ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 974–981

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Scaled rectangle diagrams can be used to visualize clinical and epidemiological data

Roger J. Marshall*

Section of Epidemiology and Biostatistics, School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, New Zealand

Accepted 5 January 2005

Abstract

Objective: To illustrate scaled rectangle diagrams as a method for displaying clinical and epidemiological attributes (such as symptoms, signs, results of marker tests, disease, or risk factors). These are quantitative Venn diagrams, but using rectangles instead of circles.

Study Design and Setting: The method is illustrated through examples from various data sets with different types of clinical information.

Results: Examples drawing on studies of lung disease, rheumatic fever, blood pressure, lipid levels, sudden infant death syndrome, and low birth weight illustrate the different types of relationships between variables that the scaled rectangle approach can reveal (e.g., high- and low-risk groups; dependent, independent, or co-occurring attributes; effects from choice of cutoff; cumulative distributions; and case-control attributes).

Conclusion: Scaled rectangle diagrams are a novel way to display clinical data. They show clearly the relative frequency of clinical attributes and the extent to which they are shared characteristics. Features are revealed that might otherwise not have been appreciated. © 2005 Elsevier Inc. All rights reserved.

Keywords: Diagnostic logic; Hypertension; Case–control; Clinical data; Statistics

1. Introduction

Venn diagrams have often been suggested in clinical research to symbolically show how features such as clinical signs, symptoms, tests, and staging coincide. Feinstein [1] pioneered their use. Some authors have tried scaling Venn diagrams so that the circles and their intersections are drawn in proportion to frequencies of occurrence. Feinstein [1], however, noted the technical difficulty in creating such proportional Venn diagrams and observed that, although they may demonstrate quantitative relationships, the qualitative aspects of the relationships may be less clear than with a purely symbolic nonproportional Venn diagram. In an earlier publication [2], I have shown that the technical difficulty is reduced somewhat by using rectangles instead of circles. The resulting scaled rectangle diagram often show both qualitative and quantitative information reasonably well.

Here, examples of the analysis of clinical and epidemiological data will illustrate the use of scaled rectangle diagrams. With examples used as case studies, some of the ideas I had earlier presented [2] will be advanced in a number of ways. For example, despite my original suggestion that four attributes may represent a limit [2], it is sometimes possible to represent more. New criteria for fitting scaled rectangle diagrams use a penalizing function to avoid narrow rectangles with a high length-to-breadth ratio. Color or shading can be used to represent the intensity of an additional variable, and scaled rectangle diagrams can be used to represent combinations of attributes and high- and low-risk subgroups. In some examples, scaled rectangle diagrams will be compared with Feinstein's original Venn diagrams. Other examples draw on studies of hypertension, lipids, sudden infant death syndrome, and low birth weight to illustrate the method.

2. Theory and methods

Suppose there exist *attributes* (e.g., signs, symptoms, or tests) that are measurable on patients and there is a data set with recorded presence of attributes in a sample of patients. The relative frequency, or prevalence, of each attribute can be computed and rectangles drawn with area proportional to prevalence. A scaled rectangle diagram is formed by positioning these rectangles so that not only are the areas of the rectangles proportional to attribute prevalence, but

* Corresponding author.
E-mail address: rj.marshall@auckland.ac.nz (R.J. Marshall).

0895-4356/05/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2005.01.018

also with areas of overlap that are proportional to the relative frequencies of different combinations of attributes. It is sometimes informative to place the resulting configuration inside a unit square, which represents the entire data set, or *universe*.

Creating such a diagram often is easy by simple geometric construction for just two or three attributes, and also when attributes are nested (e.g., the attributes 'systolic blood pressure > 120 mmHg' and 'systolic blood pressure > 130 mmHg'). It is more difficult with more than three non-nested attributes, which requires maximizing a measure of discrepancy, D, between frequency and area. It may not always be possible to obtain exact congruence of areas and frequencies.

