MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Are Clinicians Correct When They Believe They are Correct? Implications for Medical Decision Support

Charles Friedman[a], Guido Gatti[a], Arthur Elstein[b], Timothy Franz[c], Gwendolyn Murphy[d], Fredric Wolf[e]

[a] Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, USA
[b] Departments of Medical Education and Medicine, University of Illinois, Chicago, USA
[c] Department of Psychology, St. John Fisher University, Rochester, USA
[d] Department of Community and Family Medicine, Duke University, Durham, USA
[e] Department of Medical Education, University of Washington, Seattle, USA

Abstract

The process of clinical decision support is linked to the validity of clinicians' confidence in their judgments. Clinicians who are appropriately confident—highly confident when they are correct and less confident when they are incorrect—will access computer-based and other information resources only when they are needed. Clinicians who are consistently underconfident will rely on external resources when they are not needed. Those who are overconfident, who believe they are correct when in fact they are not, will be prone to medical errors. An extensive literature indicates a general tendency toward overconfidence in human judgment.

This study explores the relationship between confidence and "correctness", across three levels of clinical experience, in the task domain of diagnosis in internal medicine. We created detailed synopses of 36 diagnostically challenging cases and divided them into four equivalent sets of nine cases each. We asked 216 subjects at three experience levels (72 senior medical students, 72 senior medical residents, and 72 faculty attendings) to generate a differential diagnosis for each of the nine cases in one randomly-assigned set, and simultaneously to indicate their level of confidence in each of their diagnoses. We then examined the relationship between the correctness of these diagnoses (the appearance of the correct diagnosis anywhere in the hypothesis list) and these confidence judgments, for all subjects and separately for subjects at each experience level.

Results indicate a small but statistically significant relationship associating correctness with higher confidence for all subjects (Kendall's $\tau_b = -.106; p < .0001$). This statistical relationship is strongest for the students ($\tau_b = -.121; p < .001$), somewhat lesser but still significant for the faculty-level attendings ($\tau_b = -.103; p < .005$), and non-significant ($\tau_b = -.041$) for the residents. (The negative correlations are a coding artifact.) Subjects in this study showed a tendency toward underconfidence: they had low confidence in correct diagnoses more often than they had high confidence when wrong. Nonetheless, they were overconfident and thus "error prone" for 17% of cases overall.

The medical students were possibly overmatched by the difficulty of the cases, so their concordance between confidence and correctness may have resulted from an awareness that they were often guessing. The relatively low concordance seen in the residents and attendings makes a strong argument that decision support systems to reduce medical errors should include both "push" and "pull" models. In sum, these results indicate that medical decision support systems cannot rely exclusively on clinicians' perceptions of their information needs, as such perceptions will frequently be incorrect.

Keywords:

Clinical Decision Support; Judgment; Diagnosis; Decision Making; Calibration

Introduction

Decision making in medicine, as in all fields of professional practice, involves an interplay of what psychologist Donald Norman calls "knowledge in the head and knowledge in the world" [1]. Many decisions will be made based on the clinician's own personal knowledge, but other decisions will be informed by knowledge that derives from a range of external sources including printed books and journals, communications with professional colleagues, and, increasingly, a range of computer-based knowledge resources. Decision-making based on personal knowledge alone uses a minimum of time, which is a scarce resource in health care practice, but every practitioner's personal knowledge is incomplete in various ways. Decisions based on flawed, incomplete, or outdated personal knowledge can result in errors. A recent study [2] has documented that medical errors are a significant cause of morbidity and mortality, making a strong case for routine use of external knowledge resources in clinical practice.

Computer-based decision support systems (DSSs) comprise an important "external knowledge" resource for practitioners. A primary arm of the field of medical informatics is concerned with development, deployment, and evaluation of such systems. Many hundreds of publications document the range of systems that have been developed [3]. DSSs can function in synchronous mode in conjunction with an order entry system [4], in asynchronous mode triggering alerts and reminders [5], or as a "critiquer" of clinicians' plans [6]. Irrespective of their mode of operation, and as long as the ultimate decisions are made by humans and not the machines, all DSSs trigger a psychological process that requires the practitioner to combine information from the DSS with his/her own personal knowledge in reaching a decision.

Key to this process is the practitioner's confidence in his/her personal knowledge relevant to a particular clinical decision. Clinicians make decisions under uncertainty. At the time of decision-making, they do not know whether their actions are correct or incorrect. Clinicians who believe they are correct, or believe they know all they need to know to reach a decision, will be unmotivated to seek additional knowledge and unreceptive to any knowledge or suggestions a DSS presents to them, irrespective of how valid or timely that external advice is. Potential resistance to DSS advice also has underpinnings in clinician culture that assigns high status to practitioners who are personally knowledgeable [7].

