COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE XXX (2011) XXX–XXX

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Improving the medical scale predictability by the pairwise comparisons method: Evidence from a clinical data study

Tamar Kakiashvili[a], *Waldemar W. Koczkodaj*[b,*], *Marc Woodbury-Smith*[c]

[a] *Sudbury Therapy, Sudbury, Ontario, Canada*
[b] *Laurentian Univeristy, Computer Science, Sudbury, Ontario, Canada*
[c] *McMaster University, Faculty of Health Sciences, Hamilton, Ontario, Canada*

ARTICLE INFO

Article history:
Received 12 May 2011
Received in revised form
29 July 2011
Accepted 22 September 2011

Keywords:
Medical scale
Medical scale item
Pairwise comparisons
Inconsistency

ABSTRACT

In the clinical practice of psychiatry, presence or absence of particular symptoms is based on the subjective interpretation, by the clinician, of mental and behavioural descriptions offered by the patient. However, this subjectivity that characterizes the diagnostic decision making process may limit the reliability of diagnosis. In this current study, the pairwise comparisons (PC) method is used to investigate whether the psychometric properties of a medical screening questionnaire can be improved. The pilot data described herein did indeed demonstrate that modest improvements in diagnostic accuracy could be achieved using PC, and provides early evidence that the inconsistency produced by subjective clinical ratings can be reduced using this method, thus providing impetus for further investigation.

1. Introduction

In broad terms, advances in medical science have undoubtedly improved the reliability and validity of medical diagnosis, and screening for different disorders or diseases (such as breast and colon cancer) is more accurate than ever. The resolution of CT and MRI scanning has allowed body system pathology to be recognized. However, for disorders defined according to mental and behavioural symptoms, accuracy of diagnosis remains an issue, with each disorder defined according to a list of criteria that are based on the subjective interpretation, by the diagnosing clinician, of narratives offered by patients and their carers, with very little in the way of pathognomonic signs and symptoms. Compounding this is the lack of available medical investigations (such as blood tests, X-rays and so forth) that allow suspected clinical diagnoses and screening to be confirmed. Such diagnostic uncertainty is perhaps no better demonstrated than for the Autism Spectrum Disorders (ASDs), a group of genetic syndromes characterized by particular patterns of impairment in the social, communicative and behavioural domains. Strikingly, even among so-called experts, disagreement can occur in a significant number of cases. Recognizing the fact that investigations allowing improved diagnostic accuracy are a long way off, there is therefore an urgent need to develop methods to improve accuracy and reduce uncertainty in the diagnosis of ASDs and, indeed, all other mental disorders.

2. Background

Autism Spectrum Disorders are of particular significance because of their lack of pathognomonic features and vaguely defined diagnostic characteristics. ASDs, as with all other mental and behavioural disorders, are diagnosed according to core sets of criteria set out in the DSM-IV and ICD-10. For the ASDs there are three 'domains' (social interaction,

[*] *Corresponding author.* Tel.: +1 705 675 1151; fax: +1 705 673 6591.

0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.09.011

communication, and repetitive behaviours), and each domain is further expanded to include four exemplary behaviours. For example, one such behaviour in the social domain is 'difficulties forming and maintaining peer relationships', and within the communication impairment 'difficulties with to and fro conversation' without any further guidance of how to decide whether such impairments are present or not. This makes for significant diagnostic uncertainty, and is exemplified not only among the ASDs but for many of the other disorders described in similar terms in the DSM and ICD nosological systems (such as schizophrenia, anxiety disorders, mood disorders and personality disorders).

3. Design considerations

Attempts to reduce diagnostic uncertainty for mental disorders have been focused on the development of screening and diagnostic questionnaires. One such questionnaire is the Autism Behaviour Checklist of Krug, discussed subsequently in this paper, but our approach is applicable to practically all other questionnaires. However, the reliability and validity, and hence the Receiver Operating Characteristic (ROC), have been an area of concern for all of these questionnaires. In all likelihood this relates to the subjective nature of the decision making process regarding the raters responses, and therefore it is crucial that methods to address this ambiguity are made available.

The method of inconsistency driven pairwise comparison provides just the method that can overcome these issues.

