Application of a Latin square experimental design in health services research: estimation of the effects of setting clinical standards and performance review on the process and outcome of care in general practice

by

I. Nicholas Steen BSc, MSc

SUBMISSION FOR PhD DEGREE

TO THE UNIVERSITY OF NEWCASTLE UPON TYNE

THE WORK FOR THIS THESIS WAS UNDERTAKEN IN:

THE CENTRE FOR HEALTH SERVICES RESEARCH
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE
UNIVERSITY OF NEWCASTLE UPON TYNE

Date of submission: December 1997

Abstract

The North of England Study of Standards and Performance in General Practice was set up to investigate whether the setting of clinical standards was an effective way of improving clinical performance (North of England Study, 1991). Doctors from 60 training practices met in small groups to set standards of good clinical performance for five symptomatic conditions of childhood—acute cough; acute vomiting; itchy rash; bedwetting; and recurrent wheezy chest. Data on the process and outcome of care were collected both before and after standard setting process. Some of the baseline data was fed back to the doctors to enable them to evaluate their performance in the first phase of the study. The clinical standards and baseline data were disseminated to the small groups of doctors according to a Latin square design. By comparing responses obtained during the second phase of data collection (after standards had been set) with those obtained in the first, it was possible to estimate the effects of standard setting and other methods of performance review on what doctors did (the process of care) and on the resulting outcome of care for their patients.

The general analytic approach adopted was to fit generalised linear models to try and explain the variation in the observed data. Within this general framework, methods were developed for coping with a wide range of statistical problems including: heteroscedasticity; correlated binary responses; loss of orthogonality arising because of the incompleteness of many of the data sets; and overdispersion.

The setting of clinical standards was found to have influenced doctors' prescribing of drugs and was found to have had a beneficial influence on outcome of care for children suffering from recurrent wheezy chest. Implications for the design of future studies to evaluate this type of intervention in the health service are discussed.

List of contents

List of figures

List of tables

Preface

During the 1970s general practitioners who were involved in the vocational training scheme run by the Northern Regional Postgraduate Institute came to recognise the value of performance review as a form of continuing education. Many of them met in small groups to set clinical standards for a range of common conditions and as a result they became interested in whether setting clinical standards was an effective way of improving clinical performance. The North of England Study of Standards and Performance in General Practice was set up to answer this question. The author was appointed to the study team, towards the end of the study in 1989, as a research associate with responsibility for undertaking the analysis required to estimate the effects of the study interventions on the process and outcome of care provided by the participating doctors. This thesis describes that analysis.

Although the author played no part in their development, the educational component of the study and the study design are described briefly in Chapter 1 to provide context in which the analysis can be considered. Comprehensive descriptions of these aspects of the study are given in the final report -volumes I and II (North of England Study, 1990a and 1990b).

Acknowledgements

The North of England Study was undertaken by a large multi-disciplinary team. I would like to acknowledge the contribution of all members of that team. In particular I would like to offer my deepest thanks and acknowledge the invaluable support and advice offered by the other two members of the statistics group: Professor Ian Russell and my supervisor Dr Peter Avery. Other members of the study team who made noteworthy contributions during the analysis and to whom I would like to extend my thanks are Elaine McColl, Jenny Hewison, Martin Eccles and Claire Bamford. I would particularly like to thank Elaine McColl for her comments on a draft of this thesis and Martin Eccles for his advice on clinical issues relating to the analysis.

Author's declaration

Work for this thesis was undertaken whilst I was employed as a research associate at the University of Newcastle Upon Tyne. The work was funded by the Department of Health. Some of the material contained within this thesis has previously been published in the following papers:

Medical audit in general practice. I: effects on doctors' clinical behaviour for common childhood conditions. North of England Study of Standards and Performance in General Practice, *The British Medical Journal;* 1992, 304; 1480-1484.

Medical audit in general practice. II: effects on health of patients with common childhood conditions. North of England Study of Standards and Performance in General Practice, *The British Medical Journal;* 1992, 304; 1484-1488.

Chapter 1

Study design and overview of thesis

1.1 North of England Study

The North of England Study of Standards and Performance in General Practice took place between 1982 and 1990 and was the first comprehensive evaluation of performance review in British general practice (Irvine et al. 1986a and 1986b). The study had four main aims:

1. To develop methods of setting explicit standards of good performance in general practice (that is, statements of what general practitioners **should** do or achieve);

2. To compare clinical performance (that is, what general practitioners **actually** do or achieve) with these standards;

3. To estimate the effects (both on doctors' clinical behaviour and on patients' health) of setting standards and of receiving feedback;

4. To evaluate the costs and benefits of setting standards and assessing clinical performance in general practice.

The work which forms the basis of this thesis relates to the third of these aims—estimation of the effects of setting standards and other methods of performance review on process and outcome of care.

1.2 Educational component

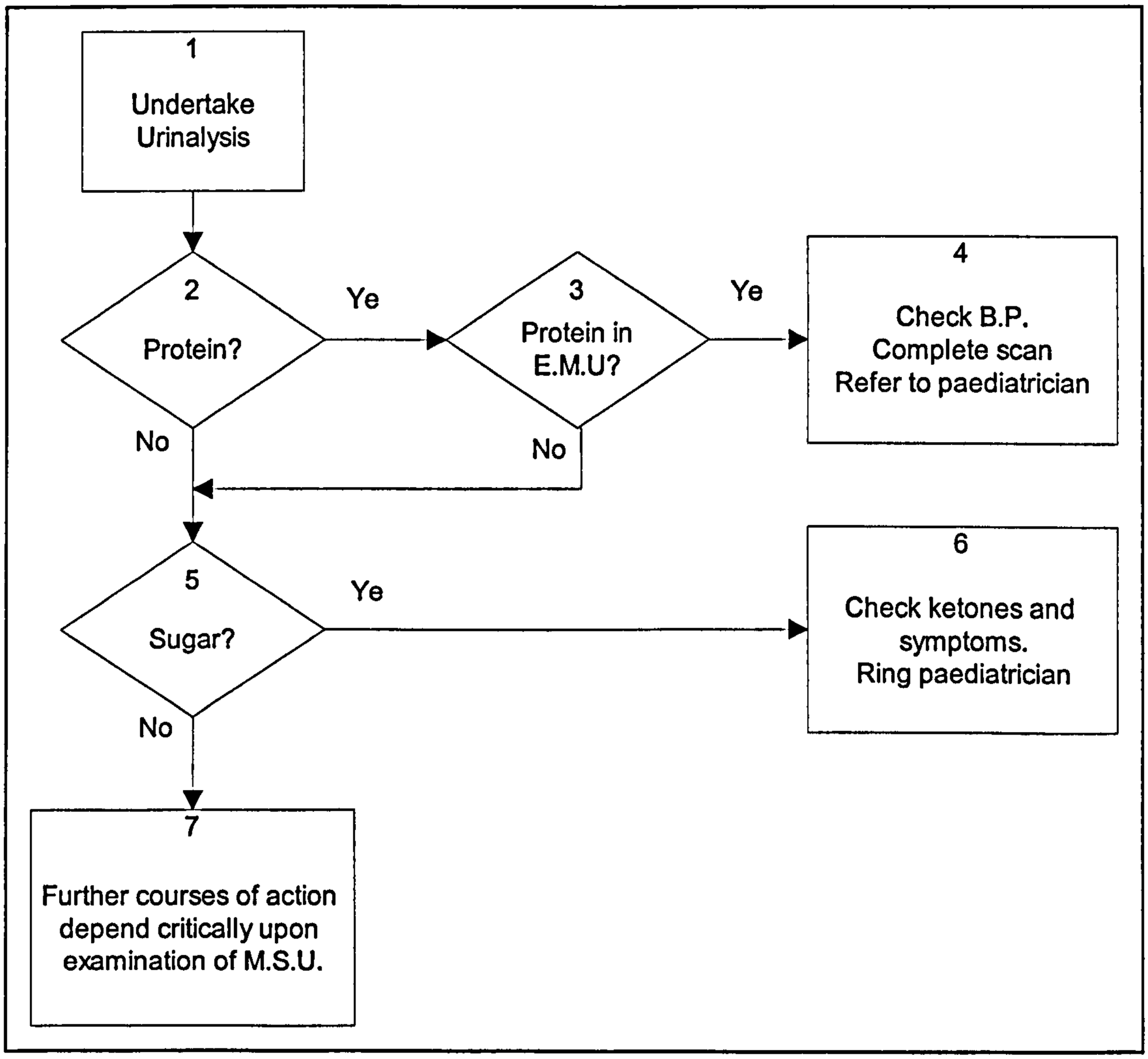
As part of their continuing post-graduate education, general practitioner trainers in the Northern Region were encouraged to take part in small group work that involved aspects of performance review. All trainers who were neither single-handed (it was anticipated that such doctors would find it difficult to attend the necessary meetings and that the burden imposed on them would be too great) nor close to retirement were invited to take part in the study. Eighty nine trainers (86% of those eligible) originally agreed and 84 (79%) completed the study; the five who withdrew all did so for reasons unconnected with the study. The doctors were divided into ten *'trainer groups'*. This allocation was essentially geographical. This facilitated communication between group members and reduced the likelihood of discussion between members of different groups.

These groups of doctors were then asked to become involved in various aspects of performance review (or medical audit as it is now known). Each group was asked to define good primary care for children with one specific symptomatic condition. They were encouraged to use algorithms (such as that given in Figure 1.1) or any other format in which the recommended course of action depended critically on the information available. The conditions (acute cough, acute vomiting, itchy rash, bedwetting and recurrent wheezy chest) were chosen by the principal investigators to span the range of skills and resources used by general practitioners. For each condition, clinical standards were set independently by two of the trainer groups.

At the time the study was being developed there was interest in the role of specialists in setting standards. A clinical standard was therefore also drawn up for each condition by one of five *'mixed groups'*. These comprised two consultant paediatricians, two experienced general practitioners and a member of the research team. Shortly before the

standards were finalised, one of the two trainer groups working on a particular condition met and exchanged standards with the corresponding mixed group. Thus five of the

Figure 1.1 Extract from clinical standard for treatment of bedwetting



trainer groups had an opportunity to re-examine and develop their draft standards in the light of paediatric input.

Altogether, five different levels of performance review were investigated (Figure 1.2). Since there were five study conditions, an individual trainer group experienced a different intervention for each one. In addition to setting a standard, trainer groups received a standard set by another trainer group for a second condition. Two other

interventions took the form of different levels of feedback on current clinical performance. The final intervention was to receive no feedback of any sort.

Figure 1.2 The study interventions

Each trainer group:

a. set a standard for one clinical condition - half of the groups received paediatric input (a1) the others did not (a2);

b. received a copy of the clinical standard set by another trainer group for a second condition;

c. received comparative feedback for a third condition, comparing their baseline performance for that condition with that of all other participating principals;

d. received descriptive feedback for a fourth condition, summarising their baseline performance for that condition; and

e. experienced no intervention for the fifth condition.

1.3 Study design

The basic design chosen for the evaluation was a before and after study consisting of three essential stages, each lasting one year in each study practice (Figure 1.3). In order to evaluate the effects of the educational interventions on the care of children it was decided to collect data on both the *process* and *outcome* of care. This data collection was undertaken before and after the setting and dissemination of clinical standards and feedback.

In evaluating the effects of any intervention in health care, one of the most difficult problems is that of inferring what would have happened in the absence of intervention.

The simple before and after study is susceptible to many sources of bias (Russell, 1983), including secular trends (e.g. in prescribing habits) and sudden external changes (e.g. introduction of the selected list of drugs). To ameliorate some of these problems, the

Figure 1.3 Before and after design of the north of England study

- **BEFORE** standard setting (Phase 1):
  - collection of baseline data on clinical performance (the process of care) and patient outcome
- **DURING** standard setting:
  - clinical standards were set for selected conditions
  - feedback of standards and baseline data to study doctors
- **AFTER** standard setting (Phase 2):
  - collection of follow-up data on clinical performance and patient outcome
  - comparison of follow-up data with baseline data

educational interventions were differentially distributed according to a replicated Latin square design (Table 1.1). The first replicate corresponds to the five trainer groups which received paediatric input; the second to the five groups which did not. The main feature of the design is that it allows an observed change in performance for a pair of trainer groups who set a standard for a particular condition to be tested simultaneously against:

(i) the change in performance of the other pairs of trainer groups who did not set a standard for that condition; and

(ii) the change in performance of that particular pair of trainer groups for the other four study conditions.

Trainer groups A to K were randomly allocated to the various audit groups. From the two trainer groups that set a standard for a particular condition, one of them was selected at random to be involved with discussion with the corresponding mixed group.

Table 1.1 Experimental design of north of England study: type of audit undertaken for each study condition by trainer groups A to K

Type of audit	Study condition				
	Acute cough	**Itchy rash**	**Acute vomiting**	**Recurrent wheeze**	**Bedwetting**
Set clinical standard:					
discussion with mixed group	G*	A	C	E	J
no discussion with mixed group	B	F	H	K	D
Receive clinical standard	J	G*	A	C	E
from another trainer group	F	H	K	D	B
Receive comparative data	E	J	G*	A	C
from all participating doctors	H	K	D	B	F
Receive comparative data	C	E	J	G*	A
from own trainer group	K	D	B	F	H
None	A	C	E	J	G*
	D	B	F	H	K

* From the top left hand corner, of the two trainer groups who set a clinical standard for acute cough, group G met the corresponding mixed group and group B worked entirely on their own; group G also received a standard for itchy rash, comparative data for vomiting and descriptive data for wheezy chest but experienced no form of audit for bedwetting

Each type of audit was equivalent for each GP in a particular trainer group. All GPs within a particular group received the same set of feedback.

1.4 Survey methods

1.4.1 Identifying children with study conditions

Two complementary methods were used to identify children with study conditions. In each practice, beginning in a random week between August 1984 and July 1985, doctors (both trainers and participating partners) were asked to identify prospectively all children consulting with them for any of the five conditions over a six week period. This was repeated, beginning in the corresponding week, for a six week period in 1986-87.

Data from a pilot study (Russell et al. 1986) indicated that, because of the low prevalence of the chronic conditions (bedwetting, itchy rash and recurrent wheezy chest), this prospective identification would not yield sufficient children with those conditions. A prevalence questionnaire was therefore sent to the parents of all children registered with the study practices. This included questions about whether the child had suffered from any of the three chronic conditions over the previous twelve months and from the two acute conditions (acute cough and vomiting) in the previous four weeks.

A sample of these children was subsequently selected for further study. Children with acute cough and vomiting were sampled randomly from those identified prospectively by doctors; children with the three chronic conditions were sampled randomly from those identified by both methods (that is prospectively by the doctor and retrospectively by the prevalence survey). In practice, the bulk of children suffering from a chronic condition were identified from the prevalence survey but unfortunately there was no variable created in the data files that indicated how a child was identified. The medical records of sampled children were marked by members of the study team to identify them.

1.4.2 Collecting data on the process of care

Pilot work indicated that medical records in British general practice serve mainly as aides-memoir; they do not provide a comprehensive account of all aspects of the process of care. Doctors (both trainers and their partners) were therefore asked to enhance the medical records of children recruited to the study so as to provide a coherent account of their care. For this purpose they were supplied with an *enhancement form*, a single sheet of A4 with five headings - diagnosis or formulation of the problem, history on which this diagnosis or formulation was based, examinations and investigations, management decisions and reasons for these management decisions. They were asked to complete an enhancement form after each relevant consultation with children suffering from acute cough or acute vomiting, until the end of the episode or illness. For children identified as suffering from one of the three chronic conditions , they were asked to enhance the records in this way after each consultation over the following 12 months. These enhancement forms were kept in special folders separate from the statutory medical records.

Both the statutory records and the enhancement forms were subsequently extracted by field-workers on the study team. This involved summarising the contents in numerical form using a coding system devised initially by the five mixed groups and revised subsequently by the study team to incorporate elements from the three clinical standards set for each condition.

1.4.3 Collecting data on the outcome of care

Data on the outcome of care were collected from samples of parents of children suffering from the study conditions through personal interviews and postal questionnaires. While *parent interviews* gathered information in considerable depth from a relatively small sample, *postal outcome questionnaires* were used to gather relatively limited information but from a large sample of parents. Due to limited

resources it was only possible to carry out interviews for three of the conditions—acute cough, bedwetting and recurrent wheezy chest. As it was possible to interview the parents of all children who suffered from bedwetting (due to the low prevalence of this condition) no postal outcome questionnaire was developed for this condition.

Outcome was assessed differently for acute and chronic conditions. Since most episodes of acute illness are short-lived and self-contained, outcome for acute cough and acute vomiting was assessed shortly after the child had consulted. For these conditions, interviews or questionnaires were administered only once. For chronic conditions, treatment regimes extend over several months; any improvement may be gradual. For bedwetting, itchy rash and recurrent wheezy chest outcome was measured twice. The initial interview or questionnaire was administered a few weeks after the child had been identified (denoted subphase A), the second a year later (denoted subphase B). This allowed the comparison of the change in outcome over the year after standard setting with the change in outcome over the same period before standard setting.

The interview schedules and postal questionnaires were developed by the study team. All data collection instruments were piloted but there was no formal assessment of their validity or reliability. Both interviews and postal questionnaires contained a set of questions to assess satisfaction with the care provided during a visit to the surgery. These questions were an anglicised adaptation of an American satisfaction scale that had been shown to be valid (Zastowny et al, 1983) in the US, although no assessment of its reliability or validity in the UK setting has been reported.

1.5 Timetable of key research activities in a single practice

The timetable for undertaking the research activities described in the previous section is given in Table 1.2. Research activities in each practice began in the same week after the second phase of data collection as they did in the first. Practices with two participating

Table 1.2 Timetable of key research activities in a single practice at the beginning of each data collection phase

Week no.	Activity
1	Start of subphase A.
3	Doctor begins to identify prospectively children presenting with study conditions. Records for children with acute conditions are enhanced for the next 6 weeks.
4	At end of week prevalence survey sent to all children registered with that practice.
10	Parents of a sample of children suffering from acute cough, bedwetting and recurrent wheezy chest are interviewed by researchers.
11	(1) Parents of a second sample of children suffering from acute cough, vomiting, itchy rash and wheezy chest are sent Postal Outcome Questionnaires. (2) Records for children with a chronic condition are enhanced for the next 12 months
15	Reminders sent to all parents who had not replied to postal questionnaire.
53	Start of subphase B
62	Parents of children in first sample with bedwetting and recurrent wheezy chest are interviewed for a second time.
63	Parents of children in second sample suffering from itchy rash or recurrent wheezy chest are sent a second Postal Outcome Questionnaire.

trainers were allocated a week to themselves; practices with one trainer were scheduled two practices to a week. Each practice had the same logical sequence of interdependent research activities.

The prevalence survey was sent out at the end of week 4. Parents of a first sample of children were interviewed in week 10; parents of a second sample of children were sent postal outcome questionnaires in week 11. For children with a chronic condition, outcome data were collected again twelve months later; parents of the first sample of

children were interviewed in week 62 and parents of the second sample of children were sent a postal questionnaire in week 63.

Data collection was completed by the end of August 1988.

1.6 Overview of thesis

The aim of this thesis is to describe the statistical aspects of the evaluation of the effects of standard setting and performance review on doctors' behaviour and the outcome of care for their patients. The key objective of the analysis was to produce interval estimates of these effects. There were a number of features inherent in the study that made it difficult to apply standard methods of analysing a Latin square. Alternative analytic strategies that attempt to overcome these problems are considered in this thesis.

The evaluation comprised an analysis of four data sets:

1. the prevalence survey (reported prevalence and consultation rates);
2. process of care (data from enhanced medical records and statutory records);
3. outcome of care 1 - interviews with parents;
4. outcome of care 2 - postal questionnaires.

An introduction to the statistical methods used during the course of the analysis is given in Chapter 2. This begins with the standard treatment of a complete Latin square with one observation per cell. Consideration is then given to methods that might be applied when the property of orthogonality is lost. This includes an introduction to the technique of generalised linear modelling which was used to analyse three of the data sets. Aspects of hypothesis testing and the modelling strategy adopted make up the remainder of the chapter.

The prevalence data set was the first to be analysed. The data set was complete (data were collected from all participating practices for each of the study conditions). It was therefore possible to analyse this data set using standard methods. This analysis is described in Chapter 3.

A primary goal of the analysis of the prevalence survey was to provide a blueprint for the analysis of the much more complex process and outcome data sets. As the analysis proceeded, however, it became apparent that there were aspects of each of the remaining three data sets that made it very difficult to apply standard methods. A description of the process data set is given in Chapter 4. It is evident from the data presented here that there was a major shortfall in the recruitment of children in the second phase of the study (after the interventions had been implemented) resulting in a loss of orthogonality. The actual analysis of the process data set using generalised linear modelling is described in Chapters 5, 6, 7 and 8 (corresponding to the recording of history, diagnosis of episode, non-drug management and drug management respectively).

The analysis of data arising from interviews with parents is described in Chapters 9 and 10. The main features of this data set which made it difficult to apply standard methods were that:

1. due to the limitation of resources interviews were only undertaken for three of the five conditions (thus the data set was not orthogonal by design),
2. for children suffering from the two chronic conditions (bedwetting and recurrent wheezy chest), parents were interviewed twice, one year apart; for children suffering from acute cough, parents were interviewed only once (this was true for both samples of children - those selected before standard setting and those selected after).

Generalised linear modelling was used to analyse clinical outcome (Chapter 9) and patient satisfaction (Chapter 10).

The final data set to be analysed was the postal outcome survey. By design, it too was incomplete. Because of the low prevalence of bedwetting, it was felt that sufficient resources would be available to interview the parents of all children suffering from this condition; questionnaires were not sent to parents of children suffering from this condition. As with interviews, for the two chronic conditions, there were two administrations of the postal survey in each phase but only there was only one administration for each of the two acute conditions. Finally the administration of the postal survey was not carried out exactly as planned; there were substantial differences between phase 1 and phase 2. Attempts were made to take these differences into account within the generalised linear modelling framework. The analysis of the postal outcome survey is reported in Chapter 11.

Many of the issues that arose during the analysis are discussed in the final chapter. There is a critical assessment of the analytic strategy adopted. The extent to which the key objective (of producing interval estimates for the effect of standard setting and performance review on the process and outcome of care) was achieved is addressed. Many of the statistical problems arose because of features that were very specific to this particular study and so generalisability of findings to other studies is limited but, where possible, recommendations have been made to assist prospective researchers who may wish to undertake a similar study.

Chapter 2

Statistical methods

2.1 Introduction

The purpose of this chapter is to provide an introduction to some of the statistical issues which arose during the evaluation of performance review. The standard analytical model for a replicated Latin square is discussed. In particular the assumptions underlying the usual methods of analysis are considered. Examples are then given to demonstrate why many of these assumptions cannot be justified when fitting the model to the main data sets compiled during the study. Alternative methods of analysis are introduced.

2.2 Classical methods of analysis

An alternative presentation (that which is usually adopted in statistical texts) of the study design is given in Figure 2.1. The number of columns (study conditions) is exactly five and, provided that a_1 and a_2 are regarded as alternatives within the same type of audit ('(a) Set clinical standard'), the number of treatments (types of audit is also five). Although the number of trainer groups is ten, pairing those groups that set a standard for the same study condition yields five rows. Since one group within each pair met a mixed group while the other did not there is a separate replicate of the Latin square for each member of the pair.

Figure 2.1 Latin square design for type of medical audit undertaken by each trainer group for each study condition

Trainer groups	Study condition				
	Acute cough	Itchy rash	Acute vomiting	Recurrent wheezy chest	Bedwetting
G, B	a_1, a_2	b, e	c, d	d, c	e, b
A, F	e, b	a_1, a_2	b, e	c, d	d, c
C, H	d, c	e, b	a_1, a_2	b, e	c, d
E, K	c, d	d, c	e, b	a_1, a_2	b, e
J, D	b, e	c, d	d, c	e, b	a_1, a_2

Key to types of medical audit:

a_1 set clinical standard and met corresponding mixed group

a_2 set clinical standard without paediatric input

b receive clinical standard set by another group

c Receive baseline data from all participating doctors

d Receive baseline data only from own trainer group

e None

Example

The top left-hand cell shows that trainer groups G and B set clinical standards for acute cough and that group G but not group B, met the corresponding mixed group.

The statistical model for this design is:

$$Y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_m + \alpha\delta_{im} + \beta\delta_{jm} + \gamma\delta_{km} + \varepsilon_{ijkm} \qquad (2.1)$$

for $i = 1, 2, 3, 4$ or 5; $j = 1, 2, 3, 4$ or 5; $k = 1, 2, 3, 4$ or 5; and $m = 1$ or 2.

where Y_{ijkm} is the value of an observed variable in row i, column j and replicate m (receiving treatment k); α_i is the effect of row i (trainer group pair); β_j is the effect of

column j; γ_k is the effect of treatment type k (audit type); δ_m is the effect of replicate m (met mixed group or not); $\alpha\delta_{im}$, $\beta\delta_{jm}$ and $\gamma\delta_{km}$ are the effects of the two way interactions between the main effects defined above; and ε_{ijkm} is a Normal independently distributed random error with $Var(\varepsilon_{ijkm}) = \sigma^2$ for all i, j, k and m.

The main effects described in equation 2.1 are orthogonal; the total sum of squares can be partitioned into components for rows, columns, treatments, replicates, the three interaction terms and an error term. If there is one observation per cell, the analysis will take the form shown in Table 2.1. Provided assumptions about the distribution of errors are justified, the null hypothesis of no main effects or no interaction effects are tested by comparing the ratio of the corresponding mean square and the error mean square with the appropriate F distribution (Cochran and Cox, 1992).

Table 2.1 Testing the null hypothesis of no main effects and no interaction effects

	Term	Degrees of freedom	Mean square	F ratio
Main effects				
A	Rows (pairs of trainer groups)	4	MS_A	MS_A/MS_E
B	Columns (study conditions)	4	MS_B	MS_B/MS_E
C	Treatments (five levels of audit)	4	MS_C	MS_C/MS_E
D	Replicate (met mixed group or not)	1	MS_D	MS_D/MS_E
Interactions				
	A by D	4	MS_{AD}	MS_{AD}/MS_E
	B by D	4	MS_{BD}	MS_{BD}/MS_E
	C by D	4	MS_{CD}	MS_{CD}/MS_E
Residual		24	MS_E	

2.3 Problems in applying classical methods

2.3.1 Initial strategy

Initially it was proposed to analyse the study data sets using the methods described above. In order to obtain a single observation corresponding to each of the cells in Figure 2.1, it was decided to:

1. aggregate responses from individual trainers within each trainer group for each condition in each phase of the study.

2. subtract the aggregated response in phase 2 from the aggregated response in phase 1 to produce a measure of change which would then form the observed variable Y_{ijkm} in equation 2.1.

A number of problems were encountered when this approach was implemented. These are described below. In the analysis of the prevalence survey, reported in the next chapter, it was possible to find solutions to problems that arose but there were additional features of the process and outcome data sets that made it necessary to consider alternative methods of analysis.

2.3.2 Distributional assumptions

The variance ratio test described above (section 2.2) is only valid if the residual errors are normally distributed. For many of the variables of interest in the study this was not the case. A number of variables were in the form of counts (for example, the number of items of present history that were recorded) or proportions (for example the proportion of children referred to hospital) and were not normally distributed. Many variables had skewed distributions (for example, most parents were very satisfied with the care provided during a consultation; just a few were not). A number of approaches to overcome this problem were considered. One alternative was to process the data in

such a way as to normalise the response variable (Armitage and Berry, 1987, page 360). Various transformations were considered. In addition, by averaging the data over all the participating doctors within a particular cell, as application of the Central Limit Theorem would suggest (Mendenhall et al. 1990), good approximations to the Normal distribution were often obtained.

Another approach was to consider alternative error structures for the data. In particular, through the process of fitting generalised linear models (McCullagh and Nelder, 1989), Binomial and Poisson error structures were adopted. This is discussed below (section 2.4).

2.3.3 Homogeneity of variance

In the model defined by equation (2.1) the errors are assumed to be identically distributed. The variance ratio tests described above (section 2.2) are not valid unless the errors have the same variance across all the cells in the Latin square. For a number of variables of interest this was found not to be the case. In particular, for observations in the form of a proportion, P, the variance, given by:

$$\text{Var}(P) = \frac{\theta(1-\theta)}{n},$$

depends upon two parameters - n, the Binomial denominator and θ, the proportion of individuals with the attribute of interest in the population from which the observation was drawn. θ often varied across conditions. For example, the proportion of children referred to hospital was very different for each condition.

Another mechanism which led to inequality in the variance was the generation of cell statistics from unequal samples. The trainer groups were not all exactly the same size. There was large variation in the list size of participating practices. Prevalence and

consultation rates were very different for the five conditions. Thus when data from individual doctors were combined to form a single cell specific statistics the standard errors associated with those statistics were not equal.

Again different approaches to the problem were investigated. Variance stabilising transformations were considered. For example, in the case of observations in the form of a proportion, P, the arc sine transformation, $\psi(P) = \arcsin\sqrt{P}$, was considered.

The variance of the transformed variable, given by:

$$\mathrm{Var}[\arcsin(\sqrt{P})] = 4n^{-1}.$$

(Dobson, 1983, pages 83-84) depends only upon n; it is independent of θ. Thus if the denominator in each cell is the same, the transformed variable may be used in the type of analysis described above. Unfortunately, as mentioned above, n tended to vary between doctors and trainer groups.

Some of the problems associated with lack of homogeneity of variance can be overcome within the framework of generalised linear modelling. This issue is considered below.

2.3.4 Missing data

Missing data arose from two sources. Firstly, by design, some of the data sets were incomplete. Due to resource limitations, outcome data were only collected by interviews with parents for three of the conditions (acute cough, bedwetting and recurrent wheezy chest). Thus two complete columns are missing from the Latin square for that data set. Similarly, because we were able to interview the parents of all children identified as having bedwetting, no postal questionnaires were sent out for that condition. Thus the postal outcome questionnaire data set is also unbalanced.

Data were also missing for non systematic reasons. Doctors sometimes failed to identify their full quota of cases (this was particularly true for the least prevalent condition - bedwetting) or failed to complete enhancement forms. About 20 per cent of the postal outcome questionnaires were not returned and fieldworkers were unable to interview a similar proportion of parents who had been targeted.

In the case of missing rows, missing columns or missing cells, the main impact on the classical mode of analysis is the loss of orthogonality; the main effects are now confounded. The total sum of squares can no longer be partitioned into component parts as in Table 2.1. Modifications to classical methods in order to deal with some of these problems have been devised. Yates (1936) describes the analysis required when a single row or column is missing; the case when more than one row or column is missing is considered by Yates and Hale (1939). Most standard textbooks on experimental design deal with modifications to the analysis in the case of missing cells (e.g. Cochran and Cox, 1992, pages 125-126).

The effect of missing individual observations depends upon the method of analysis. If the observations within a cell are aggregated to form a single statistic, the effect of the missing data is to increase the standard error associated with that statistic. This may lead to the problems of lack of homogeneity of variance described above. If the observations are analysed as repeated measures within a cell, unequal numbers of observations in each cell result in a loss of orthogonality. Again modifications to standard analytic methods are required. Snedecor and Cochran (1967, pages 277-279) describe modifications required in a one-way analysis of variance with samples of unequal size.

2.3.5 Repeated measures

When process data were collected, observations were made at two points in time—once before and once after standard setting (Chapter 1, section 1.4.2). When outcome data were collected, observations were made at two points in time for the acute conditions and four points in time for the chronic conditions (Chapter 1, section 1.4.3). In general, different children were sampled before and after standard setting (although some overlap was built in to the two samples). Outcome data for the chronic conditions comprised two sets of matched pairs of observations (one pair corresponding to the year before standard setting, the other to the year after).

If two observations are generated in each cell of the Latin square such that one corresponds to the period before standard setting and the other corresponds to the period after, there are two methods of analysis that can be undertaken. Firstly, it is possible to take advantage of the natural pairing of the two sets of observations and calculate their difference. The exact form of the analysis of these differences will depend upon their underlying statistical distributions. The main advantage of this method is that the analysis produces direct estimates of the size of any changes that have occurred between the two data collection phases.

The second method is to regard the two sets of observations as separate replicates of the Latin square. The difference between the two data collection phases can then be included in the analysis as a fixed effect. An advantage of this method is that it is easier to include covariates in the analysis that differed between the two data collection phases. It is also easier to take into account changes in sample size that occur between the two phases. Both approaches were considered.

2.4 Generalised linear modelling

2.4.1 Model fitting

In order to overcome some of the problems described above, which were inherent in a number of the study data sets, an alternative approach was adopted. To determine which systematic effects were important in explaining the responses, a strategy of model fitting was developed. The statistical package GLIM (GLIM Working Party, 1987) was used to fit sequences of generalised linear models to the data (Aitkin et al, 1989). For binary variables a Binomial error structure was used with a logit link function; for most continuous variables a Normal error distribution with an identity link function was used. For variables in the form of counts a Poisson error structure with a log link function was employed.

2.4.2 Model assessment

GLIM measures the goodness of fit of a model to the actual data by the deviance. This quantity is proportional to twice the difference between the maximum likelihood achievable and that achieved by the model under investigation. By adding systematic components in turn a series of nested models is generated. The improvement made by adding each term can be assessed by the reduction in deviance. When the effects are not orthogonal, the improvement in fit obtained by adding an extra term depends upon those effects already included in the model. Consequently several sequences of nested models must be considered before selecting the one that best represents the data.

For a Normal error structure, the GLIM deviance is identical to the residual sum of squares for the fitted model. When a new term is added, the ratio:

$$\frac{\text{(Change in SS)/(Change in df)}}{\text{(Residual SS)/(Residual df)}}$$

(where df denotes degrees of freedom and the residual sum of squares is that corresponding to the model which includes the new term) is compared with the appropriate F distribution to assess whether the improvement is statistically significant.

For Binomial and Poisson error structures, changes in deviance have an asymptotic chi-squared distribution. When a term is added to a model there is a reduction in both the residual deviance and the residual degrees of freedom. The relative importance of an effect is assessed by comparing the reduction in residual deviance with a χ^2 distribution whose degrees of freedom are equal to the change in residual degrees of freedom which resulted from adding the extra term. Large changes in deviance relative to the χ^2 distribution imply a significant effect of that factor.

In practice there were a number of response variables for which a large residual deviance indicated overdispersion in the data (Cox and Snell, 1989). Overdispersion can arise in a number of ways, one of the most common being clustering (Stigler, 1986, pages 229-238). For some of our outcome variables we could identify the practice which the child attended but not the particular doctor with whom the child usually consulted. While it was possible to allow for variation between practices in the models fitted, it was not always possible to allow for variation between doctors within a practice. But it is likely that responses from parents who consult the same doctor, for example about satisfaction with the consultation, are correlated. McCullagh and Nelder (1989) recommend the adoption of a constant dispersion factor to allow for overdispersion arising from such a mechanism. Under this assumption, if Y is the number of positive responses, the variance of Y is $n\theta(1-\theta)\phi$ where $n\theta(1-\theta)$ is the nominal Binomial variance and ϕ is the dispersion parameter which is independent of sample size.

It is possible to obtain an estimate of ϕ by fitting a full model and dividing the residual scaled deviance by the residual degrees of freedom. Improvement in model fit can be assessed by dividing the reduction in scaled deviance by the dispersion parameter and comparing this with an appropriate χ^2 distribution. In practice, fitting the full model was often computationally very intensive and a procedure proposed by Baker and Nelder (1978), in which the ratio:

$$\frac{\text{(Change in deviance)/(Change in df)}}{\text{(Residual deviance)/(Residual df)}}$$

is compared with an F distribution was employed (this is analogous to the variance ratio test for a Normal distribution described above). The residual deviance is the deviance corresponding to the fuller of the two models (that is the model with the additional term included). This procedure will be referred to as a deviance ratio test.

2.4.3 Weighted least squares

One technique considered for overcoming the lack of homogeneity of variance in some of the response variables was to use weighted least squares. Armitage and Berry (1987, pages 194 to 196) recommend that in a regression analysis of cell means, the weights should be proportional to the reciprocal of their variance. In the case of variables which are normally distributed, the weights are simply the number of observations used to determine the mean in each cell. The weighted least squares procedure is easily undertaken in GLIM by using the weight directive.

2.4.4 Parameterization

It often happens that there is not enough information in the data to determine uniquely all the parameters specified by the model formula. This always happens when more than one categorical independent variable is included in the model or when a single

categorical variable is included with a term representing the grand mean. As an example, consider the one way analysis of variance model in which the value of some outcome variable, Y, for patient i, depends only upon a single factor, A, with J levels. There are several possible formulations of the model. The simplest is

$$E(Y_{ij}) = \mu_j, \qquad j = 1,.....,J.$$

If the expected outcome for patient i is simply the mean value of that outcome for all patients in group j and is equal to μ_j. When the model is fitted, there are J independent parameters to be estimated ($\mu_1, \mu_2, \ldots, \mu_J$).

But the formulation used by most statistical packages is:

$$E(Y_{ij}) = \mu + \alpha_j, \qquad j = 1,.....,J.$$

where μ is some sort of overall effect and α_j is an additional effect due to factor A. For this formulation there are J+1 parameters ($\mu, \alpha_1, \alpha_2, ..., \alpha_J$) but only J of them are linearly independent. The package GLIM invokes corner point parameterization—α_1 is set to 0 so that α_j measures the difference between the first and jth levels of a factor and μ represents the effect of the first level. In other packages, such as SPSS (SPSS, 1990), the α_j's are constrained to sum to zero:

$$\sum_{j=1}^{J} \alpha_j = 0$$

If the design is balanced (that is, there is an equal number of patients in each group) μ will be the grand mean (or average outcome for all patients) and the α_j's represent the difference between the mean outcome for patients in group i and the grand mean. The α_j's are referred to as deviation coefficients.

Both methods of parameterization were used in the analysis. The main use of deviation coefficients has been to help interpret models in which there is significant variation between doctors.

2.4.5 Missing data

Missing data were one of the main factors that led to the adoption of generalised linear modelling in preference to using the classical approach described in Section 2.2. The presence of missing data influenced the analysis in four very specific ways.

Firstly, most of the data sets tended to be less complete after the intervention (in phase 2) than before. Rather than attempt to calculate a change score in the dependent variables (which might not always be possible), it was usually considered appropriate to include observations from each phase separately in the analysis and include differences between phases as a fixed effect within the linear component of the model. This was not exclusively the case—for some variables both approaches were used and the results compared.

Secondly, the level of missing data was not consistent across doctors (in the case of the process data sets) or practices (in the case of the outcome data sets) within each trainer group. In addition, differences between doctors and practices were not consistent over time. It is possible that differences over time may arise because of this effect. (For example if a doctor who tends to refer more patients than average is missing from phase 2, it might appear that the referral rate has dropped.) To allow for these effects variation between doctors (for the process variables) and variation between practices (for the outcome variables) were entered as fixed effects in the model in addition to variation between trainer groups.

On occasions there were systematic reasons why data were omitted. For example the postal outcome survey for some of the conditions was delayed in one of the data collection periods. This meant that differences between surveys would be difficult to interpret as questionnaires had been sent out a different times in the year. The use of covariates to try and allow for such effects was investigated. In this particular example, the use of temperature as a covariate was considered.

Finally, at a more general level, missing data were one of the major causes of imbalance in the data sets which resulted in the confounding of many of the effects of interest. Thus it was necessary to fit the effects in different orders in order to assess their relative importance using the methods described in Section 2.4.2.

2.5. Hypothesis testing

2.5.1. Effects of the interventions

The original premise underlying the analysis was that the intervention would cause doctors to change their behaviour and that this would lead to changes in the outcome of care for their patients. As the analysis progressed, various specific hypotheses were considered.

Initially it was anticipated that all the interventions might affect doctors' behaviour but that the size of the effect would be different for each intervention. The best way to fit this effect when observations from phases 1 and 2 were included separately in the analysis was carefully considered. It was recognised that fitting a simple five level factor corresponding to a treatment effect in the original Latin square design was not optimal as this would produce an 'average effect' across the two phases of the study and there could be no treatment effects in phase 1 prior to the intervention. Fitting an interaction between intervention and phase was also problematic. It is not clear whether

the test of an intervention effect should be based on adding the interaction term with the main effects already included in the model (the test would then be based on a reduction of 4 in the residual degrees of freedom) or on the effect of adding one or both of the main effects together with the interaction (in which case the test of model improvement would be based on a reduction of up to 9 in the residual degrees of freedom). In either case, the final model would include unnecessary parameters representing differences in phase 1 due to the intervention. One of the effects of including redundant terms in the model is to affect the estimates of the other parameters - in particular the standard errors of other parameters tends to increase (Aitkin et al, 1989).

The solution adopted was to create a 'composite' variable 'AUDT'. This took the value 1 for every observation in phase 1 (prior to the intervention) and for observations in phase 2 (after the intervention) corresponding to cases where no intervention was received. For the remaining observations in phase 2, the new variable took a value between 2 and 5 as shown in Table 2.2 depending on the intervention received. Testing the improvement in fit when this variable was added to the model represented a direct test of the hypothesis of interest.

Table 2.2 Values taken by variable 'AUDT' for each intervention in phases 1 and 2

		Occasion	
	Intervention	**Phase 1**	**Phase 2**
a	Set clinical standard	1	5
b	Received clinical standard from another group	1	4
c	Received comparative data for all doctors	1	3
d	Received baseline data for own trainer group only	1	2
e	None	1	1

Clearly this variable is confounded with time. Any change in behaviour due to the interventions would cause a significant 'phase' effect; conversely any change in behaviour due to some underlying trend over time might result in a significant 'intervention' effect. It is important to fit these effects in different orders in order to evaluate their relative importance. If the 'AUDT' term was significant even after allowing for a change over time, this would be regarded as strong evidence of a genuine intervention effect.

A small number of analyses were based on differences between observations made in phase 1 and observations made in phase 2. In these cases fitting an intervention term was straightforward—involving just a single five level factor representing the five levels of the intervention.

2.5.2. Effects of setting clinical standards

During the course of the study it emerged that a number of the interventions were unlikely to have had a large effect on doctors' behaviour. In interviews with trainers (North of England Study 1990c) 70 percent of the doctors reported that receiving a standard set by another group (the second intervention) had not been helpful. Even among the remaining 30 percent, many described the benefits as limited and only one trainer reported a change in practice as a result of receiving a standard.

Similarly the majority of trainers found the comparative and descriptive feedback (the third and fourth interventions) to be of little value. Only 16 percent of them reported actually using it. Furthermore, two years after the intervention, a significant number of trainers had no recollection of ever having received the feedback, perhaps an indication of its lack of impact.

In contrast, most of the trainers felt that setting a standard had been helpful (although just under half reported that they did not think that standard setting had influenced their practice). As a result of these findings, it was felt appropriate to consider the hypothesis that setting clinical standards had caused changes in doctors' behaviour and patient outcome but that the other interventions had had no effect. To test this effect, the variable 'PSTD', which took the value 2 for observations in phase 2 that corresponded to doctors who set clinical standards and 1 for all other observations, was created. On adding this term to the model there was a reduction of only one in the residual degrees of freedom compared with a reduction of four when adding the full intervention term 'AUDT'.

2.5.3. Condition by standard setting interactions

Implicit in the original choice of the Latin square study design was the assumption that any effects due to the interventions would be uniform across all five conditions. (In a Latin square with just one observation per cell, it is not possible to test for interactions between the main effects). During the course of the study, however, it became clear that such an assumption was not appropriate. It was possible that setting a standard for one condition might have a much greater influence on clinical behaviour than setting a standard for another condition. Indeed the proportion of doctors who reported finding standard setting useful varied across conditions. By adopting the approach described, in which the unit of analysis is the consultation (for the process data set) or the patient (for the outcome data sets) it is actually possible to examine whether the interventions had effects specific to each condition. The most efficient way of testing for this effect is to fit the composite variables 'CPST' (Table 2.3) and 'CAUD' (Table 2.4) corresponding to condition specific effects of standard setting and condition specific effects of all interventions respectively.

