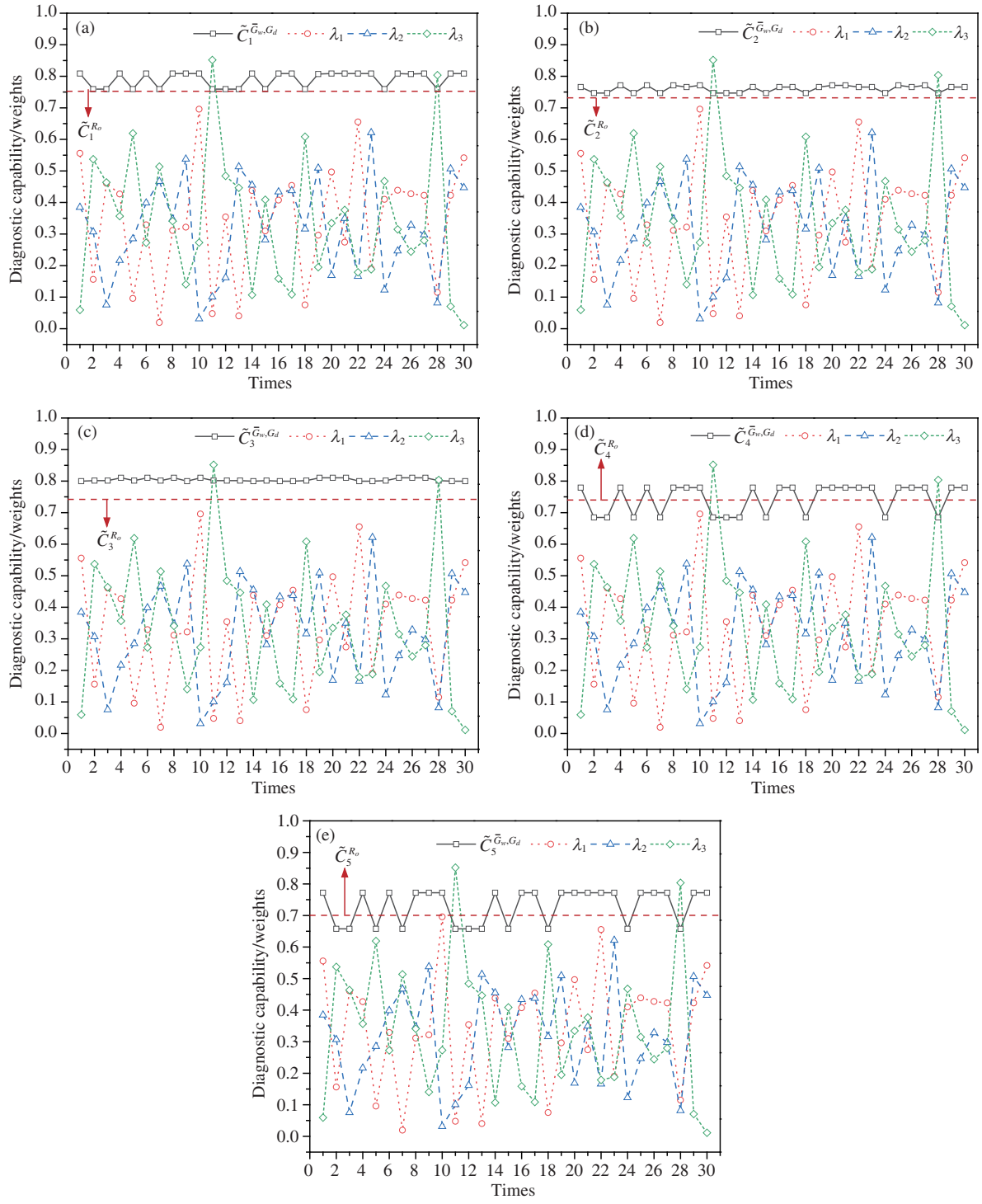

Figure 2 (Color online) Movement of $\widetilde{C}_k^{G_w,G_d}$ $(k=1,\ldots,5)$ with random λ_j $(j=1,2,3)$. (a) $\widetilde{t}_1$; (b) $\widetilde{t}_2$; (c) $\widetilde{t}_3$; (d) $\widetilde{t}_4$; (e) $\widetilde{t}_5$.

the performance of nodule classification derived from group-recommended diagnoses is compared with that derived from overall diagnoses.

