PUFF: An Expert System for Interpretation of Pulmonary Function Data

by

Janice S. Aikins, John C. Kunz, Edward H. Shortliffe, and

Robert J. Fallat

Departments of Medicine and Computer Science

Stanford University
Stanford, CA 94305

PUFF: An Expert **System for Interpretation of Pulmonary** Function Data

Janice S. Aikins*
John C. Kunz*
Edward H. Shortliffe*
Robert J. Fallat**

"Heuristic Programming Project
Departments of Medicine and Computer Science
Stanford University
Stanford, California 94305

****Pacific Medical Center**
2200 Webster Street
San Francisco, California 94115

Running head title: PUFF

Address correspondence to Janice S. Aikins, current address: Computer Research Center, Hewlett-Packard, 1601 Page Mill Road, Palo Alto, California 94304

Abstract

The application of Artificial Intelligence techniques to real-world problems has produced promising research results, but seldom has a system become a useful tool in its domain of expertise. Notable exceptions are the DENDRAL [1] and MOLGEN [2] systems. This paper describes PUFF, a program that interprets lung function test data and has become a working tool in the pulmonary physiology lab of a large hospital. Elements of the problem that paved the way for its success are examined, as are significant limitations of the solution that warrant further study.

1 Introduction

Researchers in the field of Artificial Intelligence (AI) are just beginning to produce systems that capture the specialized knowledge of experts and that use this knowledge to perform difficult tasks. Although the technology is still rather new, a small set of programs now exist as "tools" useful for building these so-called "expert systems". This paper describes an expert system, called PUFF, which was built using EMYCIN, a generalization of an earlier medical system named MYCIN. The task chosen for PUFF is described briefly, and the rationale for the appropriateness of this choice is presented. PUFF was initially developed on the SUMEX computer, a large research machine at Stanford University, and was later rewritten in a production version to run on the hospital% own mini-computer. We describe here the history of the PUFF project and its current status, including observations about its limitations and successes. We also take a brief look at the knowledge representation and control structure used for the SUMEX version of the system. Finally, the results of a formal evaluation of the production version of PUFF are presented.

2 Task

PUFF interprets measurements from respiratory tests administered to patients in the pulmonary (lung) function laboratory, at Pacific Medical Center in San Francisco. The laboratory includes equipment designed to measure the volume of the lungs, the ability of the patient to move air into and out of the lungs, and the ability of the lungs to get oxygen into the blood and carbon dioxide out.' The pulmonary physiologist interprets these measurements in order to determine the presence and severity of lung disease in the patient. An example of such measurements and an interpretation statement are shown in

Fig. 1. The test measurements listed in the top half of the figure are collected by the laboratory equipment.   The pulmonary physiologist then dictates the interpretation statements to be included in a typewritten report. All of the measurements are given as a percent of the predicted values for a normal patient of the same sex, height, and weight. The interpretation and final diagnosis are a summary of the reasoning about the combinations of measurements obtained in the lung tests.

3 Rationale

PUFF's task is to interpret such a set of pulmonary function (PF) test results, and to produce a set of interpretation statements and a diagnosis for the patient. The problem of developing an automated pulmonary function interpretation system was chosen for several reasons:

(1) The interpretation of pulmonary function tests Is a problem that occurs daily in most hospitals, so a computer program that captures the expertise involved in interpreting these tests, and that can assist in providing interpretations, fills a practical need.

(2) The biomedica researchers at Pacific Medical Center (PMC) were interested in the problem and were eager to work with us on developing a solution. It was possible that such a system could enhance the effectiveness of patient care and the pulmonary physician% efficiency. In addition, solution of this relatively simple interpretation problem could identify possibilities for further research into more difficult Interpretation tasks.

(3) PF data interpretation was a problem which the Artificial Intelligence researchers were particularly interested in solving in order to demonstrate the generality and power of

expert system techniques. Putting a system into clinical use would contribute to the credibility of those techniques, and also would show their promise and limitations in clinical practice. Earlier AI programs had demonstrated competence, but their use had required large amounts of professional time simply for data input. Puff, however, produced PF interpretations automatically without the necessity for user interaction. Thus we hoped that PUFF would be used by the clinical staff.

