Decision Style in a Clinical Reasoning Corpus

Limor Hochberg[1] **Cecilia O. Alm**[1] **Esa M. Rantanen**[1]
Caroline M. DeLong[1] **Anne Haake**[2]
1 College of Liberal Arts 2 College of Computing & Information Sciences
Rochester Institute of Technology
lxh6513|coagla|emrgsh|cmdgsh|anne.haake@rit.edu

Abstract

The dual process model (Evans, 2008) posits two types of decision-making, which may be ordered on a continuum from *intuitive* to *analytical* (Hammond, 1981). This work uses a dataset of narrated image-based clinical reasoning, collected from physicians as they diagnosed dermatological cases presented as images. Two annotators with training in cognitive psychology assigned each narrative a rating on a four-point decision scale, from intuitive to analytical. This work discusses the annotation study, and makes contributions for resource creation methodology and analysis in the clinical domain.

1 Introduction

Physicians make numerous diagnoses daily, and consequently clinical decision-making strategies are much discussed (e.g., Norman, 2009; Croskerry, 2003, 2009). Dual process theory proposes that decision-making may be broadly categorized as *intuitive* or *analytical* (Kahneman & Frederick, 2002; Stanovich & West, 2000). Further, scholars argue that decision-making may be ordered on a continuum, with intuitive and analytical at each pole (Hamm, 1988; Hammond, 1981).

Determining the decision strategies used by physicians is of interest because certain styles may be more appropriate for particular tasks (Hammond, 1981), and better suited for expert physicians rather than those in training (Norman, 2009). Language use can provide insight into physician decision style, as linguistic content reflects cognitive processes (Pennebaker & King, 1999).

While most clinical corpora focus on patients or conditions, physician diagnostic narratives have been successfully annotated for conceptual units (e.g., identifying medical morphology or a differential diagnosis), by Womack et al. (2013) and McCoy et al. (2012). Crowley et al. (2013) created an instructional system to detect cognitive biases in clinical decision-making, while Coderre et al. (2003) used protocol analysis on think-aloud diagnostic narratives, and found that features of intuitive reasoning implied diagnostic accuracy.

In this study, speech data were collected from physicians as they diagnosed dermatological cases presented to them as images. Physician verbalizations were annotated for decision style on a four-point scale from intuitive to analytical (Figure 1). Importantly, cognitive psychologists were brought into the loop for decision style annotation, to take advantage of their expertise in decision theory.

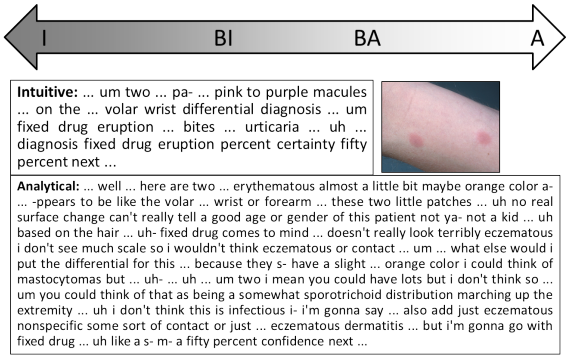

Figure 1: The decision-making continuum, showing the four-point rating scale. The example narratives were by two physicians for the same image (used with permission from Logical Images, Inc.), both correct in diagnosis. *(I=Intuitive, BI=Both-Intuitive, BA=Both-Analytical, A=Analytical).*

This work describes a thorough methodology applied in annotating a corpus of diagnostic narratives for decision style. The corpus is a unique resource – the first of its kind – for studying and modeling clinical decision style or for developing instructional systems for training clinicians to assess their reasoning processes.

This study attempts to capture empirically decision-making constructs that are much-

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 83–87,
Baltimore, Maryland USA, June 26-27 2014.

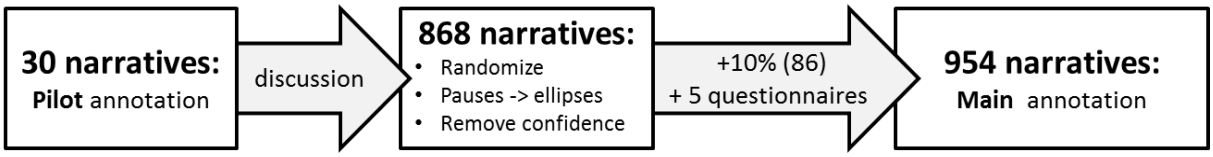

Figure 2: **Overview of annotation methodology.** Conclusions from the pilot study enhanced the main annotation study. To ensure high-quality annotation, narratives appeared in random order, and 10% (86) of narratives were duplicated and evenly distributed in the annotation data, to later assess intra-annotator reliability. Questionnaires were also interspersed at 5 equal intervals to study annotator strategy.

discussed theoretically. Thus, it responds to the need for investigating subjective natural language phenomena (Alm, 2011). The annotated corpus is a springboard for decision research in medicine, as well as other mission-critical domains in which good decisions save lives, time, and money.

