Artifice for Databases that do not Include all Known Diseases and Clinical Data

CARLOS FEDER, MD, Internal Medicine; TOMAS FEDER, PhD, Computer Sciences

Most previous computerized diagnosis programs do not include in their databases all the currently known diseases and corresponding clinical data (symptoms, physical signs, and diagnostic tests and procedures). This limitation introduces inaccuracies in calculation of differential diagnoses probabilities with almost any method applied; still worse, the diagnosis of the excluded disease will never be included in the differential diagnosis list. This is why computer programs based on Bayes formula or other methods, but circumscribed to a restricted area of diseases—such as congenital cardiopathies or nephropathies—are inherently inaccurate. To overcome this problem, we devised an artifice that compensates for the mentioned deficiency.

This paper is part of our complete medical diagnostic system, described in detail in our book [2].

We summarize here only basic concepts of previous publications; for better understanding of this paper, the reader is encouraged to consult these publications.

Sensitivity (S) is the cornerstone of our diagnostic system. A practical way to calculate S of a specific clinical datum for a given disease is to determine statistically the fraction of patients afflicted by this disease who manifest the clinical datum:

$$Sensitivity = \frac{Number\ of\ disease\ cases\ manifesting\ the\ clinical\ datum}{Total\ number\ of\ disease\ cases}$$

Positive predictive value (PP value) is the best index to determine the strength with which a specific clinical datum present in a patient supports a specific diagnosis. Our algorithm calculates PP value with the following equation:

$$PP\ value\ i = \frac{Si}{S1 + \cdots + Si + \cdots + Sn} \quad (1)$$

Where PP value i = positive predictive value of the clinical datum for the disease i under consideration

Si = sensitivity of the clinical datum for the disease i under consideration

S1 ... Sn = sensitivities of the same clinical datum for corresponding diseases

Disease model, as defined in our system, is an abstract concept that comprises all clinical data that can be manifested by all patients with a specific disease. A single patient typically never manifests all clinical data that the disease potentially can provoke. Integration of a specific disease model with all of its possible manifestations requires statistical study of a large patient population. *Each clinical form, stage, degree, or complication of a disease has its own disease model.* Because death and iatrogenic diseases are diagnoses that must be established clinically, the corresponding disease models must also be created.

Probability (P) of a diagnosis is calculated with our novel **mini-max procedure**, core of our diagnostic system, considering PP value of clinical data present (favoring corresponding diagnosis) and S of clinical data absent (disfavoring diagnosis). These values are processed by a specific formula:

$$Pi = \frac{PP\ value\ i\ (1 - Si)}{PP\ value\ 1\ (1 - S1) + \ \ldots + \ PP\ value\ i\ (1 - Si) + \ \ldots + \ PP\ value\ n\ (1 - Sn)} \qquad (2)$$

Where P_i = probability of a diagnosis i

$PP\ value_i$ = positive predictive value of the clinical datum present

S_i = sensitivity of the clinical datum absent

$PP\ value_1 \ldots PP\ value_i \ \ldots \ PP\ value_n$ = positive predictive value of the same clinical datum present for each respective diagnosis in the differential diagnosis list

$S_1 \ldots S_i \ldots S_n$ = sensitivity of the clinical datum absent for each respective diagnosis in the differential diagnosis list

We confirmed that mini-max procedure is superior to Bayes formula and other probabilistic or rating methods. Detailed explanations and examples can be found in our previous publication [2].

Cost to obtain a clinical datum involves, in our context, not only expense but also risk and discomfort resulting from the required test or procedure. We assign to each clinical datum one of four overall cost categories: no cost (clinical data typically obtained through medical history and physical examination), small cost (*e.g.*, obtained through routine laboratory analysis, ECG, and other ancillary studies), intermediate cost (*e.g.*, colonoscopy, lymph node excision biopsy), and great cost (*e.g.*, liver biopsy, laparoscopy, laparotomy). **Benefit** of a clinical datum is measured by the magnitude of change it produces in the probability (P) of the respective diagnosis, in turn depending on the magnitude of PP value of clinical data present, which increase P, and the magnitude of S of clinical data absent, which decrease P. The mini-max procedure calculates these P for corresponding diagnoses.

