# Production of Artificial "Case Histories" by using a Small Computer

F. T. DE DOMBAL, JANE C. HORROCKS, J. R. STANILAND, P. J. GUILLOU

*British Medical Journal*, 1971, 2, 578–581


## Summary

**This paper describes a method of producing artificial "case histories" by using probability theory and clinical data from a series of 600 patients with acute abdominal pain. A series of 12 such cases were distributed to clinicians, medical students, medical secretaries and technicians, and members of the general public. For each "case" most clinicians concurred with the intended diagnosis. So did the medical secretaries and technicians; indeed this group were more confident of their chosen diagnoses than were the clinicians.**

**It is suggested that clinicians are concerned to a large extent with the consequences of a diagnosis as well as its accuracy, and are motivated to some degree by a fear of the consequences of failure. They may be justified in adopting this policy, for when "errors" in diagnosis are harshly penalized the clinicians were infinitely more effective than any of the other groups.**


## Introduction

We have already drawn attention to the problems of teaching clinical diagnosis where increasing numbers of students face a relatively static population of teachers and patients, and have suggested that simulation techniques might play some part in alleviating this.[1 2] Nevertheless, in further studies[3] we have shown that the simulation of clinical diagnosis is not without its problems—particularly where a computer-based system is used in an on-line real-time mode—and have suggested that other simulation techniques and other methods of computer usage might profitably be explored.

This paper describes one such experiment in which a small computer was used to generate a series of artificial case histories on the basis of random numbers and using probabilities supplied by us after a survey of 600 patients. The "cases" were presented for comment and diagnosis to several volunteers (surgeons, medical students, technical staff, and members of the general public), and the resulting data are set out below and some tentative conclusions put forward.

## Method of Generating Case Histories

We decided to concentrate on a small subset of six diseases which generally present with abdominal pain of acute onset and to display a fixed set of clinical attributes for each case (Table I). A total of 100 patients suffered from each of the six diseases chosen for study, which were: acute appendicitis, acute diverticular disease, perforated peptic ulcer, acute cholecystitis, acute small-bowel obstruction, and non-specific abdominal pain. The patients in the last-mentioned group were admitted to hospital with abdominal pain for which no apparent cause was found before their discharge. They were all followed for six months without recurrence of the pain. Our choice of both diseases considered and attributes displayed was quite arbitrary, and could readily be changed if desired.

TABLE I—*Attributes Displayed in Each Case History*

| | | |
|---|---|---|
| Presenting complaint (pain) | | Site at onset<br>Site at present<br>Severity<br>Type<br>Duration<br>Aggravating/relieving factors |
| Other symptoms | | Nausea/vomiting<br>Appetite<br>Previous indigestion<br>Bowels<br>Micturition<br>Periods (where appropriate) |
| Previous history | | Previous similar pain<br>Previous surgery |
| Physical examination | General | Mood, colour<br>Pulse, temperature, B.P., respiration |
| | Abdomen | Movement<br>Distension<br>Scars<br>Tenderness/rebound<br>Guarding/rigidity<br>Masses<br>Bowel sounds<br>Rectal examination |

The clinical data fed into the computer-based system were derived from a "data base" of clinical information compiled from the series of 600 patients suffering from one or other of the diseases shown above. A total of 35 variables was recorded for each patient (Table I) so that the data base of clinical information contained about 20,000 items of data.[4] Two alternative methods of generating case histories were used (Tables II and III). In the first "stereotypes" were produced for each of the various diseases—that is, we considered each attribute in turn and displayed its most likely state in that disease. For example, when attempting to display a case of appendicitis this was more commonly found in males than in females, and most commonly in the second decade of life. Thus the stereotype of a case of appendicitis would be a boy aged 15, that of perforated duodenal ulcer a man aged 45, that of diverticulitis a woman aged 75 (see Table II), and so on.

