Abstract. The increasingly frequent application of formal methods, including algorithms and computer programs, to processes that are ordinarily viewed as judgmental seems to be a source of both promise and unease for physicians. A consideration of some of these judgments, and some of which may be carried out as a computation. My purpose here is to identify the latter cases. The empirical evidence suggests that such a demarcation is feasible. The most important question now is: "What should we do with these human beings?" Until this more fully explored, tension between the two advocates will persist. (N

Example, the unanswerable question is: Is artificial intelligence governed by rules that can be analyzed in terms of what the computer does and what the computer does is the whole matter of the scope of this discussion. Earlier portions of the paper referred to the time of the computer to the determination of the diagnosis. The consultation must be fully ex- 

SPECIAL ARTICLE

CLINICAL JUDGMENT AND COMPUTERS

Marsden S. Blois, Ph.D., M.D.

Abstract The increasingly frequent application of formal methods, including algorithms and computer programs, to processes that are ordinarily viewed as judgmental seems to be a source of both promise and unease for physicians. A consideration of some of these methods suggests that it may be helpful to attempt to distinguish carefully between judgment and computation. Medical care involves a complex of inferential processes, any of which may be performed as judgments, and some of which may be carried out as a computation. My purpose here is to identify the latter cases. The empirical evidence suggests that such a demarcation is feasible. The most important question appears not to be “Where can we use computers?” but “Where must we use human beings?” Until this matter is more thoroughly explored, tension between physicians and computer advocates will persist. (N Engl J Med. 1980; 303:192-7.)

Whenever the topic of medical judgment is discussed by physicians and computer scientists, some of the physicians will be apt to point out that medicine is as much an art as science, and that rules cannot replace intuition and instinct. The computer scientists may then reply that unless medicine is pure magic, clinical judgment (by which I mean the decision-making processes that we use in the absence of explicit rules or appropriate tools) must be based on inferences made from patient data and prior knowledge. They will correctly point out that certain types of knowledge can be represented in a computer system, and that whether operating as a logic engine or as a calculating machine, the computer performs more accurately and frequently more rapidly than the human being. The debate then commonly halts at this point, which is unfortunate because it is precisely here that several important issues should be raised.

When these same physicians are back with their medical students, they will probably be found describing the rules for doing this or that, while insisting that medicine is a science, not an art. The computer scientist, back in his office may resume the task of attempting to represent in his program the difference between the meanings of the words “up” and “down,” so that his program will “understand” the sentence, “The baby threw up on the down coverer.”

Serious and interesting issues underlie this debate. Many physicians are in a quandary because they do not know quite what to expect from computers. Perhaps more important, a perceived threat by the computer may have caused some physicians to underestimate the relative worth of their own various skills.

Analysis of the notion of clinical judgment and of some of the proposed contributions of computers to judgmental processes in general involves issues at several different levels. All these issues cannot be explored here. There are, for example, the unanswered questions of how much of judgment is computable and of whether thought itself is governed by rules. These issues have recently been analyzed in terms of what the computer cannot do and what the computer should not do. Other writers find the whole matter simply depressing. To restrict the scope of this discussion, let us consider only the earlier portions of the medical-care process: those occurring from the time of the first patient-physician encounter to the determination of a tentative or working diagnosis.

When a new patient steps into the consultation room, the physician’s attention must be fully extended. At that moment, the maximum cognitive span is required of the physician. The primary-care physician must be prepared to deal with the possible presence of one or more of several thousand diseases or medical “conditions” and with an unimaginably large number of presentations not representing disease. The results of a preliminary conversation, history taking, physical examination, and perhaps some laboratory tests or special examinations will be used to reduce this enormous set of possibilities to a small number of probable states of affairs: the differential diagnosis. To which of these separate acts, then, are we to apply the term “clinical judgment”? To the selection of the next question, to the search for a particular sign, to the ordering of a test, or to the final choice among the alternatives? It might seem equally applicable to all of them.

