Determining and visualising at-risk groups in case-control data

RJ MARSHALL
Department of Community Health, University of Auckland, New Zealand

Background Case-control research is often exploratory; to determine factors that increase risk. Often, regression methods are used to determine combinations of risk factors that predispose to excess risk. Recently, tree-based methods have also been proposed. Both have limitations. An alternative approach is suggested, based on a search algorithm to identify at-risk subgroups.

Methods Statistical methods to determine and visualise at-risk sub-groups in case-control studies are presented. The method of determining sub-groups — search partition analysis (SPAN) — searches among different Boolean combinations of risk factors. Sub-groups that have been identified are visualised by scaled rectangle diagrams. These show the size of sub-groups and the extent to which they overlap.

Results Theory is presented for applying the method to case-control data. The methods are illustrated by analysis of three case-control studies: one on sudden infant death syndrome, a second on heart disease and a third on child pedestrian injuries.

Conclusions The methods provide a useful alternative to regression and tree-based analysis. They demarcate sub-groups that, in the three examples, are easy to interpret and would not have been found by other methods. Scaled rectangle diagrams are a useful way to visualise the results.

Keywords epidemiology, case-control study, risk factors, odds-ratios

Introduction
The purpose of much epidemiological research is to identify groups of individuals who are at risk. Model building, for example logistic and Poisson regression, have been widely used for this purpose, but more exploratory methods have been suggested recently, in particular, tree-based analysis. Tree analysis appears to use no restrictive model, as is imposed by regression, and, by a process of data subdivision, homogeneous sub-groups are identified and are defined by logical combinations of risk factors, rather than by arithmetic-weighted averages of risk factors. Tree-based methods are now available in different software packages. However, trees also have limitations that arise primarily from the hierarchical nature of their construction.

Nevertheless, when the objective is to find who is at especially high (or low) risk, it seems reasonable to explore the data to see which logical, rather than arithmetic, combinations do present different risk levels. One method that has been proposed, as an alternative to tree-analysis, is search partition analysis (SPAN). It is non-hierarchical and enumerates all combinations of risk factors, subject to certain constraints.

The purpose of this paper is to illustrate the SPAN method for analysis of epidemiological data, in particular case-control data, and to introduce concepts concerning sub-groups and how these may be visualised.

Methods
SPAN
In case-control and cohort investigations, information on a number of different risk factors may be collected. These may be measured on binary, continuous, nominal, or ordinal scales. In SPAN, as in tree-based analysis, information on risk factors is summarised by attributes associated to variables. If serum cholesterol concentration is measured, for example, an attribute that represents 'raised serum cholesterol above 6 mmol/l' could be defined.

A Boolean combination of attributes induces a binary partition of the risk-factor space into A and 'not A' (the complement A); the partition also induces a corresponding split of the actual data. In SPAN, different Boolean attribute combinations are generated and combinations that split the data into two groups that are most homogeneous with respect to an outcome measure y are deemed 'best'. In epidemiological work y is typically a binary indicator of the presence of disease.

Since the number of ways of combining attributes can be vast, SPAN searches for a Boolean combination
subject to certain restrictions. First, Boolean expressions are restricted in terms of their complexity by imposing sensible restrictions to enable a feasible size search. Specifically, Boolean expressions of the general disjunctive form

\[ A = I_1 \lor I_2 \lor \ldots \lor I_q \]  

are generated where each \( I_j \) is the conjunction of \( p_j \) attributes. For example, expressions (3), (5) and (7) (below) are of this general form. Both \( p_j \) and \( q \) are typically small, < 3 or 4. Another way the Boolean expressions are restricted is by only allowing the attributes that are used in (1) to be drawn from a restricted pool of \( m \) attributes, where typically \( m \) is not large, say \( m < 15 \). This pool of attributes, denoted \( T_m \), may be chosen from factors that are known to present a risk, or from a preliminary analysis of the data. In practice, I have argued\(^5\)-\(^8\) that it is usually sensible for this pool to contain only ‘positive’ attributes, that is, attributes which are positively associated with \( y \). For example, ‘raised serum cholesterol’ is a positive attribute for heart disease. Similarly, low HDL cholesterol would be considered a positive attribute for that disease. If \( C \) and \( H \) denote raised serum cholesterol and low HDL cholesterol respectively, both \( C \) and \( H \) would be put into \( T_m \).

