Modeling Empiric Antibiotic Therapy Evaluation of QID

Homer Warner, Jr*, MS, Larry Reimer+, MD, David Suvinier+, MD,
Lili Li@, MS, Marilyn Nelson@, RN

First Consulting Group*
Division of Infectious Disease+, University of Utah Medical Center
@Sunquest Information Systems, Inc.
Salt Lake City, Utah

ABSTRACT

At AMIA 1997, we reported on the design and development of a new computer-based tool, called QID, for empiric antibiotic decision support. QID was designed to help physicians identify the antibiotic regimens with the highest probability of covering the pathogens that are most likely to be present in individual patients. QID creates a list of antibiotics, ordered by potential benefit in treatment, for a patient with a suspected infection before culture results are available. Since our initial publication, a "before and after" study has been done using 20 internal medicine residents and the same number of internal medicine attendings. In order to test the hypothesis that physician's would make more appropriate empiric antibiotic choices with the aid of QID, we chose University of Utah physicians and had each evaluate four infectious disease cases that were abstracted from medical record infectious disease cases. Immediately following their initial review and determination of antibiotic therapy for each case, the study participants were presented with QID's antibiotic recommendations on the same case to see if this information would change their initial drug regimen. The tool was shown to have a greater impact on the most difficult cases but statistically improved scores overall (p<.001). Details of our study design and results are presented.

INTRODUCTION

Physicians frequently need to prescribe antibiotic therapy before bacterial culture results and antibiotic susceptibility tests are available. Empiric therapy should be chosen on the basis of recent and local information about the most likely pathogens, the risks associated with making the wrong choice, and the potential gain in health quality for each antibiotic. Additional and secondary factors are drug toxicity and cost. Each of these factors is represented in our QID decision support model and described in our initial paper [1].

Inspired by early concepts in MYCIN [2] and by the success of the LDS hospital Antibiotic Assistant [3], our goal in building QID was to:

1) build a modular system that would function independent of any particular host information systems,
2) use local knowledge about disease/organism prevalence,
3) use probabilistic reasoning , Iliad [4], to calculate a differential diagnosis,
4) take advantage of indigenous susceptibility patterns, and
5) account for the risk of not treating (using a construct for measuring "Good Days of life Saved" or GDS). GDS is similar to other "net benefit" constructs described in the literature [5-6].

The QID program first calculates a differential diagnosis from an infectious disease knowledge base that we have built over the last three years which runs on Iliad's inference engine. QID then calculates the maximum GDS for each of the most likely disease/organisms (Prob(Disease)) assuming optimal antibiotic coverage. Using local antibiogram epidemiology from the last 15 months, QID multiplies the effectiveness (% susceptible) of each antibiotic by the maximum GDS score for each of the most likely disease/organisms. This algorithm generates a list of antibiotics ordered by GDS and displayed along with the toxicity and cost/24 hours of treatment for each drug. Stated another way the formula: Antibiotic GDS = Sum across most likely diseases of (Prob(Disease) x Optimal GDS score x Antibiotic Susceptibility) produces a list of antibiotics that cover the largest fraction of potential GDS.

In order to validate this decision support model and before attempting to further refine and implement QID in a clinical setting we wanted to first measure the value of the information content produced by the program. Our goal in designing QID was not to compete with but aid physicians in making better

1091—8280/99/$5.00

empiric antibiotic choices. This goal influenced our study design and will eventually guide the direction we take future development.

METHODS

Evans et.al. in 1994 reported a 17% (77% to 94%) improvement in empiric treatment effectiveness against eventual pathogens, as physicians used their Antibiotic Assistant in a study performed at LDS Hospital [2]. Expecting similar gains in performance, we used Evan's baseline numbers in calculating our study sample size. However, in addition to judging how effectively suggested antibiotic regimens did in covering isolated pathogen(s), as was done in the Evan's study, we also used infectious disease specialists to judge how well each study physician's antibiotics compared to what the infectious disease (ID) specialists would have done with the same information. This additional evaluation approach seemed appropriate given our goal to raise overall physician performance to the level of an infectious disease specialist through the use of a decision support tool. It also required agreement between the judging specialists regarding optimal antibiotic therapy.