In general, suppose that a diagram to represent q attributes is required. Let θ be the set of coordinates of the q rectangles. If diagrams are constructed to ensure that each attribute rectangle has an area exactly proportional to frequency, then there are 3 q-values in θ, given that three values determine a rectangle's position (the x,y coordinates of one vertex and the length of one side). There are 2^q possible combinations of attributes, or *cells*. Let $r_i = f_i / n$ be the relative frequency of cell i ($i = 1, \ldots, 2^q$) and let $a_i(\theta)$ be an area representing this combination. Different criteria for measuring discrepancy D between cell area and frequency can be proposed. I have suggested [2] the sum of the discrepancies between $a_i(\theta)$ and r_i (i.e., $D = \sum_i |a_i(\theta) - r_i|$). Alternative measures are squared discrepancies (least squares) $D = \sum_i (a_i(\theta) - r_i)^2$, or log-likelihood $D = -\sum_i f_i \log a_i(\theta)$. Creating, or *fitting*, a scaled rectangle diagram requires minimizing D (or, in the case of log-likelihood, maximizing D) with respect to θ. It is important to note that, although exact congruence between cell frequency (f_i) and area (a_i) may not be obtained, the construction ensures that areas of each corresponding attribute rectangle are exactly congruent. The absolute discrepancy measure, expressed as a percentage (i.e., $E = \sum_i |a_i(\theta) - r_i| \times 100$) is a measure of cell error; where nonzero, it is reported in the examples.

Because visual perception of the areas of narrow (or *thin*) rectangles is harder than for more square rectangles, an attempt to penalize for narrow rectangles can be introduced. Specifically, suppose a *thinness* measure of a rectangle is the length-to-breadth ratio. Let t be the measure of the thinnest of the q rectangles. Then, penalizing the congruence measure D by minimizing Dt^a for a parameter a attaches less weight to configurations with thin rectangles. From experience, using $a = 0.05$ together with the minimum absolute discrepancy criterion for congruence seems generally to avoid narrow rectangles while not unduly inflating mismatch between area and frequency. The log-likelihood criteria may give rectangles closer to square, but at the expense of discrepancy error.

Software to create scaled rectangle diagrams is available in the SPAN package [3]. It allows each of the three D measures. The scaled rectangle diagrams presented here are cut-and-pasted computer images, direct from the software (although in some instances the text of the labels has been edited).

3. Examples

3.1. Feinstein's lung disease example

Feinstein [1] describes a sample of 175 patients with lung disease. Patients were classified with the characteristics emphysema (EMPH), mucous gland hyperplasia (MPH), 'blue and bloated' (BB), 'pink and puffy' (PP), and with neither BB nor PP (NN). Feinstein presented two attempts to draw a quantitative Venn diagram; the better of these is reproduced in Fig. 1a. Figure 1b is a scaled rectangle diagram for these data. It shows the relative sizes of the subgroups more accurately than Feinstein's original diagram. For example, EMPH is larger than the MGH groups and the

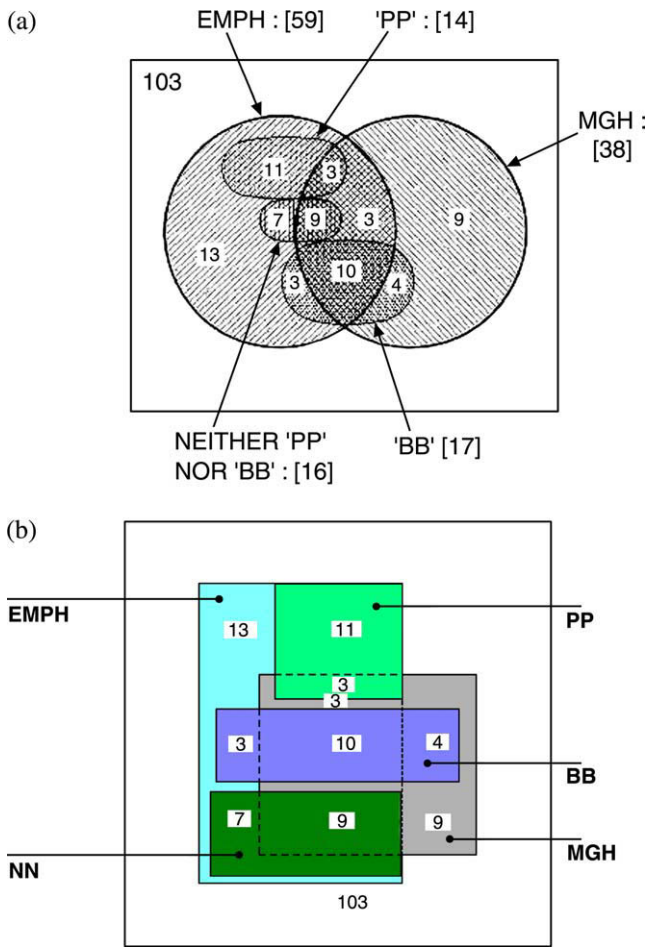

Fig. 1. Sample of 175 patients with lung disease. Attributes are: emphysema (EMPH), mucous gland hyperplasia (MGH), 'blue and bloated' (BB), 'pink and puffy' (PP), and neither BB nor PP (NN). (a) Feinstein's [1] original Fig. 29 (reproduced with permission of the publisher, Lippincott Williams & Wilkins). (b) Corresponding (exactly) scaled rectangle diagram.

