

Displaying clinical data relationships using scaled rectangle diagrams

Roger J. Marshall*

Department of Community Health, University of Auckland, New Zealand

SUMMARY

A method is presented to draw rectangles to represent categorical data relationships. The idea is an adaptation of a scaled Venn diagram. Rectangles are drawn with area proportional to the frequency of categories and the rectangles are positioned to overlap each other so that the areas of overlap are in proportion to the joint frequencies of the characteristics. The diagrams are especially useful to illustrate symptom co-occurrence. Copyright © 2001 John Wiley & Sons, Ltd.

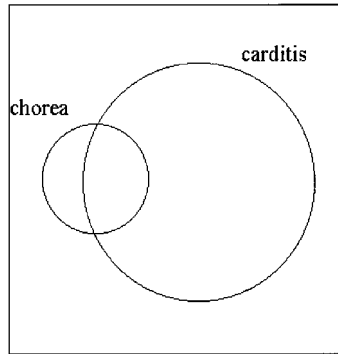
1. INTRODUCTION

Many clinicians appreciate the importance of logic and logical principles in clinical judgement and often use Venn diagrams to show how signs, symptoms and disease co-occur [1, 2]. The purely symbolic set representation of a Venn diagram is sometimes enhanced by attempting to draw circles proportional, in area, to the frequency of the characteristics they represent and, further, overlapping them by an amount proportional to their co-frequency. In this way the diagram conveys statistical as well as symbolic information. Some examples are in references [3–5] and Figure 1 shows an example for the occurrence of carditis and chorea in a sample of 271 acute rheumatic fever patients [5].

When the two circles are placed inside a unit square represents all patients, the diagram is a visualization of a two-by-two contingency table. This simple representation does not seem to be widely appreciated. Other graphical representations of contingency tables include mosaic plots [6–8], parquet diagrams [9], association plots [10] and rose diagrams [11, 12]. In these, emphasis is placed on representing the frequency of the *cells* of the tables. In a mosaic plot cells are represented by discrete rectangles which the viewer has to mentally sum to appreciate marginal frequencies. On the other hand, a scaled Venn diagram places emphasis on the *marginal* frequencies, by the circles, and the different areas within them show the cell frequencies.

Both mosaic and Venn diagram plots rely on the ability of the eye and mind to translate an area into a frequency. In practice ‘there are considerable ambiguities in how people perceive

*Correspondence to: Roger J. Marshall, Department of Community Health, University of Auckland, New Zealand.



	chorea	no chorea	
carditis	14	115	129
no carditis	11	131	142
	25	146	271

Figure 1. Scaled Venn diagram representing subsequent occurrence and co-occurrence of carditis and chorea in a sample of 271 patients acute rheumatic fever. The corresponding 2×2 table that it represents is also shown.

a two-dimensional surface and then convert that perception into a one-dimensional number' [13], but nevertheless the idea is appealing. Shape often determines how well area comparisons are perceived. It is generally accepted that compact shapes are perceived to be smaller than diverse ones of the same area [14]. In Figure 1 the frequency of the *absence* of both chorea and carditis is more difficult to perceive than the other frequencies because of its irregular shape.

Drawing a scaled Venn diagram with two circles is simple, but it is not always possible to enclose them in a unit square. For instance, when all individuals have at least one of the two characteristics it is impossible, since the two circles themselves represent the entire data. Further, extending the circle Venn diagram to three or more characteristics becomes difficult. It is generally not possible to correctly scale and overlap circles so that areas are in proportion to frequencies. Also approximations are often not easy to achieve. With four characteristics it is not even possible to draw and overlap four circles inside a square so as to create the required 16 potentially non-empty cells of the 2^4 table; only 14 are possible.

There is, however, little especially compelling about using circles in Venn diagrams; other shapes can be considered. In this paper I propose using rectangles. Provided rectangles are all oriented in the same direction, areas of intersecting regions are either simply rectangular or made up of rectangular sections which makes computations easy.

The idea is to construct a scaled rectangle diagram (SRD) in which the marginal frequency is represented by a rectangle, drawn with an area proportional to frequency, and positioned to intersect other rectangles so that the overlapping areas are also in proportion to joint frequencies. When there are q characteristics, it is hoped that in this way a representation can be constructed which shows both the non-zero cell frequencies of a 2^q contingency table and also the q marginal frequencies of the characteristics.