Best cost-benefit clinical data are recommended at each diagnostic stage to be investigated next in a patient, based on mini-max procedure, that predicts, based on probabilistic calculations, which set of such data would end the diagnostic quest, more efficiently and at lowest cost. Recommended best cost-benefit clinical data are typically quite numerous, mandating the need to heuristically reduce its number; this is achieved in part by certain **parameters** described elsewhere [2] that can be set at empirically values by the user. A tradeoff exists in each of these parameters: moving the value in one direction may significantly reduce the number of recommended data, but reducing also slightly the accuracy of diagnostic result, and *vice versa*. The effects that these parameters have on the recommendation of best cost-benefit clinical data are shown in diverse output files mentioned later.

Our program confirmed the importance of the *exhaustiveness condition* for calculating probability (P) of diagnoses, which states that to obtain accurate results all known diseases must be included in the database. Because we were not able to integrate such an extensive database on our own, we resorted to an artifice, creating a fictitious disease model that we called OTHER DISEASES in addition to the limited number of disease models that actually integrate our prototype model. This OTHER DISEASES model represents all other known diseases (estimated at several thousands). Without this artifice, the computer program interpreted some irrelevant clinical datum, for example faint heart sounds, as exclusive for pericarditis with effusion, simply because this clinical datum was not listed in the remaining limited number of disease models. Without OTHER DISEASES, equation 1 that calculates PP value of the mentioned clinical datum, had S = 0.50 in the numerator and S = 0.50 in the denominator being S = 0 for

all other diagnosis, yielding a PP value = 1.00 resulting in an improperly confirmatory P = 1.00 for the diagnosis of pericarditis. By creating OTHER DISEASES model and including in its long clinical data list the clinical datum "faint heart sounds" with a great S, we precluded this situation to occur. This great S added to the denominator of equation 1, reduces considerably the PP value of the mentioned clinical datum for pericarditis and P of this diagnosis to a non-confirmatory level.

However, at this point, another problem surfaced. In OTHER DISEASES, when assigning a great S to a clinical datum (*e.g.*, S = 1.00) that happens to be absent for other diagnoses, this S will integrate the corresponding terms in the denominator of equation 2 that calculates P of diagnoses. The corresponding term [PP value (1–S)] = [PP value (1–1)] = 0. One or more terms equaling 0 in the denominator will incorrectly increase considerably P of the diagnosis being processed. To neutralize this untoward effect, we had to create an extra OTHER DISEASES SAME model in addition to the OTHER DISEASES model, repeating in both models the same clinical data but assigning to each corresponding S half of its original value. Because these S values are added in the denominator of equation 1 that calculates PP values, the resulting PP value of a specific clinical datum for a specific diagnosis with half S value in both models, will be the same as with only OTHER DISEASES with S equal to the original entire value. However, an excessively great P is precluded by processing both mentioned models, because now the term [PP value (1–S)] = [PP value (1–0.50)] will yield a greater value and will appear twice in denominator of equation 2. Our program hides OTHER DISEASES and OTHER DISEASES SAME diagnoses form showing in the differential diagnosis list and other output files.

In summary, OTHER DISEASES and OTHER DISEASES SAME represent fictitious competing diagnoses, which temporarily replace the real competing diagnoses, not yet included in the database. The more frequently and the greater the estimated number of non-included diagnoses manifest a clinical datum, the greater must be the estimated S value assigned to this clinical datum in OTHER DISEASES, to counterbalance its confirmatory power for the included diagnosis being processed.

COMMENTS

Exhaustiveness is a condition of Bayes formula and several other methods of calculating probability of diagnoses; if violated, the results are inaccurate and diagnoses are missed. This condition requires that all currently known diseases and corresponding clinical data must be included in the database and processed. This task can hardly be accomplished by a single researcher; it requires the cooperation of a team of seasoned medical specialists, as it requires S estimation for all clinical data corresponding to each known disease. Our diagnostic prototype program proved that until we have available such exhaustive database, a limited amount of diagnoses can be processed, with somewhat less but still satisfactory accuracy, resourcing to OTHER DISEASES and OTHER DISEASES SAME artifices, representing all not yet included diseases, as described in this paper.