As indicated in [68], the area under the receiver operating characteristic (ROC) curve (AUC) is a commonly accepted measure used in comparing the performance of classification rules in clinical applications. It is adopted to compare the performances of nodule classification based on the group-recommended and overall diagnoses. The concepts associated with ROC curve include true positive, false negative, false

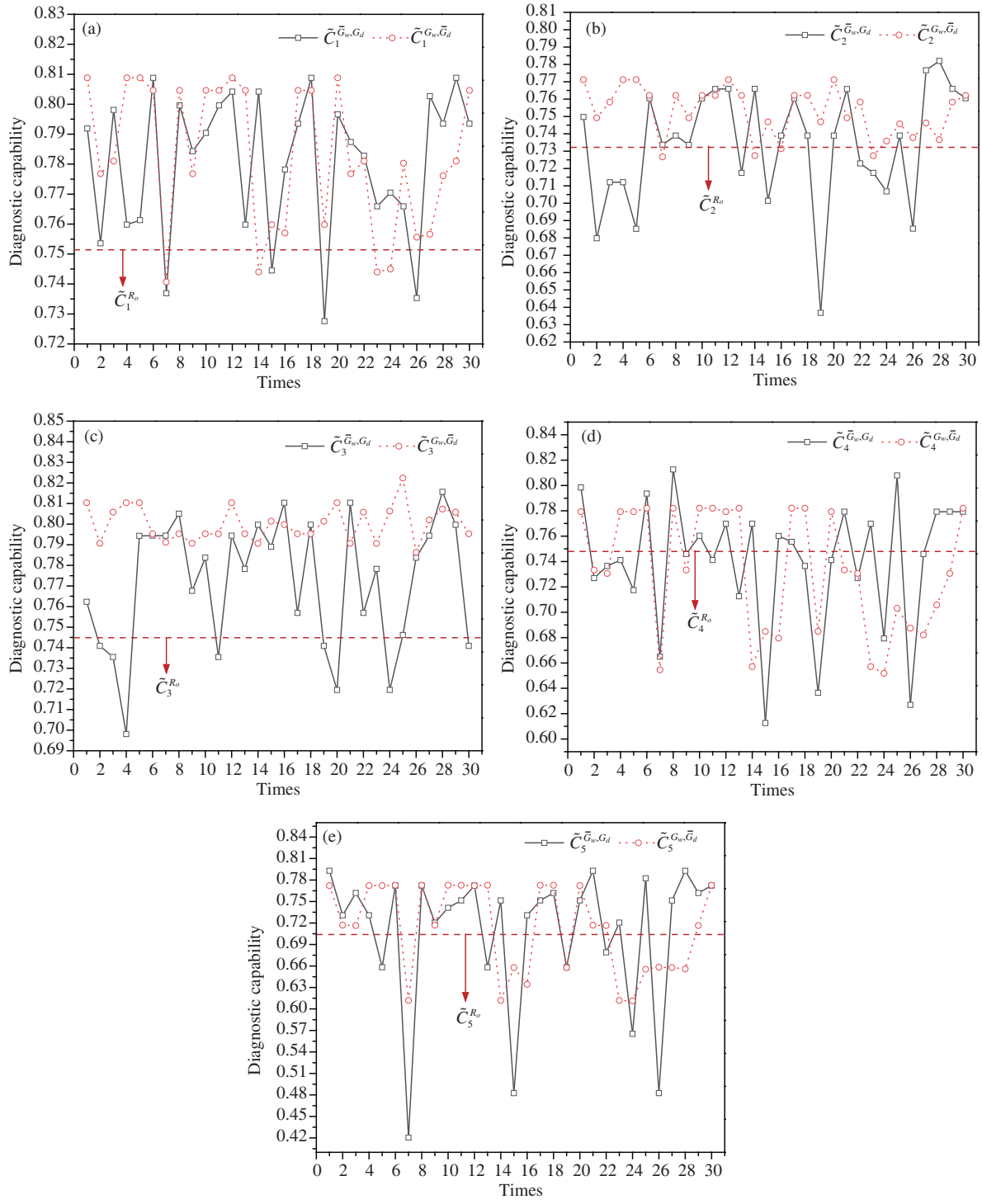

Figure 3 (Color online) Comparison between $\widetilde{C}_k^{\bar{G}_w,G_d}$ and $\widetilde{C}_k^{G_w,\bar{G}_d}$ $(k=1,\ldots,5)$. (a) $\widetilde{t}_1$; (b) $\widetilde{t}_2$; (c) $\widetilde{t}_3$; (d) $\widetilde{t}_4$; (e) $\widetilde{t}_5$.

positive, true negative, false positive rate (FPR), and true positive rate (TPR), in which the first four concepts for the diagnosis of thyroid nodules are explained in [68]. Based on these concepts, ROC curves from the group-recommended and overall diagnoses will be constructed to make a performance comparison between them.

As presented in Subsection 5.2, in the process of improving the diagnostic capabilities of radiologists $\widetilde{t}_k$ $(k=1,\ldots,5)$, the group-recommended TIRADS categories of thyroid nodules for the radiologists are obtained. As suggested by the third author, the group-recommended TIRADS category that is one of

{TIRADS 4B-1, TIRADS 4B-2, TIRADS 4B-3, TIRADS 4C, TIRADS 5} indicates a malignant nodule; while the group-recommended TIRADS category that is one of {TIRADS 3, TIRADS 4A-1, TIRADS 4A-2} indicates a benign nodule. The division of TIRADS 4B and TIRADS 4A is to facilitate radiologists' precise diagnosis of thyroid nodules and differentiate the diagnostic capabilities of different radiologists. In clinical practice, {TIRADS 3, TIRADS 4A, TIRADS 4B, TIRADS 4C, TIRADS 5} is sufficient for the binary classification of thyroid nodules. To address nodule classification, each group-recommended TIRADS category of a thyroid nodule is changed to one of {TIRADS 3, TIRADS 4A, TIRADS 4B, TIRADS 4C, TIRADS 5}. Similarly, each overall diagnosis of thyroid nodule provided by radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$ is changed to one of {TIRADS 