Table 2.3 Condition specific effects of standard setting: values taken by variable 'CPST' for each condition in phases 1 and 2

Condition	All observations in phase 1	Observations in phase 2 corresponding to doctors who set standards	All other observations in phase 2
Cough	1	2	1
Acute vomit	1	3	1
Bedwetting	1	4	1
Itch rash	1	5	1
Wheezy chest	1	6	1

Table 2.4 Condition specific effects of the intervention: values of term 'CAUD' corresponding to different values of 'AUDT' for each condition

	Value of term 'AUDT' associated with the observation				
Condition	**1**	**2**	**3**	**4**	**5**
Cough	1	2	7	12	17
Acute vomit	1	3	8	13	18
Bedwetting	1	4	9	14	19
Itchy rash	1	5	10	15	20
Wheezy chest	1	6	11	16	21

2.5.4 Mixed group effects

One of the two trainer groups that set a standard for a particular condition met a 'mixed' group that included specialists in that particular area. The purpose of the meeting was to discuss the standard. Afterwards the trainer group was given the opportunity to

revise the standard as they felt appropriate. Thus meeting a mixed group was likely to have an effect on only one of the interventions—the setting of standards. A composite variable 'MIXD' was created that took the value 2 for observations in phase 2 corresponding to trainers who had met a mixed group and set a standard for that condition and the value 1 for all other observations.

Again, as it was possible that recommendations in the standards might differ from condition to condition, consideration was given as to how best to test for condition specific effects of meeting a mixed group. The variable 'CMIX' which took the values set out in Table 2.5 was created.

Table 2.5 Condition specific effects of meeting a mixed group: values taken by the variable CMIX

Condition	Values of CMIX	
	Consultations in phase 2 with a doctor who set a standard for the condition and also met with the respective mixed group	All other consultations
Cough	2	1
Vomit	3	1
Bedwetting	4	1
Itchy rash	5	1
Wheezy chest	6	1

2.6. Modelling strategy

2.6.1. Fitting covariates and non-intervention effects

The first step in each analysis was to consider the association between the dependent variable and each of the independent variables. This preliminary (univariable) analysis

was used to inform the order in which terms were fitted in the generalised linear model. Generally the first terms to be fitted corresponded to rows and columns of the original Latin square design. For most of the dependent variables, variation between study conditions was significant. This is not surprising—one would expect many of the process measures such as referral rates and examination rates and outcome variables such as whether the child was still suffering from the condition to vary from condition to condition. The term 'COND' representing these differences was usually entered into the model first. Then variation between trainer groups (TGRP) and variation between either doctors (DOCT) (for process variables) or practices (PRAC) (for outcome variables) were considered. If these terms were significant they were retained in the model.

The next step was usually to consider any covariates that the preliminary analysis suggested might be influencing the dependent variable. These included variables such as temperature and time delay between consultation and interview. At this stage changes between phases 1 and 2 were also investigated. Finally interactions between significant terms already included in the model were considered.

2.6.2. Fitting intervention effects

There were six possible intervention effects that could be considered corresponding to the six terms described above: PSTD (a uniform standard setting effect); CPST (a condition specific standard setting effect); AUDT (separate effects for each intervention but each intervention has a uniform effect across the five conditions); CAUD (condition specific effects for each intervention;); MIXD (effect of specialist input when setting a standard); and CMIX (a condition specific effect of specialist input when setting a standard). For each dependent variable, it would have been possible to test all six hypotheses by adding each of these five terms to the model but it was felt that this was not acceptable statistically. The problem with multiple testing is that the probability of

obtaining a significant result purely by chance is greatly increased; the type 1 error rate is greatly inflated (Bland, 1995). In this case the problem was exacerbated by having a large number of dependent variables to consider.

The first step in dealing with this problem was to consider which of the six hypotheses should be considered as the most important. It was recognised that all the interventions may have caused the doctors to change their behaviour despite their perception that they had not modified their behaviour as a result of a number of the interventions. But on balance it was felt that setting a clinical standard was a much stronger intervention than the others and was the one likely to produce the largest effect. The following hypothesis testing strategy was devised.

1. Initially there would be a test of a uniform effect of standard setting by fitting the term PSTD. It was recognised that this term was confounded particularly with phase and thus was fitted with and without the prior inclusion of differences between phases in the model. Similarly different sequences of model fitting were considered when the standard setting term was considered to be confounded with other variables of interest.

2. A condition specific effect of standard setting (CPST) would be considered either when a uniform effect of standard setting had been observed or if clinicians involved with project had good reason to believe that effects of standard setting would be condition specific. As a result of this second criterion, condition specific effects were investigated for drug management and all the outcome variables. In the case of drug management clinicians were asked to predict the direction of the effects of standard setting prior to the analysis being undertaken. It was recognised that there may be confounding between a condition specific effect of standard setting and the

main effects of differences between conditions (COND) and differences between phases (PHAS) and the interaction between these two main effects (COND·PHAS). If a potential condition specific effect of standard setting was identified, the significance of the effect was assessed with and without the inclusion of the interaction term COND·PHAS in the model.

3. A mixed group effect was investigated only if a standard setting effect (either uniform or condition specific) had been found. Fitting a condition specific mixed group effect was felt to be appropriate when there was evidence of a condition specific effect of standard setting.

4. Similarly, possible effects of the other interventions were considered only if an effect that could be attributed to standard setting was found first.

2.6.3. Significance levels

In view of the large number of dependent variables considered, a significance level of one percent was used to test for effects of standard setting. When considering whether to include other effects and covariates in the model the criteria were much less stringent. Such effects were usually retained if they were significant at the five percent level and a number of models were considered that included terms, potentially confounded with standard setting, even if the associated significance levels exceeded five percent.

To give an indication of the magnitude of an effect, confidence intervals have been quoted. If an effect which corresponds to a term with one degree of freedom (e.g. the difference between phase 1 and phase 2) is significant at the one percent level, then 99% confidence intervals are quoted. Where there are a number of effects of interest associated with a composite variable, 95% confidence intervals are given. For example, there may be a condition specific effect of standard setting that is significant at the one

percent level but one would not necessarily expect the effect for each individual condition to be significant at that level. In such cases 95% confidence intervals are quoted. Finally, interval estimates are given for the effects of standard setting even when the effect is not significant—these are 95% confidence intervals.

2.6.4. Missing data

The issue of systematically missing data (e.g. the lack of postal outcome questionnaires for bedwetting) has been dealt with above. The case where all data are missing for a complete consultation or child (e.g. when doctors failed to enhance their quota of medical records) has also been addressed. Missing data also arose because responses to postal questionnaires and (to a lesser extent) responses to interview schedules, were sometimes incomplete. Generally the level of missing data was very low. No imputation of missing values was undertaken. If a dependent variable was missing, there was no observation corresponding to that particular child for that particular period of data collection. If one or more of the independent variables was missing the strategy was a little more complex and was designed to facilitate the fitting of models in GLIM (a package which, at the time of the analysis, had very limited facilities to handle missing data). For each dependent variable a preliminary (univariable) analysis was undertaken to examine the relationship between it and each of the potentially important independent variables (Section 2.6.1). If there was evidence of a significant association, the independent variable was selected for inclusion in a multiple regression type analysis. Listwise deletion across all the selected variables was then employed before the generalised linear modelling was undertaken. This meant that, for the outcome data sets, the number of cases in each analysis was not always exactly the same.

For the process variables, the level of partially missing data was almost non-existent. It was generally possible to sort out any queries at the data cleaning and validation stages.

Thus if a medical record was abstracted all the variables usually had valid values. Most of the analyses were based on the same number of cases.

2.7 Variables used in the analysis of process and outcome data

2.7.1 Glossary

In the modelling reported in Chapters 5 to 11, variable names have been abbreviated to a maximum of four letters. The categorical variables (or factors) are listed alphabetically in Table 2.6.

Table 2.6 Variables used in the modelling of process and outcome data sets

Label	Variable	Levels	Value labels
ASTH	Asthma	2	1 = non asthmatic child; 2 = child with asthma
AUDT	Types of medical audit	5	See Table 2.2 (page 28)
CAUD	Condition specific effect of all five interventions	21	See Table 2.4 (page 31)
CHRN	Type of condition	2	1 = acute condition; 2 = chronic condition
COND	Study condition	5	1 = acute cough; 2 = vomiting; 3 = bedwetting; 4 = itchy rash; 5 = recurrent wheezy chest.
CMIX	Condition specific effect of meeting with the mixed group	6	See Table 2.5 (page 32)
CPST	Condition specific effect of standard setting	6	See Table 2.3 (page 31)
DOCT	Study doctors (trainers)	84	
INOC	Initial outcome (interviews and postal questionnaires)	2	Response to initial interview or questionnaire (in subphase A): 1 = failure; 2 = success.
INTV	Interviewer effect	8	Eight different interviewers
MIXD	Mixed group effect	2	Takes the value 2 for observations in phase 2 corresponding to trainers who had met a mixed group and set a standard for that condition; takes the value 1 for all other observations.

Table 2.6 continued: Abbreviations used in reporting the results of the modelling

Label	Variable	Levels	Value labels
PHAS	Phase of study	2	1 = phase 1; 2 = phase 2.
PRAC	Study practice	64	
PSTD	Setting a clinical standard	2	Takes the value 2 for observations in phase 2 corresponding to consultations for a condtion for which the GP set a standard and 1 otherwise.
RCRD	Type of medical record	2	1 = enhanced record; 2 = routine (statutory) record.
STND	Standard setting	2	1 = other doctors; 2 = doctors who set standards (also applies to observations in phase 1).
TGRP	Trainer group	10	1 = A; 2 = B; 3 = C; 4 = D; 5 = E; 6 = F; 7 = G; 8 = H; 9 = J; 10 = K. (see page 6, Table 1.1)
WZAU	Wheeze specific effects of all interventions	5	This term was used to test the hypothesis that the interventions were only effective for one condition - wheezy chest.
WZMX	Wheeze specific mixed group effect	2	1 = other observation; 2 = observation in phase 2 corresponding to a consultation for wheezy chest with a trainer who set a standard and met the corresponding mixed group for that condition.
WZST	Wheeze standard setting effect	2	1 = other observation; 2 = observation in phase 2 corresponding to a consultation for wheezy chest with a trainer who set a standard for that condition.

A number of continuous covariates were also included in the modelling. These included:

- AGE age of the child in years;
- CTMP an effect of temperature on outcome of children with acute cough;
- LLAG log transformation of the time delay between the consultation and administration of the interview or questionnaire;
- TEMP Mean minimum temperature in the month before the postal questionnaire was sent out.

2.7.2 Interpretation of effects

The interpretation of a significant effect depends upon the nature of the dependent variable. In most of the analyses of the process variables, observations from phases 1 and 2 were included separately. In this case, significant variation between conditions, for example, (represented by the variable COND) would indicate that, overall, the value of the dependent variable for each condition was not the same. (If the dependent variable was referral rate, the result would indicate that referral rates were not the same for each condition). In analyses where the dependent variable was the difference between observations in the two phases (for example changes in referral rates) a significant effect of variation between conditions would have a different interpretation—namely that the *change* in referral rates between phases 1 and 2 was not the same for each condition.

Chapter 3

Prevalence survey

3.1 Introduction

As part of the process of identifying children eligible to participate in the study, a postal survey was sent to parents of all children aged ten or under who were registered with a general practice in the Northern Region of England. Recipients were asked to provide, for a named child, information relating to the prevalence of and consultation for five common, symptomatic conditions—acute cough, acute vomiting, bedwetting, itchy rash and recurrent wheezy chest. Following identification of subjects, data concerning the provision of paediatric care was collected from doctors and data relating to the health of the child was collected from a sample of parents. The complete process was carried out both before and after the setting and dissemination of clinical standards was undertaken by selected groups of doctors.

In this chapter, the analysis of the data arising from the prevalence survey is reported. The objectives of the analysis were:

(i) to develop methods and a strategy for analysing all the main data sets compiled during the study;

(ii) to investigate changes in reported prevalence and reported consultation behaviour for the five conditions during the period of the study;

(iii) estimate the effect, if any, of the audit on the reported prevalence and consultation rates of those conditions.

While the main hypothesised changes arising from standard setting and performance review pertained most directly to doctors' behaviour and improvements in child health, it was considered appropriate first to examine prevalence rates and consultation behaviour. Any large changes would be likely to influence subsequent analysis of process and outcome.

It is also conceivable that the intervention itself might have affected either reported prevalence or consultation rates. Some of the clinical standards which were set included educational objectives such as, "to help the child and family understand the natural course of asthma", as well as statements relating to disease management— "[under given circumstances] review the rash in twenty four hours". Any changes in patient education might alter the perception of the parent as to whether or not their child suffered from a particular condition; and changes relating to the discharge of a patient might affect consultation rates.

3.2 Survey methods

The prevalence survey took the form of a questionnaire mailed to the parents of all 76 000 children aged ten or under and registered with the 62 participating practices. Each of the questionnaires was attached to a covering letter from the child's practice. Parents were asked to complete the questionnaire, detach it from the covering letter and return it in the reply-paid envelope provided. The questionnaire consisted of five questions.

In Question 1, parents were asked to identify which doctor their child usually consulted. As the prevalence survey was sent to the parents of all children registered with each

practice, it was possible to make comparisons between trainers (who were involved with standard setting) and their partners (who were not).

The second question asked for the age of the child in years. Where there was more than one child in a family, by comparing reported ages with dates of birth obtained from the practice registers, a check could be made that the correct questionnaire had been completed for each child.

Question 3 asked whether, during the previous four weeks, the child had either of the two acute conditions—cough or vomiting. In each case, if the answer was yes, the parents were then asked if he or she had consulted one of the doctors in the practice about the condition.

Question 4 asked about the three chronic conditions. Parents were asked to identify whether, during the previous twelve months, their child had itchy rash, or bedwetting, or at least two attacks of wheezy chest. For each condition, if the answer was yes, parents were asked whether or not they had consulted one of the doctors working in the practice.

The fifth question asked about whether the child was receiving regular treatment for the three chronic conditions.

3.3 Response rates

Reminders were sent to all non-respondents 16 days after the initial mailing. Of the 76 000 questionnaires posted in each phase, 7% could not be delivered, usually because the patient's address had changed. Of the 70 000 presumed correctly delivered in each phase, 91% were returned completed in the first phase, before standard setting; and

87% in the second phase, after standard setting. These response rates were considered to be satisfactory.

3.4 Estimation of prevalence and consultation rates

The first step in the analysis was to define prevalence and consultation rates. The prevalence rate for a given condition was defined as the number of children reported by their parents as suffering from that condition during the four weeks (for acute conditions) or twelve months (for chronic conditions) before receiving the questionnaire, expressed as a proportion of the total number of children yielding a valid response; the consultation rate was defined as the number of children reported by their parents as consulting their general practitioner for a given condition divided by the number of children reported as suffering from that condition. Two sets of prevalence and consultation rates were calculated for each participating practice - one set for the children whose parents reported that they usually saw a trainer, i.e. one of those participating general practitioners who had set a standard; and the other set for the remaining children in the practice—those who usually saw one of the partners.

Three methods of combining the practice specific rates to obtain a summary statistic corresponding to each cell of the Latin square described in Chapter 2 were considered. First the practice specific rates within a cell were simply averaged. Secondly, the practice specific rates were weighted in proportion to their denominators (which range from 40 to 1200 when calculating prevalence rates). It was felt that the first method might give too much weight to the practices with small denominators (resulting in greater standard errors for the estimated rates); and that the second method would give too much weight to the larger practices (whose parameter estimates are in principle no more important those of smaller practices). A compromise was considered - the practice specific rates were weighted in proportion to the square roots of their denominators.

Preliminary analysis of variance carried out on prevalence data collected during the first phase of the study (before standard setting) gave almost identical results for each method. The third method was then adopted as it was felt to be conceptually the most appropriate.

Since the prevalence survey was conducted both before and after standard setting, these calculations generated four rates for each cell - two prevalence rates and two consultation rates. As the main interest was in determining whether there had been any changes in prevalence and consultation rates, advantage was taken of this natural pairing by analysing the change in these rates between the before and after phases.

3.5 Statistical modelling

The statistical model used in a classical analysis of the Latin square set out in Figure 2.1 is given by equation 2.1. In the case of the prevalence survey, data corresponding to partners who did not take participate directly in the educational intervention, were also collected. For both prevalence and consultation rates, these yielded two further replicates of the Latin square. The full model that can actually be fitted is therefore:

$$\begin{aligned} Y_{ijklm} = {} & \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \omega_m \\ & + \alpha\delta_{il} + \beta\delta_{jl} + \gamma\delta_{kl} + \alpha\omega_{im} + \beta\omega_{jm} + \gamma\omega_{km} + \delta\omega_{lm} \\ & + \alpha\delta\omega_{ilm} + \beta\delta\omega_{jlm} + \gamma\delta\omega_{klm} \\ & + \varepsilon_{ijklm} \end{aligned} \qquad (3.1)$$

for $i = 1, 2, 3, 4$ or 5; $j = 1, 2, 3, 4$ or 5; $k = 1, 2, 3, 4$ or 5; $l = 1$ or 2 and $m = 1$ or 2 where Y_{ijklm} is the change between the two phases in the prevalence or consultation rates in row i, column j, replicate l and doctor type m (receiving treatment k); α_i is the effect of row i (trainer group pair); β_j is the effect of column j (study condition); γ_k is the effect of treatment k (audit type); δ_l is the effect of replicate l (met mixed group or not); ω_m is the effect of doctor type m (trainer or partner); $\alpha\delta_{il}$ to $\delta\omega_{lm}$ are the effects

of the two-way interactions between the main effects defined above; $\alpha\delta\omega_{ilm}$ to $\gamma\delta\omega_{klm}$ are the effects of the three-way interactions between the main effects defined above; and ε_{ijklm} is a Normal independently-distributed random error with Var $(\varepsilon_{ijklm}) = \sigma^2$ for all i, j, k, l and m.

The analysis is very similar to that of a simple Latin square. It consists of partitioning the total sum of squares of the 100 (i.e. 5 x 5 x 2 x 2) cell-specific changes in rates into components for the grand mean (i.e. the overall change between phases), each of the five main effects, each of the ten interaction terms and a residual component for error. The components for the main effects of rows, columns and treatments and all the interactions except that between replicate and doctor type all have four degrees of freedom. The components for the grand mean, the main effects of replicate and doctor type, and the remaining interaction each have one degree of freedom, leaving 48 degrees of freedom for error. The null hypotheses of no main effects or no interaction effects are tested by dividing the corresponding mean square by the error mean square and comparing the resulting variance ratio with the F(x, 48) distribution (where x is 4 or 1 as appropriate).

The adequacy of the model was investigated by plotting the observed residuals, given by:

$$e_{ijklm} = Y_{ijklm} - y_{ijklm} \qquad (3.2)$$

where Y_{ijklm} is the observed change in prevalence or consultation rates; and y_{ijklm} is the fitted change from model (3.1). The Latin square structure of the data means that the observed residuals are partially constrained. Model (3.1) implies that the analysis should estimate row, column and treatment effects for each of four separate Latin

squares l = 1 or 2 and m = 1 or 2). By considering the estimates of these effects as linear combinations of the Y_{ijklm} 's, we can show that:

$$\mathrm{Var}\left(e_{ijklm}\right) = \sigma^2\left(1-\frac{1}{r}\right)\left(1-\frac{2}{2}\right) \qquad (3.3)$$

for an r × r Latin square. In this analysis, therefore, $\mathrm{Var}(e_{ijklm})$ is equal to $12\sigma^2/25$ for all possible values of i, j, k, l and m.

The parameter, σ^2, can be estimated by the residual mean square s^2 obtained after fitting the model. By dividing each observed residual by $\sqrt{0.48}$ we obtain standardised residuals that follow a Standard Normal Distribution if the assumptions of the model are correct. We can then test these standardised residuals for significant outliers using the method given by Cook and Prescott (1981). It can be shown that the upper bound for the probability that the magnitude of the largest residual exceeds a value t_{max} is given by:

$$\alpha = n\,\Pr\left[F > d^2(n-p-1)/(1-d^2)\right] \qquad (3.4)$$

where n is the number of residuals, p is the number of parameters being estimated, $d = t_{max}/\sqrt{(n-p)}$, and F is a random variable that follows an F distribution with 1 and (n-p-1) degrees of freedom. Thus for the model defined by equation (3.1):

$$\alpha = 100\,\Pr\left[F > 47d^2/(1-d^2)\right] \qquad (3.5)$$

3.6 Changes in prevalence rates

The results of the initial analysis of changes in reported prevalence rates are given in Table 3.1 . In this analysis the within cell differences in reported prevalence rates are modelled directly. There was a fall, significant at the 0.1 percent level, in prevalence rates between phases 1 and 2 as indicated by the large mean square for the grand mean. The variations in changes among pairs of trainer groups (rows of the Latin square in Figure 2.1) were significant at the 5% level but when the sums of squares for sources A,

D and A x D were combined to yield the sum of squares among all ten trainer groups, this was not significant.

The variation in changes among the five study conditions was significant at the 1% level. Since we are looking at changes in prevalence rates between the two phases of data collection, this implies that the fall in prevalence noted above was not consistent for all five conditions.

Table 3.1 Latin square analysis of changes in reported prevalence rates

	Source of variation	Degrees of freedom	Mean square	F	Probability
G	Grand mean (general change between phases 1 and 2.	1	67.5	31.5	< 0.001
A	Among pairs of trainer groups	4	5.6	2.61	0.05
B	Among different study conditions	4	8.4	3.91	0.01
C	Among different types of medical audit	4	2.8	1.31	0.28
D	Between trainer groups that met mixed groups and those that did not	1	0.3	0.13	0.72
E	Between trainers and their partners	1	1.0	0.43	0.52
	Interactions (A, B, C × D; A, B, C, D × E; A, B, C × D × E)	37	2.4	1.13	0.34
	Residual error	48	2.1		
	Total	100	3.3		

The variation in changes among different types of medical audit was not significant. This indicates that there was no evidence that either standard setting or receiving data had any effect on the prevalence of the study conditions as perceived by parents. There was also no difference in prevalence between children who usually saw trainers and those who usually saw their partners. None of the interaction terms was significant; these terms have been pooled in Table 3.1.

Figure 3.1 **Prevalence rates, initial analysis: plot of standardised residuals against expected normal scores**

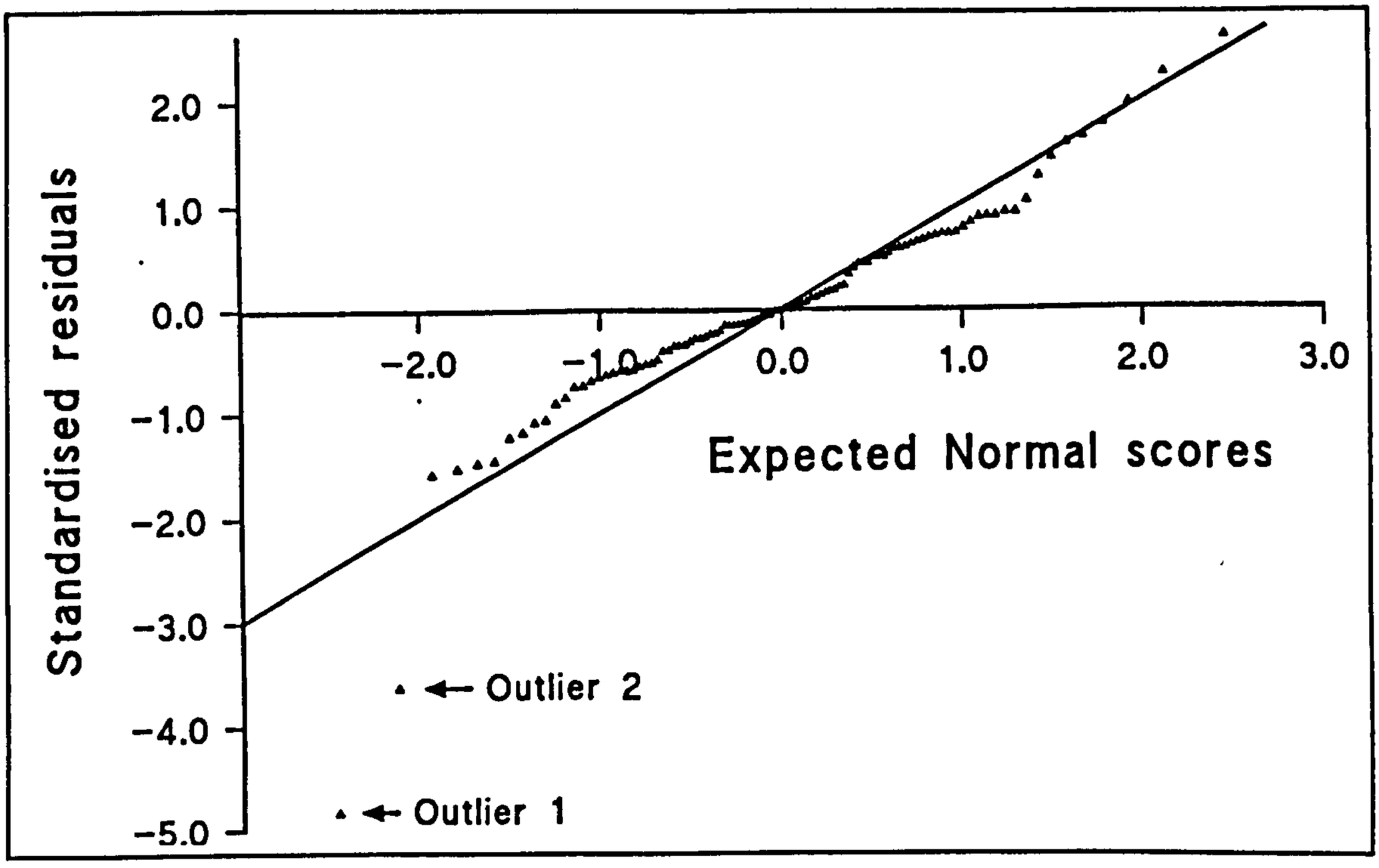
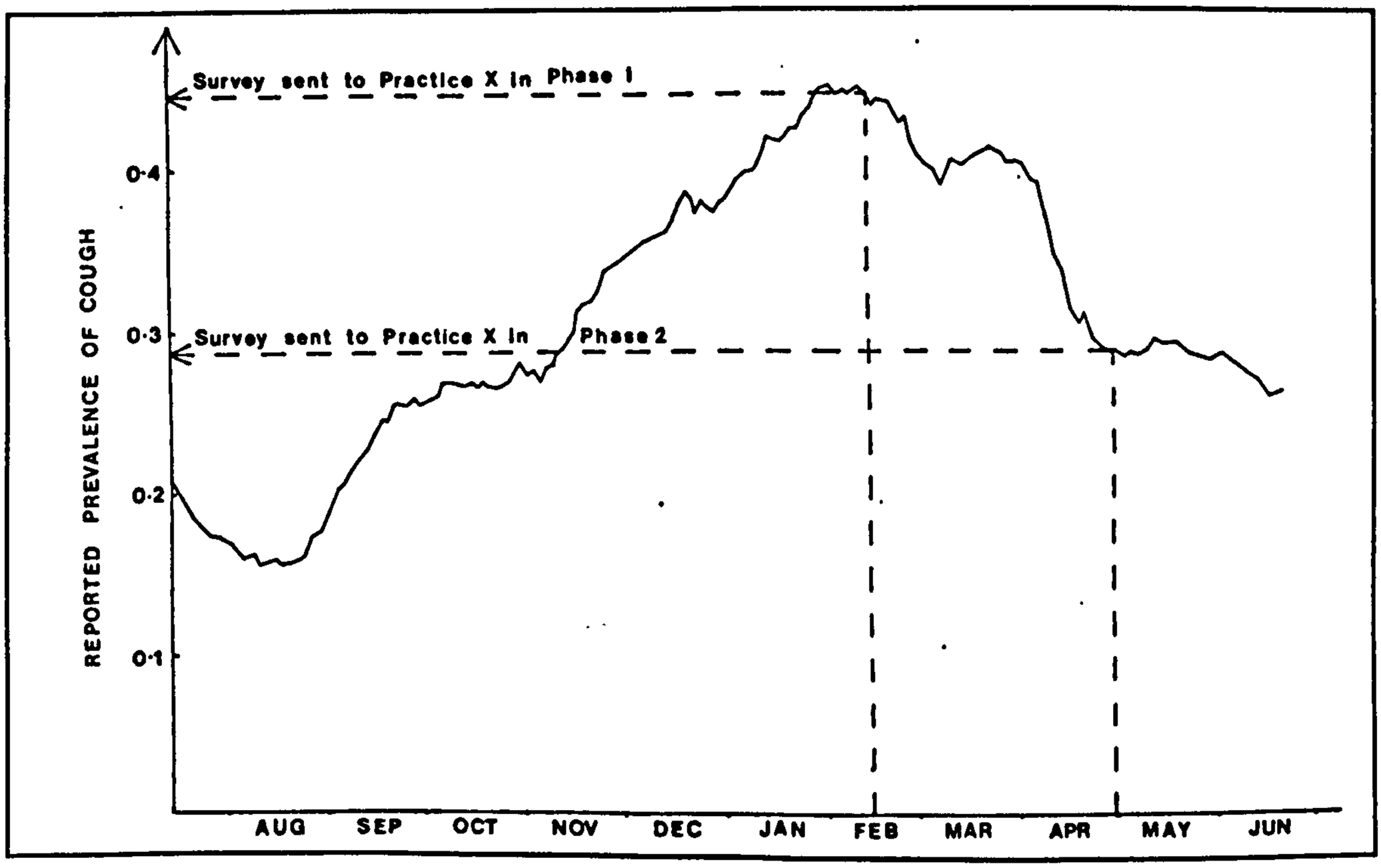


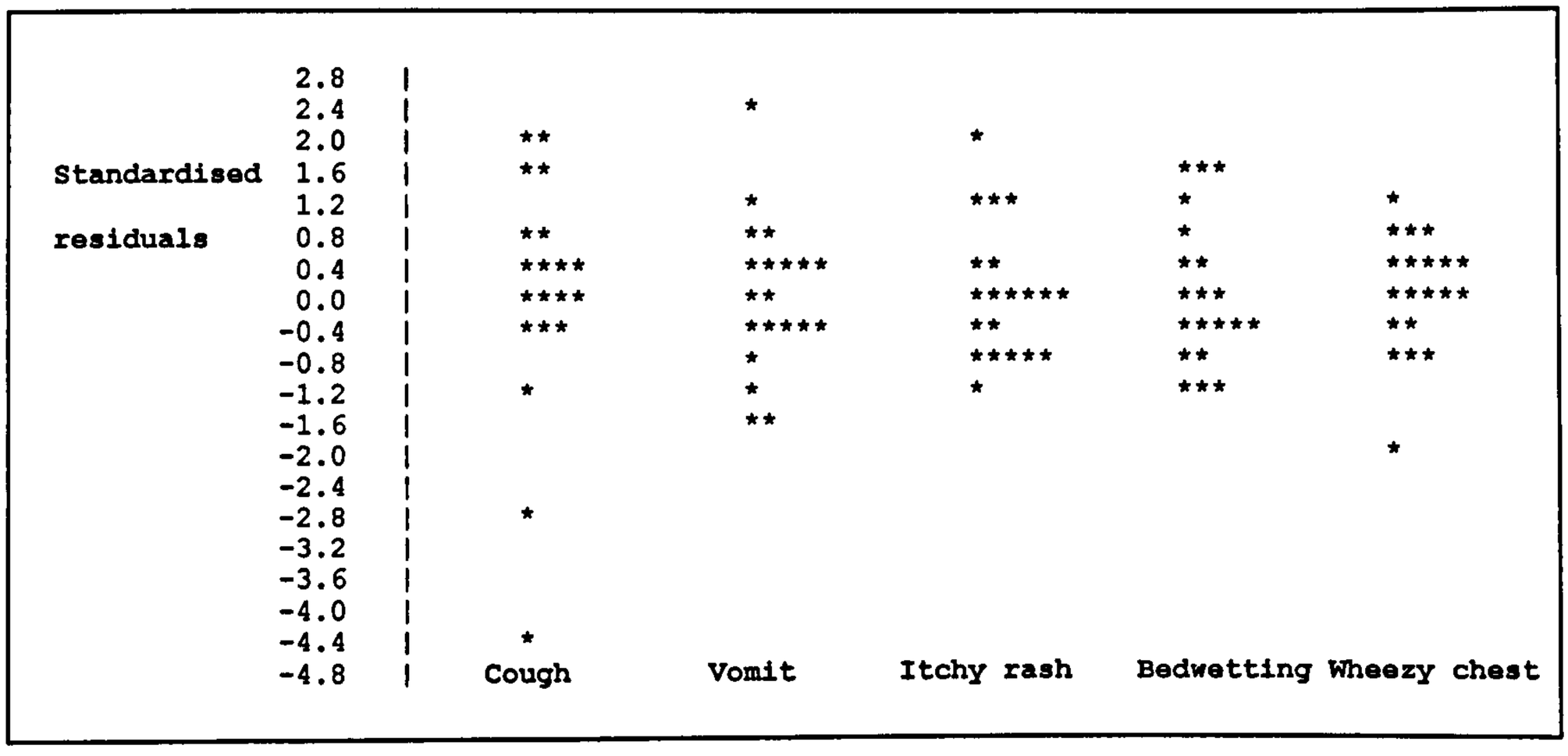
Figure 3.2 **Annual variation in reported prevalence of cough**



Examination of residuals was used to check the appropriateness of the final model. In the plot of ordered standardised residuals against expected Normal scores (Figure 3.1), most of the residuals lie very close to a straight line, thus indicating a close fit to a Normal distribution of errors. However, there are two obvious outliers. Both corresponded to changes in cough prevalence rates for one particular trainer group—one for children consulting the trainers themselves, the other for children consulting their partners. A breakdown of the data by the individual practices within that group shows that cough prevalence for one specific practice more than halved in the second phase of data collection. This was caused by failure of the practice computer in the second phase, resulting in a delay in providing a list of children registered with the practice. Although the postal survey was sent out at the beginning of February in the first phase, it was delayed until the end of April in the second phase - a time lag of nearly three months. Figure 3.2, derived from the entire postal survey and showing cough prevalence through the year, explains the anomaly: there is a peak at the beginning of February and a trough at the end of April.

Further examination of the distribution of residuals by condition (Figure 3.3) indicates a degree of heteroscedasticity within the data. The variability within the cough residuals was much greater than those corresponding to bedwetting for example. The assumption that the errors in the model defined by equation (3.1) have equal variance is not justified.

Figure 3.3 Prevalence rates: distribution of residuals by condition



To eliminate the two outliers, the prevalence rates for cough were recalculated by excluding data from the atypical practice. To overcome the lack of homogeneity of variance across conditions an alternative model was considered. Changes in $\log_e$(prevalence rates) were used rather than simple differences. The Latin square remains complete and orthogonal after these adjustments.

The results of the revised analysis are shown in Table 3.2. The non significant main effects (apart from differences between pairs of trainer groups) and interaction terms

Table 3.2 Analysis of differences in log prevalence rates

	Source of variation	Degrees of freedom	Mean square	F	Probability
G	Grand mean (general change between phases 1 and 2.	1	0.47	46.1	< 0.001
A	Among pairs of trainer groups	4	0.012	1.19	0.32
B	Among different study conditions	4	0.081	7.98	< 0.001
	Residual error	91	0.010		
	Total	100	0.018		

have been pooled with the residual error.

As before there was a general reduction in prevalence between phases 1 and 2. Variation between conditions is now highly significant but the variation between the five pairs of trainer groups is no longer significant. The final choice of model is one in which changes in prevalence (or log [prevalence]) can be explained in terms of an overall change in prevalence which is not consistent across conditions.

Changes between phase 1 and phase 2 in the prevalence of each condition are given in Table 3.3. There was a significant fall in the prevalence of acute cough, bedwetting and recurrent wheezy chest (a feature of the very large sample size is that quite modest changes in prevalence are statistically significant). Changes in the prevalence of the other two conditions were not significant. One might expect changing meteorological factors to affect the prevalence of the two respiratory conditions but it is not clear why there should be a general fall in the prevalence of bedwetting during the period of the study. (The number of children born each year is not constant and it was felt that the observed data would be consistent with a greater proportion of older children in phase 2. However the average age of children was almost identical in both phases of the study.)

Table 3.3 Changes in reported prevalence rates between phases 1 and 2

Study condition	Initial prevalence rate (%)	Final prevalence rate (%)	Significance level of change
Acute cough	31.5	30.9	0.02
Acute vomiting	9.4	9.6	0.22
Bedwetting	8.6	8.1	<0.001
Itchy rash	14.2	14.4	0.31
Wheezy chest	12.6	12.0	0.006

With the exception of a possible outlier the standardised residuals now show a close fit to a Standard Normal Distribution (Figures 3.4) and the distribution of residuals across conditions is possibly more consistent (Figures 3.5).

Figure 3.4 Log$_e$ [prevalence rates]: ordered plot of standardised residuals

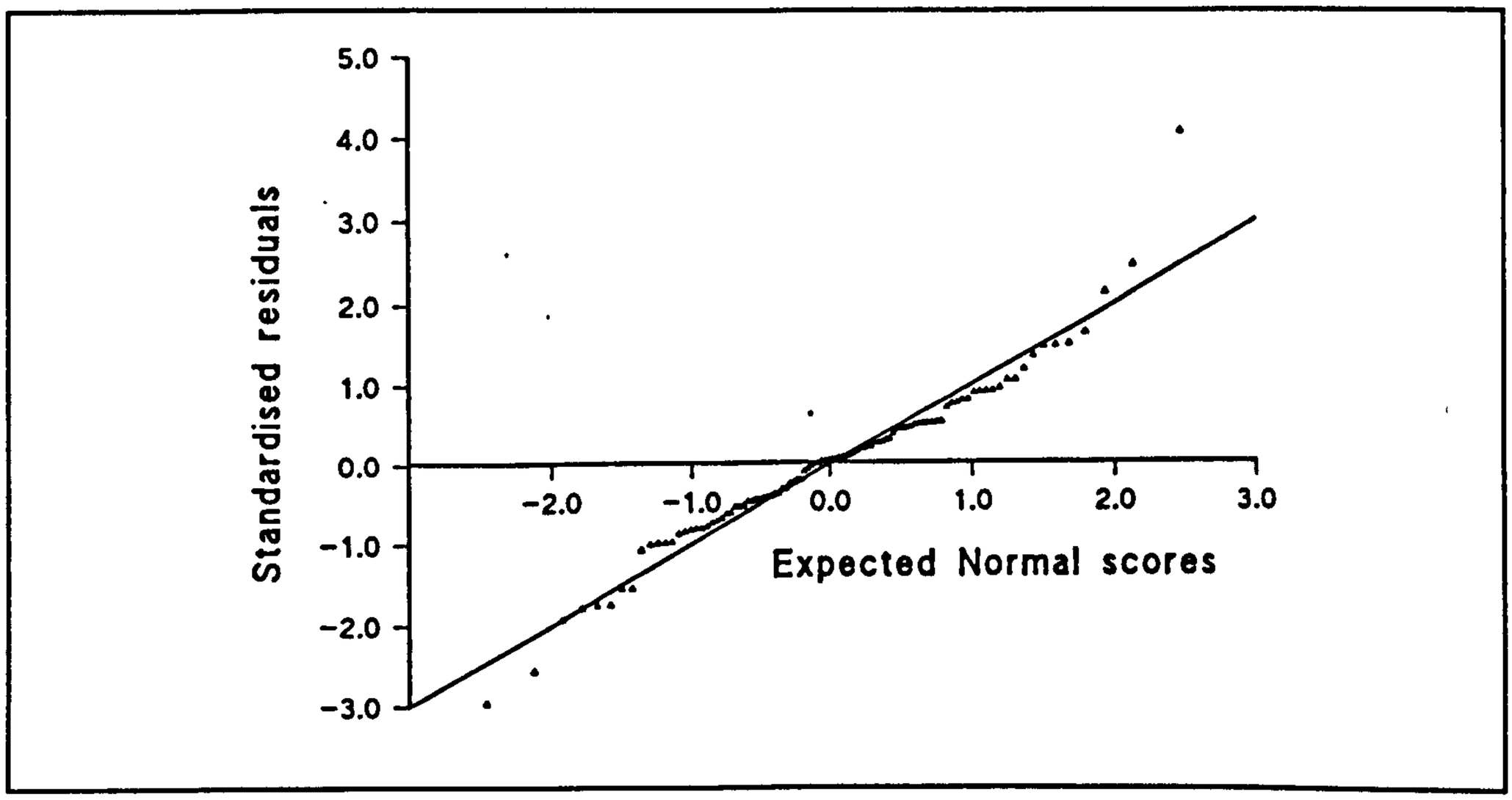
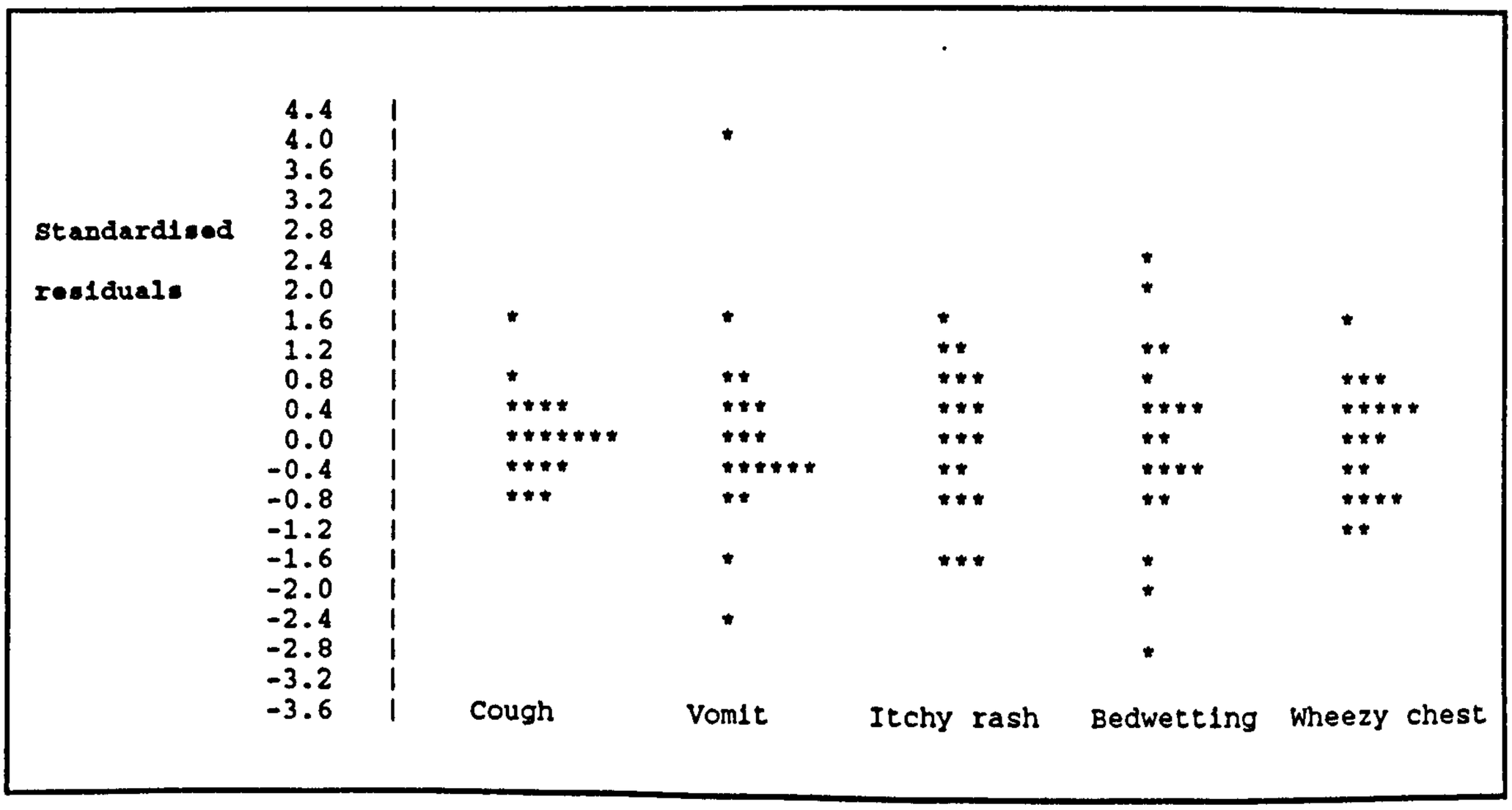


Figure 3.5 Distribution of standardised residuals for differences in log prevalence rates by study condition



The maximum absolute studentised residual is 4.06 and using equation (3.5) the upper bound for the probability that this value is exceeded by chance is less than 10^{-4}. The residual corresponds to the change in prevalence of acute vomiting for a particular set of partners. So far, it has not been possible to account for the marked increase in prevalence among children seen by those doctors.