To unpack the complex psychological processes that underlie clinical decision support, and ultimately may determine how successful DSSs will be in reducing errors, it is important to understand the validity of clinicians' beliefs about their personal knowledge. In this respect, Table 1 depicts the four possible states of concordance/discordance between belief and reality. The two "diagonal" cells represent concordance. While the "true positive" cell represents the ideal situation, the "true negative" cell is also salutary in the sense that a clinician who is truly wrong, but also unconfident, will likely seek help from an external resource. Of the two "off diagonal" cells, the "false negative" cell corresponding to underconfidence is the more salutary. A clinician who really is correct but has low confidence will be motivated to seek information that will likely confirm an intent to act correctly. However, it is possible that a consultation with an external resource can talk a clinician out of the correct decision. The "false positive" cell corresponding to overconfidence is of greatest concern. A clinician with high confidence in a decision that is really incorrect is in a psychological state we might call "error prone". He/she will not be motivated to seek information that could correct a flawed judgment or decision. The decision support strategy for such cases must be one that pushes information to clinicians who have not asked for it.

Table 1 – States of concordance and discordance between correctness and confidence

Clinician Confidence	Decision is Correct	Decision is Incorrect
High	"True Positive" Region of Appropriate Confidence	"False Positive" Region of Overconfidence
Low	"False Negative" Region of Underconfidence	"True Negative" Region of Appropriate Confidence

An extensive literature in psychology [8,9] and medical decision-making [10,11] addresses the relationship between subjective assessments, analogous to confidence in this study, and objective reality, analogous to the correctness of a diagnosis. This work has emphasized studies of "calibration" in which subjects are asked to assign subjective probabilities to events whose true probabilities are known or estimated from data, across a range of "true" probabilities. A general finding from these studies is that humans are overconfident across a range of tasks and levels of expertise; their subjective probabilities are higher than the corresponding true values [9]. Studies related to medical diagnosis have typically required clinicians to estimate the probability of a specific disease over a series of case presentations [10,11]. Overconfidence is reflected in clinicians' beliefs that the target disease is more likely than objective data would suggest. A study similar to the present one in design and intent [12] found that medical students tended toward underconfidence in their diagnostic judgments when classifying abnormal heart rhythms. To the best of our knowledge, no study has examined the relationship between confidence and correctness in diagnoses of complex clinical cases, across a range of diseases and subjects' levels of clinical experience. The advent of clinical decision support systems, coupled with the contemporary concern about medical errors, takes this issue out of the psychology laboratory and into the day-to-day practice of medicine.

Motivated by the central roles confidence and correctness play in clinical decision support, this work addresses the following specific questions:

1. In the domain of diagnosis in internal medicine, what is the relationship between clinicians' confidence in their diagnoses and the correctness of these diagnoses;

and does the relationship between confidence and correctness depend on clinicians' levels of experience ranging from medical student to attending physician?

2. When their perceptions are mismatched, do clinicians tend toward overconfidence or underconfidence, and does this tendency depend on level of clinical experience?

Materials and Methods

To address these questions, we employed a dataset created for a study of diagnostic decision support systems [13]. We developed for this study detailed written synopses of 36 diagnostically challenging cases from patient records at three US medical centers: the University of Illinois at Chicago, the University of Michigan, and the University of North Carolina. Each institution contributed twelve cases, each with a firmly-established final diagnosis. The 2-4 page case synopses did not contain results of "definitive" tests that would have made the correct diagnosis obvious to most or all clinicians. The cases were divided into four approximately equivalent sets balanced by institution, pathophysiology, organ systems, and rated difficulty. Each case set therefore contained nine cases, three from each institution.

We then recruited to the study 216 subjects from these same institutions: 72 fourth year medical students, 72 second- and third-year internal medicine residents, and 72 general internists with faculty appointments and at least two years of post-residency experience (mean: 11 years). Recruitment was balanced so that each institution contributed 24 subjects at each level of experience. Each subject was randomly assigned to work the nine cases comprising one of the four case sets, so each subject worked three cases from his/her own institution and six from the other institutions. We employed a "two pass" protocol whereby each subject worked the assigned cases first without, and then with, assistance from a computer-based decision support system: either ILIAD or QMR. On each pass through each case, subjects generated a diagnostic hypothesis set with up to six items. After generating their diagnostic hypotheses, subjects were asked: "How likely is it that you would seek assistance in establishing a diagnosis for this case?" They responded using a 1-4 scale with anchor points of "1" representing "unlikely" (and thus high confidence in the subject's diagnosis) and "4" representing "likely" (and thus low confidence). After deleting cases with missing data, the final dataset for this work consisted of 1911 cases completed by 215 subjects.

