Measuring Agreement on Linguistic Expressions in Medical Treatment Scenarios

J. Navarro[a], C. Wagner[ab], U. Aickelin[c], L. Green[d], R. Ashford[d]
[a]Lab for Uncertainty in Data and Decision Making (LUCID) and Horizon,
School of Computer Science, University of Nottingham, Nottingham, UK
[b]Institute of Computing & Cybersystems, Michigan Technological University
[c]University of Nottingham, Ningbo, China
[d]Nottingham University Hospitals, UK
Email: {psxfjn,christian.wagner,uwe.aickelin}@nottingham.ac.uk
Email: lynsey.green@nuh.nhs.uk, rashford@nhs.net

***Abstract*—Quality of life assessment represents a key process of deciding treatment success and viability. As such, patients' perceptions of their functional status and well-being are important inputs for impairment assessment. Given that patient completed questionnaires are often used to assess patient status and determine future treatment options, it is important to know the level of agreement of the words used by patients and different groups of medical professionals. In this paper, we propose a measure called the Agreement Ratio which provides a ratio of overall agreement when modelling words through Fuzzy Sets (FSs). The measure has been specifically designed for assessing this agreement in fuzzy sets which are generated from data such as patient responses. The measure relies on using the Jaccard Similarity Measure for comparing the different levels of agreement in the FSs generated. Synthetic examples are provided in order to show how to calculate the measure for given Fuzzy Sets. An application to real-world data is provided as well as a discussion about the results and the potential of the proposed measure.**

***Index Terms*—Survey data, Computing with Words, Interval Agreement Approach, Similarity, Questionnaires.**

I. Introduction

In the context of medical treatment, capturing patients' and medical professionals' perceptions of functional status is an important instrument to consider when evaluating possible outcomes after treatment intervention (e.g., job modifications, use of assistive devices, etc.) [1]. In this context, it is important to be aware of the uncertainty associated to the words (linguistic descriptors) used by the stakeholders (e.g., physiotherapists, surgeons, patients, etc.): this includes variability in people's perceptions throughout the day, experience, professional background, etc.

In [2], Zadeh introduced the Computing With Words (CWW) paradigm in which, according to him, words and propositions from natural language are used as objects of computation. As such, this paradigm focuses on narrowing the differences between human reasoning and computing by allowing the manipulation of different and usually imprecise perceptions [3].

This work was partially funded by the RCUK grant EP/M02315X/1 From Human Data to Personal Experience and Prototyping Open Innovation Models for ICT-Enabled Manufacturing in Food and Packaging EP/K014234/2.

Surveying groups of people enables the capture of uncertainty through intervals on areas of interest, which is an important resource for capturing the variations of perceptions among those surveyed. A number of methods have been developed in order to capture stakeholder perceptions of words/concepts expressed through interval-based surveys, including [4], [5], [6]. Basically, they rely on allowing participants to express their uncertainty about a given response by providing an interval. Such participants' intervals are subsequently used to generate a Fuzzy Set model (which depends of the method employed and the types of uncertainty being modelled) representing the overall perception (or a subset of the participants) of the initial word surveyed.

There are a number of measures for Type-1 Fuzzy Set (T1 FS) agreement models which have shown to be useful for several purposes. Similarity measures [7] for example, are functions which indicates the degree to which two FSs are similar. The Jaccard similarity measure [8], has been applied to both relate and compare word models to concept models in different contexts [9], [10]. In [11], an exploration of attributes (e.g., *Support Size*, *Height*, *Spread*, *Core Size* and *Fuzziness*) obtained from T1 FSs agreement models was performed. Such exploration found that additional information related to the consensus can be extracted with regard to traditional statistical measures. However, a direct measure of agreement among participants expressing how well conceived a given word is in an specific context where imprecise descriptors are being used as a basis to assess patients' health conditions has not been reported yet.

In this paper, we focus on presenting an Agreement Ratio measure which aims to provide a measure for the inter-participant agreement on perception of words (linguistic descriptors) using T1 FSs derived from interval-based surveys. This paper is structured as follows. Section II provides background on a questionnaire called Toronto Extremity Salvage Score (TESS) for the assessment of impairment in which linguistic descriptors are a key element for expressing patients' perceptions, T1 FSs and modelling of inter-participant

uncertainty using the Interval Agreement Approach. Section III introduces the motivation behind the proposal of this measure and details about its practical implementation. Section IV presents a series of numeric examples for different data followed by the application to real world data obtained from patients and medical professionals using an interval-valued questionnaire. Finally, Section V provides a discussion of the results presented and the application of the proposed measure on different types of T1 FSs, while Section VI presents the conclusions and future work derived from this proposed measure.

II. Background

The following introduces the TESS questionnaire used to assess patient' functional status, which will serve as the context for the experiments presented later in the paper. It is followed by a brief overview of T1 and Interval Type-2 FSs and two methods employed to model the agreement among a group of stakeholders using FSs. Both approaches will be used in the paper to generate FSs from data which in turn will be evaluated using the proposed measure.

