MAKING THE MOST OF RECONSIDER: An Evaluation of Input Strategies

Stuart Nelson, MD, Shane Hoffman, Hemanth Kanekal, and Andre Varna, MD

Depts. of Internal Medicine and Community Medicine,
State University of New York at Stony Brook
Stony Brook, New York

ABSTRACT

During the performance of an evaluation of a diagnostic prompting program, RECONSIDER, on consecutive first admissions at a tertiary care hospital, the question arose if there were input strategies which insured better performance of the program. The program generates differential diagnoses from patient findings by comparing these findings to disease descriptions in its library. The patient data may be entered under a general or under specific contexts. All positive findings were abstracted for each case and used as a source of input data. Each case was entered by different users who were free to choose which items to enter and under what context to enter them. Each program run generated a ranked differential diagnosis list, and the rank number of the actual diagnosis was defined as the score for that run. Nonparametric, multiple regression analysis, and defined strategy evaluations of the multiple versions of each case were performed. An optimal strategy for input remains to be determined; the ability to use various options seems to be important.

INTRODUCTION

RECONSIDER is a diagnostic prompting program, that is, a computer program which is intended to provide a clinician with a list of possible diagnoses to be considered in an individual case. The program, as previously described, provides a list of possible diagnoses after being presented with terms describing the case. A sample run is demonstrated in Appendix A.

The program's knowledge base consists of the descriptions of 3262 diseases in the book Current Medical Information and Terminology, 4th Ed., and a synonym dictionary. Each disease described is assigned to one, or sometimes two, organ systems. The disease description is broken into contexts, e.g., signs, symptoms, etiology, laboratory, under which the appropriate portions of the disease description occurs. The user enters patient findings (terms), either as individual words or as phrases. The program searches for these terms and their synonyms in the disease descriptions. The user may confine the search on a given term to a specific organ system or to specific contexts of the disease description as desired. These contexts vary from the general ("all") to the specific ("symptoms", "signs", "laboratory", etc.). A complete list of contexts is shown in Table 1.

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTENTIAL CHOICES FOR RECONSIDER</strong></td>
</tr>
<tr>
<td><strong>Contexts</strong></td>
</tr>
<tr>
<td>all</td>
</tr>
<tr>
<td>symptoms or signs</td>
</tr>
<tr>
<td>symptoms</td>
</tr>
<tr>
<td>signs</td>
</tr>
<tr>
<td>laboratory</td>
</tr>
<tr>
<td>X-ray</td>
</tr>
<tr>
<td>pathology</td>
</tr>
<tr>
<td>etiology</td>
</tr>
<tr>
<td>complications</td>
</tr>
<tr>
<td>additional terms</td>
</tr>
<tr>
<td>special sense</td>
</tr>
</tbody>
</table>

A "selectivity score" is assigned for each term such that if the term were to occur in the description of every disease its value would be 0 and if in only one disease the value would be 1. Every disease matching on at least one term is considered a possible diagnosis; the list of possible diagnoses is given in rank according to the sum of the selectivity scores.

A clinical evaluation of the RECONSIDER program was carried out over consecutive first admissions to the medical service of the University Hospital at the State University of New York at Stony Brook. All positive findings abstracted from the patients' chart at the time of admission was sent to the University of California at San Francisco, where the abstract was used as a source of input data for the program. Seven persons (enterers) each created one or more versions of each case, selecting which items were to be entered and under what context each item should be entered. For each version, a list of 40 possible diagnoses was generated and returned to the evaluator. After the patient was discharged from the hospital, the chart of that patient was reviewed to ascertain what diagnoses were proven. For each differential, the ordinal position of the correct diagnosis, the coded identity of the enterer, and the number of items entered under each context was recorded.

It was observed that the performance on a given case was quite variable between versions,
thus the question arose if there were any factors which predicted the better performance of the program. This question took the form of two separate hypotheses: first, that the individual entering the terms might have an effect upon performance (by reason of training, clinical experience, or prior experience with the program); and second, that the choice of context under which items were entered might systematically affect performance.

The investigation of these hypotheses was explored in three different ways. The results of the previously performed clinical trial of the program were evaluated using both nonparametric and parametric statistical methods, and trials of the program using defined strategies for choosing contexts were run using input data from the same cases.

**METHODS.**

**Nonparametric Evaluation.**

Cases with a single diagnosis proven during the hospitalization were used to investigate both the variability between the enterers and the effects of entering terms under different contexts. The ordinal ranking of the correct diagnosis was used as the score; if all versions of the case had achieved the same score, the case was disregarded. In all, there were 39 cases available for analysis, with seven or more versions of each case.

**Enterers.**

A pairwise comparison of the performance of each of the enterers was performed using a Wilcoxon signed rank test for symmetry around zero. Where an enterer had completed more than one version of a given case, one version was chosen at random. All 21 possible pairs were analyzed in this way.