3.7 Changes in consultation rates

Results of the analysis of changes in reported consultation rates are given in Table 3.4. Trainers and their partners are again considered separately and the interaction terms (none of which were significant) have been pooled. In contrast to the findings about reported prevalence, the small mean square for the grand mean shows that there was no significant changes between phases. Variations in changes in consultation rates among pairs of trainer groups are significant at the 1% level, but there is no significant

Table 3.4 Latin square analysis of changes in reported consultation rates

	Source of variation	Degrees of freedom	Mean square	F	Probability
G	Grand mean (general change between phases 1 and 2.	1	1.7	0.12	0.73
A	Among pairs of trainer groups	4	72	4.87	0.002
B	Among different study conditions	4	13	0.85	0.50
C	Among different types of medical audit	4	18	1.23	0.31
D	Between trainer groups that met mixed groups and those that did not	1	6.4	0.44	0.51
E	Between trainers and their partners	1	111	7.54	0.008
	Interactions (A, B, C × D; A, B, C, D × E; A, B, C × D × E)	37	13	0.85	0.69
	Residual error	48	15		
	Total	100	17.3		

difference in changes between those trainer groups who met mixed groups and those who did not. When we combine the sums of squares from sources A, D and A x D, we find that the sum of squares among all ten trainer groups (denoted in GLIM by GROUPS) is significant at the 1% level. In contrast, there is no variation among different conditions or different levels of feedback, indicating that standard setting had no effect on consultation rates.

However, there was a significant difference in changes between trainers and their partners (denoted in GLIM by the term PARTNERS). Consultation rates for partners fell significantly between phase 1 and phase 2 by 1.2 percentage points; those for trainers rose, although not significantly, by 0.9 percentage points. The difference in these figures is 2.11 with 99% confidence interval [0.17 to 4.04]. It has not been possible to identify the reasons for this differential change. So far it has been possible to show that this change is not an artefact of the movement of patients between doctors or of our collection of additional data from samples of children who usually saw trainers rather than their partners. The observed difference cannot be attributed to changes brought about by the study.

Fitting just the two significant effects (GROUPS, PARTNERS) and examining the model parameter estimates revealed that the estimate for one trainer group (Group Y) was much larger in magnitude than those for the other nine groups. For Group Y, consultation rates had fallen by 4.6 percentage points; for the other nine groups consultation rates had increased very slightly by 0.4%. The difference between Group Y and the others was 4.95% with 99% confidence interval from to 1.72% to 8.17%. A contrast representing this difference (denoted as GROUP_Y) was created. The terms of interest were then fitted sequentially (Table 3.5). Variation between partners was included first; the improvement in fit being significant at the 1% level.

Table 3.5 Stepwise analysis of changes in reported consultation rates

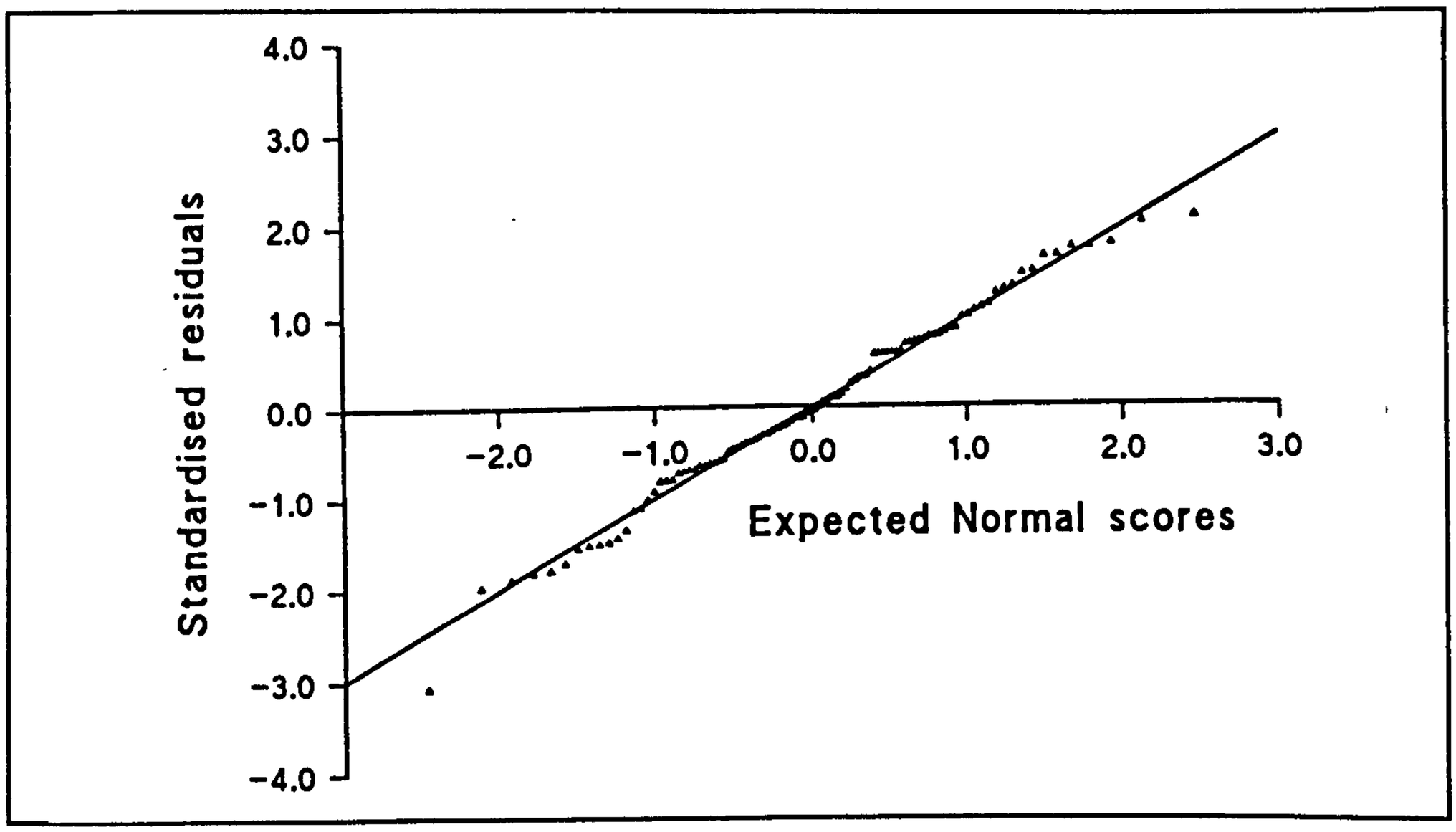
GLIM model	Residual degrees of freedom	Residual sum of squares	Mean sum of squares for extra term	Significance level of extra term
Grand mean	99	1701		
GM + PARTNERS	98	1590	111	0.010
GM + PARTNERS + GROUP_Y	97	1370	220	< 0.001
GM + PARTNERS + GROUP_Y + GROUPS	89	1212	20	0.19

The improvement on fitting the next term, GROUP_Y, was also highly significant but when variation between the remaining nine groups was added, the fit of the model improved very little. The implication of this is that variation in changes among trainer groups is mainly due to the difference in changes between one particular trainer group (Group Y) and the rest. Group Y comprised practices in the environs of the nuclear fuel reprocessing plant at Sellafield, West Cumbria. As noted above, consultation rates were lower (by approximately 5%) in the second phase of data collection than in the first.

These very large changes in consultation rates are almost certainly explained by a period of intense concern among parents of young children in West Cumbria caused by a television programme late in 1983 (Black, 1984). This programme highlighted an apparent increase in the incidence of leukaemia in children living near Sellafield. The height of the resulting controversy coincided with our first phase of data collection late in 1984. It is not surprising that parents were more prone to take their children to the doctor then than during the second phase of data collection two years later, when the controversy had subsided.

The plot of standardised residuals against expected Normal score (Figure 3.6) shows a close fit to the a straight line; the assumption that the residuals are Normally distributed is a reasonable one to make.

Figure 3.6 Reported consultations rates: plot of standardised residuals



3.8 Discussion

Both the prevalence and consultation rate data sets were complete. Data were available for every practice in the study for each clinical condition. These data were combined to produce summary statistics corresponding to the cells of the Latin square set out in Chapter 1. The model defined by equation 3.1 was then fitted. The main findings were a general reduction in reported prevalence of the study conditions between the two surveys and a difference in consultation behaviour for patients attending trainers and their partners. Reasons for these differences are not clear but the analysis suggested that they are not an effect of either standard setting or the other methods of medical audit.

The design of the study and method of analysis was powerful enough to permit the detection of two fairly obscure changes. First, an apparent change in the prevalence of acute cough within just one of the 64 participating practices was detected. This turned out to be the spurious by-product of the failure of the practice computer. Secondly, a change in children's consultation patterns within five other practices—all located near

Sellafield, a nuclear fuel reprocessing plant in West Cumbria—was identified. This change probably resulted from public concern arising from the apparent increase in the incidence of leukaemia near the plant.

No effects of standard setting or medical audit on perceived prevalence or reported consultation rates were noted. It was felt that any such changes must be small in comparison with the two incidental changes noted above. In addition, the results were consistent with expectations (no changes due to the intervention had really been expected) and were accepted.

At the time when this analysis had been completed, it was felt that most of the objectives of the analysis set out in section 3.1 had been achieved. Changes in reported prevalence and reported consultation rates had been extensively investigated and the two relatively obscure effects described above had been identified. Further analysis of these data sets concentrated on attempting to explain the significant main effects described earlier and investigating the effect of omitting one or more columns from the data sets—simulating the situation in which data were not available for one or more of the conditions. On the basis of these investigations, it was felt that the method of analysis employed could usefully serve as a blueprint for the analysis of the process and outcome data sets (one of the key objectives of the analysis). In retrospect this assessment was probably overly optimistic. Indeed there were a number of hypotheses of interest that were not properly tested during the analysis reported in this chapter. These issues will be addressed in more depth in the final discussion chapter.

Chapter 4

Process of care 1: the sample and content of consultations

4.1 Identification of cases

Children with acute cough and vomiting were identified prospectively by trainers and participating partners during a six week period. Children with the three chronic conditions were either retrospectively identified through the prevalence survey or prospectively by study doctors during the six week period mentioned above. The medical records of children thus identified had enhancement flags inserted to alert study doctors. In each phase of the study, study doctors were requested to complete an enhancement form (Figure 4.1) whenever, during a designated period, an identified child consulted for one of the study conditions. Since the acute conditions typically result in one or two consultations over a short period of time, the designated period of enhancement for these conditions was the duration of the episode. The chronic conditions, on the other hand can give rise to consultations over many months; thus the period of enhancement for these conditions was twelve months.

In the six week period of prospective identification, trainers were asked to identify at least ten and most 20 children suffering from the each of the acute conditions and to leave the records of one in five of these children unenhanced. Participating partners were also asked to identify children, the exact number depending on the number of

Figure 4.1a Enhancement form (front)

NORTHERN REGIONAL STUDY OF
STANDARDS AND PERFORMANCE IN GENERAL PRACTICE

ENHANCEMENT FORM

Child's name ..... Sex .....

Date of birth ..... Doctor .....

Place, time day and date
of consultation .....

Diagnosis or formulation of problem: *If diagnosis provisional state most likely and other alternatives being considered*

.....

.....

.....

Information on which this diagnosis or formulation is based:

(1) History .....

.....

.....

.....

.....

.....

(2) Examination and investigations .....

.....

.....

.....

.....

partners; again the records of one in five were to remain unenhanced. For the chronic conditions a maximum of sixteen records per condition per doctor unit (where a doctor unit is either an individual trainer or all the participating partners within a practice) was flagged for enhancement.

Figure 4.1b Enhancement form (back)

Management decisions

.....

.....

.....

.....

.....

.....

.....

.....

.....

Reasons for management decisions

.....

.....

.....

.....

.....

.....

.....

.....

.....

4.2 Sampling of identified children for abstraction

It was thought desirable that doctors should not know exactly which records would be analysed. Medical records were therefore abstracted for only a sample of the children identified. It was originally intended to sample ten children per study condition per doctor unit (eight of whom were to have had their records enhanced) but the target for conditions for which trainer(s) in the practice had not set a standard was reduced to five children per doctor unit (four of whom were to have had their records enhanced). The

rational for this is given in a paper published shortly after the end of the study (North of England Study, 1992a):

> *"To calculate the sample size needed we judged that a general improvement of 10% in compliance with a standard would be clinically significant. From a pilot study we estimated that we should therefore abstract the records of 10 children per condition per phase for each trainer and 10 records from his or her partners taken together. However interviews with trainers suggested that standard setting would be much more effective in stimulating change than the other three types of audit. We therefore reduced our abstraction targets for each condition for which the trainer had not set a standard to five records per phase. This maintained the power of the study to detect changes arising from standard setting while reducing its power to detect other changes."*

From this it is unclear why any reduction in the targets was necessary but the reason may have been related to the availability of resources. The net effect of this sampling strategy was to diminish the power of the study to detect changes due to the other interventions. This reinforces the decision described in Chapter 2 to investigate the effects of standard setting first before considering the effects of the other interventions.

Finally, for each child, only details relating to the first consultation were abstracted.

4.3 Selection of records for analysis

The analysis reported in this thesis was restricted to the 84 participating general practitioner trainers who completed the study. Although both trainers and their partners were asked to enhance medical records, the analyses presented in this thesis are based solely on consultations with trainers. There are two reasons for this. Firstly it is possible that partners were less committed to the study (they did not have the incentive of being involved in the small group work) and many did not complete enhancement forms and, secondly, there was a much greater turnover of partners during the period of the study

(many changed jobs or retired or joined the practice after the study started) making it very difficult to interpret any changes that might be observed.

Field-workers had abstracted and coded information corresponding to 3466 initial consultations in an episode of care (North of England Study 1990b). Nine hundred and six of these consultations were with children had suffered from acute cough, 680 with children with acute vomiting, 341 with children with bedwetting, 730 with children with itchy rash and 809 with children with recurrent wheezy chest (Table 4.1).

Table 4.1 Number and type of records analysed by study condition and phase

Study condition	Phase 1			Phase 2			Total
	Enhance-ment forms	Statutory records	Total in Phase 1	Enhance-ment forms	Statutory records	Total in Phase 2	
Acute cough	389	82	471	359	76	438	906
Acute vomiting	332	65	397	246	37	284	680
Bedwetting	163	34	197	54	90	144	341
Itchy rash	341	80	421	186	123	310	730
Wheezy chest	357	71	428	287	94	380	809
Total	1582	332	1914	1132	420	1552	3466

At the end of each surgery doctors had been expected to enhance the medical records of children that they had identified, by recording, on the specially designed enhancement forms (Section 4.1), additional information relevant to the case but not usually entered in the statutory medical record. To check whether this process of enhancement itself generated any change in performance, the medical record of every fifth child was to have been left unenhanced. Trainers achieved the target of enhancing four out of five records in phase 1 but, for two of the conditions there was an appreciable shortfall in phase 2. Trainers enhanced the records of only 60 percent of the children that they had identified as having itchy rash and only 38 percent of the children with bedwetting. Furthermore

the number of children identified in phase 2 was much less than the number identified in phase 1. These changes have been attributed to diminishing enthusiasm of the participants who had been asked to enhance for two full years in all—an onerous task (North of England Study, 1990c).

A more detailed breakdown of the number of enhanced records corresponding to the cells of the Latin square (depicted in Figure 2.1) is given in Table 4.2. The number of doctors in each trainer group is given in the second column. For any given doctor, field-workers had sampled proportionally more children for the condition for which that doctor had set a standard. If sufficient cases had been identified and had the records been enhanced, field-workers should have sampled and abstracted (for each doctor) eight records for the condition for which the doctor set a standard and four records for each of the other four conditions.

Using this information in conjunction with the data in Table 4.2 we can assess the extent to which doctors in individual trainer groups were successful in identifying cases and enhancing medical records. The shortfall of cases in phase 2, which is particularly noticeable for bedwetting, is observed across all ten trainer groups.

The main implication of the very different numbers in each of the cells in Table 4.2 is that the design is unbalanced. The main effects are confounded. In a standard analysis of variance type analysis, it is not possible to determine uniquely how much of the variability in the data is due to differences between conditions, how much is due to variation between doctors and how much is due to the educational interventions. This issue needed to be addressed during the analysis.

Table 4.2 Number of enhanced records by trainer group, study condition, phase and replicate of Latin square

Trainer group	Number of doctors	Study condition									
		Acute cough		Acute vomiting		Bedwetting		Itchy rash		Wheezy chest	
		Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
Doctors who met mixed group											
A	9	36	37	35	26	18	5	66*	18*	36	29
C	9	36	36	51*	42*	12	3	32	13	32	29
E	9	36	34	32	25	22	9	35	15	64*	46*
G	6	48*	40*	21	18	21	7	20	17	24	18
J	9	36	30	31	19	23*	11*	30	22	34	27
Doctors who did not meet mixed group											
B	12	82*	73*	43	26	17	3	38	26	38	35
D	8	28	28	26	25	14*	1*	21	18	29	23
F	9	35	32	33	24	20	8	55*	28*	35	26
H	7	28	25	40*	25*	8	3	21	17	28	23
K	6	24	24	20	16	8	4	23	12	37*	31*

Notes:

* An asterisk indicates that the doctors in this row set a clinical standard for the condition in this column.

4.4 Comparison of enhancement forms and statutory records.

In general, enhancement forms contained more items that could be allocated a code than did statutory records—a median of 20 elements compared with a median of six (Table 4.3)

Table 4.3 Content of abstracted records by record type

Category of information*	Enhancement forms			Statutory records		
	Median number of items of information	Number (%) of records with no information (n=2714)		Median number of items of information	Number (%) of records with no information (n=752)	
Diagnosis	1	58	(2·1)	1	204	(27·1)
History	7	41	(1·5)	2	170	(22·6)
Examinations and investigations	4	67	(2·5)	2	232	(30·9)
Management decisions	3	18	(0·7)	1	86	(11·4)
Reasons for management	3	155	(5·7)	0	671	(89·2)
Entire record	20	0	(0·0)	6	0	(0·0)

* Categories correspond to the five sections of the form used by doctors to enhance the medical records

As a result of being asked to complete a structured form (Figures 4.1a and 4.1b—which asked explicitly for information on each of the five areas specified in Table 4.3), doctors have recorded more information about the consultation than they would have done normally. The difference was least for the recording of management decisions—which included prescribed and advised drug therapy—but greatest for the recording of the reasons for taking those decisions. It may be that doctors are not in the habit of justifying in writing their management plans; 89 percent of statutory records contained no explicit reasons for management.

4.5 Content of consultations

The content of consultations was coded and analysed under 16 headings (Table 4.4). Six related to history: social history, family and genetic history, previous medical history,

Table 4.4 Percentage of abstracted records containing specific information by record type

Type of information	Type of record	
	Enhancement forms (n = 2714)	Statutory records (n = 752)
History		
Social history	31·3	6·1
Family and genetic history	24·2	5·1
Previous medical history	47·1	10·2
Previous diagnoses	9·5	1·9
Previous non-drug management	10·7	2·9
Previous drug management	28·7	8·4
Diagnosis of current episode		
History of presenting illness	95·0	73·0
Examination findings	95·7	62·5
Investigations	16·3	14·2
Record of diagnosis	97·9	72·9
Management of episode		
Advice, information and explanation	59·1	17·4
Doctor actions	4·8	2·3
Drug management	85·2	76·1
Follow-up decisions	51·0	13·2
Referral decisions	6·6	4·8
Reasons for management	94·3	10·8

previous diagnoses, previous non-drug management, and previous drug management; four to the diagnosis of current episode: history, examinations, investigations and recorded diagnosis; and the remaining six to the management of that episode: advice and explanation, other doctor actions, drug management, follow up decisions, referral decisions and reasons for management.

In general, there were a number of choices regarding the way in which each of these variables could be analysed. For all variables, one possibility was to look at whether or not any items were recorded in the medical records under that particular heading. Another alternative was to analyse the number of items recorded under that heading. To some extent the choice of analysis was guided by the expectations of those members of the project team involved in the design and implementation of the educational intervention. There was a general expectation that taking part in the standard setting process would encourage doctors to record more items of information. Another expectation was that the information would be more detailed. For example, it was felt that standard setting might lead to greater diagnostic precision—doctors might use the precise term asthma rather than the more vague 'wheezy chest'. For other variables, expectations varied from condition to condition. When considering changes to drug management, it was necessary to refer to the clinical standards set for each condition. For some conditions the prescribing of antibiotics was appropriate in certain circumstances; standards for other conditions suggested that the level of prescribing of antibiotics should fall.

Another issue in the analysis of the effects of the educational interventions was the extent to which information from statutory records could be used. As noted above, more information was recorded on enhancement forms than in statutory records. This might indicate that information from statutory records is unreliable—just because results of an examination were not recorded does not mean that an examination was not carried out. However in view of the shortfall in the enhancement of medical records described above, it was felt worthwhile to investigate whether the additional information could be used.

The analysis of each aspect of the content of consultation is described in detail in the following four chapters. However, two findings were common to nearly all of the

analyses; content of consultation varied from condition to condition and there was wide variation in behaviour from doctor to doctor.

Table 4.5 Percentage of enhancement forms containing specific information by study condition

Type of information	Study condition				
	Acute cough (n = 748)	Acute vomiting (n = 578)	Bedwetting (n = 217)	Itchy rash (n = 527)	Wheezy chest (n = 644)
Social history	31·4	38·2	47·0	24·9	25·2
Family and genetic history	24·2	20·6	28·6	24·7	25·3
Previous medical history	35·6	27·9	56·2	52·9	70·3
Previous diagnoses	8·2	8·3	6·9	7·6	14·8
Previous non-drug management	6·1	10·6	40·6	9·1	7·3
Previous drug management	19·1	12·8	27·2	32·3	52·0
History of presenting illness	99·9	99·8	89·4	89·0	91·6
Examination findings	99·7	99·5	73·3	94·9	95·7
Investigations	6·0	7·1	73·3	6·3	25·6
Diagnosis	98·0	98·6	99·5	98·1	96·3
Advice, information and explanation	63·9	86·3	71·4	46·1	35·6
Doctor actions	2·7	1·9	9·2	6·5	7·0
Drug management	87·6	69·0	63·1	94·1	97·2
Follow-up decisions	47·6	60·7	51·2	37·0	57·8
Referral decisions	4·3	6·4	22·1	6·3	4·5
Reasons for management	97·1	97·6	86·6	89·4	94·7

4.6 Variation between study conditions

In Table 4.5, the proportion of consultations for which a particular piece of information was recorded on an enhancement form, is broken down by study condition. The large differences between conditions can generally be explained by considering the particular

nature of each condition and is discussed in the relevant sections of the following chapters.

4.7 Variation between doctors

There was considerable variation between doctors in the frequency with which particular aspects of care were recorded. This was true across all study conditions and for both enhancement forms and statutory records. It cannot be established, from data collected during this study, how much of this observed variation is due to differences in recording style and how much to real differences in the care provided. As discussed earlier, just because an activity is not recorded on the medical record does not necessarily mean that it was not done.

It was important to take into account both sources of variation (between study conditions and between doctors) when investigating the effects of standard setting.

4.8 Presentation of analysis

Data collected can be conveniently categorised into three areas (Table 4.4): the recording of histories; the diagnosis of the current episode; and the management of the episode. The analysis of the recording of histories is reported in Chapter 5; the analysis of the diagnosis of the current episode is reported in Chapter 6. Because there was a considerable amount of information relating to drug management, it was natural to divide the analysis of the management of the current episode into two sections; analysis of non-drug management is reported in Chapter 7, the analysis of drug management is reported in Chapter 8.

In each section, the choice of variables to be analysed was made by the research team; they closely relate to the categories of information listed in Table 4.5.

Chapter 5

Process of care 2: recording of histories

5.1 Introduction

The analysis of the recording of histories is reported in this chapter. For each dependent variable a preliminary univariable analysis was undertaken. This was used in conjunction with the information presented in Chapter 4 to inform the subsequent modelling (summaries of the raw data analysed in this chapter can be found in Tables 4.4 and 4.5). Consideration was given to the most appropriate form of the dependent variable—binary (e.g. was an item of history recorded or not); poisson (e.g. a count of the number of items of history); or continuous (e.g. a summary statistic such as the mean level of item recording). The preliminary analysis was also used to inform the choice of potential covariates (such as age) to include in the modelling.

Each dependent variable was then analysed using the strategy set out in Chapter 2. Data from enhancement forms and statutory records were analysed separately. For the first few variables the results of the modelling are given in full. For subsequent variables full details are only given if there is an interesting feature of the data or interesting aspect of the modelling that has not occurred before. Instead the results of the modelling are summarised. For all variables, data were available from both enhanced records and routine medical records. As the amount of information recorded in each type of record was very different, data from the two sources were analysed separately.

In general the analyses were based on the 2714 enhanced medical records and 752 statutory medical records described in Chapter 4 (Tables 4.1 through 4.5). There were seven enhanced records corresponding to children for whom no age was recorded. Data from these records were therefore excluded when preliminary (univariable) analysis indicated a relationship between the dependent variable and the age of the child.

An overview of the effects of standard setting on the recording of histories is given at the end of the chapter.

5.2 Social history

5.2.1 Descriptive statistics

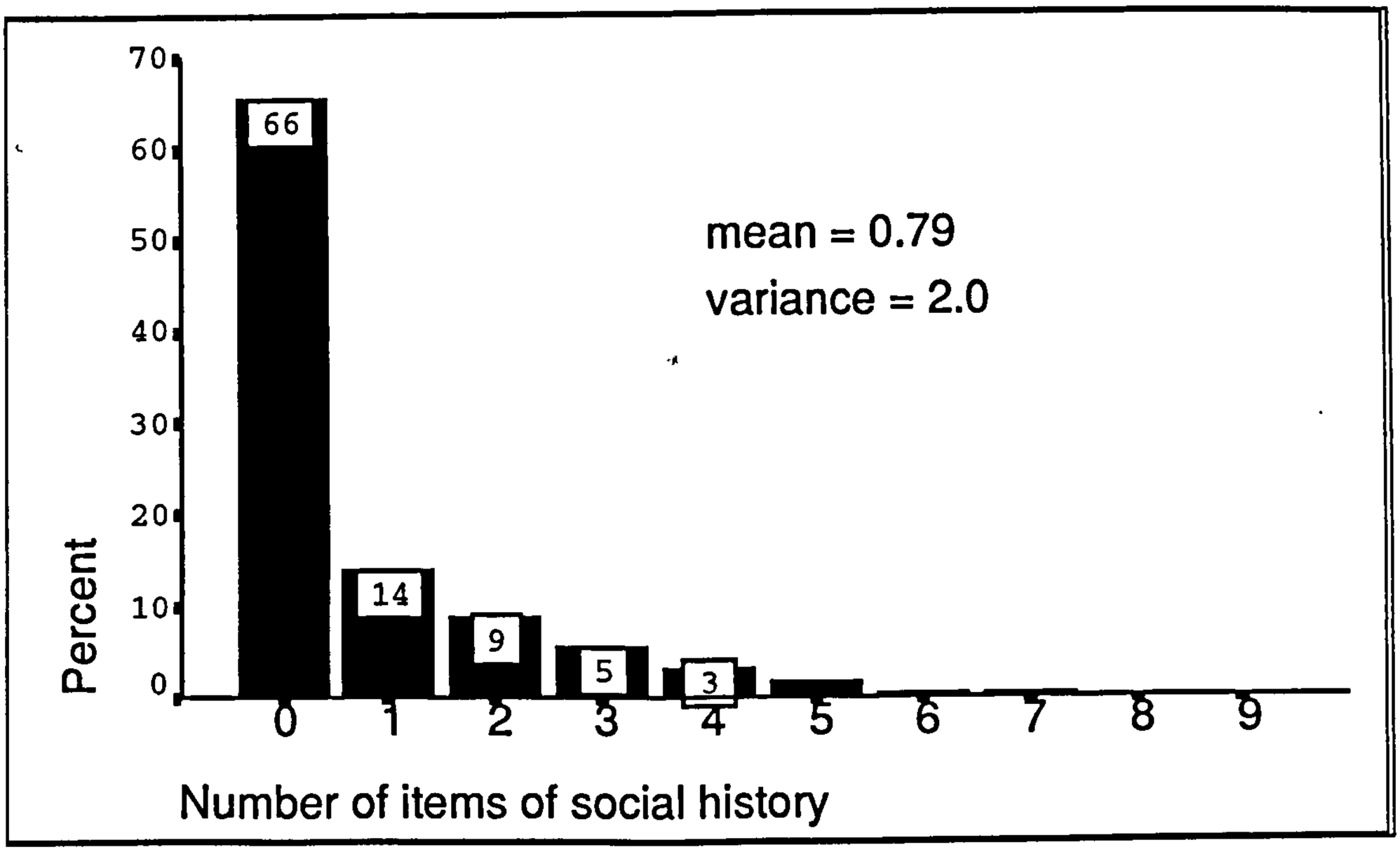
Items of social history were recorded more often on enhancement forms than on statutory records (Table 5.1). Social history was recorded most often for bedwetters, where the doctors tended to give an account of factors within the family that might have precipitated the bedwetting (such as marital strife or the birth of a sibling) and to describe sleeping and sanitary arrangements. Social history was also recorded for one-third of those consulting for acute vomiting, frequently including a profile of the child, parents and family.

The distribution of the number of items of social history given on enhancement forms in phase 1 is given in Figure 5.1. As the data are in the form of counts it is natural to investigate whether the distribution might be Poisson. The mean number of items of social history across all enhancement forms was 0·79; very much smaller than the variance which was 2·0. Responses clearly do not follow a Poisson distribution. The probability of an enhancement form, selected at random, having at least one item of social history is 0·34; the conditional probability of an enhancement form containing a second item of social history given that it contains at least one item is 0·59. Items of

Table 5.1 **Percentage of records in which an item of social history was recorded by study condition, phase and type of record. (The denominator is given in brackets)**

Study condition	Phase 1		Phase 2	
	Enhancement forms	Statutory records	Enhancement forms	Statutory records
Acute cough	36·0 (n = 389)	9·8 (n = 82)	26·5 (n = 359)	3·9 (n = 76)
Acute vomiting	41·9 (n = 332)	6·2 (n = 65)	33·3 (n = 246)	2·7 (n = 37)
Bedwetting	46·0 (n = 163)	14·7 (n = 34)	50·0 (n = 54)	16·7 (n = 90)
Itchy rash	25·5 (n = 341)	0·0 (n = 80)	23·7 (n = 186)	1·6 (n = 123)
Wheezy chest	29·1 (n = 357)	5·6 (n = 71)	20·2 (n = 287)	4·3 (n = 94)

Figure 5.1 **Distribution of the number of items of social history recorded on enhancement forms in phase 1**



social history tend to be recorded in clusters; in cases where doctors give an account of social history, they tend to record a number of items rather than just one. This would suggest that an analysis of the number of items of social history recorded would be

difficult to undertake and interpret. It was felt appropriate to analyse the presence of one or more items of social history as a binary variable.

5.2.2 Analysis of enhancement forms

The initial analysis was restricted to data abstracted from enhancement forms. Because of the problems doctors encountered in identifying and enhancing the records of children, there were a different number of cases associated with the cells of the Latin square set out in Table 2.1. These ranged from 1 to 82 (Table 4.2). It was therefore decided to analyse the data using techniques of generalised linear modelling described in Chapter 2 (Section 2.5). The key point of this method is that we make the individual record the unit of analysis. Effectively we then have one observation per cell; this helps us get round some of the problems associated with missing data.

Preliminary analysis suggested that there may be considerable variation in doctors' behaviour. In order to obtain valid estimates of the effects of standard setting it is essential to allow for variation between doctors. If doctors had been randomly assigned to trainer groups it would have been natural to include variation between doctors as a random effect (although at the time this analysis was undertaken there were no readily available packages for doing this if the dependent variable was binary). In practice allocation was systematic. Indeed some trainer groups had been formed prior to the start of the study for purposes that were unconnected with the study and trainers were allocated to the remaining groups on a geographical basis. It was therefore decided in all analyses to include variation between doctors as a fixed effect.

The results of this multiple logistic regression are reported in Table 5.2. Fitting the grand mean (model 1) leaves a residual deviance of 3376 with 2713 degrees of freedom. It is natural to include differences between the five study conditions (model 2) at an early stage in the statistical modelling; large differences were noted in the preliminary

univariable analysis and can be explained by the diverse nature of the conditions. The reduction in the residual deviance is 58·1 for a loss of four degrees of freedom. Comparing the reduction in residual deviance with a χ^2_4 distribution, the improvement is

Table 5.2 Proportion of enhancement forms containing one or more items of social history: model selection

Model Number	Specification	Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	Probability p
1	GM	3376	2713			
2	GM + COND	3318	2709	58·1	4	<0·001
3	GM + COND + TGRP	3187	2700	131·9	9	<0·001
4	GM + COND + TGRP + DOCT	2647	2626	539·6	74	<0·001
5	GM + COND + DOCT	2647	2626			
6	GM + COND + DOCT + PSTD	2639	2625	8.0	1	0·005
7	GM + COND + DOCT + PHAS	2621	2625	25·9	1	<0·001
8	GM + COND + DOCT + PHAS + COND·PHAS	2613	2621	8·3	4	0·08
9	GM + COND + DOCT + PHAS + PSTD	2621	2624	0·6	1	0·44

significant at the one percent level. Overdispersion is not a problem in this analysis; the ratio of the residual deviance to residual degrees of freedom for model 2 is only 1·13 which is close to unity. Consequently the conservative deviance ratio test and the less conservative likelihood ratio test described in Chapter 2 give identical results for this analysis.

Fitting differences between trainer groups (model 3) produces a significant improvement in fit—a reduction in deviance of 131·9 for a loss of nine degrees of freedom. The effect of differences between trainer groups is nested within another main effect—that of variation between doctors; the ten trainer groups are constituted from the 84 trainers. Thus when differences between doctors are added (model 4), there is a reduction of only

74 in the residual degrees of freedom (83 degrees of freedom for doctors less the nine degrees of freedom between groups, already included in the model). The improvement obtained by adding variation between doctors is significant at the one percent level. Model 4 can be expressed more simply by missing the nested effect 'TGRP' to give model 5.

There was a significant difference between phases (model 7). This is consistent with the data in Table 5.1 where there appears to be a reduction in the recording of social history in phase 2. There is some indication in that table that the reduction might be different for different conditions. We can test to see if this is the case by fitting an interaction between phases and conditions (model 8). The improvement upon adding this interaction was not significant suggesting that the changes observed between phases 1 and 2 were consistent across conditions.

There was evidence from interviews with the trainers (North of England Study 1990c) that, of the five levels of feedback (Figure 1.2), setting a clinical standard was by far the most likely to change clinical practice (more specific details were given in Chapter 2, Section 2.5.2). Trainers reported that the other levels of feedback had very little effect on what they did. Therefore it was decided, in the first instance, to test a binary variable comparing standard setters against non standard setters rather than an ordinal variable comparing all five levels of feedback. As standard setting was expected to cause changes only in phase 2, this was achieved by creating a new variable, PSTD, which took the value 2 when the consultation occurred in phase 2 with a trainer who set a standard for that condition and 1 otherwise. As identified in Chapter 2, this variable is confounded with phase—standard setting might result in a significant difference between phases but equally a change due to a trend over time might result in an apparent standard setting effect. It is necessary to fit PSTD with and without phase in the model.

Comparing model 7 with model 6 suggests that there was an effect of standard setting; the reduction in deviance is 8.0 for the loss of 1 degree of freedom. This reduction is significant at the one percent level although the improvement made on fitting PSTD is not as large as that made by fitting a phase effect. If we fit a standard setting effect after first allowing for a difference between phases, (compare model 9 with model 7) the reduction in deviance is only 0·6 for a loss of one degree of freedom. In model 7 we are fitting a difference between all observations in phase 1 and all observations in phase 2. In model 9 by including both PHAS and PSTD we are effectively fitting a separate phase effect for standard setters and non standard setters. In going from model 7 to model 9 we are estimating one extra parameter which represents the difference between the change in recording of social history between phases 1 and 2 for standard setters and the change in the recording of social history for non-standard setters. The improvement in fit of model compared with model 7 is not significant which indicates that both standard setters and non-standard setters changed their behaviour in exactly the same way between phases 1 and 2.

In this instance there were no *a priori* reasons for fitting further effects. There was no evidence, for example, to suggest that standard setting would lead to changes in the recording of social history that were greater for one particular condition than for any of the others. The model that best represents the recording of social history on enhancement forms, therefore, is one which allows for differences between conditions, differences between doctors and a difference between phases 1 and 2 (model 6).

The mathematical specification of model 6 is

$$\text{logit}\,\pi_{ijk} = \log_e\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \mu + \alpha_i + \beta_j + \gamma_k \qquad (5.1)$$

where: π_{ijk} is the underlying probability of doctor j recording at least one item of social history on an enhancement form for a child consulting with condition i in phase k;

α_i is the effect of condition i (i = 1, 2, 3, 4, 5);

β_j is the effect of doctor j (j = 1, 2, 3, ..., 84);

and γ_k is the effect of phase k (k = 1, 2).

A commonly used measure of the relative likelihood of a success (in this case the recording of an item of social history) in two groups is the odds ratio $\frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}$.

Using equation 5.1, for any given value of i and j, the difference in the log odds of an item of social history being recorded in phases 1 and 2 is given by:

$$\log_e\left(\frac{\pi_{ij2}}{1-\pi_{ij2}}\right)-\log_e\left(\frac{\pi_{ij1}}{1-\pi_{ij1}}\right)=\gamma_2-\gamma_1$$

Rearranging this equation gives: $$\frac{\pi_{ij2}(1-\pi_{ij1})}{\pi_{ij1}(1-\pi_{ij2})}=\exp\{\gamma_2-\gamma_1\} \qquad (5.2)$$

The package, GLIM, provides a maximum likelihood estimate of the difference $\{\gamma_2-\gamma_1\}$ and an associated standard error. These can be used in conjunction with equation 5.2 to provide an estimate of the odds ratio with an associated confidence interval. The odds ratio of an item being recorded in phase 2 relative to phase 1, controlling for variation between conditions and doctors, is 0·60 with 99% confidence interval, 0·46 to 0·78.

Similarly we can use model 9 to estimate the change in recording behaviour for standard setters and non-standard setters. For standard setters the odds ratio was 0.54 with 99% confidence interval, 0.35 to 0.83; for non-standard setters the odds ratio was 0.62 with 99% odds ratio 0.46 to 0.83. The difference in behaviour between standard setters and

non-standard setters was not significant. (This was formally tested by considering the improvement in fit when adding the term 'PSTD' to model 7 to obtain model 9.)

5.2.3 Analysis of statutory records

Fitting the grand mean (model 1, Table 5.3) results in a residual deviance of 346·2 with 751 residual degrees of freedom. The grand mean model (in which the probability of at least one item of social history being recorded on a statutory record is constant for all doctors, for each condition) appears to fit the data quite well. This is because doctors routinely record very few items of social history in statutory records; the probability is therefore very low. It is a common problem with binary data that the residual deviance may not be χ^2, however large n. However, the change in deviance upon adding terms still approximates well to a χ^2 with the requisite degrees of freedom (McCullagh and

Table 5.3 Proportion of statutory records containing one or more items of social history: model selection

Model Number	Specification	Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	Probability p
1	GM	346·2	751			
2	GM + COND	315·8	747	30·4	4	<0·001
3	GM + COND + TGRP	294·2	738	21·6	9	0·01
4	GM + COND + TGRP + DOCT	212·4	664	81·8	74	0·25
5	GM + COND + DOCT	212·4	664			
6	GM + COND + DOCT + PHAS	211·9	663	0·5	1	0·48
7	GM + COND + TGRP + PHAS	293·6	737	0·6	1	0·44
8	GM + COND + TGRP + PSTD	293·8	737	0·4	1	0·53

Nelder 1989, pages 119 and 122). There were significant differences between conditions (model 2) and trainer groups (model 3). Fitting differences between doctors (model 4) produces a reduction in deviance of 81·8 for the loss of 74 degrees of freedom.

Comparing this with the percentage points of a χ^2_{74} distribution would indicate that the improvement is not significant. But, as a percentage of the residual deviance, the reduction in deviance has been quite large. However, whether or not we leave the effect of doctors in the model, fitting a difference between phases (models 6 and 7) results in an improvement in fit that is very small. This would suggest that, overall, the difference in the recording of social history in statutory medical records between phases 1 and 2 is not significant.

Finally there is no evidence that standard setting had any effect on the recording of social history in statutory records (model 8).

The results of the analysis of statutory records are not completely consistent with those obtained from the analysis of enhancement forms. In particular the reduction in the recording of information that was observed on enhancement forms was not observed on statutory records. One possibility is that the reduction in recording observed on enhancement forms might be attributable to the diminishing enthusiasm for enhancement that was suspected earlier (Section 4.1).

5.3 Family and genetic history

5.3.1 Descriptive statistics

The level of recording of family history was about the same for all conditions, although the content varied. For the two acute conditions, a note of whether other family members were similarly affected was quite common; for bedwetting, history of enuresis in siblings and parents was often recorded; and for itchy rash and recurrent wheezy chest, information typically included whether there was any family history of related atopic conditions. Less information seemed to be recorded for older children perhaps

because this information had been entered into the medical records during consultations prior to the start of the study.

As with items of social history, items of family and genetic history tended to be recorded in clusters. The same analytic strategy was adopted; a binary variable corresponding to the presence or absence of at least one recorded item was created.

5.3.2 Analysis of enhancement forms

The proportion of enhancement forms containing items of family and genetic history was much greater than the proportion of statutory records containing items of social history (Table 4.5). Data from the two sets of forms were therefore analysed separately. A selection of models is given in Table 5.4

Table 5.4 Proportion of enhancement forms containing items of family and genetic history: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	Probability p
Number	Specification					
1	GM	2993	2706			
2	GM + COND	2987	2702	6·7	4	0·15
3	GM + TGRP	2897	2693	96·5	9	<0·001
4	GM + TGRP + DOCT	2523	2623	374·0	74	<0·001
5	GM + DOCT	2523	2623			
6	GM + DOCT + PHAS	2523	2622	0·3	1	0·58
7	GM + DOCT + PSTD	2522	2622	1·3	1	0·25
8	GM + DOCT + AGE	2517	2662	5·6	1	0·02
9	GM + DOCT + AGE + COND	2507	2618	10·4	4	0·03
10	GM + DOCT + AGE + COND + PSTD	2506	2617	1·0	1	0·32

Differences between doctors (model 4) were highly significant; individual doctors differed in the amount of social history that they recorded. When entered as the first

variable in the model (model 2), variation between study conditions was not significant. But after allowing for variation between doctors and an age effect (model 9), the variation was significant at the five percent level. Although we are adopting a significance level of one percent for testing for an effect of standard setting, consideration must be given to models that include possible explanatory variables and covariates that do not quite reach this level of significance. This is because of the potential problem of confounding arising from the unbalanced design. By not including potential explanatory variables we may miss effects present in the data. In this case the effect of standard setting has been tested with and without adjusting for the effects of study conditions and age (models 7 and 10). In neither case was the effect significant.

An interesting finding is that, unlike the recording of social history, there was no drop between phases 1 and 2 in the proportion of enhancement forms containing family and genetic history (model 6). It is not clear why there should be such a difference between the two types of history. Age (model 8) was significant at the five percent level. There is some evidence to suggest that less family history is recorded for older children.

5.3.3 Analysis of statutory records

On average 5·1 percent of statutory records contained an item of family and genetic history. The only effect that was significant was that of age of child and then only at the five percent level of significance. The regression coefficient of age (on the logistic scale, with age measured in years) was -0·13 (with a 95 percent confidence interval from -0·24 to -0·02); less family and genetic history was recorded for older children. There was no discernible effect that could be attributed to standard setting.

5.4 Previous medical history

5.4.1 Descriptive statistics

General information on previous medical history was recorded more often for the three chronic conditions. Children with a chronic illness typically consult repeatedly over a protracted period of time, allowing the doctor to build up a history of their condition. Not surprisingly such information was less likely to be recorded for very young children.