(4) PF data interpretation was a problem which was large enough to be interesting (the biomedical researchers did not know how to solve it, and the AI researchers did not know whether their techniques would be appropriate) and small enough that a pilot project of several months' duration could concretely demonstrate the feasibility of a longer development effort. Furthermore, the amount of domain-specific knowledge involved in pulmonary function testing is limited enough to make it feasible to acquire, understand, and represent that knowledge.

(5) The domain of pulmonary physiology is a circumscribed field: the data needed to interpret patient status are available from the patient's history and from measurements taken in a single laboratory. Other large bodies of knowledge are not required in order to produce accurate diagnoses of pulmonary disease in the patient.*

(6) All the data used in the laboratory at PMC were already available in a computer; the computer data were known to be accurate, reliable, and relevant to the interpretation task. The clinical staff in the PF lab were already receptive to the use of computers within their clinical routines.

(7) Pulmonary physiologists who interpret test measurements tend to phrase their

interpretations similarly from one case to the next. One goal of PUFF was to generate reports from a set of prototypical interpretation statements, thus saving the staff a great deal of tedious work. The staff themselves would not be displaced by this tool because their expertise still would be necessary to verify PUFF's output, to handle unexpected complex cases, and to correct interpretations that they felt were inaccurate.

4   Project History and Status

This research developed from work done on the MYCIN system [3]. That program used a knowledge base of production rules [4] to perform infectious disease consultations. PUFF was initially built using a generalization of the MYCIN system called EMYCIN [5]. EMYCIN, or "EssentialMYCIN", consists of the domain-independent features of MYCIN, principally the rule interpreter, explanation, and knowledge acquisition modules [6]. it provides a mechanism for representing domain-specific knowledge in the form of production rules, and for performing consultations in that domain. Just as MYCIN consists of EMYCIN plus a set of facts and rules about the diagnosis and therapy of infectious diseases, PUFF is comprised *of* the EMYCIN programs plus a pulmonary disease knowledge base.

EMYCIN (and hence the EMYCIN version of PUFF) is written in INTERLISP [7] and runs on a DEC Ki-10 at the Stanford SUMEX-AIM computer facility. in order to run PUFF on a POP-1 1 at Pacific Medical Center, a second version of the program was created after the EMYCIN version had been refined. This was done by translating the production rules into procedures and writing them in the BASIC language. Conversion to BASIC was an advantage because the POP-1 1 was located on the same site as the laboratory, and its schedule could be easily controlled to support production operation by the system users. However, as a

result of the conversion, the production and development versions of PUFF became incompatible, and modifications made to one system were sometimes difficult to make in the other.

The POP-1 1 version is now routinely used in the pulmonary function laboratory and provides lung test interpretations for about ten patients daily. Since the system became operational in 1979, it has interpreted the results of over 4000 cases. The BASIC code is currently being converted again so that it will run on a dedicated personal computer.

The form of the interpretations generated by PUFF is shown in Fig. 2. This 'report is for the same patient as in Fig. 1, seen several years later. As in the typed report, the pulmonary function test data are set forth, followed by the interpretation statements and a pulmonary function diagnosis. The pulmonary physiologist checks the PUFF report and, if necessary, the interpretation is edited on-line prior to printing the final report for physician signature and entry into the patient record. Approximately 85% of the reports generated are accepted without modifications. The change made to most others simply adds a statement suggesting that the patient's physician compare the interpretation with tests taken during previous visits. For example, statements such as *"These* test results are *consistent with those of previous visits"* o r *"These test results show considerable Improvement over those in the previous visit"* might be made. PUFF was not designed to represent knowledge about multiple visits, so this kind of statement must be added by the pulmonary physician.

6 Observations

PUFF is a practical assistant to the pulmonary physiologist, and thus is a satisfactory

and exciting result of the research done with production rule consultation systems. PUFF's performance is good enough that it is used daily in clinical service, and it has the support of both the hospital staff and its administration. However, limitations are recognized in the following areas:[3]

- ▌representation of prototypical patterns,

- ▌addition or modification of rules to represent knowledge not previously encoded,

- ▌alteration of the order in which information is requested during the consultation, and

- ▌explanation of system performance.