Subjective computational modeling is particularly challenging because often, no real 'ground truth' is available. Decision style is such a *fuzzy* concept, lacking clear boundaries (Hampton, 1998), and its recognition develops in psychologists over time, via exposure to knowledge and practice in cognitive psychology. Interpreting fuzzy decision categories also depends on mental models which lack strong intersubjective agreement. This is the nature, and challenge, of capturing understandings that emerge organically.

This work's contributions include (1) presenting a distinct clinical resource, (2) introducing a robust method for fuzzy clinical annotation tasks, (3) analyzing the annotated data comprehensively, and (4) devising a new metric that links annotated behavior to clinicians' decision-making profiles.

2 Corpus Description

In an experimental data-collection setting, 29 physicians (18 residents, 11 attendings) narrated their diagnostic thought process while inspecting 30 clinical images of dermatological cases, for a total of 868[1] narratives. Physicians described observations, differential and final diagnoses, and confidence (out of 100%) in their final diagnosis. Later, narratives were assessed for correctness (based on final diagnoses), and image cases were evaluated for difficulty by a dermatologist.

3 Corpus Annotation of Decision Style

The corpus was annotated for decision style in a pilot study and then a main annotation study (Figure 2).[2] Two annotators with graduate training in cognitive psychology independently rated each narrative on a four-point scale from *intuitive* to *analytical* (Figure 1). The two middle labels reflect the presence of both styles, with intuitive (*BI*) or analytical (*BA*) reasoning being more prominent. Since analytical reasoning involves detailed examination of alternatives, annotators were asked to avoid using length as a proxy for decision style.

After the pilot, the annotators jointly discussed disagreements with one researcher. Inter-annotator reliability, measured by linear weighted kappa (Cohen, 1968), was 0.4 before and 0.8 after resolution; the latter score may be an upper bound on agreement for clinical decision-making annotation. As both annotators reported using physician-provided confidence to judge decision style, in subsequent annotation confidence mentions had been removed if they appeared after the final diagnosis (most narratives), or, if intermixed with diagnostic reasoning, replaced with dashes. Finally, silent pauses[3] were coded as ellipses to aid in the human parsing of the narratives.

4 Quantative Annotation Analysis

Table 1 shows the annotator rating distributions.[4]

	I	**BI**	**BA**	**A**
A1	89	314	340	124
A2	149	329	262	127

Table 1: The distribution of ratings across the 4-point decision scale. *I=Intuitive, BI=Both-Intuitive, BA=Both-Analytical, A=Analytical; A1=Annotator 1, A2=Annotator 2; N=867.*

Though Annotator 1's ratings skew slightly more analytical than Annotator 2, a Kolmogorov-

[1] Two physicians skipped 1 image during data collection.

[2] Within a reasonable time frame, the annotations will be made publicly available as part of a corpus release.

[3] Above around 0.3 seconds (see Lövgren & Doorn, 2005).

[4] N = 867 after excluding a narrative that, during annotation, was deemed too brief for decision style labeling.

Factor	A1 (Avg)	A1 (SD)	A2 (Avg)	A2 (SD)
Switching between decision styles	1.0	0.0	3.6	0.9
Timing of switch between decision styles	1.6	0.5	4.2	0.4
Silent pauses (...)	2.0	0.0	3.6	0.5
Filled pauses (e.g. *uh, um*)	2.0	0.7	3.6	0.5
Rel. (similarity) of final & differential diagnosis	2.8	0.4	3.2	0.8
Use of logical rules and inference	3.2	0.8	2.2	0.4
False starts (in speech)	3.4	0.9	2.4	0.9
Automatic vs. controlled processing	3.4	0.5	4.0	0.0
Holistic vs. sequential processing	3.6	0.5	4.4	0.5
No. of diagnoses in differential diagnoses	4.0	0.0	1.6	0.5
Word choice	4.0	0.7	2.6	0.5
Rel. (similarity) of final & first-mentioned diagnosis	4.0	0.0	4.0	0.0
Perceived attitude	4.0	0.7	4.0	0.0
Rel. timing of differential diagnosis in the narrative	4.2	0.8	2.8	0.8
Degree of associative (vs. linear, ordered) processing	4.2	0.4	3.8	0.4
Use of justification (e.g. *X because Y*)	4.2	0.4	4.0	0.0
Perceived confidence	4.4	0.5	4.2	0.4

Table 3: Annotators rated each of the listed factors as to how often they were used in annotation, on a 5-point Likert scale from *for no narratives* (1) to *for all narratives (5)*. (Some factors slightly reworded.)

