# Chapter 2

## Methodology

This chapter describes the methods used in the study. The methodology comprises the following components.

1. Study Design
2. Population and sample
3. Variables
4. Data Collection and Data Management
5. Statistical Methods
6. Graphical Methods

### 2.1 Study Design

A cross-sectional survey was used in this study because it is the most appropriate method for collecting data in order to assess the risk factors related to acute diarrhea disease in children aged under 5 years in Pattani Province.

### 2.2 Population and Samples

#### 2.2.1 Sampling Frame

Random sampling was used for selecting house of children in the sample for this study, described as follows.

To get information about the population, all children under 5 years old, irrespective of diarrhea disease, over the period of 1 year (1 January – 31 December 1999) in Pattani Province were used.

Samples used in this study were 220 children under 5 years old who have ever and never been sick with diarrhea disease over the period of 1 year (1 January - 31 December 1999), in Pattani Province. The following multi stage-sampling technique was used for selecting the samples.

1. There were two different groups of districts selected for sampling:

*A. The group of districts, which had the highest incidence of diarrhea disease, including six districts : Tung Yangdaeng, Mae Lan, Panarehk, Mai Kaen, Nong Jig, and Mayo.*

*B. The group of districts, which had the lowest incidence of diarrhea disease, including six districts : Muang, Yaring, Yarang, Kok Pho, Sai Buri and Kar Por.*

2. One district from the first group was sampling by the random sampling technique. This district was Panarehk, where the incidence rate was 4033 per 100,000 population.

