Computer Diagnosis of Goiters.
The Optimal Size of Optimal Subsymptomatologies

A. Bouckaert\textsuperscript{1,2}

Received July 1973; revised March 1974

Optimization for diagnostic recognition rate was performed for subsets of symptoms of various sizes. The diagnostic problem was the recognition and identification of thyroid diseases. Unbiased evaluation of performance was obtained and the extent of the bias in other evaluation methods was determined. Interdependence of symptoms was shown to be a negligible nuisance in the application of Bayesian inference to the present data. An optimal size of optimized subsets of symptoms was observed. A comparison with sequential diagnosis shows that the two procedures are different, although they are related, and that the optimality of subsets is sensitive to departures from their composition.

1. INTRODUCTION

In a previous part of this study\textsuperscript{(3)} we had suggested that restricted subsets of symptoms could be used for the diagnostic screening of goiters. Our impression at the time was that it would be easier to achieve a good diagnostic accuracy by direct application of Bayes' rule of inference to subsymptomatologies rather than by the application of an empirical Bayesian rule—assuming the mutual independence of symptoms—to the whole symptomatology available.

In order to check if this impression was correct, diagnostic efficiencies will be determined in the present study for a subset of the same data, taking into account: (a) the size of the subsymptomatologies; (b) the symptom

\textsuperscript{1} Laboratory of Physiopathology, Faculty of Medicine, Kinshasa, Zaire.

\textsuperscript{2} Present address: Faculty of Medicine, Catholic University of Louvain, Brussels, Belgium.
composition of the subsymptomatologies; (c) the endorsement or the rejection of the hypothesis of symptom mutual independence; (d) the method of evaluation of the diagnostic efficiency.

2. MATERIAL AND METHODS

2.1. Data

The sample of 86 goitrous patients used in the previous parts of this study was retained for the present experiments. Nine physical symptoms were selected for the data matrix: (i) time elapsed since the beginning of the complaints, (ii) weight loss, (iii) nervousness, (iv) thermophobia, (v) surface of the goiter, (vi) consistency of the goiter, (vii) thrill over the thyroid region, (viii) exophthalmia, and (ix) lymphadenopathy. The patients whose record did not include all nine symptoms were discarded. Sixty-six cases were thus included in the final sample, with the following distribution of diagnoses: 14 cases of thyroid cancer, 27 cases of toxic goiter, 7 cases of nontoxic nodular goiter, 10 cases of nontoxic diffuse goiter, and 8 cases of toxic adenoma.

2.2. Probabilities

The estimation of a priori and conditional probabilities was made according to the method suggested by Bailey\(^{(a)}\) for small-sized samples, i.e.,

\[
P(D_j) = \frac{[N(D_j) + 1]}{(J + k)}
\]

\[
P(S_i/D_j) = \frac{[N(S_i/D_j) + 1]}{[N(D_j) + m]}
\]

where:

- \(P(D_j)\) is the a priori probability of diagnosis \(D_j\).
- \(P(S_i/D_j)\) is the conditional probability of symptom—or subsymptomatology—\(S_i\) for diagnosis \(D_j\).
- \(N(D_j)\) is the number of observed cases with diagnosis \(D_j\).
- \(N(S_i/D_j)\) is the number of cases where \(S_i\) and \(D_j\) are found to occur simultaneously.
- \(J\) is the total number of cases.
- \(k\) is the number of diagnoses.
- \(m\) is the number of distinct values of \(S_i\). For the time since the beginning of the complaints, the surface of the goiter, and its consistency \(m = 3\). For all the other symptoms \(m = 2\). For a subsymptomatology including, e.g., four binary symptoms, \(m = 16\). For the whole symptomatology \(m = 1728\).