3, TIRADS 4A, TIRADS 4B, TIRADS 4C, TIRADS 5}. After transformation, the categories {TIRADS 4B, TIRADS 4C, TIRADS 5} mean malignant nodules and the categories {TIRADS 3, TIRADS 4A} mean benign nodules. Because each category corresponds to a cancer risk interval, two variables c_r^- and c_r^+ such that $0.25 \leqslant c_r^- \leqslant 1$ and $0.75 \leqslant c_r^+ \leqslant 1$ are used as two thresholds to implement the binary classification of thyroid nodules. If the lower bound of the cancer risk interval of the group-recommended TIRADS category (or the overall diagnosis) is larger than or equal to c_r^- and the upper bound is larger than or equal to c_r^+, then the nodule is judged to be malignant; otherwise, it is judged to be benign. Assume that the two variables c_r^- and c_r^+ are changed from 1 to 0.25 with a step of $(1-0.25)/100$ and from 1 to 0.75 with a step of $(1-0.75)/100$, respectively. With this assumption, using the group-recommended TIRADS categories (or the overall diagnoses) of thyroid nodules for radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$, 100 groups of FPRs and TPRs are computed to form the radiologists' ROC curves. The curves of each radiologist constructed from the group-recommended TIRADS categories and from the overall diagnoses are plotted in one figure to facilitate the performance comparison. Figure 4 shows the ROC curves of radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$.

To compare ROC curve constructed from the group-recommended TIRADS categories with ROC curve from the overall diagnoses for radiologists $\widetilde{t}_k$ $(k = 1, \ldots, 5)$, the curves' associated AUCs are calculated as (S_G^k, S_O^k) $(k = 1, \ldots, 5) = \{(0.0693, 0.0008), (0.0881, 0.0004), (0.0431, 0.0008), (0.0375, 0.0002), (0.0429, 0)\}$. It is easy to see that $(S_G^k > S_O^k)$ $(k = 1, \ldots, 5)$. Consequently, compared with the nodule classification derived from the overall diagnoses, the nodule classification derived from the group-recommended TIRADS categories possesses higher performance. This finding indicates that the application of the proposed method to the diagnosis of thyroid nodules can improve radiologists' diagnostic accuracy.