CONCLUSIONS

Our algorithm and program, although somewhat complex, is straightforward, especially when compared to other attempts in this field. It emulates a clinician's diagnostic reasoning. It is logical and mathematically simple. Bayes formula is used with modifications, because it is unable to process properly interdependent clinical data (as are most symptoms) and concurrent diseases. To facilitate implementation and updating of the algorithm, we tend to avoid complicated tools of artificial intelligence, such as causal, hierarchical, and probabilistic trees and networks. The algorithm freely uses heuristic procedures, so as to preclude excessive proliferation of clinical data and diagnoses. It promises

to be user friendly because it is expressed in natural language, is rational, and readily understandable. Determination of accurate sensitivity of clinical data and integration of clinical entities into complex clinical presentation models will be labor-intensive. A complete database with all known diseases, clinical data, clinical presentations, and other information can be created; this major task will require a dedicated team of medical specialists.

REFERENCES

[1] FEDER C. Computerized Medical Diagnosis: A Novel Solution to an Old Problem. Infinity Publishing, West Conshohocken, Pennsylvania, 2006

[2] FEDER C, and FEDER T. A Practical Computer Program that Diagnoses Diseases in Actual Patients. Infinity Publishing, West Conshohocken, Pennsylvania, 2008

LEDLEY RS and LUSTED LB. Reasoning Foundations of Medical Diagnosis. Science, 130 (9): 9-21, July 3, 1959

HENRION M, PRADHAN M, DEL FAVERO B, HUANG K, and O'RORKE P: Why is diagnosis using belief networks insensitive to imprecision in probabilities? Twelfth Conference on Uncertainty in Artificial Intelligence, Portland, OR, 446-454. 1996. SMI-96-0637

MYERS JD, POPLE HE, and MILLER RA. INTERNIST: Can Artificial Intelligence Help? In: Connelly, Benson, Burke, Fenderson, eds. Clinical Decisions and Laboratory Use. Minneapolis: University of Minnesota Press, 1982: 251-269

LUSTED LB. Introduction to Medical Decision Making. Springfield, Illinois, Charles C Thomas, 1968

LUSTED LB. Twenty Years of Medical Decision Making Studies. CH1480-3/79/0000-0004$00.75. 1979 IEEE

POPLE HE. Heuristic Methods for Imposing Structure on Ill-structured Problems: The Structuring of Medical Diagnosis. In: Szolovits P, ed. Artificial Intelligence in Medicine, AAAS Symposium Series, Boulder, Colorado: West-view Press, 1982: 119-185

MARTIN J: Computer Data-Base Organization. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1977

CHAKRAVARTY S and SHAHAR Y. A Constraint-Based Specification of Periodic Patterns in Time-Oriented Data. Sixth International Workshop on Temporal Representation and Reasoning (TIME-99), Orlando, FL, 29-40. 1999. SMI-1999-0766

ALTMAN RB. AI in Medicine: The Spectrum of Challenges from Managed Care to Molecular Medicine. AI Magazine 20(3):67-77, 1999. SMI-1999-0770

BLEICH HL. Computer-Based Consultation: Electrolyte and Acid-Base Disorders. The American Journal of Medicine 53: 285-291, November 1972

BLOIS MS, TUTTLE MS, and SHERERTZ DD. RECONSIDER: A Program for Generating Differential Diagnoses. IEEE: 263-268, 1981

CHAKRAVARTY S and SHAHAR Y. Acquisition and Analysis of Periodic Patterns in Time-Oriented Clinical Data. 2000. SMI-2000-0822

DE DOMBAL FT, LEAPER DJ, STANILAND JR, McCANN AP, and HORROCKS JC. Computer-Aided Diagnosis of Acute Abdominal Pain. British Medical Journal 2: 9-13, 1972

ESHELMAN L, EHRET D, McDERMOTT JP, and TAN M (1987.) MOLE: A Tenacious Knowledge Acquisition Tool. International Journal of Man-Machine Studies, 26: 41-54

GENNARI J, MUSEN MA, FERGERSON RW, GROSSO WE, CRUBEZY M, ERIKSSON H, NOY NF, TU SW: The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. 2002. SMI-2002-0943

GINI RA and FEDER C. Informatica Clinica: Presente y Futuro. La Semana Medica (Argentina) 157: 113-128, 1980

GORRY GA, KASSIRER JP, ESSIG A, and SCHWARTZ WB. Decision Analysis as the Basis for Computer-Aided Management of Acute Renal Failure. The American Journal of Medicine 55: 473-484, 1973

GORRY GA, PAUKER SG, and SCHWARTZ WB. The Diagnostic Importance of the Normal Finding. The New England Journal of Medicine 486-489, March 2, 1978