As an experimental design we chose a "before and after" study where the same physicians would evaluate similar randomly selected infectious disease cases without the decision support tool and then again immediately afterwards with the aid of the information generated by QID. This approach controlled for inter-physician variability in prescribing patterns and eliminated the temporal problems introduced when there is long interval between measurements. Patient charts (inpatients only) were selected from the University of Utah Medical Center that met the following inclusion criteria:

a) The discharge diagnoses had to include an infection from the list shown in Figure 1, since the knowledge base used for QID only included hospital acquired diseases.
b) Only charts for patients discharged between 1992 and 1997 were included since QID is built from a knowledge base of the most recent antibiotics and sensitivity data.

Some charts were eliminated from the study because the patient chart was illegible or because there was no culture data in the chart. Many charts were abstracted and used for testing the model over a two-year period but nine separate charts were abstracted especially for the study. Since we decided not to use surgeons in the study (due to difficulty in getting them to participate), the wound infection cases were eliminated and we ended up using three bacteremia cases, three pneumonia cases, two UTI/cystitis cases, and one meningitis case.

Diseases of interest:	% Occurrence
1. Wound Infection	27%
2. UTI/Pyelonephritis	21%
3. Bacteremia	16%
4. Pneumonia	14%
5. Endocarditis	12%
6. Peritonitis	6%
7. Encephalitis	2%
8. Meningitis	2%

Figure 1
(Distribution of Nosocomial diseases, 1997, University of Utah Medical Center)

Two groups of 'non-infectious disease' physicians from the University of Utah were recruited into the study – 20 internal medicine residents (3rd and 4th year) and 20 internal medicine attendings. Each physician was paid $50 for participating.

Each participant was presented with abstracts from four of the nine abstracted cased and asked to analyze the meningitis case and one of each of the randomly assigned pneumonia, bacteremia, and cystitis cases. Each was asked to carefully study the case and note their suggested antibiotic regimen on the bottom of the page. This completed, each was then given QID's recommendations printed on paper (partially shown in Figure 2) along with a short questionnaire. Participants were asked to make any modifications to their initial antibiotic regimen if influenced to do so by this new information. Additional questions on the questionnaire attempted to collect feedback on influencing factors regarding reasons for change. A total of 160 cases were collected during the study that spanned 30 days. Of the 160, twenty-one were eliminated either because the judges could not read the handwriting or because the participant didn't follow directions in filling one of the two forms.

Computer Generated Drug List

Class	Antibiotic	% Coverage	Tox	Cost/24
Fluoroquinolone	Ofloxacin	80%	2	$5.28
TMP/SMX	Sulfa & Trimeth	72%	3	$2.76
Vacomycin	Vacomycin	61%	4	$12.80
Lincomycin	Clindamycin	45%	2	$27.00

Figure 2 (Part of QID output)

Physician antibiotic "before and after" choices were judged by two ID specialists (gold standard). These experts evaluated each of the 139 cases independently using the same scoring mechanism. A scale from 0 to 100% was used as a distance measure

against the gold standard. A score of 0 meant that the antibiotics chosen were completely ineffective against the mostly likely pathogens while a score of 100 meant that the chosen regimen was expected to have 100% coverage against the most likely pathogens. Scores in between indicated partial coverage.

RESULTS

There are two sets of results. The first analysis measures how well study participants' choice of antibiotics matched with the empiric choices of the infectious disease specialists. The second analysis measures how well each participant did in covering the eventual pathogen as isolated by the laboratory. The same ID specialists judged both analyses. Each infectious disease specialist (rater 1 and rater 2) scored all 139 cases independently.

The before and after results of the first analysis in shown in Figure 3 as a stratified distribution across cases of before and after scores as the mean of both raters. Rater 1 scored the participants performance without QID (termed "before") at an average score of 68.3% and participant's performance using the output from QID (termed "after") at an average score of 77%, an 8.7% increase. By comparison, rater 2 scored the average "before" performance at 88.2% and the "after" performance at 93.1%, for an increase of 4.9%. When combining the two raters the mean difference in scores from "before" to "after" was 6.8% (78.2% to 85%).