Data for this study derive only from subjects' first pass through each case, since the immediate focus of this investigation is clinicians' confidence in their own diagnoses when unassisted by a computer system. We assigned a binary score of correct-incorrect to the diagnostic hypothesis set offered by each subject in each case. We did this using procedures reported earlier [13], counting as correct a case in which the correct diagnosis, or a very closely related disease, appeared anywhere in the subject's hypothesis set. The measure of clinician confidence was the "1-4" response to the question about seeking assistance, as described above. Since subjects did not receive any feedback on their diagnoses until they had completed work on all assigned cases, subjects offered their confidence judgments without any definitive knowledge of whether their diagnoses were, in fact, correct. (Most subjects received no feedback at all.)

To address the first research question, we cross-tabulated binary case correctness scores with (four) levels of confidence, doing this first across all subjects and then separately for each experience level. We employed non-parametric methods, using Kendall's τ_b as a measure of association, to test the significance of each relationship. To address the second question, we dichotomized the confidence scores, such that responses of "1" and "2" were associated with high confidence and the clinicians' *belief* that their diagnosis was correct. Responses of "3" and "4" were associated with low confidence and clinicians' *belief* that their diagnosis was incorrect. This enabled us to populate a 2 x 2 contingency table, isomorphic to that portrayed Table 1, representing the concordance between correctness and confidence for all cases and separately for subjects at each experience level. We computed an overconfidence coefficient equal to the fraction of all cases in which clinicians' confidence was high but their diagnosis was incorrect, and an underconfidence coefficient equal to the fraction of all cases in which confidence was low but the diagnosis was correct.

In these studies, we employed the case as the unit of analysis because this approach conveys the most readily interpretable portrayal of the results. Because multiple cases were completed by each subject, the results for each case are not statistically independent. We therefore conducted additional analyses with correction for the nesting of cases in subjects. The results of these additional analyses are essentially unchanged from those reported below.

Results

Table 2 displays the relationship between correctness of diagnosis and levels of confidence for all subjects and separately for each experience level. (Although the analyses for the first research question are based on the four-level confidence scale, Table 2 displays dichotomized confidence scores in the interest of brevity.) The difficulty of these cases is evident from the result that 760 of 1911 (40%) were correctly diagnosed by our sample. In accord with expectations, attendings (49% correct) outperformed the residents (44% correct) who outperformed the students (26% correct). The difficulty of the cases is also reflected in the skewness of the confidence ratings to the lower end. This confirms that the subjects collectively were aware that the cases were hard.

Addressing the first research question, the Kendall τ_b measure of association between the binary measure of correctness and the four level measure of confidence is computed to be -.106 ($p < .0001$). The negative coefficient is an artifact of coding and reflects the expected polarity of the relationship. There is thus a small but significant tendency for subjects to be more confident when, unbeknownst to them, their diagnoses were correct. Separately for each level of training, Kendall coefficients are: for students $\tau_b = -.121$ (n = 645 cases; $p < .001$), for residents $\tau_b = -.041$ (n = 638; NS), and for attendings $\tau_b = -.103$ (n = 628; $p < .005$). So the strongest relationship is seen in the students; a relationship lesser in magnitude, but still statistically significant, is seen in the attendings. There is no statistical relationship between correctness and confidence observed in the residents.

Table 2 – Cross-tabulation of correctness and confidence for all subjects and each experience level

Experience Level	Diagnosis	Confidence: High	Low	Total
All Subjects	Correct	271	489	760
	Incorrect	318	833	1151
	Total	589	1322	1911
Students	Correct	54	114	168
	Incorrect	104	373	477
	Total	158	487	645
Residents	Correct	103	178	281
	Incorrect	125	232	357
	Total	228	410	638
Attendings (Faculty)	Correct	114	197	311
	Incorrect	89	228	317
	Total	203	425	628

With reference to the second research question, Table 2 summarizes the case frequencies for which clinicians at each level were correctly confident—where confidence was in accord with correctness—as well as frequencies where they were overconfident and underconfident. Students were balanced almost equally between overconfidence (16% of cases) and underconfidence (18% of cases). Their more experienced colleagues were more often underconfident (28% of cases for residents, 31% for faculty) than overconfident (19% of cases for residents, 14% for faculty).

Discussion

It is most instructive, in interpreting these results, to separate the medical students' results from those of their more experienced colleagues. Recognizing that these cases were very difficult, students' diagnoses were incorrect for 74% of cases overall. The students were probably overmatched by many of these cases. Subjects who were overmatched in a case, and thus guessing at a diagnosis, were probably quite aware that they were overmatched. For such cases, one would expect high alignment between perception and reality, and the greater concordance seen in the students may be an artifact of the students' substantial difficulty with these challenging cases. A better estimate of the concordance between confidence and correctness for the students might be obtained by challenging the students with cases they could diagnose approximately 50% of the time, making the diagnostic task as difficult for them as it was for the faculty and residents.

