Dealing with brittleness in the design of expert systems for immunohematology

S.A. Guerlain, P.J. Smith, J.H. Ohradowich, S. Rudmann, P. Storm, J.W. Smith, and J. Surbely

In recent years, there has been increased discussion about the potential of expert systems to support medical decision-making tasks, including applications in clinical laboratory settings. This study provides data regarding the cognitive errors that technologists make on an important problem-solving task: the identification of antibodies in a patient’s blood. It explores alternative designs for expert systems developed to reduce such errors. It also evaluates the effects of these alternative designs on the ability of the user to effectively stay “in the loop,” applying their own expertise and judgment while using the computer as a tool to assist with their analyses. A pilot study was conducted involving 32 certified medical technologists, which compared two alternative roles for the computer: (1) use of the computer to automatically complete subtasks upon request, and (2) use of the computer as a monitoring device to critique technologists as they completed the analyses themselves. The system design that automatically completed subtasks for the technologist induced a 29 percent increase in errors relative to the design that critiqued technologists as they completed the analyses themselves. Immunohematology 1996:12:101-107.

There are a number of potential causes of human error when attempting complex problem-solving tasks. Slips,¹ cognitive biases,² fallible heuristic strategies,³ misconceptions, and lack of knowledge can all lead to incorrect solutions.

Past studies have shown that the practice of antibody identification is susceptible to all of these sources of error.⁴⁻⁶ The challenge, however, is finding effective solutions. One approach is improved education.⁷ A second approach is administrative help, such as having an expert review all antibody identifications.⁴ A third approach is to provide technologists with assistance from a computerized decisionsupport system.

Although all three of these approaches merit serious consideration, this paper focuses on the third—the design of “intelligent” tools to assist in supporting technologists in the laboratory.

System Design Issues

From a human factor’s perspective, there is clear evidence that the role the computer plays in supporting a person’s decision making can have a major impact on the cognitive processes of that person, and consequently, on the errors that person is likely to make.³¹ The degree to which the person is “in the loop” during the initial hypothesis generation and testing activities can establish the degree of reliance on the computer. A computer can also influence a person’s cognitive processes, changing that person’s focus when looking at the raw data and influencing that person’s interpretation of the data. Once the computer has erroneously suggested, for example, that the answer is “anti-E,” the technologist can no longer “see” the data as he or she would have before receiving the computer’s conclusion. Cognitive biases as well as attention, memory, and perceptual processes are all influenced by the computer’s suggested answer.²

One important design decision in developing an expert system to assist in the laboratory is the selection of an effective role for that system. Given the same knowledge base or set of expertise, there are several alternative ways to design the role of such a system. At one extreme is complete automation. In the middle is an expert consulting system, which automatically generates an answer for review by the human practitioner, who then decides how to proceed. At the other extreme is a critiquing system in which the human practitioner completes the analysis and records intermediate and final answers in the computer. The computer in essence “looks over the shoulder” of the practitioner and critiques these intermediate and final answers, cautioning the person when it detects suspicious intermediate and final conclusions.

Research Issues

Because of the human factors issues outlined above, one of the critical questions to be addressed is:

What role should the computer play so that, in working together with the user as a “cooperative system,” errors are minimized?

The assumptions are:

1. Without computer assistance, users make certain
types of errors. Given knowledge of these errors, an expert system can be designed to help in detecting or avoiding them.

2. Computer system designers, like users, are fallible. They are not likely to anticipate all of the scenarios that will arise in the laboratory, nor can they envision all of the behaviors of the complex knowledge base that they have created. They also cannot fully eliminate the possibility of implementation errors in building the system. Hence, the computer system will be brittle (producing erroneous answers when encountering scenarios unanticipated by the designer) and will make errors under certain circumstances.

Therefore, a system design must be found that ensures the person will be effective in detecting and compensating for the computer system designer's errors.

More specifically, the study reported below asked the question:

How does the role of an expert system influence the conclusions of the user under circumstances in which the computer's knowledge might be inadequate?