5.4.2 Analysis of enhancement forms

Presence of one or more items of previous medical history was analysed as a binary variable. The logistic model that best represented the data was one that included variation between doctors, variation between conditions and age of child as a linear trend (the regression coefficient was 0·07 with 95 percent confidence interval: 0·03 to 0·10). There were no differences in the recording of previous medical history between phase 1 and phase 2 and there were no effects of standard setting.

5.4.3 Analysis of statutory records

The analysis of statutory records gave results that were consistent with the analysis of enhancement forms-there was significant variation between doctors and conditions and some evidence of an age effect (but the significance of the age effect was only seven percent).

5.5 Previous diagnoses

5.5.1 Descriptive statistics

Previous diagnoses were recorded on just under 10 percent of enhancement forms and two percent of statutory records (Table 4.4). They were recorded most often for children consulting with recurrent wheezy chest (Table 4.5) who were frequently

described as 'known asthmatic'. There were very few instances of more than one previous diagnosis being recorded. The recording of previous diagnoses was therefore analysed as a binary variable.

5.5.2 Analysis

When enhancement forms were analysed, the only significant effects were those of variation between conditions and variation between doctors. There were no differences between phases 1 and 2, and no differences between doctors who had set a standard for a particular condition and those who had not. When statutory records were analysed, the only significant effect was that of variation between study conditions.

5.6 Previous non-drug management

A record of previous non-drug management, by parent or doctor, was generally rare but most common for bedwetters (Table 4.5). The presence of a record of previous non-drug management was analysed as a binary variable. For both data recorded on enhancement forms and data recorded in statutory records, there was significant variation between study conditions. There was significant variation between doctors in the information recorded on enhancement forms only. In neither case was there a difference between phases nor was there any effect that could be attributed to standard setting.

5.7 Previous drug management

Previous drug management, including responses to treatment, was noted more frequently than previous non-drug management. For the chronic conditions, information on drugs prescribed or advised at previous consultations was often included, while for acute conditions a record of medicine purchased and administered by the parent was not unusual. The presence of any items of previous drug management was analysed as

binary variable. Analysis of data recorded on enhancement forms indicated significant variation between conditions, significant variation between clinicians and a significant age effect—more information was recorded for older children (regression coefficient, 0·05 with 95% confidence interval 0·01 to 0·08). The proportion of enhancement forms containing an item of previous drug management ranged from just under 13 percent for acute vomiting to 52 percent for recurrent wheezy chest (Table 4.5). When data form statutory records were analysed the only significant effect was that of differences between conditions.

5.8 Effects of standard setting on the recording of histories

5.8.1 Assuming a uniform effect across all five conditions

Although standard setting did not have an effect on the recording of histories that was statistically significant, it is still instructive to look at estimates of the magnitude of the effect. Ninety five percent confidence intervals for the effect of standard setting on the

Table 5.5 Effect of standard setting on the recording of histories

Variable	Model	Odds ratio	95% confidence interval
Social history	GM+COND+DR+PHAS+PSTD	0·87	(0·60 to 1·24)
Family and genetic history	GM+COND+DR+AGE+PSTD	0·85	(0·61 to 1·17)
Previous medical history	GM+COND+DR+AGE+PSTD	1·03	(0·78 to 1·36)
Previous diagnosis	GM+COND+DR+PSTD	1·06	(0·69 to 1·64)
Previous non-drug management	GM+COND+DR+PSTD	0·65	(0·40 to 1·08)
Previous drug management	GM+COND+DR+AGE+PSTD	1·26	(0·92 to 1·71)

recording of histories are given in Table 5.5. These are based on the analysis of enhancement forms only. In each case the model on which the estimates are based is specified.

The 95% confidence intervals appear to be fairly wide but it is difficult to say whether they could include an effect that could be regarded as clinically significant. These results may be slightly easier to interpret if the inverse logit transformation:

$$\frac{e^{\mu+x}}{1+e^{\mu+x}} - \frac{e^{\mu}}{1+e^{\mu}}$$

is used to generate crude estimates of the effect of standard setting in terms of a change in proportion. In this expression, μ is an estimate of the parameter corresponding to the grand mean and x is an estimate of the parameter associated with the effects of standard setting. By replacing alternately x with the estimated lower and upper confidence limits for this parameter, an interval estimate of the effect of standard setting is generated. (Proportions have been multiplied by 100 so that results are given in terms of percentages).

For the first variable in Table 5.5, the transformation indicates that standard setting generated a change of between -9·8 and +4·8 percentage points in the recording of social history (this would be in addition to any change due to a trend over time). For the other history variables in Table 5.5, approximate 95% confidence intervals for the change due to the effects of standard setting were: family and genetic history between -7·9 and +3·0 percentage points; previous medical history between -6·1 and +7·7 percentage points; recording of a previous diagnosis between -2·7 and +5·2 percentage points; recording of previous non-drug management between -6·1 and +0·8 percentage points; and the recording of previous drug management -1·7 and +12·1 percentage points. The width of these intervals is variable. In general, the intervals are widest for variables where the proportion of records containing a relevant item of history is close to 0·5 (an intrinsic

feature of the binomial distribution). Whether these confidence intervals include changes that might be regarded as clinically significant is a matter for conjecture.

5.8.2 Assuming effects specific to each condition

The criteria for fitting a condition specific effect of standard setting were not met for any of the history variables and so the corresponding hypothesis that standard setting affected different conditions in different ways was not tested. But for the purposes of

Table 5.6 Ninety five percent confidence intervals for the effect of standard setting on the recording of histories for each condition

Variable	Cough	Vomit	Bedwetting	Itchy rash	Wheezy chest
Social history	0·45 to 1·54	0·36 to 1·49	0·20 to 2·69	0·56 to 2·62	0·39 to 2·03
Family and genetic history	0·57 to 1·71	0·23 to 1·27	0·41 to 5·10	0·44 to 2·06	0·30 to 1·49
Previous medical history	0·61 to 1·66	0·60 to 2·13	0·35 to 4·71	0·82 to 4·18	0·37 to 1·19
Previous diagnosis	0·48 to 2·56	0·03 to 1·76	0·13 to 10·2	0·99 to 6·31	0·42 to 1·97
Previous non-drug management	0·04 to 0·93	0·16 to 1·52	0·42 to 6·92	0·35 to 2·90	0·21 to 2·78
Previous drug management	0·63 to 2·04	0·16 to 1·25	0·21 to 3·66	1·14 to 4·70	0·85 to 2·80

evaluating the power of the study to detect differences, it is instructive to fit the term CPST as defined in Chapter 2. The resulting models can then be used to generate estimates of the effects of standard setting on each condition. Ninety five percent confidence intervals for these effects are given in Table 5.6. Again the confidence intervals are fairly wide. Subject specific estimates of the effects of standard setting are based on comparatively few observations. These results suggest that the power of the

study to detect condition specific changes of standard setting in a binary variable is fairly low.

Chapter 6

Process of care 3: diagnosis of current episode

6.1 Introduction

In this chapter, the analysis of variables relating to the diagnosis of the current episode is described. In general these items were recorded more frequently than the items of history described in the previous chapter and there was greater scope in the choice of method of analysis. Many of the variables have been analysed using more than one approach. Results of the alternative analyses are compared. In particular, for a number of variables, there was a choice between analysing the *change* in behaviour (of an individual doctor) between phase 1 and phase 2 (there is therefore just one observation per doctor per condition in the analysis) or including the behaviour (of the doctor) in each phase separately (there are two observations—one in each phase—per doctor per condition in the analysis). The former of these two methods will be referred to as an *analysis of differences*; the latter will, for convenience, be referred to as a *repeated measures analysis* although only univariate statistical tests have been undertaken (multivariate tests that are usually associated with a repeated measures analysis have not been undertaken).

As in the previous chapter, a preliminary analysis of each variable, used to inform the modelling process, is described. The analyses of the first variables are given in full. For

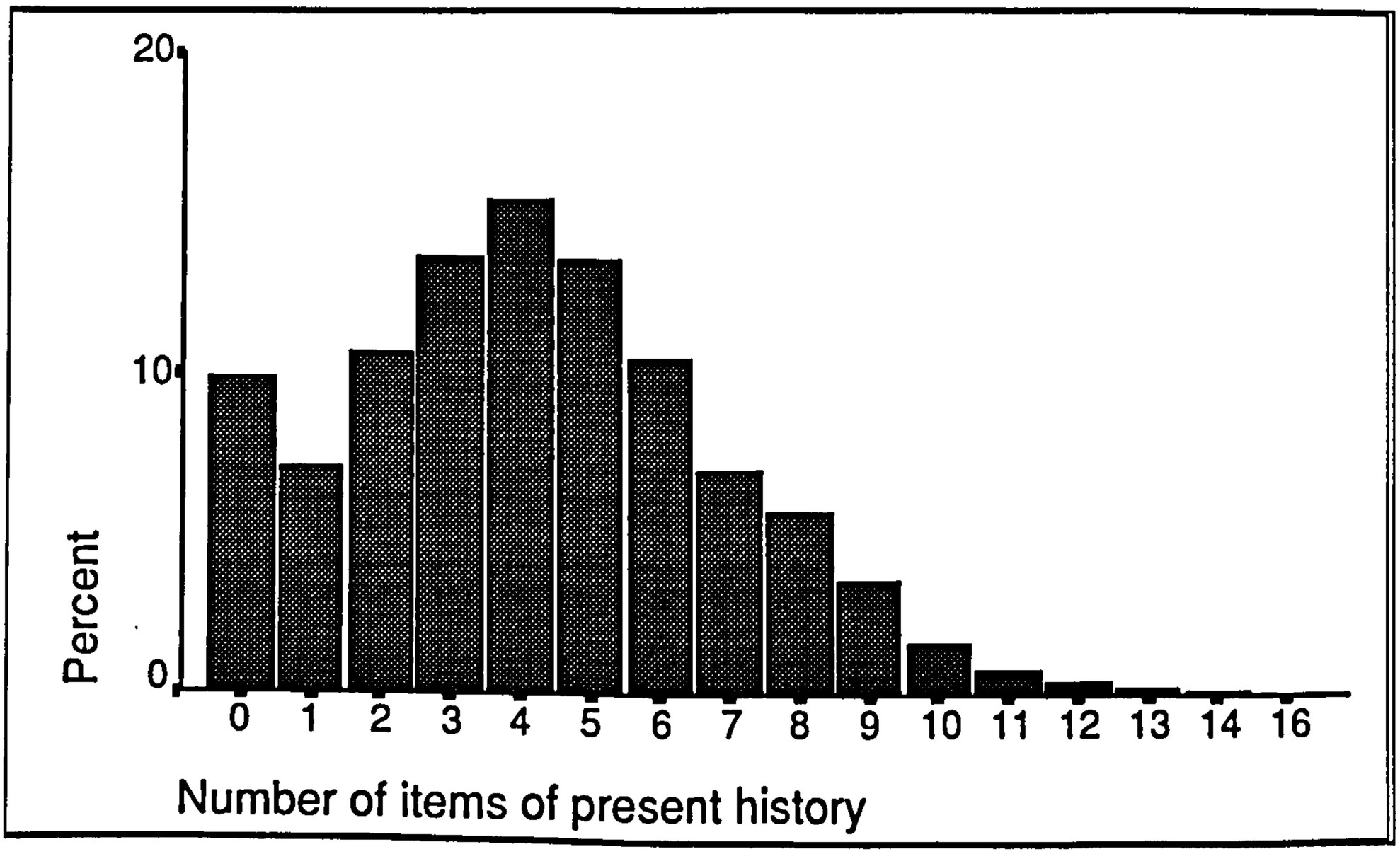
subsequent variables, summaries of the analyses only are given unless there are features of the modelling that have not been encountered before.

Again the analyses are based on 2714 enhanced medical records and 752 routine medical records (described in Tables 4.1, 4.2, 4.3, 4.4 and 4.5). There was very little missing data, thus in the analysis of a binary variable there are usually 2713 residual degrees of freedom after fitting the grand mean. The main exception to this is when age is included as a covariate. There were seven enhanced medical records corresponding to children for whom age was not recorded. Data from these records were excluded listwise from these analyses (the residual degrees of freedom after fitting the grand mean was 2706).

6.2 Current medical history (items of history relating to the presenting illness)

6.2.1 Descriptive statistics

Figure 6.1 Frequency distribution of the number of items of current medical history recorded on enhancement forms



Information on current history was present on 95 percent of enhancement forms (Table 4.4) and was most likely to be recorded for the two acute conditions (Table 4.5). The distribution of the number of items of present history is given in Figure 6.1. A number of methods of analysing these data were considered within the framework of the generalised linear model.

6.2.2 Poisson error structure

As these data take the form of a series of counts, it is natural to try to fit a model with a Poisson error structure. The number of items of current medical history, R, on an enhancement form is assumed to be modelled by

$$\Pr(R = r|\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad \lambda > 0, \qquad r = 0,1,2,....$$

where λ is the mean number of items. A log link function was used such that $\log \lambda = \beta'x$. The results of this analysis are given in Table 6.1.

Table 6.1 Number of items of current medical history recorded on enhancement forms: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	Mean deviance	p
Number	Specification						
1	GM	4403	2713				
2	GM + COND	3739	2709	664·9	4	166·2	<0·001
3	GM + COND + TGRP	3705	2700	33·4	9	3·7	0·004
4	GM + COND + TGRP + DOCT	3155	2626	550·2	74	7·4	<0·001
5	GM + COND + DOCT	3155	2626				
6	GM + COND + DOCT + PHAS	3155	2625	0·1	1	0·1	0·77
7	GM + COND + DOCT + PSTD	3154	2625	0·5	1	0·5	0·52

Fitting the grand mean (model 1) gives a residual deviance of 4403 with 2713 residual degrees of freedom. Fitting variation between conditions (model 2) results in a reduction of 664·9 in the residual deviance for a loss of four degrees of freedom. In this case there is some evidence of overdispersion in the data. The residual deviance is much larger than the residual degrees of freedom. We therefore opt for the more conservative deviance ratio test described in Chapter 2, to assess improvement in fit. The improvement is highly significant. Although variation between trainer groups was only significant at the five percent level (model 3), there was in fact large variation between individual doctors (model 4). There were no other significant effects. In particular, there was no difference in recording between phases 1 and 2 (model 6) and no effect of standard setting (model 7).

Residual plots can be used to check the fit of the model and to look for violations of the assumptions made in fitting it. Anscombe (1961) proposed defining residuals that followed an approximate Normal distribution. McCullagh and Nelder (1989, page 38) show that the transformation that both "normalises" the probability function and also stabilises the variance is:

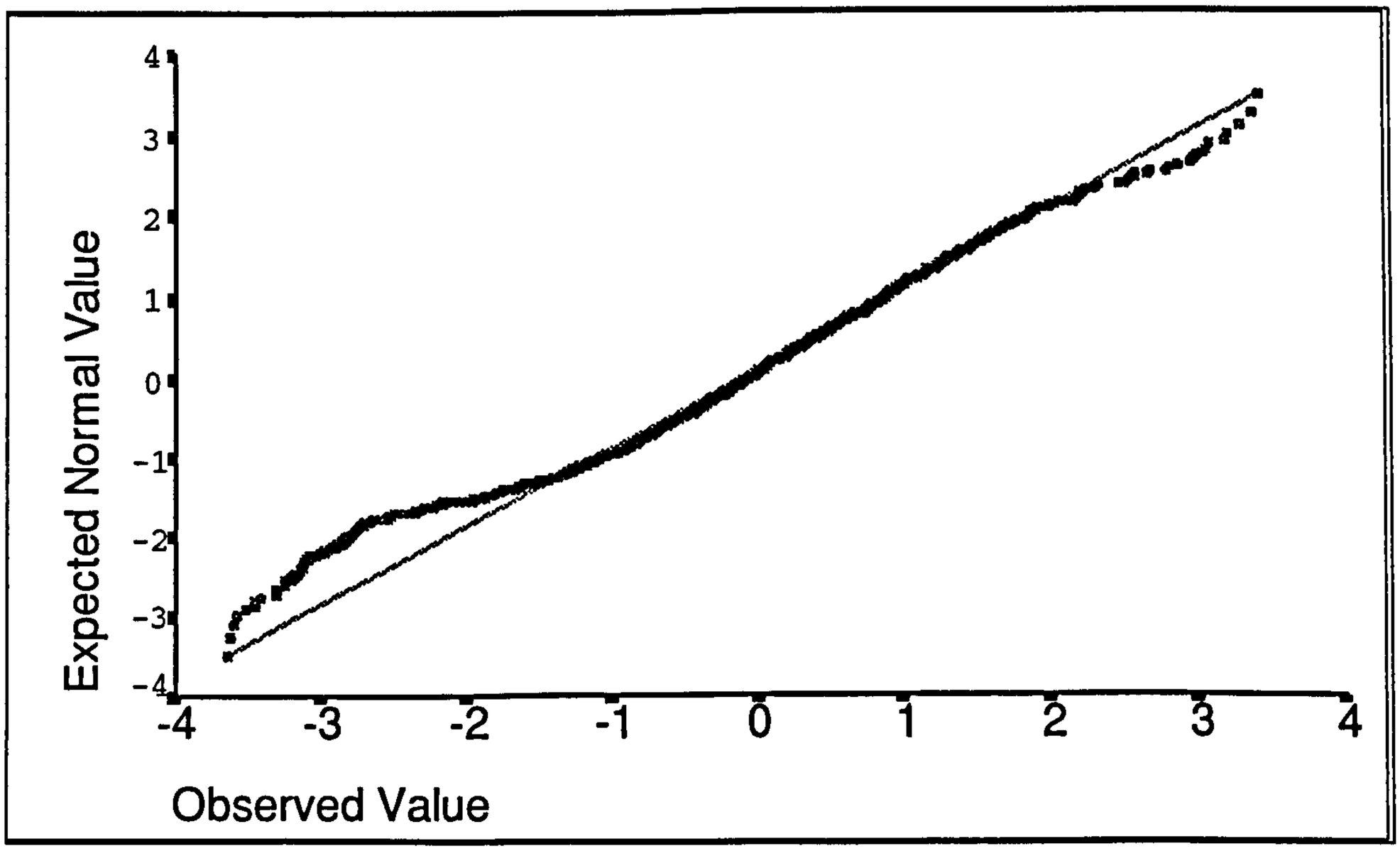
$$r_i = \frac{\frac{3}{2}\left(y_i^{2/3} - \hat{\mu}_i^{2/3}\right)}{\hat{\mu}_i^{1/6}}$$

where the y_i's are the observed values and the $\hat{\mu}_i$'s are the fitted values. As these plots should follow an approximate Normal distribution a Normal (or quantile-quantile) plot can be used to check the model. The residuals are sorted into ascending order and are plotted against the expected order statistics of a Normal sample. McCullagh and Nelder (1989, page 407) suggest that the expected order statistics be calculated as

$$\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right)$$

where Φ^{-1} is the inverse Normal cumulative density function. The quantile-quantile plot for model 5 is shown in Figure 6.2 Departure from a straight line is noticeable at each end of the plot. This corresponds to the two extremes of the distribution of the number of items of current history; there were more enhancement forms with no recorded

Figure 6.2 History of presenting condition: a Normal quantile-quantile plot of Anscombe residuals assuming a Poisson error structure

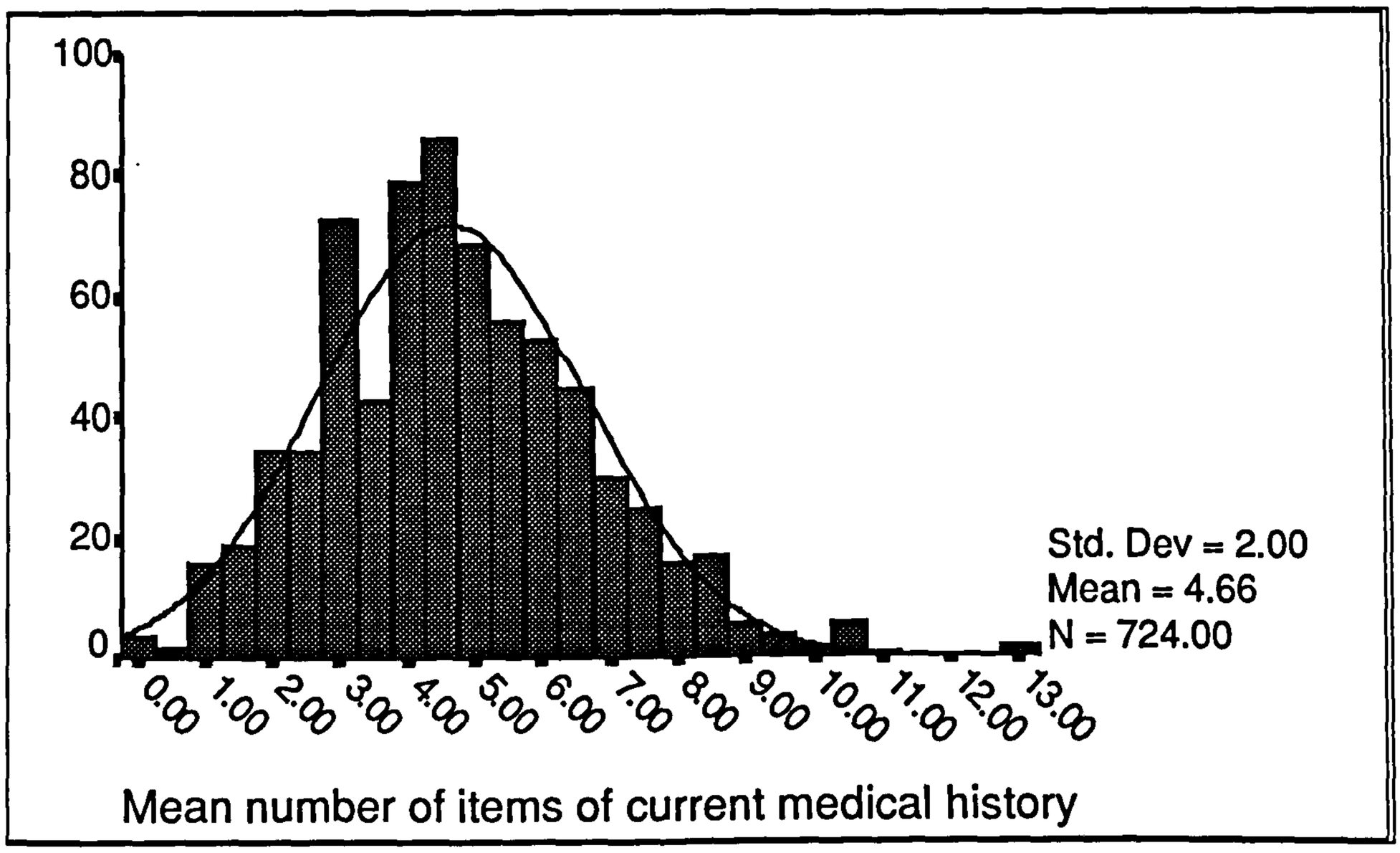


histories and more enhancement forms with a large number of histories than would be expected if the distribution was truly Poisson.

6.2.3 Normal error structure - analysis of cell means

An alternative approach to analysing this data is to model the mean number of items of current medical history for each doctor-condition combination recorded on enhancement form in each phase of the study. The distribution of these means is given in Figure 6.3. It would seem reasonable to model these data using a Normal error structure.

Figure 6.3 Distribution of mean number of items of current medical history recorded on enhancement forms by individual doctors for each condition



Two approaches were considered. The first was to repeat the analysis described above but using cell means rather than counts. This is reported in Table 6.2.

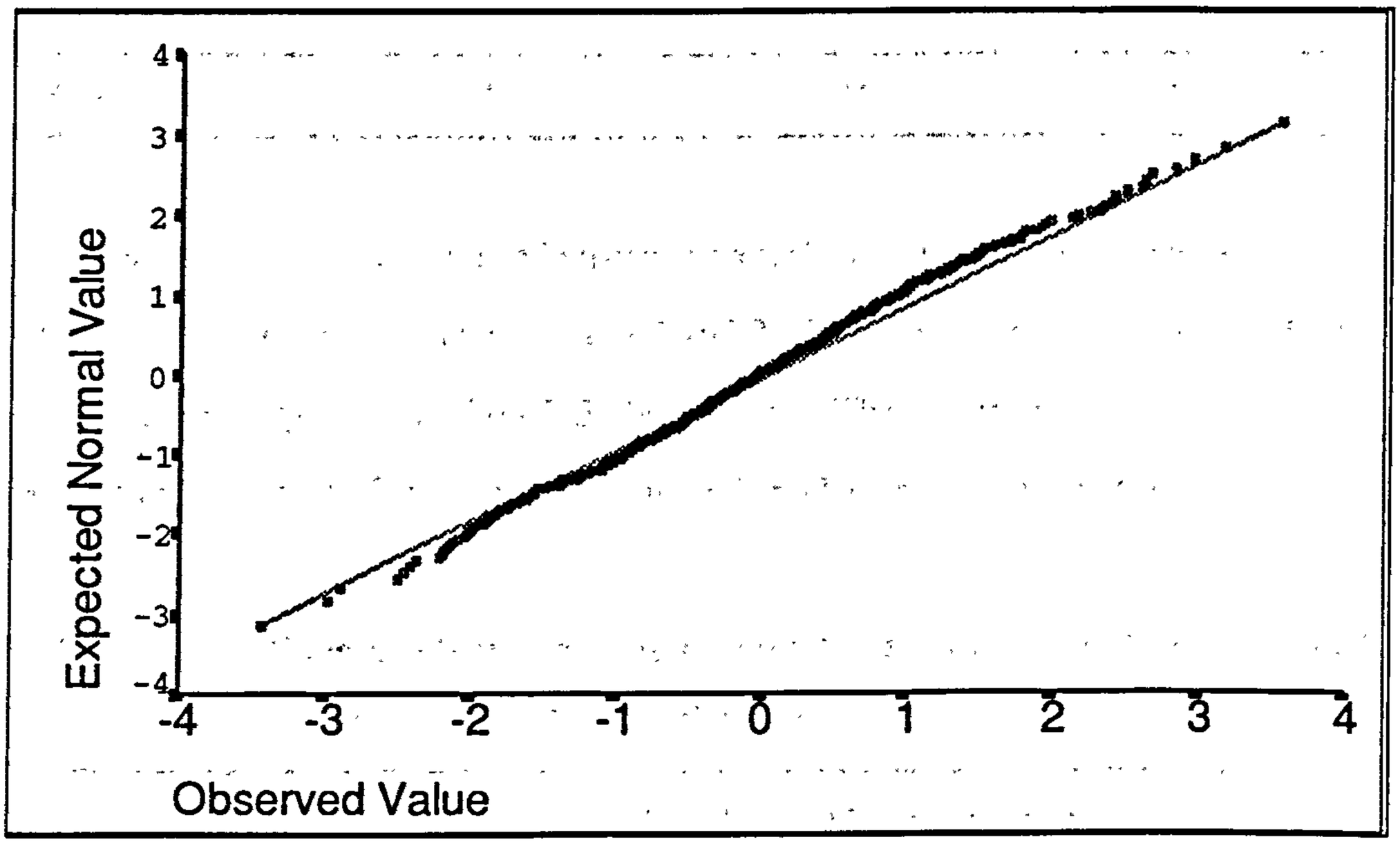
Table 6.2 Mean number of items of current medical history: model selection

Model		Residual sum of squares	Residual degrees of freedom	Change in sum of squares	Change in degrees of freedom	Mean square	p
Number	Specification						
1	GM	2905	723				
2	GM + COND	1992	719	913	4	228·3	<0·001
3	GM + COND + DOCT	1269	636	723·0	83	8·7	<0·001
4	GM + COND + DOCT + PHAS	1268	635	1·3	1	1·3	0·42
5	GM + COND + DOCT + PSTD	1268	635	1·5	1	1·5	0·39

The results are the same as before—there were differences between conditions, there was significant variation among doctors, there was no differences between phases and no

observed effect of setting standards. The ordered quantile-quantile plot of standardised residuals (Figure 6.4) indicates that the distributional assumptions are justified. The Normal probability model fits the data quite well.

Figure 6.4 Mean number of items of current medical history: a Normal quantile-quantile plot of standardised residuals



6.2.4 Normal error structure - analysis of differences

The second approach that was adopted, was to model directly the differences between phases 1 and 2. A key objective of the study was to estimate the effects that standard setting had on the recording of process data. Any changes brought about by standard setting would have occurred between phases 1 and 2. It therefore seemed natural to analyse the difference between the mean score in phase 1 and the mean score in phase 2.

The main problem with this approach was the reduced level of enhancement in phase 2 that we noted earlier (see Table 4.2). The number of doctors who identified children (and provided the data shown in that table) is given in Table 6.3

Table 6.3 Number of doctors who enhanced medical records by study condition and phase

Study condition	Phase 1	Phase 2	Both
Acute cough	83	81	80
Acute vomiting	81	72	71
Bedwetting	64	35	30
Itchy rash	79	69	68
Wheezy chest	83	77	76
Total	390	334	325

In order to calculate a difference in mean scores, doctors need to have enhanced records for a particular condition in both phases 1 and 2. Thus we make use of only a subset of the data; the analysis is based on 325 observations. The differences in mean scores between phases 1 and 2 were modelled assuming a Normal error structure (Table 6.4).

Table 6.4 Difference between phases 1 and 2 in the mean number of items of current medical history: model selection

Model		Residual sum of squares	Residual degrees of freedom	Change in sum of squares	Change in degrees of freedom	Mean sum of squares	p
Number	Specification						
1	GM	971·2	324				
2	GM + COND	936·7	320	34·5	4	8·6	0·02
3	GM + COND + DOCT	671·5	240	264·9	80	3·3	0·17
4	GM + CHRN	938·7	323	32·5	1	32·5	<0·001
5	GM + CHRN + COND	936·7	320	2·0	3	0·7	0·88
6	GM + CHRN + STND	938·4	322	0·3	1	0·3	0·75

Variation between conditions was significant at the five percent level (model 2); there was no significant variation between doctors (model 3). The GLIM parameter estimates and associated standard errors for model 2 are given in Table 6.5

Table 6.5 Maximum likelihood estimates of the parameters in model GM + COND

Parameter	GLIM estimate	Standard error
Grand mean		
GM	0·34	0·19
Study condition		
Acute cough	0	
Acute vomiting	-0·16	0·28
Bedwetting	-0·87	0·36
Itchy rash	-0·64	0·28
Wheezy chest	-0·71	0·27

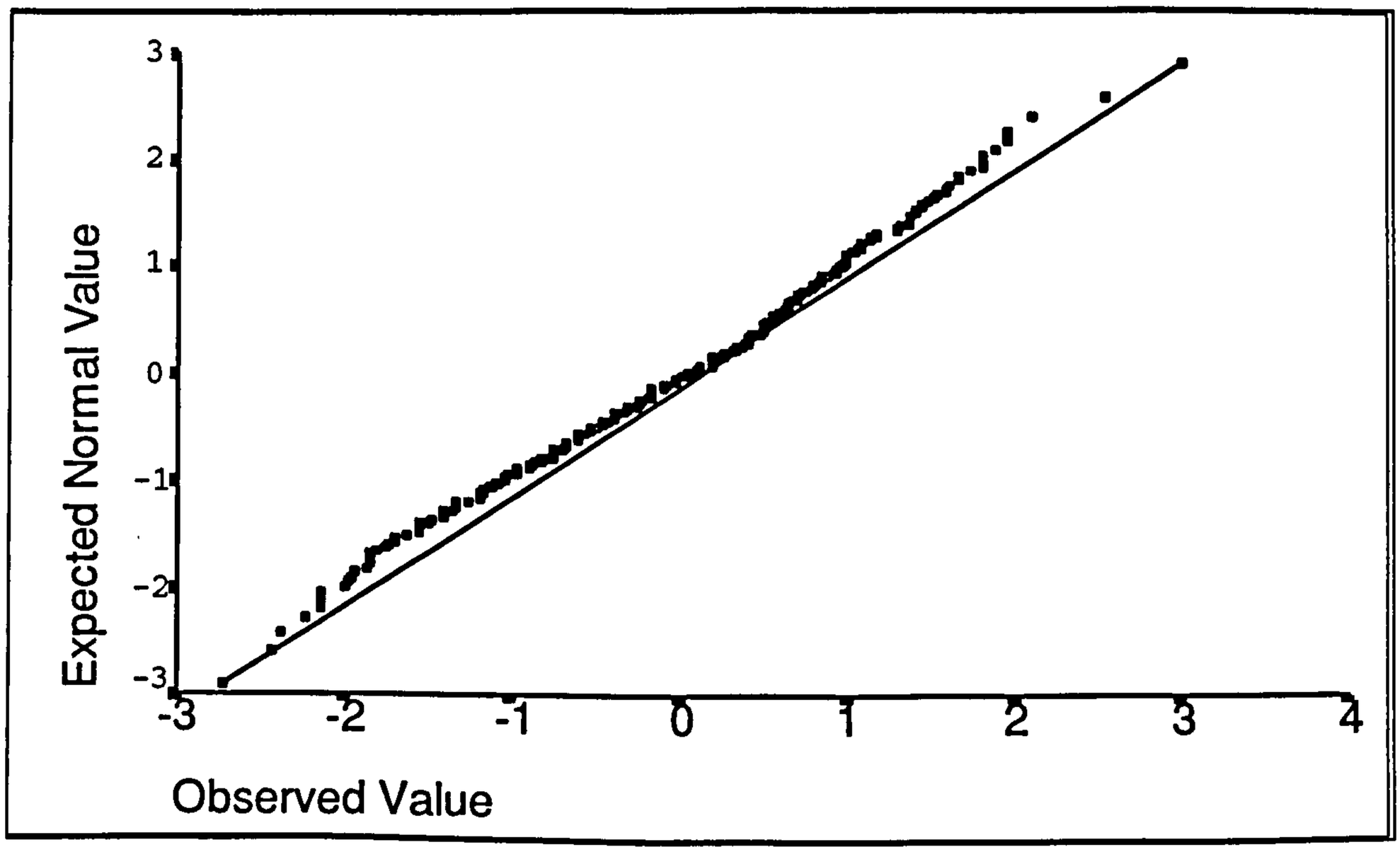
In GLIM, one of the parameters for each main effect is set to zero (Chapter 2, section 2.4.4). In this case it is the parameter corresponding to consultations for acute cough which is set to zero. The other parameters are shown relative to this one. Thus for acute cough, the estimated change in mean score is 0·34 (corresponding to the grand mean + 0 = 0·34) items of current medical history; for acute vomiting the estimated change is 0·34 -0·16 = 0·18. Similarly, the estimated changes for the other conditions are: bedwetting—a decrease of 0·53 items; itchy rash—a decrease of 0·30 items; and wheezy chest—a decrease of 0·37 items of current medical history.

These estimates suggest that most of the variation between conditions could be explained by a difference between the acute conditions and the chronic conditions. For the two acute conditions, the model would suggest that there was a slight increase in the mean number of items of current medical history recorded; for the three chronic conditions there was a slight drop. We can test this hypothesis by fitting a contrast (represented by the new variable "CHRN") which is set to 1 for the two acute conditions and 2 for the three chronic conditions—model 4. The resulting drop in the residual sum of squares was 32·5 for the loss of one degree of freedom—an improvement in fit that is

significant at the 0·1 percent level. Fitting variation between the individual conditions (model 5) now results in a reduction of only 2·0 for the loss of the three degrees of freedom; the corresponding F test is clearly not significant. The variation between conditions can be explained by a difference between the two acute conditions and the three chronic conditions.

As in the two previous analyses there was no evidence of any effect of standard setting (model 6). No other effects were significant. The model that best represents this data is simply one that allows for a difference between acute conditions and chronic conditions. A Normal quantile-quantile plot of standardised residuals (Figure 6.5) shows a close fit to a straight line; this indicates that the assumption of a Normal error structure is reasonable.

Figure 6.5 Change in the mean number of items of current medical history: a Normal quantile-quantile plot of standardised residuals



6.2.5 Comparison of methods

Each of the three methods of analysis described above has specific advantages and disadvantages. Examination of the plots of residuals indicates that the distributional assumptions appear to have been less good for the initial analysis of counts than for the two subsequent analysis that involved application of a Normal error structure. One of the main advantages of the analysis of counts is that covariates that relate to particular consultations (such as age of child and gender of child) can be incorporated into the model. It is less easy to adjust for these effects when the data is aggregated to provide a summary statistic for each doctor. In this case, however, univariate analyses suggested that the recording of history relating to the presenting condition was not affected by any of these particular covariates.

When mean scores were analysed, the choice was between either treating the data from the two phases as repeated measures or analysing the difference in means between the two phases. The former permitted the inclusion of all the available data in the analysis but, because data for some doctors are available in only one of the phases, the design is more unbalanced and there is a greater level of confounding between the variables which had to be investigated during the modelling process. The second approach allowed advantage to be taken of the natural pairing of the data; differences between the two phases were modelled directly. This is attractive because any effects of standard setting will be to cause some differences between phases 1 and 2. Although no effects of standard setting could be identified, this analysis indicated that there might be some differences between the acute and chronic conditions between the two data collection phases.

This difference was not detected by either of the other two analyses because the slight increase in recording of history for the two acute conditions was offset by the slight reduction for each of the three chronic conditions. Overall there was no difference

between the two phases (model 6, Table 6.1 and model 4, Table 6.2). To detect different changes for each condition in the first two analyses, it is necessary to fit an interaction. We can model the different change for acute and chronic conditions between phases 1 and 2 by fitting the interaction 'CHRN·PHAS' (Table 6.6).

Table 6.6 Fitting a different change for acute and chronic conditions between phases 1 and 2

Model		Residual scaled* deviance	Residual degrees of freedom	Change in scaled* deviance	Change in degrees of freedom	Mean scaled* deviance	p
Number	Specification						
Number of items of current medical history (Poisson error model)							
1	GM + COND + DOCT	3155	2626				
2	GM + COND + DOCT + PHAS + CHRN·PHAS	3147	2624	8·1	2	4·0	0·03
Mean number of items of current medical history (Normal error model)							
3	GM + COND + DOCT	1269	636				
4	GM + COND + DOCT + PHAS + CHRN·PHAS	1252	634	17·6	2	8·8	0·01

* In the analysis of counts the scale parameter is 1 (the scaled deviance is simply the deviance as defined in Chapter 2); in the analysis of mean scores the scaled deviance is the sum of squares.

The interaction term cannot be fitted without also fitting a main effect representing the differences between phases. In Table 6.6 both the main effect and interaction term are added to the model which includes terms representing variation between conditions and variation between doctors. In the analysis of counts, the reduction in the deviance was 8·1 for a loss of two degrees of freedom (model 2). Allowing for overdispersion this represents an improvement that is significant at around the five percent level. A similar improvement is obtained by inclusion of the interaction term when modelling the mean number of items of medical history relating to the presenting condition (model 4). In both cases the difference between all five conditions has already been included in the model. There is a loss of two degrees of freedom because two additional parameters

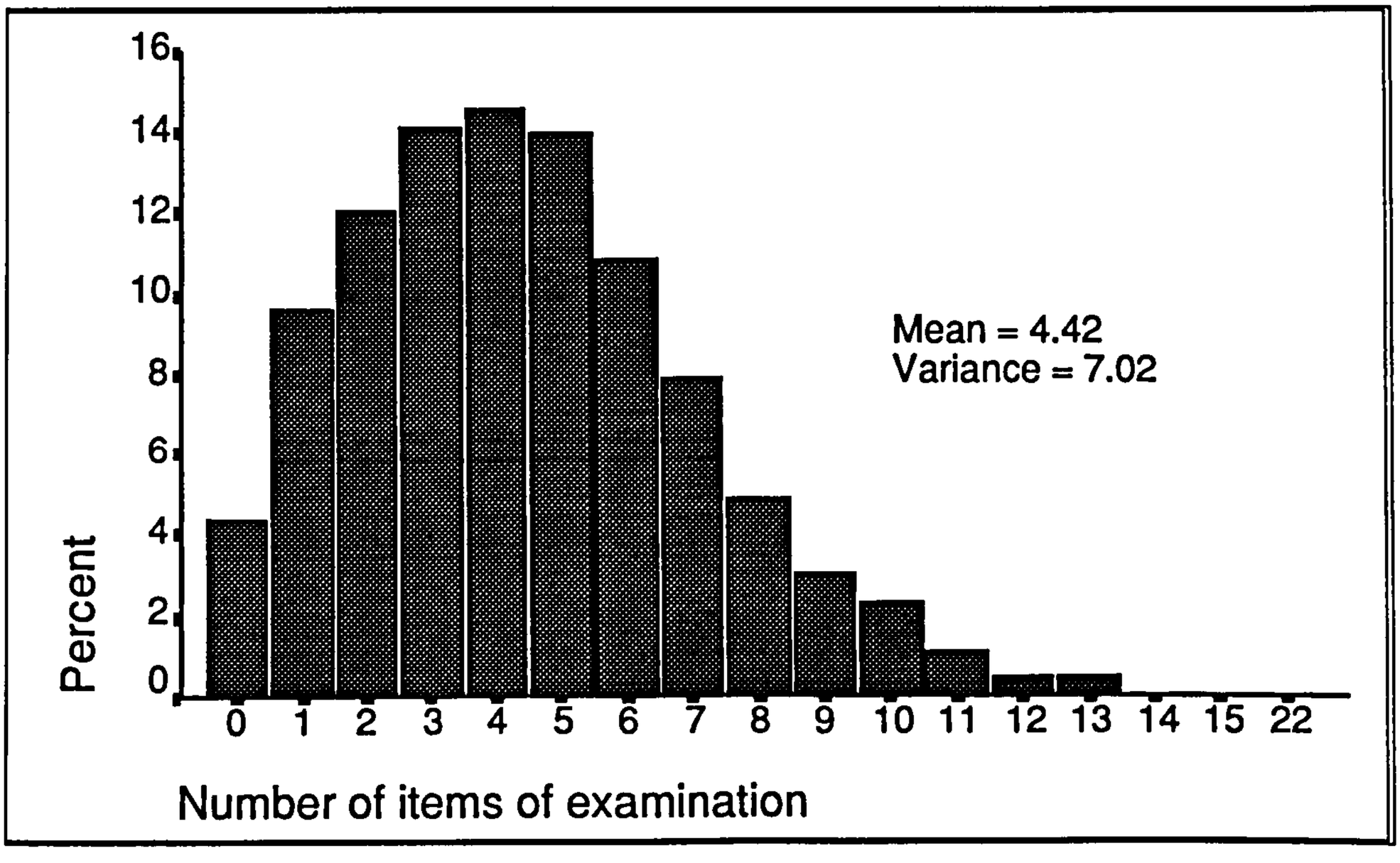
need to be estimated. These correspond to an estimate of the change for acute conditions between phases 1 and 2 and an estimate of the change for chronic conditions between phases 1 and 2. (Equivalently there is one degree of freedom associated with the main effect PHAS and one degree of freedom associated with the interaction CHRN·PHAS).

The tests given above are not directly equivalent to the test for a chronic condition effect in the analysis of differences. The equivalent test is a test of the difference between the change for chronic conditions and the change for acute conditions based on one degree of freedom. This is achieved by looking at the improvement obtained by fitting the pure interaction term CHRN·PHAS with the main effect PHAS already incorporated in the model. In the analysis of counts the reduction in deviance is 8·0; in the analysis of mean scores the reduction in the residual sums of squares is 16·3. In both cases there is a reduction of one in the residual degrees of freedom; in both cases the improvement in the fit of the model is significant at the one percent level. The three different analyses give consistent results.

6.3 Examination findings

Examination findings were most likely to be recorded for the two acute conditions (Table 4.5). The most frequently recorded items related to examination of the abdomen (72·4% of consultations for children with acute vomiting) and auscultation of the chest (92·1% of consultations with children with acute cough). The average number of items of examination recorded on enhancement forms ranged from 1·8 for bedwetting to 5·6 for acute vomiting. The distribution of the number of items recorded for all consultations is given in Figure 6.6

Figure 6.6 Items of examination recorded on enhancement forms: frequency distribution



The distribution is similar to the distribution of recorded items of current medical history described in the preceding section (Figure 6.3).