three nested PP, BB, and NN groups are each similar in area. The intersection of the EMPH and MGH is larger than suggested by Fig. 1a.

3.2. Feinstein's rheumatic fever example

Figure 2a is a reproduction of a scaled Venn diagram, from Feinstein's Fig. 17 [1], showing prognostic attributes in a group of 441 patients with acute rheumatic fever. The relative frequency that is written in each cell of the diagram represents prognosis, that is, the proportion of patients with clinical evidence of acute rheumatic fever 8 years later. A corresponding scaled rectangle diagram is shown in Fig. 2b. Instead of shading to discriminate the individual attribute rectangles, as in Fig. 1a, here shading is done according to the value of the 8-year prognosis in each cell, so that the shading represents the proportions that Feinstein wrote in each cell. It is sometimes useful to shade the diagrams in this way—that is, by the intensity of some other variable that may be related to the rectangle attributes. Here it shows quite clearly that severe carditis confers the poorest long-term prognosis.

Unlike the example in Fig. 1, all individuals in the data set possess at least one of the five attributes, so that it is unnecessary to place the configuration in a unit square, because there is no longer any white space. There is some error between cell area and frequency in both Figs. 2a and 2b, but the errors in Fig. 2b are much less. For example, 'chorea' is considerably less common than is indicated by Fig. 2a and relative sizes of 'chorea' and 'severe carditis' are quite incorrect. 'Arthralgia' in Fig. 2a is drawn showing the possibility that it occurred alone, even though there were no patients with just arthralgia. The constraints of trying to construct a diagram with circles presumably made this