An Empirical Evaluation

To address this general question, and to help guide the design of an expert system to assist blood banks, a study was conducted evaluating the effects of system design on overall performance. Two system designs were compared. Although the underlying knowledge base was identical for both systems, one computer design automatically completed a problem-solving task for the user, while in the other design the user completed the task and the system watched in a critiquing mode. The details of this study are given below.

Antibody identification assistant #1 (AIDA1): critiquing user performance

The first system, Antibody Identification Assistant #1 (AIDA1), provided online access to pertinent data for the task of antibody identification and allowed technologists to mark intermediate and final conclusions, similar to the way they would perform the task using paper and pencil. For each of the cases built into the system, the practitioner performed the antibody identification process by asking the computer (via a pull-down menu option) to show test results and other pertinent information, including relevant facts about the patient's medical history (Fig. 1).

Users could make annotations on the data sheets by selecting from a set of color-coded "markers," available as buttons along the top of the screen, and then clicking on red cells of interest using the mouse. Rows, columns, and red cells could be highlighted, and antibodies could be marked as either "ruled-out," "unlikely," "likely," or "confirmed." The colors for these markers were chosen to correspond with the danger of introducing that antigen into the patient's bloodstream. For example, a confirmed antibody was marked red, indicating that a transfusion reaction could occur if blood containing that antigen was given to the patient.

The data sheets used in AIDA1 were the same as those currently used in paper form in some laboratories. The organization of the display is shown in Figure 2.

For the purpose of this experiment, AIDA1 had a single error-detection tool embedded in the design. It monitored the technologist's performance for errors in the rule-out procedure (ruling out possible antibodies based on nonreacting test cells). The system used its expert rules to critique the user's performance for possible sips
or mistakes, such as ruling out an antibody when the respective antigen was present on a reacting test cell. Thus, each time the user marked an antibody as ruled-out, the system checked its rule base to evaluate the appropriateness of that inference. If the system's heuristics did not agree with the user's inference, then an error message was displayed in a dialogue box with an explanation (see Fig. 5). The user could then correct his or her "error" by clicking on a button to undo the rule-out. Alternatively, the technologist could click on another button to leave it the way it was, essentially overriding the system.

![Sample error message for an invalid rule-out.](image)

Fig. 3. Sample error message for an invalid rule-out.

**Antibody identification assistant #2 (AIDA2): automatic completion of subtasks**

The second system not only had the passive rule-out detector embedded as in AIDA1, but it also had a function that allowed the user to request the computer to automatically complete all rule-outs based on nonreacting cells. Thus, AIDA2 had the same knowledge base for rule-out as did AIDA1 but would actually complete the task of rule-out upon request by the user. This automatic rule-out function offered the potential to save time and effort for the technologist. The literature reviewed earlier, however, suggests that this function could induce new errors on cases in which the computer was incompetent (i.e., in cases in which the technologist needed to apply his or her own expertise and override the computer's conclusions).<sup>3,5,8</sup>

**Materials and Methods**

By comparing performances of technologists using AIDA1 and AIDA2, this study was designed to provide objective data regarding the existence and seriousness of the cognitive bias predicted by the authors of this study. It was also designed to provide insights into the errors made by technologists when unaided by the expert system.

**Subjects**

Thirty-two blood bankers from ten blood bank laboratories were each tested on either AIDA1 or AIDA2. Subjects were randomly assigned to one of two groups. Subjects ranged in experience from newly hired to over 30 years of experience.

**Experimental design**

All subjects were tested on the same ten cases, presented in the same order. The first three cases were used to give both groups the same initial training on how the system functioned and establish a benchmark on their current performance strategies (i.e., did they rule-out, did they do antigen typing, etc.). Subjects could use the system to request test results, highlight and mark the panels as they might normally do with a paper form, and input their answers for each case. No feedback was given as to their problem-solving technique, nor were they told if their final answers were correct.

Depending on which of the two groups the subject was assigned, the system was then enhanced to include functions corresponding to either the AIDA1 system or the AIDA2 system. Cases 4 and 5 were used to train subjects with the additional features and Cases 6-10 were used to test and compare performances when using the two systems.

