Statistical Computing and Graphics

A Diamond-Shaped Equiponderant Graphical Display of the Effects of Two Categorical Predictors on Continuous Outcomes

Xiuhong LI, Jennifer M. BUECHNER, Patrick M. TARWATER, and Alvaro MUÑOZ

Three-dimensional (3-D) bar graphs and their current 2-D alternatives have certain limitations when the primary objective is to provide an *equal* representation of the effects of two predictors on an outcome. This article proposes a graphing methodology (in the shape of a diamond) that projects 3-D bar graphs into 2-D whereby the third dimension is replaced with a polygon whose area and middle vertical and horizontal lengths represent the outcome. The proposed graphical representation is invariant to rotations and avoids outcomes in categories being concealed by others. This article shows several applications of our proposal to a variety of data types (e.g., proportions, incidence rates, relative risks), which can be easily implemented with available software.

KEY WORDS: Epidemiology; Graphical methods; Statistics; Three-dimensional bar graphs; Two-dimensional projections.

1. INTRODUCTION

Three-dimensional (3-D) bar graphs are commonly used in biomedical studies to portray how two categorical variables (predictors, risk factors) jointly contribute to an outcome (Klag et al. 1996; Huang et al. 1997; Farahmand et al. 2000). However, most 3-D bar graphs fail to achieve the desired feature of *equally* representing the relationships between the outcome variable and each of two predictors, including fixing (conditioning on) one predictor to examine the relationship of the other predictor and the outcome. In addition, 3-D bar graphs are prone to misinterpretation and misperception, and are limited to data that exhibit nonoverlapping trends (i.e., none of the outcomes in joint categories is concealed by others), as addressed by Cleveland and McGill (1984), Wilkinson (1999), and Harris (1999).

To overcome these shortcomings, two-dimensional (2-D) alternatives have been used, such as mosaic (Hartigan and Kleiner 1981, 1984; Friendly 1994; Wilkinson 1999), grouped bar graphs (Tufte 1983), grouped dot plots and framed rectangle charts (Cleveland and McGill 1984), and Trellis display (Becker, Cleveland, and Shyu 1996). However, none of these 2-D displays can *equally* present the relationships between a continuous outcome and each of two categorical predictors in a single plot.

This article proposes a graphing methodology that projects 3-D bar graphs into 2-D whereby the third dimension is replaced with a polygon whose area and middle vertical and horizontal lengths represent the outcome. The proposed graphical representation is invariant to rotations and avoids outcomes in categories being concealed by others. Therefore, our method circumvents limitations of both 3-D bar graphs and current 2-D alternatives, while preserving a desired feature of 3-D bar graphs.

2. METHODS

The key idea in our proposal is to replace the parallelepiped volume in 3-D bar graphs with a polygon in each cell of the 2-D grid defined by categories of the two predictors. To achieve equal representation for two predictors, we choose to use square cells rotated 45° clockwise to construct the grid resembling a diamond (Figure 1). The result is a diamond square cell that not only produces a more aesthetically pleasing graph, but also facilitates the correct display of data because it gives both predictors equal importance. Once the square cells are established, the volumes of the parallelepipeds are compressed into the areas of a polygon within each cell (i.e., three dimensions are projected down to two dimensions). In addition, given that the perception of areas is more difficult than the perception of lengths (Cleveland and McGill 1984), we aim to select a polygon whose middle vertical and horizontal lengths also represent the outcome.

Our proposal is suitable for a variety of data types (e.g., proportions, incidence rates, relative risks). To facilitate the description of our proposal we use the case where the outcome is proportional data. First of all, to achieve equal representation of the two predictors, we restrict attention to polygons centered around the middle of the squares in the grid defined by the categories of two predictors. Figure 1 shows 2-D representations of proportional data ($p = 0.10, 0.25, 0.50$, and 0.75) using five methods (A, B, C, D, and E). The upper left-hand panel corresponds to simply providing the values of the percentages of the areas of the four square cells in each panel occupied by the

Xiuhong Li is Senior Statistical Programmer/Data Analyst, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Room E-7005, Baltimore, MD 21205. Jennifer M. Buechner is Programmer Analyst, University of Wisconsin–Madison, Department of Biostatistics and Medical Informatics, Room 239–WARF Building, 610 Walnut Street, Madison, WI 53726. Patrick M. Tarwater is Assistant Professor, University of Texas, School of Public Health, 1100 N. Stanton, Suite 110, El Paso, TX 79902. Alvaro Muñoz is Professor, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Room E-7008, Baltimore, MD 21205 (E-mail: amunoz@jhsph.edu). This work was supported by grants UO1-AI-35043 and UO1-AI-42590 from the U.S. National Institute of Allergy and Infectious Disease, and PO1-ES-06052 from the National Institute of Environmental Health Sciences. The authors are grateful to David George for editorial assistance.