4. Description of method

The method of pairwise comparison has a long history in decision making since Fechner used it as a scientific method in 1860. In 1927, Thurstone specified it as a 'law of comparative judgments'. It represents a method by which the many factors that underlie subjectively defined criteria (or choices) can be processed, and is intuitive in as much as decisions are reduced to a two at a time rather than all at once approach. In 1977, Saaty published his seminal paper [9]. The introduction of hierarchy and an inconsistency index made pairwise comparisons method fit for real-life applications. This has particular practical significance for those situations where direct measurements are impractical or impossible, with no standard against which a decision can be compared (e.g., social skills have no clear yardstick for measuring them).

We take for granted standards that are in common use, such as measurements of length, mass and temperature. In medical terms, body temperature can be accurately measured and fever diagnosed. Similarly, plasma haemoglobin can be measured against a standard and anaemia can be confidently diagnosed. However, how do we judge the boundaries between normal and pathological social interaction? More importantly, how do we identify the relative importance of different characteristics that are seen in a disorder in terms of their diagnostic relevance? Setting the relative importance of different features seen in Autism Spectrum Disorders (ASDs) is no trivial task. However, in this paper we will demonstrate how such decisions, crucial to diagnosis, can be effectively achieved using the pairwise approach (see Appendix A for a more detailed description of the PC approach). Some of the computer side of this approach were previously demonstrated in [4,5] but this is the first time where clinical data are used.

For busy readers who may not have time to familiarize themselves with all the references and who may be a bit intimidated by the mathematical look of Appendix A, the following informal description of the main algorithms may be useful. A square pairwise comparisons matrix is created from the partial assessments provided by a psychiatrist. The matrix is analyzed for inconsistencies. These inconsistencies needs to be decreased by presenting them to psychiatrist. System detects the most inconsistent triads (creating a cycle) of the psychiatrist's own assessments of the relative importance of two items of the ABC scale. The exhaustive search is used as there are maximum 21 elements to search. For each triad, the formula: $ii := \min(|1-(a_{ij}/a_{ik}a_{kj})|, |1-(a_{ik}a_{kj}/a_{ij})|)$ is used (the explanation of ii is in Appendix A; a_ij represents relative ratios of two compared items of ABC scale). Subsequently, geometric means of rows of the matrix are computed and normalized. The normalized vector is used as relative weights for items. A hierarchical structure is used to restrict the number of pairs (seven is commonly assumed for the limit giving $7 \times (7-1)/2$ hence 21 pairs).

The aim of this study was to examine the impact on the screening properties of the Autism Behaviour Checklist (ABC) by applying expert driven pairwise comparison to create a vector of weights that reduce the inconsistency inherent in the subjective nature of rater response, and then examine the properties with and without the weights being applied. We hypothesized that the weighted version of the questionnaire would demonstrate superior psychometric properties.

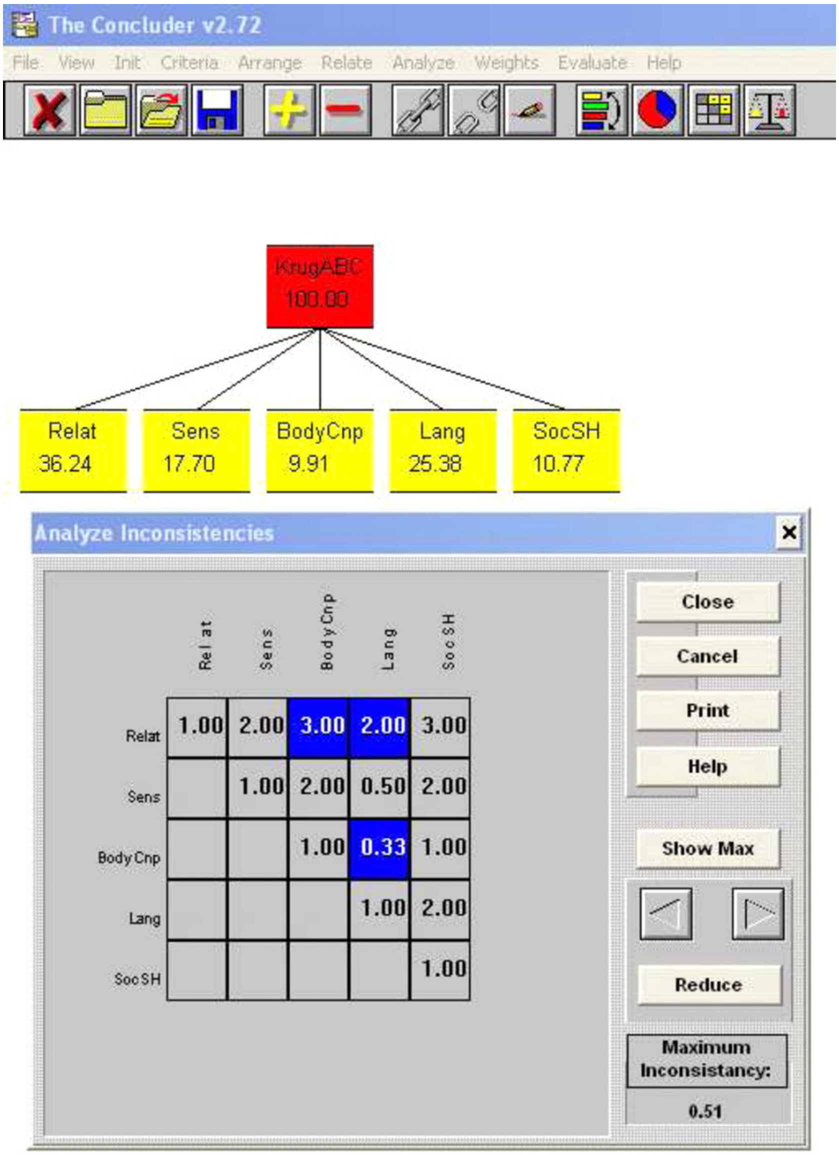
As part of ongoing collaborative genetics studies of the ASDs across Canada and internationally, a number of screening questionnaires have been completed on families with one or more child with autism. One such screening questionnaire is the Krug Autism Behaviour Checklist.

