AN EXCEL PROGRAM FOR CALCULATING AND PLOTTING RECEIVER-OPERATOR CHARACTERISTIC (ROC) CURVES, HISTOGRAMS AND DESCRIPTIVE STATISTICS

GYÖRGY GÖRÖG*
Navix Inc., 542 Flynn Road, Lincoln Technology Park, Camarillo, CA 93012, U.S.A.

(Received 6 July 1993; in revised form 28 December 1993; received for publication 21 February 1994)

Abstract—Receiver-operator characteristic (ROC) curves have been used increasingly to assess the performance of clinical laboratory tests and to determine suitable positive/negative threshold values. The Microsoft Excel 4.0 program Plot.ROC was designed to provide formatted, publication quality output with minimal user involvement. It asks for a few parameters at the beginning and then runs autonomously, creating ROC charts, histograms and descriptive statistics for each test chosen, and a cumulative ROC chart to compare tests. Output can easily be formatted further using Excel's graphics and text formatting capabilities.

INTRODUCTION

ROC curves have been used increasingly to assess the performance of clinical laboratory tests and to determine suitable threshold values [1–3]. This graphical method is especially suitable to compare the performance of two or more tests, or the same test applied under different circumstances (e.g. at different times after a myocardial infarct). The ROC curve is a plot of sensitivity against 1-specificity as the positive/negative cut-off point is varied.

Among the computer programs available for the construction of ROC curves is a Lotus spreadsheet [4] that requires the user to enter cut-off values manually. The author created a spreadsheet program (macro) that provides ROC curves, histograms and summary statistics automatically. The spreadsheet Excel (Microsoft Corporation) was chosen because it runs under Windows on IBM PC compatibles and the Macintosh as well, and its excellent formatting and graphics capabilities, together with its large number of built-in statistical functions, make it a valuable tool to analyze and present laboratory data.

The program Plot.ROC was designed to provide formatted, publication quality output with minimal user involvement. It asks for a few parameters at the beginning and then runs autonomously, creating ROC charts, histograms and descriptive statistics for each test chosen, and a cumulative ROC chart to compare tests. Output can easily be formatted further using Excel’s graphics and text formatting capabilities. Although the Excel macro language is not an easy programming tool, extensive comments have been provided to allow users to modify the program to their needs.

COMPUTER PROGRAM

The program requires Excel 4.0 and is provided in the macro sheet ROC.XLM. For histograms and descriptive statistics, the Excel Analysis Tools should be installed.

* Present address: Laborexpert Ltd., H-1015 Budapest, Csaloga'ny u. 22–24, Hungary.
Fig. 1. ROC curves generated by the Excel program Plot.ROC showing the relative usefulness of different tumor markers.

Patient data should be provided in the active Excel worksheet. Excel is capable of importing data directly from dBASE and related databases, and most other spreadsheets like Lotus. However, some of these programs (notably dBASE) do not distinguish between true zero values and missing values; to overcome this difficulty, an option has been provided to treat zeros as missing values. Patient data should start in the upper left cell of the sheet, and form a contiguous area (no empty rows). Preferably the sheet should not contain any other data; the program clears and uses for output 40 columns right to the database. This limits the number of tests to around 200. The number of patients is limited by Excel to 16 383. The first row should contain the field names. One of the fields should contain the codes for normal and diseased patients (here termed Normals and Pathols); the respective codes should be typed in the dialog box appearing shortly after the program has been started by pressing ctrl-R. Any (and any number) of the other fields can be chosen for analysis. Histograms and descriptive statistics are optional.

Plot.ROC first checks whether the codes given as Normals and Pathols exist in the column designated as the code field. Then it starts to vary the positive/negative cutoff (threshold) starting from the minimal test value found until the maximal one is reached. Ten values will be within the Normals range and ten between the maximal Normals value and the maximal Pathols value; this way the whole range of specificity and sensitivity is covered evenly. For each cut-off value, true and false negatives and positives are counted; sensitivity, specificity, and test efficiency are calculated. The ROC curve (sensitivity vs 1-specificity) is plotted, and the ROC curve of each test is added to the cumulative ROC chart (Fig. 1). The area under the curve is estimated using a simple trapezoidal approximation [5].

If histograms were requested, the program extracts Normals and Pathols from the database (treating zeros as missing values), specifies 20 categories (can be changed in ROC.XLM) and calls the standard Excel function HISTOGRAM. The output of this function is modified and plotted to allow a direct comparison of the normal and pathological values; a bar histogram is displayed.

If descriptive statistics was requested, the standard Excel function DESCR is called; its output is formatted for an easier comparison of normal and pathological values.

**EXAMPLE**

To demonstrate the usefulness of the program, a relatively large database (courtesy of Dr B. Schumann) was used. It contained data of 382 healthy donors and 2065 tumor
Excel program

patients (442 with ovary carcinoma, 256 with bladder carcinoma, 885 with lung cancer, 235 with gastric cancer and 247 with mammary carcinoma), all tested for six tumor markers (lipid bound sialic acid, neopterine, carcinoembryonic antigen, putrescine, spermidine and spermine). The database was maintained in dBASE III and was opened from Excel directly. All tests were requested for ROC; output was complete in 4 min on a 25 MHz 486DX IBM PC compatible. With histograms and descriptive statistics, the run time was 12 min. As shown in Fig. 1, LSA (lipid bound sialic acid), a general tumor marker [6], provided the best discrimination between healthy donors and cancer patients.

The program is available from the author in an IBM PC format.

SUMMARY

In summary, the article describes a computer program that creates ROC curves, statistics and histograms of clinical laboratory data. The program was written in Microsoft Excel 4.0. It requires minimal user involvement and produces publication quality graphics output.

REFERENCES


About the Author—GYÖRGY GÖRÖG graduated in 1980 as a biologist in Budapest. He has worked in the fields of cellular immunology (T-cell differentiation) and immunochemistry (homogeneous immunoassay) at various basic and applied research institutions in Hungary and Italy. He received his Ph.D. (C.Sc.) in 1992. He is currently a Project Leader, Immunoassay Development at Navix Inc.; his interests include the immunodiagnosis of infectious diseases and malignancies (tumor markers). Mr Görög is the author of a number of statistical and data handling computer programs, among them one for kidney transplantation (HLA matching, survival statistics, etc.) and another for decision support for the clinical laboratory (learning systems, pattern recognition).