A. Toronto Extremity Salvage Score (TESS)

The Toronto Extremity Salvage Score (TESS) is a disease-specific measure developed for patients undergoing limb preservation surgery for tumours of the extremities [1]. It is a patient-completed questionnaire with questions framed to ask about the difficulty experienced performing daily activities over the last week aimed to monitor the effects of therapeutic interventions. TESS is commonly administered at four time points: the first session (which is commonly before surgery) and 12, 18 and 24 months from then on. The TESS consists of 30 and 29 items for lower limb and upper limb cases respectively with items such as the the one shown in Fig. 1. As can be seen, difficulty is rated on a 5-point Likert-type

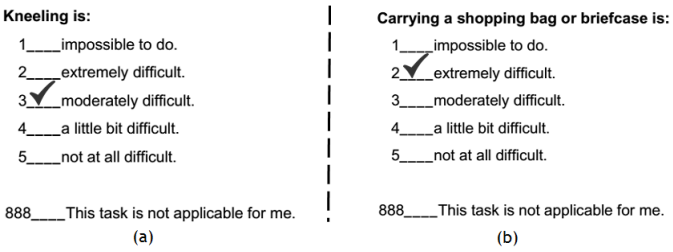


Fig. 1: Two sample TESS items: (a) item taken from the lower extremity questionnaire, (b) item taken from the upper extremity questionnaire.

scale ranging from "*not at all difficult*" to "*impossible to do*". Commonly, after having been completed by the patient, the whole set of answers is used to generate a standardized score ranging from 0 to 100. This evaluated TESS is finally analysed by surgeons/physiotherapists in order to measure changes in physical functions over time.

B. Type-1 Fuzzy Sets

Fuzzy Sets are sets in which, unlike in traditional set theory, the membership of each element is a number in the interval $[0, 1]$. Given a universe of discourse X, a FS A is represented as a set of ordered pairs of an element x and its membership value within A, denoted by $\mu_A(x)$, i.e.

$$A = (x, \mu_A(x))|x \in X \tag{1}$$

1) Alpha-cuts: Alpha-cuts (or α-cuts) are an important concept in FSs, given that a FS A can also be represented as a collection of its α-cuts [12]. An α-cut of a FS A is a crisp set defined as

$$A_\alpha = \{x|\mu_A(x) \geq \alpha, \alpha \in [0, 1]\} \tag{2}$$

2) Fuzziness: The measure of fuzziness is a function $f(A)$ which assigns a non-negative real number to a given FS A expressing the degree to which the boundary of a A is not sharp[13]. Fuzziness of a FS is then defined as

$$f(A) = \sum_{x \in X} (1 - |2\mu_A(x) - 1|) \tag{3}$$

The presented measure of fuzziness satisfies three essential requirements for fuzziness measures.

1) $f(A) = 0$ if and only if A is a crisp FS (see Fig. 2a).
2) $f(A)$ attains its maximum value if and only if $\mu_A(x) = 0.5, \forall x \in X$.
3) $f(A) \leq f(B)$ when A is "sharper" than B, i.e., $\mu_A(x) \leq \mu_B(x)$ when $\mu_B(x) \leq 0.5$ and $\mu_A(x) \geq \mu_B(x)$ when $\mu_B(x) \geq 0.5$ for all $x \in X$.

C. Interval Agreement Approach

The Interval Agreement Approach (IAA) was introduced in [4] as a method for generating FSs from surveys in which answers are given as interval-valued data representing uncertainty in people's opinions/perceptions. It is built on top of the work presented in [14], where an agreement-based method [15] of capturing interval-valued survey data is demonstrated.

The IAA considers two types of intervals in the process of capturing responses: crisp (no uncertainty about the interval endpoints) and uncertain (each endpoint modelled itself as a crisp interval). It considers two types of uncertainty to be modelled through different dimensions of the resultant FSs, namely inter-source (variation among a group of participants) and intra-source (variation in the opinion of a particular participant). Depending on the data, the IAA can generate:

- Type-1 FSs. When crisp intervals and either inter- or intra-source uncertainty are modelled in the primary degree of membership by combining multiple intervals,
- Interval Type-2. When uncertain intervals and also, either inter- or intra-source uncertainty is modelled in the primary degree of membership by combining multiple intervals,
- General Type-2 FS based on zSlices [16]. In this case, both inter- and intra-source uncertainty are being modelled through the primary and secondary degrees of membership.