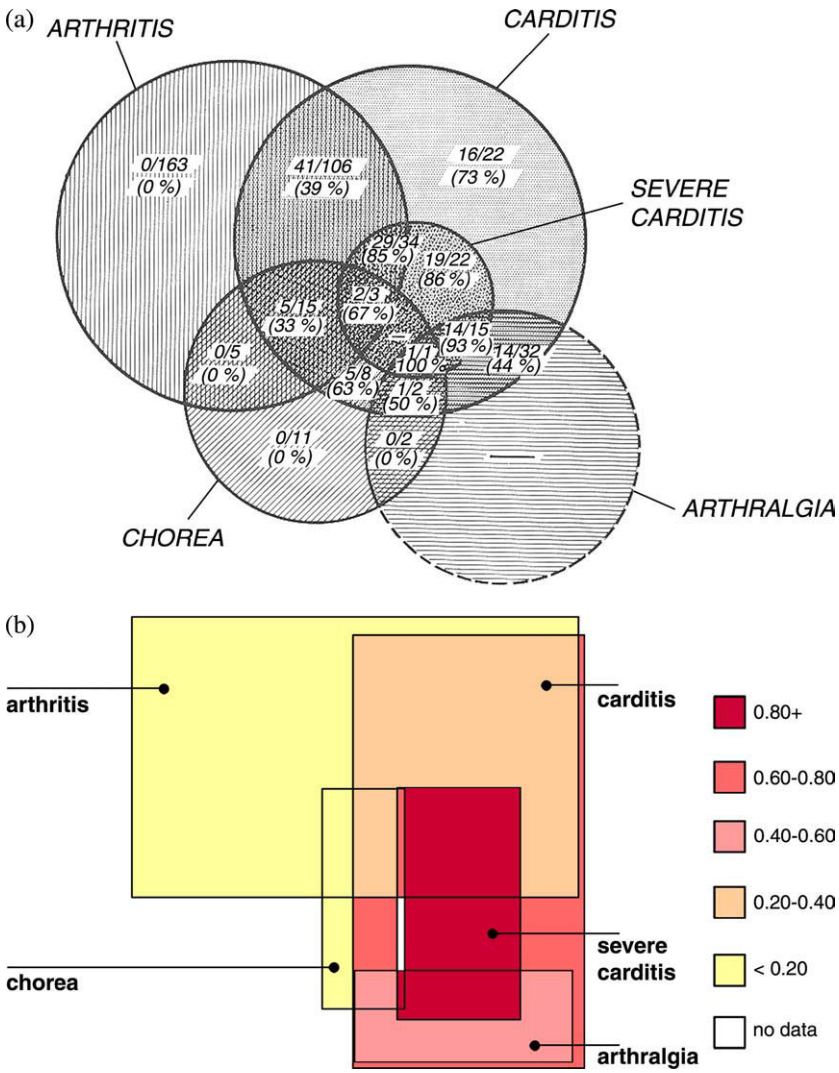

Fig. 2. (a) Scaled Venn diagram of patients with acute rheumatic fever (reproduced from Feinstein's [1] Fig. 17, with permission of the publisher, Lippincott Williams & Wilkins). Values indicate proportion with clinical evidence of rheumatic fever after 8 years. (b) Corresponding scaled rectangle diagram (E = 3.7%), shaded according to that same proportion.

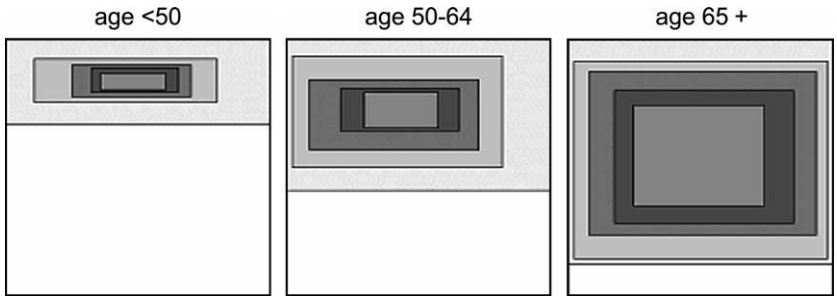

Fig. 3. Frequencies of elevated systolic blood pressure with increasing age in a sample of Auckland adults [4], with cutoffs at the five levels 120, 130, 140, 150, and 160 mmHg.

unavoidable. Although there is an exact correspondence between area and frequency of the attribute rectangles in the scaled rectangle diagram, there is some cell error. For example, there are no observations at all in one cell in the center of the diagram.

3.3. Cutoffs for elevated systolic blood pressure

One way that scaled rectangle diagrams can be used is to show the cumulative distribution of a variable. For example, Fig. 3 provides a way to visualize the shrinking prevalence of 'elevated systolic blood pressure' with increasing cutoffs. Data are of systolic blood pressure (BP) in 2,485 men and women recorded in a cross-sectional study conducted in Auckland, New Zealand [4]. Obviously, because the group with elevated systolic BP $>$ 130 mmHg is a subset of those with elevated BP $>$ 120 mmHg, and so on, the rectangles must be nested. The diagrams in Fig. 3 show the prevalence of elevated systolic BP at different ages and, looking across the three diagrams, they demonstrate how prevalence, for each cutoff, increases with age.

3.4. Diastolic and systolic blood pressure

Using the same data as in section 3.3, the joint occurrence of elevated systolic BP and elevated diastolic BP is shown in Fig. 4 (for all ages combined). It demonstrates that elevated values of diastolic BP tend to occur in conjunction with elevated systolic, but elevated systolic BP does not necessarily coincide with elevated diastolic. For example, a definition of hypertension based only on 'systolic BP $>$ 140 mmHg' would determine almost the same group of people as a definition based on 'systolic BP $>$ 140 or diastolic BP $>$ 90'. With a diastolic BP cutoff equal to 80 mmHg, however, there are many people who do not have elevated systolic BP.