Autism Behaviour Checklist (ABC): this is a 57 item checklist developed to identify school aged children with autism through a series of questions related to autism like behaviours divided into five categories: (1) sensory, (2) relating, (3) body and object use, (4) language, and (5) social and self-help. Raters are asked to indicate the presence or absence of each behaviour, and in this way total and subdomain scores are generated. The psychometric properties have been previously investigated, with generally reasonable, although variable, validity and reliability being demonstrated.

Importantly, although Krug originally identified weights for items according to their importance, it is unclear whether these weights have been used consistently in studies examining its psychometric properties which may be one explanation for the variation of results seen.

4.1. *Procedure*

The ABC was completed by the primary caregiver for each child recruited into the study. In addition, clinician derived clinical diagnoses were assigned to children by a best



	Relat	Sens	BodyCnp	Lang	SocSH
Relat	1.00	2.00	3.00	2.00	3.00
Sens		1.00	2.00	0.50	2.00
Body Cnp			1.00	0.33	1.00
Lang				1.00	2.00
SocSH					1.00

Maximum Inconsistancy: 0.51

Fig. 1 – Model and pairwise comparisons matrix.

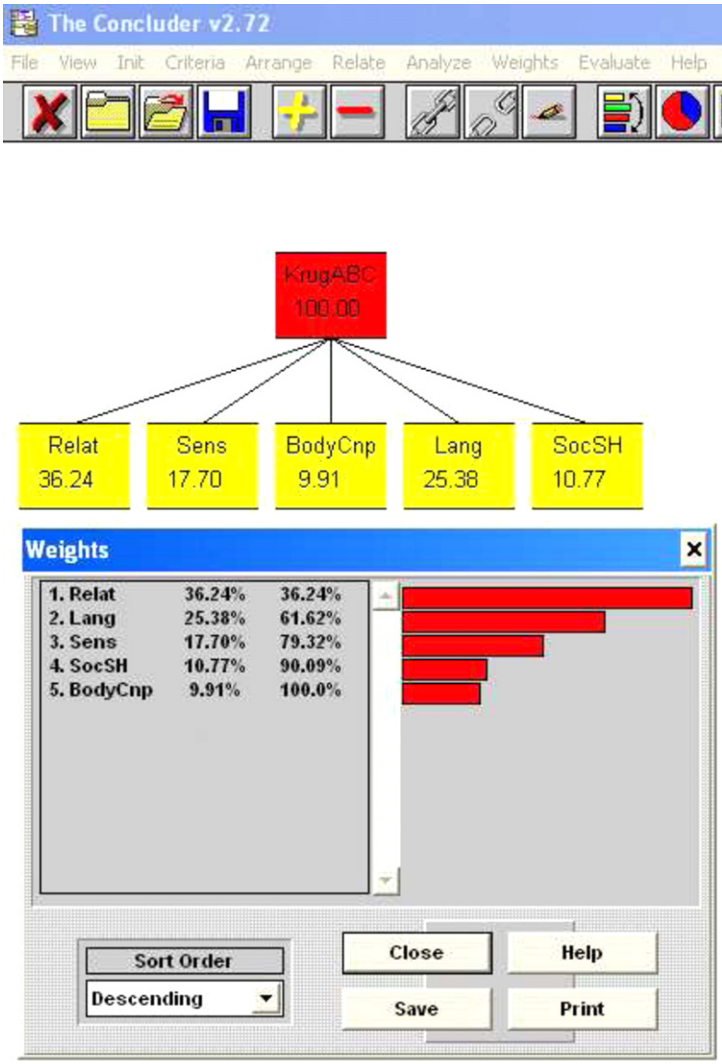
estimates consensus approach. All diagnoses were carried out blind to each child's ABC score (Fig. 1).

