Antimicrobial Selection by a Computer; A Blinded Evaluation by Infectious Diseases Experts. Victor L. Yu, Lawrence M. Fagan, Sharon M. Wraith, et al., Sep 1979

Reprinted with permission from the Journal of the American Medical Association, Vol.242, No.12, pp.1279-1282.

Special Communications

HPP-78-21

Antimicrobial Selection by a Computer

A Blinded Evaluation by Infectious Diseases Experts

Victor L. Yu, MD; Lawrence M. Fagan; Sharon M. Wraith; William J. Clancey; A. Carlisle Scott, MS; John Hannigan, MS; Robert L. Blum, MD; Bruce G. Buchanan, PhD; Stanley N. Cohen, MD

• An evaluation of a computer-based consultation system called MYCIN was made. Eight independent evaluators with special expertise in the management of meningitis compared MYCIN's choice of antimicrobials with the choices of nine human prescribers for ten test cases of meningitis. MYCIN received an acceptability rating of 65% by the evaluators; the corresponding ratings for acceptability of the regimen prescribed by the five faculty specialists ranged from 42.5% to 62.5%. The system never failed to cover a treatable pathogen while demonstrating efficiency in minimizing the number of antimicrobials prescribed. The study design may be useful in assessing the performance of other computer-based clinical decision-making systems.

(*JAMA* 242:1279-1282, 1979)

DURING the last two decades, many computer programs have been developed to assist physicians in the diagnosis or treatment of a variety of medical disorders.[1] However, to our knowledge, the medical accuracy of these programs has not undergone clinical evaluation by independent experts. We present a comparison of the performance of a computer-based consultation system with the performance of clinicians. The task evaluated was the selection of antimicrobials for cases of infectious meningitis before the causative agent had been identified.

The computer program, MYCIN, provides advice for the diagnosis of diseases and the treatment of patients with infectious diseases.[2,3] During the last five years, MYCIN's extensive knowledge base and its therapy-selection process have been developed by infectious disease specialists and clinical pharmacologists and encoded by collaborating computer scientists. The program's task is a complicated one; it must decide whether and how to treat a patient, often in the absence of microbiological information. It must allow for the possibility that any important piece of information might be unknown, unreliable, or uncertain. In deciding which organisms should be covered by therapy, MYCIN takes into account the specific clinical situations (eg, trauma, neurosurgery), host factors (eg, age, immunosuppression), and the possible presence of unusual pathogens (eg, *Francisella tularensis*, *Candida* non-*albicans*). In selecting antimicrobial therapy, the system considers antimicrobial factors (eg, organism susceptibility, synergistic combinations) and relative contraindications (eg, patient allergies, poor response to prior therapy).

When knowlege about a new area of infectious disease is incorporated into MYCIN's knowledge base, the system's performance is evaluated to determine whether its therapeutic regimens are as reliable as the regimens that an infectious diseases specialist would recommend. An evaluation of the system's ability to diagnose and treat patients with bacteremia yielded encouraging results.[4] The results of that study, however, were difficult to interpret because of the potential bias occurring in an unblinded study and the disagreement among the infectious diseases specialists as to the optimal therapeutic regimen for each of the test cases.

Our study design enables us to compare MYCIN's performance with that of clinicians in a blinded fashion. This study design involved a two-

From the Division of Clinical Pharmacology and Infectious Diseases, Department of Medicine, Stanford University Medical Center, Stanford, Calif. Dr Yu is now with the University of Pittsburgh School of Medicine.

Reprint requests to University of Pittsburgh School of Medicine, Department of Medicine, 968 Scaife Hall, Pittsburgh, PA 15261 (Dr Yu).

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.

Materials and Methods

Ten patients with infectious meningitis were selected by a physician who was not acquainted with MYCIN's methods or knowledge base pertaining to meningitis. All of the patients had been hospitalized at a county hospital affiliated with Stanford University, 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. A detailed summary of each case was compiled. The summary included the history, physical examination, laboratory data, and the hospital course before therapeutic intervention. These summaries were used to run the MYCIN consultations. Only the information contained in the summaries was used as input to MYCIN.

An abbreviated summary of the ten test cases is presented.

Report of Cases

Case 1.—A 26-year-old man had a one-day history of chills and headache. Nuchal rigidity was present. A lumbar puncture showed a protein level of 95 mg/dL and a WBC count of 235/cu mm, with 98% polymorphonuclear leukocytes. The patient had an allergy to penicillin as manifested by acute angioneurotic edema.

Case 2.—A 3-year-old girl had a one-day history of fever and headache. Myringotomy tubes were inserted one year previously for chronic otitis media. Nuchal rigidity and petechiae were present. A lumbar puncture disclosed a protein level of 25 mg/dL and a WBC count of 110/cu mm, with 96% polymorphonuclear leukocytes.

Case 3.—A 1-day-old term male infant was noted to have nuchal rigidity. Gram-positive cocci in pairs, clusters, and chains were seen. The prenatal course was complicated by cervical gonorrhea in the mother five months previously.