**Context Variables.**

A similar strategy was adopted for evaluation of the choice of context and terms. Variables representing the number of items entered under each context were established, a single variable was used for all contexts which represented single organ systems. A last variable, that of the total number of items entered, was also used. Context variables were compared for each user pair and standardized for outcome. In this manner the addition or subtraction of individual items from a given context could be associated with a better or worse outcome. The Wilcoxon test was stratified by case according to the method described by Varma and a normal distribution test was computed for each variable. The stratified Wilcoxon test was used to minimize the effect of the differences between cases, and to concentrate the comparison to that of differences between versions.

**Parametric Evaluation.**

A stepwise multiple regression analysis was performed using the number of terms entered under each context, the total number of terms entered, and a dummy variable representing the enterer as independent variables. As dependent variables four different functions were used. The functions are shown in Table 2. All versions of the 42 cases with a single proven diagnosis were used, a total of 326 versions.

**TABLE 2**

**REGRESSION ANALYSIS DEPENDENT VARIABLES**

| $X_1(n,i)$ | $X_2(n,i) = P(n,i) - Q(n)$ 
| $X_3(n,i) = [X_1(n,i)]^2$ | $X_4(n,i) = X_3(n,i) + [X_2(n,i)]^2$ |

Where

$P(n,i)$ = Ordinal rank of correct diagnosis in version $i$ of case $n$

$Q(n)$ = Minimum value of $P(n,i)$ for all versions $i$ of case $n$

**Defined Strategy Evaluation**

Input items from the first 35 cases of the clinical evaluation were entered according to the following defined protocols:

1. Disease attributes (including intact phrases) were taken from the case abstract form, and entered directly under the most appropriate specific context. If a given attribute did not occur in the designated context, it was re-entered under a broader context.

2. Disease attributes were uniformly entered under "all" (the broadest context), and the phrases were decomposed into individual words and entered under "all."*

3. As in 2, except that the phrases were entered intact.

4. As in 1, except that symptoms and signs were entered in the combined context "ss."

5. As in 3, except that "emphasis" was permitted. Attributes thought by the enterer to be of greater significance could be entered twice or more, using different contexts, and thereby emphasizing the effects of these attributes. Phrases, for example, could be entered intact, and then again as single words. The "all" category was used with rare exceptions.

For each of these strategies, two runs were performed; the first using all the terms that had been provided, and the second using only a few terms selected by the enterer.

**RESULTS**

**Nonparametric Evaluation**

The enterer comparisons are presented in
Table 3 with standard normal deviates (z) calculated for each possible pair. A negative z value indicates that enterer I performed better than enterer II. Although the results were not statistically significant, enterers F and G performed, on average, better than the other enterers, and were close in their compared performance.

<table>
<thead>
<tr>
<th>Enterer I</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.132</td>
<td>-0.865</td>
<td>-0.386</td>
<td>0.517</td>
<td>0.538</td>
<td>-0.495</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-1.32</td>
<td>-0.502</td>
<td>0.406</td>
<td>0.171</td>
<td>-0.371</td>
<td>0.861</td>
<td>1.16</td>
</tr>
<tr>
<td>C</td>
<td>0.940</td>
<td>0.386</td>
<td>-0.368</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-0.022</td>
<td>-1.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-1.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The z values for the variables associated with choices of context are shown in Table 4. A negative value indicates that a better outcome was associated with adding more items under that context.

### Table 4

<table>
<thead>
<tr>
<th>Context Variable</th>
<th>Average Number Items</th>
<th>z Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>3.58</td>
<td>-1.62</td>
</tr>
<tr>
<td>symptoms or signs</td>
<td>5.14</td>
<td>-0.50</td>
</tr>
<tr>
<td>symptoms</td>
<td>0.23</td>
<td>-1.87</td>
</tr>
<tr>
<td>signs</td>
<td>0.22</td>
<td>-0.95</td>
</tr>
<tr>
<td>lab</td>
<td>3.13</td>
<td>0.92</td>
</tr>
<tr>
<td>X-ray</td>
<td>0.38</td>
<td>-1.29</td>
</tr>
<tr>
<td>pathology</td>
<td>No entries</td>
<td></td>
</tr>
<tr>
<td>specific organ system</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td>other (etiology, etc.)</td>
<td>0.31</td>
<td>1.21</td>
</tr>
<tr>
<td>total number of terms</td>
<td>11.62</td>
<td>2.37</td>
</tr>
</tbody>
</table>

*Possibly spurious, used by only 3 of 7 enterers

### Parametric Evaluation

There was no difference in the results of the analysis regardless of the dependent variable function used. There was a positive correlation between the total number of items entered and a better result. There was a negative correlation with the number of terms entered under the context "X-ray". No significant correlation could be found with other variables, including that of enterers.

### Defined Strategy Evaluation

Eleven of the 35 cases were unsuitable for analysis, either because no diagnosis was achieved during hospitalization, or because the diagnosis was established before admission. 40 diagnoses occurring in 24 cases were scored. Performance was considered successful if the program suggested the correct diagnosis within the first 40 of its differential. Results of the defined strategy evaluation are shown in Table 5. During the clinical evaluation, the comparable performance on the same cases was 38 successful suggestions in at least one of seven or more versions of each of the 24 cases.