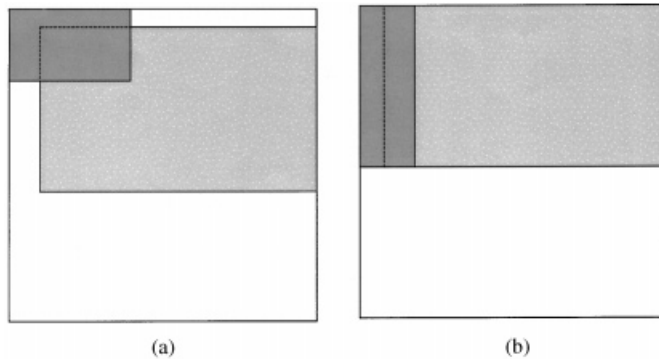


Figure 2. Scaled rectangle diagrams of data in Figure 1.

In Section 2 I give some examples of the method and, in Section 3, I outline how the diagrams can be constructed.

2. EXAMPLES

2.1. Rheumatic fever

Using the data illustrated in Figure 1, an SRD is shown in Figure 2(a). The position of the two rectangles inside the unit square is, of course, arbitrary, as is the extent to which the rectangles are offset. Figure 2(b) is an alternative representation in which they are not offset and which may be preferable as *all* areas are rectangles; both the marginal and cell frequencies are rectangular. However, a drawback is that the edges of the two marginal rectangles coincide, so losing some appreciation of the two distinct characteristics.

Other characteristics of the patients in the rheumatic fever data set [5] included subsequent residual rheumatic heart disease and presentation with arthritis. Figure 3 shows an SRD for carditis and chorea together with the occurrence of residual rheumatic heart disease. Although these diagrams are intended to ‘visualize’ frequencies, it is sometimes also helpful to write frequencies inside the cells, as in Figure 3. I have also included the raw 2^3 table in Figure 3. The SRD clearly shows that the residual heart disease occurred entirely in patients presenting with carditis. This feature is not obvious from the 2^3 table or from the corresponding mosaic plot in Figure 4, in which zero cell frequencies are shown by the vertical lines.

Finally, for these data Figure 5 shows an SRD of the three previous characteristics – carditis, chorea and residual rheumatic heart disease – plus a fourth: presenting with arthritis. Again residual heart disease is seen to occur entirely in patients with carditis. Most patients present with arthritis, but few of them develop residual heart disease. Notice that in this diagram a value $E = 2.7$ per cent is given. This figure is a measure of discrepancy between the representation of areas as frequencies and will be formally defined below. $E = 0$ per cent indicates exact correspondence, as in Figure 3. In Figure 5 the error arises primarily because there are no patients who do not possess at least one of the characteristics, but it does not seem possible to configure four rectangles within a unit square and not have a residual ‘white

	Carditis=yes		carditis=no	
	Rheum_hd		Rheum_hd	
	Yes	no	yes	no
Chorea=yes	5	9	0	11
Chorea=no	49	66	0	131

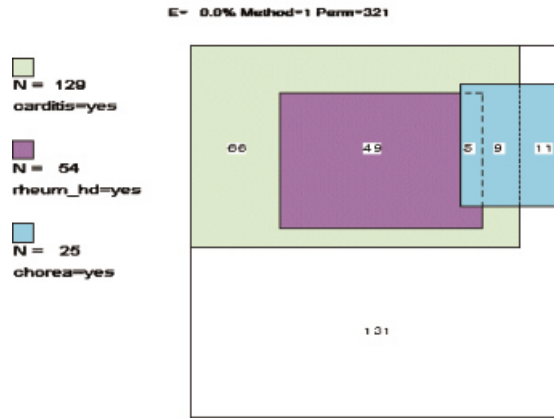


Figure 3. SRD of three characteristics: presence of carditis; presence of chorea, and subsequent rheumatic heart disease.