3.5. Lipid distribution

Serum lipid levels were also measured in the Auckland study used in examples 3.3 and 3.4. The relationship between

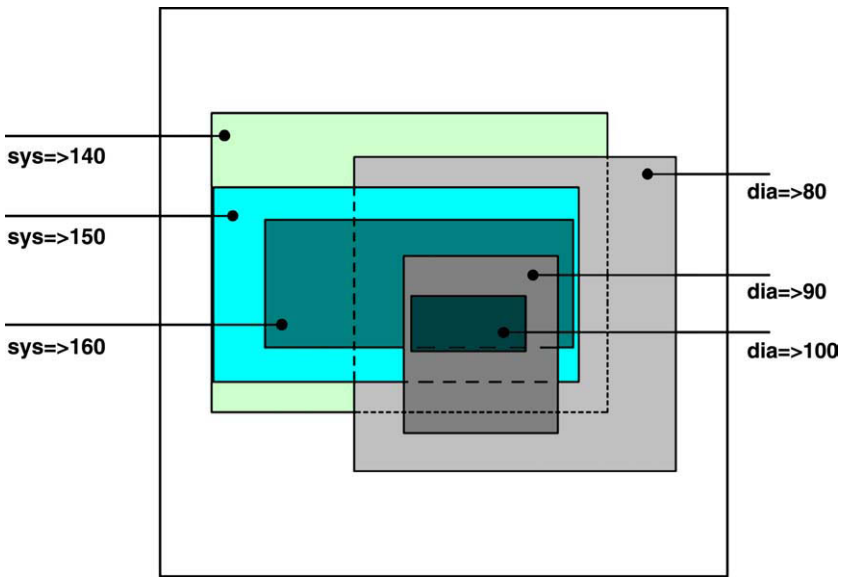

Fig. 4. Different cutoffs for elevated diastolic and elevated systolic blood pressure (E = 0.5%) in the Auckland sample [4].

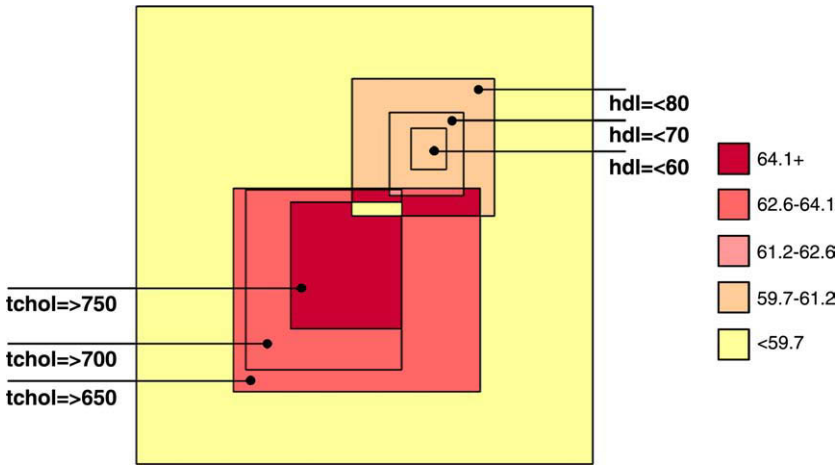

Fig. 5. HDL and total cholesterol (tchol) levels in the Auckland sample [4], shaded according to average age ($E = 1\%$).

total cholesterol and high-density lipoprotein (HDL) cholesterol is examined in Fig. 5. Because low values of HDL are considered adverse, rectangles for HDL are drawn for values below selected cutoffs. The diagram shows that low HDL and high total cholesterol tend not to co-occur often, suggesting independence. If, for example, total cholesterol > 650 and HDL < 80 are independent attributes, the expected overlap of the total cholesterol > 650 and HDL < 80 rectangles would be ~2.3% of the unit square. In Fig. 5, the intersection is ~1.8%, not too different from expected ($\chi^2 = 2.45$, $P = 0.11$).

Figure 5 is shaded according to the average age of respondents in each cell of the diagram. It shows that the average age of people with low HDL, but without elevated total cholesterol, is relatively young. People with elevated total cholesterol, or with both elevated total cholesterol and low HDL, tend to be older. The youngest in the sample are people with low total cholesterol and elevated HDL (the surrounding area). There is, however, an anomalous cell with total cholesterol over 750 mmol/L and HDL between 70 and 80 mmol/L (light cell in the center of the diagram). It is probably random variability.

3.6. Sudden infant death syndrome cases and controls

It is sometimes useful to compare scaled rectangle diagrams in different groups of people (as in Fig. 3). Another example is to compare active and control groups in a clinical trial or epidemiological investigation. Figure 6 shows, in a maternal case-control study of sudden infant death syndrome (SIDS) in New Zealand [5], the occurrence of low birth weight (<2.5 kg) and ethnic subgroups (Maori and Pacific Islander) in cases and controls. Comparing the two diagrams for cases and controls clearly demonstrates the overrepresentation of Maori and low birth weight among cases; however, Pacific Islanders are under-represented among cases and controls. Because a person cannot (in the definitions utilized) be a Maori and Pacific Islander at the same time, the rectangles for these two groups must not, of course, overlap.