6 Conclusion

A large amount of diversified information and knowledge in the era of informatization drives us to depend on group expertise and experience to handle real problems. Many GDM-related studies have been conducted to generate group-satisfactory solutions. Among the factors that influence the generation of group-satisfactory solutions in these studies, the weight of each criterion and the weight of each expert are two principal ones. Traditional methods mainly use the subjective preferences of experts or objective decision matrices to determine criterion weights and experts' weights, and pay little attention to the learning of the two types of weights from historical decision data. Such learning is beneficial for the generation of the two types of weights and further the generation of GDM solutions that are consistent with the preferences of experts characterized by their historical decision data.

To help generate group-satisfactory solutions based on historical decision data, this paper proposes a data-driven MCGDM method with positive interval numbers. In this paper, to facilitate the learning of criterion weights and experts' weights from historical decision data, existing distance measures between positive interval numbers are analyzed to highlight the selected distance measure. With the assumption that experts' weights are positively correlated to their capabilities to make correct decisions, the weight of each expert is learned from historical overall assessments and the corresponding gold standards based on the selected distance measure between positive interval numbers. To learn criterion weights, it is assumed that the larger the weight of a criterion, the larger the similarity between the assessment on the criterion and the overall assessment. A set of historical assessments on criteria and historical overall assessment can generate a set of criterion weights under this assumption. Based on all sets of learned criterion weights,