In this paper, we are focusing on the agreement ratio among stakeholders (inter-participant uncertainty). Such uncertainty is captured through crisp intervals and consequently, it is modelled by employing the IAA to generate T1 FSs. We provide a brief review of generating T1 FSs using the IAA below:

Consider N (closed) intervals $\bar{A}_i = [l_{\bar{A}_i}, r_{\bar{A}_i}]$, $i \in \{1, ..., N\}$ to be modelled as a T1 FS A where the intervals are delimited by $l_{\bar{A}_i}$ and $r_{\bar{A}_i}$. The membership function of A (denoted by μ_A) given in (4).

$$\begin{aligned}\mu(A) = y_1 / \bigcup_{i_1=1}^{N} \bar{A}_{i_1} \\ + y_2 / \left(\bigcup_{i_1=1}^{N-1} \bigcup_{i_2=i_1+1}^{N} \left(\bar{A}_{i_1} \cap \bar{A}_{i_2} \right) \right) \\ + y_3 / \left(\bigcup_{i_1=1}^{N-2} \bigcup_{i_2=i_1+1}^{N-1} \bigcup_{i_3=i_2+1}^{N} \left(\bar{A}_{i_1} \cap \bar{A}_{i_2} \cap \bar{A}_{i_3} \right) \right) \\ + \cdots \\ + y_N / \left(\bigcup_{i_1=1}^{1} \cdots \bigcup_{i_N=N}^{N} \left(\bar{A}_{i_1} \cap \cdots \cap \bar{A}_{i_N} \right) \right),\end{aligned} \tag{4}$$

where $y_i = \frac{i}{N}$ and $/$ refers to the common notation of membership, not division. For practical applications, (4) can be calculated in a recursive and discrete manner by formulating the function as

$$\mu_A(x') = \frac{\left(\sum_{i=1}^{N} \mu_{\bar{A}_i}(x') \right)}{N}, \tag{5}$$

$$\text{where: } \mu_{\bar{A}_i}(x') = \begin{cases} 1 & l_{\bar{A}_i} \leq x' \leq r_{\bar{A}_i} \\ 0 & else \end{cases}.$$

III. Agreement Ratio

This section proposes a method of generating a useful value for the analysis of linguistic information represented as FSs, called the Agreement Ratio. Section III-A reviews the aims and motivation behind the proposed agreement ratio, followed in Section III-B by an in-depth description of the measure.

A. Motivation

The use of words as a means of communication between patient-medical staff is a natural way of expressing perceptions. However, the challenge of dealing with different interpretations of words ("words mean different things to different people") leads to looking for a standardised vocabulary, as has been suggested by the European Union [17]. Therefore, there is a need for a means to analysing how similar the understanding of key words is across stakeholders in patient treatment.

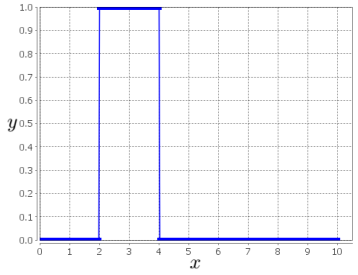
The aim of the agreement ratio is to provide a number contained within the interval $[0, 1]$ representing the extent of agreement among a group of surveyed stakeholders (without discarding particular responses) whose responses are modelled in the given FS. Therefore, a method for generating data-driven FS models, considering the stakeholders opinions is key. As such, we have considered FSs generated with the IAA and EIA methods to analyse the agreement obtained.

Two key assumptions for the measure are:

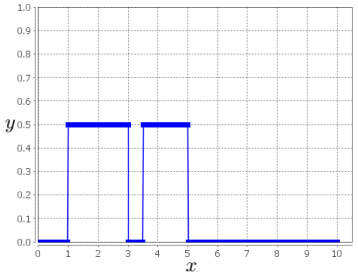
1) *Agreement* is when 2 or more sources coincide in a given point/value/opinion.
2) The more opinions overlap at a particular region, the stronger the agreement is conceived.

Initially, we began developing the measure considering that in the IAA given the inter-participant variation is reflected directly in the primary degree of membership y (or μ). To illustrate, consider two simple contrasting cases in which two intervals/participants ($N = 2$) are used to create a FS:

- Two identical intervals. An agreement ratio must be equal to 1 given that all stakeholders (intervals) totally agree (overlap).
- Two disjoint intervals. An agreement ratio must be equal to 0 given that there are no regions in which the stakeholders agree (Fig. (2b)).



(a)



(b)

Fig. 2: FSs generated from the intervals $\overline{A}_1 = [2, 4]$ and $\overline{A}_2 = [2, 4]$ for Fig. (2a), and the intervals $\overline{A}_1 = [1, 3]$ and $\overline{A}_2 = [3.5, 5]$ for Fig. (2b).

From these initial cases, it can be seen that at the highest level of membership (we will refer to it as y_N), there might be one or more intervals representing regions where the N intervals agree, at the y_{N-1} level, there might be one or more overlapping regions where at least $N - 1$ intervals agree and so on. Thus, if a proportion of each of the y levels contained with the next lower level is calculated, then a ratio representing their quantitative relation. For example, in the FS depicted in Fig. 2a, the length of the agreement interval at the y_2 level is 2, which is equal to the one at the y_1 level and thus, the

relation can be represented as $\frac{2}{2} = 1$. For the FS of figure 2b, such relation between the length at level y_2 is 0 since there is not any region where both intervals overlapped and the length at level y_1 is $2 + 1.5 = 3.5$ can be represented as $\frac{0}{3.5} = 0$. Moreover, considering cases with more intervals to analyse, regions with higher agreement over others with less must contribute to the ratio with "higher relevance".