3. Twenty-two out of 52 villages in Panarehk District were selected for this study using the random sampling technique.

The schematic diagram of the random stratification method is shown in Figure 2.1

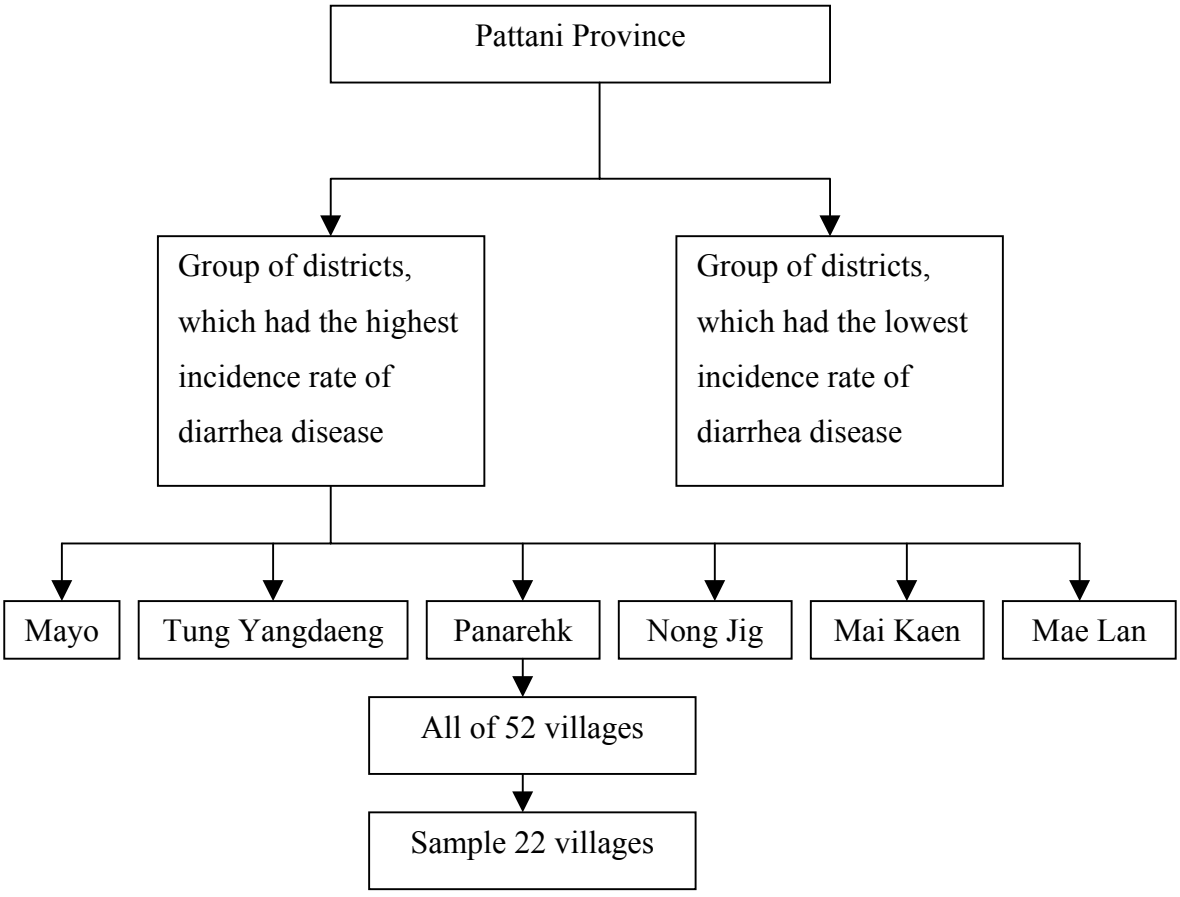



*Figure 2.1 Schematic diagram of random stratification method*

The distributions of the 22 villages of Panarehk District sampled are shown in Table 2.1.

| Name of sub-District | Number of village | Total |
|---|---|---|
| Panarehk | 2, 4 | 2 |
| Bannoog | 1, 2, 3, 6 | 4 |
| Don | 1, 2, 3, 4, 5, 6 | 6 |
| Thanam | 5 | 1 |
| Koog Kra Beu | 1, 3 | 2 |
| Nambo | 1, 2, 3, 4 | 4 |
| Ban Klang | 3, 6 | 2 |
| Thakam | 2 | 1 |
| Total | | 22 |

*Table 2.1 Distribution of the of Pananehk District sampled*

### 2.2.2 Sample size

The sample size needed to obtain a specified precision is estimated using the following equation (McNeil, D. 1996)

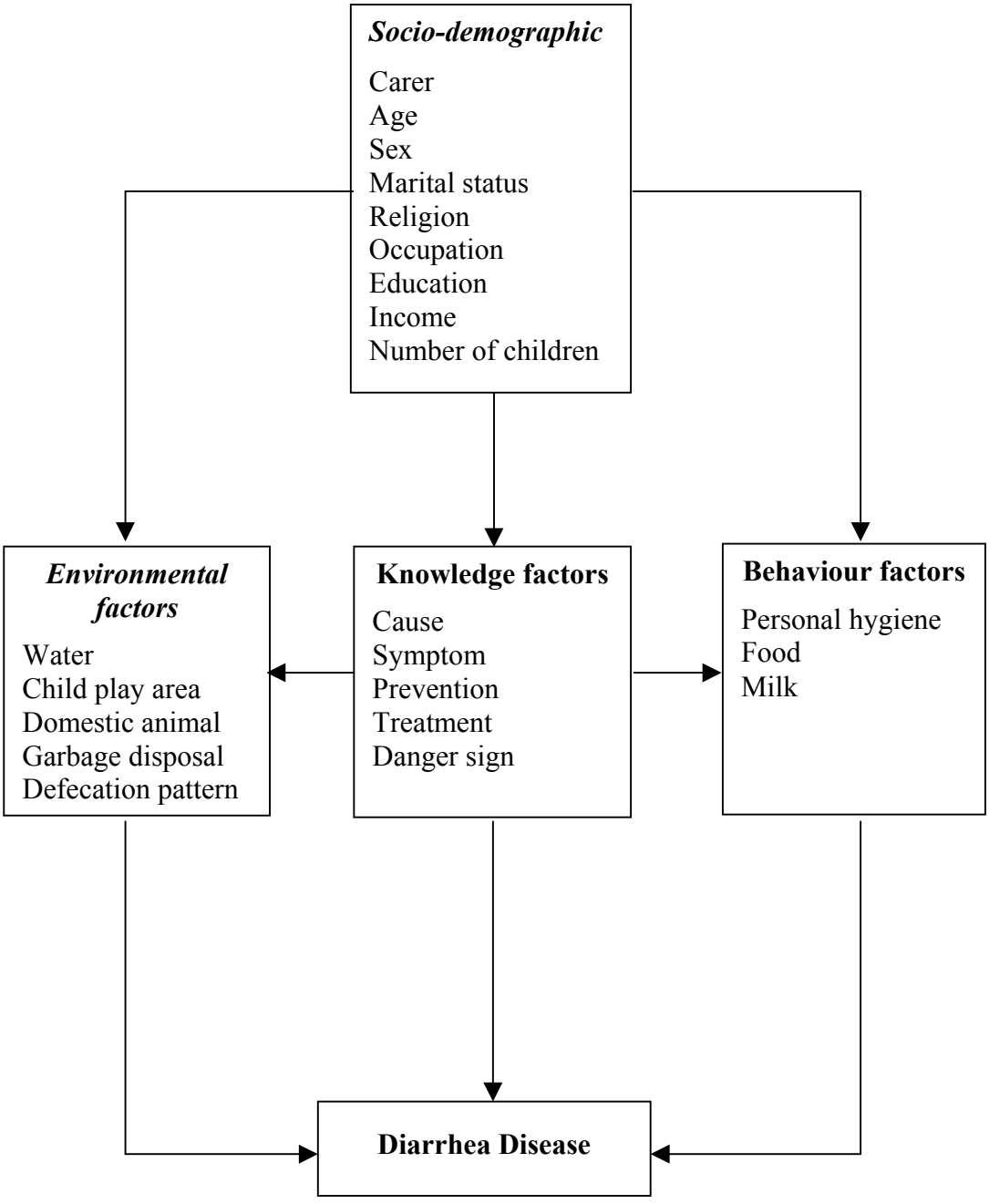
$$\mathrm{n} = Z^2_{\alpha/2} P \frac{(1-P)}{d^2} \qquad (2.1)$$