The first point refers to the fact that many cases can be viewed as relatively simple variations of typical patterns. PUFF does not recognize that a case fits a typical pattern, nor can it recognize that a case differs in some important way from typical patterns. As a result, PUFF's explanations of its diagnoses lack some of the richness of explanation that physicians can use when a case meets, or fails to meet, the expectations of a prototypical case. The medical knowledge in PUFF is encoded as "rules". Rules encode relatively small and independent bodies of domain knowledge. The rule formalism makes modification of the program's knowledge much easier than when that knowledge is embedded in computer code. Howe-vet, additions or modifications to the rules as referred to in the second point have caused difficulties because changes to one rule sometimes affect the behavior of other rules in unanticipated ways. The last two points apply only to the EMYCIN version of PUFF which runs interactively in a consultation style, question and answer mode with the user. in that system, questions are sometimes asked in an unusual order, and explanations of both the final interpretation, and of the questions being asked of the user, need to be improved.

Even though PUFF does exhibit certain limitations, the representation of pulmonary knowledge as production rules allows the encoding of interpretive expertise which previously was difficult to define because it is heuristic knowledge of the expert. EMYCIN on the SUMEX computer provided an excellent environment for acquiring, encoding, and debugging this expertise. However, it would have been inefficient and somewhat impractical to use the EMYCIN version of PUFF in a hospital setting. The simplicity of EMYCIN's reasoning process made the translation into BASIC procedures a feasible task, thus allowing the hospital% own computer staff to take over maintenance of the system.

The BASIC version of PUFF runs in "batch" mode and does not require interaction with a physician. We-believe that this system was readily accepted by the pulmonary staff for several reasons: First, the program's interpretations are consistently accurate. Second, explanations of diagnoses are appropriately detailed so that the user has confidence in the accuracy of correct diagnoses and enough information with which to recognize and modify incorrect diagnoses. Third, less physician time is required to produce consistently, high quality reports using the system than is required to analyze and dictate case reports without it. Finally, the program is well integrated into the routine of the laboratory; its use requires very little extra technician effort.

6 Overview of EMYCIN-PUFF

6.1 Knowledge Representation

The knowledge base of the EMYCIN-PUFF system consists of (a) a set of 64 *production* rules dealing with the interpretation of pulmonary function tests and (b) a set of

59 *clinical parameters.* **The production version (BASIC-PUFF) has been extended to include 400 production rules and 76 clinical parameters. The clinical parameters in EMYCIN-PUFF represent pulmonary function test results (e.g.,** *TOTAL LUNG CAPACITY* **and** *RESIDUAL VOLUME),* **patient data (e.g.,** *AGE* **and** *REFERRAL DIAGNOSIS),* **and data which are derived from the rules (e.g.,** *FINDINGS* **associated with a disease and** *SUBTYPES* **associated with the disease). There may be auxiliary information associated with the clinical parameters, such as a list of expected values and an English translation used in communicating with the user.**

The production rules operate on associative *(attribute-object-value)* **triples, where the attributes are the clinical parameters, the object is the patient, and the values are given by the patient data and lung test results. Questions are asked during the consultation in an attempt to fill in values for the parameters.**

The production rules consist of one or more "premise" clauses followed by one or more "action" clauses. Each premise is a conjunction of predicates operating on associative triples in the knowledge base. A sample PUFF production rule is shown in Fig. 3.

The rules are coded internally in LISP. The user of the system sees the production rules in their English form which is shown first in the figure. The English version is generated automatically from templates, as is described in [5].

6.2 Control Structure

The EMYCIN-PUFF control structure is primarily a goal-directed, backward chaining of production rules. The goal of the system at any time is to determine a value for a given clinical parameter. To conclude a value for that clinical parameter, it tries a pre-computed

list of rules whose actions conclude values for the clinical parameter (refer to [5] for details).

if the rules fail to conclude a value for a parameter, a question is then asked of the user in order to obtain that value. An **exception to this process occurs for parameters labeled ASKFIRST parameters. These represent information generally known by the user, such as results of pulmonary** function tests. **For these parameters it is more efficient simply to ask a consultation question than to attempt to infer the information by means of rules.**[4]

7 Evaluation of the BASIC-PUFF Performance System

The knowledge base from the original performance version of PUFF was tested on 107 cases chosen from files in the pulmonary function laboratory at Pacific Medical Center. Those 107 cases formed a representative sample of the various pulmonary diseases, their degrees and subtypes. Modifications were made to the knowledge base and the cases were tried again. This iteration continued until our collaborating expert was satisfied that the system's interpretations agreed with his own. At this point the system was "frozen" and a new set of 144 cases was selected and interpreted by the system. Ail 144 cases also were interpreted separately by two pulmonary physiologists (the expert working with us and a physician from a different medical center).