Let us represent this judgmental process by a funnel or horn (Fig. 1), with its large diameter at the onset of the process and its small end at the conclusion. The decreasing diameter of the funnel may be taken to represent the shrinking cognitive span required of the physician during this process. At the start, the physician’s breadth of comprehension must extend to the totality of the everyday world. That is, he or she must have at least an average acquaintance with the world and must be able to exercise common sense. The necessity for this breadth of comprehension is frequently overlooked as an ingredient in medical judgment. However, in dealing with computer programs, nothing can be taken for granted. And
Figure 1. The Cognitive Span Required during Diagnosis.
See text for discussion.

clinical judgment counts for little unless it rests on a
firm base of ordinary human judgment.

Whereas nearly everything is possible at the begin-
ing of a patient’s visit, the field of possibilities then
becomes progressively restricted. As more informa-
tion is obtained, the possibilities are reduced until
only a relatively small number of potential disorders
remain. The cognitive universe, in which further fact
finding and inference takes place and in which a de-
cision must finally be made, becomes smaller, more de-
tailed, and more specific. More important, a sense of
structure emerges. The alternatives not only become
fewer, but they are sharpened, and relations among
them appear.

Although it seems fitting to apply the term “clinical
judgment” to the inferential processes occurring be-
tween Point A and Point B in Figure 1, these processes
appear to be quite different in nature. Some physi-
cians, if pressed about these differences, might reply
that the reasoning processes involved near Point B are
the ones most demanding of their abilities, that in this
portion of the overall process their unique skills and
special knowledge are most called for, and that here
computers would be least useful. I will argue that the
contrary seems actually to be the case.

Let me begin with two examples that are referable
to medical events in the neighborhood of Point B and
represent problems frequently faced by physicians.
One is that of managing the treatment of a patient
with acid-base or electrolyte abnormalities. These
particular problems lie in an area of medicine in
which the causal mechanisms are fairly well under-
stood and the clinical management of the condition
has in part been reduced to the solution of certain
problems in applied physical biochemistry. The
necessary tasks here are the taking of a history, the ob-
servation of some clinical signs, the determination of
particular laboratory data, the substitution of these
values into appropriate formulas, and the solution of
these formulas. The patient may well have other prob-
lems or symptoms that are not described by these for-
mulas, and the physician must also deal with them.

But a physician who is serious about correcting the
acid-base or electrolyte disturbance would do well to
use the formulas (at least as a general guide to diag-
nosis and management) since they seem best to re-

present current medical understanding of the problem.

To facilitate the proper use of these formulas and
rules, Bleich has developed computer programs that
request and accept certain laboratory and clinical
data, solve the problems (with use of the formulas),
and then provide the physician with appropriate ad-
vice. Bleich’s programs go beyond the arithmetic
itself: they may request further data, suggest a dif-
ferential diagnosis, make therapeutic suggestions, and
provide the basis for the recommendations. It may be
asked whether such computer programs perform as
well as the physician. Since both the program and the
physician employ the same formulas (if they do not
always approach the problem identically), perhaps we
could phrase the matter differently and ask, “Do phy-
sicians solve these formulas (follow these rules) as well
as the computer?” Since neither mathematicians nor
engineers would claim to be able to calculate more reli-
ably than computers, it would seem strange to find
physicians who would make such a claim.

Another clinical problem is that of choosing the
proper doses of drugs, particularly of certain relatively
toxic ones. Kinetic models of drug distribution (re-
presented as sets of equations) have been developed
and incorporated into programs that, when provided
with appropriate input data, such as patient weight,
renal function, and amount and timing of the previous drug doses, will then calculate the dose necessary to achieve a desired plasma level. Such a program has been refined through the use of additional techniques, including feedback, in the attempt to deal with idiosyncrasies among patients. When such systems are evaluated, their performance usually appears to be better than the unaided physician's judgment. This should come as no surprise if one grants that, despite certain limitations that we will discuss later, computers do calculate extremely well. Programs of this type have become known as “expert” or “consultant” systems. Other such programs, employing different techniques, provide clinical advice in dealing with acute renal failure and antimicrobial therapy.