One of the implications of including just positive attributes in \( T_m \) is that Boolean expressions of the form of (1) are ‘regular’, in that they are in terms of only positive attributes. This means that generating combinations such as \( C \& H \), that is, raised serum cholesterol and raised HDL cholesterol, are precluded. Such combinations, in which negative and positive attributes combine, seem unlikely to be biologically important. In tree analysis, however, such combinations are, by virtue of the splitting process, generally an inherent feature that often makes interpretation difficult\(^5\) (also see the Discussion below).

If analysts are unwilling to commit themselves to specifying an attribute as positive, or if it is felt that an attribute may change risk in different ways, depending on how it is combined with other attributes, it is possible to include both an attribute and its complement in the attribute set \( T_m \). For example, allowing both ‘male’ and ‘female’ as attributes, or both ‘normal’ and ‘abnormal’ to be in \( T_m \). This, of course, makes the search more extensive.

**Sub-groups**

The combinations \( I_1, I_2 \), etc., in (1) can be considered as defining intrinsic sub-groups\(^7\) in \( A \), which are not necessarily mutually exclusive. Their size, and the extent to which they overlap, can be visualised using scaled rectangle diagrams. These are a means to display categorical cross-classifications\(^9\), and are readily adapted to show sub-groups. For case-control data it is instructive to present these diagrams separately for cases and for controls, so that the distinctions between the prevalence of sub-groups in cases and controls becomes immediately apparent. By expressing \( A \) in disjunctive form, as in expressions (4), (6) and (8) below, sub-groups can also be identified in \( A \), which can be thought of as combinations of negative attributes that define low risk sub-groups.

**Criterion for best partition**

The criterion for choosing the ‘best’ partition can be made in different ways. The reduction in diversity of \( y \), by splitting into \( A \) and \( \tilde{A} \) is a simple and appropriate criterion; it is used, for example, in classification and regression trees (CART)\(^10\). Specifically, if \( d(P) = P(1 - P) \) is diversity for a proportion \( P \), the reduction in diversity is:

\[ G = d(P_A) - P_Ad(P_{A1}) - P_{\tilde{A}}d(P_{\tilde{A}1}) \]

where \( P_A \) and \( P_{\tilde{A}} \) are the proportions in \( A \) and \( \tilde{A} \), and \( P_{A1} \) and \( P_{\tilde{A}1} \) are, respectively, proportions of \( y = 1 \) in the sample as a whole, in \( A \) and in \( \tilde{A} \).

In case-control sampling, it is not obvious that \( G \) is appropriate, since \( y \) is pre-determined by the case-control mix; it is more natural to consider the diversity of a binary indicator variable, for example \( z_A \), which is one for individuals who possess \( A \). The objective is to find \( A \) such that the case and control series are as homogeneous as possible with respect to \( z_A \). The appropriate criterion is therefore the reduction in the diversity of \( z_A \) in the case and control series, that is,

\[ G' = d(P_A) - P_Ad(P_{A1}) - P_{\tilde{A}}d(P_{\tilde{A}1}) \]

where \( P_{A1} \) and \( P_{\tilde{A}1} \) are the proportion of individuals with \( A \) in the case and control series respectively, and \( P_A \) and \( P_{\tilde{A}} \) are the proportions of cases and controls. It can be shown, that (2) reduces to simply:

\[ G' = d(P_A)(P_{A1} - P_{\tilde{A}1})^2 \]

so that maximising \( G' \) is tantamount to maximising the difference between the proportion of cases and controls possessing \( A \). Note that it can also be shown that

\[ G' = \frac{d(P_A)}{d(P_{\tilde{A}})} \]

so that, as \( d(P_{\tilde{A}}) \) is fixed, \( G' \) is equivalent to ‘balancing’ \( G \), that is, multiplying \( G \) by \( d(P_A) = P_A P_{\tilde{A}} \) to try to ensure that numbers in \( A \) and \( \tilde{A} \) are not disparate\(^6\), a device also used by CART to avoid ‘end cut preference’\(^10\) in splits.

**Iterating a search**

An iterative procedure\(^7\) can also be implemented to expand the extent of a search: an optimal partition is
first found which itself defines a new attribute. This is then added to the set $T_m$ for a subsequent search.

**Complexity penalising**

In practice, sub-group definitions need to be kept simple. A simple partition is preferable to a more complex one with only marginally larger $G'$. Complexity, $c$, can be measured as the total number of high and low risk sub-groups, minus one. The upper envelope of a plot of $G'$ versus $c$ for generated partitions of a search is known as the `complexity hull' and can be used to indicate the extent to which gains in $G'$ are obtained by increasing complexity. A trade-off between the two can generally be judged by inspecting the plot. Formally, complexity penalising can be used, whereby $G'$ is penalised using $G' - cf$ for some value $f$.