The results of the comparison of interpretations by each diagnostician are presented in the table in Fig. 4. The table compares "close" agreement in diagnosing the severity of the disease, where "close" is defined as differing by at most one degree of severity. Thus, for example, diagnoses of mild (degree=1) and moderate (degree=2) are considered close,

while mild and severe (degree=3) are not. Further, a diagnosis of *normal* is not considered to be close to a diagnosis of a mild degree of any disease.

The table shows that that the overall rate of agreement between the two physiologists on the diagnoses of disease was 92%. The agreement between PUFF and the physician who served as the expert to develop the PUFF knowledge base (MD-1 in the table) was 96%. Finally, the agreement between PUFF and the physician who had no part in the development of the PUFF knowledge base (MD-2) was 89%. Fig. 5 shows the distribution of diagnoses by each diagnostician, The number of diagnoses made by each diagnostician does not total 144 because patients were often diagnosed as having more than one disease.

8 Conclusions

The PUFF research has demonstrated that if the task, domain, and researchers are carefully matched, then the application of existing techniques can result in a system which successfully performs a moderately complicated task of medical diagnosis. Success of the program can be measured not only in terms of the system's technical performance, but equally importantly, by the ease and practicality of the system's day-to-day use in the lab fdr which it was designed, Rule-based representation allowed easy codification and later modification of expertise, and the simplicity of the rule interpreter in the INTERLISP version facilitated translation into BASIC and implementation on the hospital's own POP-1 1 machine. Using EMYCIN allowed the researchers to move quickly from a point where they found it difficult even to describe the diagnostic process to a point where a simple diagnostic model was implemented. Having a diagnostic model allowed them to focus on individual issues in

order to improve that model.   Although PUFF does not itself represent new Artificial
Intelligence techniques, its success is a testimonial for EMYCIN. In addition, its simplicity
has facilitated careful analysis of EMYCIN's rule representation and control structure and
has led to other productive research efforts ([8] and [9]).

Acknowledgements

Legends to Figures

Figure 1. **Verbatim copy of pulmonary function report dictated by physician**

Figure 2. **Pulmonary function report generated by POP-Y 1 version of PUFF**

Figure 3. A PUFF production rule in English and LISP versions

Figure 4. Summary of percent agreement in 144 cases

Figure 6. Number of diagnoses by each diagnostician for 144 cases

PRESBYTERIAN HOSPITAL OF PMC
CLAY AND BUCHANAN, BOX 7999
SAN FRANCISCO, CA. 94128
PULMONARY FUNCTION LAB

WT 40.8 KG, HT 161 CM, AGE 65   SEX F
REFERRAL DX-
***TEST DATE 5-13-76

	PREDICTED (+/-SD)	OBSER(%PRED)	POST DILATION OBSER(XPRED)	
INSPIR VITAL CAP (IVC) L	2.7(0.6)	2.3 (83)		
RESIDUAL VOL (RV) L	2.0(0.1)	3.8 (193)	3.1	(154)
FUNC RESID CAP (FRC) L	2.9(0.3)	4.6 (158)	3.9	(136)
TOTAL LUNG CAP (TLC) L	4.7(0.7)	6.1 (129)	5.5	(116)
RV/TLC %	42.	62.	55.	
FORCED EXPIR VOL (FEV1) L	2.3(0.5)	1.5 (66)	1.6	(71)
FORCED VITAL CAP (FVC) L	2.7(0.6)	2.3 (85)	2.4	(88)
FEV1 /FVC %	82.	64.	66.	
FORCE EXP FLOW 200-1200L/S	3.6(0.8)	1.8	1.9	
FORCED EXP FLOW 25-753 L/S	2.6(0.5)	0.7	0.7	
FORCED INS FLOW 200-1200L/S	2.5(0.5)	2.5	3.4	
AIRWAY RESISTANCE(RAW) (TLC= 6.1) 2.5		1.5	2.2	
DF CAP-HGB=14.5 (DSBCO)(TLC= 4.8) 23.		17.4 (72)		

****
INTERPRETATION

The vital CAPACITY is low, the residual volume is high as is the total lung
capacity, indicating air trapping and overinflation. This is consistent
with a moderately severe degree of airway obstruction as indicated by the
low FEV1, low peak flow rates and curvature to the flow volume loop.
Following isoproterenol aerosol there is virtually no change.