 DOI: 10.1198/0003130031883

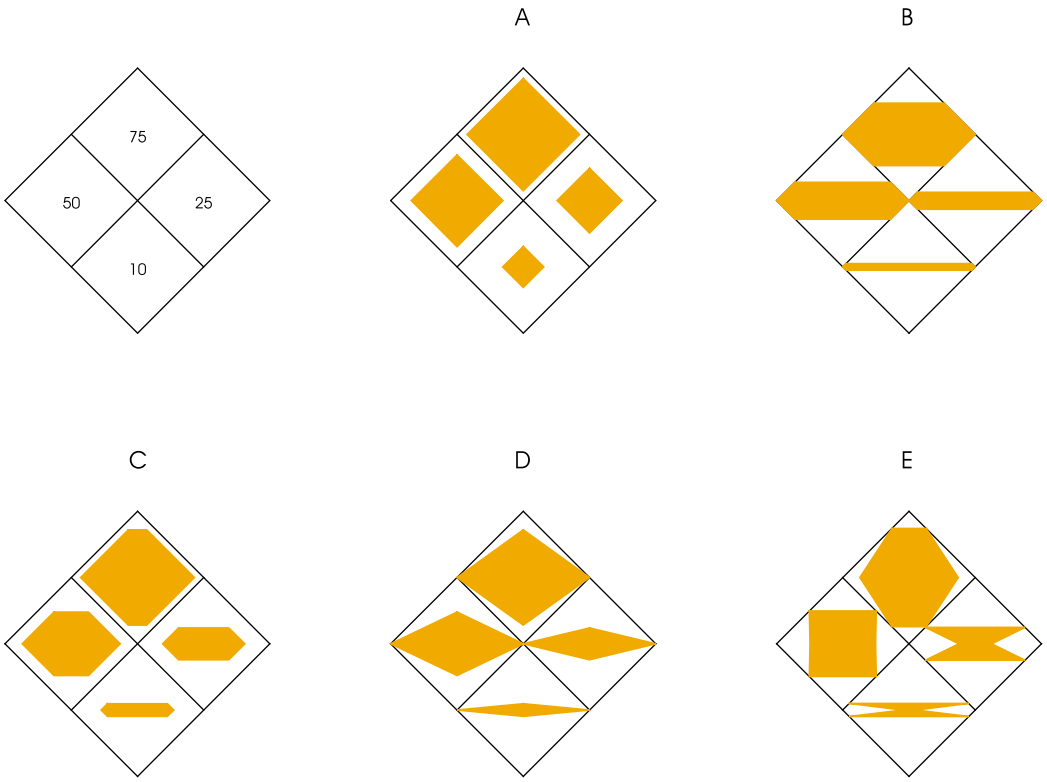

Figure 1. Representations of proportional data (p = 0.10, 0.25, 0.50, and 0.75) as shaded areas that represent the percentages (10, 25, 50, and 75) of the area of each square cell using five alternative methods (A, B, C, D, and E).

shaded polygons. Each of the five methods provides a graphical representation that is invariant to rotation and avoids outcomes being concealed by others. We used S-Plus to develop a function for drawing the square cells of the grid and the shaded areas defining the polygons. In our function, p represents the proportion describing the outcome in a given cell and ℓ is the length of the diagonal for each square cell. The shaded polygons in Figure 1 are constructed based on three attributes: (1) the middle vertical length of the polygons $(V(p))$; (2) the middle horizontal length of the polygons $(H(p))$; and (3) the horizontal length of the top and bottom of the polygons $(h(p))$. The areas of the shaded polygons relative to the area of the square cell are given by $V(p)[H(p) + h(p)]$.

Table 1 illustrates the characteristics of the shaded polygons for all five methods in Figure 1 when ℓ is assumed to be one. Although it is apparent from Table 1 that all five methods correctly display the shaded polygons' area as the proportion p of the square cells' area, only methods C, D, and E correctly display the polygons' middle vertical lengths as p of the square's diagonal height too. Because the eye is familiar with distinguishing differences in displacement, the inclusion of length assists in training one to properly associate numerical values with areas. In method E the middle horizontal length $H(p)$ is also equal to p, but the shapes are not homogeneous with those corresponding to $p < 0.5$ not being convex resulting in a confusing visual effect. In method C, the middle horizontal length of the shaded area is a linear one-to-one function of $p(= 0.5 + 0.5p)$. The same cannot be said for method D, where $H(p)$ is always equal to the horizontal diagonal of the square cell. Another notable property of method C (also shared by method A) is that all shaded areas are disjoint from those in the neighboring cells. This is in marked contrast to methods B, D, and E, where shaded areas of two categories may collide, thus making perception difficult. Therefore, we conclude that method C is the preferable approach because its attributes have maximal number of linear relationships with p, and shaded polygons have constant shape. We also indicate the numeric value of outcome as a label at the center of shaded polygons to further facilitate the graphical display.