The first two examples above concern processes that are deterministic in the sense that their behavior can be predicted by known chemical or physical laws. If a relevant law is known and it can be represented by a suitable formula, a computation would seem generally to yield a more accurate prediction than would estimation or intuition. But there are also nondeterministic (stochastic) processes that, on average, can be predicted by formulas, whence insurance companies and gambling casinos are profitable.

In 1954, Paul Meehl, a clinical psychologist, reviewed a number of different studies in which the performance of a clinician (usually a psychologist) was compared with that of some statistical procedure. These studies were concerned with a variety of tasks, including psychologic diagnosis, prediction of the performance of students entering college or of student aviators, and prediction of recidivism among parolees. In each instance, certain test data or historical data were available, and a statistical procedure (frequently as simple as a linear regression model) was then applied after it had been empirically fitted to previous information on the particular task. The overall result of Meehl’s review was that in each of the 24 comparison studies (later increased to 35), the statistical, or as Meehl termed it “actuarial,” method performed as well as, or better than, the human predictors. The results of Meehl’s study were so perplexing that many clinicians and apparently puzzled Meehl himself. We seem instinctively to have a high regard for careful judgment. How could a mindless formula compete with it?

As Elstein" has commented, these findings appear puzzling, contrary to our intuition, and controversial. Meehl’s actuarial methods set standards of performance that could be approached or equaled but not generally surpassed by groups of human predictors. Once an optimum prediction formula has been found, obtaining the correct solution seems to be the best that can be done. There is no level of performance in calculating that is superior to achieving correctness.

More recently, medical diagnostic algorithms based on a number of techniques (statistical methods such as pattern matching, clustering procedures, decision rules, and production rules) have been incorporated into computer programs. These programs have been designed to use observational data and to choose the proper alternative from a fixed and usually small set. Such diagnostic programs have been used for congenital heart disease, thyroid disease, glaucoma, medical disorders, and the acute abdomen. A program of the last type, developed by de Dombal, has been evaluated in the emergency rooms of several hospitals. It has been claimed to perform better than house staff and even slightly better than consultants when the same clinical information is made available to both the program and a clinician. Performance such as this has been viewed by many physicians either with skepticism or as a prospective threat. Such results again raise the question of the role of clinical judgment.

In considering this matter, it is essential to appreciate that these predictions or decision programs (like those cited by Meehl) deal with activities located near Point B of Figure 1. The most ambitious of these programs, INTERNIST, selects from a set of a few hundred medical disorders and would thus stand slightly to the left of the others if they were represented in Figure 1, but it would still stand well to the right of Point A. All these computer programs operate in small and well-structured task domains that were initially organized (formalized) by a human judge functioning earlier at Point A. This feature is nearly always overlooked in discussions of computer diagnosis. Questions such as “Do we run the acute-abdomen program or the chest-pain program?” or even “Shall we run a diagnostic program?” must first be decided by human beings. A great deal of human information processing must take place before the job is turned over to the computer. When Dr. (as Meehl’s results) "The ultimate ignorance is surely to discover that the vast experience and formal training of the clinician results in judgments no better than the simplest possible formula," he repeats this error. By failing to recognize where the judgmental process begins, he confuses a portion of a process (the final choice between well-defined alternatives) for an entire process.

But let us go back and examine the informational state of affairs at Point A. How many facts are there here? Asking this question is equivalent to asking how many facts would have to be listed if one were to attempt a complete description of the world. Researchers in artificial intelligence, cognitive psychologists, and philosophers have studied this matter in much more detail than can be recounted here. Among the optimists, workers in artificial intelligence such as Minsky have estimated that something like a million “facts” would be needed to achieve “great intelligence” in an artificial-intelligence system. Pauker et al. have estimated that the core knowledge of internal medicine involves a million “facts” and that if the medical subspecialties were included, this figure
might rise to two million. Earlier, Russell and then Wittgenstein (in the *Tractatus*) had proposed that the world might be susceptible to description in terms of "atomic statements," propositions so primitive that they require no further explanation. This hope (for logical atomism) was subsequently abandoned by most philosophers, including Wittgenstein, and it is now generally agreed that there is no way to avoid an infinite regress if one attempts to describe the world in this way. In other words, any statement or proposition raises further questions that require still more explanations and lead to still more questions ad infinitum. The everyday world, it appears, cannot be described by a small number of "primitive" statements (there being no such things) or by any finite number of any kind of statements.  