Area Under the Curve (AUC) represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The improvement may seem small in terms of percentage, but it is nonetheless an improvement and in the scale of a country may be translated into thousands. The estimated number of autistic children in Canada is 70,000 so 0.53% is equivalent to 370 children who could be potentially misdiagnosed and their lives ruined by both false positive and false negative. The figure may not look impressive although but taking into account that according to the International Programs Center (see [11]), U.S. Census Bureau, the total population of the World, projected to 04/17/11 at 11:44 UTC (EST+5) is 6,912,739,707 (Fig. 2). According to the World Health Organization (see [10]): "One in four patients visiting a health service has at least one mental, neurological or behavioural disorder but most of these disorders are neither diagnosed nor treated" hence is 25% can be assumed to illustrate the potential gain for the estimated number of better diagnosed patients. It would be 9,159,380 (that is, 6, 912, 739, 707 × 0.25 × 0.0053). So, it may be worth further studies (to avoid the potential of diagnosis over 9 million people) since the use of the proposed method does not require the new data collections by medical trials (usually an expensive undertaking) but just better processing these data.

5. Status report

Table 1 shows the ROC AUC analysis results for the original and improved data. Area under fitted curve (Az) has increased from 0.8088 for the original data to 0.8131 (which is 0.529% relative improvement) for the PC enhanced data while the estimated standard error has decreased 0.0346–0.0336 for this case. Trapezoidal (Wilcoxon) area has increased from 0.8091

Please cite this article in press as: T. Kakiashvili, et al., Improving the medical scale predictability by the pairwise comparisons method: Evidence from a clinical data study, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.09.011



1. Relat	36.24%	36.24%
2. Lang	25.38%	61.62%
3. Sens	17.70%	79.32%
4. SocSH	10.77%	90.09%
5. BodyCnp	9.91%	100.0%

Fig. 2 – Model and results.

Table 1 – ROC AUC analysis results for the original and improved data.

Final ROC estimates for the original data
$A = 1.1950$ std. error $(A) = 0.1683$
$B = 0.9334$ std. error $(B) = 0.1288$
Correlation $(A, B) = 0.1554$
ROC AUC:
Area under fitted curve (Az) = 0.8088
Estimated std. error = 0.0346
Trapezoidal (Wilcoxon) area = 0.8091
Estimated std. error = 0.0291

Final ROC estimates for the PC enhanced data
$A = 1.2022$ std. error $(A) = 0.1648$
$B = 0.9095$ std. error $(B) = 0.1240$
Correlation $(A, B) = 0.1638$
ROC AUC:
Area under fitted curve (Az) = 0.8131
Estimated std. error = 0.0336
Trapezoidal (Wilcoxon) area = 0.8125
Estimated std. error = 0.0287

for the original data to 0.8125 PC enhanced data while the estimated standard error has decreased 0.0291–0.0287 for this case.

The used consistency-driven pairwise comparisons method was implemented in C++during the late 1990s as a part of a project commissioned by the Ontario Ministry of Northern Development and Mines for hazard rating of abandoned mines. The software system (The Concluder) was used by authors to create the presented model. The improvement of predictability has taken place and the method performs well. However, the data quality was a little problematic and contributed to the modesty of results. In clinical practice, it is uncommon to have many questions unanswered. In fact, once disorder is recognized, a psychiatrist may not even bother with the completion of the remaining part of the questionnaire. Some answers may be hard to read and fast data entry into computer also introduces errors. We believe that these issues are unfortunately inherent in most questionnaires and survey based research, and are not peculiar only to our study.

