Title

Parallel Belief Updating in Sequential Diagnostic Reasoning

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36)

ISSN
1069-7977

Authors
Jahn, Georg
Stahnke, Rebekka
Rebitschek, Felix G.

Publication Date
2014

Peer reviewed

Parallel Belief Updating in Sequential Diagnostic Reasoning

Georg Jahn (georg.jahn@uni-greifswald.de)
University of Greifswald, Department of Psychology
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

Rebekka Stahnke (rebekka.stahnke@staff.hu-berlin.de)
Humboldt-Universität zu Berlin, Department of Education Studies
Geschwister-Scholl-Str. 7, D-10117 Berlin, Germany

Felix G. Rebitschek (felix.rebitschek@uni-greifswald.de)
University of Greifswald, Department of Psychology
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

Abstract

In sequential diagnostic reasoning the goal is to determine the most likely cause for a number of sequentially observed effects. Potential hypotheses are narrowed down by integrating the cumulating observed evidence leading to the selection of one among several hypotheses. In the reported diagnostic reasoning experiment, thirty-eight participants were tested with quasi-medical problems consisting of four sequentially presented symptoms with four candidate diagnostic hypotheses. We used ambiguous sequences that could be equally caused by two chemicals to investigate possible order effects and explicitly highlighted alternative hypotheses by using a stepwise rating procedure that also enabled us to compare participants' ratings with belief updating in a Bayes net. Even though alternatives were explicitly highlighted, participants were biased towards the initial hypothesis in a pair of equally supported hypotheses. We conclude that ambiguous symptom sets and non-diagnostic symptoms invite biased symptom processing and can produce primacy effects even in a step-by-step procedure.

Keywords: Diagnostic reasoning, Belief updating, Probabilistic inference, Order effects

Introduction

Diagnostic reasoning is a case of information integration. The task is to infer the most likely cause of observed symptoms. Often in medical diagnosis, the symptoms are probabilistic cues to their possible causes and do not suggest just a single diagnosis. Instead, symptoms usually have several possible causes and trigger the generation of multiple diagnostic hypotheses that are tested and updated during subsequent symptom processing (Thomas, Dougherty, Sprenger, & Harbison, 2008; Weber, Böckenholt, Hilton, & Wallace, 1993). The final diagnosis is the result of integrating symptom information. In the reported experiment, we studied the parallel updating of multiple diagnostic hypotheses during the processing of symptoms that each supported more than one diagnostic hypothesis. Symptom sequences that finally support two diagnoses equally should result in equal proportions of final diagnoses according to the normative standard of Bayesian belief updating. By collecting continuous belief ratings, we traced deviations from Bayesian updating and found evidence for symptom processing biased towards the leading hypothesis even in task conditions that are considered to induce no bias or an opposite bias (Hogarth & Einhorn, 1992).

The initial hypothesis or the set of initial hypotheses triggered by early symptoms can bias the processing of subsequent symptoms (Hagmayer & Kostopoulou, 2013; Jahn & Braatz, 2014; Kostopoulou, Russo, Keenan, Delaney, & Douiri, 2012; Rebitschek, Scholz, Bocklisch, Krems, & Jahn, 2012). The support that later encountered symptoms provide for the focal hypothesis is emphasized and their support for alternative hypotheses is considered less than would be appropriate. Such biased symptom processing favors the hypothesis that is strongly supported by early symptoms and consequently strengthens the weight of early symptoms. A strong weight of early symptoms constitutes a primacy effect.

Primacy effects have been observed in diagnostic reasoning with ambiguous symptom sequences. However, in a procedure that requires step-by-step belief ratings, there are reasons to expect unbiased integration or a recency effect rather than a primacy effect (Catena, Maldonado, Megías, & Freese, 2002; Hogarth & Einhorn, 1992; Rebitschek et al., 2012). The procedure of step-by-step belief ratings prompts ratings of the current status of diagnoses after each symptom presentation. Thus, participants are reminded of alternative diagnoses after each symptom. Second, the ratings prolong the retention interval for earlier symptoms and may interfere with the rehearsal of earlier symptoms. Consequently, the memory representation of later symptoms could be stronger and the relative weight of later symptoms could increase. Finally, with step-by-step belief ratings symptom integration cannot be delayed. An intermediate integration takes place after each symptom and the current status of diagnostic hypotheses could function as an anchor (Catena et al., 2002). The influence of a late symptom in adjusting an anchor could be stronger than the symptom's contribution when it is part of a larger set of symptoms that are integrated.