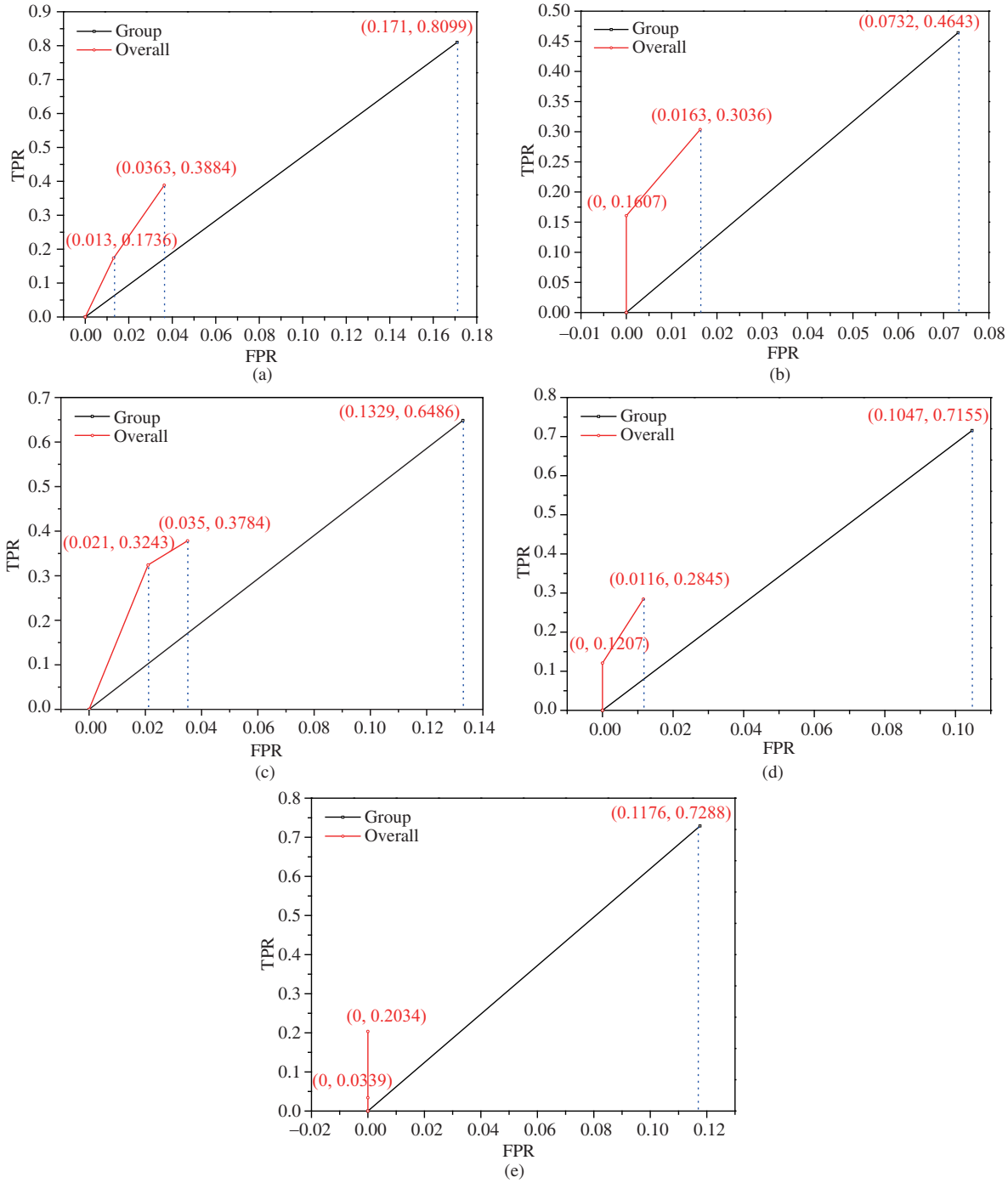

Figure 4 Comparison between ROC curves from the group-recommended TIRADS categories and that from the overall diagnose for radiologists. (a) $\widetilde{t}_1$; (b) $\widetilde{t}_2$; (c) $\widetilde{t}_3$; (d) $\widetilde{t}_4$; (e) $\widetilde{t}_5$.

an optimization model is constructed to generate a representative set of criterion weights and the unique optimal solution to the model is theoretically found. By involving the learning of criterion weights and experts' weights, the MCGDM process of the proposed method is presented and analyzed. The proposed method is used to help improve the capabilities of radiologists in diagnosing thyroid nodules based on the historical examination reports and the corresponding pathologic findings collected from the tertiary hospital located in Hefei, Anhui Province, China. In addition to the criterion weights of the radiologists with high diagnostic capabilities, their distributions of the overall diagnoses are verified to help improve the radiologists with low diagnostic capabilities.

The theoretical contributions of this paper include: (1) a data-driven MCGDM method with interval numbers is proposed; (2) expert weights are learned from historical overall assessments and the corresponding gold standards; (3) criterion weights are learned from historical assessments on criteria and historical overall assessments, in which the obtainment of the weights and their uniqueness are theoretically proved; and (4) the two processes of the proposed method for its two purposes are developed, in which one is to compare alternatives and the other is to make an overall conclusion. The application of the proposed method to helping improve the diagnostic capabilities of radiologists highlights the practical contributions of this paper.

When the individual diagnoses of other diseases on each criterion and the overall diagnoses of the diseases are represented by interval numbers, there exists a linearly weighted relationship between the individual diagnoses and the overall diagnoses, and the historical data are available, the proposed method can be applied in the auxiliary diagnosis of the diseases without any change. Only the criteria for diagnosing the diseases need to be identified based on the radiologist's expertise and experience. If the first two conditions are not satisfied in the diagnosis of other diseases, the proposed method needs to be simply extended, in which the similarity between the expressions of the diagnoses is required to be reconstructed and the combination of the individual diagnoses is required to be reconsidered. In the future study, the proposed method will be extended to be applied in the auxiliary diagnosis of other diseases, in which the scalability of the proposed method will be evaluated.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 71622003, 71571060, 71690235, 71690230, 71521001).