Case 4.—A 44-year-old woman with hyperthyroidism had a one-day history of fever and headache. Physical examination results showed a comatose woman with purpura on her extremities. A lumbar puncture showed a protein level of 400 mg/dL and a WBC count of 3,800/cu mm, with 95% polymorphonuclear leukocytes.

Case 5.—A 24-year-old woman had facial swelling, neck stiffness, and headache. She had squeezed a boil on her face two days previously and had noted subsequent pain and swelling in that area. A lumbar puncture showed a protein level of 200 mg/dL and a WBC count of 3,000/cu mm, with 95% polymorphonuclear leukocytes.

Case 6.—A 42-year-old woman had a three-day history of headache, nausea, and fever. Physical examination results showed nuchal rigidity and coma. A lumbar puncture showed a protein level of 580 mg/dL and a WBC count of 374/cu mm, with 66% polymorphonuclear leukocytes.

Case 7.—A 6-year-old boy underwent neurosurgery. Three days later fever and neck stiffness occurred. The left tympanic membrane was bloody. A lumbar puncture showed Gram positive cocci in pairs.

Case 8.—A 2-day-old male infant, the product of 30 weeks' gestation, was noted to have nuchal rigidity. Gram's stain showed Gram-positive rods.

Case 9.—A 42-year-old woman had a one-week history of headache. A lumbar puncture showed a protein level of 265 mg/dL and a WBC count of 810/cu mm, with 91% lymphocytes.

Case 10.—Meningitis developed while a 73-year-old woman with rheumatoid arthritis was in the hospital. An India ink stain showed rare budding yeasts.

The pathogens ultimately isolated from the aforementioned cases were (case 1) virus, (2) virus, (3) *Streptococcus* group B, (4) *Neisseria meningitidis*, (5) *Staphylococcus aureus*, (6) *Listeria monocytogenes*, (7) *Str pneumoniae*, (8) *L monocytogenes*, (9) *Mycobacterium tuberculosis*, and (10) *Cryptococcus neoformans*.

These 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 disease, to one senior resident in medicine, and to one senior medical student. The resident and student had just completed a six-week rotation in infectious diseases. None of these people were associated with the MYCIN project. These seven physicians and the student were asked to prescribe an antimicrobial therapy regimen for each case on the basis of the information in the summary. There were no restrictions concerning the use of textbooks or any other reference materials.

Ten prescriptions were compiled for each case: that actually prescribed by the treating physician at the county hospital, that recommended by MYCIN and those recommended by the medical student and the seven Stanford physicians. In the remainder of our report, MYCIN, the student, and the eight physicians will be referred to as prescribers.

The second phase of the evaluation involved eight infectious diseases specialists at institutions other than Stanford, hereafter referred to as "evaluators," who had published clinical reports dealing with the management of 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 (ten each by ten prescribers) were classified by each evaluator into the following categories: (1) Equivalent—the recommendation was identical to or equivalent to the evaluator's own recommendation (eg, selection of nafcillin in case 5 was judged equivalent to oxacillin therapy); (2) Acceptable alternative—the recommendation was different, but the evaluator considered it to be an acceptable alternative (eg, the selection of ampicillin in case 4 was considered an acceptable alternative to penicillin); and (3) Not acceptable—the evaluator found the recommendation inappropriate (eg, the recommendation of chloramphenicol and ampicillin for case 9 was considered unacceptable by all evaluators who prescribed antituberculosis therapy).

The 800 assessments (100 each by eight evaluators) were analyzed as follows. A one-way analysis of variance 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 the attainment of significance.[5] A similar analysis of variance was used to measure evaluator variability.

Results

The evaluators' ratings for each prescriber are shown in the Table, column 1. Since there were eight evaluators and ten 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, 62.5% to 42.5%). 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=3.2865 with 9 and 70 df;

Ratings of Antimicrobial Selection Based on Evaluator Rating and Etiologic Diagnosis			
Prescribers	No. (%) of Cases in Which Therapy Rated Acceptable* by an Evaluator (n=80)	No. (%) of Cases in Which Therapy 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 diseases 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 acceptable alternative.

$P<.01$).

Consensus among evaluators was measured by determining the number of cases (n=10) in which the prescriber received a rating of acceptable therapy by the majority (five or more) of experts (Table, column 2). 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 by the majority of evaluators in three cases. MYCIN prescribed penicillin for case 4 (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 "not acceptable." MYCIN prescribed penicillin for case 3 (group B *Streptococcus*); however, most evaluators selected ampicillin and gentamicin as initial therapy. MYCIN prescribed penicillin for case 6 (*Listeria*); however, most evaluators used various combinations of two drugs.