To summarize, these reasons to expect unbiased integration or recency effects – the saliency of alternatives, memory dynamics favoring late evidence, and contrast effects in anchoring and adjustment – postulate processes counteracting a known tendency to bias symptom processing towards the initially leading hypothesis. We used a quasi-medical diagnostic reasoning task (Meder & Mayrhofer, 2013; Mehlhorn, Taatgen, Lebiere, & Krems, 2011), with which a bias favoring the leading hypothesis had been demonstrated several times before (Jahn & Braatz, 2014; Rebitschek et al., 2012), and tested whether step-by-step belief ratings could overcome this bias.

Experiment

Participants were put in the role of a physician diagnosing which chemical had affected patients presenting with certain symptoms. First, they learned about four chemicals and the symptom categories that each could cause (Table 1 and Table 2). Then, they worked through a series of diagnostic reasoning items consisting of four symptoms each. There were non-diagnostic symptoms (x-symptoms) and symptoms that could be caused by two chemicals but with different causal strengths. Symptoms strongly suggesting one and weakly suggesting another chemical are denoted Ab (strongly suggesting A and weakly suggesting B) and Ba (strongly suggesting B and weakly suggesting A). Sequences with equal support for two chemicals are listed as item type AB in Table 3.

After each symptom, participants rated the current probability of each chemical as the cause of the symptoms seen so far. These step-by-step belief ratings are compared with posterior probabilities computed in a Bayes net and can indicate biased symptom processing. Proportions of final diagnoses indicate biased symptom processing if they deviate from .5 for sequences with equal support.

Method

Participants. Thirty-eight students of the University of Greifswald (21 female, 17 male) with a mean age of 23.2 years (*SD* = 3.2) took part in the experiment and were included in the analysis. Of eight additional participants, six did not complete the experiment and two produced disproportionately many errors (36% and 53% diagnoses that were not supported by any diagnostic symptom).

Materials. In preparation for the diagnostic reasoning task, participants learned about four chemicals and the symptoms that each chemical could cause. There were six symptom classes each containing two symptoms that are listed in Table 1. We used symptom classes encompassing symptoms to limit the complexity of the causal structure to be learned while still ensuring a sufficient variety of symptom sequences to be constructed from symptoms.

The strength with which a chemical caused symptoms from a certain class was either strong or weak. These levels of causal strength were communicated to participants as relative frequencies in verbal and pictorial form. For example, weak symptoms were presented as caused in “3 out of 10 patients”. This relative frequency was additionally visualized by a row of stick-figures illustrating how many of 10 patients being affected by the respective chemical show symptoms from the respective class: 3 red and 7 black.

Table 1: Symptom classes and symptoms

Symptom Class	Symptoms	
Eyes	Eyelid swelling	Lacrimation
Respiration	Cough	Difficult breathing
Skin	Acid burn	Rash
Neurological	Paralysis	Speech disorder
Circulatory Pr.	Sweating	Swoon
Pain	Twinge	Sting

Note. Original materials were in German.

Each chemical had one strong and three weak symptom classes (see Table 2). These were presented in separate rows on a screen during the learning phase. For example, such a screen for the R chemical read: The chemical R is gasiform. It causes eyes-symptoms in 9 out of ten patients. <9 red stick figures, 1 black stick figure>. It causes respiration-symptoms in 3 out of 10 patients. <3 red, 7 black>. It causes circulatory problems in 3 out of 10 patients. <3 red, 7 black> It causes pain-symptoms in 3 out of 10 patients. <3 red, 7 black>

As apparent in Table 2, *circulatory problems* and *pain* were non-diagnostic symptom classes. Symptoms from these classes are denoted “x” in the following. The remaining four symptom classes were each caused strongly by one and weakly by a second chemical (columns 3 and 4 in Table 2). For example, *skin*-symptoms were strongly caused by the W-chemical, but only weakly by the K-chemical. Such symptoms are denoted “Ab” (strong for A, weak for B) or “Ba” (strong for B, weak for A) in the following.