By contrast, residents and faculty correctly diagnosed 44% and 49% of cases respectively. Because these two more experienced groups are directly responsible for patient care, and were much less likely to be overmatched by these cases, findings for these groups take on a different interpretation and greater clinical significance. In both groups, the relationship between "being correct" and "thinking you are correct" was quite low, and in the residents it does not exceed chance expectations. The greater concordance exhibited in the faculty suggests that awareness of one's correctness may be an important component of clinical experience. On cases for which these clinicians' correctness and confidence were discordant, both groups showed a tendency toward underconfidence, which is at variance with previously reported work on clinician calibration. Nonetheless, in 19% of cases for residents and 14% for faculty, these clinicians were overconfident: they placed high credence in a diagnosis that was in fact wrong, which is the "error prone" cell of primary concern in Table 1.

Limitations of this study include restriction of the task to diagnosis in internal medicine. Diagnosis, which may be more or less difficult or engage different cognitive processes in other clinical disciplines, may generate different levels of concordance in those disciplines. Differences in results may also been seen in other clinical tasks such as determination of appropriate therapy for a problem already diagnosed. The cases, chosen to be very difficult and with definitive findings excluded, certainly generated lower rates of accurate diagnosis than is typically seen in routine clinical practice. Were the cases in this study more routine in nature, this may have affected the measured levels of concordance between confidence and correctness. In addition, this study was conducted in a laboratory setting, using written case synopses, to provide experimental precision and control. While the synopses contained very large amounts of clinical information, the task environments for these subjects was not the task environment of routine patient care.

Conclusion

These results carry numerous implications for the design of decision support systems. Decision support systems that rely on a "pull metaphor", where clinicians recognize that they need external information about a case and then take action to find this information, assume that clinicians' perceptions of their "correctness" are accurate. The results for residents and faculty indicate that there is a role for "pull" systems in decision support since these subjects were wrong, and accurately unconfident, in 36% of cases. While there is no guarantee that clinicians would actually seek assistance for these cases, in this study they indicated a clear intention to do so. The fractions of cases on which residents were overconfident (19% overall) indicates an additional need for systems that "push" information "just in time" to clinicians who have not necessarily asked for it [14]. In sum, these results suggest that effective decision support systems cannot rely on clinicians to perceive a need for external information, as these perceptions are too often inaccurate.

Acknowledgments

This work was supported by Grant R01-LM-05630 from the U.S. National Library of Medicine.

References

[1] Norman DA. The Design of Everyday Things. New York: Doubleday, 1990.

[2] Kohn LT, Corrigan JM, and Donaldson MS, eds. *To Err is Human: Building a Safer Health System.* Washington: National Academy Press, 2000.

[3] Miller RA. Medical diagnostic decision support systems--past, present, and future. *J Am Med Inform Assoc* 1994: 1: 8-27.

[4] Evans RS, Pestotnik SL, Classen DC, Clemmer TP, Weaver LK, Orme JF, Lloyd JF, and Burke JP. A computer-assisted management program for antibiotics and other antiinfective agents, *N Eng J Med* 1998: 338: 232-238.

[5] Wagner MM, Pankaskie M, Hogan W, Tsui FC, Eisenstadt SA, Rodriguez E, and Vries JK. Clinical event monitoring at the University of Pittsburgh. *Proceedings AMIA Fall Symposium* 1997: 188-92.

[6] Miller PL. Building an expert critiquing system: ESSENTIAL-ATTENDING. *Meth Inf Med* 1986: 25: 71-78.

[7] Weaver RR. *Computers and Medical Knowledge: The Diffusion of Decision Support Technology.* Boulder, CO: Westview Press, 1991.

[8] Tversky A and Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974: 185: 1124-1131.

[9] Lichtenstein S and Fischhoff B. Do those who know more also know more about how much they know? *Org Beh and Human Perf* 1977: 20: 159-183.

[10] Christensen-Szalanski JJ and Bushyhead JB. Physicians' use of probabilistic information in a real clinical setting. *J Exp Psych* 1981: 7: 928-935.

[11] Tierney WM, Fitzgerald J, McHenry R, Roth BJ, Psaty B, Stump DL, and Anderson FK. Physicians' estimates of the probability of myocardial infarction in emergency room patients with chest pain. *Med Dec Making* 1986: 6: 12-17.

[12] Mann D. The relationship between diagnostic accuracy and confidence in medical students. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, 1993.

[13] Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, and Abraham V. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: A multisite study of 2 systems. *JAMA* 1999: 282: 1851-1856.

[14] Chueh H and Barnett GO. "Just in time" clinical information. *Acad Med* 1997: 72: 512-517.

Address for Correspondence

Center for Biomedical Informatics
University of Pittsburgh
8084 Forbes Tower
Pittsburgh, PA 15213
USA
cpf@cbmi.upmc.edu