Using these cases as basis we can proceed to generalise and propose a method for calculating an agreement ratio for a FS in Section III-B.

B. Method

Let $\bar{A}_n$, $n \in \{1, ..., N\}$ be a set of intervals $\bar{A}_i = \left\{l_{\bar{A}_i}, r_{\bar{A}_i}\right\}$. The IAA uses the set of intervals to model the overall agreement through a FS based model where the membership value of each $x \in X$ accounts for the ratio of a given x contained in the set of intervals.

An agreement ratio γ is obtained from a T1 FS with the following equation:

$$\gamma(A) = \left(y_N \left(\frac{\left|\bar{A}\right|_N}{\left|\bar{A}\right|_{N-1}} \right) + \cdots + y_2 \left(\frac{\left|\bar{A}\right|_2}{\left|\bar{A}\right|_1} \right) \right) / \sum_{i=2}^{N} y_i \quad (6)$$

where $0 \leq \gamma \leq 1$, / refers to division and $y_i = \frac{i}{N}$ weights the relation between immediate agreement y levels in question. It can be noticed that the lowest level y_1 is not being used because the agreement is conceived when 2 or more intervals overlap. Also, we use $\left|\bar{A}\right|_i$ to represent the total length(s) of the set(s) of intervals with all possible *i*-tuple intersection of intervals associated to the y_i agreement level. For example, the length $\left|\bar{A}\right|_1$ is equal to the length of the support of A since it is the union of all intervals whereas $\left|\bar{A}\right|_N$ is equal to the length of the intersection of all intervals. Finally, the overall summation is divided by the sum of "weights" so the final ratio is normalised to a number in the range [0,1].

The term $\left|\bar{A}\right|_N$ can be represented as described in (7).

$$\left|\bar{A}\right|_N = \sum_{i_1=1}^{1} \cdots \sum_{i_N=N}^{N} \left|\bar{A}_{i_1} \cap \cdots \cap \bar{A}_{i_N}\right| \quad (7)$$

Considering that such calculations can involve handling a considerable number of combinations to compute as the number of participants/intervals increases, (7) can be estimated through "discretisations" in practical applications by using alpha cuts (instead of the so-called y agreement levels) such as described in Algorithm1. A consideration for the calculation of the γ measure using alpha-cuts is: if IAA generated FS models are used, then the number of alpha cuts can be chosen to be equal to the number of intervals/participants; if any other method is used, then it depends of the number of desired "discretisations". It should be noticed that, the use of alpha cuts can allow the Agreement Ratio to be applied in any T1 FS regardless the method employed to generate it from intervals. However, as stated in the motivation, the degree of membership of the FS in question is assumed to express the extent of agreement among the surveyed stakeholders. This assumption, allows the proposed measure to take advantage of different FSs (normal or non-normal, convex or non-convex) from an interpretative point of view, in order to show so, we will analyse different FSs shapes and provide the results in next section.

Algorithm 1 Estimation of lengths based on α-cuts

```
procedure AlphaLength(α, A)  ▷ The sum of lengths
    l ← 0, r ← 0
    N ← # of discretisations
    discretise(x)            ▷ discretise domain x_1, ..., x_i, ...x_N
    b ←false                 ▷ Boolean for detection of intervals
    for i = 1 to N do
        y_i ← μ_A(x_i)
        if y_i < α then
            y_i ← 0
            if b = true then
                r ← x_{i-1}
                addCut(l, r)     ▷ Add the detected interval
            b ← false
        else
            if b =false then
                l ← x_i
            b ← true
    if b = true then
        r ← x_N
        addCut(l, r)
    for each αCut j do
        length ← length + (r_j − l_j)        ▷ Interval Size
    return length
```

IV. Results

In this section, we present synthetic examples of application of the proposed agreement ratio considering diverse types of T1 MFs and finally, its application to real-world data obtained from three groups of people involved in a pilot study. For the real-world application, we present the results of using models obtained from the IAA.

A. Synthetic Examples 1 using convex FSs

Consider the FSs depicted in Fig. 3 where no assumptions about the method used to generate them has been made other than, that the membership axis represents the level of agreement among the participants. Thus, one of the FSs depicted has been chosen to be non-normal deliberately so values from the calculated Agreement ratio can be contrasted. For comparison purposes, we also consider two common shapes for FS membership functions: triangular and trapezoidal. For the FS *A*, the calculated agreement ratio using 10 alpha cuts is $\gamma(A) = 0.5904$ whereas for the FS *B* $\gamma(B)$ is 0.9578. Note that as expected, for the FS *A* the agreement is considerably smaller than for FS *B* since its shape is *sharper* and *shorter*

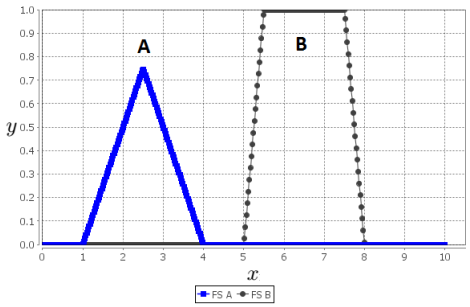


Fig. 3: Two convex Fuzzy sets A (non-normal) and B (normal).

due to the differences in the weighted comparisons of alpha-cut lengths.