In this formula $Z_{\alpha/2}$ is the critical value for standardized normal distribution corresponding to a two-tail probability $\alpha$ . The parameter *P* is the probability of an adverse outcome. The value *d* is half of width of the 100(1-$\alpha$) % confidence interval.

An estimate of *P* is 0.17, the outcome being defined as sick with diarrhea disease children under 5 years in Pattani Province. $\alpha$ is 0.05 and *d* is 0.05. The number of mother and child carers to be sampled in of the study should thus be

$$\mathrm{n} = 1.96^2 \times 0.17 \frac{(1-0.17)}{.05^2} \qquad \mathrm{n} = 220$$

The schematic diagram for this study is shown in Figure 2.2




*Fiqure 2.2 Schematic diagram of variables of interest*

## 2.3 Variables

***Determinants***

- *Socio-demographic factors* : carers, age, sex, marital status, religion, occupation, income and education of carers.

- *Environmental factors* : water, child play area, domestic animal, garbage disposal and defecation pattern.

- *Knowledge factors* : causes of diarrhea disease, symptoms, prevention, treatment and danger sign.

- *Behaviour factors* : personal hygiene, food and milk.

***Outcome variable***

- Sick with diarrhea disease in children aged under 5 years.

## 2.4 Data Collection and Data Management

***Data collection***

To collect data, a questionnaire was used in order to get information about risk factors related to acute diarrhea disease in children under 5 years old in Pattani Province, for a year (1 January – 31 December 1999) This questionnaire included four main factors as follows.

*1. Socio-demographic factors.*
*2. Environmental factors.*
*3. Knowledge factors about diarrhea disease.*
*4. Behavior factors for prevention of diarrhea disease.*

The methodology used random sampling of one of every four houses starting from the centre of the village. It there was no child in the selected house, the next house was sampled, and 10 samples were collected from each of 22 villages. (Random number generator was used to select house).

***Data management***

*Database design using Microsoft Access program.*

*1. ER-diagram consists of 8 tables.*

Table 1: data of socio-demographic factors including index, carer, age, status, religion, occupation, education, income, number of children, number of sick, number of death.