Analysis of the mean number of items of examination recorded by a doctor for children consulting with a particular condition was undertaken (Table 6.7). The repeated measures analysis, including phase as a fixed effect (models 1 to 7) and the analysis of differences (models 8 to 12) gave consistent results.

There were differences between conditions (model 2) and significant variation among doctors (model 3). There was also a difference between phases 1 and 2 (model 5). This phase difference was different for each condition resulting in a significant condition by phase interaction (model 6) in the first analysis and a simple condition effect (model 9) in the second. There were no discernible effects of standard setting (models 4, 7, 10 and 12).

Table 6.7 Mean number of items of examination: model selection

Model		Residual sum of squares	Residual degrees of freedom	Change in sum of squares	Change in degrees of freedom	Mean square	p
Number	Specification						
	Repeated measures analysis						
1	GM	3235	723				
2	GM + COND	2258	719	976·9	4	244·2	<0·001
3	GM + COND + DOCT	1159	636	1098·9	83	7·3	<0·001
4	GM + COND + DOCT + PSTD	1158	635	1·1	1	1·1	0·44
5	GM + COND + DOCT + PHAS	1145	635	14·1	1	14·1	0·005
6	GM + COND + DOCT + PHAS + COND·PHAS	1120	631	24·8	4	6·2	0·007
7	GM + COND + DOCT + PHAS + COND·PHAS + PSTD	1120	630	0·1	1	0·1	0·81
	Analysis of differences						
8	GM	1035·1	324				
9	GM + COND	986·3	320	48·8	4	12·2	0·004
10	GM + COND + STND	981·4	319	4·9	1	4·9	0·21
11	GM + COND + DOCT	649·7	240	336·6	80	4·2	0·006
12	GM + COND + DOCT + STND	644·1	239	5·6	1	5·6	0·15

The mean number of items of examination recorded on enhancement forms is broken down by condition and phase in Table 6.8. The last column gives the estimated change for each condition between phases 1 and 2. These estimates are obtained from model 6 and are thus adjusted to take into account variation between doctors. There was a significant fall in the recording of information relating to examinations for three of the conditions—acute cough, acute vomiting and recurrent wheezy chest. For the other two conditions, bedwetting and itchy rash the change was not significant.

Table 6.8 Mean number of items recorded on enhancement forms by condition and phase

Study condition	Phase 1		Phase 2		Estimated change (with 95% confidence interval)*
	Mean	Standard deviation	Mean	Standard deviation	
Acute cough	5·03 (n = 389)	2·27	4·50 (n = 359)	2·22	-0·61 (-1·02 to -0·20)
Acute vomiting	5·79 (n = 332)	2·50	5·42 (n = 246)	2·63	-0·52 (-0·95 to -0·09)
Bedwetting	1·67 (n = 163)	1·92	2·04 (n = 54)	2·08	0·40 (-0·16 to 0·96)
Itchy rash	4·01 (n = 341)	2·62	4·26 (n = 186)	2·47	0·14 (-0·29 to 0·58)
Wheezy chest	4·24 (n = 357)	2·81	3·85 (n = 287)	2·72	-0·49 (-0·90 to -0·07)

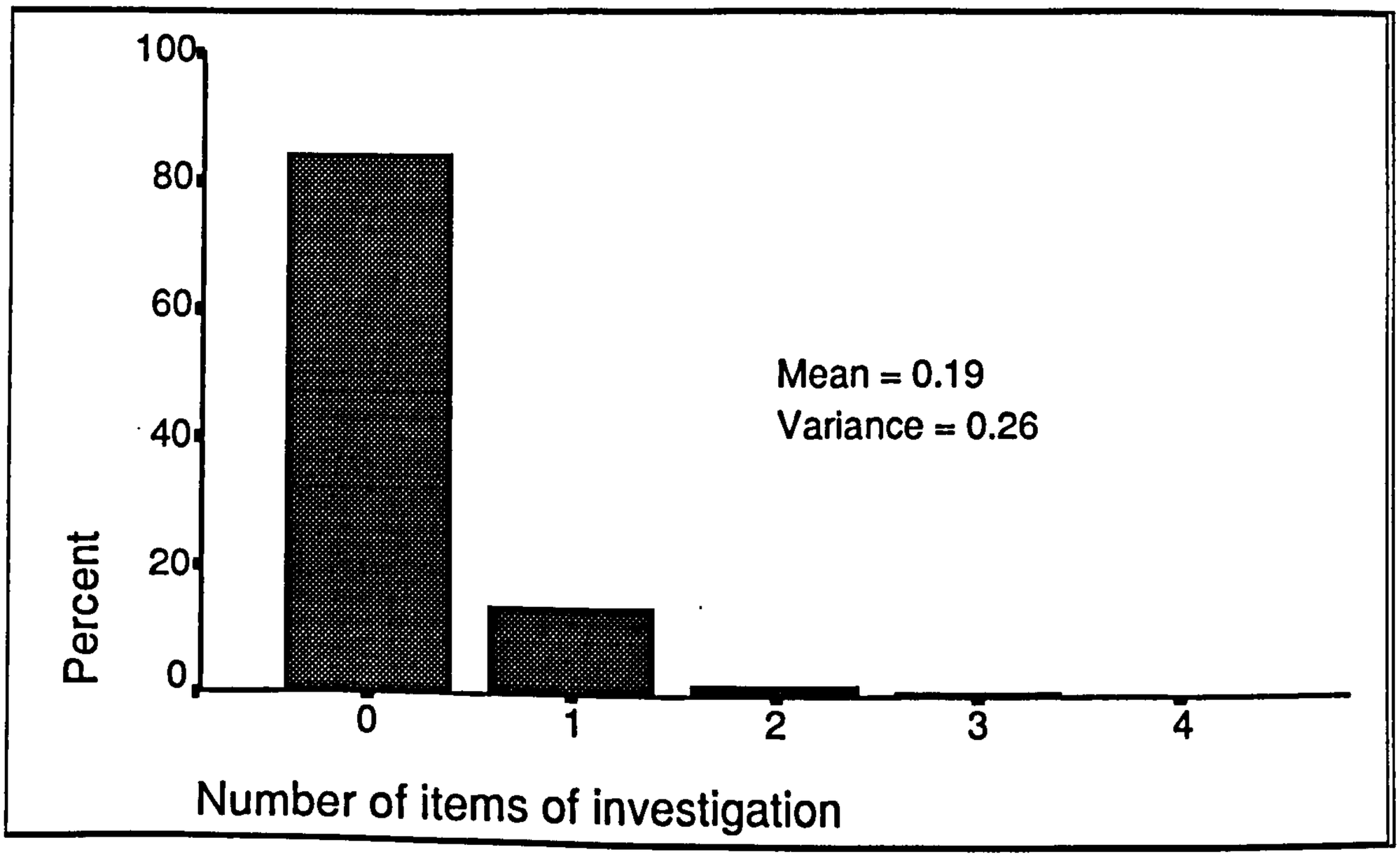
Note:

* The estimated change and associated confidence intervals are based on the GLIM model which takes into account variation between doctors (in addition to variation between conditions and variation between phases)

6.4 Investigations

Planned investigations or results of previous tests were most frequently reported for

Figure 6.7 Items relating to investigations: frequency distribution



bedwetting (urinalysis, urine microscopy or culture) and for recurrent wheezy chest (peak flow measurements or chest X-rays). Overall, no investigations were recorded for more than 80 percent of consultations (Figure 6.7). It was decided to analyse the presence of one or more investigations as a binary variable. There was significant variation between conditions and significant variation among doctors. There was also a significant age effect: more investigations were undertaken for older children. There was no difference between phases 1 and 2 and there was no difference between trainers who set standards and those who did not.

6.5 Diagnosis

6.5.1 Descriptive statistics

A diagnosis or formulation of the problem was given on 98 percent of enhancement forms (Table 4.4). The most common diagnoses for each condition are given in Table 6.9. There was considerable overlap between the two respiratory conditions, with cough (not otherwise specified), asthma and upper respiratory tract infection accounting for 43 percent of all diagnoses for children consulting with acute cough and 57 percent of those identified for recurrent wheezy chest. For children presenting with eczema, a diagnosis of asthma or hay fever, or any other atopic condition was coded as 'relevant'; this might explain the relatively high proportion of 'other relevant diagnosis' for itchy rash.

A number of diagnoses were felt not to be relevant to the presenting condition. In these cases the child has presented with concomitant illnesses. These diagnoses were most common for bedwetting, suggesting that parents may have used the other problem as an 'entry ticket'.

Table 6.9 Diagnoses most commonly recorded on enhancement forms by study condition

Study condition	Diagnosis	Percentage of children
Acute cough	Cough (nos)	29·2
	Upper respiratory tract infection (nos)	29·2
	Coryza	13·1
	Infective respiratory condition (nos)	9·9
	Asthma	9·5
Acute vomiting	Vomiting	38·2
	Gastro-enteritis (nos)	22·6
	Infective non gastrointestinal condition (nos)	14·2
	Viral gastro-enteritis	10·4
	Upper respiratory tract infection	9·5
Bedwetting	Nocturnal enuresis	33·6
	Enuresis (nos)	32·3
	Non-relevant condition	25·8
	Primary enuresis	12·9
	Urinary tract infection	11·5
Itchy rash	Eczema	42·2
	Non-relevant condition	15·9
	Other relevant condition	12·3
	Itchy rash (nos)	7·2
	Atopic eczema	5·9
Recurrent wheezy chest	Asthma	64·3
	Wheezy chest	13·2
	Upper respiratory tract infection (nos)	12·0
	Chest infection	7·3
	Cough (nos)	7·0

It was felt that setting standards might have two possible effects on the recording of diagnoses. Firstly, the number of diagnoses given might increase. Secondly, standard setting might lead to greater diagnostic precision, for example the use of the precise term

Figure 6.8 Mean number of diagnoses for each condition for each doctor: frequency distribution

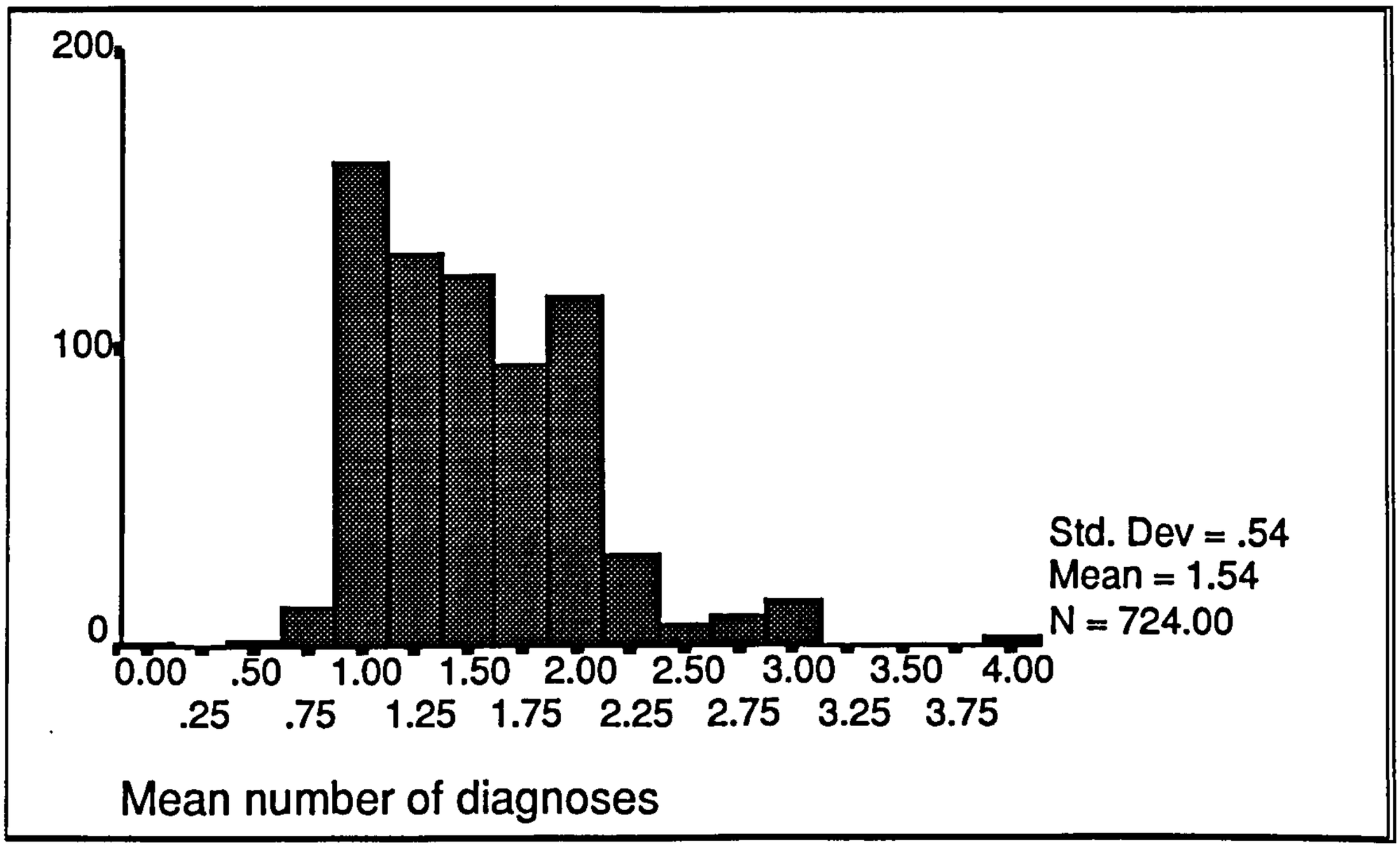
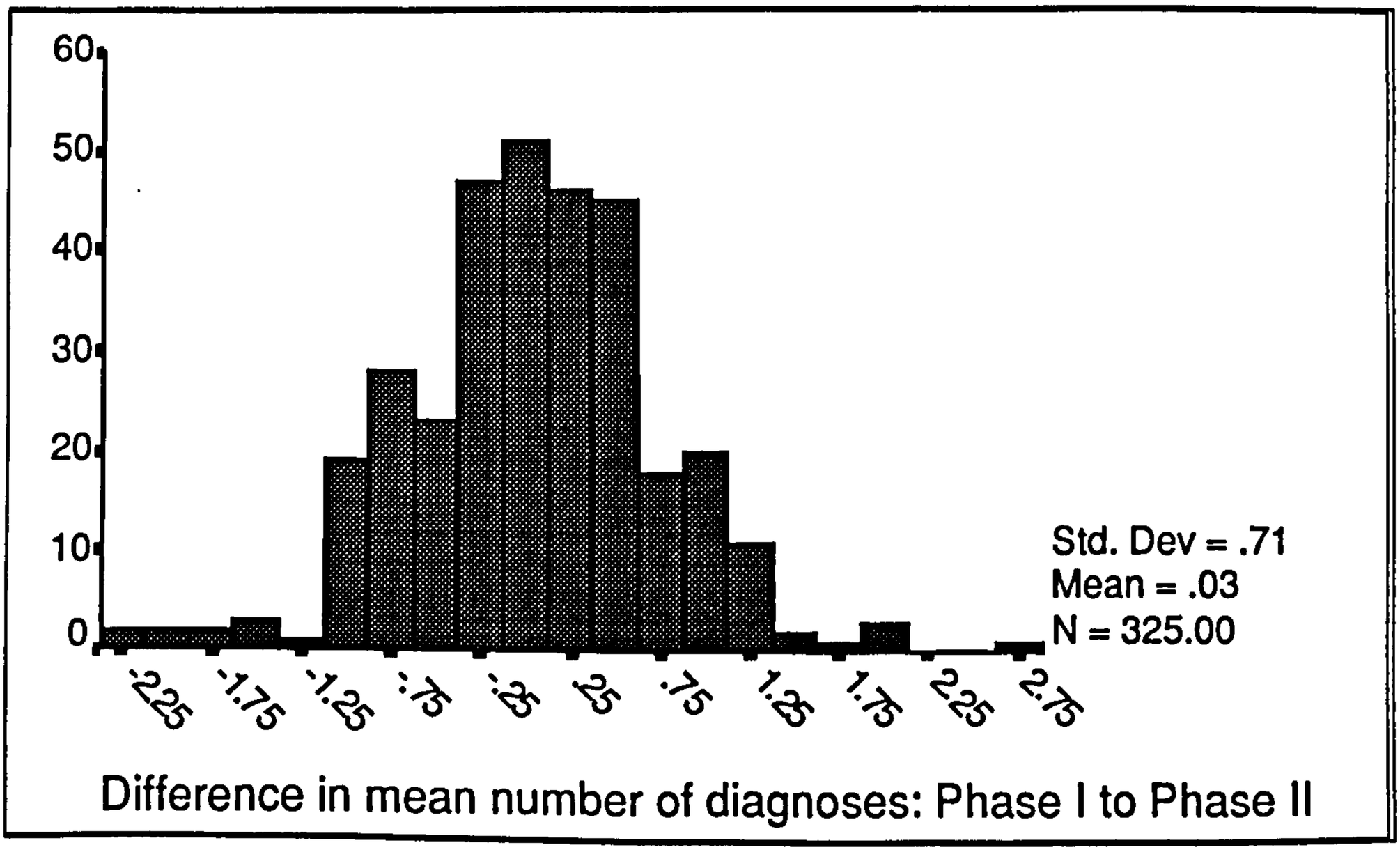


Figure 6.9 Difference between the mean number of diagnoses in phases 1 and 2: frequency distribution



'asthma' rather than the more vague 'wheezy chest'. Both hypothesised effects were investigated.

6.5.2 Number of diagnoses

The distributions of (i) the mean number of diagnoses and (ii) the difference in means between phases 1 and 2 for particular doctors for each condition are given in Figures 6.8 and 6.9.

A repeated measures analysis and an analysis of differences were undertaken assuming a Normal error structure. (Table 6.10). The results of the two analyses were consistent.

Table 6.10 Mean number of diagnoses: model selection

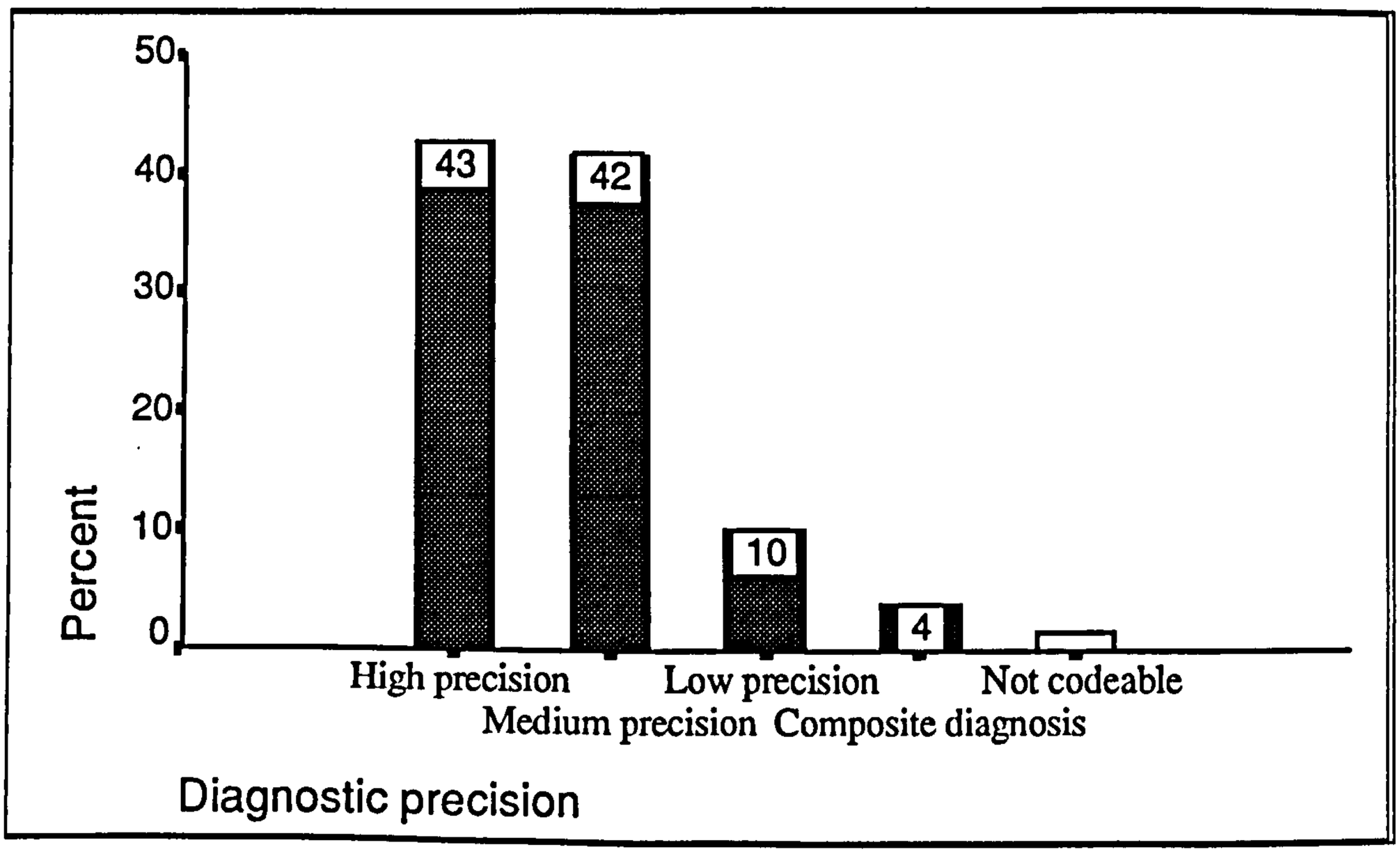
Model		Residual sum of squares	Residual degrees of freedom	Change in sum of squares	Change in degrees of freedom	Mean square	p
Number	Specification						
	Repeated measures analysis						
1	GM	210·8	723				
2	GM + COND	207·9	719	2·9	4	0·73	0·04
3	GM + COND + DOCT	170·9	636	37·0	83	0·45	<0·001
4	GM + COND + DOCT + PSTD	170·7	635	0·1	1	0·14	0·54
5	GM + COND + DOCT + PHAS	170·3	635	0·6	1	0·56	0·14
6	GM + COND + DOCT + PHAS + COND·PHAS	169·2	631	1·1	4	0·28	0·39
	Analysis of differences						
7	GM	163·9	324				
8	GM + COND	159·4	320	4·5	4	1·12	0·06
9	GM + DOCT	121·3	244	42·6	80	0·53	0·34
10	GM + STND	162·8	323	1·1	1	1·1	0·15

There were differences in the mean number of diagnoses recorded for each condition (model 2) but these differences were the same in both phases (model 6 in the first analysis; model 8 in the second). Variation between phases was not significant (model 5) and there were no observed effects of standard setting (models 4 and 10).

6.5.3 Diagnostic precision

Diagnoses were classified into four categories - *high precision* (for example, asthma; salmonella; atopic eczema; and primary enuresis), *medium precision* (for example, bronchitis; gastro-enteritis; eczema; and enuresis), *low precision* (for example, cough; gastritis; itchy rash; and incontinence) and *composite* (for example, infective respiratory condition; gastrointestinal condition; bullous disorders; and psychological disorder or behaviour problem). Most diagnoses recorded on enhancement forms were either high or medium precision (Figure 6.10).

Figure 6.10 Precision of diagnoses recorded on enhancement forms



The distribution of responses suggested that the analysis should focus on whether or not a diagnosis from the "high precision" category was recorded for a consultation. Logistic regression indicated significant variation between conditions; high precision diagnoses were given most frequently to children consulting for recurrent wheezy chest—a function of the high proportion of diagnoses of 'asthma' in this group. There was significant variation between doctors and a significant age effect with more precise diagnoses being given for children aged two and over. But there was no significant change in the proportion of high precision diagnoses between the two phases; this finding was true of both doctors who set standards and doctors who did not.

6.6 Effects of standard setting

The effects of standard-setting on the diagnosis of the current episode are summarised in Table 6.9. Three alternative methods were used to analyse the amount of current medical history recorded on enhancement forms (Section 6.1). Each analysis was based on a slightly different number of observations, which is the main reason why the width of the three confidence intervals given in Table 6.9 is not the same for each method. All observations were included in the first analysis (an analysis of counts assuming a poisson error structure). The 95% confidence interval for the increase in the number of items or current medical history per enhancement form due to standard setting was between -0·35 and +0·20. The third method was based on fewest observations and gave the largest confidence interval (between -0·46 and +0·61 items of history) for the effects of standard setting. As one would expect there is considerable overlap between the three confidence intervals—they are each based on the same data set. The mid point of each interval is not the same but one would not necessarily expect this to be the case. In the first analysis equal weight was given to each enhancement form so effectively most weight was given to those doctors who enhanced the most records. In the other two analyses

doctors were given equal weight provided that they had enhanced records in both phases of the study for a particular condition.

The width of the confidence intervals for the number of items relating to examination findings were similar to those for items of current medical history. The width of the confidence intervals for the number of diagnoses recorded was slightly smaller, perhaps reflecting the smaller variability in this variable—there was greater consistency among doctors in the recording of diagnoses.

Table 6.11 Effects of standard setting on the recording of diagnosis of the current episode on enhancement forms

Variable and method of analysis	95% confidence interval for change attributable to standard-setting
Current medical history	
Number of items per enhancement form	
analysis of counts (Section 6.2.2)	-0·35 to +0·20
Number of items per enhancement form per doctor	
analysis of repeated measures (Section 6.2.3)	-0·58 to +0·26
analysis of cell differences (Section 6.2.4)	-0·46 to +0·61
Examination findings	
Number of items per enhancement form per doctor	
analysis of repeated measures (Section 6.3)	-0·39 to +0·46
analysis of cell differences (Section 6.3)	-0·19 to +0·87
Investigations	
Was investigation recorded?	
analysis of a binary variable (Section 6.4)	0·77* to 1.83*
Diagnoses	
Number of items per enhancement form per doctor	
analysis of repeated measures (Section 6.5.2)	-0·11 to 0·21
analysis of cell differences (Section 6.5.2)	-0·37 to 0·09
Diagnostic precision	
Was diagnosis given with high precision?	
analysis of a binary variable (Section 6.5.3)	0·75* to 1·32*

* For binary variables, the results take the form of an odds ratio - the odds of a successful outcome for a standard-setter divided by the odds of a successful outcome for non-standard-setters.

Confidence intervals for the two binary variables seem fairly large. In percentage terms they represent a change in recording of investigations by between -3·2 and +10·0 percentage points and an increase in the recording diagnoses, at the most precise level, of between -6·9 and +6·9 percentage points.

It is a matter of clinical judgement whether any of the confidence intervals given in Table 6.11 include changes in behaviour that might be regarded as clinically significant.

Chapter 7

Process of care 4: non-drug management

7.1 Introduction

This chapter contains a report of the investigation into how standard setting affected non-drug management. The categories under which non-drug management was analysed are given in Table 7.1. These categories correspond to those developed by the study team (with specialist support) during the process of data abstraction. A breakdown of the proportion of enhancement forms containing each type of decision by study condition is included in the table. Clinicians predicted that, for each condition, the setting of clinical standards would lead to an increase in the recording of such decisions. There was no compelling reason to fit a condition specific standard setting effect in preference to a uniform effect across all five study conditions. (Although the predicted change in behaviour is in a specified direction it was decided that all statistical tests should be two sided. The information about the direction of change was not used to modify the tests undertaken.)

This chapter follows the format of the two previous chapters. Analyses are given in full only when there are interesting statistical points to be considered. Again the term repeated measures analysis does not imply any use of multivariate tests. A summary of the effects of standard setting is provided at the end of the chapter.

Table 7.1 Percentage of enhancement forms containing specific items of non-drug management by study condition

Recorded management decision	Study condition				
	Acute cough (n=748)	Acute vomiting (n=578)	Bed-wetting (n=217)	Itchy rash (n=527)	Wheezy chest (n=644)
Advice, information and explanation					
General items	10·0	49·3	54·4	19·5	6·8
Condition specific items	60·6	64·0	46·5	34·5	32·1
Any type of item	63·9	86·3	71·4	46·1	35·6
Doctor action					
Any action	2·7	1·9	9·2	6·5	7·0
Decision relating to follow up					
Definite follow up	19·8	20·4	47·9	18·4	42·2
No definite follow up	29·1	41·7	3·7	18·8	16·6
Any follow up decision	47·6	60·7	51·2	37·0	57·8
Decision relating to referral					
Definite referral	3·1	4·7	16·6	4·0	2·8
No definite referral	1·2	1·9	6·5	2·7	1·7
Any referral decision	4·3	6·4	22·1	6·3	4·5

7.2 Advice, information and explanation

Items relating to advice, information and explanation were categorised as either being specific to the presenting condition or being of a general nature. Three variables were considered—the number of general items; the number of condition specific items and the total number of items. General advice or explanation were more likely to be recorded for the two acute conditions. Doctors often recorded instructions on cooling and on controlling food and fluid intake and explained the nature of the condition and its management. Explanation of the prevalence of bedwetting and its prognosis were also relatively common. Condition specific advice and explanations were more likely to be

recorded for acute vomiting where advice on dietary restriction was frequently given; and for bedwetting, where doctors frequently recommended 'lifting' the child and using star charts or buzzer alarms. The results of the analyses of the three variables were very similar. The only analysis that is presented in detail is that of the total number of items (Table 7.2).

The presence of an item of advice was analysed as a binary variable. The mean number of items of advice was also modelled using the two methods described earlier. The three analyses yielded consistent results. There was significant variation between conditions (models 2 and 9) but the way in which they differed was the same in each phase (models 12 and 15). There was significant variation between doctors (models 3 and 10) and there was a slight drop in the number of items recorded between phases 1 and 2 (models 5 and 11). But there was no evidence of any effects of standard setting on the recording of items of advice and explanation (models 6, 13 and 18).

It is reassuring that, although there were slightly different assumptions made in each case, the results of all three analyses were consistent. If the results concerning the factors influencing the recording of items of advice had been dependent on the method of analysis, one would not have had very much confidence in those results. That the analyses give consistent results might be regarded as evidence of robustness.

Table 7.2 Items of advice: model selection

Model		Residual scaled deviance*	degrees of freedom	Change in scaled deviance*	Change in degrees of freedom	Mean scaled deviance*	p
No	Specification						
	Binary analysis: presence of an item of advice						
1	GM	3662	2706				
2	GM + COND	3256	2702	405·4	4	101·4	<0·001
3	GM + COND + DOCT	2833	2619	423·4	83	5·1	<0·001
4	GM + COND + DOCT + PSTD	2830	2618	2·8	1	2·8	0·11
5	GM + COND + DOCT + PHAS	2828	2618	5·1	1	5·1	0·03
6	GM + COND + DOCT + PHAS + PSTD	2327	2617	0·6	1	0·6	0·41
7	GM + COND + DOCT + PHAS + AGE	2817	2617	10·4	1	10·4	0·002
	Repeated measures analysis: mean number of items of advice						
8	GM	865·4	723				
9	GM + COND	655·0	719	210·4	4	52·6	<0·001
10	GM + COND + DOCT	403·5	636	251·5	83	4·8	<0·001
11	GM + COND + DOCT + PHAS	399·0	635	4·5	1	4·5	0·007
12	GM + COND + DOCT + PHAS + COND·PHAS	398·0	631	0·9	4	0·2	0·84
13	GM + COND + DOCT + PHAS + PSTD	397·9	634	1·1	1	1·1	0·19
	Analysis of differences						
14	GM	349·8	324				
15	GM + COND	346·25	320	3·5	4	0·8	0·52
16	GM + STND	349·6	323	0·2	1	0·2	0·67
17	GM + DOCT	239·3	244	110·5	80	1·4	0·03
18	GM + DOCT + STND	238·8	243	0·5	1	0·5	0·48

* for the binary dependent variable, the scale parameter was 1 (the scaled deviance is simple the deviance as defined in Chapter 2); for the other two analyses, which assumed a Normal error distribution, the residual scaled deviance is the residual sum of squares

7.3 Doctor actions

Information on actions by the doctors was recorded on 4·8 percent of enhancement forms. Again the presence of such information was analysed as a binary variable. There was significant variation between doctors and significant variation between conditions. Information was recorded more frequently for the chronic conditions. Actions included pricking and cauterising lesions for itchy rash, writing to social services for a laundry allowance for those consulting with bedwetting, and testing inhaler technique for children with recurrent wheezy chest. There was also an age effect—more information was recorded for older children. There was no difference between phases 1 and 2 and there was no difference between doctors who set standards and those who did not.

The reason why more doctor actions were recorded for older children is unclear. One possible reason is that older children are better able to communicate with the doctor than young children and are thus able to make the doctor aware of any functional limitations (either at home or at school) that arise as a result of their condition. This might result in more actions being taken by the doctor.

7.4 Decisions to follow up

Decisions relating to follow up were split into two types: those where the doctor made plans to see the child again; and those where the doctor recorded a decision not to see the child again. Each was analysed separately. There was an emphasis of the importance of follow up throughout most of the standards. It was therefore expected that the number of decisions to review the child would increase and the number of decisions to discharge the child would decrease.

7.4.1 Decisions to review

There was significant variation between study conditions. Definite plans for reviews were more likely to be recorded for bedwetting and recurrent wheezy chest (Table 7.1). These are conditions where it may be necessary to monitor the child's progress and response to management over a protracted period. There was a difference between enhancement forms and statutory records—decisions to follow up were recorded on 27 percent of enhancement forms but only 10 percent of statutory records. Variation between doctors was significant but there was no difference between phase 1 and phase 2 and there was no evidence that standard setting influenced the review rate.

7.4.2 Decisions to discharge

Decisions to discharge the child or to see the child again only if its condition deteriorated or if there was no response to treatment were recorded on 25 percent of enhancement forms but only three percent of statutory records. Data from the two sets of records were therefore analysed separately (Table 7.3).

There was significant variation between study conditions (model 2). Decisions to discharge were most often recorded for the two acute conditions (Table 7.1). Many children presenting with these conditions may be suffering from self-limiting illnesses which do not warrant further contact with the doctor.

There was significant variation between doctors (model 3) and some evidence of a phase effect (model 4) although the improvement in fit is significant only at the five percent level. If this effect is retained in the model, the improvement obtained by fitting an effect of standard setting (model 5) is not significant. But if the term representing the phase effect is first taken out of the model, the improvement obtained by fitting a standard setting effect is now significant at the one percent level (model 6). The two effects (a

Table 7.3 Decisions to discharge: analysis of enhancement forms—model selection

Model		Residual deviance	degrees of freedom	Change in deviance	Change in degrees of freedom	p
No	Specification					
1	GM	3034	2706			
2	GM + COND	2842	2702	192·5	4	<0·001
3	GM + COND + DOCT	2433	2619	408·7	83	<0·001
4	GM + COND + DOCT + PHAS	2427	2618	6·2	1	0·012
5	GM + COND + DOCT + PHAS + PSTD	2423	2617	3·3	1	0·07
6	GM + COND + DOCT + PSTD	2426	2618	7·3	1	0·007
7	GM + COND + DOCT + PSTD + PHAS	2423	2617	2·2	1	0·14
8	GM + COND + DOCT + PSTD + CPST	2420	2614	4·9	4	0·30
9	GM + COND + DOCT + PSTD + MIXD	2421	2616	3·9	1	0·05
10	GM + COND + DOCT + PSTD + AUDT	2423	2615	2·4	3	0·49
11	GM + COND + DOCT + PSTD + CAUD	2407	2599	18·1	19	0·52
12	GM + COND + DOCT + PHAS + PSTD + MIXD	2419	2916	3.9	1	0·05
13	GM + COND + DOCT + PHAS + PSTD + AUDT	2423	2614	0.4	3	0·94
14	GM + COND + DOCT + PHAS + PSTD + CAUD	2407	2598	16.1	19	0·65

difference between phase 1 and phase 2 and a difference between standard setters and non-standard setters) are clearly confounded. Much of the difference between phases 1 and 2 can be explained by a difference between consultations in phase 2 with a doctor who set a standard and all other consultations. Model 6 better fits the data than model 4 but the two models are not nested and it is difficult to quantify the difference between the two models. The raw data corresponding to the relevant consultations are presented in Table 7.4.

In general, except for consultations for itchy rash, there was a tendency for doctors who set standards to record fewer decisions to discharge patients in phase 2. The tendency for other doctors to record fewer decisions to discharge is less noticeable—there seems to be a slight fall in recorded decisions for the two of the three chronic conditions. This would tend to suggest that most of the phase effect (the reduction in recorded decisions to discharge between phases 1 and 2) is due to the performance of those doctors who set standards. The GLIM models indicate that there may be a genuine standard setting effect but because of the confounding described above, it is difficult to be certain.

Table 7.4 Percentage of enhancement forms on which a decision to discharge was recorded by study condition, phase and whether doctor set standard for that condition

Study condition	Percentage of consultations at which a decision to discharge was recorded (the denominator is given in brackets)					
	All consultations in phase 1		Consultations in phase 2 with a doctor who set a standard for that condition		Consultations in phase 2 with other doctors	
Acute cough	31·9	(n = 389)	24·8	(n = 113)	26·8	(n = 246)
Acute vomiting	42·8	(n = 332)	34·3	(n = 67)	42·5	(n = 179)
Bedwetting	4·3	(n = 163)	0	(n = 12)	2·4	(n = 42)
Itchy rash	18·2	(n = 341)	19·6	(n = 46)	20·0	(n = 140)
Wheezy chest	19·9	(n = 357)	3·9	(n = 77)	15·7	(n = 210)

As there was a potential effect of standard setting, possible condition specific effects of standard setting and effects of the other interventions were considered (as prescribed by the modelling strategy—Chapter 2, Section 2.6).

To fit a condition specific effect of standard setting the term 'CPST' (as defined in Table 2.3) is added to model 6 to generate model 8. The residual deviance falls by 4.9 for the loss of four degrees of freedom (this variable has five degrees of freedom

associated with it but one of them has already been accounted for by the prior inclusion of a general standard setting effect 'PSTD'). The improvement in fit is not significant; any change due to standard setting would appear to be consistent across conditions. Ninety five percent confidence intervals for the effects of standard setting on each condition are given in Table 7.5.

Table 7.5 Ninety five percent confidence intervals for the effect of standard setting on the recording of decisions to discharge by condition

Study condition	Odds ratio	95% confidence interval
Acute cough	0·82	(0·47 to 1·40)
Acute vomit	0·61	(0·32 to 1·05)
Bedwetting	0·03	(0·00 to 131)
Itchy rash	0·78	(0·33 to 1·85)
Wheezy chest	0·24	(0·07 to 0·83)
All conditions	0·71	(0·56 to 0·90)

These results are consistent with the raw data given in Table 7.4. The odds ratios in Table 7.5 however take into account variation between doctors. The confidence interval for bedwetting is very large reflecting the small number of children identified with that condition in phase 2 (none of the 12 identified children had a decision to discharge recorded). The largest effect seems to be for consultations with children with wheezy chest but this may just be due to chance. There was no prior evidence to suggest that setting a standard for wheeze should have a different effect to setting a standard for other conditions so no further hypothesis testing was undertaken.

To investigate the effect of specialist input during the standard setting process the term 'MIXD' is entered into the model. This variable was set to two for consultations in phase 2 with a doctor who had set a standard for that condition and belonged to a group

who had met the appropriate mixed group; and was set to one otherwise. The significance level associated with the improvement in fit (model 9) was around five percent. Parameter estimates suggested that doctors who met mixed groups were more likely to record decisions to discharge (the odds ratio was 2.0 with 95% confidence interval from 0·99 to 4·1). It was certainly not anticipated that meeting a mixed group would have an effect in this direction. In any case, as a significance level of one percent had been specified for testing possible effects of the intervention (Chapter 2, Section 2.6.3), this result was not regarded as significant evidence of a mixed group effect.

Next, the other interventions were considered. The variable 'AUDT' had five levels (Table 2.2). A value of one was assigned to all consultations in phase 1 and to consultations in phase 2 where the doctors had received no feedback at all for that particular condition. Values two, three, four and five were assigned to consultations in phase 2 corresponding to the four aspects of performance review described in Figure 1.2. Fitting this effect (model 10) caused a reduction in the deviance of 2·4 for the loss of three degrees of freedom. (A term representing the effect of standard setting had already been included in the model so one of the four degrees of freedom associated with the term 'AUDT' had already been accounted for.) The improvement was not significant; there was no evidence of any effect of the other interventions on the recording of decisions to discharge. For interest, 95% confidence intervals for the effect of each intervention are given in Table 7.6.

The confidence interval for the effect of standard setting is much narrow than the interval for the other interventions. This reflects the decision to sample proportionally more consultations for the condition for which a doctor set a standard (Chapter 4, Section 4.1). The confidence intervals for the other interventions are remarkably similar. It is possible that the other conditions had an effect on the recording of discharge

decisions but that the study has insufficient power to detect them (the odds ratio of 0·61 for the effect of standard setting falls within the other three confidence intervals). It is a matter of clinical judgement as to whether the confidence intervals in Table 7.6 might include changes in recording that could be regarded as clinically significant.

Table 7.6 Ninety five percent confidence intervals for the effects of the interventions on the recording of decisions to discharge on enhancement forms

Intervention	Odds ratio*	95% confidence interval
Receive comparative data from own trainer group	0·83	(0·55 to 1·24)
Receive comparative data from all participating doctors	0·83	(0·56 to 1·22)
Receive clinical standard from another trainer group	0·81	(0·54 to 1·21)
Set clinical standard	0·61	(0·43 to 0·.86)

* Odds ratios are based on the model GM + COND + DR + AUDT. The odds making up the denominator correspond to the recording of discharge decisions for all consultations in phase 1 and consultations in phase 2 with doctors who received *no* intervention for that condition.

The effect 'CAUD' (defined in Table 2.4) was fitted to allow the magnitude and direction of the effect to vary from condition to condition (Table 7.3, model 11). The improvement in fit was not significant. Finally, the effects of the other interventions were considered after adjusting for differences between phases (as there was some evidence from model 4 that phase might be an important variable). The improvement obtained by fitting the intervention terms in models 12, 13 and 14 was comparable to that obtained when the same terms were added without adjusting for phase in models 9, 10 and 11.

The results of the analysis of enhancement forms are slightly ambiguous. There is an apparent effect of standard setting but it is possible that the effect is simply the result of a general change between the two data collection phases. It is possible to look at statutory

records to see whether the same models adequately explain the observed data. A breakdown of the proportion of statutory records which contained a record relating to a decision to discharge is given in Table 7.7. The main problem with this data is that for all conditions, except acute vomiting, very few such decisions were recorded. Although the data may be consistent with a reduction in recording of decisions to discharge that is due to standard setting, the numbers of cases involved are too small to provide good evidence of this.

Table 7.7 Percentage of statutory records on which a decision to discharge was recorded by study condition, phase and whether doctor set standard for that condition

Study condition	Proportion of consultations at which a decision to discharge was recorded (the denominator is given in brackets)					
	All consultations in phase 1		Consultations in phase 2 with a doctor who set a standard for that condition		Consultations in phase 2 with other doctors	
Acute cough	4·9	(n = 82)	0	(n = 25)	2·0	(n = 51)
Acute vomiting	10·8	(n = 65)	11·1	(n = 9)	14·3	(n = 28)
Bedwetting	0	(n = 34)	0	(n = 24)	1·5	(n = 66)
Itchy rash	1·3	(n = 80)	0	(n = 38)	3·5	(n = 85)
Wheezy chest	0	(n = 71)	0	(n = 26)	0	(n = 68)

As a result the model GM + COND (model 2, Table 7.8) fits the data very well. The residual deviance is only 169·2 while there are 745 residual degrees of freedom. Fitting variation between doctors (model 3) results in a reduction of 79·9 in the residual deviance for the loss of 83 degrees of freedom. This improvement is not significant when the reduction is compared with the percentage points of a χ^2_{83} distribution. But in percentage terms the deviance has been reduced by 47 percent by including variation between doctors and there is a reasonable argument for retaining variation between doctors in the model. The improvements obtained by fitting a phase and a standard

Table 7.8 Decisions to discharge: analysis of statutory records—model selection

Model No	Specification	Residual deviance	degrees of freedom	Change in deviance	Change in degrees of freedom	p
1	GM	198·6	749			
2	GM + COND	169·2	745	29·4	4	<0·001
3	GM + COND + DOCT	89·3	662	79·9	83	0·58
4	GM + COND + DOCT + PHAS	88·8	661	0·5	1	0·48
5	GM + COND + DOCT PSTD	86·2	661	3·2	1	0·07
6	GM + COND + PHAS	169·2	744	<0·01	1	0·95
7	GM + COND + PSTD	167·4	744	1·8	1	0·18

setting effect are investigated both with and without allowing for variation between doctors (models 4 and 5 and models 6 and 7). In neither case is either of the effects significant; the evidence is inconclusive.