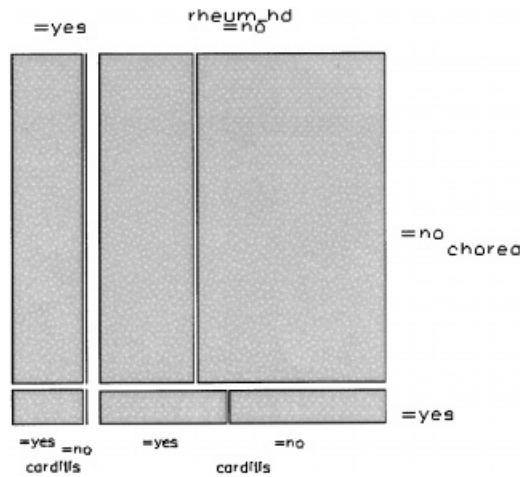


Figure 4. Mosaic plot corresponding to Figure 4.

space' cell for the absence of all four characteristics. One of the problems with creating SRDs for $q > 2$ is that it is not always possible to construct a diagram in which area and frequency are exactly congruent.

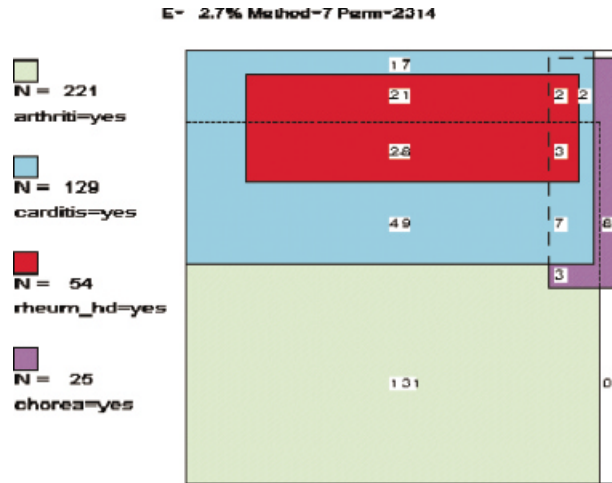


Figure 5. SRD of four characteristics: presence of carditis; presence of chorea; presence of arthritis; and subsequent rheumatic heart disease.

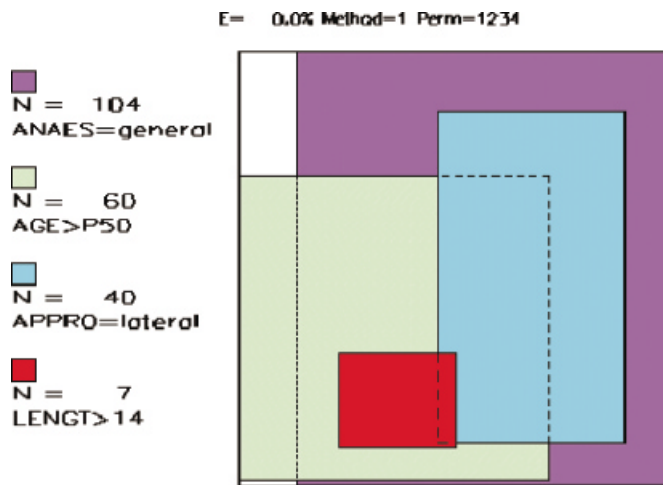


Figure 6. SRD of four characteristics of hip operation data.

2.2. Hip operations

Data on 120 hip operations are displayed in Figure 6. It shows an SRD for four characteristics: general anaesthetic (ANAES = general); lateral surgical approach (APPRO = lateral); age of patient (above the 50th percentile; AGE > P50), and length of stay in hospital more than 14 days (LENGT > 14). In this example an exact representation is obtained ($E = 0$ per cent). It clearly shows that nearly all operations required a general anaesthetic and all longer stay patients were in the older age group.

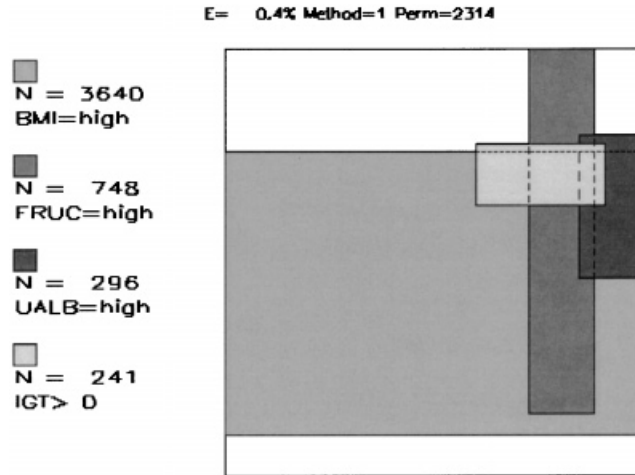


Figure 7. SRD of four characteristics of impaired glucose tolerance data.

