31

An Evaluation of MYCIN's Advice

Victor L. Yu, Lawrence M. Fagan, Sharon Wraith Bennett, William J. Clancey, A. Carlisle Scott, John F. Hannigan, Robert L. Blum, Bruce G. Buchanan, and Stanley N. Cohen

A number of computer programs have been developed to assist physicians with diagnostic or treatment decisions, and many of them are potentially very useful tools. However, few systems have undergone evaluation by independent experts. We present here a comparison of the performance of MYCIN with the performance of clinicians. The task evaluated was the selection of antimicrobials for cases of acute infectious meningitis before the causative agent was identified.

MYCIN was originally developed in the domain of bacteremias and then expanded to include meningitis. Its task is a complicated one; it must decide whether and how to treat a patient, often in the absence of microbiological evidence. It must allow for the possibility that any important piece of information might be unknown or uncertain. In deciding which organisms should be covered by therapy, it must take into account specific clinical situations (e.g., trauma, neurosurgery), host factors (e.g., immunosuppression, age), and the possible presence of unusual pathogens (e.g., *F. tularensis* or *Candida nonalbicans*). In selecting optimal antimicrobial therapy to cover all of the most likely organisms, the system must consider antimicrobial factors (e.g., efficacy, organism susceptibility) and relative contraindications (e.g., patient allergies, poor response to prior therapy).

When knowledge about a new area of infectious disease is incorporated into MYCIN's knowledge base, the system's performance is evaluated

This chapter is an edited version of an article originally appearing in *Journal of the American Medical Association* 242: 1279–1282 (1979).

to show that its therapeutic regimens are as reliable as those that an infectious disease specialist would recommend. An evaluation of the system's ability to diagnose and treat patients with bacteremia yielded encouraging results (Yu et al., 1979a). The results of that study, however, were difficult to interpret because of the potential bias in an unblinded study and the disagreement among the infectious disease specialists as to the optimal therapeutic regimen for each of the test cases.

The current study design enabled us to compare MYCIN's performance with that of clinicians in a blinded fashion. This study involved a two-phase evaluation. In the first phase, several *prescribers*, including MYCIN, prescribed therapy for the test cases. In the second phase of the evaluation, prominent infectious disease specialists, the *evaluators*, assessed these prescriptions without knowing the identity of the prescribers or knowing that one of them was a computer program.[1]

31.1 Materials and Methods

Ten patients with infectious meningitis were selected by a physician who was not acquainted with MYCIN's methods or with its knowledge base pertaining to meningitis. All of the patients had been hospitalized at a county hospital affiliated with Stanford, were identified by retrospective chart review, and were diagnostically challenging. Two criteria for case selection ensured that the ten cases would be of diverse origin: there were to be no more than three cases of viral meningitis, and there was to be at least one case from each of four categories, tuberculous, fungal, viral, and bacterial (including at least one with positive gram stain of the cerebrospinal fluid and at least one with negative gram stain). A detailed clinical summary of each case was compiled. The summary included the history, physical examination, laboratory data, and the hospital course prior to therapeutic intervention. These summaries were used to run the MYCIN consultations. Only the information contained in the summaries was used as input to MYCIN, and no modifications were made to the program.

These same summaries were presented to five faculty members in the Division of Infectious Diseases in the Departments of Medicine and Pediatrics at Stanford University, to one senior postdoctoral fellow in infectious diseases, to one senior resident in medicine, and to one senior medical student. The resident and student had just completed a six-week rotation

[1]We wish to thank the following infectious diseases specialists who participated in this study: Donald Armstrong, M.D.; John E. Bennet, M.D.; Ralph D. Feigin, M.D.; Allan Lavetter, M.D.; Phillip J. Lerner, M.D.; George H. McCracken, Jr., M.D.; Thomas C. Merigan, M.D.; James J. Rahal, M.D.; Jack S. Remington, M.D.; William S. Robinson, M.D.; Penelope J. Shackelford, M.D.; Paul F. Wehrle, M.D.; and Anne S. Yeager, M.D.

in infectious diseases. None of these individuals was associated with the MYCIN project. The seven Stanford physicians and the medical student were asked to prescribe an antimicrobial therapy regimen for each case based on the information in the summary. If they chose not to prescribe antimicrobials, they were requested to specify which laboratory tests (if any) they would recommend for determining the infectious etiology. There were no restrictions concerning the use of textbooks or any other reference materials, nor were any time limits set for completion of the prescriptions.