7.5 Decisions to refer

Decisions relating to referral were analysed in a similar way to those relating to follow up. A distinction was made between decisions to refer the child to see someone else and decisions not to refer the child. A definite decision to refer was recorded most often for bedwetters (Table 7.1). Some practices appeared to have a policy of referring bedwetters to a health visitor, continence advisor or paediatrician for management. However the highest proportion of decisions against referral was also for bedwetting. It is possible that doctors felt the need to state explicitly their decision not to refer in the face of pressure from parents or some local policy. The GLIM analyses showed that there were no significant differences between phase 1 and phase 2 and that there were no effects that could be attributed to setting clinical standards.

7.6 Reasons for non-drug management

The study team found that reasons for management decision were the most difficult elements of data to code (North of England Study, 1990b). This was due in part to doctors' uncertainty as to what should be recorded in this section of the enhancement form. Many doctors simply provided additional subjective or objective data on the patient, rather than a justification for their plan of action. Furthermore, it was in this section that the greatest differences in style from doctor to doctor were observed. It was felt that these data were the least reliable of the process data and that a detailed analysis was not warranted.

During a preliminary analysis, reasons for management were grouped into a number of broad categories. The most commonly specified reason for non-drug management (specified on 23 percent of enhancement forms) was to monitor the condition. These were often recorded in conjunction with a decision to follow up. Characteristics of the carer were also important in determining how to manage the condition and were mentioned on 18 percent of enhancement forms. Judgements about the carer's competence and intelligence (for example, 'mother sensible', 'mum of low intelligence' and 'feckless mother') frequently underlay doctors decisions. So too did considerations of the levels of parental anxiety which were given on 11 percent of enhancement forms. Health education, including general statements such as 'encouraging self-help' was cited in eight percent of consultations and more general considerations of whether or not the child was 'unwell' were mentioned in seven percent of consultations.

7.7 Effects of standard setting on the recording of non-drug management

Most of the dependent variables analysed in this chapter were binary—either a particular piece of information was recorded on enhancement forms or not. The mean number of items of advice however was also analysed assuming a normal error structure using a

repeated measures analysis and an analysis of differences (Section 7.2). The parameter estimates of model 13 in Table 7.2 can be used to estimate the effects of standard setting based on the analysis of repeated measures. Ninety five percent confidence intervals for this estimate suggest that standard setting caused a change of between -0·40 and +0·11 in the mean number of items of advice recorded per doctor per enhancement form. A corresponding estimate based on the analysis of differences (using the parameter estimates of model 18, Table 7.2) suggests that standard setting caused a change of between -0·22 and +0·43 in the mean number of items of advice per doctor per enhancement form.

Estimates of the effects of standard setting on all the binary measures of non-drug management are given in Table 7.9. Some of these confidence intervals are fairly wide (particularly those relating to referral decisions) and may well include effects that might be regarded as clinically significant. If so this would indicate that the study as implemented had low power to detect clinically significant changes in binary process variables.

Table 7.9 Effects of standard setting on the recording of non-drug management

Variable	Model	Odds ratio	99% confidence interval
Advice	GM+COND+DOCT+PHAS+AGE+PSTD	0·89	(0·63 to 1·18)
Doctor actions	GM+COND+DOCT+AGE+PSTD	1·28	(0·71 to 2·29)
Decisions to review	GM+COND+DOCT+PHAS+PSTD	1·14	(0·84 to 1·54)
Decisions to discharge	1: GM+COND+DOCT+PSTD	0·64	(0·45 to 0·89)
	2: GM+COND+DOCT+PHAS+PSTD	0·71	(0·49 to 1·03)
Decisions to refer	GM+COND+DOCT+PSTD	0·61	(0·28 to 1·30)
Decisions not to refer	GM+COND+PSTD	1·21	(0·53 to 2.71)

Estimates of the effect of standard setting on the recording of decisions to discharge are based on two different models. This is because the effect of standard setting and the effect of a simple change over time are confounded. In the first model which also allows for differences between conditions and difference between doctors the effect of standard setting appears to be significant. But if we first allow for a global change between phase 1 and phase 2 the effect is no longer significant (the 95% confidence interval corresponding to the second model includes the value "1"). Both models lead to similar estimates of the odds ratio but the confidence interval generated by the second model is a little wider than that generated by the first.

The other interventions were not found to have any effect on the recording of decisions to discharge. (Section 7.4.2).

Chapter 8

Process of care 5: drug management

8.1 Introduction

Information relating to drug management was recorded on 85 percent of enhancement forms and 76 percent of statutory records (Table 4.4). There was considerable variation between conditions in the number of drugs advised or prescribed (Table 8.1)

Table 8.1 Percentage of consultations in which drugs were prescribed or advised by study condition

Number of drugs	Acute cough (n = 906)	Acute vomiting (n = 680)	Bedwetting (n = 341)	Itchy rash (n = 730)	Wheezy chest (n = 809)	Total (n = 3466)
0	23·0	42·7	56·9	14·1	6·8	24·6
1	52·8	44·2	41·3	55·1	41·0	47·7
2	20·5	11·5	1·8	20·9	36·0	20·6
3 or more	3·7	1·5	0·0	9·8	16·2	7·2

Children presenting for bedwetting were least likely to be given a prescription or advice about use of over the counter medicines. In contrast, over 93 percent of consultations for recurrent wheezy chest and almost 86 percent of consultations for itchy rash resulted in one or more drugs being prescribed or advised. Three or more drugs were prescribed at many consultations—particularly for recurrent wheezy chest (typically bronchodilators

or oral steroids in conjunction with antibiotics and inhaled steroids), and itchy rash (typically drugs for treating rash in combination with others for symptom relief).

Clinicians on the study team decided that it would be appropriate to categorise drugs under the headings given in Table 8.2. Antibiotics were prescribed for over a quarter of all children. They were administered most frequently to those presenting with a respiratory condition, and were given least often to children suffering from itchy rash, perhaps reflecting either a paucity of infective causes or a failure of doctors to identify such cases.

Table 8.2 Percentage of consultations in which drugs were prescribed by type of drugs and study condition

Type of drugs	Acute cough (n = 906)	Acute vomiting (n = 680)	Bedwetting (n = 341)	Itchy rash (n = 730)	Wheezy chest (n = 808*)
Antibiotics	40·8	23·4	10·0	7·7	33·9
Other therapeutic drug	12·8	16·0	31·1	59·7	68·2
Analgesic/antipyretic	21·2	20·9	1·2	2·2	4·7
Other symptom relief drug	22·8	2·9	0·6	37·5	7·7
Prophylactic drug	0·8	0·3	0	0·3	24·0
Miscellaneous preparation	2·9	4·8	1·8	4·8	8·2

Note:
* One case deleted due to missing data

The most frequently prescribed category of drug was that of therapeutic drugs other than antibiotics, given to 38 percent of all children. This type of medication was used most commonly for recurrent wheezy chest, in the form of oral steroids and bronchodilators; these drugs were less frequently prescribed for acute cough. These other therapeutic drugs were used at almost 60 percent of consultations for itchy rash; here they include

topical corticosteroids, antifungals and preparations specific to the treatment of particular types of rash, for example benzyl benzoate. Just over 31 percent of consultations for bedwetting resulted in the prescription of a non-antibiotic drug with therapeutic action; almost all tricyclic antidepressants. Drugs to treat acute vomiting, mainly oral rehydration preparations, were used at 16 percent of consultations with this condition.

Drugs with analgesic and antipyretic properties, such as paracetamol and aspirin were used most often for children presenting with the two acute conditions. Fever is more likely in episodes of acute illness. An analysis of the reasons for using this category of drug shows that relief of pyrexia was the most important. The alleviation of pain and symptom relief were also mentioned quite frequently.

Drugs to relieve other specific symptoms were used most often for acute cough and itchy rash. Over 38 percent of consultations for itchy rash resulted in the prescription or recommendation of symptom relief drugs—mainly emollient and barrier creams, bath additives and antihistamine preparations. Cough linctus and suppressants, and nasal decongestants were suggested more often for acute cough than for recurrent wheezy chest. The use of drugs to relieve symptoms of acute vomiting was unusual, and was mainly confined to antacids and kaolin mixtures.

The use of prophylactic medication was almost entirely confined to recurrent wheezy chest, in the form of drugs like inhaled steroids and beclomethasone. Miscellaneous preparations, mainly dressings, foodstuffs and 'home remedies' such as lemon and honey were advised at almost five percent of consultations.

8.2 Testing for an effect of standard setting

The clinicians on the project team anticipated that standard setting might not influence the prescribing of drugs in exactly the same way for all five study conditions. Reference to the actual standards supported this view. For example, the two standards for acute cough and the two corresponding standards for recurrent wheezy chest cautioned against the indiscriminate prescribing of antibiotics. In contrast the standards for itchy rash advocated the use of antibiotics for infected eczema and impetigo. None of the four standards for acute vomiting or bedwetting gave unequivocal advice about use of antibiotics. In this case it is not appropriate to fit a simple effect—changes in one condition might cancel out changes in another resulting in the overall effect not being significant. It is necessary to fit a separate effect for each condition. This is most easily achieved by fitting the composite term CPST (defined in Table 2.3) after variation between conditions (COND) has already been included. For each condition, estimates are given for the effect of standard setting on that condition.

8.3 Prescription of antibiotics

Whether an antibiotic was described at a consultation was analysed as a binary variable. Preliminary analysis indicated close agreement between information recorded on enhancement forms and information recorded in statutory records. On this basis, it was felt that it ought to be possible to include data from both sources in this analysis. The results of the modelling are reported in Table 8.3.

Fitting the grand mean (model 1) leaves a residual deviance of 3953 with 3462 degrees of freedom. As large differences had been noted between conditions (see above) this effect was incorporated into the model at an early stage (model 2). To check that there were no difference between types of medical record, record type (RCRD) was entered into the model (model 3). The improvement in the fit of the model was not significant.

Record type can therefore be safely omitted from the model; it was felt appropriate to retain data from both enhancement forms and statutory records in the analysis.

Table 8.3 Prescription of antibiotics: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	3953	3462			
2	GM + COND	3615	3458	338·3	4	<0·001
3	GM + COND + RCRD	3615	3457	< 0·1	1	0·82
4	GM + COND + DOCT	3361	3375	254·0	83	<0·001
5	GM + COND + DOCT + PHAS	3360	3374	0·8	1	0·39
6	GM + COND + DOCT + PSTD	3360	3374	0·8	1	0·38
7	GM + COND + DOCT + CPST	3343	3370	18·6	5	0·002
8	GM + COND + DOCT + PHAS + COND·PHAS	3353	3370			
9	GM + COND + DOCT + PHAS + COND·PHAS + CPST	3335	3365	18.0	5	0·003
10	GM + COND + DOCT + CPST + MIXD	3341	3369	1·6	1	0·20
11	GM + COND + DOCT + CPST + CMIX	3339	3365	4·0	5	0·55
12	GM + COND + DOCT + CPST + CAUD	3318	3355	24·4	15	0·06

The effect of differences between doctors (model 4) was highly significant. There was no overall change in prescribing behaviour between phases 1 and 2 (model 5). There was no overall change that could be attributed to standard setting (model 6) but when, as described above, a different change for each condition is included (model 7) the improvement in fit is significant at the one percent level. This would indicate that the size and magnitude of the effects of standard setting varies from condition to condition.

In the same way that a change due to setting a standard may be confounded with an overall change between phases 1 and 2, a condition specific effect of standard setting may be confounded with condition specific changes between phases 1 and 2. To see

whether the possible effects of standard setting can be attributed to such changes we first fit the condition specific change over time (model 8) and then add the condition specific effect of standard setting (model 9). The improvement is still significant at the 1% level; the effects of standard setting cannot be explained by condition specific changes over time.

Having established that there is strong evidence that setting clinical standards has had an effect on the way doctors prescribe antibiotics, it is appropriate to go on to consider the other aspects of the intervention. The next step is to fit a mixed group effect (represented by the term MIXD). This variable is set to two for consultations in phase 2 with a doctor who had set a standard for that condition and belonged to a group who had met the appropriate mixed group and is set to one otherwise. The improvement in fit (model 9) was not significant. As the standards recommend different actions for different treatments it is probably more appropriate to consider a condition specific mixed group effect rather than a general one. MIXD was therefore removed from the model and replace by the variable CMIX defined in Table 2.5. The improvement in fit (model 11) was still not significant—there was no evidence of a condition specific effect of specialist input when setting a standard. (This is consistent with the observation that the recommendations were similar in each of the standards.)

Finally, the other interventions (the different levels of feedback) were considered. Just as in the case of standard setting, the clinicians attached to the study team felt that the other interventions would have a different effect on each condition. (They felt that it would be most unlikely that receiving feedback would cause a uniform increase in the prescription of antibiotics across all conditions, for example.) It was therefore decided to fit CAUD (defined in Table 2.4) which represented a condition specific effect of each of the four major interventions (model 12). As a condition specific effect of standard setting had already been incorporated in the model the reduction in residual degrees of freedom was

only 15 (the additional parameters estimated corresponded to the effects of the other three non-trivial interventions for each of the five conditions). The reduction in residual deviance of 24.4 was not significant at the 1% level.

The model that best represented this aspect of drug management was, therefore, GM + COND + DOCT + CPST. This model was used to generate estimates of the effects of standard setting on the prescription of antibiotics. The odds ratio of antibiotic prescribing after standard setting controlling for variation between doctors is given for each condition in Table 8.4. The results are consistent with the recommendations given in the clinical standards which are described above. The prescription of antibiotics was increased for itchy rash and reduced for acute cough. There was also some evidence of a reduction in their prescription for wheezy chest; the level of significance associated with this reduction was just over five percent.

Table 8.4 The effect of standard setting on the prescription of antibiotics for each study condition

Study condition	Consultations after doctor had set standard for study condition	All other consultations (those before doctor had set standard for study condition and those with doctors who had set standards for one of the other four study conditions)	Odds ratio of antibiotic prescribing after standard setting	Odds ratio (and 95% confidence interval) of antibiotic prescribing after standard setting controlling for variation between doctors
Acute cough	27·2% (n=138)	41·9% (n=767)	0·76	0·62 (0·40 to 0·96)
Acute vomiting	28·9% (n=76)	22·7% (n=604)	1·38	1·22 (0·67 to 2·21)
Bedwetting	11·1% (n=36)	9·8% (n=305)	1·15	1·23 (0·38 to 3·93)
Itchy rash	16·9% (n=83)	6·5% (n=646)	2·92	3·41 (1·69 to 6·87)
Wheezy chest	23·3 % (n=103)	35·5% (n=705)	0·55	0·61 (0·35 to 1·05)

8.4 Prescription of other therapeutic drugs

Of the 15 standards (10 set by trainer groups and five by mixed groups), all six for respiratory conditions (acute cough and recurrent wheezy chest) advocated the use of bronchodilators for children with wheeze or persistent cough. The three standards for acute vomiting advocated use of oral rehydration fluids. The three standards for bedwetting cautioned against the prescribing of tricyclic antidepressants except as a last resort. Finally the three standards for itchy rash advocated drugs such as benzyl benzoate for scabies but cautioned against indiscriminate prescribing of steroids for mild eczema. As in the prescribing of antibiotics, a general effect of standard setting across all conditions is unlikely. Any effects are likely to be specific to a particular condition. It is therefore appropriate to fit a different effect for each condition. The analysis of whether this type of drug was prescribed is presented in Table 8.5.

Table 8.5 Prescription of other therapeutic drugs: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	4600	3462			
2	GM + COND	3708	3458	892·2	4	<0·001
3	GM + COND + RCRD	3705	3457	3·4	1	0·07
4	GM + COND + PHAS	3687	3457	21·7	1	<0·001
5	GM + COND + PHAS + CPST	3675	3452	11·8	5	0·04
6	GM + COND + DOCT	3556	3375	151·9	83	<0·001
7	GM + COND + DOCT + PHAS	3532	3374	23·8	1	<0·001
8	GM + COND + DOCT + PHAS + COND·PHAS	3529	3370	2·6	4	0·62
9	GM + COND + DOCT + PHAS + CPST	3522	3369	10·1	5	0·07
10	GM + COND + DOCT + PHAS + CMIX	3513	3364	9.6	5	0.09
11	GM + COND + DOCT + PHAS + CAUD	3508	3354	13.9	15	0.53

As noted above (Table 8.2) the prescription of other therapeutic drugs varied from condition to condition (model 2). The difference between enhancement forms and

statutory records was not significant (model 3)—the analysis of the combined data set is therefore reported here. There was a significant difference between phases (models 4 and 7)—more of these other therapeutic drugs were prescribed or advised in phase 2 than in phase 1. This difference seemed to be consistent across all five conditions (model 8). There was significant variation among doctors (model 6).

When a separate effect of standard setting was fitted for each condition, the level of significance associated with the effect depended upon whether or not variation between doctors was included in the model. Because not all doctors managed to identify their full quota of children for each condition the effects of standard setting and variation between doctors are confounded. Whichever effect is fitted first accounts for some of the variation due to the other. The significance level was just under five percent before fitting variation between doctors (model 5) but increased to just over five percent after (model 9). The parameter estimates associated with this effect are given in Table 8.6.

Table 8.6 GLIM estimates of the effect of standard setting on the prescription of other therapeutic drugs

	Before fitting variation between doctors		After fitting variation between doctors	
Study condition	**log (odds)**	**standard error**	**log (odds)**	**standard error**
Acute cough	0·10	0·26	0·22	0·29
Acute vomiting	0·58	0·29	0·36	0·31
Bedwetting	-1·12	0·47	-1·28	0·49
Itchy rash	0·02	0·25	0·08	0·27
Wheezy chest	0·21	0·25	0·10	0·27

One of the main features of the table is the fairly large standard errors associated with each of the estimates. This is primarily a consequence of the fairly small sample sizes involved. The raw data and the number of cases in each group are given in Table 8.7.

Table 8.7 Percentage of consultations in which other therapeutic drugs were prescribed before and after standard setting by study condition

Study condition	Before standard setting: All doctors	After standard setting: Doctors who set standard for that condition	After standard setting: Other doctors	Odds ratio (with 95% confidence interval) of therapeutic prescription after standard setting adjusting for variation between doctors
Acute cough	10·4 (n = 471)	15·9 (n = 138)	15·2 (n = 296)	1·24 (0·71 to 2·18)
Acute vomiting	12·3 (n = 397)	27·6 (n = 76)	18·8 (n = 207)	1·44 (0·78 to 2·65)
Bedwetting	30·5 (n = 197)	16·7 (n = 36)	37·0 (n = 108)	0·28 (0·11 to 0·73)
Itchy rash	56·5 (n = 421)	65·1 (n = 83)	63·6 (n = 225)	1·08 (0·63 to 1·86)
Wheezy chest	64·3 (n = 428)	75·7 (n = 103)	71·5 (n = 277)	1·11 (0·64 to 1·91)

The evidence is inconclusive. There have been substantial changes between phases 1 and 2 in the recording of the prescription of other therapeutic drugs. The odds ratio of prescribing other therapeutic drugs in phase 2, adjusting for variation between doctors and variation between conditions was 1·49 with 95 percent confidence interval, 1·24 to 1·79. But the difference between doctors who set standards and those who did not was significant only for consultations for bedwetting; doctors who had set standards for bedwetting reduced their prescribing of other therapeutic drugs by a much greater amount than other doctors.

It has been argued (North of England Study, 1992a) that the test of a standard setting effect described above, is overly conservative. A separate effect has been fitted for each condition without taking into account the expected direction of change. A test in which a consistent proportional change in the recording of therapeutic prescribing is modelled was proposed. The test consisted of fitting a change in the log odds (for doctors who set a standard in phase 2) of a constant multiplied by: +1 for acute cough, acute vomiting and recurrent wheezy chest; 0 for itchy rash and -1 for bedwetting. When a term representing this change is included in the model, the improvement in fit is significant at the one percent level if variation between doctors is not included but just over one

percent if the variation between doctors is already included. But this model performs no better than one in which we fit a change for doctors who set a standard for bedwetting only and no change (other than the general change between phases 1 and 2 already included) for the other four conditions. Although the observed changes in prescribing are consistent with the standards that were set, the only clear evidence for a change is for doctors who set a standard for bedwetting.

There is a further reason why the original test of a standard setting effect is likely to be conservative. Examination of the standards for acute cough and recurrent wheezy chest reveals marked similarities between the two conditions. It is therefore probable that setting a standard for one respiratory condition is very likely to have affected the way a doctor treats children consulting for the other. It is quite possible that the treatment of some of the 'controls' has been affected in this way. It would be very difficult to take into account these sorts of effects given the design of the study. Implicit in the design was the assumption that such effects would not occur. Similarly it is impossible to determine whether any global change across all four conditions is a general effect arising from the process of setting a standard for the care of children or whether it arises as a result of some general trend in the study population.

Although it the evidence for an effect of standard setting is inconclusive, it was felt appropriate to consider the effects of the other interventions. It was thought that any such effects would be specific to each condition. There was no effect of meeting a mixed group (Table 8.5, model 10) or of the other types of medical audit (model 11).

8.5 Prescription of antipyretic and analgesic drugs

The recording of whether an antipyretic or analgesic drug was advised or prescribed was much more common on enhancement forms than statutory records. This may be because

for many of the drugs in these categories no prescription is necessary; they are available over the counter. There was significant variation between study conditions (as is clear from Table 8.2) and significant variation between doctors. There was no significant change from phase 1 to phase 2. Nor was there any difference in the behaviour of those doctors who set standards and those who did not.

8.6 Prescription of other symptom relief drugs

There was a significant reduction from phase 1 to phase 2 in the proportion of consultations at which other symptom relief drugs were advised or prescribed. This decrease was observed for all five conditions and was consistent with the advice, given in many of the standards, that indiscriminate use of this type of drug should be avoided. But the decrease in prescribing rates by standard setters did not differ significantly from that for non standard setters. The change in behaviour cannot therefore be attributed to the intervention.

8.7 Prescription of prophylactic drugs

The prescribing of prophylactic drugs was restricted mostly to children with recurrent wheezy chest. The agreement between enhancement forms and statutory records in the recording of these drugs was good. Variation between doctors was not significant and there was no change between phase 1 and phase 2. There was no difference between doctors who set standards and those who did not.

8.8 Effects of standard setting

The effects of setting-standards on the use of antibiotics and other therapeutic drugs has been covered in detail in Sections 8.3 and 8.4. Ninety five percent confidence intervals

for the effects of standard setting on the use of the remaining three categories of drugs are given in Table 8.8.

Table 8.8 Ninety five percent confidence for the effects of setting clinical standards on the recorded use of antipyretic and analgesic drugs, other symptom relief drugs and prophylactic drugs by study condition

Type of drug	Cough	Vomit	Bedwetting	Itchy rash	Wheezy chest
Antipyretic and analgesic	0·63 to 1·83	0·30 to 1·53	1·58 to 115	0·07 to 5·05	0·22 to 2·90
Other symptom relief	0·39 to 1·34	0·00 to 97·3	0·00 to 7×10^{8}	0·57 to 3·04	0·14 to 1·30
Prophylactic	0·93 to 19·1	0·00 to 4×10^{9}	0·00 to 2×10^{17}	0·00 to 1×10^{9}	0·67 to 1·74

Some of these confidence intervals for the odds of prescribing after standard setting divided by the odds of prescribing before standard setting are very wide. In general the widest intervals correspond to the cells in Table 8.2 where the level of prescribing of the drug for the condition corresponding to that cell was very small. These confidence intervals are therefore based on a relatively small amount of information.

Chapter 9

Outcome of care 1: clinical outcome—evidence from interviews with parents

9.1 Introduction

Data on outcome of care were collected from parents of children suffering from the study conditions by two means—face to face interviews conducted by members of the study team and postal questionnaire (Chapter 1, section 1.4.3). The analyses of data arising from interviews are reported in this chapter (clinical outcome) and the next (parents' satisfaction with the care their child received). The analysis of the data derived from the postal questionnaires is reported in Chapter 11.

As described in Chapter 1, interviews were used for only three of the conditions—acute cough bedwetting and recurrent wheezy chest. The interview schedules were very detailed and typically took between 30 and 60 minutes to administer (mean length = 41 minutes). Many of the questions were highly specific to the nature of the condition. For example, for recurrent wheezy chest there were questions that related to seasonal triggers of asthma and the bedwetting schedule included questions relating to laundry costs incurred because of the child's condition. For the purpose of assessing the effects of the intervention it was decided to concentrate on measures of outcome that took approximately the same form in each schedule. Examination of the schedules revealed

only two such measures—clinical outcome and parents' satisfaction with the care that their child received.

Analysis of the process data sets has indicated that any effects of standard setting on doctors' behaviour tended to be condition specific. For example, the way in which doctors changed their prescribing of antibiotics was not the same for each condition. For this reason it was decided that it would be appropriate to look for condition specific effects of standard setting on clinical outcome.

9.2 The sample

During the initial rounds of interviews in each phase (approximately ten weeks after each prevalence survey—subphases 1A and 2A) field-workers conducted a total of 1791 interviews—454 for acute cough, 616 for bedwetting and 721 for recurrent wheezy chest (Table 9.1). These responses represented 82 percent of parents sampled and 92 percent of those who were actually contacted. Of those patients interviewed for the two chronic conditions, 82 percent were interviewed again 12 months later (subphases 1B and 2B) yielding a total of 2888 interviews.

A full analysis of response rates is given in the final report (North of England Study, 1990b). The main findings were that there were no significant differences in response rates between phases 1 and 2, no differences between study conditions and no differences between interviewers. Failure to achieve an interview was largely due to interviewers' inability to contact parents rather than parents' unwillingness to co-operate with the study.

Table 9.1 Number of interviews conducted by study condition

		Study condition			
Phase	**Subphase**	**Acute cough**	**Bedwetting**	**Wheezy chest**	**Total**
1	A	237	321	385	943
1	B	-	266	314	580
2	A	217	295	336	848
2	B	-	229	288	517
Total		454	1111	1323	2888

In most of the analyses reported in this chapter, for children suffering from a chronic condition, only matched responses from parents who were interviewed in both the A and B subphases were considered (a total of 2648 interviews [237 + 217 + 2×(266 + 229 + 314 + 288)] corresponding to 1551 children [237 + 217 + 266 + 229 + 314 + 288]). A small number of children (about five percent of those suffering from bedwetting and recurrent wheezy chest) were sampled both before and after standard setting (that is, in both phases 1 and 2). Otherwise the sample contained different children. Because of the small number of cases involved this source of repeated measures was ignored.

9.3 Assessment of outcome

Clinical outcome was assessed by the extent to which the child was reported to have recovered or was not troubled by his or her condition at the time of the interview. Two measures of clinical outcome were available. The first was a single, summarising, categorical measure for each condition; the second was a potentially more sensitive graded measure that was developed for the two chronic conditions.

9.4 Categorical measure of clinical outcome

9.4.1 Descriptive statistics

The first assessment of clinical outcome was based on yes/no questions which asked about the condition of the child at the time of the interview or (in the case of recurrent wheezy chest) the condition of the child in the month prior to the interview. The particular questions asked are given in Table 9.2. For two of the conditions, acute cough and bedwetting there was just a single question of the form, "Does your child still have the condition now?" For recurrent wheezy chest there were four different questions relating to clinical outcome. It was desired to form a single measure of outcome from these questions that could be included in an analysis of the effects of standard setting along with data from the other two conditions. The clinical members of the study team recommended that a composite measure should be formed based on just three of the questions relating to clinical outcome. A successful outcome was recorded if the parents had responded no to the three questions relating to waking at night, breathlessness and wheezing. A yes response to any of these questions was regarded as an unsuccessful outcome.

There are one or two points of interest relating to the data presented in Table 9.2. For each of the chronic conditions there is a large difference between the number of successful responses in the A and B subphases. This is as expected; some of the children will be 'growing out' of the condition and there may be some improvement due to the treatment of the condition. The second point of note is that, for children with recurrent wheezy chest, there appear to be fewer successful responses in subphase 2A than in subphase 1A. This might indicate that those children identified in phase 2 had more severe wheeze than those identified in phase 1 or that there may be some environmental factor such as climate that has caused the observed difference.

Table 9.2 Categorical measure of clinical outcome: frequency distribution of responses (percentage of cases in each category) in each subphase

Condition and measure of outcome	Response	Subphase 1A	1B	2A	2B
Acute cough		(n=237)		(n=217)	
Has X got the cough now?	Yes	77·2		77·4	
	No	21·1		22·1	
	Missing	1·9		0·6	
Bedwetting		(n=321)	(n=266)	(n=295)	(n=229)
Is X still wetting now?	Yes	80·4	59·0	77·3	60·3
	No	18·4	40·6	22·3	39·7
	Missing	1·2	0·4	0·3	0·0
Recurrent wheezy chest		(n=385)	(n=314)	(n=336)	(n=288)
In the last month has X ever been breathless at all?	Yes	30·1	29·8	38·4	32·6
	No	69·6	69·8	61·3	67·4
	Missing	0·3	0·3	0·3	0·0
Has s/he wheezed in the last month?	Yes	39·7	37·8	45·8	36·8
	No	59·7	61·9	53·9	62·8
	Missing	0·5	0·3	0·3	0·3
Has s/he coughed at all in the last month?	Yes	63·6	54·0	69·6	58·2
	No	35·8	45·4	29·5	42·0
	Missing	0·6	0·6	0·9	0·0
During the last month, has your child been woken in the night with his/her chest trouble?	Yes	29·9	28·6	34·5	27·1
	No	68·8	71·4	64·9	72·6
	Missing	1·4	0·0	0·6	0·3
Composite: has child either been breathless, woken at night or wheezed during the last month?	Yes	52·7	51·4	59·8	50·3
	No	45·2	48·3	39·6	49·0
	Missing	2·1	0·3	0·6	0·7

9.4.2 Time elapsed between consultation and interview

Univariable analysis indicated that there were a number of covariates of interest. Some of these were very striking. Children suffering from acute cough were identified when they consulted a doctor participating in the study. This may have been at any time within a four week window. Subsequently parents were interviewed at a time that was convenient for themselves and the interviewer assigned to them. For these reasons there was considerable variation in the length time that elapsed between the consultation and

the interview. The length of this time interval greatly affected the probability that a child was still coughing at the time of the interview (Table 9.3).

Table 9.3 Proportion of children still coughing by length of interval between the consultation and the interview

Variable: time lag between consultation and interview		Number (and proportion, θ) of children no longer coughing		Number (and proportion, 1-θ) of children still coughing		$\log\left(\frac{\theta}{1-\theta}\right)$
Value	Label					Log odds of a success
1	0 - 14 days	48	(53·3%)	42	(46·7%)	0·13
2	15 - 21 days	78	(77·2%)	23	(22·8%)	1·21
3	21 - 30 days	87	(84·5%)	16	(15·5%)	1·72
4	More than 30 days	114	(91·2%)	11	(8·8%)	2·90

Of those children whose parents were interviewed within two weeks of the consultation just under 50 percent were still coughing. The corresponding figure for children whose parents were interviewed more than a month after the consultation was less than ten percent.

The exact length of elapsed time was not available. It was necessary to use the ordinal variable given in Table 9.3. Logistic regression analysis was used to help decide how the variable could be used most appropriately in any analyses (Table 9.4). Three choices were considered: (i) to include elapsed time as a categorical variable with four factors (LAG4); (ii) to treat the variable as interval and then fit a linear trend (LAG1); and (iii) treat the variable as interval, transform it and then fit the transformed variable as a linear effect. A log transformation (LLAG) is the one considered here.

Fitting the four level factor (model 2) yields a reduction in the deviance of 44·9 for the loss of 3 degrees of freedom—a large improvement in fit. But fitting the linear trend

Table 9.4 Proportion of children no longer coughing: fitting elapsed time

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	441·1	418			
2	GM + LAG4	396·2	415	44·9	3	<0·001
3	GM + LAG1	397·7	417	43·4	1	<0·001
4	GM + LAG1 + LAG4	396·2	415	1·5	2	0·22
5	GM + LLAG	396·5	417	44·6	1	<0·001

(model 3) explains nearly all the variation that could be attributed to the full four factor variable. Clearly the deviation from linearity (model 4) was not significant. Finally fitting the log transformed variable (model 5) results in a reduction in deviance of 44·6—just 0·3 less than the improvement obtained by fitting the full four level factor model. The choice of models lies between model 3 and model 5. They are not nested so a formal comparison is not possible but both would probably serve equally well. A fairly arbitrary decision was taken to fit the log transformed variable in full analyses as it produced a slightly larger reduction in deviance than the untransformed variable. In retrospect it may have been better to fit the simpler variable.

For the two chronic conditions, the time lapse between the consultation and interview did not prove to be an important factor. Children do tend to 'grow out' of both bedwetting and asthma but the length of time this involves is large in comparison with the length of elapsed time considered here. The variable LLAG was set to an arbitrary constant value for these children. Provided there is a term representing the difference between cough and the two chronic conditions already included in the model the choice of value of the constant makes no difference to the model fitting process. The actual reduction in deviance obtained when fitting the term LLAG is exactly the same for all values of the constant and is equal to the reduction in deviance obtained when the term is fitted to a reduced data set containing only cough cases. Similarly the value of the

logistic regression coefficient corresponding to LLAG does not depend upon the value of the constant chosen. In GLIM, which uses corner point parameterization, the only parameter that is affected by the choice of constant is that corresponding to the term "GM" (this parameter corresponds to the value of an observation that takes the value 1 for all factors and the value zero for all continuous variables and would be equivalent to the grand mean in a package that generated deviation coefficients). Estimates of the effects of standard setting are not affected. In this analysis the value of LLAG was set to zero for observations corresponding to a child with either bedwetting or wheezy chest.

9.4.3 Children with recurrent wheezy chest with a diagnosis of asthma

Children identified as having recurrent wheezy chest were suffering from a range of medical conditions. Of these, asthma was by far the most common. Nearly 70 percent of parents either believed or suspected that their child might have asthma. These children were much more likely to have either been wheezy, woken at night or been breathless in the month preceding the interview than the remaining children (67 percent compared with 41 percent). All standards set for recurrent wheezy chest included components that were specific to children with asthma (North of England Study, 1990d). It was felt that it would be appropriate to take differences between children with a diagnosis of asthma and other children with wheezy chest into account when analysing the effects of standard setting.

The proportion of children who had (or were suspected of having) asthma increased from 60 percent in phase 1 to 75 percent in phase 2. This increase was significant at the one percent level and was the same for those trainers who set standards and those who did not. It is not clear whether the observed increase was due to environmental factors or was a result of some underlying trend.

9.4.4 Analysis of initial interviews

The results of the logistic regression analysis of initial outcome are described in Table 9.5. This is based on the interviews administered in subphase A (in the case of bedwetting and wheezy chest, only those children for whom a second interview was administered in subphase B were included).

Table 9.5 Binary measure of initial outcome: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	2060·0	1508			
2	GM + COND	1652·2	1506	407·8	2	<0·001
3	GM + COND + LLAG	1607·6	1505	44·6	1	<0·001
4	GM + COND + LLAG + INTV	1595·5	1498	12·1	7	0·10
5	GM + COND + LLAG + PHAS	1605·9	1504	1·7	1	0·19
6	GM + COND + LLAG + PSTD	1606·4	1504	1·2	1	0·28
7	GM + COND + LLAG + CSTD	1606·3	1504	1·2	1	0·27
8	GM + COND + LLAG + ASTH	1572·1	1504	35·5	1	<0·001
9	GM + COND + LLAG + ASTH + TGRP	1559·2	1495	12·9	9	0·17
10	GM + COND + LLAG + ASTH + PRAC	1494·3	1443	77·1	61	0·08
11	GM + COND + LLAG + ASTH + INTV	1558·2	1497	13·9	7	0·05
12	GM + COND + LLAG + ASTH + PHAS	1571·8	1503	0·4	1	0·55
13	GM + COND + LLAG + ASTH + PSTD	1570·9	1503	1·2	1	0·26
14	GM + COND + LLAG + ASTH + CSTD	1570·9	1503	1·2	1	0·27

Fitting the Grand Mean (model 1) leaves a residual deviance of 2060.0 with 1508 degrees of freedom. It is natural to fit the large differences between conditions noted earlier (Table 9.2) at an early stage (model 2). The ratio of the residual deviance to the residual degrees of freedom after this term has been added is 1.1, which is close enough to unity to suggest that overdispersion is not a problem in this analysis. Changes in deviance are therefore compared with the appropriate chi-squared distribution in order to

assess improvement in fit as each term is added. In this case the reduction of 407·8 in the deviance for the loss of only two degrees of freedom represents an improvement in fit which is significant at the 0·1 percent level.

It was noted earlier that, for children with acute cough, the length of time between the consultation and interview affected the probability of a successful outcome. It would appear sensible to allow for this effect early on in the model fitting process. Fitting 'LLAG' (model 3) generates a big improvement in fit. The second covariate that was mentioned earlier was whether or not a child with recurrent wheezy chest was believed to have asthma. The improvement in fit when this term was added (model 8) was significant at the 0·1 percent level. Children with asthma or suspected asthma were less likely to have a successful initial outcome. Because of the lack of orthogonality in the design many of the effects are confounded. Some of the variation that is actually due to a new term might already have been accounted for when fitting a previous term. It was therefore necessary to consider fitting terms in different orders. A number of terms have been fitted before and after allowing for differences between suspected asthmatics and non-asthmatics.

Variation between interviewers was not significant (models 4, and 11). Even though interviewers underwent extensive training to ensure high inter-rater reliability (Streiner and Norman, 1989) it is important to check that there is no systematic differences between them. There was no difference in initial outcome between phases 1 and 2 (models 5, and 12).

Differences between trainer groups were not significant (model 9) and there were no differences between individual practices (model 10). In the analyses reported in the previous chapters relating to the process of care, it was always possible to allow for variation between individual doctors. For the outcome data sets, when there was more

than one trainer in a practice, it was not always possible to match the interview with a particular trainer. Therefore we fit variation between the 62 practices rather than variation between the 84 doctors.

Finally it is important to see whether there are any differences between interviews corresponding to children who consulted with trainers after they had set a clinical standard for a particular condition and all other interviews. Fitting this effect (represented by the term 'PSTD') did not produce a significant improvement in fit (models 6, and 13). For the two chronic conditions, the clinicians on the team did not expect there to be any change in initial clinical outcome due to standard setting. It was felt that the most likely effect of the intervention would be to influence the outcome as a result of changes in the management of the condition over a period of time. A change in final outcome would be much more likely than change in initial outcome. For children with acute cough, it was expected that any improvement in outcome would occur immediately. The parents of these children were therefore only interviewed once (Chapter 1, section 1.4.3). A term 'CSTD', representing an improvement for children in Phase 2 who saw a trainer who has sent a standard for acute cough was therefore fitted to the data (models 7 and 14). The improvement was not significant.

The model that best represents initial clinical outcome is one that allows for differences between conditions, differences between suspected asthmatics and other children with recurrent wheezy chest and, for children with acute cough, the length of time between the consultation and the interview (model 8). The raw data corresponding to this model are given in Table 9.6. The results are consistent with the univariable analyses (sections 9.4.2 and 9.4.3).

Table 9.6 Initial clinical outcome by condition and type of case

Condition	Type of case	Number* and percentage of cases reporting a successful outcome		Number* and percentage of cases reporting an unsuccessful outcome	
Acute Cough	0 -14 days between consultation and interview	48	(53·5)	42	(46·7)
	15-21 days between consultation and interview	78	(77·2)	23	(22·8)
	21-30 days between consultation and interview	87	(84·5)	16	(15·5)
	more than 30 days between consultation and interview	114	(91·2)	11	(8·8)
	all cases	*327*	*(78·0)*	*92*	*(22·0)*
Bedwetting	*all cases*	*66*	*(14·7)*	*384*	*(85·3)*
Recurrent wheezy chest	non-asthmatics	114	(59·1)	79	(40·9)
	suspected asthmatics	135	(33·3)	270	(66·7)
	all cases	*249*	*(41·6)*	*349*	*(58·4)*

* The table only includes cases where there were no missing values for any of the variables included in the logistic regression analysis described above.

9.4.5 Analysis of final interviews

The analysis of final interview is presented in Table 9.7. The analysis is based on the interviews administered in subphase B for the two chronic conditions and interviews administered in subphase A for acute cough.

The three variables that had had most effect on initial outcome were fitted first (models 2 to 4). These were all highly significant—final outcome varied across the three conditions, was associated with the delay between the consultation and interview for children with cough, and was poorer for children who were suspected asthmatics. As in the analysis of initial interviews, there were no differences between trainer groups (model 5) and no differences between practices (model 6). There was no overall change

Table 9.7 Binary measure of final outcome: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	2076·6	1504			
2	GM + COND	1926·1	1502	150·5	2	<0·001
3	GM + COND + LLAG	1881·5	1501	44·6	1	<0·001
4	GM + COND + LLAG + ASTH	1851·8	1500	29·7	1	<0·001
5	GM + COND + LLAG + ASTH + TGRP	1845·7	1491	6·1	9	0·73
6	GM + COND + LLAG + ASTH + PRAC	1785·3	1439	66·5	61	0·29
7	GM + COND + LLAG + ASTH + PHAS	1851·5	1499	0·3	1	0·58
8	GM + COND + LLAG + ASTH + PSTD	1851·8	1499	0·0	1	0·89
9	GM + COND + LLAG + ASTH + CPST	1849·0	1497	2·8	3	0·42

between the two phases of data collection (model 7) and there was no discernible effect of standard setting (models 8 and 9).

9.4.6 Change in outcome

For the two chronic conditions it is appropriate to consider change in outcome between the two interviews. A simple way of doing this is to include initial outcome as an explanatory variable in the logistic model. This is represented by the term 'INOC' in Table 9.8. After fitting a difference in final outcome between conditions (model 2) and a difference between asthmatics and other children with wheezy chest (model 3) initial outcome is included as possible predictor of final outcome (model 4). The reduction in residual deviance is very large—it falls by 81·2 for the loss of just one degree of freedom. There appears to be a strong link between initial outcome and final outcome. This association would appear to be different for each condition—the improvement upon adding the interaction term 'COND·INOC' is significant at the one percent level (model 6). After allowing for this interaction, the interaction between initial outcome and whether a child with wheezy chest has asthma is no longer significant (model 7).

Table 9.8 Binary measure of final outcome for children with bedwetting and recurrent wheezy chest (including initial outcome as an explanatory variable): model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	1493·5	1085			
2	GM + COND	1485·0	1084	8·5	1	0·004
3	GM + COND + ASTH	1455·4	1083	29·7	1	<0·001
4	GM + COND + ASTH + INOC	1374·2	1082	81·2	1	<0·001
5	GM + COND + ASTH + INOC ASTH·INOC	1368·9	1081	5·3	1	0·02
6	GM + COND + ASTH + INOC + COND·INOC	1363·9	1081	10·3	1	0·001
7	GM + COND + ASTH + INOC + COND·INOC + ASTH·INOC	1363·7	1080	0·2	1	0·65
8	GM + COND + ASTH + INOC + COND·INOC + PHAS	1363·2	1080	0·7	1	0·40
9	GM + COND + ASTH + INOC + COND·INOC + PSTD	1363·9	1080	0·1	1	0·75
10	GM + COND + ASTH + INOC + COND·INOC + CPST	1363·1	1079	0·9	2	0·64

There was no difference in final outcome between the two data collection phases (model 8) and again there was no discernible effect of standard setting (models 9 and 10).

Model 6 is selected as the one that best represents the data. The raw data broken down by the terms appearing in this model are given in Table 9.9. Of those children with bedwetting who had not wet the bed in the month preceding the initial interview, just under 20% had had problems with bedwetting in the month preceding the final interview. In contrast, of the 134 children suspected of having asthma but who recorded no chest problems in the month preceding the initial interview, more than 40% reported at least one chest problem in the month preceding the final interview. Asthma tends to be

Table 9.9 Proportion of children for whom a successful response (no bedwetting or no chest problem) was recorded in the final interview by study condition and outcome recorded at the initial interview

	Study Condition		
		Recurrent wheezy chest	
Initial outcome	**Bedwetting**	**Asthmatics**	**Other children**
Problem reported at first interview	140/422 (33·1%)	89/268 (33·2%)	39/79 (49·4%)
No problem reported at first interview	56/69 (81·2%)	76/134 (56·7%)	86/114 (75·4%)

cyclical in nature—children can have 'good' months and 'bad' months depending upon the absence or presence of triggers.