**Software**

The SPAN method and creation of scaled rectangle diagrams were implemented using SPAN Windows software written by the author. For details and downloads see: http://www.auckland.ac.nz/mch/span/span.htm.

**Results**

**Example 1: Sudden infant death syndrome (SIDS) study**

A nationwide case-control study of SIDS has been carried out in New Zealand. Three controls were selected for each of about 500 SIDS cases, and information about the household, mother and child were collected by questionnaire and from hospital records.

SPAN analysis was implemented to identify combinations of risk factors that discriminate mothers at high risk from those with low risk. By considering each risk factor individually, an attribute set of the best $m = 12$, according to $G'$, were selected. These were, in decreasing $G'$ value: smoking, prone sleeping, unmarried, Maori ethnicity, low birth-weight $< 3000$ g, age at first pregnancy $< 25$, low socioeconomic status, receiving welfare benefit, age $< 25$, gestational age $< 38$ weeks, no breast feeding, and bed-sharing. These form $T_{12}$. For attributes formed from numeric variables (e.g. birth-weight), an optimal cut-point was chosen, from a set specified in advance, by maximising $G'$. For example, the optimal cut-point of 3000 g for birth-weight was the best of the set 1600, 1800, ..., 3400, 3600. The iterative procedure for generating partitions and complexity penalising was adopted. The initial search was done by generating all partitions of the form of (1) with $q \leq 2$ and $p_i \leq 2$. The best partition on this search was used to create a new attribute, which augments $T_m$ for a subsequent search. This process was then iterated until convergence, that is, the best partition on the final cycle is identical to that on the previous one. The generated complexity hull is shown in Figure 1. The point on the hull at complexity $c = 4$ seems a reasonable compromise that offsets increasing complexity against increasing $G'$. Formally, it is the optimal complexity penalised partition for parameter $\beta = 0.00036$, that is, it maximises the vertical distance $G' - c\beta$ to the line with slope $\beta = 0.00036$.

The partition is:

$$A = (S \land L) \lor (S \land B) \lor P$$

where $S$ indicates that the mother was a smoker, $B$ is `bed-sharing', $P$ is `sleeps prone', and $L$ is low birth-weight ($< 3000$ g). In words, rule (3) is any baby whose mother is either a smoker with a low birth-weight baby ($S \land L$), or a smoker who bed-shares ($S \land B$), or places baby in a prone position ($P$).

Any baby whose mother does not fall in the high risk category, is in the low risk category, $A$, defined (simply by switching $\land$ and $\lor$, and taking complements of each attribute). The expression can be re-expressed in disjunctive form as:

$$\bar{A} = (\bar{S} \land \bar{P}) \lor (\bar{L} \land \bar{B} \land \bar{P})$$

As this example was a case-control study, odds ratios (OR) can be used to estimate relative risks (RR). The OR of $A$ relative to $\bar{A}$ is 7.7 [95% confidence interval (CI): 5.7–10.3]. This compares with individual ORs: 4.25 (95% CI: 3.35–5.36) for smoking alone, 3.7 (95% CI: 2.93–4.67) for prone sleeping, 3.05 (95% CI: 2.47–3.87) for low birth-weight and 2.70 (95% CI: 2.04–3.59) for bed-sharing.

![Fig. 1 Upper envelope (complexity hull) of $G'$, c plot during the search. Scattered plotting positions are of partitions generated on the third iteration. The point on the hull at complexity $c = 4$ is optimal with respect to the line of slope $\beta = 0.00036$.](image-url)
From (3) there are three intrinsic high-risk subgroups within A: \( S \& L \), \( S \& B \), and \( P \). Scaled rectangle diagrams representing these are shown in Figure 2 for cases and for controls. Each rectangle is scaled in area according to the number of individuals in each subgroup, as well as the extent of overlap. Comparing the case and control diagrams, it is quite clear how much more common the sub-group combinations are in cases than in controls. They also show the extent to which they intersect each other, and that most of the cases were in one or other of the sub-groups.

The estimated RR of each of the sub-groups relative to the low risk group as a whole are 15.3 (95% CI: 10.8–22.0) for \( S \& L \), 18.3 (95% CI: 12.0–27.2) for \( S \& B \) and 7.3 (95% CI: 5.4–9.8) for \( P \). With \( A \) written as in (4), there are two intrinsic low risk sub-groups: \( S \& P \) and \( L \& B \). The OR of each of these relative to \( A \) as a whole are 0.11 (95% CI: 0.08–0.15) and 0.11(0.9–0.17).