A single diagnostic reasoning item consisted of a sequence of four symptoms, for example: *acid burn*, *paralysis*, *swoon*, and *speech disorder* (Ab_Ba_x_Ba). This sequence belongs to the ABB item type because it contains one Ab-symptom and two Ba-symptoms. Table 3 shows the three item types (AAB, AB, and ABB) that each comprised three symptom sequences.

Table 2: The chemicals and the symptom classes that each could cause

Chemical	Group	In 9 out of 10 patients	In 3 out of 10 patients	In 3 out of 10 patients	In 3 out of 10 patients
R	Gas	Eyes	Respiration	Circulatory Pr.	Pain
B	Gas	Respiration	Eyes	Circulatory Pr.	Pain
W	Fluid	Skin	Neurological	Circulatory Pr.	Pain
K	Fluid	Neurological	Skin	Circulatory Pr.	Pain

Note. Original materials were in German.

The symptom sequences in Table 3 were used with each of the chemicals in the A-role and the remaining chemical from the same group in the B-role. All possible assignments of symptoms to item types were constructed with the restriction that no single symptom was repeated in a symptom sequence.

Table 3: Item types and symptom sequences

Item type	Symptom sequence
AAB	Ab_x_Ab_Ba
	Ab_Ab_x_Ba
	Ab_Ab_Ba_x
AB	Ab_Ba_x_x
	Ab_x_Ba_x
	Ab_x_x_Ba
ABB	Ab_x_Ba_Ba
	Ab_Ba_x_Ba
	Ab_Ba_Ba_x

Bayesian posterior probabilities. For comparing the sequential belief ratings with normative reference values, the causal structure of the scenario was implemented in a Bayes net. The causal model (Figure 1) reflects the structure presented in Table 2. The chemicals as candidates for the unknown root cause were defined as mutually exclusive. The four potential states of the unknown root cause spread to the diagnostic and non-diagnostic symptom classes. The symptom classes as the effects were mutually independent but not mutually exclusive.

The node of the root cause was modeled with four states corresponding to the four chemicals R, B, W, K. The prior probabilities of the chemicals (states of the root cause) were set as equal and the probabilities of the symptoms' presences given the different chemicals were fixed as depicted in the boxes in Figure 1. Under the specific parameterization, the posterior probabilities take on values of 0, .25, .5, and .75 (Figure 3).

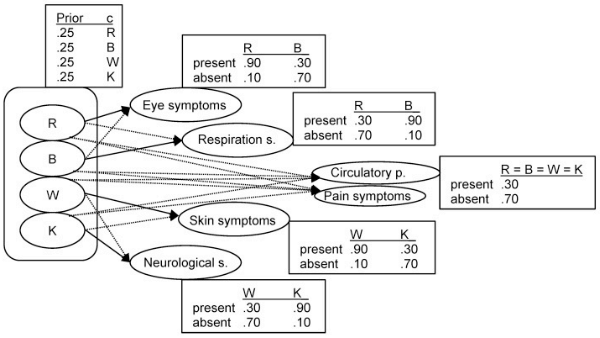

Figure 1. Bayesian causal model including the states of the root cause, the diagnostic and non-diagnostic effects (symptom classes), and respective parameter settings.

Procedure. At the beginning of the learning phase, participants were instructed that their task would be to determine the cause of a patient's symptoms. They were told that the patients are workers in a chemical plant that processes four chemicals. Each patient was affected by exactly one of those chemicals. Participants should determine which chemical most likely had caused a patient's symptoms.

First, they studied a screen explaining which symptoms belong to which symptom class (Table 1) and worked through test trials until the set of twelve symptoms was once assigned to symptom classes without errors. Then, participants were told that each chemical caused one of the six symptom classes almost always and a second symptom class occasionally. They were further told that two symptom classes are caused occasionally by all of the chemicals.