B. Synthetic Example 2 using Gaussian FSs

Consider the FSs G_1,G_2 and G_3 with Gaussian membership functions depicted in Fig. 4. These FSs are both normal and convex which, from the perspective of the assumptions made in order to develop the measure, indicates that all of the intervals/participants have agreed in a region. They have been arbitrarily chosen to have the same mean ($m = 5$) but different standard deviations ($\sigma_1 = 0.1$, $\sigma_2 = 1.0$, $\sigma_3 = 2.0$). Again, no assumptions about the method employed to generate them have been made. By using 10 alpha cuts, the calculated agreement

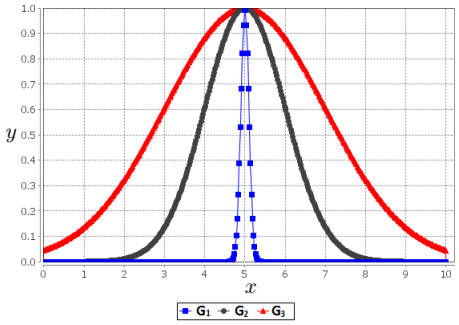


Fig. 4: Three Fuzzy sets with Gaussian membership functions.

ratio for the three FSs is $\gamma(G_1) \approx \gamma(G_2) \approx \gamma(G_3) \approx 0.6518$ and similar results can be found by using different numbers of alpha-cuts. Further discussion of these results are found on Section V.

C. Synthetic Example 1 using IAA generated FS

Consider the FS C depicted in Fig. 5 generated by two intervals $\bar{C}_1 = [2, 4]$ and $\bar{C}_2 = [2.5, 3.5]$. The agreement ratio $\gamma(C)$ is calculated by dividing the total length of the union of all combinations of intervals where there is at least an intersection of 2 intervals (y_2) by the union (y_1) of all intervals.

$$\gamma(C) = 1\left(\frac{1}{2}\right)/1 = 0.5$$

Note that in this example with 2 intervals, it can be simply considered as the length of the intersection of both intervals divided by the length of the union.

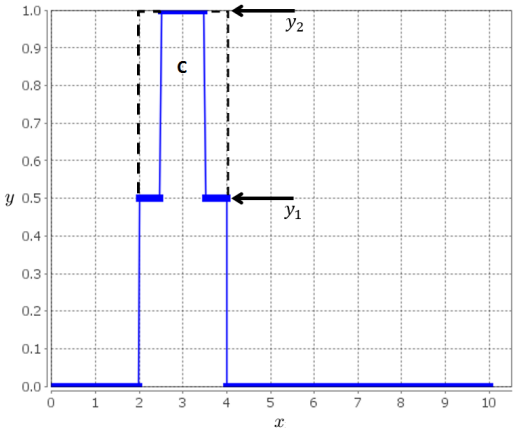


Fig. 5: Fuzzy set generated from 2 intervals

D. Synthetic Example 2 using IAA (non-convex)

Consider the FS A depicted in Fig. 6 generated by the intervals $\bar{D}_1 = [2, 5]$ and $\bar{D}_2 = [3, 5]$, $\bar{D}_3 = [6, 8]$ and $\bar{D}_4 = [3, 7]$. The agreement ratio $\gamma(D)$ is obtained by adding the weighted (y_4, y_3 and y_2) similarities between the lengths of the union of combinations of intervals where there is at least 4 and 3, 3 and 2, and 2 and 1 intervals respectively:

$$\gamma(D) = \frac{1\left(\frac{0}{2}\right) + \frac{3}{4}\left(\frac{2}{3}\right) + \frac{2}{4}\left(\frac{3}{6}\right)}{1 + \frac{3}{4} + \frac{2}{4}} = 0.333$$