GREENS RA, PELEG M, BOSWALA AA, TU S, PATEL VL, and SHORTLIFFE EH. Sharable Computer-Based Clinical Practice Guidelines: Rationale, Obstacles, Approaches, and Prospects. Medinfo, London, UK, 2001. SMI-2001-0860

HUANG K, HENRION M: Efficient Search-Based Inference for Noisy-OR Belief Networks: Top Epsilon. Proceedings of the Twelfth Conference of Uncertainty in Artificial Intelligence, 325-331. Aug 1996, Portland, OR. SMI-96-0640

ILIAD® 4.5 Diagnostic and Reference Tool for Physicians and Medical Professionals. User Guide. 1998

JAAKKOLA TS, JORDAN MI: Variational Probabilistic Inference and the QMR-DT Network. Sun May 9, 16:22:01 PDT 1999

LUDWIG DW. INFERNET – A Computer-Based System for Modeling Medical Knowledge and Clinical Inference. Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care: 243-249, November 1981

MIDDLETON B, SHWE M, HECKERMAN D, HENRION M, HORVITZ E, LEHMANN H, & COOPER G: Probabilistic Diagnosis using a reformulation of the INTERNIST-1/QMR Knowledge Base II. Evaluation of Diagnostic Performance. Section on Medical Informatics Technical report SMI-90-0329, Stanford University, 1990

MILLER RA, POPLE HE, and MYERS JD. INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. The New England Journal of Medicine, 1982: 468-476

MUSSEN MA, GENNARI JH, and WONG WW: A Rational Reconstruction of INTERNIST-I Using PROTÉGÉ-II. Nineteenth Annual Symposium on Computer Applications in Medial Care. New Orleans, LA, 289-293. 1995. SMI-95-0574

MUSEN MA: Modeling for Decision Support. Section on Medical Informatics. Stanford University School of Medicine. Stanford, CA 94305-5479. SMI-98-0739

NOY NF and MUSEN MA. Ontology Versioning as an Element of an Ontology-Management Framework. Stanford Medical Informatics, Stanford University, 251 Campus Drive, Stanford, CA 94305, USA. SMI-2003-0961. March 31, 2003

PATRICK EA: Decision Analysis in Medicine: Methods and Applications. CRC Press, West Palm Beach, FL, 1979

PERLROTH MG, and WEILAND DJ. Fifty Diseases: Fifty Diagnoses. Year Book Medical Publishers, 1981

POPLE HE, MYERS JD, and MILLER RA: Dialog: A Model of Diagnostic Logic for Internal Medicine. Fourth International Joint Conference on Artificial Intelligence. Tbilisi, Georgia, URRS, 3-8 September 1975, Volume Two. 1975: 848-855

PRADHAN M, DAGUM P: Optimal Monte Carlo Estimation of Belief Network Inference. Proceedings of the Twelfth Conference of Uncertainty in Artificial Intelligence, 446-453. Aug 1996, Portland, OR. SMI-96-0638

SHORTLIFE EH. Computer-Based Medical Consultations: MYCIN. American Elsevier Publishing Company, 1976

SHORTLIFFE EH. The Next Generation Internet and Health Care: A Civics Lesson for the Informatics Community. In C.G. Chute, Ed., 1998 AMIA Annual Symposium, Orlando, FL, 8-14. 1998. SMI-98-0730

SHWE M, MIDDLETON B, HECKERMAN D, HENRION M, HORVITZ E, LEHMANN H, & COOPER G. Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base I. The Probabilistic Model and Inference Algorithms. Methods of Information in Medicine, 30(4):241-255, 1991. SMI-90-0296

SZOLOVITS P and PAUKER SG. Categorical and Probabilistic Reasoning in Medical Diagnosis. Artificial Intelligence. 11: 115-144, 1978

TU SW et al: Modeling Guidelines for Integration into Clinical Workflow. Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. SMI-2003-0970

VAN BEMMEL JH, MUSSEN MA (eds): Medical Informatics. Springer, 1997

WEINSTEIN MC and FINEBERG HV: Clinical Decision Analysis. W. B. Saunders Company, 1980

WEISS S, KULIKOWSKI CA, and SAFIR A. Glaucoma Consultation by Computer. Comput. Biol. Med. 8: 25-40, 1978