**Ten cases used to test AIDA**

A crucial part of evaluating a cooperative problem-solving system like AIDA is to use a set of tasks that test the range of scenarios that might be encountered in practice. Therefore, in testing AIDA, a case was used for which the computer's automatic rule-out strategy failed. (This case [Case 8] was a weak anti-D, which did not react with all of the test cells.)

The characteristics of the ten cases are summarized below:

**Case 1.** One antibody, anti-C, reacting strongly.

**Case 2.** Two antibodies, anti-Le<sup>a</sup> reacting only on immediate spin (IS), and anti-Fy<sup>a</sup> reacting by indirect antiglobulin test (IAT) and showing a dosage effect.

**Case 3.** Two antibodies, anti-D reacting at IS, LSS...
37°C, and IAT, and anti-K reacting by IAT. The D-screening cells masked anti-K but panel work demonstrated both.

Case 4. (Used to introduce subjects to the additional features of the system to which they were assigned.) One antibody, anti-M. The two screening cells showed variable reactions at IS, LISS 37°C, and IAT, suggesting two antibodies. Panel work demonstrated at anti-M showing a dosage effect.

Case 5. (The second case used to train subjects on the use of their version of AIDA.) Two antibodies, anti-Lw reactive at IS, LISS 37°C, and IAT, and anti-K reactive by IAT. Antibody screen indicated the presence of two antibodies, confirmed by the panel.

Case 6. (First test case comparing two versions of AIDA.) One antibody, anti-Fy^a, showing a dosage effect. Both screening cells reacted the same by IAT. The panel, however, showed the dosage effect. It was predicted that for this case, AIDA2 might be slightly faster for the user because much of the ruling out was done automatically.

Case 7. (An easy case.) One antibody, anti-K, reacting strongly. Only one of two screening cells reacted, allowing some antibodies to be ruled out prior to the panel work. All other antibodies could be ruled out by the panel and the patient tested negative for the K antigen. This was another case in which AIDA2 users should solve the case faster because the automatic functions do the ruling out.

Case 8. (A real patient's case used to test how users cope with the "brittleness" of this expert system.) This case was a weak, newly-forming anti-D in a 36-week pregnant, Rh-negative woman. These facts suggest the possible use of prophylactic RhIg. However, only two of five D+ cells were reacting weakly. Since an Rh antibody was the greatest possibility, a different technique such as using enzymes should be tried to enhance reactivity. It was predicted that only those users with enough a priori knowledge might recognize important clues leading to the right answer. Neither version of AIDA identified those clues. Therefore, it was predicted that AIDA2 would be detrimental to those users who might normally notice such clues but would ignore them when the system automatically ruled out anti-D.

Case 9. (Another real and deceptively simple case of two antibodies, anti-E and anti-K, looking like one antibody, anti-Fy^b.) The panel reactions seemed consistent with the pattern of an anti-Fy^b showing dosage. However, some Fy(b+) cells did not demonstrate the dosage. Users not following through with ruling out and antigen typing concluded that anti-Fy^b was the answer. However, the patient was Fy(b+) with anti-E and anti-K.

Case 10. (This case was used to test for completeness of rule-out.) There were two antibodies, anti-Fy^a and anti-E. The anti-Fy^a masked anti-E. The only way to discover anti-E was to test additional Fy(a-) cells positive for antigens masked by the anti-Fy^a pattern. Alternatively, enzymes could be used to destroy the Fy^a antigen on the panel cells. For this case, it was predicted that AIDA2 users would fare better because of the ease in using the automatic rule-out feature.

Results

There were five major findings that resulted from the testing of AIDA:

1. There was a significant savings in time when using AIDA2, the version with automated sorting, moving, and ruling out.
2. Both versions of the system reduced errors (mistakes and slips) through their feedback and error checking.
3. If the weak-D case (the case in which the computer rules out the right answer) is not included in the analysis, those subjects using the AIDA1 system missed antibodies in over twice as many cases as those subjects using AIDA2. However, on the weak-D case, a significant effect in the opposite direction occurred—more AIDA2 users missed the correct antibody (p < .10).
4. Many of the technologists lacked relevant domain knowledge and used strategies that were questionable, and there was evidence to show that some technologists did not know when and why their strategies might fail.
5. There was evidence of technologists learning from the computer support tools in both AIDA1 and AIDA2, suggesting that having such support tools can provide embedded training.