Our methods can easily be extended for outcomes ranging from 0 to any positive value (e.g., incidence rates, relative risks,

Table 1. Characteristics of Five Methods for Portraying Three-Dimensional Bar Graphs in Two-Dimensions

Attributes of shaded polygons	*A*	*B*	*C*	*D*	*E*
Middle vertical length = $V(p)$	$\sqrt{p}$	$1 - \sqrt{1-p}$	p	p	p
Middle horizontal length = $H(p)$	$\sqrt{p}$	1	$0.5 + 0.5p$	1	p
Top and bottom length = $h(p)$	0	$\sqrt{1-p}$	$0.5 - 0.5p$	0	$1-p$
Shaded polygon's to square's area = $[H(p) + h(p)]V(p)$	p	p	p	p	p

NOTE: To simplify the formulas, it is assumed that each square cell's diagonal length ℓ is equal to 1.

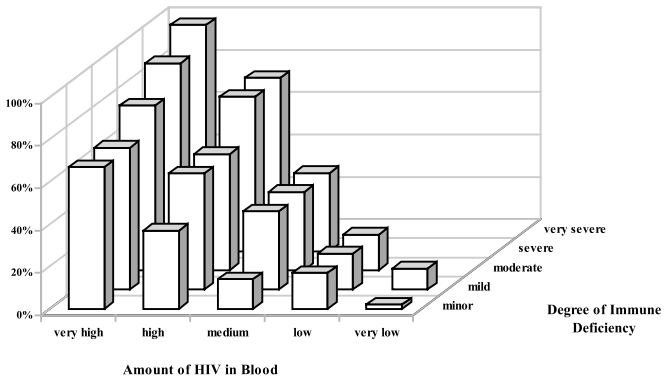

Figure 2. Likelihood of developing AIDS in six years based on the amount of HIV in the blood (very high: >30,000; high: 10,001–30,000; medium: 3,001–10,000; low: 501–3,000; and very low: ≤ 500 copies/ml) as well as the degree of immune deficiency (very severe: ≤ 200; severe: 201–350; moderate: 351–500; mild: 501–750; and minor >750 cells/mm^3).

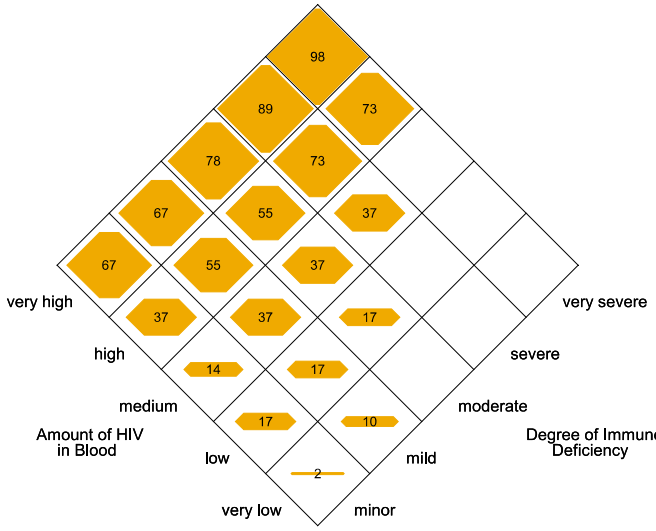

Figure 3. Two-dimensional approach for depicting the likelihood of developing AIDS in six years according to the amount of HIV in the blood and the degree of immune deficiency. Cells with no or very few individuals are blank.

odds ratios). First, dividing each value of the outcome by the highest one, we convert them into proportional values. These derived proportions are used to plot polygons that represent the relative magnitude of outcomes. The cell with the lowest value has the smallest shaded area. This provides a measure of the dispersion (i.e., discrimination) of risk in the population. The larger the difference between the lowest and the highest values, the lower the value of their ratio and the wider the coverage of the shaded areas in the interval from 0 to 1. Instead of derived proportions, the exact numeric values of the outcome are used as the labels at the center of the polygons to conserve the information inherent in the original scale of the outcome.

3. APPLICATIONS

3.1 Graph of Proportions: Likelihood of Developing AIDS in Six Years According to the Amount of HIV in the Blood and Degree of Immune Deficiency

Figure 2 is a 3-D bar graph depicting the likelihood of developing AIDS in six years based on level of HIV in the blood and degree of immune deficiency, as reported by Mellors et al. (1997). The amount of HIV in a patient's blood is measured by plasma HIV-RNA levels (i.e., viral load), an indicator of how many copies of HIV are present in a milliliter of plasma (e.g., very high: >30,000; high: 10,001–30,000; medium: 3,001–10,000; low: 501–3,000; and very low: ≤ 500 copies/ml). The degree of immune deficiency is measured by the CD4$^+$ T-lymphocyte count (e.g., very severe: ≤ 200; severe: 201–350; moderate: 351–500; mild: 501–750; and minor >750 cells/mm^3). These T lymphocytes help the body fight infections and are depleted in HIV patients. Categories that do not have a parallelepiped drawn (e.g., severe immune deficiency with a very low blood viral load) are those for which there are none or few individuals with the characteristics defined by the cell.