Ten prescriptions were compiled for each case: that actually given to the patient by the treating physicians at the county hospital, the recommendation made by MYCIN, and the recommendations of the medical student and of the seven Stanford physicians. In the remainder of this chapter, MYCIN, the medical student, and the eight physicians will be referred to as *prescribers.*

The second phase of the evaluation involved eight infectious disease specialists at institutions other than Stanford, hereafter referred to as *evaluators,* who had published clinical reports dealing with the management of infectious meningitis. They were given the clinical summary and the set of ten prescriptions for each of the ten cases. The prescriptions were placed in random order and in a standardized format to disguise the identities of the individual prescribers. The evaluators were asked to make their own recommendations for each case and then to assess the ten prescriptions. The 100 prescriptions (10 each by 10 prescribers) were classified by each evaluator into the following categories:

Equivalent: the recommendation was identical to or equivalent to the evaluator's own recommendation (e.g., treatment of one patient with nafcillin was judged equivalent to the use of oxacillin);

Acceptable alternative: the recommendation was different from the evaluator's, but he considered it to be an acceptable alternative (e.g., the selection of ampicillin in one case was considered to be an acceptable alternative to penicillin);

Not acceptable: the evaluator found the recommendation unacceptable or inappropriate (e.g., the recommendation of chloramphenicol and ampicillin in one case was considered to be unacceptable by all evaluators who thought the patient had tuberculosis and who prescribed antituberculous therapy).

The 800 assessments (100 each by 8 evaluators) were analyzed as follows. A one-way analysis of variance (ANOVA) was used to analyze the overall difference effects between MYCIN and the other prescribers. The Tukey studentized range test was used to demonstrate individual differences between prescribers following attainment of significance. A similar analysis of variance was used to measure evaluator variability.

TABLE 31-1 Ratings of Antimicrobial Selection Based on Evaluator Rating and Etiologic Diagnosis

Prescribers	*No. (%) of items in which therapy was rated acceptable* by an evaluator (n = 80)*	*No. (%) of items in which therapy was rated acceptable* by majority of evaluators (n = 10)*	*No. of cases in which therapy failed to cover a treatable pathogen (n = 10)*
MYCIN	52 (65)	7 (70)	0
Faculty-1	50 (62.5)	5 (50)	1
Faculty-2	48 (60)	5 (50)	1
Infectious disease fellow	48 (60)	5 (50)	1
Faculty-3	46 (57.5)	4 (40)	0
Actual therapy	46 (57.5)	7 (70)	0
Faculty-4	44 (55)	5 (50)	0
Resident	36 (45)	3 (30)	1
Faculty-5	34 (42.5)	3 (30)	0
Student	24 (30)	1 (10)	3

*Therapy was classified as acceptable if an evaluator rated it as equivalent or as an acceptable alternative.

31.2 Results

The evaluators' ratings of each prescriber are shown in the second column of Table 31-1. Since there were 8 evaluators and 10 cases, each prescriber received 80 ratings from the evaluators. Sixty-five percent of MYCIN's prescriptions were rated as acceptable by the evaluators. The corresponding mean rating for the five faculty specialists was 55.5% (range, 42.5% to 62.5%). A significant difference was found among the prescribers; the hypothesis that each of the prescribers was rated equally by the evaluators is rejected (standard F test, $F = 3.29$ with 9 and 70 df; $p < 0.01$).