Score Percent	Mean Before	Mean After
0%	3	3
10%	2	0
20%	4	1
30%	5	1
40%	3	3
50%	11	6
60%	11	7
70%	12	10
80%	22	21
90%	24	32
100%	40	58

Figure 3

Since the distributions of the scores for both raters appears to be skewed we could not assume normality. Therefore, we used a Wilcoxin non-parametric statistic along with a paired t-test to test the null hypothesis that there is no difference in physician performance with or without the decision support aid. Both the paired t-test and Wilcoxin produced p-values < .001. The number of positive differences was 57, the number of tied ranks was 77 (i.e., either the before score was 100% or the after score did not improve over the before score). There were also 5 negative differences where the participants score decreased by revising his/her initial antibiotic choice(s). The antibiotic(s) that QID recommended covered the isolated pathogen(s) in 8 of 9 test cases as judged by our two raters. In the one case QID missed, the organism (*E. cloacae*) was not included in the program's knowledge base. This is the most likely explanation for the 5 negative differences.

The results of the second analysis (against isolated pathogen), are shown in Figure 4. The overall "before" and "after" means for the second analysis were 66.2% and 75.2% respectively – an average improvement of 9.0%. The Wilcoxin rank-sum produced 7 negative, 30 positive, and 102 tied ranks. Both parametric (paired T-test) and non-parametric tests showed significant improvement at $p<.001$. As in the first analysis, the greatest improvement in the second analysis (21.5%, $p=.006$) was found in the most difficult case (i.e., meningitis).

Score Percent	Mean Before	Mean After
0%	7	6
10%	19	5
20%	14	18
30%	0	0
40%	0	0
50%	8	4
60%	4	3
70%	0	2
80%	10	10
90%	44	37
100%	33	54

Figure 4

When considering the use of human judges as a "gold standard" it is important to measure their degree of agreement. This was done using the Kappa statistic which focuses attention on the cells along the diagonal of a RxC contingency table (Figure 5). The numbers along the diagonal represent assignments agreed upon by the two raters. The numbers in each of the cells in Figure 5 represent counts of how <u>often</u> each rater scored before/after improvement. The five cases that had lower after than before scores were counted as <u>no</u> improvement.

	Rater 2		
Rater 1	No improvement	Improvement	Total
No improvement	65	22	87
Improvement	22	30	52
Total	87	52	139

Figure 5

This data yields a Kappa of .33 with $p < .001$, thus rejecting the null hypothesis that agreement between the two raters was no better than that predicted by chance. Where there is disagreement, as measured by the numbers off the diagonal, there is perfect symmetry implying random variation or variation due to chance (in other words the disagreement is not systematic). In 30 of the 139 cases both raters agreed that the participant improved using the decision support tool. Moreover, 74 of the 139 cases or 53% improved using the tool at judged by at least one of the raters.

When eliminating the 16 cases where there was no room for improvement (i.e., cases with a "before" score of 100%) the mean difference in before and after scores improved from 6.8% to 7.7%. The before and after differences were most striking when we isolated the most difficult or rare cases for analysis. Most would expect relatively good before scores for frequently seen cases (e.g., urinary tract infection (UTI). In our study, UTI's produced only small percentage improvement gains, but when isolating the 31 rarely seen meningitis cases the average difference in scores was 19.2% (66.3% to 85.5%).

Since we had 20 residents and 20 attendings (who were not ID specialists) in the study, we also looked at empiric antibiotic performance by education level, independent of any decision support. To do this we compared how well each group scored independent of QID (the "before" score only). Using the Mann Whitney U statistic (a non-parametric equivalent to the two sample t-test) the mean ranks were almost identical between the two groups (67.07 vs. 66.88) with p=.977. This suggests that education level (R3/R4 vs. attending) for our test group had no bearing on the effectiveness of empiric antibiotic prescribing.

Using a five point Likert scale, with 0 meaning "not influenced at all" and 5 meaning "highly influenced," three additional follow-up questions were asked of study participants for cases where they changed their treatment regimen after looking at QID's recommendations:

1) What was the impact of the GDS scoring metric on your decision making?
2) How much did knowing about drug toxicity play a part in your changing your drug regimen?
3) How much did knowing about drug cost per 24 hours influence your decision?

Of the 42% responders to these questions, 93% (54 out of 58) indicated that GDS strongly influenced their decision to change antibiotics whereas only 16% (9 out of 58) said that either drug toxicity or cost played a significant role in their antibiotic choice.