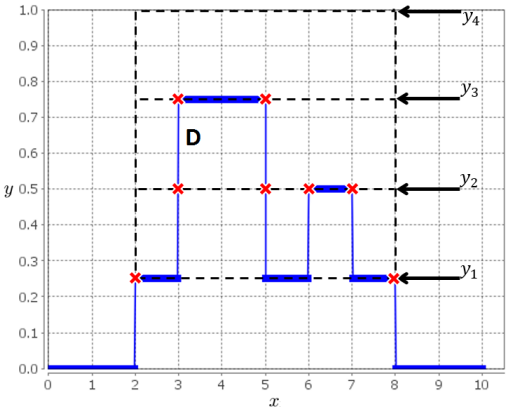


Fig. 6: Fuzzy set generated from 4 intervals

As can be seen in the above example (marked with red crosses), the length at the y_4 level is 0, at the y_3 level is 2, at the y_2 level is 3 and at the y_1 level is 6. Note that at the y_2 level there are 2 intervals which have to be added. Now lets consider the FS E depicted in Fig. 7 created using the next intervals: $\bar{E}_1 = [2, 5]$ and $\bar{E}_2 = [3, 5]$, $\bar{E}_3 = [4, 6]$ and $\bar{E}_4 = [3, 7]$. Note that they are almost the same intervals than for FS D but except one and such difference allows the generated FS to have a region with total agreement, i.e., the interval $[4, 5]$. As such, we can expect a higher agreement

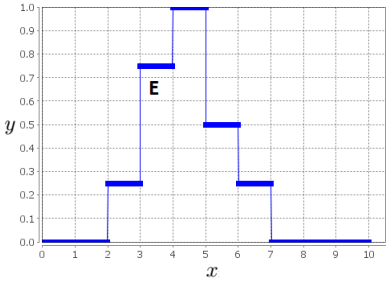


Fig. 7: Fuzzy set E generated from 4 intervals

ratio when comparing $\gamma(E)$ to $\gamma(D)$. Calculations using (6) and (7) are shown below:

$$\gamma(E) = \frac{1\left(\frac{1}{2}\right)+\frac{3}{4}\left(\frac{2}{3}\right)+\frac{2}{4}\left(\frac{3}{5}\right)}{1+\frac{3}{4}+\frac{2}{4}} = \frac{4}{7} \approx 0.5778$$

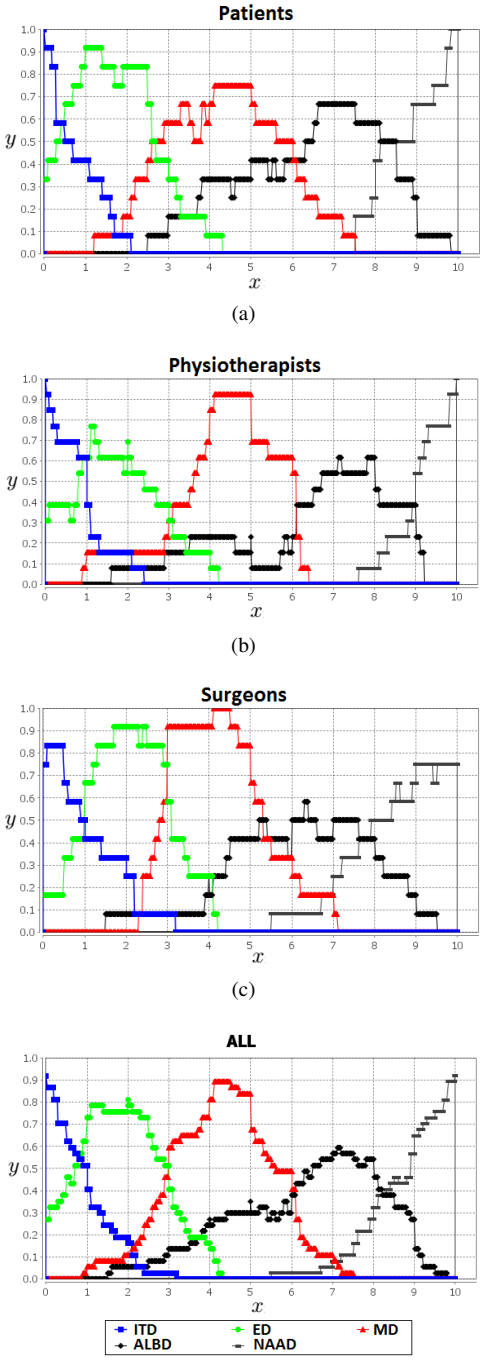
E. Application to TESS Data using the IAA

In our previous work [10], we described a process of interval-valued data collection from different groups of people involved in assessment of function following sarcoma surgery, namely: Patients, Physiotherapists, Surgeons and a fourth one created from the combined responses from both Physiotherapists and Surgeons (PS) which together represent the body of "medical professionals".

We surveyed thirty-seven participants (12 sarcoma surgeons, 13 physiotherapists and 12 patients undergoing lower limb salvage surgery) on 5 linguistic terms used to describe the extent of difficulty to perform daily activities: "*impossible to do*","*extremely difficult*", "*moderately difficult*", "*a little bit difficult*", and "*not at all difficult*". Subsequently, we used the gathered intervals in order to generate T1 FSs by using the IAA and calculated basic FS attributes (e.g., height, centroid) and their respective agreement ratio using (6) and Algorithm 1 (see Table I). For comparison purposes, we provide the measure of *fuzziness*, which captures the vagueness of a FS as described in Section II-B2.