TABLE II—*Method of Displaying Stereotype of Diverticulitis*

| Variable | Data Base (100 Cases) Indicates: | Most Commonly Found State | So Stereotype for This Disease is: | System Displays: |
|---|---|---|---|---|
| Sex | Male, 39 cases<br>Female, 61 cases | Female | Female | "Patient is female . . ." |
| Age (years) | <-20, 0 cases<br>20-39, 4 cases<br>40-49, 10 cases<br>50-59, 10 cases<br>60-69, 28 cases<br>70-79, 34 cases<br>≥80, 14 cases | Between 70 and 79 | Between 70 and 79 | "aged 75 years . . ." |

**University Department of Surgery, General Infirmary, Leeds**
F. T. DE DOMBAL, M.D., F.R.C.S., Senior Lecturer in Surgery
JANE C. HORROCKS, Assistant in Programming
J. R. STANILAND, Assistant in Surgical Research
P. J. GUILLOU, B.SC., M.B., House Surgeon

The second method of generating artificial case histories made use of random numbers and probability theory. A series of random numbers was first generated, using a small desk-top computer (a Mathatronics Mathatron 848 Biostatistician) and a programme designed to generate a series of two-digit numbers by the congruential method.

TABLE III—*Method of Generating a "Case" of Appendicitis based on Probability Theory and Random Numbers*

| Variable | Data Base (100 Cases) Indicates: | So if Random Number is: | Then Display for this Case: | Actual Random Number Generated | Therefore this case is: |
|---|---|---|---|---|---|
| Sex | Male, 60 cases<br>Female, 40 cases | 1-60<br>61-100 | Male<br>Female | 79 | "Female" |
| Age | 0-9 years, 22 cases<br>10-19 years, 33 cases<br>20-29 years, 22 cases<br>30-39 years, 8 cases<br>40-49 years, 5 cases<br>50-59 years, 5 cases<br>≥60 years, 5 cases | 1-22<br>23-55<br>56-77<br>78-85<br>86-90<br>90-95<br>96-100 | aged 5<br>aged 15<br>aged 25<br>aged 35<br>aged 45<br>aged 55<br>aged 65 | 35 | "aged 15" |

The way in which these numbers are used is shown in Table III. In this instance we are attempting to generate a "patient" with appendicitis. In deciding the age and sex of the patient we first take into account the probabilities in real life, and note that of the last 100 patients in real life with appendicitis 60 were male and 40 female. Therefore, in this artificial case if the random number generated by the system is anywhere between 1 and 60 the case will be displayed as being "male," and if the random number is between 61 and 100 the case will be "female." A similar procedure applies for the age of the patient. In the example given in Table III the random numbers generated were 79 and 35; thus the case is one of a 15-year-old girl.

The type of case history generated by such methods is shown in full in Table IV. This patient shows most of the classical features of a perforated duodenal ulcer—but not all. Thus, though the patient is in severe pain, with a tender, silent, rigid abdomen, the blood pressure and pulse rate are perhaps closer to normal than might be expected in real life in the same circumstances. In fact, by using this mode of generating artificial patients it is possible (though unlikely) that from time to time some very strange patients will appear. Thus the patient in Table IV *could* have been a 15-year-old girl but, as in real life, the odds against this sort of thing happening were considerable (around 500-1 in this case).

TABLE IV—*Specimen "Case History" of 45-year-old Man*

| | | |
|---|---|---|
| Presenting complaint | Abdominal pain | Began in upper central abdomen<br>6 hours ago, now all over abdomen<br>Very severe<br>Steady<br>Aggravated by movement |
| | Other symptoms | Nausea<br>No vomiting<br>Appetite decreased<br>Bowels normal<br>Micturition normal<br>History of indigestion |
| | Previous history | Previous similar pain<br>No previous surgery |
| On physical examination | General | Mood—distressed<br>Colour pale<br>Pulse 90/min<br>B.P. 134/80<br>Temperature 98·2°F (36·8°C)<br>Respiration 22/min |
| | Abdomen | Poor movement<br>No scars<br>No distension<br>Generalized tenderness<br>No rebound<br>No guarding<br>Rigidity<br>No masses<br>Decreased bowel sounds<br>Rectal examination N.A.D. |