2.3. Impaired glucose tolerance

In a work-force survey of 5542 people, glucose tolerance was measured as well as body mass index, urinary albumin and fructosamine. Figure 7 shows an SRD for: impaired glucose tolerance ($IGT > 0$); raised body mass index ($BMI = \text{high}$); elevated urinary albumin ($UALB = \text{high}$), and raised fructosamine ($FRUC = \text{high}$). Clearly, those with impaired glucose tolerance are mostly among the individuals who are more obese, as are people with raised urinary albumin, although there is not much overlap between these groups.

3. CONSTRUCTING SCALED RECTANGLE DIAGRAMS

In general, suppose there are q characteristics. The problem is to draw q rectangles, which each represent one of the characteristics, such that their areas are proportional to the q marginal frequencies and at the same time positioned so that they overlap with areas proportional to cell frequencies of the characteristic combinations. The configuration must then be entirely enclosed in a unit square representing the whole sample.

3.1. The case $q=2$

A construction as in Figure 2(b) is always possible. Using rectangles that are offset, as in Figure 2(a), is obviously not possible when the frequency of the absence of both characteristics is zero since there is always residual white space when offset rectangles are enclosed by a unit square. The extent to which rectangles *can* be offset, as in Figure 2(a), is determined by the frequency of neither characteristic.

3.2. The case $q=3$

There are various geometric constructions that can be attempted to draw a diagram with $q=3$. One of these begins with the $q=2$ construction as in Figure 2(b) and superimposes a third rectangle.

If simple geometric construction fails, which it may do, an optimized approach, as for $q=4$ and described below, can be implemented. In this approach the discrepancy, E , between relative frequency and area is minimized with respect to the positions of the three rectangles.

In certain situations an exact construction may not be possible. One case is when (using an obvious notation) the frequency of cell 123 is zero, but the frequency of each of the three cells $\hat{1}23$, $1\hat{2}3$ and $12\hat{3}$ is non-zero. In this case it is not possible to position three rectangles so that they overlap each other but not all three together.

3.3. The case $q=4$

Except under independence (see Section 3.4), I have been unable to find a simple geometric way to construct a correctly scaled diagram for $q=4$. Instead, it can be approached as an optimization problem; to position four rectangles and enclose them by a unit square so that the discrepancy, E , between cell area and cell relative frequency is minimized. Here discrepancy can be measured by

$$E = \sum_{i=1}^{2^4} |a_i - r_i| \quad (1)$$

where r_i is the cell relative frequency and a_i is the area of cell i . Subject to the restriction that each rectangle is exactly proportional to its marginal relative frequency, there are three position parameters of each: the co-ordinates of one corner and the length of one side. Minimization is therefore with respect to 12 parameters. A surrounding rectangle of unit area can be arbitrarily positioned to contain the rectangles; if the rectangles will not fit inside a unit rectangle, a scale reduction can be done to force a fit, with concomitant inflation of E . The final optimized configuration will lie in a unit area rectangle rather than square but an area invariant transformation from an x, y co-ordinate system to an x', y' system with $x' = cx$ and $y' = y/c$, for some c , will create a unit square boundary.

The function E is generally not well behaved and does not have a unique minimum, that is, different configurations may achieve the same value. Also E tends to have local minima. However, minimization can be achieved reasonably well using Powell's [15] conjugate gradient algorithm, which does not require derivatives. To compute E requires evaluating cell areas, a_i , given corner co-ordinates of each rectangle. To do this, I superimpose an irregular grid, as in Figure 8, by extrapolating edges of the rectangles. Cell areas, a_i , are then obtained by adding together grid rectangles that lie in a cell. For example, 17 grid cells make up the white space in Figure 8.