Smirnov test showed no significant difference between the two distributions ($p = 0.77$).

	WK	%FA	%FA+ 1	*N*
A1 - A2	.43	50%	94%	867
A1 - A1	.64	67%	100%	86
A2 - A2	.43	50%	95%	86

Table 2: Inter- and intra-annotator reliability, measured by linear weighted kappa (WK), percent full agreement (%FA); and full plus within 1-point agreement (%FA+1). Intra-annotator reliability was calculated for the narratives rated twice, and inter-annotator reliability on the initial ratings.

As shown in Table 2, reliability was moderate to good (Altman, 1991), and inter-annotator agreement was well above chance (25%). Indeed, annotators were in full agreement, or agreed within one rating on the continuum, on over 90% of narratives. This pattern reveals fuzzy category boundaries but sufficient regularity so as to be measurable. This is in line with subjective natural language phenomena, and may be a consequence of imposing discrete categories on a continuum.[5] Annotator 1 had better intra-annotator reliability, perhaps due to differences in annotation strategy.

[5]Nonetheless, affect research has shown that scalar representations are not immune to variation issues (Alm, 2009).

5 Annotator Strategy Analysis

Five questionnaires evenly spaced among the narratives asked annotators to rate how often they used various factors in judging decision style (Table 3). Factors were chosen based on discussion with the annotators after the pilot, and referred to in descriptions of decision styles in the annotator instructions; the descriptions were based on characteristics of each style in the cognitive psychology literature (e.g., Evans, 2008). Factors with high variability (SD columns in Table 3) reveal changes in annotator strategy over time, and factors that may influence intra-annotator reliability.

Both annotators reported using the *rel. (similarity) of final & first-mentioned diagnosis*, as well as *perceived attitude*, *perceived confidence*, and *use of justification*, to rate most narratives. Types of *processing* were used by both sometimes; this is important since these are central to the definitions of decision style in decision-making theory.

Differences in strategies allow for the assessment of annotators' individual preferences. Annotator 1 often considered the *no. of diagnoses in the differential*, and *rel. timing of the differential*, but Annotator 2 rarely attended to them; the opposite pattern occurred with respect to *switching between decision styles*, and the *timing of the switch*.

The shared high factors reveal those consistently linked to interpreting decision style, despite

the concept's fuzzy boundaries. In contrast, the idiosyncratic high factors reveal starting points for understanding fuzzy perception, and for further calibrating inter-annotator reliability.

6 Narrative Case Study

Examining particular narratives is also instructive. Of the 86 duplicated narratives with two ratings per annotator, *extreme agreement* occurred for 22 cases (26%), meaning that all four ratings were exactly the same.[6] Figure 3 (top) shows such a case of intuitive reasoning: a quick decision without reflection or discussion of the differential. Figure 3 (middle) shows a case of analytical reasoning: consideration of alternatives and logical inference.

Agr (I): ... there's a ... brown papule with telangiectasias on the ... nasal tip ... uh the differential includes a pigmented basal cell melanoma ... nevus ... and the diagnosis is melanoma **(diagnosis incorrect)**
Agr (A): ... okay so a large ... purple ... uh ... mass ... on a face ... no ... it's on the foot ... or the ... yeah ... um ... yeah it would **depend** a lot on how well it blanches ... you want- wanna ... feel that ... um ... could be just a ... hemangioma ... could be a ... basal cell skin cancer could be a melanoma ... um uh might be one of those things you wanna ... toughen the uh ... the edge of it ... has a bit of a ... pearly look to it but i don't know if that's just again from ... being on a foot ... and uh ... and having more uh ... hydrostatic pressures there ... um -**cause** it's mostly ... uh purple ... it could be a you know angiosarcoma um but it's a little on the small side ... um ... you know common things being common go with the uh ... hemangioma ... as the number one thought ... with uh ... m- basal cell skin cancer being the second hemangioma again **(diagnosis incorrect)**
DisAgr (A, I): ... uh uh ... think we're on a foot you see some scale at the bottom makes me think there's little fungus there but ... looks like the thing that they took the picture of is a purple irregular tumor ... um ... has very ill-distinct borders with surrounding red areas ... **it's so purple it makes me think of a vascular tumor ... so i think kaposi's sarcoma is most likely** ... could be a melanoma ... could be a metastatic renal cell tumor ... my best guess is that this is kaposi's sarcoma **(diagnosis incorrect)**

Figure 3: Narratives for which annotators were in *full agreement* on *I* (top) and *A* (middle) ratings, vs. in *extreme disagreement* (bottom).