## Conduct of Study

Altogether 12 artificial patients were generated by these means—6 stereotypes (one of each disease) and 6 probability patients (again one for each diagnosis). The order in which the cases were presented was randomized, and this series of 12 cases was presented to several different groups who participated in the experiment: 15 clinicians, 12 students, 9 ancillary personnel (medical secretaries and technicians), and 6 members of the general public. Each of these subjects was asked to work through the series of 12 cases; thus in all some 504 diagnoses were attempted. For each diagnosis we asked two questions. Firstly, the most likely diagnosis (with alternatives if thought appropriate). Also, to test how confident the subject was of his chosen diagnosis, we allocated 10 "votes" for each case, to be distributed in any way the subject wished among the six diagnoses.

## Findings

*Accuracy of Diagnosis.*—To analyse the accuracy of the diagnosis we merely noted in each instance whether the subject's chosen diagnosis matched our intended one (Fig. 1). It came as something of a relief to find that the clinicians usually agreed with the intended diagnosis and that the medical students (in their final year) should record marginally less agreement. We were greatly surprised, however, by the ancillary workers, whose performances were often comparable with the students and clinicians. By contrast the members of the general public who participated scored much less agreement than the rest.

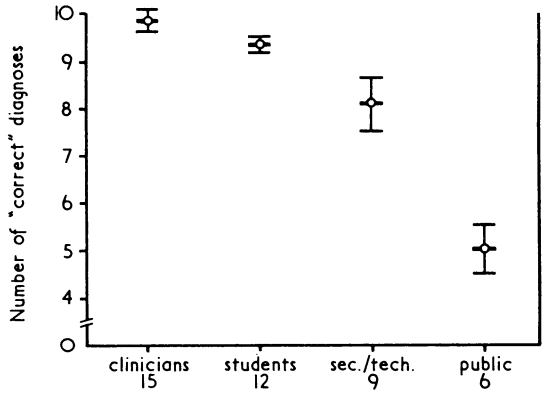


FIG. 1—Comparison of accuracy of diagnosis between groups indicated. Note that difference between clinicians and students is NOT significant ($t = 1{\cdot}59$, $n = 25$, $P > 0{\cdot}1$), nor is difference between students and secretaries and technicians ($t = 2{\cdot}05$, $n = 19$, $P > 0{\cdot}05$).

*Confidence.*—To measure the confidence with which the subjects made their diagnoses we noted the number of votes placed on the diagnoses of their choice, irrespective of whether this was a correct or incorrect choice (Fig. 2). The clinicians as a group behaved in a fairly cautious fashion, as did the students; but the secretaries, technicians, and members of the general public appeared to have few such reservations, and were often more confident of their chosen diagnoses than the clinicians or students.

*Effectiveness.*—"Effectiveness" in diagnosis is difficult to quantitate. One possible method of measuring this variable is by adding together the votes cast in each instance for the "correct" diagnosis, thereby combining both accuracy and certainty. In this respect the clinicians as a group scored higher than any other group (Fig. 3), since even when they reached a different diagnosis from that which was intended by the computer they nonetheless allocated a substantial proportion of their votes to the correct diagnosis. A clinician, for example, might consider a case intended to be appendicitis as an example

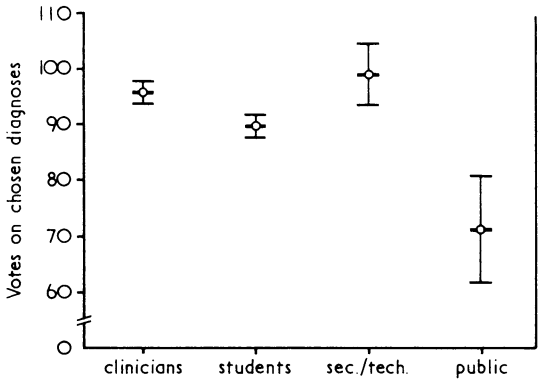


FIG. 2—Comparison of confidence in diagnosis between groups indicated. Note that secretaries and technicians are more confident of chosen diagnoses than either students or clinicians.