My experience is that convergence to a sufficiently small E depends to some extent on the initial position. One initial configuration is the exact construction that is always possible under an assumption of independence (see below). Other geometric starting configurations can be derived; these are the 'Methods' referred to in the diagram headings. One method uses the configuration in Figure 8, which shows how four rectangles can be positioned to give all 16 possible cells. Some inner cells can be determined by simple geometry so that

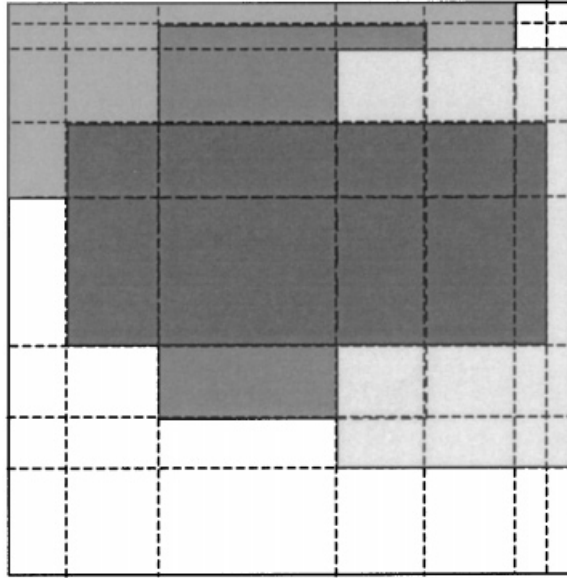


Figure 8. Superimposed grid to calculate cell areas. Also shows arrangement of four rectangles in which all 2^4 cells are displayed.

$a_i = r_i$, and the rectangles themselves completed to ensure their area is equal to each marginal relative frequency. This usually leaves other cells incorrectly scaled, but the configuration may be a good starting point. There are other ways the geometric construction can be done and, further, each can be achieved with a different permutation of the characteristics. I run through the different methods and different permutations to find the best initial configuration to start the minimization algorithm.

Sometimes it does not seem possible to construct a diagram for which E is small. In particular, this may occur when the union of all four characteristics approaches the entire sample and it is not possible to 'fit' a configuration inside the unit square. Another situation is when one or some of the inner cells are empty, as mentioned above for $q=3$.

3.4. Independence

It is shown in the Appendix that, for $q \leq 4$ at least, it is always possible to construct an exactly scaled rectangle diagram of expected counts under an assumption of complete independence. Further, the cells in this situation are all rectangular. Visual comparison of the 'observed' and 'expected' SRD may sometimes give an indication of the departure from independence, though typically clinical data are not independent. A scaled rectangle diagram of expected counts for the data in Figure 5 is shown in Figure 9. The difference between these observed and expected diagrams shows – to no great surprise – that the variables are not independent (as is confirmed by a chi-square test $P < 0.00001$).

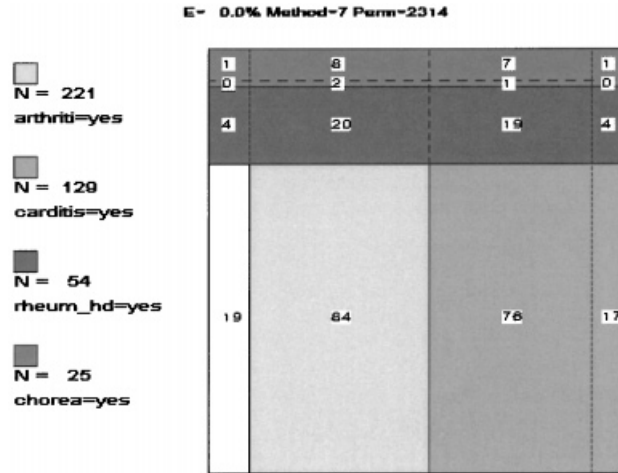


Figure 9. SRD of data in Figure 6 under assumption of complete independence, showing expected counts (rounded to nearest integer).

3.5. The case $q \geq 5$

For $q = 5$, there are up to $2^5 = 32$ non-empty cells to accommodate by positioning five rectangles. In principle the E minimization procedure that is described above could be extended, requiring optimization with respect to 15 parameters. However, it is not possible to position five rectangles and represent all $2^5 - 1 = 31$ possible cells; some error is inevitable, except in special cases where some cells are empty. Diagrams with five, or more, rectangles can look rather cluttered anyway.