Consensus among evaluators was measured by determining the number of cases ($n = 10$) in which the prescriber received a rating of acceptable from the majority (five or more) of experts (third column of Table 31-1). Seventy percent of MYCIN's therapies were rated as acceptable by a majority of the evaluators. The corresponding mean ratings for the five faculty prescribers was 44% (range, 30% to 50%). MYCIN failed to win a rating of acceptable from the majority of evaluators in three cases. MYCIN prescribed penicillin for a case of meningococcal meningitis, as did four evaluators. However, four other evaluators prescribed penicillin with chloramphenicol as initial therapy before identification of the organism, and they rated MYCIN's therapy as not acceptable. MYCIN prescribed penicillin as treatment for group B *Streptococcus;* however, most evaluators selected ampicillin and gentamicin as initial therapy. MYCIN prescribed penicillin as treatment for *Listeria;* however, most evaluators used combinations of two drugs.

There were seven instances in which prescribers selected antimicrobial therapy that failed to cover a treatable pathogen (fourth column of Table 31-1). Five instances involved a case of tuberculous meningitis in which ineffective antibacterials (ampicillin, penicillin, and chloramphenicol) or no antimicrobials were prescribed. The other two instances included a case of meningococcal meningitis where one prescriber failed to prescribe any antimicrobial therapy and a case of cryptococcal meningitis where flucytosine was prescribed in inadequate dosage as the sole therapy.

31.3 Comment

In clinical medicine it may be difficult to define precisely what constitutes appropriate therapy. Our study used two criteria for judging the appropriateness of therapy. One was simply whether or not the prescribed therapy would be effective against the offending pathogen, which was ultimately identified (fourth column of Table 31-1). Using this criterion, five prescribers (MYCIN, three faculty prescribers, and the actual therapy given the patient) gave effective therapy for all ten cases. However, this was not the sole criterion, since failure to cover other likely pathogens and the hazards of overprescribing are not considered. The second criterion used was the judgment of eight independent authorities with expertise in the management of meningitis (second and third columns of Table 31-1). Using this criterion, MYCIN received a higher rating than any of the nine human prescribers.

This shows that MYCIN's capability in the selection of antimicrobials for meningitis compares favorably with the Stanford infectious disease specialists, who themselves represent a high standard of excellence. Three of the Stanford faculty physicians would have qualified as experts in the management of meningitis by the criteria used for the selection of the national evaluators.

Of the five prescribers who never failed to cover a treatable pathogen (fourth column of Table 31-1), MYCIN and the faculty prescribers were relatively efficient and selective as to choice and number of antibiotics prescribed. In contrast, while the actual therapy prescribed by the physicians caring for the patient never failed to cover a treatable pathogen, their therapeutic strategy was to prescribe several broad-spectrum antimicrobials. In eight cases, the physicians actually caring for the patient prescribed two or three antimicrobials; in six of these eight cases, one or no antimicrobial would have sufficed. Overprescribing of antimicrobials is not necessarily undesirable, since redundant or ineffective antimicrobial therapy can be discontinued after a pathogen has been identified. However, an optimal clinical strategy attempts to limit the number and spectrum of antimicrobials prescribed to minimize toxic effects of drugs and superin-

fection while selecting antimicrobials that will still cover the likely pathogens.

The primary limitation of our investigation is the small number of cases studied. This was a practical necessity, since we had to consider the time required for the evaluators to analyze 10 complex cases and rate 100 therapy recommendations. Although only 10 patient histories were used, the selection criteria provided for diagnostically diverse and challenging cases to evaluate MYCIN's accuracy. The selection of consecutive or random cases of meningitis admitted to the hospital might have yielded a limited spectrum of meningitis cases that would not have tested fully the capabilities of either MYCIN or the Stanford physicians. In addition to our evaluation, the program has undergone extensive testing involving several hundred cases of retrospective patient histories, prospective patient cases, and literature cases of meningitis. These have confirmed its competence in determining the likely identity of the pathogen, selecting an effective drug at an appropriate dosage, and recommending further diagnostic studies (a capability not evaluated in the current study).