In the full data set (initial ratings), there were 50 cases (6%) of 2-point inter-annotator disagreement and one case of 3-point inter-annotator disagreement (Figure 3, bottom). This latter narrative was produced by an attending (experienced physician), 40% confident and incorrect in the final diagnosis. Annotator 1 rated it analytical, while Annotator 2 rated it intuitive. This is in line with Annotator 1's preference for analytical ratings (Table 1). Annotator 1 may have viewed this pattern of *observation* → *conclusion* as logical reasoning, characteristic of analytical reasoning. Annotator 2 may instead have interpreted the phrase *it's so purple it makes me think of a vascular tumor...so i think [...]* as intuitive, due to the *makes me think* comment, indicating associative reasoning, characteristic of intuitive thinking. This inter-annotator contrast may reflect Annotator 1's greater reported use of the factor *logical rules and inference* (Table 3).

[6]There were no cases where all four labels differed, further emphasizing the phenomenon's underlying regularity.

7 Physician Profiles of Decision Style

Annotations were also used to characterize physicians' preferred decision style. A decision score was calculated for each physician as follows:

$$d_p = \frac{1}{2n}\sum_{i=1}^{n}(r_{A1_i} + r_{A2_i}) \quad (1)$$

where p is a physician, r is a rating, n is total images, and $A1, A2$ the annotators. Annotators' initial ratings were summed – from 1 for *Intuitive* to 4 for *Analytical* – for all image cases for each physician, and divided by 2 times the number of images, to normalize the score to a 4-point scale. Figure 4 shows the distribution of decision scores across *residents* and experienced *attendings*.

Residents exhibit greater variability in decision style. While this might reflect that residents were the majority group, it suggests that differences in expertise are linked to decision styles; such differences hint at the potential benefits that could come from preparing clinical trainees to self-monitor their use of decision style. Interestingly, the overall distribution is skewed, with a slight preference for analytical decision-making, and especially so for attendings. This deserves future attention.

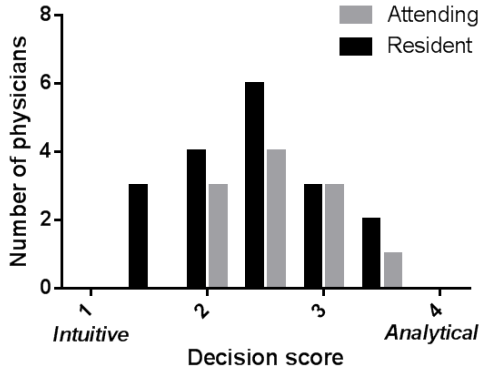

Figure 4: Decision score distribution by expertise.

8 Conclusion

This study exploited two layers of expertise: physicians produced diagnostic narratives, and trained cognitive psychologists annotated for decision style. This work also highlights the importance of understanding annotator strategy, and factors influencing annotation, when fuzzy categories are involved. Future work will examine the links between decision style, expertise, and diagnostic accuracy or difficulty.

Acknowledgements

Work supported by a CLA Faculty Dev. grant, Xerox award, and NIH award R21 LM01002901. Many thanks to annotators and reviewers.

This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Alm, C. O. (2009). *Affect in text and speech.* Saarbrücken: VDM Verlag.

Alm, C. O. (2011, June). Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 107-112). Association for Computational Linguistics.

Altman, D. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education, 37*(8), 695-703.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., ... & Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education, 18*(3), 343-363.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*(8), 775-780.

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022-1028.

Evans, J. (2008). Dual-processing accounts of reasoning, judgement and social cognition. *Annual Review of Psychology, 59*, 255-278.

Hamm, R. M. (1988). Clinical intuition and clinical analysis: Expertise and the cognitive continuum. In J. Dowie & A.S. Elstein (Eds.), *Professional judgment: A reader in clinical decision making* (pp. 78-105). Cambridge, England: Cambridge University Press.

Hammond, K. R. (1981). *Principles of organization in intuitive and analytical cognition (Report #231).* Boulder, CO: University of Colorado, Center for Research on Judgment & Policy.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*(2), 137-165.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics of intuitive judgment: Extensions and applications* (pp. 49-81). New York, NY: Cambridge University Press.

Lövgren, T., & Doorn, J. V. (2005). Influence of manipulation of short silent pause duration on speech fluency. In *Proceedings of Disfluency in Spontaneous Speech Workshop* (pp. 123-126). International Speech Communication Association.

McCoy, W., Alm, C. O., Calvelli, C., Li, R., Pelz, J. B., Shi, P., & Haake, A. (2012, July). Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the 6th Linguistic Annotation Workshop* (pp. 95-103). Association for Computational Linguistics.

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*(1), 37-49.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645-665.

Womack, K., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., and Haake, A. (2013, August). Using linguistic analysis to characterize conceptual units of thought in spoken medical narratives. In *Proceedings of Interspeech 2013* (pp. 3722-3726). International Speech Communication Association.