The number of things that can be truthfully said about any object in the world is unimaginably large, and the only way in which we can speak usefully of things is to confine ourselves to matters that are relevant. By "relevant" I mean connected to the topic or purpose with which we are engaged at a given moment. It is only through this selection of relevant attributes that we succeed in communicating at all. At Point A in Figure 1, almost all the facts that could be stated about a patient would be irrelevant to the problem of finding out what is wrong. It is the physician's task to select from this unimaginably large number of indifferent and neutral facts the ones that happen to be relevant. This situation is quite unlike the one at Point B, where the selection process is largely completed, and the physician alone may suffice, and a calculation may do the trick. Once Point B is reached in the process of clinical judgment, the irrelevant facts have been filtered off and the medical problem has become relatively well structured. But initially the funnel of required comprehension opens up widely, interfacing with the world in all its complexity.

How does the physician deal with this complexity, and how successful are computer programs at this task? Many distinguished physicians have taught that history taking is the most critical step in the entire diagnostic process, and that this phase is the one that separates the exceptional physician from less able ones. From the viewpoint of information theory, this idea seems quite plausible. Consider the manner in which we might go about measuring the difference between the diagnostic worth of a fact that is volunteered by a patient and a fact that is elicited in response to a question. In one standard compilation of diseases, 3262 diseases and "medical conditions" are listed, with each defined in terms of clinical and laboratory attributes. In our studies of this compilation, the term "nosebleed" (or its synonyms) was found to occur in the description of 27 different diseases. If a new patient whom we were seeing reported "nosebleed," our attention would be drawn to these 27 diseases, and not to the 3235 in which "nosebleed" is not considered to be a characteristic. If on the other hand, we routinely asked all patients whether they had "nosebleed," most replies would be negative; this approach would enable us to reject only 27 diseases instead of 3235. The "diagnostic" or "selection" power of a positive response is thus more than 100 times greater than a negative one for this attribute. And when patients volunteer facts, they are almost always stated in an affirmative sense. Patients do not report the symptoms that they do not have. The amount of information obtained from a random inquiry about symptoms (yielding mostly negative responses) will therefore be very small. If, in contrast, the patient is encouraged to volunteer affirmative (positive) symptoms, the information value is clearly much greater. Every practitioner knows this approach and practices it instinctively, although it is not widely recognized that this practice rests on a firm theoretical basis. By beginning in this way (with the "chief complaint") the physician equipped with both common sense and medical knowledge is provided with a means for exploiting relevance. This is the physician's method for dealing with the unimaginably complex. How well would computer programs do here?

Perhaps the only computer systems appropriate here are those designed to take the patient's medical history. When introduced some years ago, these programs appeared to offer great promise, but since then interest in them seems to have waned. Friedman and Gustafson comment,  

A graphic example [of the failure to report the reasons for unsuccessful system performance] can be found in the area of computer applications to the acquisition of Medical History Data. Numerous groups across the country have worked and published in this area, often duplicating previous efforts. However, although the great majority of these efforts have since been abandoned we could find no publications detailing the reasons for these failures.

Yet the cause for this failure does not seem too difficult to locate. In order to ask relevant questions of a patient, the physician uses his or her sense of the situation and general medical knowledge. But for a fact to be relevant, it must be relevant to something, and that would ordinarily be a diagnosis. We cannot collect relevant data without having a theory or diagnosis in mind, despite the suggestion that data collection and diagnosis can be carried out as distinct enterprises. Gorry has pointed out the limitations of programs based on decision-analysis alone and has alluded to the role of common-sense knowledge. Since then, others have proposed that artificial intelligence should be used to overcome this limitation.