Next, the chemicals R and B were studied on separate screens listing the symptom classes and their respective frequencies verbally and pictorially. Participants proceeded to testing when they felt ready.

In each test trial of the learning procedure, a symptom class was presented together with a frequency (e.g. "Pain in 3 out of 10 patients") and participants responded with the letter of the chemical that causes this symptom with this frequency or with the letter "a" for all chemicals. All pairings of symptom classes and frequencies were tested in random order and the whole set was tested until it was once answered without errors. Then, the screens for the chemicals W and K were studied and tested and finally, all four chemicals were restudied and all symptom classes with frequency pairings were tested in random order until the test was completed without errors. Learning was completed within 16.4 min on average (*SD* = 4.5).

Diagnostic reasoning. In each diagnostic reasoning trial, a sequence of four symptoms was presented. Each symptom presentation consisted of a fixation cross shown for 1s followed by a symptom that remained visible for 2s. Then, probability ratings were collected for all four chemicals on separate screens in random order. Each screen asked to enter a number between 0 and 100 to indicate in how many of 100 patients presenting the symptoms seen so far the respective chemical would be the correct diagnosis. Participants entered a number and hit return to proceed to the next screen. Editing with backspace was possible and only numbers between 0 and 100 were accepted. When the probability rating for the fourth chemical had been completed, the presentation of the next symptom started with a fixation cross. After the ratings for the fourth symptom, participants indicated their final diagnosis with the respective letter and rated their confidence for the diagnosis with number keys from 1 (very unsure) to 7 (very sure).

The first four trials were training trials and the very first trial was performed under supervision of the experimenter who ensured and explained that the ratings after each symptom should sum to 100 and that all symptoms seen so far should be considered.

After the training trials, each participant worked through the 36 possible combinations of chemicals with symptom sequences. The order of the 36 trials was pseudo-random and balanced across participants. For each trial, the actual sequence of symptoms was drawn randomly from the possible symptom assignments for this combination of symptom sequence and chemical in the A-role.

After half of the trials, participants were encouraged to pause for a couple of minutes. The whole experiment took 60 to 90 min in total.

Results

Trials that were responded to with a chemical that was not supported by any of the diagnostic symptoms (C- or D-diagnoses) were not included in the following analyses (1.7% of all trials). Furthermore, trials were dropped, in which the likelihood ratings after one of the four symptoms did sum to less than 85% or to more than 115% (5.1% of all trials with A- or B-diagnoses).

Diagnoses. The mean proportion of A-diagnoses for each symptom sequence is shown in Figure 2. Sequences of the AAB item type were mostly responded to with the A-chemical and sequences of the ABB item type were mostly responded to with the B-chemical in line with the relative support for A and B. Sequences of the AB item type revealed a primacy effect in diagnoses: A-proportions were higher than .5 for AB-items, $t(37) = 3.89$, $p < .001$, $d = 0.63$, and higher if non-diagnostic symptoms delayed the Ba-symptom in Ab_x_Ba_x and Ab_x_x_Ba sequences with *d*s of 0.52 and 0.59, respectively, than if the Ba-symptom immediately followed the Ab-symptom in the Ab_Ba_x_x sequence ($d = 0.26$).

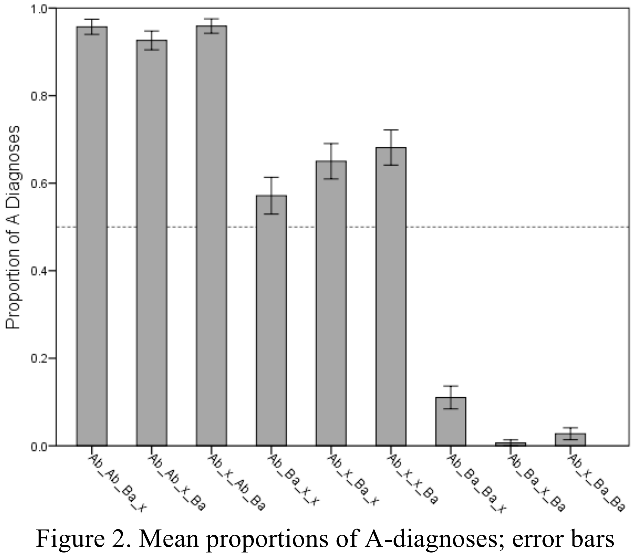

Figure 2. Mean proportions of A-diagnoses; error bars indicate standard errors.