The diffusing capacity is low indicating loss of alveolar capillary
surface.

Conclusion: Overinflation, fixed airway obstruction and low diffusing
capacity would all indicate moderately severe obstructive airway disease of
the emphysematous type. Although there is no response to bronchodilators
on this one occasion, more prolonged use may prove to be more helpful.

PULMONARY FUNCTION DIAGNOSIS: OBSTRUCTIVE AIRWAY DISEASE, MODERATELY
                                 SEVERE, EMPHYSEMATOUS TYPE

FIGURE 1.

PRESBYTERIAN HOSPITAL OF PMC
 CLAY AND BUCHANAN, BOX 7999
SAN FRANCISCO, CA. 94128
  PULMONARY FUNCTION **LAB**

WT 48.8 KG, HT 161 CM,   **AGE 69   SEX F**
REFERRAL DX-
TEST **DATE 05/13/80

	PREDICTED (+/-SD)	OBSER(%PRED)	POST DILATION OBSER(%PRED)
INSPIR VITAL CAP(IVC)L	2.7	**2.3 (86)**	**2.4 (90)**
RESIDUAL VOL (RV) L	2.8	**3.8** (188)	**3.0** (148)
TOTAL LUNG CAP (TLC) L	**4.7**	6.1 (**138**)	5.4 (115)
RV/TLC %	43.	62.	56.
FORCED EXPIR **VOL(FEV1)** L	2.2	1.5 (68)	1.6 (73)
FORCED VITAL CAP (FVC) L	2.7	2.3 **(86)**	2.4 **(90)**
FEV1 /FVC %	73.	65.	67.
PEAK EXPIR FLOW (PEF) L/S	7.1	1.8 (25)	1.9 (26)
FORCED EXP FLOW 25-75% L/S	1.8	8.7 (39)	8.7 (39)
AIRWAY RESIST(RAW) (TLC= 6.1)	0.0(0.0)	1.5	2.2

D F CAP-HGB=14.5   (TLC= 4.8) 24.   **17.4 (72)** (74%IF TLC = 4.7)
****

INTERPRETATION: ELEVATED LUNG VOLUMES INDICATE OVERINFLATION. IN ADDITION, **THE**
RV/TLC RATIO IS INCREASED, **SUGGESTING A MODERATELY SEVERE DEGREE OF** AIR **TRAPPING.**
THE FORCED VITAL CAPACITY IS **NORMAL.** **THE** FEV1/FVC RATIO AND MID-EXPIRATORY FLOW
ARE REDUCED AND THE AIRWAY RESISTANCE IS INCREASED, SUGGESTING **MODERATELY SEVERE**
AIRWAY OBSTRUCTION. FOLLOWING BRONCHODILATION, THE EXPIRED **FLOWS SHOW MODERATE**
IMPROVEMENT. HOWEVER, **THE** RESISTANCE DID **NOT** IMPROVE. **THE LOW** DIFFUSING
CAPACITY INDICATES **A LOSS OF ALVEOLAR** CAPILLARY SURFACE, WHICH IS MILD.

CONCLUSIONS: THE LOW DIFFUSING CAPACITY, IN COMBINATION **WITH** OBSTRUCTION **AND A**
HIGH **TOTAL LUNG** CAPACITY IS CONSISTENT WITH A DIAGNOSIS **OF EMPHYSEMA. ALTHOUGH**
BRONCHODILATORS **WERE** ONLY SLIGHTLY **USEFUL IN** THIS ONE CASE, PROLONGED USE MAY
PROVE TO BE BENEFICIAL TO THE PATIENT.

PULMONARY FUNCTION DIAGNOSIS:
1. **MODERATELY SEVERE OBSTRUCTIVE** AIRWAYS DISEASE.
    EMPHYSEMATOUS TYPE.