Example 2: Auckland heart study
Data from an age-frequency-matched case-control study of heart disease carried out in Auckland, New Zealand\(^{13}\) are used in this example. There were 525 cases and 227 controls. For the SPAN analysis in this example the same procedure has been used as in the SIDS example. The top 10 attributes that were used to form the attribute set \( T_{10} \) were, in descending order of \( G' \): smoking, low HDL cholesterol (< 1.05 mmol L\(^{-1}\)), high total cholesterol (> 7.5 mmol L\(^{-1}\)), alcohol abstainer, heavy drinker (daily alcohol > 20 g day\(^{-1}\)), history of hypertension, high body mass index (BMI > 25), non-regular exercise, low social class, and either abstainer or heavy drinker. The optimal partition was:

\[
A = S \text{ or } (N \& O) \text{ or } (H \& X)
\] (5)

where \( S \) is smoking, \( N \) is no alcohol, \( O \) is overweight (BMI > 25), \( H \) is hypertensive, and \( X \) is ‘does not exercise’. Its complement in disjunctive form is:

\[
\hat{A} = (\hat{S} \& \hat{N} \& \hat{H}) \text{ or } (\hat{S} \& \hat{N} \& \hat{X}) \text{ or } \\
(\hat{S} \& \hat{O} \& \hat{H}) \text{ or } (\hat{S} \& \hat{O} \& \hat{X})
\] (6)

In \( A \) there are three sub-groups: the first is the subgroup of smokers \( (S) \); the second is the non-drinking overweight \( (N \& O) \); and the third group is the non-exercising hypertensives \( (H \& X) \). The OR of these relative to \( \hat{A} \) as a whole are 4.40 (3.06–6.31), 3.64 (2.16–6.12) and 4.41 (2.67–7.27) respectively. From equation (6) there are four low risk sub-groups. Figure 3 shows the scale rectangle diagrams for the high risk groups. Clearly, \( A \) is dominated by smokers; the non-drinking overweight being a smaller group, but roughly twice as common in cases.

Example 3: Child pedestrian injury study
A third example is from a case-control study of child pedestrian accidents\(^{14}\). ‘Cases’ were road sites where a child pedestrian had been injured by being hit by a motor vehicle. Two control sites were selected at random from streets in proximity to the case site. There were 175 case sites and 350 control sites. Features of the road were measured at the sites.

Individually, the attributes that comprised \( T_{11} \) were, in decreasing order of \( G' \): flow-rate in excess of 100 vehicles per hour, presence of a bus stop, non-residential road, main arterial road, presence of a median divider, presence of pedestrian crossing, more than one traffic lane, on-street parking, presence of footpaths on both sides of the road, curvature in the road, and no verges.

A SPAN analysis, carried out as for the previous examples, led to the high risk combination:

\[
A = (F \& B) \text{ or } (F \& P) \text{ or } L
\] (7)

Fig. 2 Scaled rectangle diagrams for sub-groups identified in a case-control study of SIDS. Each rectangle is scaled in proportional to sample size and according to extent of overlap with each other. \( P \) is sleeps prone, \( S \& L \) is smokes and low birth-weight, \( S \& B \) is smokes and bed-shares.

Fig. 3 Scaled rectangle diagrams for sub-groups identified in a case-control study of coronary heart disease. \( S \) is smokes, \( N \& O \) is non-drinker and overweight, \( H \& X \) is hypertensive non-exercisers.
where $F$ is flow rate in excess of 100 vehicles per hour, $B$ is presence of a bus stop, $P$ is presence of a footpath on both sides of the road, and $L$ is more than one traffic lane.

The complement, representing low risk, is

$$
\bar{A} = (\bar{F} \& \bar{L}) \text{ or } (\bar{B} \& \bar{L} \& \bar{P})
$$

(8)

The high risk group is composed of three sub-groups: sites near a bus stop on a high flow road ($F \& B$); and sites with footpaths on high-flow-rate road ($F \& P$); sites on a two, or more lane road ($L$). The low risk group consists of two sub-groups: sites on a single-lane low-flow-rate road, ($\bar{F} \& \bar{L}$), and sites on a single-lane road with no bus-stop and no footpath ($\bar{L} \& \bar{B} \& \bar{P}$). The OR of $A$ versus $\bar{A}$ is 12.6 (7.8–20.5) and those of the three sub-groups relative to $\bar{A}$ are 17.9 (8.5–37.7) for $F \& B$, 12.3 (7.6–20.2) for ($F \& P$) and 47.5 (20.7–109.0) for $L$.