Figure 2 has three major limitations. First, the variables are not equally represented (i.e., in each cell, the lengths of the adjacent sides are unequal). The unequal axes and the angle at which the axes intersect may give readers the impression that one variable is more important than the other. A second weakness of Figure 2 is the difficulty of ascribing to the perspective suggested by the graph. Specifically, looking at the bars corresponding to a medium amount of HIV in the blood (3,001–10,000 copies/ml) and to mild (501–750 cells/mm^3), moderate (351–500 cells/mm^3), and severe (201–350 cells/mm^3) degrees of immune deficiency, it may appear that the parallelepiped closest to the back is of greater magnitude than the one immediately in front of it. However, all of these three bars represent the same value. Furthermore, attempting to determine true data values from these graphs can lead to misinterpretations; for instance, one might predict the value of the parallelepiped closest to the back to be $\sim 25\%$ when the true value is 37%. The third drawback of Figure 2 is that it is limited to data with nonoverlapping trends. Specifically, if Figure 2 had large values for outcomes associated with very low HIV in the blood and/or a minor degree of immune deficiency, smaller values adjacent to these outcomes would be blocked and thus unobservable. In addition, if large values occurred for medium amounts of HIV in the blood and/or moderate degrees of immune deficiency, outcomes behind these extremes would be out of the viewing area.

In Figure 3, the newly proposed graphing method has been used to display the same data previously presented in Figure 2. Figure 3 clearly depicts the epidemiological inferences of HIV concentration and immune deficiency on the development of AIDS in six years. The amount of HIV in a patient's blood is informative for each category of immune deficiency. On the other hand, when conditioning on the categories for the amount of HIV in the blood, the incidence of AIDS increases as immune deficiency becomes more severe. Furthermore, Figure 3 reveals that the strength of the association is stronger for HIV blood concentration compared to degree of immune deficiency. This is observable within each category of immune deficiency; for each category of amount of HIV in the blood, a different outcome exists. This is not true when looking across categories of immune deficiency for a fixed category of HIV blood concentration.

Our methods have several advantages. To begin with, the data display uses squares for cells, thus giving equal representation

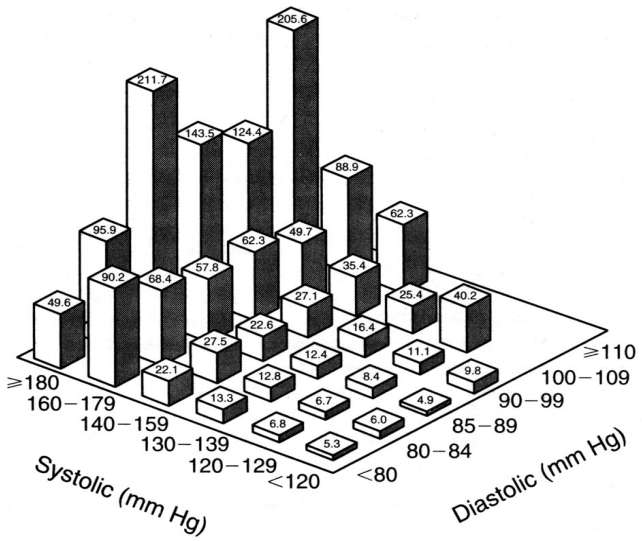

Figure 4. Age-adjusted rates of end-stage renal disease due to any cause per 100,000 person-years according to systolic and diastolic blood pressures. Reprinted with permission from Klag et al. (1996), The New England Journal of Medicine, 334, 13–18.