9.4.7 Effects of standard setting

The conclusion, based on the results of these three analyses, was that the binary measures of clinical outcome showed no effects of standard setting.

9.5 Graded measure of clinical outcome

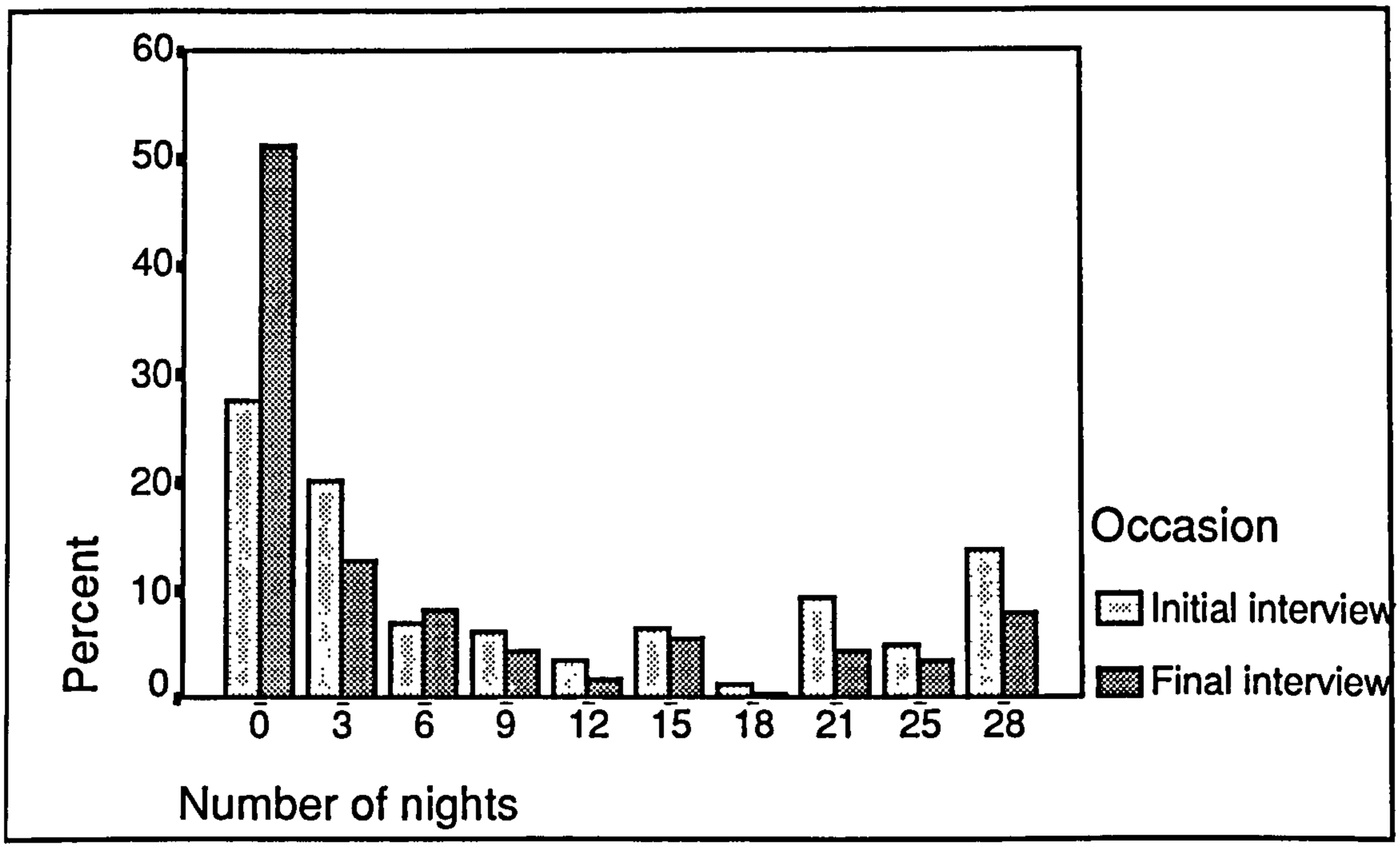
Quantitative measures of clinical outcome were developed for the two chronic conditions.

9.5.1 Developing a measure of outcome for bedwetting

For children with bedwetting, parents were asked to estimate the number of nights during the preceding four weeks on which the child had wet the bed. The data were skewed and the distribution of responses could nearly be described as U-shaped (Figure 9.1). Many children had not wet the bed at all in the previous four weeks; and a

smaller group had wet the bed on every night during that period. The distribution is clearly not Normal.

Figure 9.1 Number of nights on which the child was reported to have wet the bed during the four weeks preceding the interview by occasion



There is a marked difference between initial and final interviews. The proportion of children who did not wet the bed at all during the relevant four weeks rose from less than 30% at the time of the initial interview to just over 50% at the time of the final one and is consistent with the improvement in binary outcome reported in section 9.4.

The difference between responses to the initial and final interviews was considered as a measure of improvement (Figure 9.2). The mean reduction in the number of nights on which the child wet the bed was 3·7. The data is peaked and has extended tails with kurtosis = 1·4; the data are clearly not normally distributed (the Normal distribution has a kurtosis of zero).

Figure 9.2 **Reduction in the number of nights the child wet the bed between initial and final interviews**

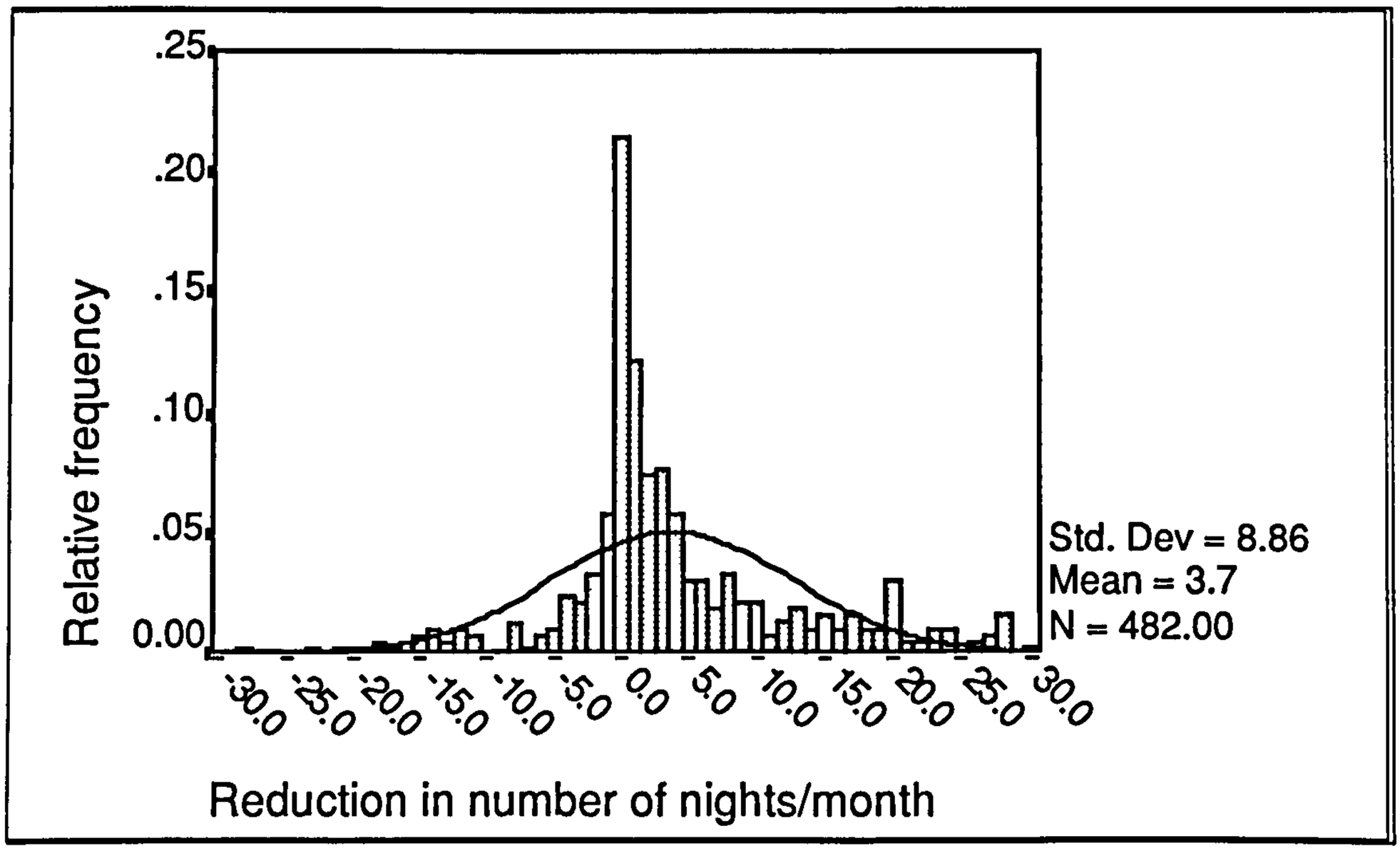
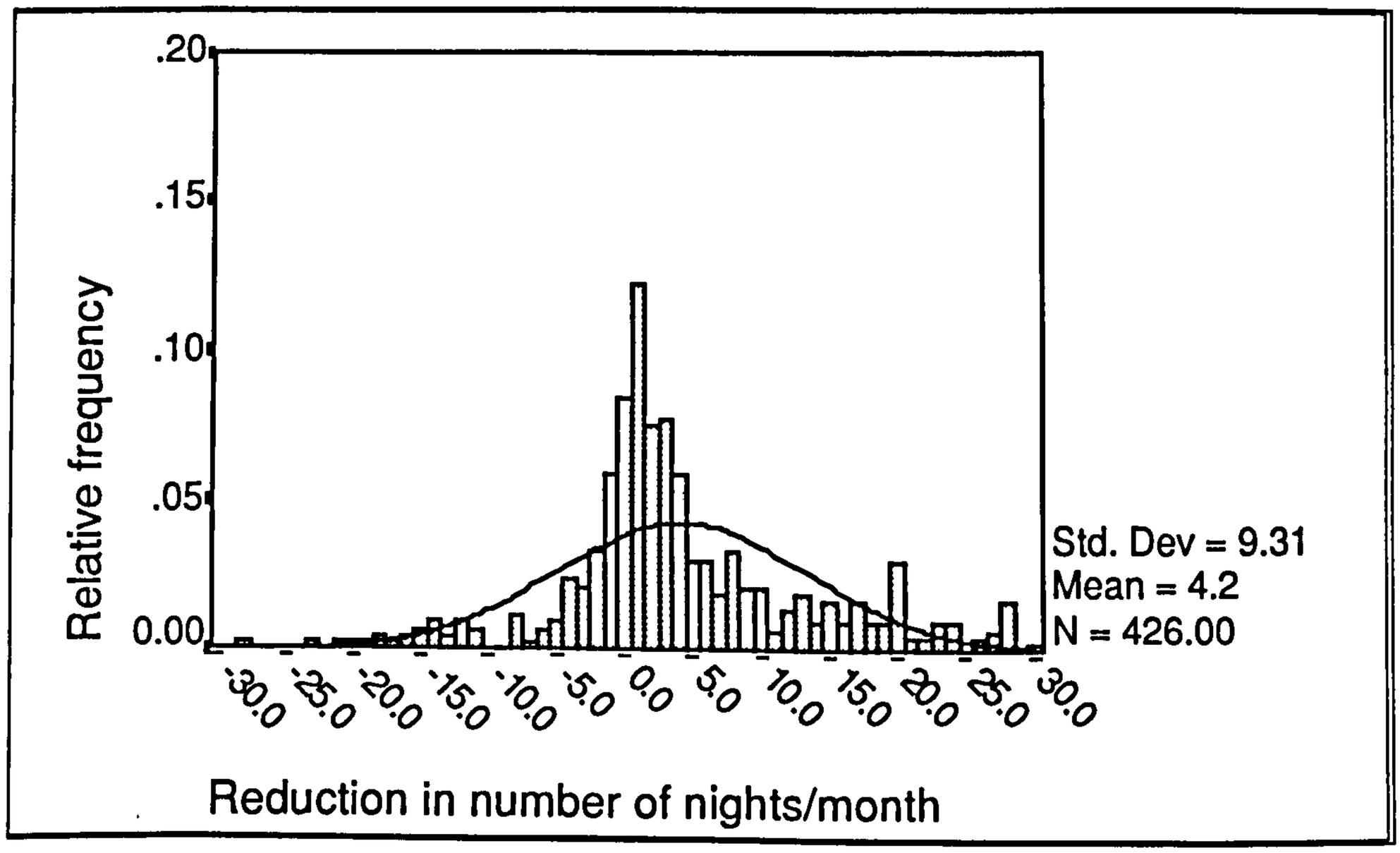


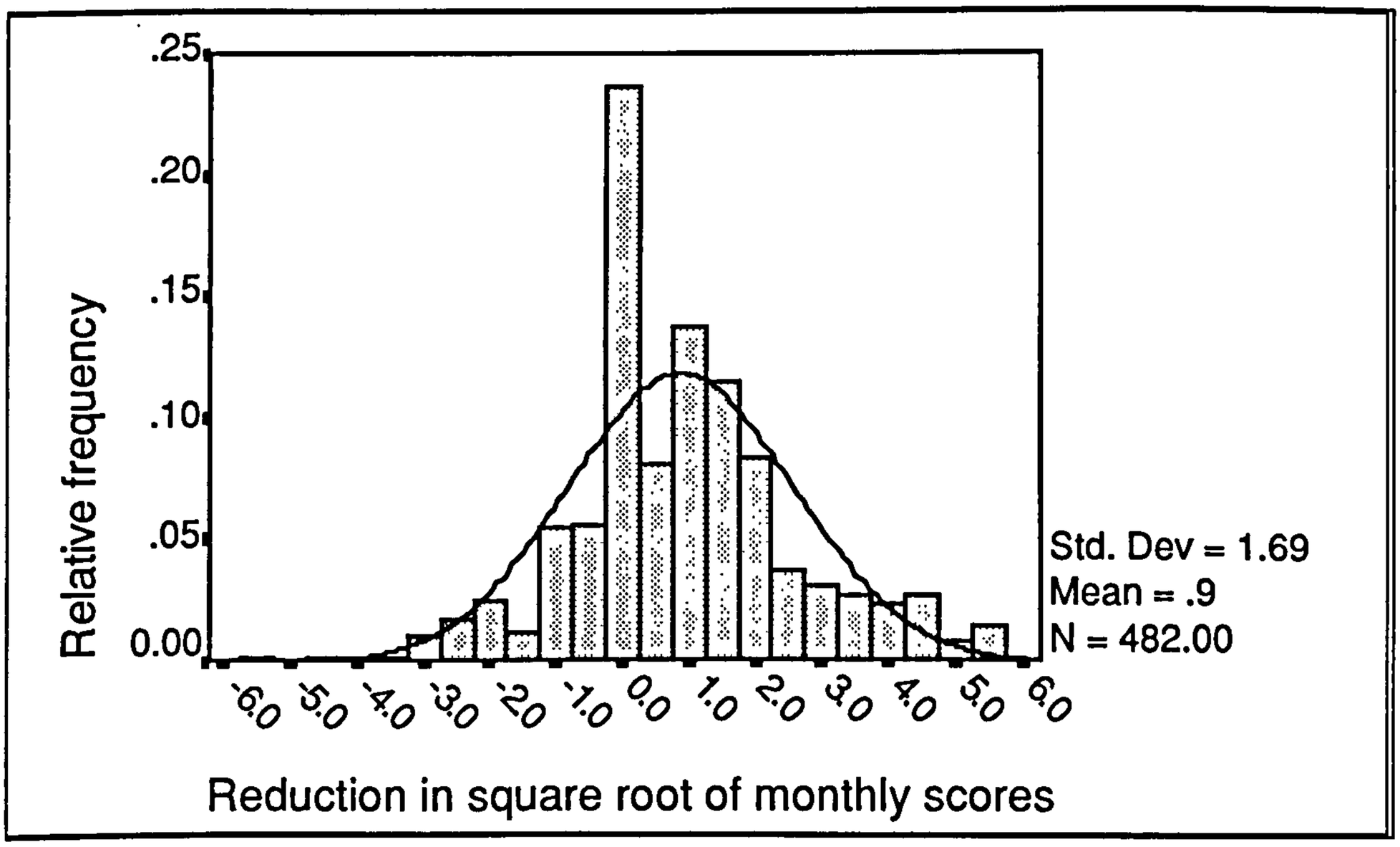
Figure 9.3 **Reduction in the number of nights the child wet the bed for those children who were reported to have wet the bed on at least one occasion**



If we remove those children for whom no wetting was reported at either interview, the kurtosis is reduced to 1·0 (Figure 9·3) but the tails of the distribution are still fairly long.

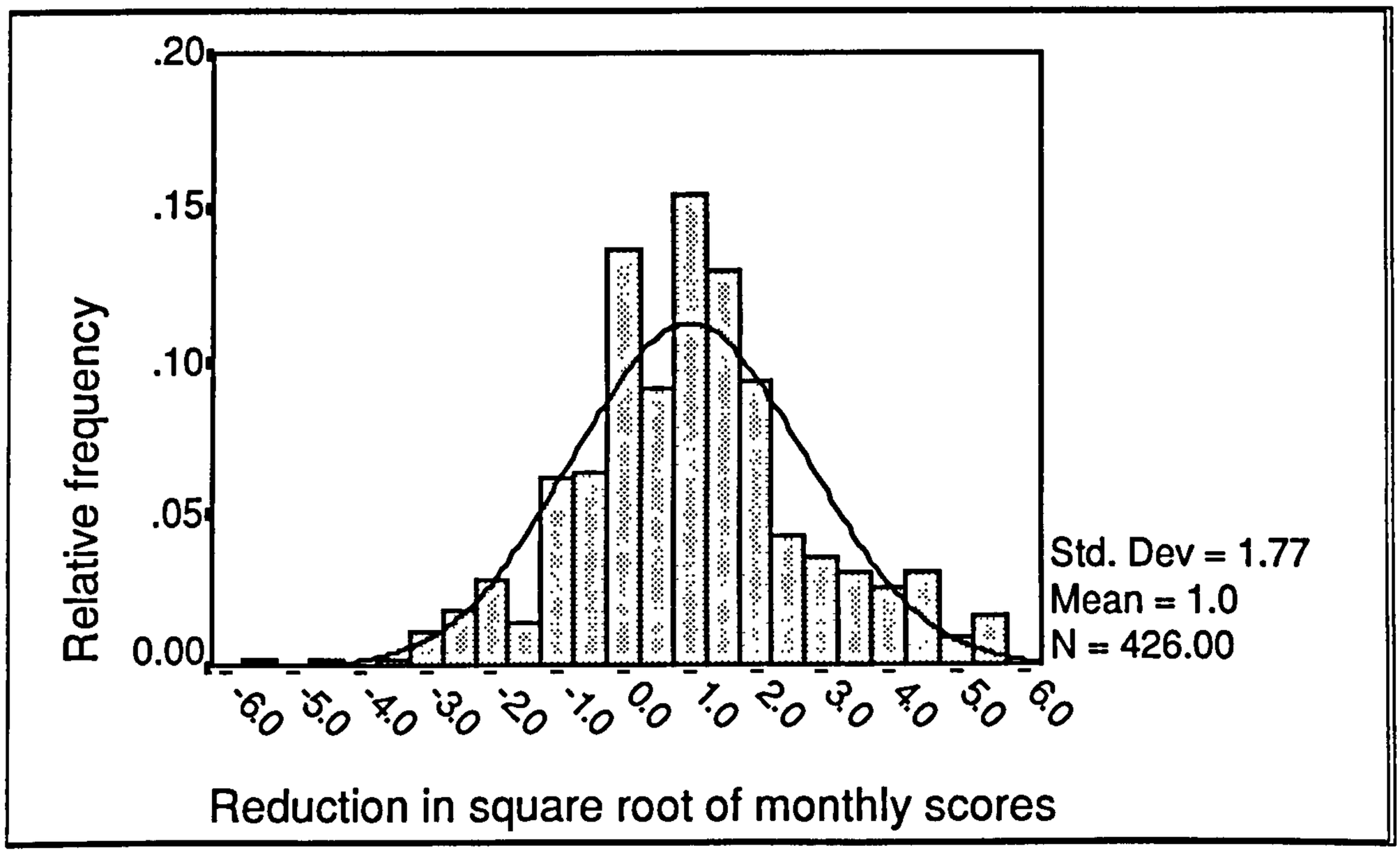
A square root transformation of the data was considered. The distributions of change scores for all children and the subset of children who were reported to be wetting at least one of the interviews are given in Figures 9.4 and 9.5 respectively. The values of kurtosis for these distributions are 0·7 and 0·5 respectively. These distributions follow more closely a Normal distribution. It may be more appropriate to use the transformed scores in regression type analyses.

Figure 9.4 Reduction in the square root of the number of nights the child wet the bed between initial and final interviews



There were also clinical reasons for considering the square root transformation. It is not clear that a reduction in wetting for one child from 30 to zero nights per month is twice as good as a reduction from 15 nights to zero for another child. It is also difficult to

Figure 9.5 Reduction in the square root of the number of nights the child wet the bed for those children who were reported to have wet the bed on at least one occasion



argue that a reduction from 25 to 20 nights is equivalent to a reduction from five to zero nights. (In each of these examples clinicians felt that more weight should be given to the second scenario; this is achieved by using a square root transformation).

Comparison of Figure 9.5 with Figure 9.4 might suggest that any distributional assumptions of normality are better met if we omit those children who had not wet the bed in the month prior to either interview (although an analysis based on only these children would not answer the research question of interest).

9.5.2 Developing a graded measure of outcome for recurrent wheezy chest

For children with recurrent wheeze, three outcome measures were considered: the number of days on which the child wheezed; the number of days on which the child was breathless; and the number of nights during which the child was disturbed in the month

preceding the interview. These three variables were correlated with each other (Table 9.10).

The degree of collinearity that these correlations imply suggests that it might be appropriate to devise a single index of clinical outcome rather than analyse each variable separately. This view was supported by a principal components analysis (Chatfield and Collins, 1980, Chapter 4). The three eigen values were 2·0, 0·6 and 0·4;

Table 9.10 Inter-correlation among wheeze clinical outcome variables

	Correlation coefficients	
Number of days breathless	0·46	
Number of days wheezy	0·54	0·61
	Number of disturbed nights	**Number of days breathless**

the first principal component explained 67% of the variation in the data. Principal components are simply linear combinations of the variables of interest. In a principal components analysis the coefficients of the first principal component are chosen in such a way that the proportion of the total variation explained by the linear combination is maximised. It is therefore natural to use this as the indicator variable of clinical outcome.

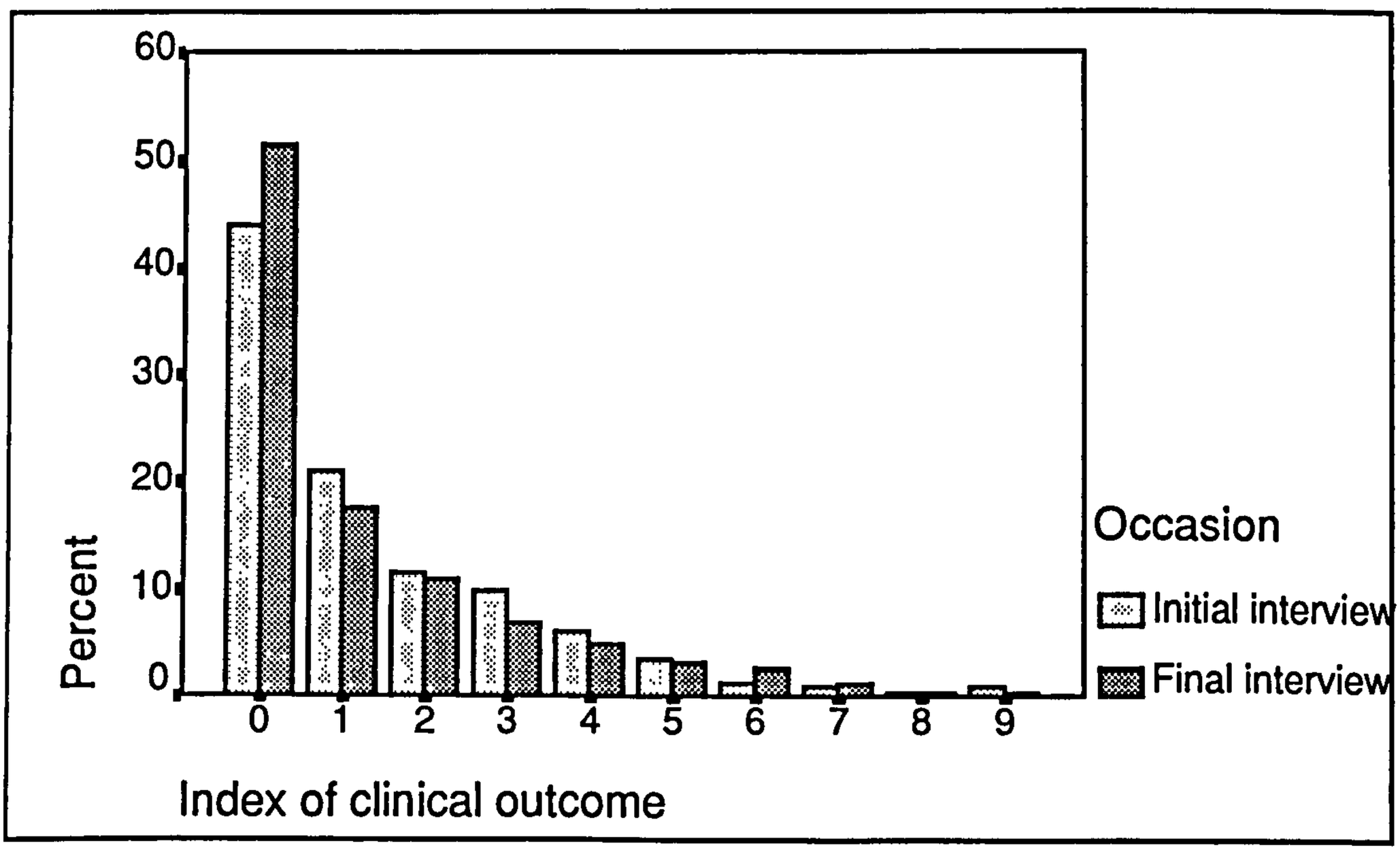
Because of the skewed nature of the data, a square root transformations of each variable was made before the principal components analysis was undertaken.

The actual index used was given by:

$$\text{Index} = 0.59 \times \sqrt{\text{days breathless}} + 0.62 \times \sqrt{\text{days wheezy}} + 0.53 \times \sqrt{\text{nights woken}} \qquad (9.1)$$

The distribution of index scores obtained from initial and final interviews is given in Figure 9.6. The distribution is similar to that of the wetting outcome variable—there is a large proportion of children who reported no chest trouble at all in the month prior to the interview and there is some reduction noticeable between the initial and final interviews. The main differences are that this reduction is not as marked and that there is a much smaller proportion of children who suffered from symptoms on every day of the month

Figure 9.6 Index of clinical outcome for wheeze: distribution of scores broken down by occasion of interview



As for bedwetting, it is natural to consider the difference between the initial and final interviews as a measure of clinical outcome. The frequency distribution of difference scores is given in Figures 9.7 and 9.8. There were a total of 555 children for whom valid responses were available for all three component variables (breathlessness, wheezing and disturbed nights) for both interviews. Of these, 162 reported no symptoms at all at

Figure 9.7 Reduction in clinical index score between initial and final interviews: frequency distribution

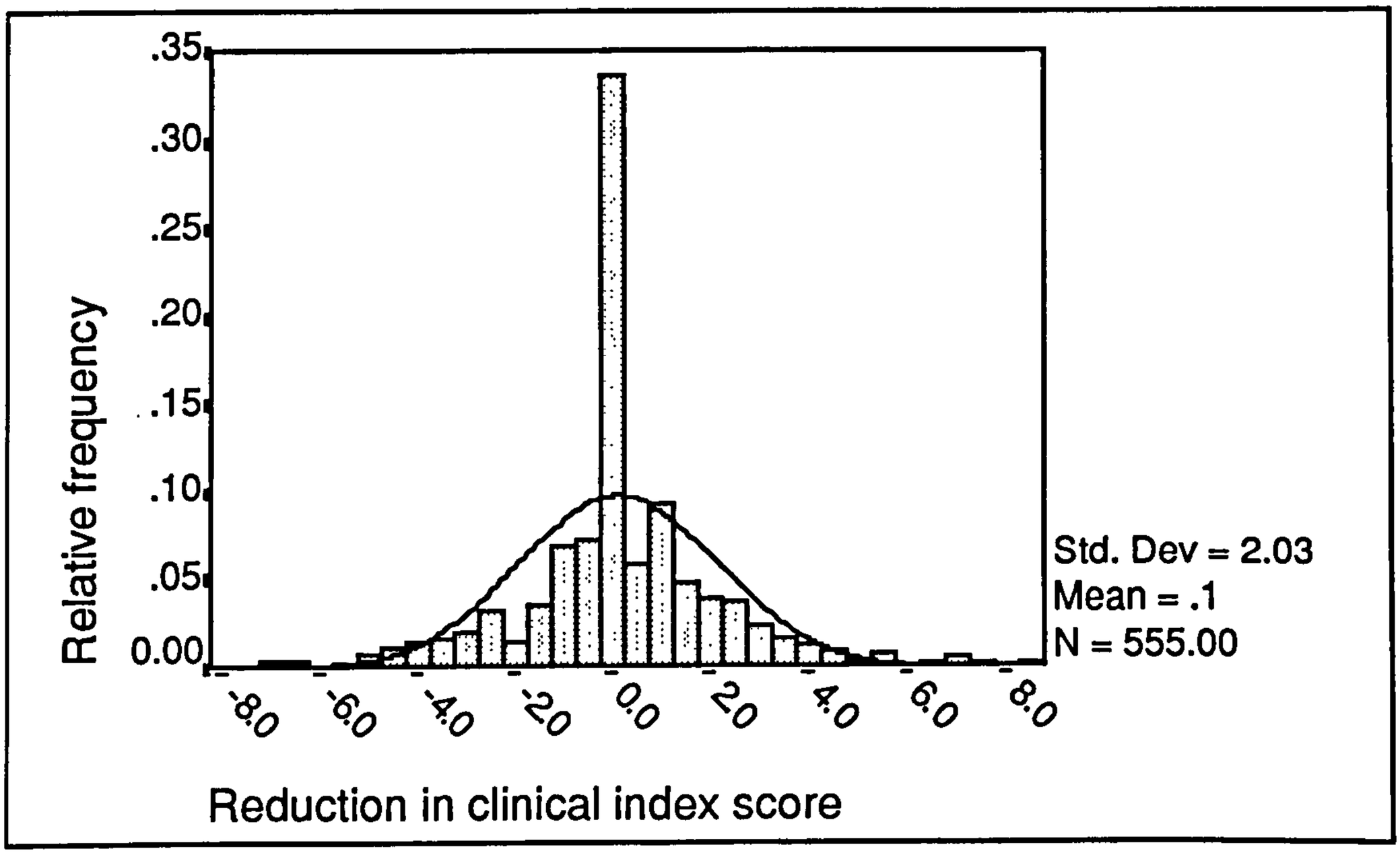
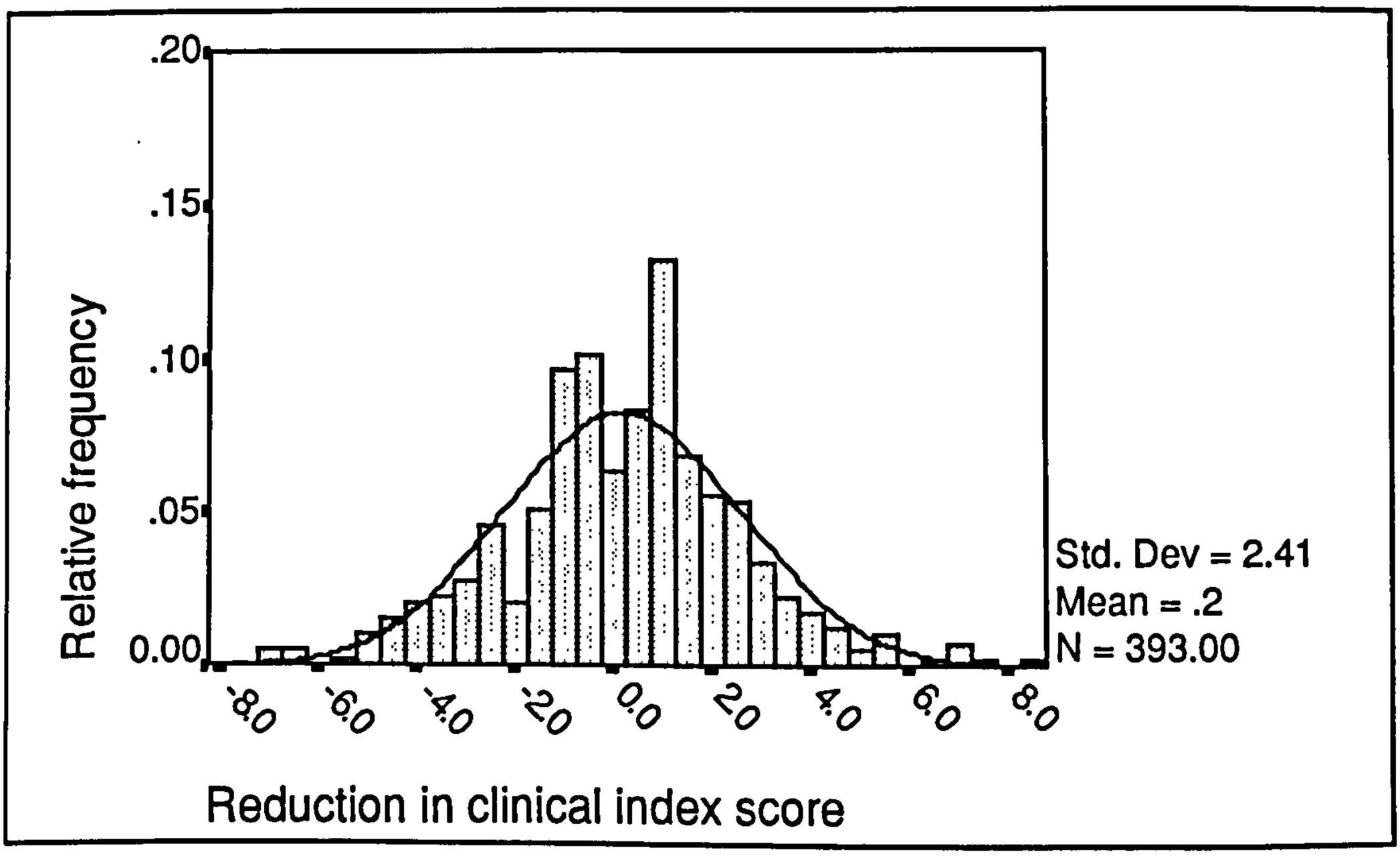


Figure 9.8 Reduction in clinical index score between initial and final interviews for those children who reported some chest trouble in the month preceding at least one of the interviews



either interview. These responses yield the prominent spike in the distribution in Figure 9.7. The mean reduction in the index score for all children was 0·14; the standard deviation was 2·03.

Again, as in the case of children with bedwetting, if we omit those children who had not had a problem in either subphase, the distribution of difference scores looks much more normal (Figure 9.8).

9.5.3 Combining the two condition specific measures in one analysis

The method of assessing clinical outcome for each condition has been defined above. The outcome variable is a measure of difference between two statistics; the first calculated from responses to the initial interview, the second calculated from responses to the final interview. Histograms of the frequency distribution of each measure suggest that general linear modelling with the assumption of a Normal error structure might be appropriate. But this statistical model assumes that all errors are identically distributed with the same nominal variance. If we exclude those children who experienced no symptoms prior to either interview, the variances of the measures as defined above are 3·13 for bedwetting and 5·81 for recurrent wheezy chest. In analogous situations in structural equation modelling, Bentler (1989, page 20) suggests that the input variables be scaled so that their variances are approximately equal. In this case the variance of the wheeze clinical outcome index is influenced by the values of the relative weights in equation 9.1. These weights were arbitrarily chosen such that their squares summed to unity. It therefore seemed sensible to scale the wheeze index such that the difference scores have a variance equal to that of the corresponding bedwetting scores. The wheeze scores were multiplied by a scale factor of 0·73.

9.5.4 Results

The results of the GLIM analysis are summarised in Table 9.11. The analysis was done for all children and then repeated for those children for whom some symptoms had been reported at at least one of the interviews. It was recognised that omitting children who had had no symptoms reported in either subphase actually changed the nature of the research question that was being addressed. In particular if one group of doctors had been particularly successful at preventing a recurrence of symptoms in those children with no symptoms in subphase A, this would not be detected by the second analysis (although preliminary multiple logistic regression analyses indicated that it was unlikely that any of the interventions had had this type of effect).

It was felt that there were two good reasons for doing both analyses. The first was a statistical one. The spikes caused by the *"double zeros"* in the distributions shown in Figures 9.4 and 9.7 result in the distributions having a kurtosis which is higher than that of a normal distribution. In his discussion of the effects of small and high values of kurtosis, Lindman (1992, p.22) concludes that *"the F-test is robust with respect to non-normality if N is large and is likely to be robust even if N is only moderately large"*. It was felt that it would be useful to be to compare the p-values from the two analyses in order to carry out some sort of check of this robustness. The second reason was clinical. With both conditions there is an age effect—children tend to 'grow out' of each condition. Graphs of prevalence rates which show this effect are given in volume 3 of the final report (North of England Study, 1990c). It is possible that a number of the children who had reported no symptoms in the first subphase (approximately 10 weeks after the administration of the prevalence survey) fell into this category. It is unlikely that the interventions could have had any effect on clinical outcome for these children. Unfortunately it is impossible to distinguish these children from those who were still suffering from the condition but whose double zero response was a result of good

control. The true effects of the interventions are likely to be somewhere between the estimates provided by each analysis.

F-tests for the comparison of nested models are reported. Both analyses gave almost identical results.

Table 9.11 Change in clinical outcome: model selection

Model	All children				Children with reported symptoms			
	ν_1	ν_2	$F_{\nu1,\nu2}$	p	ν_1	ν_2	$F_{\nu1,\nu2}$	p
1 GM								
2 GM + COND	1	1035	62·02	<0·001	1	817	47·40	<0·001
3 GM + COND + ASTH	1	1034	0·05	0·82	1	816	0·33	0·56
4 GM + COND + PRAC	61	974	1·12	0·25	61	756	1·15	0·21
5 GM + COND + PHAS	1	1034	2·59	0·11	1	816	2·39	0·12
6 GM + COND + PSTD	1	1034	3·61	0·06	1	816	3·60	0·06
7 GM + COND + CPST	2	1033	4·45	0·011	2	815	4·79	0·009
8 GM + COND + WZST	1	1034	8·74	0·003	1	816	9·44	0·002
9 GM + COND + WZST + CPST	1	1033	0·16	0·69	1	815	0·14	0·71
10 GM + COND + WZST + MIXD	1	1033	0.17	0·68	1	815	0·40	0·53
11 GM + COND + WZST + WZMX	1	1033	0.25	0·62	1	815	0·79	0·37
12 GM + COND + WZST + AUDT	4	1030	0·24	0·92	4	812	0·28	0·89
14 GM + COND + WZST + WZAU	3	1031	0·33	0·80	3	813	0·38	0·77

There was a difference between conditions (model 2) but as the outcome measure used was not derived in exactly the same way for each condition, this difference has no clinical significance. There was no difference between children with asthma and other children with wheezy chest (model 3). Differences between practices were not significant (model 4) and there were no differences between the two phases of data collection (model 5). When a standard setting effect was included (model 6) the improvement was

not quite significant at the 5% level. When a separate standard setting effect was included for each condition (model 7) the improvement was significant at around the 1% level.

Examination of the parameter estimates (Table 9.12) suggested that, for bedwetting, there was no difference between doctors who set standards and other doctors. The difference seemed to occur mainly for doctors who set a standard for wheezy chest. This hypothesis was tested by creating the variable 'WZST' which took the value 2 for observations corresponding to children who consulted (for wheezy chest) in phase 2 with a doctor who set a standard for wheezy chest and 1 otherwise. Fitting this contrast (model 8) gives an improvement significant at the 0·1% level. Adding the remaining contrast that makes up the full interaction term (model 9) offers virtually no improvement at all. This supports the hypothesis that apparent condition specific effect of standard setting noted above can be attributed to just one effect—that of setting a standard for wheezy chest.

Table 9.12 Estimates of effect of standard setting on clinical outcome by study condition generated from model 7

	Difference (and 95% confidence interval) in clinical outcome between doctors who set standards and other doctors			
Condition	**For all children**		**For subset of children**	
Bedwetting	-0·10	(-0·60, 0·40)	-0·11	(-0·70, 0·48)
Wheezy chest	+0·69	(0·23, 1·15)	+0·93	(0·33, 1·53)

As a possible effect of standard setting had been identified the effects of the other interventions were investigated. First the effect of meeting a mixed group was considered (model 10); the effect was not significant. As wheezy chest was the only

condition for which setting a standard had a significant effect, it is natural to consider the effect of meeting a mixed group on setting a standard for this particular condition. This was achieved by fitting the contrast represented by the term 'WZMX' (model 11). The effect was not significant.

Finally the effects of the other types of audit were considered. There was no general effect (model 12) and no condition specific effect of the other types of audit activity on outcome for wheezy chest (model 13).

The final model is simply GM + COND + WZST. Normal quantile-quantile plots were plotted for this model for each of the analyses (Figure 9.9). In both plots the data lie reasonably close to a straight line. There are two noticeable jumps in the first plot. These correspond to the children for whom there was no change in outcome—one for each condition. As these children were excluded from the second analysis, the second plot is somewhat smoother.

Parameter estimates corresponding to the final model indicate a significant reduction in the scaled clinical index score for those children who consulted with doctors who set standards for recurrent wheezy chest. The reduction and 95% confidence intervals estimated from each analysis were respectively 0·69 [0·23, 1·15] and 0·69 [0·33, 1·53]. To help interpret these results, the results of a univariable analysis for each component of the wheeze clinical outcome index are presented in Table 9.13.