Figure 8 depicts the FSs for the inter-patient agreement for the 5 linguistic descriptions (words) and different groups. Note that at first sight, by considering the width and height of the FSs, the term *A little bit difficult* is the most subjective and less accepted among the different groups of stakeholders. Moreover, the agreement ratio (γ) substantially facilitates performing judgements about the acceptance of the different linguistic terms in a given scenario.

From Table I, it is worthwhile to note that the linguistic terms *A little bit difficult* and *Moderately difficult* from *Surgeons* are contrastingly the less and most agreed, respectively. Such information is not inferred from other FS measures (e.g., height, centroid, support size) and can help to analyse through a numerical approach the aptness of different linguistic terms in specific contexts.



(a) (b) (c) (d)

Fig. 8: FSs modelling the word concepts from left to right: *ITD*, *ED*, *MD*, *ALBD*, and *NAAD*, generated from different sources: (a) is for patients. (b) is for Physiotherapists. (c) is for Surgeons. (d) is for the combined responses from all groups.

V. Discussion

The proposed Agreement Ratio was developed as a means of having a tool of analysis for deciding which linguistic descriptors for patients' conditions can be more apt for a given scenario. As previously mentioned, it is of a key importance

TABLE I: Agreement ratios in the context of other FS measures. Bold numbers highlight the greatest values calculated for height, fuzziness and agreement ratio.

Group	Ling. Term	Height	Centroid	Fuzziness	Agr. Rat.
Patient	ITD	**1.000**	0.686	121.667	0.669
	ED	0.917	1.711	96.167	0.652
	MD	0.750	4.356	121.167	0.433
	ALBD	0.667	6.433	136.667	0.323
	NAAD	**1.000**	9.051	91.167	0.775
Physio	ITD	**1.000**	0.727	122.308	0.629
	ED	0.769	1.767	122.462	0.368
	MD	0.923	4.312	117.231	0.733
	ALBD	0.615	6.462	**157.385**	0.242
	NAAD	**1.000**	9.279	118.154	0.701
Surgeon	ITD	0.917	0.988	129.167	0.554
	ED	0.917	2.085	92.833	0.691
	MD	**1.000**	4.289	92.333	**0.803**
	ALBD	0.583	6.126	156.333	0.185
	NAAD	0.833	8.555	118.667	0.461
ALL	ITD	0.919	0.817	141.622	0.687
	ED	0.811	1.862	106.108	0.559
	MD	0.892	4.319	129.568	0.717
	ALBD	0.595	6.346	156.432	0.280
	NAAD	0.919	8.901	137.405	0.719

to know how similar the perception of the meaning of a given word is across different stakeholders while taking into account participants' uncertainty represented through intervals. We have shown results considering synthetic examples with FSs using some of the most used types of membership function and a data-driven approach (IAA) in which it can be highlighted that:

1) The application of the proposed measure produces meaningful results when using FSs which express both, inter-source uncertainty in the domain and agreement in the membership.
2) The measure does not consider the whole scale being surveyed, but only the function domain and the proportional changes in ratio from the membership function support to the top (highest agreement). Therefore, the application of the proposed measure to Gaussian FSs produces similar values due to the Gaussian shape "smoothness" scaled through different function supports.

Regarding the application of the measure to the data obtained from the survey on TESS linguistic descriptors, we acknowledge that this is a limited sample size but if low agreement values still being obtained for *ED* and *ALBD* from a larger sample, then this may suggest that there is a potential risk of miscommunication in this scenario. Consequently, a set of different (and more unanimously understood) linguistic expressions could be sought to replace the current ones.

In the particular case of the comparison of the Agreement ratio against the measure of *fuzziness*, while both measures are similar in the sense that they both analyse the degree of "sharpness", they differ in particular at their extremes i.e., where they provide their minimum and maximum values respectively (i.e., null agreement and maximum fuzziness). That is, consider n y levels (where $n \geq 2$, y_n is the level at the top and y_1 the level at the bottom) the agreement ratio returns 0 if and only if the length of the interval(s) obtained by the α-cut at a given y_2-level is 0 and therefore there is no overlapping with the y_1 level below, while the fuzziness of a FS A attains its maximum iff $\mu_A(x) = 0.5$ for all $x \in X$ (as depicted in Fig. 9). A statistical analysis comparing fuzziness and the Agreement ratio for all stakeholder groups and linguistic descriptors using the Pearson correlation coefficient indicates a negative correlation of -0.688 with a p-value of 7.9295×10^{-4}.

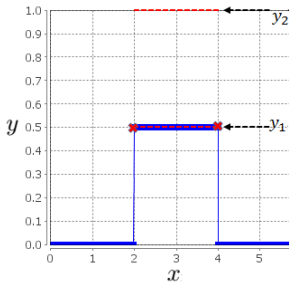


Fig. 9: Non-normal fuzzy set resulting on maximum fuzziness and lowest agreement (using 2 y levels).