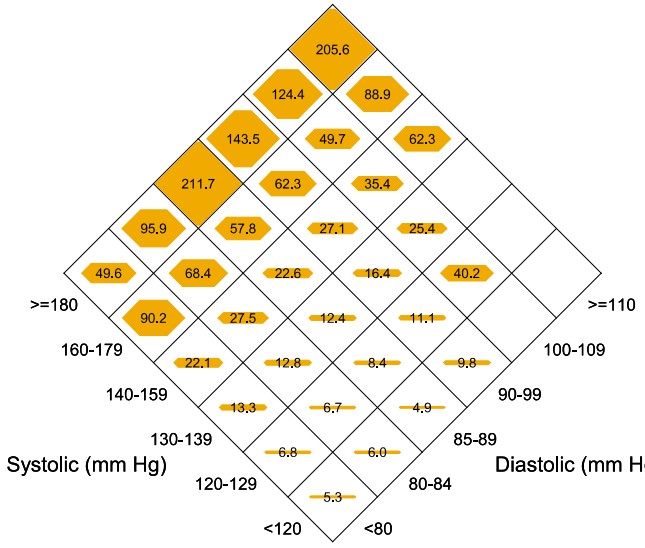

Figure 5. Two-dimensional approach for depicting the age-adjusted rates of end-stage renal disease due to any cause per 100,000 person-years according to systolic and diastolic blood pressures. Each shaded area is a proportion relative to the largest incidence rate with 211.7 being 1 (full shading). The numeric values of the rates are used as the labels. Cells with no or very few individuals are blank.

to both variables (i.e., amount of HIV in the blood and degree of immune deficiency). Readers can now interpret data with cell axes of the same magnitude and an overall base that no longer lies at an angle. Two additional advantages of Figure 3 are also quickly discerned: (1) it avoids the need of conforming to a perspective alluded to by the base of Figure 2; and (2) it facilitates identifying similar values in the third dimension (e.g., three cells with 37% for medium amount of HIV in blood). We further enhanced this feature by placing numbers at the center of each of the shaded areas thereby removing any uncertainty of the values being presented in the graph. One's perception of adjacent data with similar values is not skewed because the two-dimensionality of the graph allows it to be interpretable from any angle.

3.2 Graph of Incidence Rates: Rates of End-Stage Renal Disease Based On Systolic and Diastolic Blood Pressures

Figure 4 is a 3-D bar graph depicting the incidence of end-stage renal disease according to six categories of both systolic and diastolic blood pressures, as reported by Klag et al. (1996). The angled base deceives the readers when comparing values according to one variable for a fixed category of the other variable. It is only after looking at the numbers placed on the parallelepipeds that one can distinguish the magnitude by which these values differ. This is evident when comparing the disease rates for diastolic values of $\geq$110 mm Hg and 85–89 mm Hg for a fixed systolic of $\geq$180 mm Hg. It appears as though the height of 205.6 is greater than that of 211.7.

In Figure 5 our method was used after the data were transformed into proportions by dividing each one by the highest outcome (e.g., divide the values in Figure 4 by 211.7). Thus, the square cell with full shading represents the highest outcome and the cell with the lowest incidence ($= 4.9$ per 100,000 person-years) has the smallest shaded area ($= 4.9/211.7 = 2.3\%$). Therefore, in the case of great discrimination of risk from 4.9 to 211.7 per 100,000 person-years, the range of the shaded areas spans almost the full interval from 0 to 1. An important property of the proposed method is that the relative incidence in the raw data is preserved by the relative size of the shaded areas in the graph. Furthermore, exact values, rather than derived proportions, are placed at the center of the shaded polygon. Figure 5 clearly shows that end-stage renal disease incidence increases with systolic blood pressure for each category of diastolic blood pressure, while it does not increase with diastolic blood pressure for each category of systolic blood pressure.

4. DISCUSSION

Three-dimensional (3-D) bar graphs and their current 2-D alternatives have certain limitations when the primary objective is to provide an *equal* representation of the effects of two predictors. Here, we presented a graphing methodology that projects 3-D bar graphs into 2-D whereby the third dimension is replaced with a polygon whose area and middle vertical and horizontal lengths represent the outcome. All outcomes in joint categories are equally represented within the square cells in a 2-D grid defined by two predictors, thus circumventing certain limitations of 3-D bar graphs.

To achieve the symmetry required by an equiponderant representation of the effects of two predictors, our proposed method depicts the outcome with polygons centered around the middle of each cell in the grid defined by two predictors. An alternative representation of the outcome is to use "thermometers" (i.e., framed rectangle charts; Cleveland and McGill 1984) of a constant width at the center of each cell. However, the "thermometers" do not optimally use the space in the graph, and they are more appropriate for cartographic applications. We incorporate the primary feature exhibited by "thermometers" in our proposed method by representing the outcome with the middle vertical length of the shaded polygon. Another alternative repre-