There were seven instances in which prescribers selected antimicrobial therapy that failed to cover a treatable pathogen (Table, column 3). Five instances involved a case of tuberculosis meningitis (case 9) in which ineffective antibacterials (ampicillin, penicillin, and chloramphenicol) or no antimicrobials were prescribed. The other two instances included a case of meningococcal meningitis (case 4) where one prescriber failed to prescribe any antimicrobial therapy and a case of cryptococcal meningitis (case 10) where flucytosine was prescribed in inadequate dosage as sole therapy.

Comment

In clinical medicine it may be difficult to define precisely as to what constitutes appropriate therapy. Our study used two criteria for judging appropriateness of therapy. One was simply whether the prescribed therapy would be effective against the offending pathogen, which was ultimately identified (Table, column 3). 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 (Table, columns 1 and 2). Using this criterion, MYCIN received a higher rating than any of the ten human prescribers.

This shows that MYCIN's capability in selection of antimicrobials for meningitis compares favorably with the Stanford infectious diseases 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 using the criteria used for selection of the national evaluators.

Of the five prescribers who never failed to cover a treatable pathogen (Table, column 3), MYCIN and the faculty prescribers were relatively efficient and selective in their 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 superinfection 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 ten complex cases and rate 100 therapy recommendations. Although only ten 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 had undergone extensive testing involving several hundred cases of retrospective patient histories, prospective patient cases, and literature cases of meningitis with confirmation of 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 our 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 case 9, although the majority of evaluators agreed with MYCIN's selection of antituberculosis drugs for initial therapy, two evaluators did not and rated MYCIN's therapy "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 popularly known as "artificial intelligence." The detailed workings and structure of MYCIN are beyond the scope of our report and have been reported in detail elsewhere.[2,5] It may be useful, however, 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 probably 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 that "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, in case 9, 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 case 4 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 was 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 judgment 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, for frequent additions and modifications ensure its currentness. The meningitis knowledge base incorporates information from the most recent journal articles and current experience of an infectious diseases division. Therapy selection and dosage calculations are derived from prescribing recommendations more recent than any textbook. This was a factor in case 10, when 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 diseases experts in this study, we believe that it will be a valuable resource to 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, its impact on patient care, plus issues of cost and legal implications remain to be answered. Other capabilities of MYCIN that may assist the practicing physician include the following: (1) Identification of potential pathogens with an estimate of its likelihood in causing the disease. (2) Recommendation of antimicrobial doses 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. (3) Checking for contraindications of specific drugs, including pregnancy, liver disease, and age. (4) Graphing of predicted serum concentrations for aminoglycosides with relation to the expected minimal inhibitory concentration of the organism.[6] (5) Justification of its recommendation in response to queries by the physician.[7]

The evaluation methodology is of interest because it was implemented in an attempt to analyze clinical decisions in which there is not clearly a right or wrong answer. 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.

This project was supported by the Bureau of Health Services Research and Evaluation research grant No. HS01544.

The following infectious diseases specialists participated in this study: Donald Armstrong, MD; John E. Bennet, MD; Ralph D. Feigin, MD; Allan Lavetter, MD; Phillip J. Lerner, MD; George H. McCracken, Jr, MD; Thomas C. Merigan, MD; James J. Rahal, MD; William S. Robinson, MD; Jack S. Remington, MD; Penelope J. Shackelford, MD; Paul F. Wehrle, MD; and Anne S. Yeager, MD. Dana Ludwig, MD, assisted in the program development, and John Hannigan, MS, performed the statistical analyses for this study.

Members of the MYCIN Project included Randall Davis, PhD; William J. VanMelle, MS; Jan Aikens, MS; Edward H. Shortliffe, MD, PhD; and Stanton G. Axline, MD.

Nonproprietary Names and Trademarks of Drugs

Amphotericin B—*Fungizone.*
Flucytosine—*Ancobon.*

References

1. Wardle A, Wardle L: Computer aided diagnosis: A review of research. *Meth Inform Med* 17:15-28, 1978.

2. Shortliffe EH: *Computer-Based Medical Consultation: MYCIN.* New York, American Elsevier, 1976.

3. Wraith SM, Aikins J, Buchanan BG, et al: Computerized consultation for selection of antimicrobial therapy. *Am J Hosp Pharm* 33:1304-1308, 1976.

4. Buchanan BG, Davis R, Yu VL, et al: Rule-based medical decision making by computer, in Shires DB, Wolf H (eds): *MEDINFO 77.* Amsterdam, North-Holland Publishing Co, 1977, pp 147-150.

5. Scheffe H: *Analysis of Variance.* New York, John Wiley & Sons Inc, 1959, p 73.

6. Wraith SM: Antimicrobial therapy selection with pharmacokinetic evaluation by computerized consultation (abstract). Read before the 12th Annual American Society of Hospital Pharmacists Midyear Clinical Meeting, Atlanta, 1977.

7. Scott AC, Clancey WJ, Davis R, et al: Explanation capabilities of knowledge-based production system. *Am J Computational Linguistics,* microfiche 62, 1977.

Printed and Published in the United States of America