Savings in time with AIDA2

As predicted, there appeared to be a significant savings in time when using AIDA2. In four of the five experimental cases, the mean case time using AIDA2 was less than the mean time using AIDA1 (see Fig. 1).

Effect of certification level

An analysis of how subjects performed according to certification level suggests that Specialists in Blood
Banking (SBBs) tend to be much better than Medical Technologists (MTs) and Medical Lab Technologists (MLTs) (see Table 1). However, the sample size was too small to attain statistical significance. This is most likely because SBBs go through more rigorous training and examination than do MTs and MLTs. However, some MTs and MLTs have more experience but no formal training. This difference in expertise should not have had a great effect on the analysis for this experiment, as the number of subjects for each certification level was fairly evenly distributed across the two conditions. In fact, looking at the results across condition and certification level (Table 2), the percentages of wrong answers were fairly even.

**Effect of system role on performance**

If we look at the total number of times users missed antibodies on a case, there seemed to be no difference between the different versions of the system (19 percent misidentification rate for AIDA1 users and 17.5 percent misidentification rate for AIDA2). However, if we look only at cases in which the antibodies in AIDA2 made valid inferences (all but the weak-D case), there was a trend in favor of AIDA2. Without the weak-D case in the analysis, there was an 11.9 percent misidentification rate for AIDA1 users as opposed to a 5.6 percent misidentification rate for AIDA2 users.

In contrast, for the weak-D case, AIDA2 users had a 72 percent misidentification rate as opposed to AIDA1 users, who had a 43 percent misidentification rate. Thus, this study raises serious concerns that the automatic completion of a subtask by the computer can induce a significantly higher error rate than an equivalent system used to critique the person's performance when the computer provides misleading results.

**Analysis of existing, inefficient strategies**

Although the primary focus of this study was on the effects of the system's role on performance, it also provided insights into current weaknesses in the strategies of some technologists:

1. **Solution did not explain all reactions.**

   Eleven subjects marked answers that did not account for all of the reactions exhibited on the various test panels. For example, in Case 2, the reactions on the antibody screen strongly suggested the presence of two antibodies because one cell was reacting only at the IS phase and the other cell reacted only by IAT. However, many subjects found only one of the two antibodies because they used a test procedure that enhanced one type of reaction and inhibited the other. (Prewarming enhanced the IAT reactions and inhibited the IS reactions: running the cells at room temperature had just the opposite effect.) Finding a match based only on one of these panels left some unexplained reactions on other panels.

2. **Not using all information provided by a test result.**

   Often, users would not make all of the inferences possible given a test result. For example, when one subject ran an additional cell that came up negative, he only ruled out anti-E when other antibodies could have been ruled out on that cell as well. It is possible that, since his goal at that point was to rule out anti-E and not to rule out all antibodies possible (as it usually is when running a main panel), he never realized that the diagnostic information he collected to satisfy his current goal also
provided information that would satisfy other goals, such as solving the case as quickly and as efficiently as possible.

3. Ruling out on reactive cells.

Three users exhibited the strategy of ruling out antibodies on cells in which there were reactions but the antigen was not present. This strategy will work only if just one antibody is present in the patient. As soon as there is more than one antibody, such an strategy can rule out the correct answer.

4. Not knowing how different test procedures will affect results.

Many of the users did not know which antibodies would be affected by which tests. For instance, if one runs a warm panel certain antibody reactions will be weakened, sometimes to the point of no reaction. Therefore, those that are weakened by a test process should not be ruled out based on that panel. Yet eight of the practitioners ignored or were unaware of this and ruled out incorrectly on the various panels, sometimes causing them to rule out one or more of the answers.