Sequential likelihood ratings. Figure 3 shows means of the likelihood ratings for A- and B-chemicals, and for C- and D-chemicals after each symptom for symptom sequences of the AB item type (Ab_Ba_x_x, Ab_x_Ba_x, and Ab_x_x_Ba) plotted separately for trials answered with A (A Diagnosis) and B (B Diagnosis). Bayesian posterior probabilities are shown for comparison.

As visible in Figure 3, mean ratings after the first symptom match well with the Bayesian posterior probabilities in both trials with final A- and trials with final B-diagnoses. Right before the final diagnosis after the fourth symptom, the rating for the chemical that was subsequently chosen as the final diagnosis was generally higher than the rating for the competing alternative. Thus, final diagnoses were consistent with the last ratings.

The mean A-ratings after x-symptoms for trials answered with B (right column in Figure 3) are lower than for trials answered with A (left column) and lower than the respective Bayesian probabilities. The decrease of A-ratings after x-symptoms that occurred before a Ba-symptom in trials answered with B shows that participants did not process x-symptoms as non-diagnostic. Instead and particularly in trials with final B-diagnoses, x-symptoms increased the ratings for alternatives to A (for B, but also ratings for C and D). This shift to alternatives after x-symptoms that was more pronounced in trials with a final B-diagnosis is clearly apparent in the mean sums of C- and D-ratings listed in Table 4.

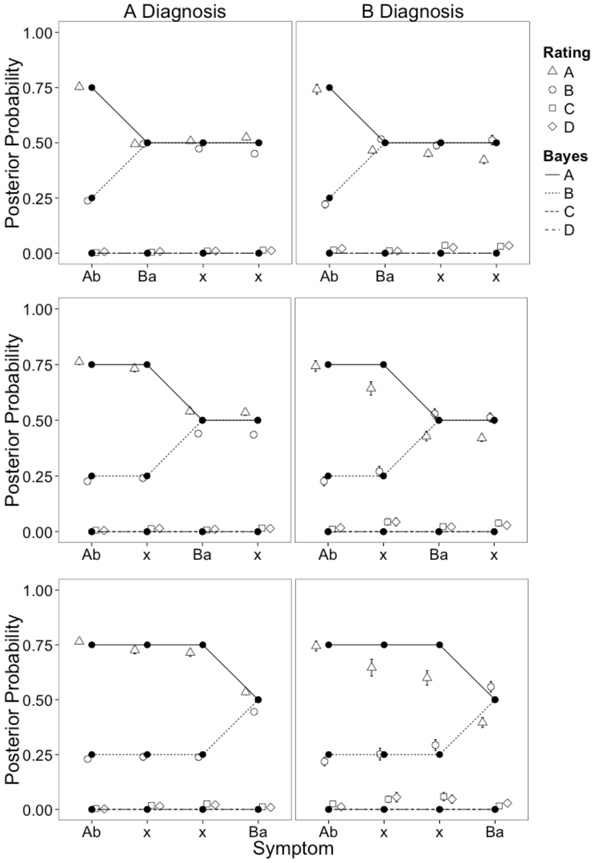

Figure 3. Mean likelihood ratings of A-, B-, C-, and D-diagnoses for the three sequences of the AB-item type with standard errors along with the posterior probabilities computed with the Bayesian causal model separately for trials finally answered with A (column A Diagnosis) and trials finally answered with B (column B Diagnosis).