3.7. Rectangles as attribute combinations, high- and low-risk subgroups

A rectangle may also be used to represent a combination of attributes. For example, the composite attribute ‘elderly hypertensive Maori’ defines a subgroup derived from age, BP, and race attributes. If one is interested in relationships between subgroups, coalescing attributes in this way is sometimes useful. Such subgroups may be of special interest, and specified by the research question, but they may also arise from statistical analysis. For example, a method (known as search partition analysis, SPAN) has been proposed to identify combinations of attributes that define high- and low-risk subgroups [6]. Visualizing these subgroups was, in fact,

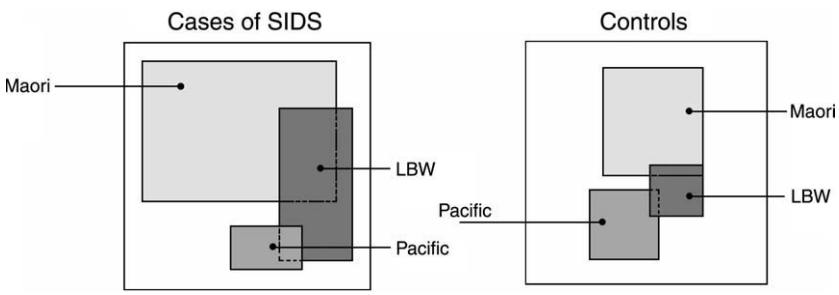

Fig. 6. Ethnic group: Maori and Pacific people and low birth weight < 2.5 kg (LBW) in SIDS cases and controls in a case-control study [5].

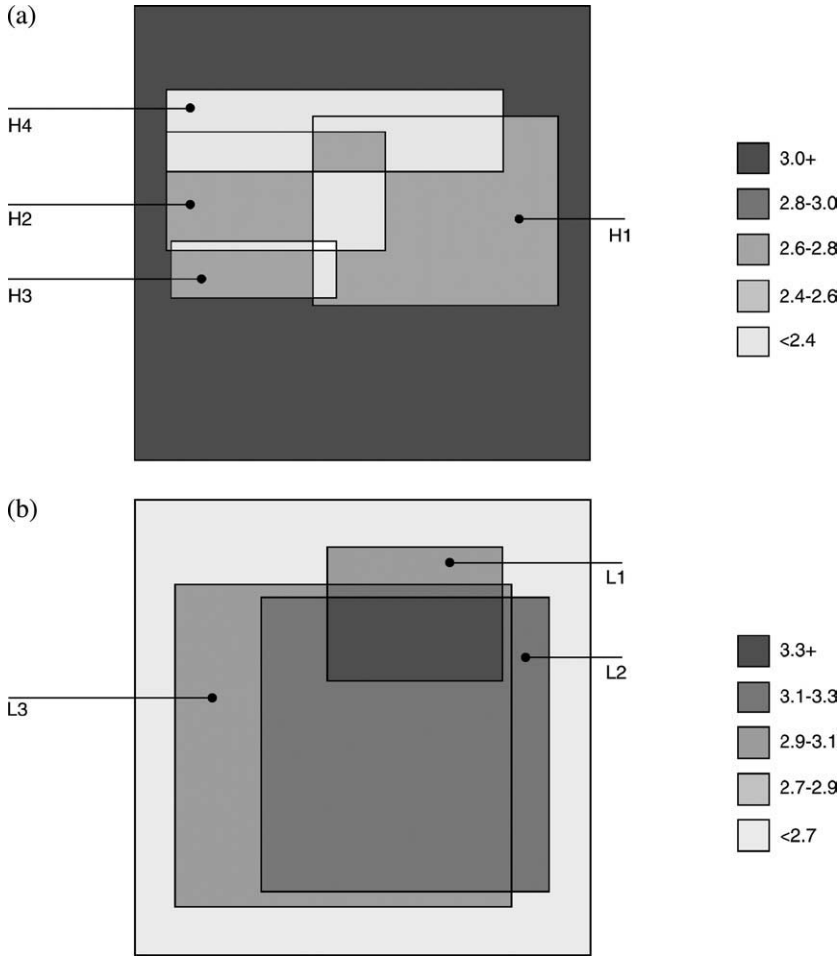

Fig. 7. High- and low-risk subgroups derived by a SPAN [6] analysis of birth weight data [7]. (a) High-risk subgroups for low birth weight shaded according to birth weight (kg). See text for definitions of H1, H2, H3, and H4. ($E = 2.0\%$).

the original impetus for developing scaled rectangle diagrams.