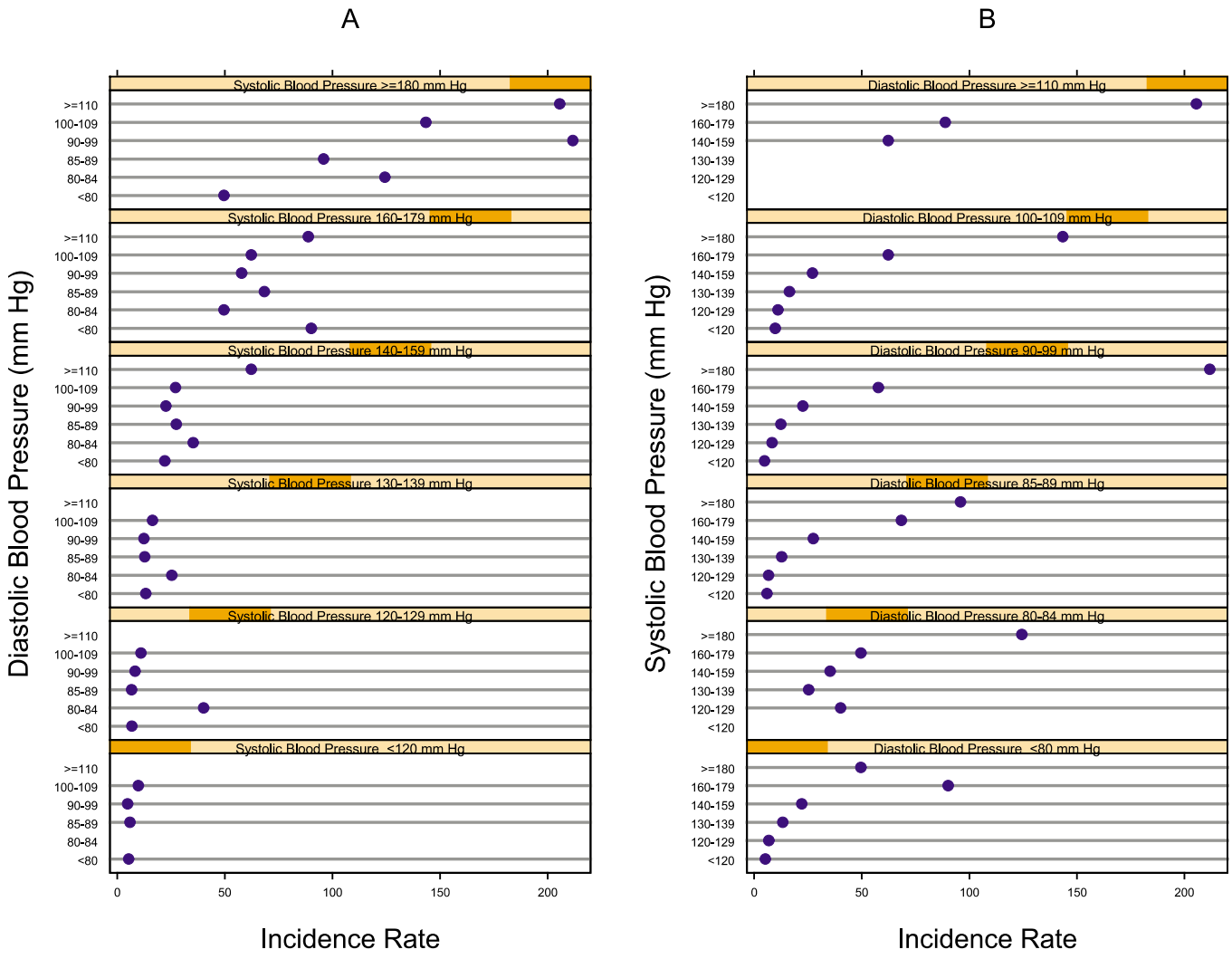

Figure 6. Trellis rendition of the age-adjusted rates of end-stage renal disease due to any cause per 100,000 person-years according to systolic and diastolic blood pressures. The A Trellis displays systolic blood pressure as the conditioning variable and diastolic blood pressure as the panel variable. The B Trellis switches the roles of the blood pressures.

sentation of the outcome is to simply shade from the bottom of each cell on up, but the shaded areas of this option will provide a distorted representation of the outcome and shaded areas in contiguous cells may collide making perception difficult.

Trellis (Becker, Cleveland, and Shyu 1996) uses a three-way rectangular array of panels to display 3-D data. The primary feature of trellis plots is the arrangement of the different panels in a form reminiscent of a garden trelliswork. Different panels of a trellis correspond to different categories of one of the predictor variables (conditioning variable), and in each panel the outcomes for categories of the other predictor variable (panel variable) are displayed. Therefore, Trellis graphs do not provide equal representation of the effects of the two predictor variables unless the roles of the conditioning and panel variables are switched, which results in the requirement of multiple Trellis displays. Indeed, Figure 6 is a Trellis rendition requiring two displays for the data depicted in Figures 4 and 5 using 3-D bar graphs and our equiponderant method, respectively. It is apparent that Figure 5 is a succinct and efficient approach to depict the data. On the other hand, if the two predictors are not equally important (e.g., exposure and confounder), Trellis graphs are preferable because the exposure is depicted as the panel variable and the confounder is depicted as the conditioning variable.