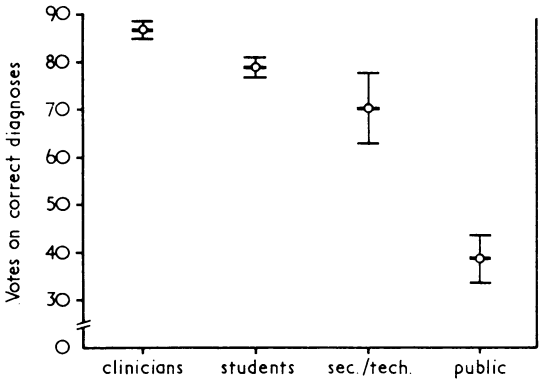


FIG. 3—Comparison of "effectiveness" of diagnosis, as measured by adding together total number of votes allocated by each individual to the 12 "correct" diagnoses. Clinicians (cf. Fig. 1) are now more effective than students ($t = 2{\cdot}20$, n = 25, $P > 0{\cdot}05$) or secretaries ($t = 2{\cdot}64$, n = 22, $P > 0{\cdot}02$).

of non-specific abdominal pain yet allocate only six votes to non-specific pain and the remaining four to appendicitis. (In the same case a secretary or technician might allocate all 10 votes to small-bowel obstruction.)

*Stereotypes v. Random Generation.*—It might have been expected that the clinicians' certainty levels would have been much higher in the stereotype cases than in those generated at random. This was not, in fact, so; the clinicians allocated a mean of 7·77 votes out of 10 to each of the stereotypes, and 6·73 to the randomly generated cases. Moreover, when we analysed the data for each individual case we were unable to find a significant correlation between the levels of certainty reached by probabilistic analysis and the certainty levels recorded by clinicians (Table V). This implies that clinicians do not "think Bayes," or that pure probability theory plays a relatively small part in their diagnostic process; but we plan further experiments to test this hypothesis.

TABLE V—*Comparison of "Stereotype" and "Randomly Generated" Cases showing Mean Votes Allocated by Each Clinician, and Probability of "Correct" Diagnosis by Computer-based Bayesian Analysis*

| Case No. | Intended Diagnosis | Clinicians'* Mean Votes | Probability by Bayesian Analysis |
|---|---|---|---|
| | *Stereotype* | | |
| 1 | Appendicitis | 9·33 | 0·999 |
| 2 | Perforated D.U. | 9·40 | 0·999 |
| 3 | Cholecystitis | 8·44 | 0·985 |
| 4 | Non-specific pain | 4·26 | 0·985 |
| 5 | Small-bowel obstruction | 9·46 | 0·999 |
| 6 | Diverticulitis | 5·67 | 0·999 |
| Total | | 7·77 | 0·997 |
| | *Random Generation* | | |
| 1 | Appendicitis | 7·80 | 0·999 |
| 2 | Perforated D.U. | 8·73 | 0·999 |
| 3 | Cholecystitis | 8·33 | 0·999 |
| 4 | Non-specific pain | 4·90 | 0·884 |
| 5 | Small-bowel obstruction | 6·26 | 0·999 |
| 6 | Diverticulitis | 4·27 | 0·994 |
| Total | | 6·73 | 0·979 |

*Comparing clinicians' votes *v.* Bayesian probability for each case, $\rho = 0{\cdot}508$ ($P > 0{\cdot}1$). No significant correlation.