5. Failure to collect independent, converging evidence.

Thirty-one percent of the users did not do antigen typing on either of the two test cases, a technique that will provide converging evidence for an answer. For those who did do antigen typing, two subjects did not use the information from a positive antigen typing to rule out the corresponding antibodies. In addition, two subjects interpreted the results of the antigen typing backward, ruling out an antibody because the patient lacked that antigen.

6. Fixating on early hypotheses.

Of those subjects who made predictions based on the information on the antibody screen, two subjects used information to such an extent that, from then on, they only considered the subset of antibodies that they had marked as likely. They only tried to find a solution from among that subset and only tried to rule out the remaining members of that subset. It was as if they had implicitly ruled out everything else on the antibody screen that did not seem likely, based on the two screen cell reactions.

Evidence of embedded training

There was also evidence of embedded training. Some technologists learned from the computer, predicting when it would and would not give them an error message. For example, the technologists would start looking to see if the antigen was homozygous or not and say, “Will this one beep at me? No. Good. It won’t like that one,” and they eventually learned to rule out on only homozygous cells.

Conclusions

Two versions of a computerized support tool were designed, each system having its own underlying design philosophy. AIDA1 was designed to be a critiquing system alone, while the AIDA2 system was designed to also have active support tools for automatic completion of subtasks. Comparing the two versions of the system showed that the AIDA2 system helped to solve cases faster.

Given the modest sample size of this initial pilot study, all conclusions need to be considered tentative and need further study. However, the results suggest that (1) many of the users lacked relevant knowledge of the field and used strategies that were questionable. There was also evidence that some users did not know when and why their investigative strategies were likely to fail; (2) if the weak-D case (the case in which the computer was incompetent), ruling out the right answer is excluded from the analysis, a sizable number of subjects using both versions made errors in their final answers (an 11.9 percent error rate for users of AIDA1 and a 5.9 percent error rate for users of AIDA2); and (3) on the weak-D case, users of AIDA2 made 29 percent more errors than users of AIDA1. Although the sample size is too small to achieve statistical significance, this result cautions us about automating functions because of possible impacts on user performance.

In considering the design of a decision-support system, three sources of human error need to be considered:

- Errors made by unaided users (the status quo)
- Errors made by the system designers and implementers
- Errors made by the aided users

Thus, a critical question in developing such a system is:

Can we reduce the first category of errors without creating significant new errors because of the latter two categories of errors?

The study reported above, combined with previous studies, suggests that significant cognitive errors are made by unaided medical technologists. Consistent with similar studies in other domains, this study also provides evidence in favor of designing critiquing sys-
tems rather than automating subtasks, if there is reason to believe that (1) the system design will be "bittie," and (2) if the avoidance of errors is more important than faster performance.

Based on this guidance, we are exploring the design of a more complete critiquing system that has a full range of expert functions to monitor user performance on the task of antibody identification. Among other things, this system will monitor all six of the areas in which we found technologists lacking in expertise in this study. A critical question in terms of the efficacy of this approach is whether such a critiquing system, working in cooperation with users rather than automating functions, can reduce the 11.9 percent error rate found for users of AIDA1 on cases when the system was competent, without producing anything similar to the 29 percent increase in errors faced by AIDA2 on cases when the system was incompetent.

Acknowledgment
This work has been supported under Grant No. HL51611-03 from the National Heart, Lung, and Blood Institute, Bethesda, MD.

References

Stephanie A. Gaertlein, Cognitive Engineering Laboratory, The Ohio State University, 210 Baker System, 1721 Neil Avenue, Columbus, OH 43210-1271; Philip J. Smith, Jodi Heintz Obnudovic, Sally Rudman, Patricia I. Strohm, Jack W Smith, and John Svirbely, Cognitive Engineering Laboratory. The Ohio State University, Columbus, OH.

---

**IMMUNOHEMATOLOGY—PAST, PRESENT, AND FUTURE**
Visit the American Red Cross exhibit at the AABB annual meeting in Orlando, Florida, in October 1996 and see Immunohematology on the World Wide Web! You will be able to Subscribe by computer, read past issues, send a letter to the editor, read the instructions to authors, and much more. Come to the American Red Cross exhibit (Booth 246) and tell us what you think of the present and future Immunohematology!