Supporting information Proofs. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Galo N R, Calache L D D R, Carpinetti L C R. A group decision approach for supplier categorization based on hesitant fuzzy and ELECTRE TRI. Int J Prod Econ, 2018, 202: 182–196
2 Qin G D, Liu X W, Pedrycz W. An extended TODIM multi-criteria group decision making method for green supplier selection in interval type-2 fuzzy environment. Eur J Oper Res, 2017, 258: 626–638
3 Cheng S H. Autocratic multiattribute group decision making for hotel location selection based on interval-valued intuitionistic fuzzy sets. Inf Sci, 2018, 427: 77–87
4 Chen L H, Ko W C, Tseng C Y. Fuzzy approaches for constructing house of quality in QFD and its applications: a group decision-making method. IEEE Trans Eng Manage, 2013, 60: 77–87
5 Ren J Z, Liang H W. Measuring the sustainability of marine fuels: a fuzzy group multi-criteria decision making approach. Transpation Res Part D-Transp Environ, 2017, 54: 12–29
6 Wu B, Yan X P, Wang Y, et al. Selection of maritime safety control options for NUC ships using a hybrid group decision-making approach. Saf Sci, 2016, 88: 108–122
7 Lu J, Ma J, Zhang G Q, et al. Theme-based comprehensive evaluation in new product development using fuzzy hierarchical criteria group decision-making method. IEEE Trans Ind Electron, 2011, 58: 2236–2246
8 Li G X, Kou G, Peng Y. A group decision making model for integrating heterogeneous information. IEEE Trans Syst Man Cybern Syst, 2018, 48: 982–992
9 Liu P D, Chen S M. Group decision making based on heronian aggregation operators of intuitionistic fuzzy numbers. IEEE Trans Cybern, 2017, 47: 2514–2530
10 Wu Q, Wu P, Zhou L G, et al. Some new Hamacher aggregation operators under single-valued neutrosophic 2-tuple linguistic environment and their applications to multi-attribute group decision making. Comput Ind Eng, 2018, 116: 144–162
11 Jana C, Senapati T, Pal M, et al. Picture fuzzy Dombi aggregation operators: application to MADM process. Appl Soft Comput, 2019, 74: 99–109
12 Fu C, Yang S L. An evidential reasoning based consensus model for multiple attribute group decision analysis problems with interval-valued group consensus requirements. Eur J Oper Res, 2012, 223: 167–176
13 Yeh C T. Existence of interval, triangular, and trapezoidal approximations of fuzzy numbers under a general condition. Fuzzy Sets Syst, 2017, 310: 1–13
14 Lima A S, de Souza J N, Moura J A B, et al. A consensus-based multicriteria group decision model for information technology management committees. IEEE Trans Eng Manage, 2018, 65: 276–292
15 Yan H B, Ma T J, Huynh V N. On qualitative multi-attribute group decision making and its consensus measure: a probability based perspective. Omega, 2017, 70: 94–117