## Discussion

### IMPLICATIONS FOR TEACHING

We have shown that artificial case histories can be produced in substantial numbers, cases which can be recognized by medical staff but not by lay persons. Moreover, this does not involve great labour or expense, and neither does it involve the use of computers in an on-line real-time mode. Our cases were all produced within two to three hours, and while use was made of a desk-top computer to generate random numbers this could easily be done by using currently available statistical textbooks instead. As regards availability of clinical data many such series exist. Our own series (which was originally analysed with a quite different purpose in mind) is set out elsewhere.[4] The cases can be produced with varying and quantifiable degrees of "difficulty" by altering the precise mode of generation. Probably such a method of teaching will never form a major part of the medical curriculum, but there are isolated occasions (such as when a patient scheduled for bedside teaching goes home or refuses permission) when a series of artificial substitutes might be useful. Indeed, we have found our series quite useful on occasion—not so much for the cases themselves as for the subsequent discussion with the students, to whom the concept of "certainty" in diagnosis is often new and intriguing.

In terms of performance it would be facile (even if true) to point out that the students behaved more like medical secretaries than like clinicians. Actually our main query is not why the students should have done "badly" (which was true only in a relative sense, since the patients were artificial) but why the technicians and secretaries should have done so well. Probably their acquisition of knowledge both by casual contacts and by exposure to stereotypes of the various diseases was far greater than we had supposed.

### IMPLICATIONS FOR DIAGNOSIS

The results from this particular experiment confirm a widely held supposition—namely, that clinicians like other human beings are conservative data processors in that they extracted less information than was inherent in the data presented to them. Thus the computer-based system was asked to analyse the 12 cases itself. It came to the correct decision in all 12 and reached a certainty and effectiveness level of 119—far higher than any of the human subjects (see Figs. 2 and 3). This was scarcely a surprise, since the system had generated the cases in the first place, but the high levels of certainty and effectiveness in the computer analysis gave rise to a further query. Was this level of performance due to a better appreciation of the probability values for each clinical attribute by the computer or due simply to the computer's ability to process a large amount of information at once? The computer-based systems therefore processed the cases again, but this time we restricted the clinical information available to the system, first to 12 items for each case and then in a further analysis to six items per case. (Within this restriction the computer was allowed to select the six most appropriate items of information for the 12 cases.)

Interestingly, when this restriction was imposed the computer-based system's performance approximated closely

as regards accuracy, certainty, and effectiveness to that of the group of clinicians. These data would seem to indicate that in real life "conservative" data processing in diagnostic-type tasks may well be due to primary inability to process large amounts of clinical data at the same time. In other words, the sort of "limited channel capacity" suggested by Miller[5] and more recently by MacRae[6] may operate. Further experiments, however, are planned in order to investigate this point.

Our findings would seem to confirm another of our suspicions —namely, that while accuracy of diagnosis is undoubtedly important *certainty* is also a major problem. Particularly striking was the fact that the clinicians were *less* certain of their chosen diagnoses than the secretaries and technicians—even though half of the cases were stereotypes of the diseases concerned. In our studies from real life[7] we have delineated three "phases" of the diagnostic process and have suggested that in one of these phases the concept of "pay-off" predominates. Thus the principal concern of the physician is not merely diagnostic accuracy but also the consequences of various alternative decisions about treatment. Hamilton[8] has aptly commented that one prime concern of many clinicians may be *a consideration of the consequences of error*, and certainly our results from the present study would tend to confirm this. The results also confirm that this policy is both justified and effective, since if a scoring system is adopted which harshly penalizes errors (such as *multiplying* together the votes allocated to each correct diagnosis) the clinicians fare infinitely better than any of the other groups (Fig. 4).