DISCUSSION

A statistically significant difference in prescribing effectiveness is encouraging but we remain concerned about the reliability of our gold standard as well as how to interpret the clinical significance of these findings. A concern we had from the beginning of the study was how to establish a reliable "gold standard" in the face of considerable antibiotic prescribing variability, even among experts [7]. A Kappa of .32, while statistically significant, is still low. However, we were unable to justify using more raters and unable to come up with a more suitable gold standard for our purposes. A second concern is that the physician will feel that their return on investment is low. In other words, they may feel that a 10% or 20% improvement in empiric decision-making is not worth the time they must invest to enter patient data into the program.

The most challenging problem facing all computer-based decision support tools is data collection. Most of the data used by QID in the diagnostic phase is not available in electronic form in existing databases. Therefore, the physician must manually enter the data. This constraint will likely relegate QID use to difficult cases, but an area of the tool's greatest impact as indicated by our results. A second challenge is in validating QID's content both in terms of what has been defined and what may be missing. We are in the process of doing sensitivity analyses on our morbidity/mortality tables but there is further work required in validating the sensitivities and specificities used in the program's diagnostic front-end as well as in overall testing of the tool.

Once the content has been validated, there are a number of very challenging "human factor" issues that include logistical (work flow), mechanical (data input mechanisms), psychological (physician comfort level and confidence in the tool), legal (will I get sued for using or not using the tool) and regulatory (FDA and software oversight) issues that must be dealt with in order to achieve market penetration. A final challenge lies in implementation approach and

marketing focus. Should implementers attempt to integrate QID into existing applications (like physician order entry) where it can eventually play a more "active" role in decision support (e.g., Antibiotic Assistant at LDS hospital [2]) or is it better placed in outpatient settings where there is less infectious disease decision support of any kind so it can be used as a consultant on difficult cases.

CONCLUSION

The optimal use of antibiotics in the empiric phase of treatment involves the decision to use antibiotics at all, the choice of which antibiotic(s) to use, and deciding how to administer the drug (e.g., dosing) once a specific regimen has been chosen. The QID decision support model provides information to aid the physician in forming a differential diagnosis and in providing a rationale for making a treatment decision that takes into account the "net benefit" of choosing one antibiotic or antibiotic combination over another. The "good days of life" (GDS) construct used in this model uses an estimate of mortality and both acute and long term morbidity in its calculation of a score that yields a quantitative coverage estimate of the most likely pathogens. Similar approaches have been used in other decision support models in order to provide appropriate weight and utility to various treatment options [6-7].

We believe that a valid construct for empiric antibiotic therapy must take into account the risk of misdiagnosing as well as provide a quantitative "goodness of fit" measure. Our use of a diagnostic front-end (Iliad) combined with a treatment expert system back-end has the potential of making an important contribution to clinical practice.

It is tempting to believe that models can be built that will answer all questions and control for potential confounding variables. This does not appear possible in most clinical settings since too much uncertainty exists. However, since our goal is not to replace the physician but to improve their decision-making capabilities, the results of our "laboratory" experiment and the performance of the initial prototype are encouraging. Further testing, refinement and evaluation approaches are required in taking the next step in understanding the full clinical utility of our model.

References

1. Warner H Jr, Blue SR, Sorenson D, Reimer L, Li L, Nelson M, Barton M, Warner H. New computer-based tools for empiric antibiotic decision support. Proc AMIA Ann Fall Symp 1997;:238-4.
2. Shortliffe EH, Axline SG, Bunan BG, Merigan TC, Cohen. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Comput Biomed Res 1973;6: 544-60.
3. Evans RS, Classen DC, Pestotnik SL, Lundsgaarde HP, Burke JP. Improving empiric antibiotic selection using computer decision support. Arch Intern Med 1994;154:878-884.
4. Warner HR, Warner HR Jr., Bouhaddou O., et.al. Iliad as an expert consultant to teach differential diagnosis. SCAMC 1988, Washington, DC, IEEE Computer Society Press, pp 371-376.
5. Warner HR, Computer-assisted medical decision-making. Academic Press, 1979, pp 143-148.
6. Kassirer JP. The principles of clinical decision making: an introduction to decision analysis. Yale J Biol Med 1976;49:149-64.
7. Buckwold, FJ, Ronald, AR. Antimicrobial misuse - effects and suggestions for control. The British Society for Antimicrobial Chemotherapy. 1979, 5, 129-136.