A well known data set concerns predictive maternal factors of low birth weight [7]. For these data, a SPAN analysis (omitting details) identifies a high risk of low birth weight combination as

H1 or H2 or H3 or H4

Each of H1, H2, H3, and H4 defines a subgroup that can represented by the rectangle of a scaled rectangle diagram, as shown in Fig. 7a. Here H1 is maternal weight $<$ 49.5 kg (109 pounds, in the source data [7]), H2 is age $<$ 30 with uterine irritability; H3 is age $<$ 30 with past history of premature labor; and H4 is black, age $<$ 30 and smokes. The cells of the diagram are shaded according to the average baby birth weight in each, clearly showing that each of these combinations confers low birth weight.

The above combination identifies low birth weight subgroups. Its complement identifies groups with higher birth weight. After some Boolean algebra, the complement can be expressed as a union of low-risk subgroups L1, L2, and L3:

L1 or L2 or L3

where L1 is maternal weight $>$49.5 kg and age $>$30 years old; L2 is $>$49.5 kg, nonsmoking, no uterine irritability, and no past history of premature labor; and L3 is $>$49.5 kg, no uterine irritability, no past history of premature labor, and not black. These low-risk groups can also be represented by a scaled rectangle diagram, as in Fig. 7b.

Note that the dark surrounding area in Fig. 7a represents precisely the same people as the area represented by the union of the three rectangles in Fig. 7b. Conversely, the light surrounding area in Fig. 7b is the same as the union of the four rectangles in Fig. 7a. Again, Fig. 7b is shaded by average birth weight so that, as expected, the groups L1, L2, and L3 confer higher birth weight, the highest for the intersection of all three groups.

4. Discussion

Clinical and epidemiological researchers are often interested in the extent to which signs, symptoms, results of tests,

risk factors, and other clinical observations co-occur. Often an appreciation of the actual extent of co-occurrence is lost by the reduction to summary measures of association, which are usually intended to convey departure from statistical independence; however, although independence is a fundamental concept, it is instructive to actually visualize the co-occurrence of attributes, and scaled rectangle diagrams provide a useful way to do so. A scaled rectangle diagram is essentially a type of Venn diagram that goes beyond the purely symbolic qualitative display often adopted in discourses of clinical logic, to also show quantitative information.

In a formal statistical sense, a scaled rectangle diagram is a display of a 2^q cross-tabulation of q binary categorical variables. It shows the nonempty cells of the table as well as the marginal totals of the q attributes. Other methods to display categorical data have been developed: for instance, mosaic plots [8–11]. These show only the cells of the table as discrete rectangles and the viewer has mentally to piece them together to appreciate the marginal attribute frequency. An alternative graphic representation of a 2^2 table is the two by two diagram [12]. this differs from both a mosaic plot and a scaled rectangle diagram, in that frequency is represented by length, rather than area.

Scaled rectangle diagrams rely on the ability of the eye and brain to correctly perceive the relative magnitude of the areas of the diagram and to 'see' them (interpret them) as frequencies. In practice, people do not generally perceive areas especially well; they tend to underestimate actual areas [13,14]. Also, perception of area depends on shape: generally, it seems easier to gauge the area of a square than a long, narrow rectangle and, for this reason, some attempt is made in the construction of scaled rectangle diagrams to avoid narrow rectangles.

Rectangles, instead of circles, are used to represent attributes because algorithms for their construction are relatively easy: all areas can be disaggregated into smaller rectangles, and those areas easily calculated [2]. Using circles would create crescents, lenses, and other odd shapes made from joined arcs of circles, whose areas would not only be difficult to compute (except perhaps by Monte Carlo computer methods), but also hard for the viewer to perceive. Another simple advantage to using rectangles instead of circles is that of coverage: with four circles it is possible to draw only a four-circle Venn diagram to cover 14 out of the potential $2^4 = 16$ cells, but four rectangles potentially allow all 16 cells.