4. DISCUSSION

Graphical representations of data are only useful if they show things that might otherwise not have been appreciated [12] and if they have the ability to display one or more ‘phenomena’ [16]. Scaled rectangle diagrams seem to achieve these requirements. They show both the frequency of characteristics and at the same time the extent to which characteristics are shared. In clinical and health research, diagrams that illustrate these features are often useful.

An SRD can be regarded, more generally, as a visualization of a 2^q contingency table. Other ways to display such data are available, the most common being the mosaic plot. Mosaic plots focus on representing cell frequencies by rectangles, the marginal frequencies being less easily appreciated, while SRDs focus on rectangles for marginal frequencies, with cell frequencies being areas within them. Both visualizations seem to be useful, and complementary.

Although it is often difficult in mosaic plots to appreciate the extent to which characteristics overlap (as in Figure 4), they do have some advantages: they are easier to construct; are always exactly scaled; are useful in log-linear model building [7]; and have a physical conceptualization [8]. In principle, they can be extended to any number of variables but, as

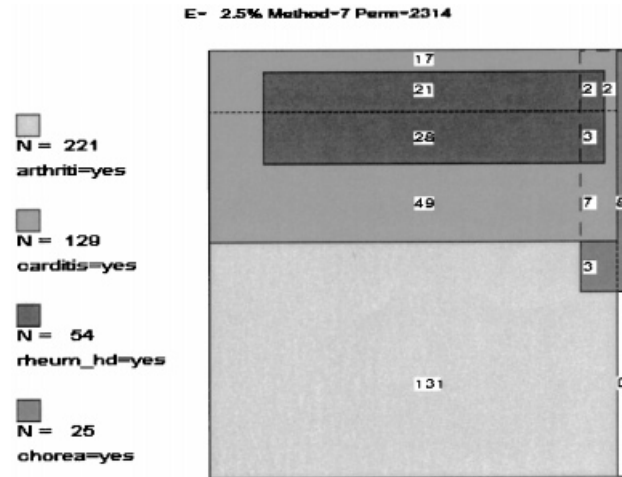


Figure 10. SRD of power transformed frequencies for data in Figure 5. Note that E is here the discrepancy of power transformed frequencies.

for SRDs, they are really only useful for four or fewer variables, despite a suggestion that six variables is a 'practical maximum' for mosaic plots [6].

Unlike mosaic plots, SRDs are restricted to binary categories. However, the binary case is so commonplace, especially in clinical applications, to warrant its own graphical representation (for example, as in rose diagrams [11]).

The main criticism of SRDs is that exact correspondence between area and frequency is not always possible, at least for $q=4$ and sometimes for $q=3$. The question is whether exact correspondence is absolutely necessary. As the mind cannot perceive exact correspondence between area and frequency, and compact shapes are generally perceived to be smaller than diverse shapes of the same area [14], small discrepancies when comparing cells that are a different shape may not be important. The intention is to convey an *impression* of the relative size of groups and the extent to which they overlap. Small discrepancies may not matter unduly in an appreciation of the whole. Usually a configuration with a discrepancy under 5 per cent can be achieved which will provide a good representation.

Nevertheless, it would be helpful to know when it is, and when it is not, possible to construct a diagram with good congruence between relative frequency and area. I have, however, been unable to elicit simple conditions, though some cases are obvious and have been alluded to. Under independence an exact representation is always possible.

Some work on perception of areas [17] has shown that, for squares at least, perceived area is an underestimate of actual area, varying as a power law with exponent about 0.9. To compensate for this effect, one could consider an SRD for power transformed frequencies. Figure 10 shows the diagram corresponding to Figure 5, for cell frequencies raised to the power $1/0.9 = 1.1$. It shows some amplification of areas for larger frequencies and attenuation of smaller ones. Assessing the merits, if any, of this kind of transformation will require further work. Another topic for investigation is the discrepancy measure E (expression (1)) that is used as an objective function in the fitting process. Alternative measures could be employed