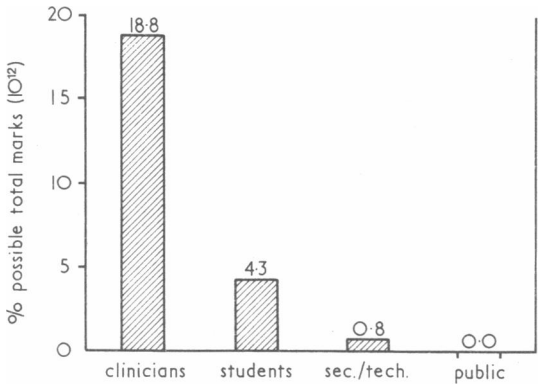


FIG. 4—As for Fig. 3 but with failure of diagnosis harshly penalized by multiplying together the votes cast for the 12 correct diagnoses. Clinicians and students are now the only groups to score over 1% of total possible votes, clinicians being four times as effective as students.

We are sincerely grateful to the clinicians, medical students, secretaries, technicians, and members of the general public whose experiences form the basis of this report. We are grateful also to Professor J. C. Goligher for his advice and encouragement throughout this study, and to Professor M. Hamilton for helpful conversations, together with the specific points mentioned in the text. Finally, one of us (J.C.H.) was aided by a grant from the Medical Research Council, which we acknowledge with gratitude.

## References

[1] de Dombal, F. T., Hartley, J. R., and Sleeman, D. H., *Lancet*, 1969, **1**, 145.
[2] de Dombal, F. T., Hartley, J. R., and Sleeman, D. H., *British Journal of Surgery*, 1969, **56**, 754.
[3] de Dombal, F. T., Horrocks, Jane C., Staniland, J. R., and Gill, P. W., 1971, *British Medical Journal*, 1971, **2**, 578.
[4] de Dombal, F. T., Horrocks, Jane C., Staniland, J. R., and Guillou, P. J., *Proceedings of the Royal Society of Medicine*, 1971, in press.
[5] Miller, G. A., *Psychological Review*, 1956, **63**, 81.
[6] Mac Rae, A. W., *Psychological Bulletin*, 1970, **73**, 112.
[7] de Dombal, F. T., in *Principles and Practice in Medical Computing*, ed. W. Lutz and L. G. Whitby. Edinburgh, Livingstone, 1971, in press.
[8] Hamilton, M., personal communication, 1970.

# *Today's Drugs*

*With the help of expert contributors we print in this section notes on drugs in common use.*

# Mucolytic Agents

Many patients with chest disease have difficulty in clearing their chests of sputum. A variety of substances have been used in an attempt to help them. These can be divided into two broad groups, though in the case of some drugs there may be some degree of overlap. Firstly there are the so-called expectorants. These are compounds that stimulate patients to cough, and many of them are also emetics. For example, some of the traditional remedies containing sodium bicarbonate and *small* doses of iodine probably act as non-specific cough stimulants and emetics and owe their action to this rather than to any specific action on the secretion of bronchial mucus. On the other hand, there are agents which "thin" the sputum or render it less viscid so that it can be more easily expectorated. These are termed "mucolytic agents," and it is with this latter class of agents that this article is concerned.

### Non-specific Remedies

The viscosity of sputum depends on its degree or hydration, which in turn depends on the degree of hydration of the patient. Many patients with chest disease become dehydrated. This results in sticky and even caked sputum. Adequate hydration, by the intravenous route if need be, can make a big difference to sputum viscosity and the ease with which it can be expectorated. Inhalations of steam may also be helpful, whether flavoured with menthol or not. Alevaire is a weak detergent which may be rather more effective.

### Chymotrypsin and Cysteine Compounds

Chymotrypsin and other enzymes have been successfully used to digest sputum in vitro. Nevertheless, their clinical use has been disappointing.

Various preparations for inhalation are available such as Lomudase.

Acetylcysteine and methylcysteine compounds are available in aerosol or oral forms. By inhalation they undoubtedly have an effect, especially on sputum volume, so much so that the manufacturers of one of them (Airbron) issue a warning that such large volumes of mucus may be mobilized