Another alternative to our proposed method is to use mosaic plots, where an outcome is represented by color-intensity scales. However, this approach has limitations for a continuous outcome. For a mosaic rendition of the data in Figures 4 to 6, one would need to categorize the incidence rate so that the number of colors becomes manageable. As a consequence, it would preclude depicting some of the differences shown in Figures 4 to 6. In addition, mosaic graphs require a legend (which usually occupies a large proportion of graphing area) with the link of the intensity of the color and the value of the outcome.

The methods presented here are not applicable when a graphical representation of the effects of two continuous predictors is desired. In this case, the use of contour plots (with lines showing equal values of outcome) is a viable and preferred option.

Our method is directly suitable for depicting proportions of events in each cell of a grid, defined by two predictors (Figures 2 and 3), and can be extended to cases where the outcome is not a proportion (i.e., an unbounded number, Figure 5). In such a case, each value of the outcome is divided by the maximum

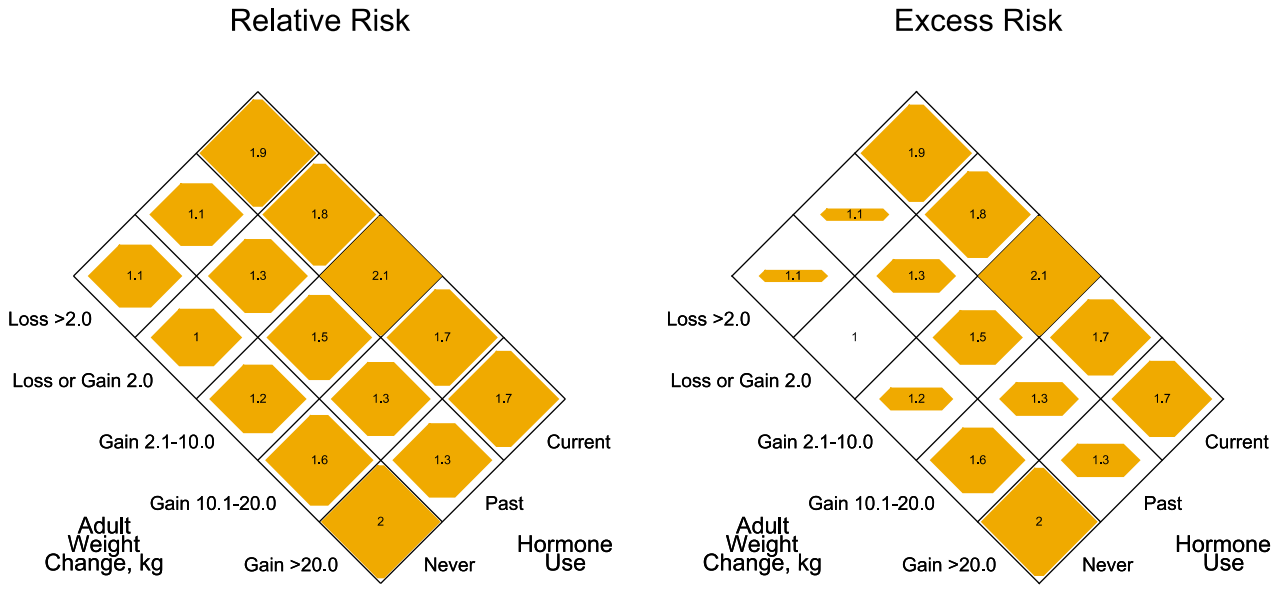

Figure 7. Relative risk of breast cancer by adult weight change and hormone use among postmenopausal women reported by Huang et al. (1997). The graph on the left panel depicts the relative risks (i.e., p_i = rr_i/highest rr) and the right panel depicts the excess risk (i.e., p_i = (rr_i − lowest rr)/(highest rr − lowest rr).

one in the dataset, thereby creating a proportion. This extension allows the ratio of two shaded areas to preserve the ratio of the relative frequencies; and the magnitude of the smallest shaded area conveys a measure of the dispersion (i.e., discrimination) of risk. Furthermore, if there were to be a predetermined maximum value (from an external source) associated with the outcome, it can be used as the normalizing value. In this case, the cell with the observed highest outcome will have a polygon corresponding to $p =$ highest observed/predetermined maximum.

Our method can also be adapted to represent relative and excess risks. Figure 7 depicts the relative risks of breast cancer by adult weight change and hormone use among postmenopausal women, as reported by Huang et al (1997). The left panel depicts the relative risks (i.e., $p_i = rr_i/\text{highest}\, rr$) and the right panel depicts the excess risk. Specifically, if rr_i denotes the relative risk for the ith cell, excess risk is represented by polygons based on the proportions $p_i = (rr_i - \text{lowest } rr)/(\text{highest } rr - \text{lowest } rr)$. One advantage of this approach is that the shaded areas span from 0 to 1 with the cell having the lowest relative risk having no polygon drawn since $p_i = 0$ (i.e., the method naturally highlights the reference category). However, the ratio of the areas of the two polygons does not preserve the ratio of the two corresponding relative risks. It may sometimes give an exaggerated impression of the cell-to-cell differences.