It is not easy to obtain clinical data with the external validation. In fast, the results are always expected to be skewed.

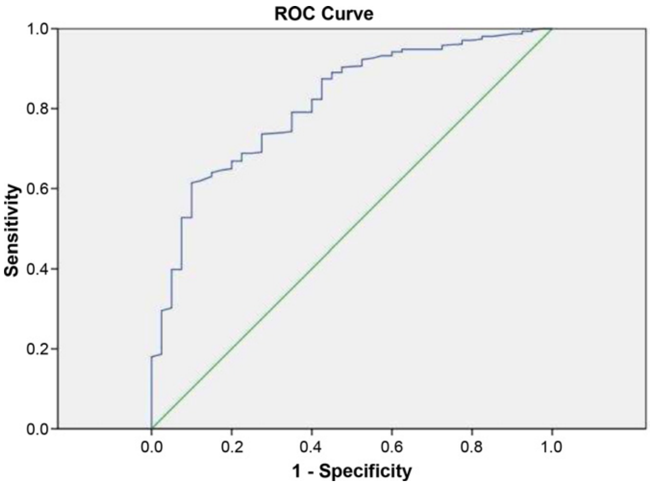


Fig. 3 – Diagonal segments are produced by ties.

In case of psychiatric clinical data, positive cases are not drawn from the general population but from the pollination of referred patients hence a borderline cases.

6. Lessons learned

An analysis of the key insights gained from the work to date, focusing on the statement of general principles that can contribute to the knowledge in the field. The statement of such lessons must be well supported by examples. The emphasis should be on the statement of principles in such a form that they will be of use to other investigators in the field.

In this data analytic study, we tested the impact of applying pairwise comparison driven weighed scores to domain scores on an Autism Spectrum Disorders screening questionnaire. As far as we are aware this is the first study to examine its effectiveness at improving the screening properties of a clinical assessment by the method which is not only fairly old, as we demonstrated earlier, and well tested in the past in other areas of application in science and engineering. Our results, although modest, were consistent with improved psychometric properties of the questionnaire examined, as evidenced by the superior AUC classifier percentage after weights were added (Fig. 3).

7. Future plans

It is also important to recognize that even small improvements can impact hugely on heath care delivery, in terms of reduced weight times and its economic impact. However, in terms of medical diagnosis and screening, the significance of the improvement may be more immediately apparent. Clinicians often struggle with the far from perfect diagnostic criteria used to diagnose mental and behavioural syndromes, and it is likely that diagnostic tests that improve diagnostic precision, such as through a blood test, are a long way off. Therefore, it is particularly exciting that the pairwise comparison method offers some optimism to facilitate the diagnostic process. By way of caution, however, the modest improvement seen also indicates that this method needs further evaluation and repetition before its wider clinical applicability can be more strongly argued. The modesty of results is a natural implication of the clinical data accuracy as well as the well researched topic (17,993 PubMed hits for 'autism' as of 2011-07-1).

In short, more studies for other scales such as the one described here, are urgently needed. The diagnostic process of developmental disorders, depression, bipolar disorder, early schizophrenia, and many other mental disorders heavily relies (in the absence of psychiatrists) on medical scales and most of them can be improved by the presented method.

Acknowledgments

Appreciation is expressed to the anonymous reviewer for his or hers kind words: "The authors presented the pilot data to show that using PC method the modest improvements in diagnostic accuracy could be achieved. It is important to observe that the pairwise comparison method offers some optimism to facilitate the medical diagnostic process."

Appendix A. Basics of pairwise comparisons

From the mathematical point of view, the pairwise comparisons method generates a matrix (say A) of ratio values (a_{ij}) of the ith entity compared with the jth entity according to a given criterion. Comparing two entities in pairs to assess which of them is preferred, or has a greater amount of some property is irreducible since having one entity compared with itself has very little or practical meaning. However, assessments often involve inconsistency, which is usually undesirable.

Making one comparison of two items at a time is simpler than simultaneously assessing *all* items of a scale according to their contribution to the overall score. However, we need a method for synthesizing these partial assessments. The pairwise comparisons method serves exactly this purpose, with the inconsistency analysis allowing us to localize the most questionable partial assessments and revise them.