Figure 4 is a visualisation of the sub-groups. It immediately shows how the very large OR for $L$ comes about, as very few control sites had more than one lane. Also, it shows that the $F \& B$ sub-group is almost wholly contained in $F \& P$.

**Discussion**

Epidemiological analysis is generally a means to identify groups who have been or are at risk. There may be no single factor that places an individual at excess risk and, given the widely accepted notion that many illnesses are caused by the conjunction of factors, it is natural to try to identify combinations of potential factors that may raise risk. This can be done within the framework of a regression model, by simply categorising the linear-risk score of a fitted model. But this may lead to a diverse group of individuals, their only common element being a similar arithmetic weighted average of risk-factor scores.

Another approach is to use tree-based methods\textsuperscript{1–4}. Although these are appealing they have disadvantages. One criticism is that the tree structure, which can be thought of as analogous to the regression model, is not a particularly good one\textsuperscript{5}. It allows combinations of risk factors that may be quite meaningless and hard to interpret.

For example, an optimal regression tree for the SIDS data in Example 1 is shown in Figure 5. It was constructed using CART 4.0 software. The leftmost node gives the lowest risk group ($\bar{S} \& \bar{P}$) that is used as reference for the calculation of the OR at each of the sub-groups indicated by the other terminal nodes. The sub-groups have varying degrees of risk relative to the baseline group.

Elements of the SPAN derived sub-groups are manifest. For example, the lowest risk node $\bar{S} \& \bar{P}$ is a SPAN sub-group (expression 4). However, with the exception of this node and the rightmost node 9, $S \& L \& P$ (smoking, low birth-weight and prone sleeping), none is particularly easy to interpret. For example, node 8 is the sub-group $S \& L \& P \& B$, that is, smokers, low birth weight, not prone, and bed shares. One might ask what 'not prone' is doing there? Is it necessary to be not prone in conjunction with smoking, bed-sharing and low birth-weight to be at risk? Should it be inferred that being placed prone in conjunction with the other three attributes does not confer a risk? Similarly, being unmarried appears to be a risk only in the smoking, not low birth-weight and prone sub-group, but with OR 2.8 ($= 19.3/6.9$), it is marginally lower than for individuals who are not in the sub-group (3.4).

Note also that splits on $P$ occur at nodes a, b and c (Figure 5). That is, prone sleeping is an effect in each of

---

**Fig. 4** Scaled rectangle diagrams for sub-groups identified in a case-control study of child pedestrian injuries. $L$ is a two, or more, lane highway, $F$ & $P$ is a high traffic-flow rate and footpaths, $F$ & $B$ is a high traffic-flow rate and presence of a bus stop.

**Fig. 5** Optimal regression tree for the SIDS data using CART software. $P$ is sleeps prone, $S$ is smokes, $L$ is low birth-weight, $B$ is bed-shares and $U$ is unmarried. Node boxes give: node number, OR (in bold relative to node 1) and sample size at the node.
the sub-groups $\tilde{S}$, $S \& \tilde{L}$ and $S \& L$, which together comprise the whole, so that prone sleeping could be interpreted as an across-the-board effect that works on its own, as it appears in the SPAN derived sub-groups (expression 3).

One intention of SPAN is to counter the drawbacks of tree analysis, but at the same time deriving risk groups that are logical, rather than arithmetic, combinations of risk factors. SPAN only achieves a broad categorisation into low and high risk. While this may be considered a crude categorisation, many epidemiological measures, for example, RR, are of this nature. Further distinctions in levels of risk become apparent when considering sub-groups within the high- and low-risk dichotomy, as in the examples presented above.

Methods for visualising risk-groups in epidemiology are not well-developed. Typically, researchers rely on tabulations. The use of scaled rectangle diagrams to visualise sub-groups is a new approach that seems to be useful. They can be used to investigate how risk factors co-occur, whether or not they demarcate especially high-risk combinations. The sub-groups identified by the SPAN procedure are not, of course, the only important combinations of risk factors. Other partitions of the data, either equally, more, or less complex, will be near the optimal, and give rise to different combinations. It is often instructive, when exploring data, to view these too.

One criticism of SPAN and of tree-based methods, is that, by searching the data, they are essentially ‘fishing’ expeditions that may turn up something, but of dubious significance. One way to check whether a partition is ‘real’ is to generate binary partitions of the data at random and check whether the partition is statistically significant against the random-sampling distribution. In the three examples presented above, there is overwhelming evidence that these are not spurious (as is, anyway, clear from the conventionally derived OR confidence intervals).

**References**