Our method can easily be implemented using commercially available software (e.g., S-Plus). To overcome some of the graphical limitations of S-Plus (e.g., the four sides of the square cell occupy part of the area of the square cell so that the visible area of the square cell is actually less than 1), we have found scaling all proportions by 1.005–0.005 (number of categories of variable 1 + number of categories of variable 2) to be useful. Using this adjustment, the graphical illustration is enhanced when p is high.

We realize that our method has its own limitations. For instance, according to Cleveland and McGill's (1984) graphical perception theory, it is not easy to perceive data via areas. Notwithstanding this difficulty, our proposed method based on areas of polygons does accomplish an equiponderant representation of the effects of two predictors on the outcome. To our knowledge, no other available method succinctly conveys the symmetry requirement by an equiponderant representation. In addition, to capitalize on the easiness of perceiving linear representations and in order to extract quantitative information more accurately and efficiently, we selected a type of polygon whose middle vertical length also represents the outcome and whose middle horizontal length is a linear one-to-one function of the outcome. Furthermore, we added numerical value at the center of each polygon to convey the actual outcome in order to increase the accuracy of perceptions, because using words, numbers, and drawing together can enhance the quality of visual presentation of data, as suggested by Tufte (1983). In addition, the numerical value at the center of each polygon serves to distinguish a cell with an outcome of zero (represented by 0 at the center and no polygon) from a cell with no or few individuals (represented by a blank cell).

Our method can be used to depict higher dimensional data. For instance, one can incorporate the fourth dimension by plotting a sequence of graphs. A case in point would be the depiction of the proportions for developing AIDS, not just at six years (Figure 3), but also at three and nine years, which can be easily accomplished by a sequence of three graphs in a single page. Such a depiction would not only allow the equiponderant representations of the effects of two predictors on the outcome but also reveal how the outcomes evolve over time. Innovative applications and further extensions of the methods proposed here will enhance the function of graphical methods in summarizing and portraying the scientific information data hold.

5. EPILOG

At the final stage of the review process, the editor invited us to consider naming our graphical method. We have welcomed the suggestion and have opted to name our new method of display

the *diamond* graph. It has the shape (and we hope the value) of a diamond. Perhaps more importantly, it is reminiscent of the baseball diamond that *The American Statistician*s equiponderantly love.

[Received December 2000. Revised March 2003.]

REFERENCES

Becker, R. A., Cleveland, W. S., and Shyu, M. J. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Statistical Graphics*, 5, 123–155.

Cleveland, W. S., and McGill R. (1984), "Graphic Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of American Statistical Association*, 79, 531–554.

Farahmand, B. Y., Michaëlsson, K., Baron, J. A., Persson, P. G., and Ljunghall, S. (2000), "Body Size and Hip Fracture Risk," *Epidemiology*, 11, 214–219.

Friendly, M. (1994), "Mosaic Displays for Multi-way Contingency Tables," *Journal of American Statistical Association*, 89, 190–200.

Harris, R. L. (1999), *Information Graphics: A Comprehensive Illustrated Reference*, New York: Oxford University Press.

Hartigan, J. A., and Kleiner, B. (1981), "Mosaics for Contingency Tables," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, New York: Springer.

——— (1984), "A Mosaic of Television Ratings," *The American Statistician*, 38, 32–35.

Huang, Z., Hankinson, S. E., Colditz, G. A., Stampfer, M. J., Hunter, D. J., Manson, J. E., Hennekens, C. H., Rosner, B., Speizer, F. E., and Willett, W. C. (1997), "Dual Effects of Weight and Weight Gain on Breast Cancer Risk," *Journal of the American Medical Association*, 278, 1407–1411.

Klag, M. J., Whelton, P. K., Randall, B. L., Neaton, J. D., Brancati, F. L., Ford, C. E., Shulman, N. B., and Stamler, J. (1996), "Blood Pressure and End-Stage Renal Disease in Men," *New England Journal of Medicine*, 334, 13–18.

Mellors, J. W., Muñoz, A., Giorgi , J. V., Margolick, J. B., Tassoni, C. J., Gupta, P., Kingsley, L. A., Todd, J. A., Saah, A. J., Detels, R., Phair, J. P., and Rinaldo, C. R. (1997), "Plasma Viral Load and CD4+ Lymphocytes as Prognostic Markers of HIV-1 Infection," *Annals of Internal Medicine*, 126, 946–954.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Wilkinson, L. (1999), *The Grammar of Graphics*, New York: Springer-Verlag.