A scale $\left[\frac{1}{c}, c\right]$ is used for 'i to j' comparisons where $c > 1$ is a not-too-large real number (often 5–9 is used in most practical applications). It is usually assumed that all the values a_{ii} on the main diagonal are 1 (the case of 'i compared with i', that is with itself) and that matrix A is *reciprocal*: $a_{ij} = (1/a_{ji})$ since 'i to j' is (or at least, is expected to be) the reciprocal of 'j to i'. (In other words, for x, $y \neq 0$, $x/y = 1/(y/x)$.) However, in practice even the reciprocity condition is not always guaranteed. For example, in blind wine testing we may conclude that *wine i* is better than *wine i* if it is served in unmarked glasses.

Since 1996, a *distance-based* adjective has been used by other researchers for the new inconsistency defined in 1993 in [6]. The distance-based adjective reflects the nature of the *inconsistency indicator*, which is defined, in essence, as a function of a distance from the nearest consistent *triad* in matrix A. Unlike the eigenvalue-based inconsistency, introduced in [9]), which is of a *global* indicator, and as such a non-identifying, the distance-based inconsistency identifies the most inconsistent triad (or triads). It is the maximum over all triads $\{a_{ik}, a_{kj}, a_{ij}\}$ of elements of A (say, with all i, j, k distinct) of their inconsistency indicators, which in turn are defined as $ii := \min(|1-(a_{ij}/a_{ik}a_{kj})|, |1-(a_{ik}a_{kj}/a_{ij})|)$.

The inconsistency indicator of A equals zero if and only if A is fully consistent as it was (in all likeliness shown for the first time in [9]. Consistent matrices correspond to the ideal situation in which we know all exact values of all properties (or at least it seems to be a reasonable assumption to make). However, a realistic situation which is complex enough, nearly always involves inconsistency and we need to deal with it. In fact, when we are able to locate it, our comparisons can be reconsidered to reduce the inconsistency in the next round.

Certainly, inconsistency is undesirable in a system. On the other hand, although this may sound strange, it is not easy (we suspect, impossible) to construct a non-trivial *fully* inconsistent system: an ideal system where everything contradicts everything else.

The distance-based inconsistency locates the most inconsistent triad (or triads). This allows the user to reconsider the assessments included in the most inconsistent triad.

$$\begin{array}{ccccc} & \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \\ \mathrm{A} & 1 & \boxed{1} & \boxed{5} & 4 \\ \mathrm{B} & 1 & 1 & \boxed{2} & 2\frac{1}{2} \\ \mathrm{C} & \frac{1}{5} & \frac{1}{2} & 1 & \frac{1}{2} \\ \mathrm{D} & \frac{1}{4} & \frac{2}{5} & 2 & 1 \end{array} \tag{A.1}$$

Changing the value 1 in the above triad to 2.5 makes this triad fully consistent since $2.5 \times 2 = 5$. Unfortunately, this is not the end of our problems since there is another triad $\left[2, 2\frac{1}{2}, \frac{1}{2}\right]$ that is inconsistent and "boxed" below:

$$\begin{array}{ccccc} & \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \\ \mathrm{A} & 1 & 2\frac{1}{2} & 5 & 4 \\ \mathrm{B} & \frac{2}{5} & 1 & \boxed{2} & \boxed{2\frac{1}{2}} \\ \mathrm{C} & \frac{1}{5} & \frac{1}{2} & 1 & \boxed{\frac{1}{2}} \\ \mathrm{D} & \frac{1}{4} & \frac{2}{5} & 2 & 1 \end{array} \tag{A.2}$$

Assume that we have good reason (coming from the knowledge domain; not from mathematics), to change the value of $2\frac{1}{2}$ to 1 It is an arbitrary decision since 2 could have been changed to 5 or $\frac{1}{2}$ to $1\frac{1}{4}$ also making this triad consistent. Only the domain knowledge (in our case, psychiatry) can determine the change of the value (or values) in a triad. However, changing 2 may not be wise since it belongs to a consistent triad altered in the previous step. In our case, the only reason why we have chosen to change $2\frac{1}{2}$ to 1 was to illustrate how the inconsistency procedure works and the reader may be disappointed to find that there is yet another triad "boxed" below which is inconsistent.