16 Yang Y, Wang X X, Xu Z S. The multiplicative consistency threshold of intuitionistic fuzzy preference relation. Inf Sci, 2019, 477: 349–368
17 Li C C, Rodríguez R M, Martinez L, et al. Consensus building with individual consistency control in group decision making. IEEE Trans Fuzzy Syst, 2019, 27: 319–332
18 Meng F Y, An Q X, Tan C Q, et al. An approach for group decision making with interval fuzzy preference relations based on additive consistency and consensus analysis. IEEE Trans Syst Man Cybern Syst, 2017, 47: 2069–2082
19 Wan S P, Wang F, Dong J Y. A three-phase method for group decision making with interval-valued intuitionistic fuzzy preference relations. IEEE Trans Fuzzy Syst, 2018, 26: 998–1010
20 Kou G, Ergu D J, Lin C S, et al. Pairwise comparison matrix in multiple criteria decision making. Tech Economic Dev Economy, 2016, 22: 738–765
21 Kou G, Ergu D J, Shang J. Enhancing data consistency in decision matrix: adapting Hadamard model to mitigate judgment contradiction. Eur J Oper Res, 2014, 236: 261–271
22 Liu B S, Shen Y H, Zhang W, et al. An interval-valued intuitionistic fuzzy principal component analysis model-based method for complex multi-attribute large-group decision-making. Eur J Oper Res, 2015, 245: 209–225
23 Wu T, Liu X W, Liu F. An interval type-2 fuzzy TOPSIS model for large scale group decision making problems with social network information. Inf Sci, 2018, 432: 392–410
24 Wu T, Liu X W. An interval type-2 fuzzy clustering solution for large-scale multiple-criteria group decision-making problems. Knowledge-Based Syst, 2016, 114: 118–127
25 Chen X, Zhang H J, Dong Y C. The fusion process with heterogeneous preference structures in group decision making: a survey. Inf Fusion, 2015, 24: 72–83
26 Tang J, Chen S M, Meng F Y. Heterogeneous group decision making in the setting of incomplete preference relations. Inf Sci, 2019, 483: 396–418
27 Wan S P, Xu J, Dong J Y. Aggregating decision information into interval-valued intuitionistic fuzzy numbers for heterogeneous multi-attribute group decision making. Knowledge-Based Syst, 2016, 113: 155–170
28 Haag F, Lienert J, Schuwirth N, et al. Identifying non-additive multi-attribute value functions based on uncertain indifference statements. Omega, 2019, 85: 49–67
29 Qin G D, Liu X W. Multi-attribute group decision making using combined ranking value under interval type-2 fuzzy environment. Inf Sci, 2015, 297: 293–315
30 Yue C. A geometric approach for ranking interval-valued intuitionistic fuzzy numbers with an application to group decision-making. Comput Industrial Eng, 2016, 102: 233–245
31 Entani T, Inuiguchi M. Pairwise comparison based interval analysis for group decision aiding with multiple criteria. Fuzzy Sets Syst, 2015, 274: 79–96
32 Fu C, Xu D L. Determining attribute weights to improve solution reliability and its application to selecting leading industries. Ann Oper Res, 2016, 245: 401–426
33 Kim J H, Ahn B S. Extended VIKOR method using incomplete criteria weights. Expert Syst Appl, 2019, 126: 124–132
34 Zhang Z, Guo C H, Martinez L. Managing multigranular linguistic distribution assessments in large-scale multiattribute group decision making. IEEE Trans Syst Man Cybern Syst, 2017, 47: 3063–3076
35 Dong Y C, Xiao J, Zhang H J, et al. Managing consensus and weights in iterative multiple-attribute group decision making. Appl Soft Comput, 2016, 48: 80–90
36 Liu B S, Shen Y H, Chen Y, et al. A two-layer weight determination method for complex multi-attribute large-group decision-making experts in a linguistic environment. Inf Fusion, 2015, 23: 156–165
37 Shi Z J, Wang X Q, Palomares I, et al. A novel consensus model for multi-attribute large-scale group decision making based on comprehensive behavior classification and adaptive weight updating. Knowledge-Based Syst, 2018, 158: 196–208
38 Pérez I J, Cabrerizo F J, Alonso S, et al. On dynamic consensus processes in group decision making problems. Inf Sci, 2018, 459: 20–35
39 Hajek P, Froelich W. Integrating TOPSIS with interval-valued intuitionistic fuzzy cognitive maps for effective group decision making. Inf Sci, 2019, 485: 394–412
40 Liu W, Li L. An approach to determining the integrated weights of decision makers based on interval number group decision matrices. Knowledge-Based Syst, 2015, 90: 92–98
41 Tambouratzis T, Canellidis V. Reward-penalty assignments and genetic algorithms for ordinal interval number group decision making. Int J Intell Syst, 2014, 29: 727–750
42 Yue Z L. Group decision making with multi-attribute interval data. Inf Fusion, 2013, 14: 551–561
43 Roberts R, Goodwin P. Weight approximations in multi-attribute decision models. J Multi-Crit Decis Anal, 2002, 11: 291–303
44 Yang G L, Yang J B, Xu D L, et al. A three-stage hybrid approach for weight assignment in MADM. Omega, 2017, 71: 93–105
45 Wang Z J, Liu F, Lin J. Fuzzy eigenvector method for obtaining normalized fuzzy weights from fuzzy pairwise comparison matrices. Fuzzy Sets Syst, 2017, 315: 26–43
46 Shirland L E, Jesse R R, Thompson R L, et al. Determining attribute weights using mathematical programming. Omega, 2003, 31: 423–437
47 Fu C, Xu D L, Xue M. Determining attribute weights for multiple attribute decision analysis with discriminating power in belief distributions. Knowledge-Based Syst, 2018, 143: 127–141
48 Barron F H, Barrett B E. Decision quality using ranked attribute weights. Manage Sci, 1996, 42: 1515–1523