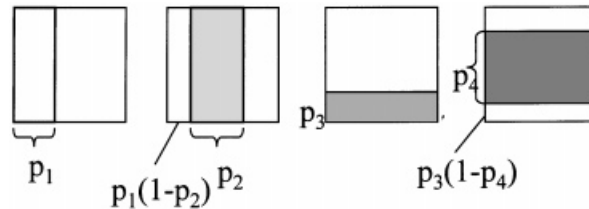


Figure A1. Construction of an SRD for $q=4$ under assumption of complete independence.

that account for how areas are perceived and, possibly, for aesthetic qualities. For instance, as discrepancies in small cells seem to be easier to discern, greater weight could be attached to small cells. Limited trials with different weights have not, however, produced benefits. Aesthetic qualities could be accounted for, as one possibility, by penalizing diagrams with long skinny rectangles.

Scaled rectangle diagrams can be drawn as a facility of search partition analysis (SPAN) software [18]. The program allows rectangles to be coloured, which gives greater clarity than the shades of grey reproduced here. The software is a Windows program that provides, among other functions, an interactive environment to draw both scaled rectangle diagrams and mosaic plots. The program is available from the web site URL <http://www.auckland.ac.nz/mch/span.htm>.

APPENDIX: DIAGRAMS OF EXPECTED COUNTS

Consider the case $q=4$ and suppose $p_j = n_j/n$, for $j=1,2,3,4$, are the marginal relative frequencies of each characteristic. Draw a separate $q=1$ SRD for each characteristic positioned as indicated by Figure A1. It is easily verified that superimposing these four SRDs produces a $q=4$ SRD for the four characteristics in which the areas are equal to cell relative frequencies under an assumption of independence. For example, the intersection of all four rectangles is a rectangular cell of size $p_1 p_2 \times p_3 p_4$.

Obviously removing any one rectangle leaves a $q=3$ diagram under independence and removing any two leaves a $q=2$ diagram under independence.

REFERENCES

1. Feinstein AR. *Clinical Judgment*. Williams and Wilkins: Baltimore, 1967.
2. Wulff HR. *Rational Diagnosis and Treatment*. Blackwell Scientific Publications: Oxford, 1979.
3. Giesecke J, Johnson J, Hawkins A, Noone A, Nicoll A, Wadsworth J, Wellings K, Field F. An estimate of the prevalence of human immunodeficiency virus infection in England and Wales by using a direct method. *Journal of the Royal Statistical Society, Series A* 1994; **157**:89–103.
4. Owen SG, Friesen WT, Roberts MS, Flux W, Francis H. A community-based survey of rheumatoid arthritis in Tasmania. In *Epidemiology in Tasmania*, King H (ed.). Brolga Press: Canberra, 1987.
5. Lusted LB. *Introduction to Medical Decision Making*. Charles C Thomas: Springfield, 1968.
6. Hartigan JA, Kleiner BA. A mosaic of television ratings. *American Statistician* 1984; **38**:32–35.
7. Friendly M. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* 1994; **89**:190–200.
8. Friendly M. Conceptual and visual models for categorical data. *American Statistician* 1995; **49**:153–160.
9. Reidwyl H, Shüpbach M. Parquet diagrams to plot contingency tables. In *Softstat '93: Advances in Statistical Software*, Faulbaum F (ed.). Gustav Fischer: New York, 1994; 293–299.

10. Cohen A. On the graphical display of the significant components of a two-way contingency table. *Communications in Statistics – Theory and Methods* 1980; **A9**:1025–1041.
11. Fienberg SE. Perspective Canada as a social report. *Social Indicators Research* 1975; **2**:151–174.
12. Wainer H. *Visual Revelations: Copernicus*, New York, 1997.
13. Tufte ER. *The Visual Display of Quantitative Information*. Graphics Press: Connecticut, 1983.
14. Foster JJ. The influence of shape on apparent area: a new demonstration. *Acta Psychologica* 1976; **40**: 103–113.
15. Powell MJD. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal* 1964; **7**:155–162.
16. Tukey JW. Data-based graphics: visual display in the decades to come. *Statistical Science* 1990; **5**:327–339.
17. Ekman G, Junge K. Psychophysical relations in visual perception of length, area and volume. *Scandinavian Journal of Psychology* 1961; **2**:1–10.
18. Marshall RJ. A program to implement a search method for identification of clinical subgroups. *Statistics in Medicine* 1995; **14**:2645–2659.