Table 2: data of environmental factors including index, source of water, water quality, place tending, pets, filthy water, keep rubbish, eliminated rubbish, defecate place, eliminate defecate.

Table 3: data of knowledge about the factors, which cause diarrhea diseases, including index, items and response.

Table 4: data of knowledge about the factors, which danger signs of diarrhea diseases, including index, items and response.

Table 5: data of behaviour of child carers for prevention of diarrhea diseases including index, items and behaviour.

Table 6: data of breast milk including index and behaviour.

Table 7: data of can milk including index and behaviour.

Table 8: data of fresh milk including index and behaviour.

*2. Association between each determinant and outcome was determined using all data to put on database for further analysis.*

***Data analysis***

*1. Statistics for descriptive analysis including*

- Percentage
- Mean
- Standard deviation

*2. Statistics for inferential analysis including*

- Odds ratio
- Chi-square
- t-test

*3. Create a model to forecast the risk factors related to acute diarrhea disease.*

- Logistic regression analysis

*4. Program for analysis data of including*

- MATLAB, Microsoft Access, Microsoft Excel

## 2.5 Statistical Methods

From the schematic diagram (Figure 2.2), the model specifies sick with diarrhea disease in children aged under 5 years as the outcome. This study focuses on the association between the outcome and the determinant. Socio-demographic factors, environmental factors, knowledge factors about diarrhea disease and behaviour factors for prevention of diarrhea disease as the determinant. However, the outcome variable is binary and the determinant variables are complicated. To simplify the preliminary analysis the outcome and the determinants are reversed. Descriptive and inferential statistics are used to analyses the variables of interest.

### *Descriptive Statistics*

The variables of interest are summarized by histograms and by mean, standard deviations, minimum and maximum values. The socio-demographic factors and environmental factors are described by percentages.

### *Univariate Analysis*

Peason,s chi-square test , 95% confidence intervals for odds ratios and t-test are used to assess the association between the determinant variables and the outcome of this study. The formulas of contingency tables (McNeil, 1998b) are as follows (X is the determinant of interest, Y is sick with diarrhea disease in children, Z is a stratification variable).

*A. 2 x 2 table*

X is the determinant and Y is the outcome. Each variable is binary (0 or 1).

The odds ratio is a measure of the strength of an association between two binary variables (i.e., in which both the outcome and the determinant are dichotomous) (McNeil, 1998a, 1998b). That describes the degree of association between two variables in different risk factors related to acute diarrhea in children aged under 5 years in Pattani Province. To illustrate the definition of the odds ratio, a two-by-two table is constructed as follows.

| | | Y = 1 | Y = 0 |
|---|---|---|---|
| X | 1 | a | b |
| | 0 | c | d |

n = a + b + c + d

The ratio of these odds is referred to as the odds ratio (McNeil 1996, 97). Thus the estimate the odds ratio is

$$\text{OR} = \frac{ad}{bc} \tag{2.2}$$

One method of testing the null hypothesis of no association between determinant and outcome is using a z-statistic z = ln(OR)/SE, where SE is the standard error of the natural logarithm of the odds ratio (McNeil 1996, 97). Its asymptotic standard error is given by

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{2.3}$$

A 95 % confidence interval is thus

$$95\ \%\ \text{CI} = \text{OR x exp}\ (\pm 1.96\ \text{SE}\ [\ln \text{OR}]) \tag{2.4}$$

Pearson's chi-square statistic is defined as

$$\chi^2 = \frac{(ab - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \tag{2.5}$$

*B. Non-stratified r×c tables*

In this study, some of variables are multi categorical. We use non-stratified r×c table to compare them. For example, X is category of clean fingernails and Y is category of sick and not sick with diarrhea disease.