Because of the diagnostic complexities of the test cases, unanimity in all eight ratings in an individual case was difficult to achieve. For example, in one case, although the majority of evaluators agreed with MYCIN's selection of antituberculous drugs for initial therapy, two evaluators did not and rated MYCIN's therapy as not acceptable. Six of the ten test cases had negative CSF smears for any organisms, so in these cases antimicrobial selection was made on a clinical basis. It is likely that if more routine cases had been selected, there would have been greater consensus among evaluators.

The techniques used by MYCIN are derived from a subfield of computer science known as artificial intelligence. It may be useful to analyze some of the factors that contributed to the program's strong performance. First, the knowledge base is extremely detailed and, for the domain of meningitis, is more comprehensive than that of most physicians. The knowledge base is derived from clinical experience of infectious disease specialists, supplemented by information gathered from several series of cases reported in the literature and from hundreds of actual cases in the medical records of three hospitals.

Second, the program is systematic in its approach to diagnosis. A popular maxim among physicians is "One has to think of the disease to recognize it." This is not a problem for the program; rare diseases are never "forgotten" once information about them has been added to the knowledge base, and risk factors for specific meningitides are systematically analyzed. For example, the duration of headache and other neurological symptoms for one week before hospital admission was a subtle clue in the diagnosis of tuberculous meningitis. The program does not overlook relevant data but also does not require complete and exact information about the patient. For example, in a case involving a patient with several complex medical

problems, the presence of purpura on physical examination was an important finding leading to the diagnosis of meningococcal meningitis. However, even if the purpura were absent or had been overlooked, MYCIN would have treated empirically for meningococcal meningitis on the basis of the patient's age and CSF analysis.

Third, since the program is based on the judgments of experienced clinicians, it reflects their understanding of the diagnostic importance of various findings. The program does not jump to conclusions on the basis of an isolated finding, nor does it neglect to ask for key pieces of information. Abnormal findings or test results are interpreted with respect to the clinical setting.

Finally, the system is up to date; frequent additions and modifications ensure its currentness. The meningitis knowledge base incorporates information from the most recent journal articles and the current experience of an infectious diseases division. Therapy selection and dosage calculations are derived from prescribing recommendations more recent than those in any textbook. (This was a factor in a case for which, at the time of this study, the recommendation of low-dose amphotericin B therapy combined with flucytosine was available only in recent issues of specialty journals.)

Because MYCIN compared favorably with infectious disease experts in this study, we believe that it could be a valuable resource for the practicing physician whose clinical experience for specific infectious diseases may be limited. The data demonstrate the program's reliability. However, further investigations in a clinical environment are warranted. Questions concerning the program's acceptability to practicing physicians and its impact on patient care, as well as issues of cost and legal implications, remain to be answered. Other capabilities of MYCIN that may assist the practicing physician include the following:

1. Identifying each of the potential pathogens with an estimate of its likelihood in causing the disease (Chapter 5).
2. Recommending antimicrobial dosages, considering weight, height, surface area, and renal function. Separate dosage regimens are given for the neonate, infant, child, and adult. Intrathecal dosage regimens are also given (Chapter 19).
3. Checking for contraindications of specific drugs, including pregnancy, liver disease, and age (Chapter 6).
4. Graphing predicted serum concentrations for aminoglycosides with relation to the expected minimal inhibitory concentration of the organism (Chapter 19).
5. Justifying its recommendation in response to queries by the physician (Chapter 18).

The methodology of the evaluation is of interest because it was developed in an attempt to analyze clinical decisions for which there is no clear right or wrong choice. Since most areas of medicine are characterized by a variety of acceptable approaches, even among experts, the technique used here may be generally useful in assessing the quality of decision making by other computer programs.