An acceptable threshold of inconsistency, for most practical applications, turns out to be $\frac{1}{3}$. This is so because one value in a triad is not more than two grades off the scale from the remaining two values. This heuristic was introduced in [6] and it seems more mathematically sound than 10% proposed in [9].

There is no need to continue decreasing the inconsistency indefinitely to zero, as only a high value of it is harmful. In fact, a zero or a small inconsistency value may indicate that artificial data were entered hastily without reconsideration of former assessments, which is an unacceptable practice.

It is important to note the difference between inaccuracy and inconsistency. For example, in a triad [2, 5, 3], a rash approach may lead us to believe that A/C should indeed be 6 since it is 2 × 3, but we do not have any reason to reject the estimation of B/C as 2.5 or A/B as 5/3. This is what inconsistency is about. It is not inaccuracy, but when used wisely, it may help to decrease inaccuracy.

The reader will notice that while the three-step inconsistency-reduction procedure performed above does not offend the common sense, it is rather *ad hoc*, hence not fully satisfactory. This remark applies both to the choices of triads to be corrected, and to the choices of the particular members of each such triad that is being modified. The algorithm analyzed in [8] (and, by extension, the present note) is more canonical with respect to the second point. In general, it replaces the triad $\{a_{ik}, a_{kj}, a_{ij}\}$ by $\{a_{ik}/r, a_{kj}/r, ra_{ij}\}$, where $r := \sqrt[3]{a_{ik}, a_{kj}/a_{ij}}$. This corresponds to subtracting from the matrix $(\log a_{uv})$ its orthogonal projection onto the direction of the skew-symmetric matrix $B = (b_{uv})$ defined by the requirement that $a_{ik} = 1 = a_{kj}$, $a_{ij} = -1$ and that all other super-diagonal entries are 0; the corresponding subspace in the context of Theorem is $U = \{X : X \text{ is an } n \times n \text{ skew-symmetric matrix such that } \operatorname{tr} BX = 0\}$. In particular, for the first triad [1, 2, 5] considered above, we have $r = 2/5$ and the corrected triad is $[\sqrt[3]{5/2}, 2\sqrt[3]{5/2}], 5\sqrt[3]{2/5}] \approx [1.36, 2.71, 3.68]$.

The method of pairwise omparisons was used by a research team, lead by Professor Waldemar W. Koczkodaj, to develop AMIS (Abandoned Mines Hazard Rating System) for the government of Ontario (The Ministry of Northern Ontario and Mines). The system ranked an abandoned mine, located in Northern Ontario, as one of the most dangerous from a public safety point of view. Its eventual collapse convinced the government that its research founding was well spent.

REFERENCES

[4] T. Kakiashvili, W.W. Koczkodaj, P. Montgomery, K. Passi, R. Tadeusiewicz, Assessing the properties of the World Health Organization's Quality of Life Index, in: Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT, IEEE Xplore, 2008, pp. 151–154.

[5] T. Kakiashvili, W.W. Koczkodaj, D. Matheson, P. Montgomery, K. Passi, F. Rybakowski, R. Tadeusiewicz, M. Woodbury-Smith, Supporting a medical diagnostic process by selected AI methods: an Asperger Syndrome case study, Studia Informatica 1 (10) (2008) 5–13.

[6] W.W. Koczkodaj, A new definition of consistency of pairwise comparisons, Mathematical and Computer Modelling 18 (7) (1993) 79–84.

[8] W.W. Koczkodaj, S.J. Szarek, On distance-based inconsistency reduction algorithms for pairwise comparisons, Logic Journal of the IGPL 18 (6) (2010) 859–869.

[9] L.T. Saaty, A scaling method for priorities in hierarchical structures, Journal of Mathematical Psychology 15 (3) (1977) 234–281.

[10] World Health Organization, WHO urges more investments, services for mental health. http://www.who.int/mental_health/who_urges_investment/en/index.html (accessed 17.04.11).

[11] U.S. Census Bureau, World POPClock Projection by International Programs Center web page http://www.census.gov/ipc/www/popclockworld.html (accessed 17.04.11).

Please cite this article in press as: T. Kakiashvili, et al., Improving the medical scale predictability by the pairwise comparisons method: Evidence from a clinical data study, Comput. Methods Programs Biomed. (2011), doi:10.1016/j.cmpb.2011.09.011