Table 4: Mean sums of C- and D-ratings after each symptom for the three sequences of the AB-item type separately for trials finally answered with A (A-Diagnosis) and trials finally answered with B (B-Diagnosis)

	A-Diagnosis		B-Diagnosis	
Order/	C+D-Rating	*N*	C+D-Rating	*N*
Symptom	*M (SE)*		*M (SE)*	
Ab	0.88 (0.65)	80	3.50 (1.48)	60
Ba	1.13 (0.64)		2.00 (0.91)	
x	1.98 (0.74)		6.13 (1.93)	
x	2.55 (0.94)		6.53 (1.88)	
Ab	1.10 (0.69)	91	2.86 (1.30)	49
x	2.81 (0.92)		8.82 (2.80)	
Ba	1.80 (0.73)		4.43 (1.69)	
x	2.97 (1.05)		6.69 (1.95)	
Ab	0.60 (0.33)	92	3.72 (1.75)	43
x	3.28 (0.95)		10.19 (3.43)	
x	4.64 (1.14)		10.51 (3.25)	
Ba	2.09 (0.67)		4.40 (1.52)	

The decrease of A-ratings after x-symptoms in trials with B-diagnoses further suggests that a tendency towards the final response developed rather early in a trial. To quantify the dependence of final diagnoses on early x-symptom processing, we computed the difference between A- and B-ratings after each symptom and tested with logistic regressions how well the AB-differences predicted the final response. The results of the logistic regressions are shown in Table 5. Note that the unit for the AB-difference was set to 10 rating points and that the clustering of trials at the level of participants was not modeled in the reported regressions.

The regression weights for the AB-difference increase across the four symptoms for all three AB-sequences. For the Ab_Ba_x_x and the Ab_x_Ba_x sequences, the prediction weights increase earlier than for the Ab_x_x_Ba sequence confirming that how Ba was processed was important for the final diagnosis. The changes in regression weights additionally confirm that the processing of non-diagnostic symptoms influenced the final diagnosis.

Discussion

Symptom sequences that contained somewhat diagnostic symptoms and non-diagnostic symptoms and that equally supported two competing diagnostic hypotheses induced symptom processing that more often favored the initially leading hypothesis. This bias towards the leading hypothesis occurred although step-by-step belief ratings highlighted alternatives and could have strengthened the weight of a later symptom supporting the competing alternative (Catena et al., 2002; Hogarth & Einhorn, 1992).

The ambiguous symptom sequences are particularly sensitive to biased symptom processing because each symptom is consistent with the favored diagnosis and can be interpreted as supporting it. The belief ratings suggest that participants indeed interpreted somewhat diagnostic symptoms that were consistent with two diagnostic hypotheses in support of the currently favored hypothesis.

Non-diagnostic symptoms increased ratings of unsupported alternatives (C and D), but less so in the more frequent trials, in which participants stayed with the initially leading hypothesis (see Figure 3 and Table 4) suggesting that non-diagnostic symptoms were rather interpreted as supporting the leading hypothesis than alternatives. Normatively, any change in ratings after non-diagnostic x-symptoms is unjustified. Yet, the attenuating effect (dilution) of non-diagnostic evidence is common (Nisbett, Zukier, & Lemley, 1981). In the present experiment, favoring the leading hypothesis resulted in a smaller dilution effect by non-diagnostic symptoms.

Missed non-diagnosticity (pseudodiagnosticity) is a known phenomenon in human diagnostic reasoning and is usually explained with missed alternative possible causes (Fischhoff & Beyth-Marom, 1983; Tversky & Koehler, 1994). In the present study, however, the repeated prompts to rate all candidate causes prevented that possible causes could be missed.

Table 5: AB-difference in ratings after each symptom as predictor of the final response (A vs. B) in sequences of the AB item type. Results of logistic regressions with the unit of the AB-difference set to 10 rating points (10%)

Order/ Symptom	Intercept	exp(β)	[95% CI]	Chi^2(1) [a]	*p*	*R²* [b]	*N*
Ab	.33	0.99	[0.86; 1.14]	0.01	.91	< .001	140
Ba	.36	1.36	[1.02; 1.82]	5.22	.02	.05	
x	.27	1.81	[1.03; 3.19]	10.24	.001	.10	
x	.30	1.85	[1.17; 2.91]	21.31	< .001	.19	
Ab	.46	1.03	[0.90; 1.18]	0.19	.66	.002	140
x	-.03	1.16	[1.02; 1.32]	5.57	.02	.05	
Ba	.59	1.52	[1.21; 1.91]	21.82	< .001	.20	
x	.63	1.83	[1.30; 2.58]	26.81	< .001	.24	
Ab	.66	1.02	[0.86; 1.20]	0.05	.83	.001	135
x	.36	1.10	[1.00; 1.23]	2.45	.12	.03	
x	-.02	1.22	[1.07; 1.39]	9.57	.002	.10	
Ba	.86	1.72	[1.33; 2.22]	32.06	< .001	.30	