49 Koksalmis E, Kabak Ö. Deriving decision makers' weights in group decision making: an overview of objective methods. Inf Fusion, 2019, 49: 146–160
50 Wang Y M, Luo Y. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making. Math Comput Model, 2010, 51: 1–12
51 He Y H, Guo H W, Jin M Z, et al. A linguistic entropy weight method and its application in linguistic multi-attribute group decision making. Nonlin Dyn, 2016, 84: 399–404
52 Şahin R, Liu P. Maximizing deviation method for neutrosophic multiple attribute decision making with incomplete weight information. Neural Comput Applic, 2016, 27: 2017–2029
53 Yue Z L. Approach to group decision making based on determining the weights of experts by using projection method. Appl Math Model, 2012, 36: 2900–2910
54 Yue C. Entropy-based weights on decision makers in group decision-making setting with hybrid preference representations. Appl Soft Comput, 2017, 60: 737–749
55 Qi X W, Liang C Y, Zhang J L. Generalized cross-entropy based group decision making with unknown expert and attribute weights under interval-valued intuitionistic fuzzy environment. Comput Ind Eng, 2015, 79: 52–64
56 Zhang X, Liu P D. Method for multiple attribute decision-making under risk with interval numbers. Int J Fuzzy Syst, 2010, 12: 237–242
57 Tran L, Duckstein L. Comparison of fuzzy numbers using a fuzzy distance measure. Fuzzy Sets Syst, 2002, 130: 331–341
58 de Carvalho F A T, Simões E C. Fuzzy clustering of interval-valued data with City-Block and Hausdorff distances. Neurocomputing, 2017, 266: 659–673
59 Zhang J, Pang J Z, Yu J F, et al. An efficient assembly retrieval method based on Hausdorff distance. Robot Comput-Integrated Manuf, 2018, 51: 103–111
60 Ramos-Guajardo A B, Grzegorzewski P. Distance-based linear discriminant analysis for interval-valued data. Inf Sci, 2016, 372: 591–607
61 Irpino A, Verde R. Dynamic clustering of interval data using a Wasserstein-based distance. Pattern Recogn Lett, 2008, 29: 1648–1658
62 Li X, Zhang S L, Zhang M, et al. Rank of interval numbers based on a new distance measure. J Xihua Univ (Nat Sci), 2008, 27: 87–90
63 Winston W L. Operations Research: Applications and Algorithms. Boston: Duxbury Press, 2003
64 Sahli Z T, Karipineni F, Hang J F, et al. The association between the ultrasonography TIRADS classification system and surgical pathology among indeterminate thyroid nodules. Surgery, 2019, 165: 69–74
65 Horvath E, Silva C F, Majlis S, et al. Prospective validation of the ultrasound based TIRADS (thyroid imaging reporting and data system) classification: results in surgically resected thyroid nodules. Eur Radiol, 2017, 27: 2619–2628
66 Kwak J Y, Han K H, Yoon J H, et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. Radiology, 2011, 260: 892–899
67 Park J Y, Lee H J, Jang H W, et al. A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma. Thyroid, 2009, 19: 1257–1264
68 Fu C, Liu W Y, Chang W J. Data-driven multiple criteria decision making for diagnosis of thyroid cancer. Ann Oper Res, 2018. doi: 10.1007/s10479-018-3093-7
69 Cappelli C, Castellano M, Pirola I, et al. The predictive value of ultrasound findings in the management of thyroid nodules. QJM Int J Medicine, 2006, 100: 29–35
70 Chan B K, Desser T S, McDougall I R, et al. Common and uncommon sonographic features of papillary thyroid carcinoma. J Ultrasound Med, 2003, 22: 1083–1090
71 Frates M C, Benson C B, Charboneau J W, et al. Management of thyroid nodules detected at US: society of radiologists in ultrasound consensus conference statement. Radiology, 2005, 237: 794–800
72 Moon W J, Jung S L, Lee J H, et al. Benign and malignant thyroid nodules: US differentiation–multicenter retrospective study. Radiology, 2008, 247: 762–770