For the most important portions of the medical history, then, current computer programs of substantial breadth, although they follow a systematic course, can at best ask rote questions and, for the most part, obtain low-worth, negative responses. Moreover, even when positive responses are obtained with these systems, clinicians have complained that they require further interpretation and that the information so obtained frequently turns out to be unimportant.
This type of performance is to be expected when the issue of relevance has been ignored. Several studies\(^7\)\(^-\)\(^9\) have confirmed that these automated history systems usually obtain "valid" information. But irrelevancy, not invalidity, is the source of the problem, and determining what is relevant in a situation requires common sense and general medical knowledge.

As far as the performance of judgment-like (as opposed to computational) processes is concerned, the everyday problems around Point A appear to be the ones with which computer programs are least useful. Indeed, it is doubtful that they have ever performed usefully here. To put it differently, for a computer program to perform successfully in this region, it would be necessary for the programmer to have developed a means of representing the universe of possibilities at this point. And at Point A this would entail a description of the world. Artificial-intelligence researchers who work with programs that attempt to "understand" natural language\(^40\) or to provide expert advice\(^11\)\(^-\)\(^14\) have devised ingenious and impressive systems. These systems appear to perform reasonably well, however, only when operating in so-called "microworlds," that is, within extremely small task domains that are near Point B and have been carefully structured and circumscribed, so that the developer can hope to provide the necessary a priori knowledge. Some of these Point B systems have been highly useful,\(^11\)\(^-\)\(^14\) and others\(^11\)\(^-\)\(^14\) have shown promise. (The differences between Point A and Point B applications appear to serve as a demarcation between the two types of research on artificial intelligence. The other and empirically the more successful applications are increasingly being described by Feigenbaum's term, "knowledge engineering."\(^4\)\(^-\)\(^4\)\(^4\) The idea that the real world at Point A can be reconstructed from a large set of microworlds has not received any support and appears on epistemological grounds to be in serious doubt.\(^2\)\(^4\) Before this attempt to put "microworlds" together to form a world becomes plausible, it will be world apart. This undertaking remains, and its success appears doubtful for the development of computer programs that can be said to behave intelligently. By "behave intelligently," I mean to operate successfully nearer to Point A than situations. The first of these approaches requires a means to represent common-sense knowledge in a computer program so that it can be applied to raw situations. So far, progress in accomplishing this goal has been scant, and the problem may, as has been suggested, be intractable.\(^2\)\(^4\) An alternative to storing such a priori knowledge would be to devise a program that is, in some important sense, capable of "learning." But the kind of learning needed includes the generalization of what has been learned. This does not seem to have been accomplished after some 30 years of effort.\(^4\)

There may be a third approach that would render the entire issue moot. This would be the discovery of yet unknown technologies — of hardware, organization, and software so radically different from anything presently known that a machine so constructed could in a sense be said to "live." By living in the world (surviving, competing, succeeding, failing, and so on) such a device might begin to acquire knowledge on its own. Whether the development of such a machine is possible and what we would call it if it were invented cannot be considered here.

The essential point is to distinguish between the nature of the situations existing at Point A, where the totality of the world must be confronted, and at Point B, where the task domain has been structured through previous human effort, an abstraction is available, and little common-sense knowledge may be required. My reason for emphasizing this difference is that the physician, who is equipped with medical knowledge through training and experience and with common sense through having lived in the world, can readily deal with the first situation. Physicians obviously perform well at Point B too, but in dealing with a growing number of specific (and computable) processes being identified in this region, physicians are unable to outperform a computer program. This trend is almost certain to continue.

It would be particularly unfortunate, therefore, if physicians mistakenly regarded their professional cognitive skills as lying in just those aspects of the diagnostic process that are, or may become, computational in nature and thereby minimized their own indispensable role in dealing with situations generally. This process is far beyond the present or foreseeable reach of the computer.\(^2\)\(^4\) It seems to be only in highly structured and formalized situations that we are presented with choices between human beings and machines; our more interesting problems deal with the selection of appropriate roles for human beings together with machines and for judgment together with computation.\(^4\) And in this selection it would not seem presumptuous to suggest that both be used for the tasks that they perform best, particularly since these tasks appear to be quite different.

I am indebted to David D. Sherertz for the development of the programs used in the studies of term usage, to Mark S. Tuttle for his assistance, and to Valerie Walters for the preparation of the manuscript.

REFERENCES