Note. [a] Likelihood ratio test. [b] Nagelkerke's R^2

The non-diagnostic symptoms were linked to supported and to unsupported alternatives. Thus, they could be interpreted as caused by the leading hypothesis and could be taken as confirming the leading hypothesis. Presumably, such a confirmation of the leading hypothesis by non-diagnostic symptoms resulted in a stronger primacy order effect (higher A-proportion in Figure 2) in the AB-sequences, in which the Ba-symptom was preceded by x-symptoms. For non-ambiguous ABB-sequences, such an effect of a preceding x-symptom presumably was annihilated by a recency effect of the second Ba-symptom in the final position.

The observed biased symptom processing of somewhat diagnostic and non-diagnostic evidence is consistent with theories postulating biased information sampling (Busemeyer & Townsend, 1993) and with theories of biased information interpretation in the construction of a coherent representation (Hagmayer & Kostopoulou, 2013; Thagard, 1989). Reviewing the symptoms for evaluating the status of alternatives can be seen as information sampling in working memory and for such sampling a bias towards earlier presented information as well as a bias towards information supporting the leading alternative is deemed possible (Busemeyer & Townsend, 1993).

In biased information interpretation, the information value of a piece of evidence is not fixed but can be modified by stressing certain aspects to attain a better fit with an overall interpretation (Kostopoulou et al., 2012; Thagard, 1989). Such biased interpretation is particularly easy with ambiguous evidence and thus, a general tendency towards coherent representations could well be the reason for the observed bias towards the initially leading hypothesis.

Our results are consistent with recently reported biased symptom processing in very similar tasks without step-by-step belief ratings (Jahn & Braatz, 2014; Rebitschek et al., 2012). Sequential belief ratings are a quite obtrusive method for process tracing. It is remarkable that symptom processing biased towards the leading diagnostic hypothesis was nonetheless confirmed. In more realistic diagnostic tasks, perfectly ambiguous symptom patterns are unlikely and if information search is possible, uncertainty will motivate for continued search. If, however, ambiguity is strong and cannot be overcome, biased symptom processing seems likely.

Acknowledgments

This research was supported by German Research Foundation (DFG) Grant JA 1761/7-1.

References

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432-459.

Catena, A., Maldonado, A., Megías, J. L., & Freese, B. (2002). Judgment frequency, belief revision, and serial processing of causal information. *The Quarterly Journal of Experimental Psychology, 55B*, 267-281.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*, 239-260.

Hagmayer, Y., & Kostopoulou, O. (2013). A probabilistic constraint satisfaction model of information distortion in diagnostic reasoning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 531-136). Austin, TX: Cognitive Science Society.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Jahn, G., & Braatz, J. (2014). Memory indexing of sequential symptom processing in diagnostic reasoning. *Cognitive Psychology, 68*, 59-97.

Kostopoulou, O., Russo, J. E., Keenan, G., Delaney, B. C., & Douiri, A. (2012). Information distortion in physicians' diagnostic judgments. *Medical Decision Making, 32*(6), 831-839.

Meder, B., & Mayrhofer, R. (2013). Sequential diagnostic reasoning with verbal information. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1014-1019). Austin, TX: Cognitive Science Society.

Mehlhorn, K., Taatgen, N. A., Lebiere, C., Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *37*, 1391-1411.

Nisbett, R., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13*, 248-277.

Rebitschek, F. G., Scholz, A., Bocklisch, F., Krems, J. F., & Jahn, G. (2012). Order effects in diagnostic reasoning with four candidate hypotheses. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 905-910). Austin, TX: Cognitive Science Society.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-467.

Thomas, R. P., Dougherty, M. R., Sprenger, A., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155-185.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. Psychological Review, *101*, 547-567.

Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1151–1164.