In some cases, there may be no way to produce a configuration that will give a scaled rectangle diagram in which cell areas are perfectly congruent to cell frequencies, although the way the diagrams are constructed ensures that areas of attribute rectangles themselves are exactly congruent. Some nonzero measure of cell error may remain [2], depending on the number and the interrelationships between attributes. Nonetheless, given that the eye and brain may not precisely judge areas, some small error in construction may not matter unduly.

There is no unique scaled rectangle diagram for a set of data. There may be an infinite number of ways of placing the rectangles to obtain a good representation of the data. In the SPAN software for creating these diagrams there is a 'nudge' button that allows a slight shift in the configuration, which may yield a more aesthetically pleasing and interpretable configuration or improve the congruence between area and frequency. Also, the software allows different starting configurations of the iterative fitting process, because convergence on a good fit is sometimes sensitive to the initial configuration. The graphics presented here are, apart from some editing of labels or shading, direct reproductions of the bitmapped images. The software allows either color or monochrome.

Sometimes the congruence between area and frequency may be unacceptable and using different starting configurations may still resist obtaining a well fitting diagram. It would be useful to know when it is and when it is not possible to obtain a good fit, and also a minimum achievable error. This is a difficult mathematical problem, a solution of which eludes the author. There are, however, situations when zero error is unlikely. When $q > 4$, there may be more data combinations than can be represented by cells of overlaid rectangles. For example, with $q = 6$ there are 64 possible data combinations. Superimposing six rectangles cannot give 64 cells, so that, if all combinations are represented in the data, error is inevitable. Where attributes are nested, the problem is lessened. Conversely, there may be no data on a certain combination but a cell representing it remains in the diagram.

The use of computers has made creating statistical graphics very easy and has opened the way for new types of data representation. Making scaled rectangle diagrams without a computer would be a tedious business. A good graphic is said to be one that reveals data features that would otherwise not have been appreciated [15,16]. Judged by this measure, scaled rectangle diagrams can be a useful addition to standard ways of visualizing data.

The SPAN software to produce scaled rectangle diagrams can be downloaded as shareware from http://www.auckland.ac.nz/mch/span.

References

[1] Feinstein AR. Clinical judgment. Baltimore, MD: Williams & Wilkins; 1967.

[2] Marshall RJ. Displaying clinical data relationships with scaled rectangle diagrams. Stat Med 2001;20:1077–88.

[3] Marshall RJ. Search Partition Analysis. SPAN software home page [Internet]. 1999–2001. Available at: http://www.auckland.ac.nz/mch/span.

[4] Bullen C, Simmons G, Trye P, Lay-Yee R, Bonita R, Jackson R. Cardiovascular disease risk factors in 65–84 year old men and women: results from the Auckland University Heart and Health Study. N Z Med J 1998;111:4–7.

[5] Mitchell EA, Scragg R, Stewart AW, Becroft DM, Taylor BJ, Ford RP, Hassall IB, Barry DM, Allen EM, Roberts AP. Results from the first year of the New Zealand cot death study. N Z Med J 1991;104:71–6.
[6] Marshall RJ. Determining and visualising at-risk groups in case-control data. J Epidemiol Biostat 2001;6:343–8.
[7] Hosmer DW, Lemeshow S. Applied logistic regression. 1st ed. New York: Wiley 1990.
[8] Hartigan JA, Kleiner B. A mosaic of television ratings. Am Stat 1984;38:32–5.
[9] Friendly M. Mosaic displays for multi-way contingency tables. J Am Stat Assoc 1994;89:190–200.
[10] Friendly M. Conceptual and visual models for categorical data. Am Stat 1995;49:153–60.
[11] Reidwyl H, Shüpbach M. Parquet diagrams to plot contingency tables. In: F. Faulbaum, editor. Softstat '93: Advances in Statistical Software. New York: Gustav Fischer; 1993:293–9.
[12] Johnson KM. The two by two diagram: a graphical truth table. J Clin Epidemiol 1999;52:1073–82.
[13] Foster JJ. The influence of shape on apparent area: a new demonstration. Acta Psychol 1976;40:103–13.
[14] Ekman G, Junge K. Psychophysical relations in visual perception of length, area and volume. Scand J Psychol 1961;2:1–10.
[15] Tufte ER. The visual display of quantitative information. 1st ed. Cheshire, CT: Graphics Press; 1983.
[16] Wainer H. Visual revelations. New York: Copernicus; 1997.