Table 5

<table>
<thead>
<tr>
<th>Strategy</th>
<th>All Terms</th>
<th>Selected Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>12</td>
</tr>
</tbody>
</table>

### Discussion

The individual selecting items to be entered may have an effect on the outcome. Enterers F and G were two individuals with the greatest degrees of prior clinical experience. In other respects, such as prior familiarity with the program, they were not comparable. Although the data did not reach the level of statistical significance, it suggests the possibility that clinical experience by the user has a role in predicting outcome.

There does appear to be some correlation between better performance and the total number of items entered. This result is not surprising. The better the descriptions of the case, the more likely that the correct diagnosis will be recognized. Whether this effect will hold true in cases with multiple diagnoses remains to be seen.

The negative effect of entering terms under the context "X-ray" seen on regression analysis was not confirmed in the nonparametric evaluation. That context was used infrequently, further experience with using it may be instructive.

Overall, the use of specific contexts does not seem to have a predictable effect. This result is not intuitively obvious. In theory, a term entered under a more specific context, because of its higher selectivity score, should raise the diagnosis to a higher ranking in the differential list. That it does not predictably do so might reflect either difficulties in selecting an appropriate context, or inconsistencies in the treatment of terms by the knowledge base. A third possibility is that the numerical effect of the selectivity score is small, and that the higher selectivity score is overwhelmed by the effect of the total number of matches. Each of these possibilities has some validity. The effect of clinical experience, which might have an effect on choices of strategy and context, has been mentioned. Inconsistency in the treatment of certain terms in the OMIT has been noted earlier.

The most striking aspect of the results is the relatively poor performance of the program using defined strategies. Compared to the performance in the clinical evaluation, no one single strategy is as successful as choosing among
the multiple alternative strategies. This unusual feature, the ability to enter terms under contexts, is clearly an important one.

These studies were intended to elucidate factors which could improve the performance of the program. Continued efforts will be necessary to find methods of selecting strategy which will improve the performance of the program and its successors.

ACKNOWLEDGEMENTS.

We thank the seven individuals who entered the abstracted data. They are Mark S. Tuttle, Marsden Blois, MD, Mark Erlbaum, MD, Ryo Kim, MD, Peter Harrison, MD, Pauline Velez, Bernard Winkelmann, MD, and Dale Yamashita, MD.

REFERENCES


APPENDIX A

A 78 year old woman was admitted for hemoptysis, pleuritic chest pain, cough and weight loss. On physical exam she was noted to have axillary adenopathy, rib tenderness, decreased breath sounds at the bases, rales, and a systolic murmur. Admission labs revealed hematuria and a thrombocytosis, the chest X-ray showed bilateral pleural effusions.

Enter terms: xr/chest, pleural, effusions
Enter terms: ss/hemoptysis, pleuritic, chest, pain, cough, weight loss
Enter terms: ss/rib, breath, sounds, rales, axillary, adenopathy
Enter terms: ss/systolic murmur
Enter terms: all/thrombocytosis, hematuria

Computing scores for X-ray terms
Finished chest, selectivity = 0.991
Finished pleural, selectivity = 0.992
Finished effusions, selectivity = 0.995

Computing scores for Signs or Symptoms terms
Finished hemoptysis, selectivity = 0.979
Finished pleuritic, selectivity = 0.995
Finished chest, selectivity = 0.929
Finished pain, selectivity = 0.991
Finished cough, selectivity = 0.923
Finished weight loss, selectivity = 0.941
Finished rib, selectivity = 0.991
Finished breath, selectivity = 0.905
Finished sounds, selectivity = 0.944
Finished rales, selectivity = 0.971
Finished axillary, selectivity = 0.983
Finished adenopathy, selectivity = 0.971
Finished systolic murmur, selectivity = 0.983

Computing scores for all terms
Finished thrombocytosis, selectivity = 0.999
Finished hematuria, selectivity = 0.954

Sorting totalled disease scores...

1847 diseases in list
17.035 maximum total score
1 9.196 coccidioidomycosis
2 8.220 pneumonia, eosinophilic, chronic
3 8.177 nocardiosis
4 7.583 tuberculosis, pulmonary
5 7.299 pneumonia, primary atypical
6 7.261 mediastinum, lymphosarcoma
7 7.211 histoplasmosis
8 7.182 sarcoidosis, pulmonary
9 7.182 lung, gangrene
10 6.603 emphysema, pulmonary
11 6.346 diaphragmatis
12 6.307 pneumonia, klebsiella
13 6.240 lung, echinococcosis
14 6.240 bronchiectasis
15 6.240 lung, abscess
16 6.240 bronchus, injury
17 6.213 cadmium, toxicity
18 6.211 lung, carcinoma, bronchogenic
19 5.687 atelectasis, compression
20 5.666 atrial septal defect

The patient was found to have active pulmonary tuberculosis.