FIGURE 2,

RULE811

If: 1) **A: The mmf/mmf-predicted ratio is between 35 and 45, and**
        B: The fvc/fvc-predicted ratio is greater than 88, or
    2) **A: The mmf/mmf-predicted ratio is between 25 and 35, and**
      B: The fvc/fvc-predicted ratio is less than 88

Then: 1) **There is suggestive evidence (.5) that the degree of obstructive airways disease as indicated by the MMF is moderate, and**
     2) **It is definite (1.0) that the following is one of the findings about the diagnosis of obstructive airways disease: Reduced mid-expiratory flow indicates moderate airway obstruction.**

PREMISE: **[$AND ($OR (SAND (BETWEEN*** (VAL1 CNTXT **MMF) 35 45)**
                             (GREATERP***** (VAL1 **CNTXT FVC) 88))**
                    (SAND (BETWEEN* (VAL1 CNTXT MMF) 25 35)
                    (LESSP***** (VAL1 **CNTXT FVC) 80]**

ACTION: **(DO-ALL (CONCLUDE CNTXT DEG-MMF MODERATE TALLY 588)**
               (CONCLUDETEXT CNTXT FINDINGS-OAD
                              (TEXT $MMF/FVC2**) TALLY 1888))**

FIGURE 3.

PERCENT AGREEMENT

DIAGNOSIS	MD-1 **MD-2**	MD-1 **PUFF**	**MD-2** **PUFF**
NORMAL	92	95	**92**
OAD	94	99	**94**
RLD	92	99	85
DD	98	91	85
TOTAL **(S.D.)**	92 (1.63)	**96** (3.83)	89 (4.69)

Diseases: Normal=Normal Pulmonary Function
OAD=Obstructive Airways Disease
RLD=Restrictive Lung Disease
DD=Diffusion Defect

FIGURE 4,

DIAGNOSTICIAN

DIAGNOSIS	MD-1	**MD-2**	**PUFF**
NORMAL	31	26	38
OAD	79	85	89
RLD	52	45	55
DD	53	35	52

5.

Footnotes

1. **Measurements include spirometry and, optionally, body plethysmography, single-breath CO diffusion capacity, and arterial blood gases. Measurements can be made at rest, following inhalation of a bronchodiiator, and during exercise.**

2. **This was a problem in MYCIN, a related system for determining the diagnosis and therapy for infectious disease cases. The results produced by the system often suffered because it lacked knowledge** about **related diseases that were also present in the patient.**

3. **Many of these problems are also present in other rule-based systems; they motivated the development of the experimental CENTAUR** system **[8].**

4. **in the BASIC version of PUFF implemented at PMC, all of the test data is known ahead of time so that "asking a question" merely** entails **retrieving another datum from a stored file.**

References

[1] Buchanan, B.G., and Feigenbaum, E. A. DENDRAL and Meta-DENDRAL: Their Applications Dimension. *Artificial Intelligence 1 I(*1,2)(1978), pp. 6-24.

[2] Friedland, P. *Knowledge-based Experiment Design in Molecular Genetics. Proceedings of the Sixth international Joint Conference on Artificial Intelligence,* 1979, pp. 285-287.

[3] Shortliffe, E. H. *Computer-Based Medical Consultations: MYCIN.* New York: American-Eisevier, 1976.

[4] Davis, R., and King, J. An Overview of Production Systems; In *Machine Intelligence* (E. W. Elcock and D. Michie, Eds.), Vol. 8, pp.300-332. New York: Wiley & Sons, 1977.

[5] vanMelle, W. *EMYCIN: A Domain-independent Production-rule System for Consultation Programs.* STAN-CS-80-820, Stanford University, June 1980.

[6] Shortliffe, E. H., Davis, R., Buchanan, B., Axline, S., Green, C., Cohen, S. Computer-based Consultations In Clinical Therapeutics -- Explanation and Rule Acquisition Capabilities of the MYCIN System. *Computers and Biomedical Research, 8* (1975), pp. 303-320.

[7] Teitelman, W. *INTERLISP Reference Manual.* Xerox Palo Alto Research Center, Palo Alto, Ca., October 1978.

[8] Aikins, J. S. *Prototypes and Product ion Rules: A Knowledge Representation for Computer Consultations.* STAN-CS-80-8 14, Stanford University, August 1980.

[9] Smith, D.E., and Clayton, J. E. *A Frame-based Production System Architecture. Proceedings of the First Annual National Conference on Artificial Intelligence,* 1980, pp. 164-I 66.