Assume X is nominal (1,2,…,r), Y is binary (1,2). For each category (X), seeing from a two-by-two table by aggregating counts (McNeil, 1998b, 205). Thus

| | | Y | |
|---|---|---|---|
| | | 1 | 2 |
| X | 1 | $a_{11}$ | $a_{12}$ |
| | 2 | $a_{21}$ | $a_{22}$ |
| | 3 | $a_{31}$ | $a_{32}$ |

Thus the estimate the odds ratio (OR) is

$$\mathrm{OR_{ij}} = \frac{a_{ij} d_{ij}}{b_{ij} c_{ij}} \tag{2.6}$$

Where $\mathrm{b_{ij}} = \sum_{j=1}^{2} a_{ij} - a_{ij}$, $\mathrm{c_{ij}} = \sum_{j=1}^{r} a_{ij} - a_{ij}$, $\mathrm{d_{ij}} = n - a_{ij} - b_{ij} - c_{ij}$, $\mathrm{n} = \sum_{i=1}^{r}\sum_{j=1}^{2} a_{ij}$

The standard error of the natural logarithm of the odds ratio is given by the same formula as for the two-by-two table. In general, the association is composed of $r$ odds ratios, but only r-1 of them are independent.

The standard error is given by

$$SE(\ln OR_{ij}) = \sqrt{\frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}} \tag{2.7}$$

A 95 % confidence interval is thus

$$95\ \%\ \mathrm{CI} = \mathrm{OR} \times \exp(\pm 1.96\ \mathrm{SE}[\ln \mathrm{OR}]) \tag{2.8}$$

Pearson's chi-square statistic for independence (i.e., no association) is an *r* x *2* table define as

$$x^2 = \sum_{i=1}^{r} \frac{(a_i - \hat{a}_i)^2}{\hat{a}i} + \sum_{i=1}^{r} \frac{(b_i - \hat{b}_i)^2}{\hat{b}_i} \tag{2.9}$$

where $\hat{a}_i = \dfrac{(a_i + b_i)\sum_{k=1}^{r} a_k}{\sum_{k=1}^{r}(a_k + b_k)}$ and $\hat{b}_i = \dfrac{(a_i + b_i)\sum_{k=1}^{r} b_k}{\sum_{k=1}^{r}(a_k + b_k)}$

When the null hypothesis of the independence is true, this has a chi-squared distribution with *r-1* degree of freedom. (McNeil 1998b, 205).

***Two-sample t test***

Two-sample *t test* was used to compare the mean of determinant is of the binary type and the outcome variable is continuously varying. The appropriate method is based on either a two-sample *t test* (if the data in the two determinant groups are independent samples) or a paired *t tests* (if the samples are matched). And so it goes.

For this study we used two-sample *t test*, because two determinant groups (sick and not sick with diarrhea disease) are independent samples and outcome age is continuously variables

The null hypothesis that two populations mean are equal may be expressed as

$$\mathrm{H_o}: \ \mu_1 = \mu_2 \tag{2.10}$$

where $\mu_1$ and $\mu_2$ are the respective population means.

If samples of size $n_1$ and $n_2$ are taken from the two populations, giving sample means $\bar{y}_1$ and $\bar{y}_2$, a *t* statistic may be used to test this null hypothesis, just as a *t* statistic is used to test a hypothesis for a single population mean. The two-sample *t* statistic takes the from.

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.11}$$

In this formula (McNeil 1996, 44)., $s$ is the pooled sample standard deviation, defined be subtracting the sample mean from each sample to give a set of $n_1 + n_2$ residuals, dividing the sum of the squares of these residuals by $n_1 + n_2 - 2$ and taking the square root of the result. It $s_1 + s_2$ denote the standard deviations of the two samples, respectively, it must be shown that the pooled sample standard deviation is given be the formula.

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2.12}$$

A p-value is now obtainable from the table of the two-tailed *t* distribution with $n_1 + n_2$ - *2* degrees of freedom. (McNeil, 19996, 43).