VI. Conclusions and Future Work

In this study, we proposed a simple method to obtain an agreement ratio focused on inter-participant agreement through FSs generated from a data-driven approach, namely the IAA. We provided synthetic examples to show the calculations and also the results of the measure on a real world dataset obtained from different groups of people involved in a medical assessment scenario in which perceptions are key. The results show that the proposed measure can provide directly a means of evaluating the aptness of a Fuzzy Set representing a word in a given group over others. This measure has an important potential in several medical-patient intercommunication scenarios in which differences in background and context may produce misleading /assessments interpretations among different groups.

We foresee the proposed measure's usefulness in practical scenarios in which decision based on linguistic assessments are needed. For example, it can be useful to analyse a codebook with potential linguistic terms as candidates in which it is needed to avoid ambiguity as much as possible, e.g., by grouping/ranking similar terms using a defined criterion (centroid, etc.) and selecting those with the highest agreement ratio γ. Another application can be to use the agreement ratio to measure the level of consensus and allow discussion of the results among the stakeholders and repeat the survey process until more considerable agreement ratios are obtained. Although the measure proposed in this paper has only been designed for T1 FSs, we have already explored the extension of the measure to T2 FSs which will be presented in a future publication. The

extension is focused on enabling the application of the measure to other common FS generation techniques which generate T2 FSs (e.g., the Enhanced Interval Approach). We also plan to develop a more detailed methodology for the selection of words for CWW engines based on the proposed agreement ratio and explore its results in comparison to other approaches.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of Josie McCulloch for all her beneficial writing assistance and proof reading on the paper.

REFERENCES

[1] A. M. Davis, J. G. Wright, J. I. Williams, C. Bombardier, A. Griffin, and R. S. Bell, "Development of a measure of physical function for patients with bone and soft tissue sarcoma," *Quality of Life Research*, vol. 5, no. 5, pp. 508–516, 1996.

[2] L. A. Zadeh, "Fuzzy logic equals Computing with words," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 103–111, 1996.

[3] F. Herrera, S. Alonso, F. Chiclana, and E. Herrera-Viedma, "Computing with words in decision making: Foundations, trends and prospects," *Fuzzy Optimization and Decision Making*, vol. 8, pp. 337–364, 2009.

[4] C. Wagner, S. Miller, J. Garibaldi, D. Anderson, and T. Havens, "From Interval-Valued Data to General Type-2 Fuzzy Sets," *IEEE Transactions on Fuzzy Systems*, vol. 23, pp. 248–269, 2014.

[5] S. Coupland, J. M. Mendel, and D. Wu, "Enhanced Interval Approach for encoding words into interval type-2 fuzzy sets and convergence of the word FOUs," in *International Conference on Fuzzy Systems*. IEEE, 2010, pp. 1–8.

[6] L. Feilong and J. M. Mendel, "Encoding Words Into Interval Type-2 Fuzzy Sets Using an Interval Approach," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 6, pp. 1503–1521, 2008.

[7] M. Setnes, R. Babuska, U. Kaymak, and H. R. van Nauta Lemke, "Similarity measures in fuzzy rule base simplification." *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 28, no. 3, pp. 376–86, 1998.

[8] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bulletin de la Société vaudoise des sciences naturelles*, vol. Volume 5, no. 163, 1908.

[9] C. Wagner, S. Miller, and J. M. Garibaldi, "Similarity based applications for data-driven concept and word models based on type-1 and type-2 fuzzy sets," *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–9, 7 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6622466

[10] J. Navarro, C. Wagner, U. Aickelin, L. Green, and R. Ashford, "Exploring Differences in Interpretation of Words Essential in Medical Treatment by Patients and Medical Professionals," in *IEEE World Congress On Computational Intelligence*, Vancouver, Canada, 2016.

[11] S. Miller, C. Wagner, and J. M. Garibaldi, "Exploring Statistical Attributes Obtained from Fuzzy Agreement Models," in *IEEE International Conference on Fuzzy Systems*, 2014.

[12] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-I," *Information sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[13] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, 1995.

[14] S. Miller, C. Wagner, J. M. Garibaldi, and S. Appleby, "Constructing general type-2 fuzzy sets from interval-valued data," *IEEE International Conference on Fuzzy Systems*, pp. 10–15, 2012.

[15] C. Wagner and H. Hagras, "Employing zSlices Based General Type-2 Fuzzy Sets to Model Multi Level Agreement," *IEEE SSCI 2011: Symposium Series on Computational Intelligence - T2FUZZ 2011: 2011 IEEE Symposium on Advances in Type-2 Fuzzy Logic Systems*, pp. 50–57, 2011.

[16] ——, "Toward General Type-2 Fuzzy Logic Systems Based on zSlices," *Fuzzy Systems, IEEE Transactions on*, vol. 18, no. 4, pp. 637–660, 2010.

[17] D. C. Berry, D. K. Raynor, P. Knapp, and E. Bersellini, "Patients' understanding of risk associated with medication use: impact of European Commission guidelines and other risk scales." *Drug safety : an international journal of medical toxicology and drug experience*, vol. 26, no. 1, pp. 1–11, 2003.