***Logistic Regression***

Multiple logistic regression analysis is used for adjusting the association between determinant variables and sick with diarrhea disease. Logistic regression is a method of analysis that gives particularly simple representation for logarithm of the odd ratio association with risk factors, and when fitted to data involving binary outcome and determinant, it automatically provides estimates of odds ratio and confidence intervals for specific combinations of the risk factors. (McNeil, 19996< 125). The method is defined as

$$\ln\left(\frac{P}{1-P}\right) = a + b_1x_1 + b_2x_2 + .... + b_px_p \tag{2.13}$$

In this formula, *P* denotes the probability of occurrence of the outcome variables and $\{x_j\}$represents the $j^{th}$ determinant variable, *a* is the constant coefficient, and $\{b_j\}$ is the set of regression coefficients. This equation may be inverted to give an expression for the probability *P* (McNeil 1996, 129).

$$P = \frac{1}{1 + \exp(-a - \sum_{j=1}^{p} b_j x_j)} \tag{2.14}$$

Logistic regression is in common use in epidemiological studies, which are mostly concerned with a dichotomous or binary outcome. It describes the relationship between a dichotomous response a set of continuous or categorical variables. The difference between the logistic regression model and the linear regression model is that the outcome variables is required to be dichotomous for the regression model rather than continuous for the linear regression model. For a set of predictor variables $x_1, x_{2,\ldots}, x_n$ (McNeil 1996, 138). the logistic regression model they take the form :

$$\ln(\frac{P}{1-P}) = \alpha + \sum_{i=1}^{p} \beta_i x_i \tag{2.15}$$

Where $P$ denotes probability of occurrence of the outcome and $x_i$ is determinant, its formula can be shown as

$$P[Y=1] = \frac{\exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)} \tag{2.16}$$

Using the logistic regression model for the data arising from a two-by-two table, we suppose $x_i = 1$ or 0 (i = 1, 2, …, p) which the value of determinant X=( $x_1, x_{2,\ldots}, x_n$ ) is taken to be 1 (exposure) and 0 (no exposure). (McNeil, 1996, 138). Thus the logistic regression model (1) can be written as follows

$$\ln\{\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)}\} = \alpha + \beta \tag{2.17}$$

$$\ln\{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}\} = \alpha \tag{2.18}$$

The equations (2.18) and (2.19) actually are the (natural) logarithms of the odds for the outcome given the exposed (X=1) and non-exposed (x=0), respectively. After exponentialting each equation, the odds for the exposure and non-exposure group can

be written as $\exp(\alpha + \beta)$ and $\exp(\alpha)$ respectively. The odds ratio therefore is obtained from the simple formula. (McNeil, 19966, 138).

$$\text{OR} = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta) \tag{2.19}$$

## 2.6 Graphical Methods

Matlab version 5 ( Hanselman and Littlefield, 1997) and Asp (McNeil, 1998) were used for graphical presentation and statistical anslysis.

### *Histogram*

Histograms and statistical summaries of a set of variables that are determinants, outcome, or other variables graph the data. Odd ratio plots graph the associations between the outcome variables and the determinants of interest. The logistic regression in the multivariate analysis is graphed by logistic regression printouts, which may include standardized residuals plots.

Histograms with statistical summaries of raw data for all of variables represents the distribution and summaries including the size, mean, standard deviation, minimum and maximum of a set of data. A histogram presents the data as bar extending away from the axis representing independent variable. The length of each bar is determinant by the value of the dependent variable.

### *Odds Ratio Plot*

Graph of the odd ratios and 95% confidence intervals can be used to present the association between two nominal categorical variables. The association between the outcome variable and the determinant of interest is investigated by an odd ratio, which provides a useful measure. The graph of an odd ratio includes a 95% confidence interval. Confidence intervals may be graphed using line interval. The dot on the line interval is the estimated odd ratio. For an odd ratio, the null value is conventionally taken to be 1, corresponding to equal risks of an outcome in two comparison groups.

This corresponds to a null value of 0 for the difference between two the population proportions under the null hypothesis represent by the dotted line. The p-value shown at the top is the overall (Pearson's chi-square) test of no relationship between the determinant and the outcome. Additionally, the p-value shown in the horizontal panels of the graph may be used to assess the associations between the outcome and a set of binary determinants obtained by aggregating the counts for the unspecified levels of the determinant. The homogeneity test is used to tell if the association could be the same in the different strata, small p-values providing evidence to the contrary.