The mean reduction in the square root of the number of days breathless for children consulting in phase 2 with doctors who standard setters was 0·68 (with 95% confidence interval from 0·16 to 1·20); the corresponding reduction for other children was -0·02 (with 95% confidence interval -0·15 to 0·11). The difference between the two groups was 0·70 with 95% confidence interval (0·25 to 1·15). Children who consulted with

Figure 9.9 Normal Q-Q plots of standardised residuals for final model in each analysis

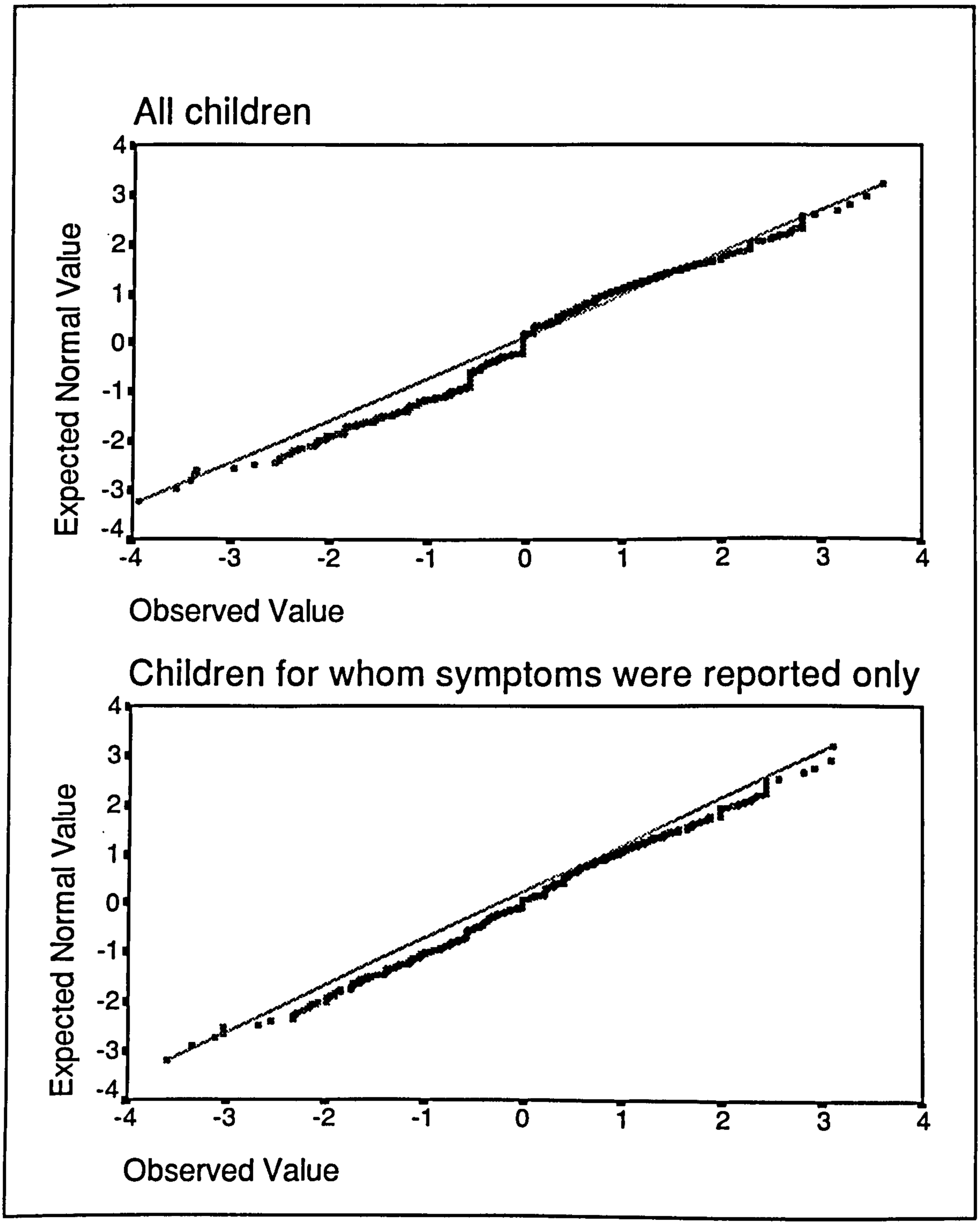


Table 9.13 Components of clinical outcome for children with recurrent wheezy chest: comparison of children who consulted in phase 2 with doctors who set a standard with other children

Outcome variable (Reduction in the square root of :)	Group means: Children in phase 2 who consulted with a doctor who set a standard for wheezy chest (n=50)	Group means: Other children (n=505)	Difference between group means	95% confidence interval for difference
Days breathless	0·68	-0·02	0·70	(0·26, 1·15)
Days wheezy	0·69	0·00	0·68	(0·22, 1·15)
Nights woken	0·34	0·12	0·22	(-0·17, 0·61)

doctors who set standards demonstrated a larger reduction in breathlessness than other children. Similarly these children experienced a greater reduction in the square root of the number of days they were breathless. The difference between the two groups was 0·69 with 95% confidence interval from 0·22 to 1·15. But there was no corresponding difference in the square root of the number of disturbed nights.

9.6 Summary

The analysis of the graded measure of clinical outcome has indicated that there has been improvement in clinical outcome for patients with recurrent wheezy chest that may be attributed to the setting of clinical standards. That this improvement was not detected by the binary measure of outcome (section 9.4) possibly reflects the poorer sensitivity of that measure.

Chapter 10

Outcome of care 2: parents' satisfaction with care—evidence from interviews

10.1 Introduction

At the time when the interview schedules were being drawn up (1983/84), the study team were unable to find an existing, valid and reliable measure of satisfaction with medical care that had been used in the British general practice. It was therefore decided to adapt an instrument developed and validated in the United States (Roughmann et al 1979, Zastowny et al, 1983). to assess parents' satisfaction with the care that their child had received. The instrument comprised 12 items relating to accessibility, availability and quality of health care. As a result of pilot testing, some of the language was anglicised and an extra response category was added to item 1 to reflect the fact that not all English general practice surgeries operated an appointment system. The distribution of responses obtained at initial interviews by study condition is given in Table 10.1. In general, responses were highly skewed—parents tended to be very satisfied with the care that their child received.

The authors suggested that the instrument could be used in the form of two 6 item scales (Table 10.2) employing simple summation of the item scores to obtain scale totals. To determine if this method of scoring satisfaction was appropriate in this study, a number of investigations were undertaken.

Table 10.1 Satisfaction scale: frequency of responses at initial interviews

No	Item	Response	score	Frequency (%) Acute cough (n=454)	Bed-wetting (n = 495)	Wheezy chest (n = 602)
1	Except in cases of emergency how difficult is it to get an appointment with (your child's) doctor?	Very difficult	0	9·9	7·7	10·8
		Difficult	1	25·1	30·7	27·7
		Not difficult	2	54·0	52·1	53·2
		No appointments		2·2	4·0	3·8
		Missing		8·8	5·5	4·5
2	Is it easy, difficult or very difficult for you to get to (your child's) GP during surgery hours?	Easy	2	73·8	79·8	82·9
		Difficult	1	17·0	12·9	10·6
		Very difficult	0	2·9	2·2	2·5
		Missing		6·4	5·1	4·0
3	After getting there do you feel that the time that you have to wait to see the doctor is much too long, too long, or not too long?	Much too long	0	14·5	9·9	11·8
		Too long	1	28·0	26·3	23·9
		Not too long	2	50·7	58·4	59·6
		Missing		6·8	5·5	4·7
4	Once you see the doctor does s/he usually spend enough time with (your child) or not enough time?	Enough	2	77·1	85·3	88·2
		Not enough	1	15·4	9·1	7·5
		Missing		7·5	5·7	4·3
5	In your opinion, how concerned is the doctor about (your child) as a person. Is s/he very concerned, concerned or not very concerned?	Very concerned	2	33·9	41·8	43·2
		Concerned	1	48·0	42·0	47·3
		Not very concerned	0	11·0	8·9	5·3
		Missing		7·0	7·3	4·2
6	How careful is the doctor when s/he examines your child - very careful, careful or not careful?	Very careful	2	57·7	59·8	66·8
		Careful	1	32·6	29·9	27·7
		Not very careful	0	2·9	2·1	1·0
		Missing		6·8	8·1	4·5
7	How willing is the doctor to listen when you tell him/her about your child's health - very willing, willing or not very willing?	Very willing	2	56·4	65·5	66·3
		Willing	1	30·2	25·3	24·9
		Not very willing	0	6·6	3·4	4·7
		Missing		6·8	5·9	4·2
8	Do you usually feel that the doctor usually gives you enough information about your child's health, or would you like more?	Enough	2	56·2	69·7	67·6
		Not enough	1	37·0	23·6	28·2
		Missing		6·8	6·7	4·2
	In the last 12 months, when you were there with your child, were there times when you had a bad experience with a doctor and felt.....					
9	that the doctor just tried to get rid of you?	Yes	0	7·0	1·6	3·3
		No	2	85·2	83·2	89·4
		Missing		7·7	15·2	7·3
10	that the doctor or other people there didn't care about you?	Yes	0	15·0	8·3	9·6
		No	2	77·5	76·4	82·7
		Missing		7·5	15·4	7·6
11	that the doctor was too busy to spend enough time with you?	Yes	0	12·8	6·3	6·6
		No	2	79·7	78·0	85·4
		Missing		7·5	15·8	8·0
12	Have you had any other bad experiences at the practice over the last year?	Yes	0	24·7	15·4	19·1
		No	2	68·3	69·3	72·9
		Missing		7·0	15·4	8·0

10.2 Reliability of satisfaction scales

First the internal reliability of the two scales was assessed using Chronbach's alpha (Chronbach 1951). This took the value 0·59 for scale 1 and 0·71 for scale 2. These values are not particularly high given the generally high level of satisfaction overall. In addition, it is not clear that the groupings form two distinct conceptual aspects of satisfaction with care. In particular, the internal reliability of the second scale can be increased by omitting the first item (item number 2) which relates to access to medical care (Table 10.2). It is not clear why it should be included with the remaining items that relate to personal contact with the doctor.

Table 10.2 Internal reliability of the two patient satisfaction scales

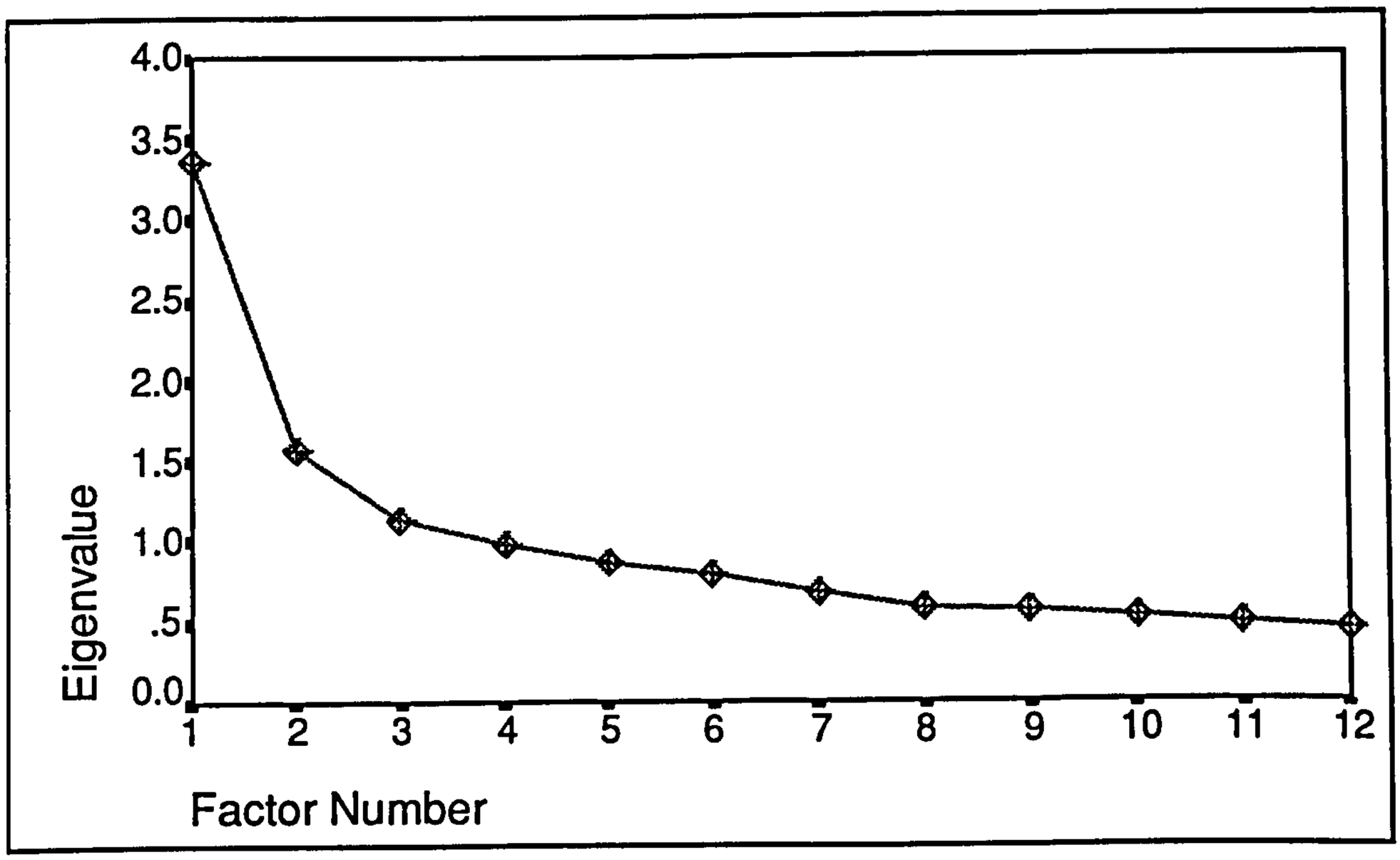
Scale 1				Scale 2			
Item no.	Item content	Corrected item-total correlation	α if item deleted	Item no.	Item content	Corrected item-total correlation	α if item deleted
1	Getting appointment	0·24	0·55	2	Getting there	0·05	0·77
3	Waiting for care	0·26	0·54	4	Spend enough time	0·54	0·66
9	Get rid of you	0·34	0·52	5	How concerned	0·58	0·62
10	Did not care	0·44	0·46	6	How careful	0·52	0·64
11	Too busy	0·35	0·51	7	Willing to listen	0·59	0·61
12	Other bad experience	0·28	0·54	8	Enough information	0·47	0·66
α = 0·59				α = 0·71			

10.3 Principal components analysis of satisfaction items

A principal components analysis of all twelve items of the satisfaction instrument was undertaken. There were three eigen values that exceeded unity. This suggested that a rotation of the first three components to see if homogeneous groupings of variables could be obtained might be worthwhile (Kim and Mueller, 1994, page 111). Examination of the scree plot (Figure 10.1), however suggested that a two factor

solution might be preferred. Orthogonal rotations of the first two and the first three principal components were carried out (Table 10.3).

Figure 10.1 Scree plot of eigen values



In both solutions, the first factor is based around items 4 to 8 which all relate to communication with the general practitioner. In both cases item 11 (was doctor too busy?) cross-loaded onto this factor. In the two factor solution, all the remaining items loaded onto the second factor. In the three factor solution, the remaining items split to load onto two separate factors. The last four items, which relate to adverse occurrences at the practice, formed factor 2 and items 1 to 3, which relate to access to health care, formed the third.

The three factor solution is perhaps the easiest to interpret. The internal reliability of the three factors were respectively 0·77, 0·56 and 0·38. The internal reliability of the last factor in particular was very low. It is not clear that it could be used as a reliable index of satisfaction with access to care.

Table 10.3 Principal component analysis: two and three factor solutions - rotated factor loadings*

Item	Item content	Two factor solution		Three factor solution		
		Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1	Getting appointment		0·61			0·71
2	Getting there		0·37			0·57
3	Waiting for care		0·45			0·64
4	Spend enough time	0·71		0·70		
5	How concerned	0·77		0·77		
6	How careful	0·70		0·70		
7	Willing to listen	0·77		0·78		
8	Enough information	0·61		0·62		
9	Get rid of you	0·35	0·44		0·69	
10	Did not care		0·66		0·76	
11	Too busy	0·55	0·43	0·45	0·58	
12	Other bad experience		0·45		0·55	
Internal reliability		0·78†	0·51‡	0·77‡	0·56	0·38

* Factor loadings less than 0·25 have been omitted
† Item 9 not included when calculating Chronbach's alpha
‡ Item 11 not included when calculating Chronbach's alpha

10.4 Standard setting and satisfaction

Before proceeding with the analysis thought was given about the mechanisms by which standard setting might influence satisfaction. At a basic level it is possible to hypothesise that if standard setting improves the health of the child, this might be reflected in higher levels of satisfaction with care. Consideration was then given to the three components of satisfaction corresponding to the three factor solution derived above. A number of standards included educational components which addressed issues such as the provision of more information to the parent about their child's condition. As factor 1 included

items relating to satisfaction with the amount of information provided, any change in this area might be reflected in an improvement in the factor score. The other items in factor 1, which relate to care provided by the GP, could be affected through similar mechanisms. These improvements are likely to be specific to the child's condition, arising from changes in the management of that condition.

It is possible that standard setting could influence the items that make up factors 2 and 3. Doctors were free to consider all aspects of care which might include changes to the appointment system or changes to the waiting room. The problem with many such changes is that they are unlikely to be restricted to children with the specific condition for which the child consulted. It is probable that they will affect the care given to the children with the control conditions. The study was not designed to detect such changes. The study was designed only to detect changes specific to the condition for which the doctor set a standard. It was felt that there was more likelihood of detecting a change in factor 1 than in the other two factors. Although factors 2 and 3 (adverse occurrences at the practice and access to care) were analysed for completeness, only the analysis of factor 1 (satisfaction with care during the consultation) is reported in full in this thesis.

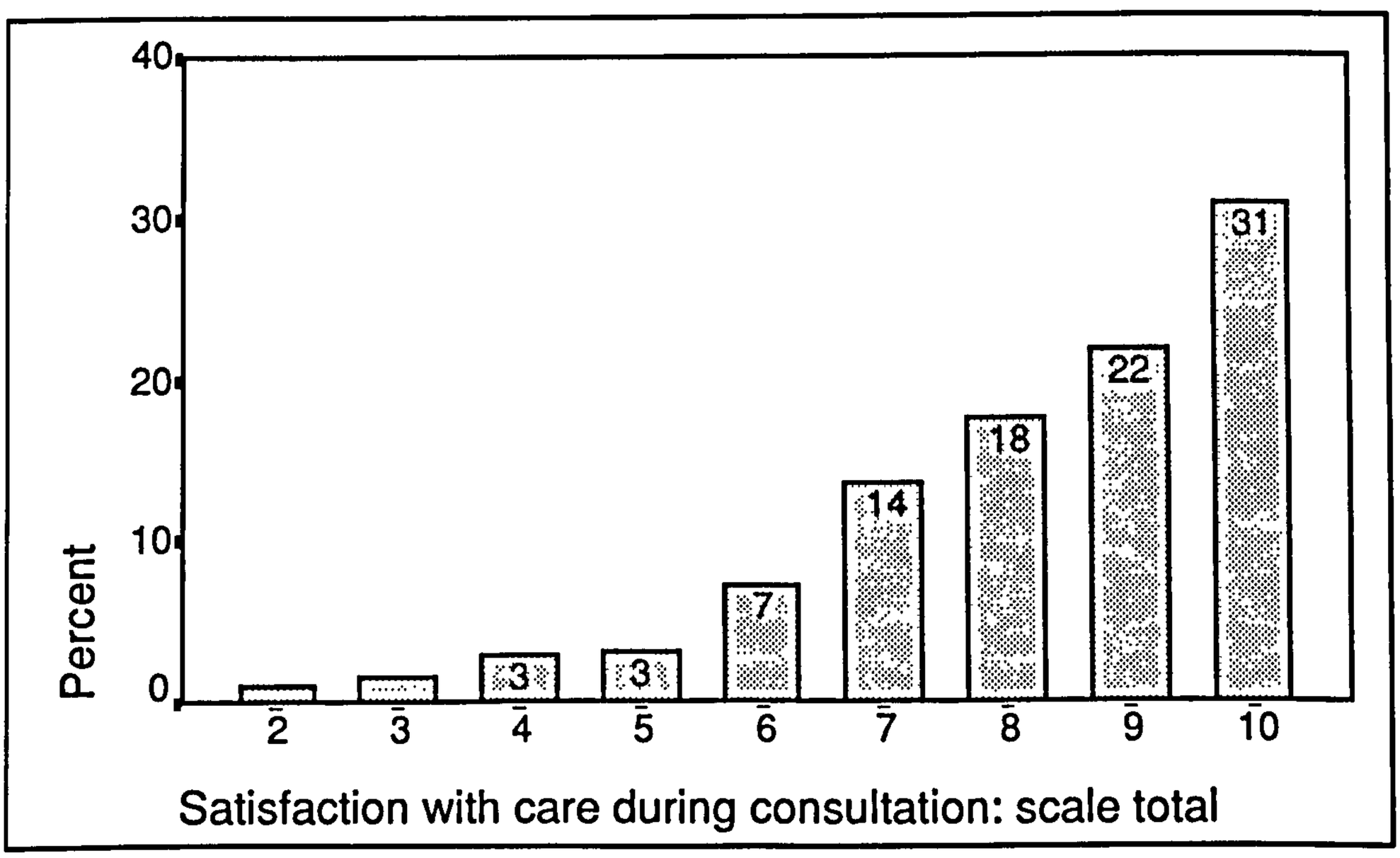
10.5 Satisfaction with the care delivered during the consultation

10.5.1 Frequency distribution

Using the scoring system for items given in Table 10.1, the first index of satisfaction (the sum of the items that make up the first factor) has a possible range of scores from 2 to 10. The frequency distribution of index scores at the time of the initial interviews is given in Figure 10.2. The scores are highly skewed. More than 30% of parents reported that were very satisfied with all aspects of the consultation. The scale clearly suffers from ceiling effects (Streiner and Norman, 1989). With so many respondents recording

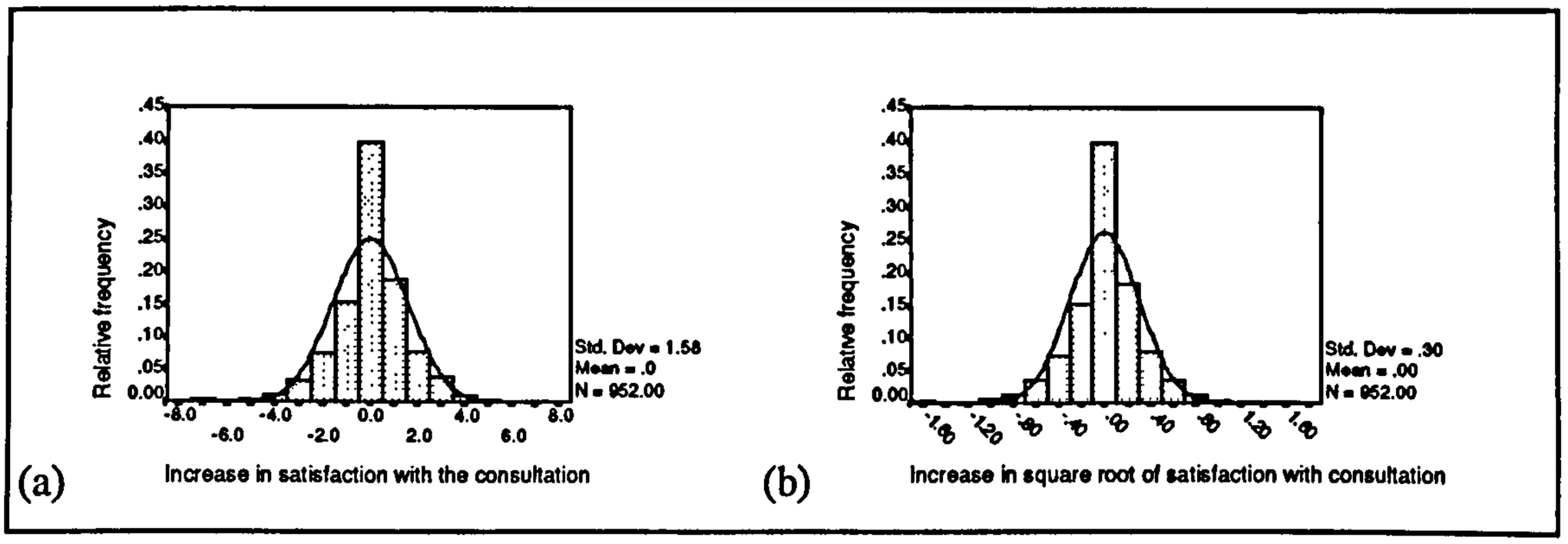
the maximum possible score at the initial interview, the potential of the scale to pick up any improvement is limited. The mean score was 8·2; the standard deviation of scores was 1·8. The distribution is clearly not Normal. If individual satisfaction scores are included in analyses of variance or Normal regression analyses it is likely that the underlying distributional assumptions will be violated.

Figure 10.2 Satisfaction with care delivered during the consultation: frequency distribution of responses at initial interview



For the two chronic conditions, change in satisfaction between initial and final interviews can also be considered as an outcome variable. The frequency distribution of responses is shown in Figure 10.3, Diagram (a). There was very little change in satisfaction. The distributions was very peaked about a mean of zero. A square root transformation was also considered (Diagram(b)) but it appeared to offer very little advantage over using simple differences in satisfaction score.

Figure 10.3 Change in satisfaction between initial and final interviews: frequency distribution of responses

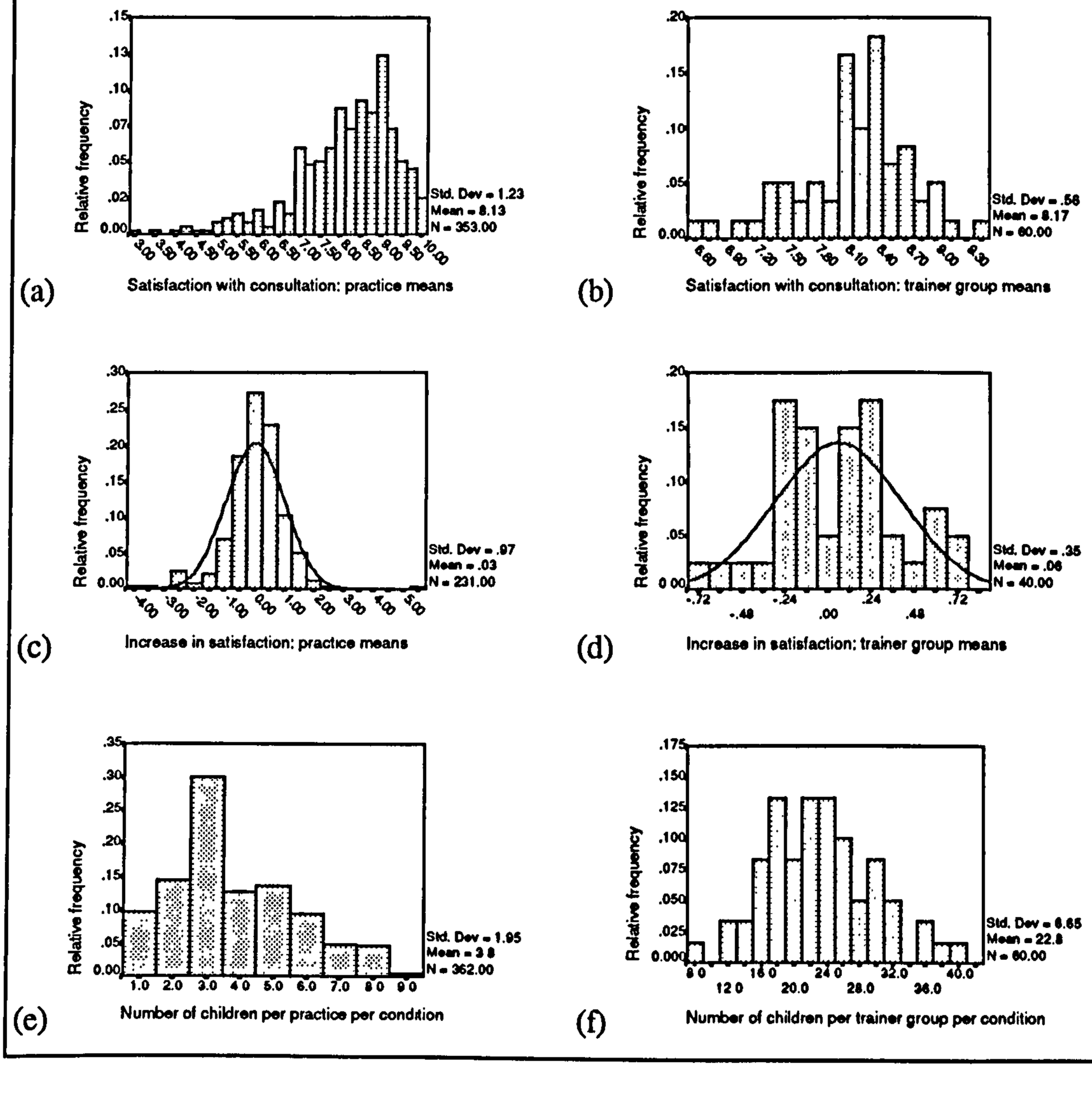


As an alternative to analysing responses from individuals, it is possible to average the satisfaction scores across groups of children. If this is done it is no longer possible to consider covariates such as age and gender of child but distributional assumptions may be more appropriate. For each condition, the mean satisfaction score was calculated (i) for all children registered with a particular practice and (ii) for all children registered with practices within a particular trainer group. The frequency distributions of responses are given in Figure 10.4.

Diagrams (a) and (b) are the means of the responses obtained at initial interviews. In both cases the mean level of satisfaction with the consultation was just over 8. In diagram (a) the left hand tail of the distribution was still fairly long. The reason for this is probably the small number of children interviewed in some practices for some of the conditions. The number of children interviewed in a practice for a particular condition ranged from 1 to 9—Diagram (e). The observations in the tail of the distribution are likely to correspond to groups including just one or two children whose parents reported that they were not satisfied with the care received. When averaging over trainer groups the number of children within a group ranged from 8 to 40—Diagram (f). As one would expect from application of the Central Limit Theorem, the tails in the distribution

of mean scores for trainer groups—Diagram (b)—are much shorter than those in Diagram (a).

Figure 10.4 Satisfaction with the consultation: mean scores for practices and trainer groups



For bedwetting and recurrent wheezy chest, the frequency distributions of the mean increase in satisfaction scores are shown for practices and trainer groups in Diagrams (c) and (d) respectively. Both distributions have a mean of approximately zero.

Studying the plots in Figure 10.4, it is not clear that there is any great advantage in analysing by trainer group rather than analysing by practice. The main advantage of analysing by practice is that we are better able to allow for the clustering of responses from parents within each practice. As each item in the scale asks about satisfaction with care provided by the doctor, it is likely that responses from parents whose children consulted a particular doctor will be correlated. Data is not available that would allow a doctor based analysis; analysis by the practice is the closest that can be achieved.

Diagrams (a) and (e) suggests that there will be some heteroscedasticity in the data. The standard error associated with each of the cell means will be a function of the number of observations in that cell. The normal assumption that errors are identically distributed with a constant variance is likely to be violated. The model GM + PRAC + COND (variation between practices and variation between conditions) was fitted to the data obtained from initial interviews. The histogram of standardised residuals (Figure 10.5) indicates that although the overall fit to a Normal distribution is reasonable there are a number of residuals in the left hand tail of the distribution that are larger than would be expected.

Weighted least squares was then used to fit the same model. The histogram of standardised residuals (Figure 10.6) now shows a very close fit to a Normal distribution. The number of very large residuals is considerably reduced. In the analysis reported below weighted least squares estimation was used to allow for the heteroscedasticity in the data noted above.

Figure 10.5 Unweighted regression: histogram of standardised residuals

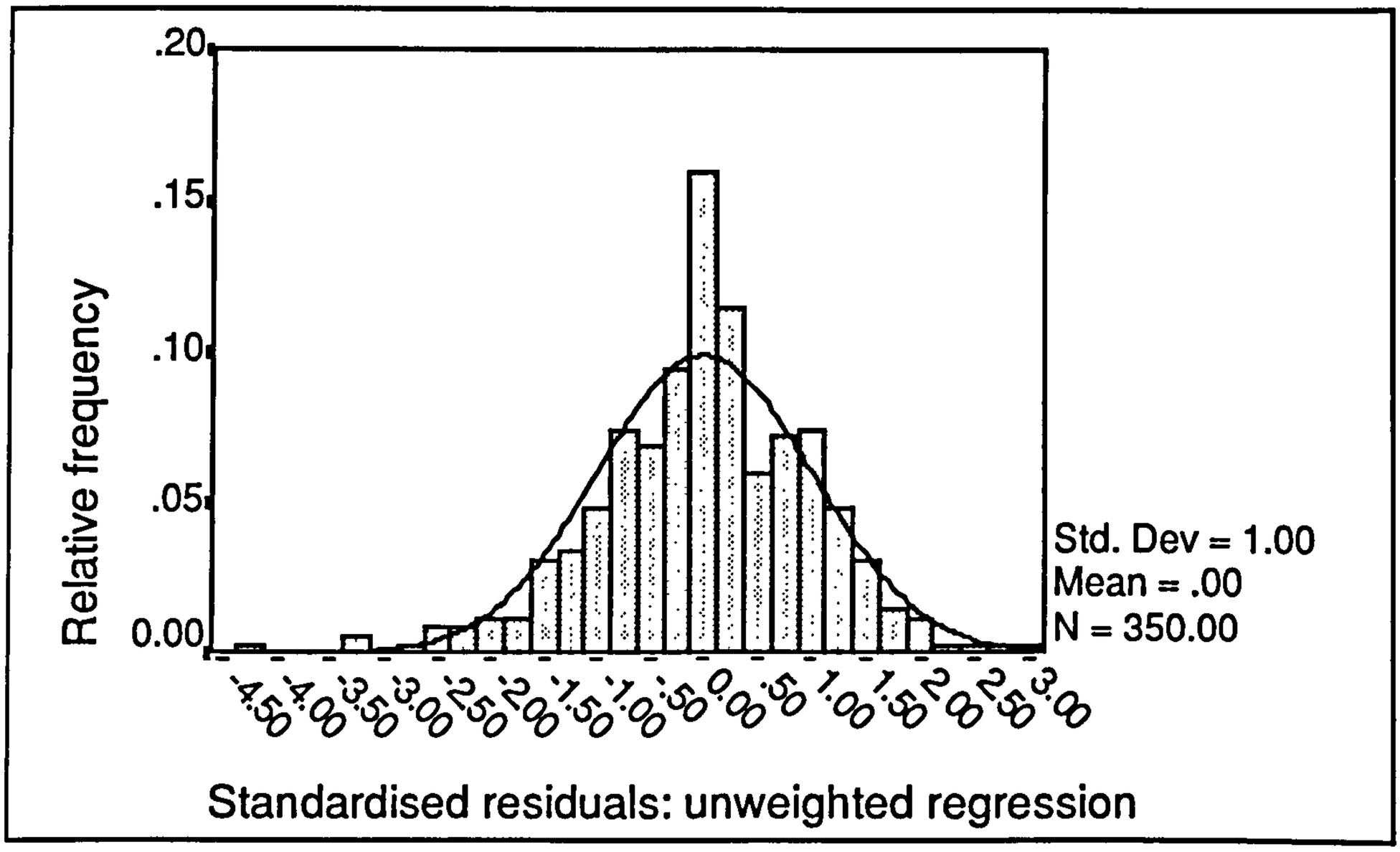
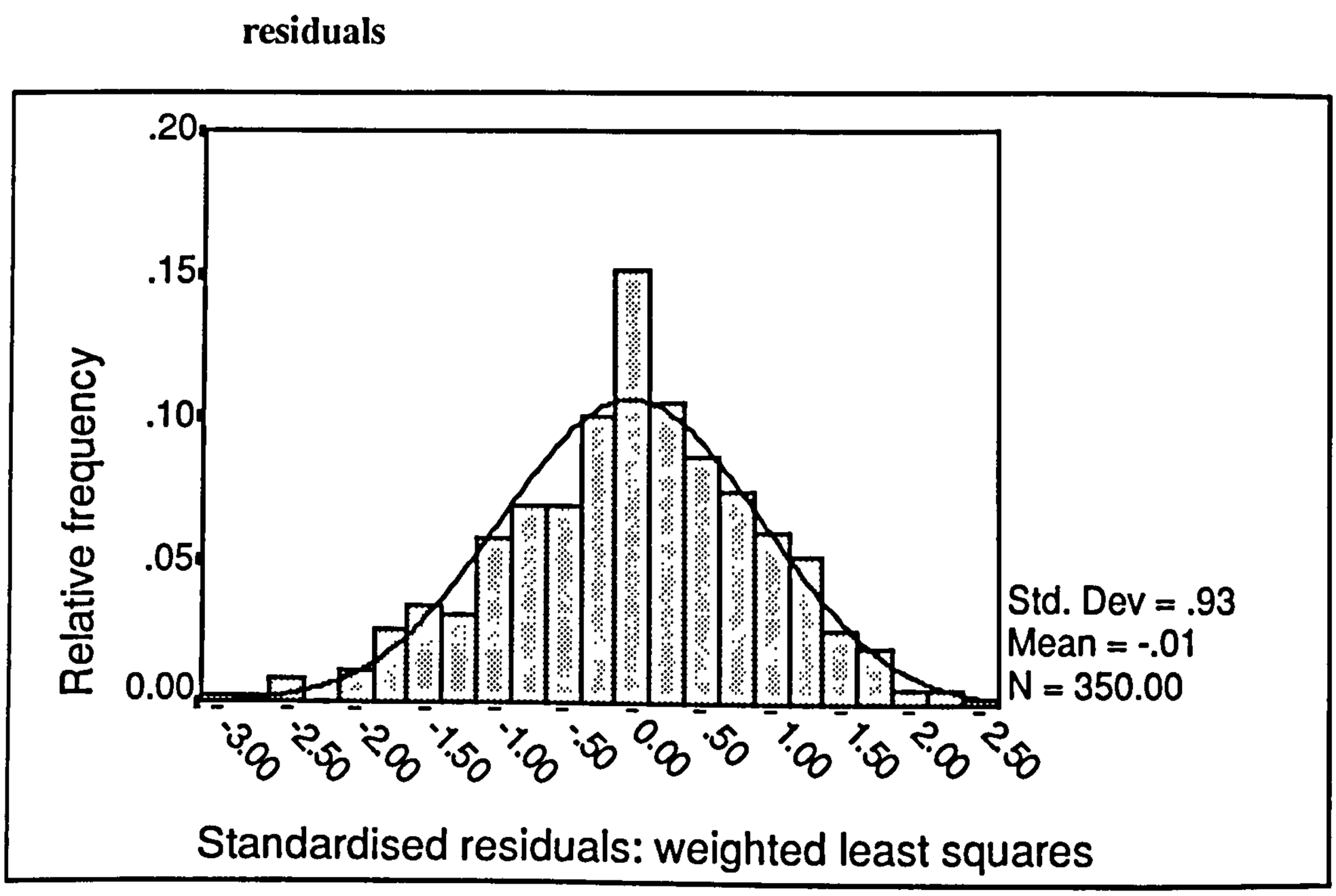


Figure 10.6 Weighted least squares estimation: histogram of standardised residuals



10.5.2 Results of analysis

The analysis of satisfaction with the care delivered during the consultation is shown in

Table 10.4 Satisfaction with care during the consultation: model selection

Model		Residual sum of squares	Residual degrees of freedom	Change in sum of squares	Change in degrees of freedom	Mean square	p
Number	Specification						
	Initial Interviews*						
1	GM	1620·8	349				
2	GM + TGRP	1549·3	340	71·5	9	7·95	0·08
3	GM + PRAC	993·3	289	627·5	60	10·5	<0·001
4	GM + PRAC + COND	913·5	287	79·8	2	39·9	<0·001
5	GM + PRAC + COND + PHAS	906·9	286	6·7	1	6·7	0·15
6	GM + PRAC + COND + PSTD	908·1	286	5·4	1	5·4	0·19
7	GM + PRAC + COND + CPST	907·8	284	5·8	3	1·9	0·61
	Final Interviews*						
8	GM	1824·5	349				
9	GM + TGRP	1741·8	340	82·7	9	9·2	0·07
10	GM + PRAC	1040·8	289	783·7	60	13·1	0·001
11	GM + PRAC + COND	960·2	287	80·6	2	40·3	0·001
12	GM + PRAC + COND + PSTD	960·2	286	0·0	1	0·01	0·96
13	GM + PRAC + COND + CPST	958·7	284	1·5	3	0·5	0·93
14	GM + PRAC + COND + PHAS	941·4	286	18·8	1	18·8	0·02
15	GM + PRAC + COND + PHAS + PSTD	938·9	285	2·5	1	2·5	0·38
	Change between initial and final interviews						
16	GM	655·2	230				
17	GM + TGRP	618·5	221	36·7	9	4·1	0·17
18	GM + PRAC	477·8	170	177·5	60	3·0	0·39
19	GM + COND	651·5	229	3·7	1	3·7	0·26
20	GM + PHAS	650·9	229	4·3	1	4·3	0·22
21	GM + PSTD	645·3	229	9·9	1	9·9	0·06
22	GM + CPST	641·5	227	13·8	3	4·6	0·18

* For acute cough, only one interview was conducted. Responses were included in both the analysis of initial satisfaction and the analysis of final satisfaction

Table 10.4. Responses recorded during initial interviews were analysed first.

Although variation between trainer groups was not significant (model 2) there was significant variation between practices (model 3). This was expected for the reasons given above—the questions asked about the care given by specified doctors. The only other significant effect was variation between conditions (model 4). Parents of children consulting with the two chronic conditions were slightly more satisfied than parents of children who consulted for acute cough. There was no evidence of any difference between the two phases of data collection (model 5) and no evidence of any effects of standard setting (models 6 and 7).

Analysis of responses from the final interviews gave very similar results. There was significant variation between practices (model 10) and study conditions (model 11). Although there was some evidence of a slight change between phases 1 and 2 (model 14 —the effect was significant only at the 5% level) there was no evidence of any effects of standard setting (models 12, 13 and 15).

In the final analysis, for the two chronic conditions, of change in satisfaction between initial and final interviews there were no significant effects at all. The p-value associated with fitting the standard setting term (model 21) is 0·06 but this is not close to the one percent level that would be regarded as evidence of a significant effect. A 99 percent confidence interval for the effect of standard setting, generated form model 21, indicates a change of between -0·84 and +0·13 units on the satisfaction scale. The mean satisfaction score is approximately 8, so the lower limit of this interval would represent a *decrease* in satisfaction score of around 10 percent. It is possible that this might be regarded as a clinically significant change.

10.6 Satisfaction with access to care and satisfaction with the practice.

The items which formed factors 2 and 3 (in the analysis reported earlier) were summated to form two more indices of satisfaction: satisfaction with the practice and satisfaction with access to care respectively. In both cases initial satisfaction, final satisfaction and change in satisfaction were analysed. For both indices, in the analysis of initial and final satisfaction there was significant variation between practices. This finding was expected. The practices were located in varied geographical locations; there were both urban and rural practices. Similarly, there was wide variation in the internal organisation of the practices. There was some evidence that parents of children suffering from bedwetting and recurrent wheezy chest were more satisfied than parents of children suffering from acute cough. There was no evidence of any effects of standard setting.

10.7 Effects of standard setting on parents' satisfaction with care

Ninety five percent confidence intervals for the effects of standard setting on parental satisfaction are given in Table 10.5.

Table 10.5 Effect of standard setting on parents' satisfaction with care

Measure	95% confidence intervals for:	
	Final satisfaction	Change in satisfaction
Satisfaction with consultation	(-0·53 to 0·21)	(-0·72 to 0·02)
Satisfaction with access to care	(-0·04 to 0·42)	(-0·03 to 0·52)
Satisfaction with the practice	(-0·02 to 0·62)	(-0·67 to 0·20)

These confidence intervals are fairly small. Satisfaction with the consultation is measured on a 10 point scale (Figure 10.2). The problem with this type of scale is assessing the magnitude of change that might be regarded as being clinically significant. But any change that may be attributed to standard setting is small compared to the

between subject variation that occurs in this scale. Similar results hold for the other two satisfaction scales.

Chapter 11

Outcome of care 3: health outcomes—evidence from postal questionnaires

11.1 Introduction

The effects of standard setting and other types of medical audit on the outcome of care for children, as reported by parents in a postal survey of their experiences and opinions, is reported in this chapter. Postal outcome questionnaires (POQs) were used to collect limited data on episodes of illness from parents of a large sample child patients. These postal questionnaires were designed to complement the more detail information collected by the parent interview survey. Postal outcome questionnaires were developed for only four of the five study conditions: acute cough, acute vomiting, itchy rash and recurrent wheezy chest. Because of the low prevalence of the fifth condition, it was decided to interview all the parents of children suffering from bedwetting. It was not therefore necessary to develop a postal questionnaire for this condition.

The administration of the postal questionnaires was very similar to that of the interview survey (described in Chapter 9). For the two acute conditions questionnaires were sent out once before standard setting and once after standard setting. For the two chronic conditions questionnaires were administered twice before and twice after standard setting. (The planned timetable for these activities is set out in Table 1.2.) In this way process and outcome for the two chronic conditions could be monitored over time, so

that long-term regimes and gradual improvements could be monitored. Only parents who normally saw a trainer were sent a questionnaire. It was impossible to be sure, however, that the child patient saw only that trainer, so the data reported in this chapter relates to the practice rather than to the individual doctor.

Each postal questionnaire was designed as a self-completion document. Each questionnaire had five sections: condition-specific and functional outcomes; past history; consultation; perceived causes of the condition; and parental satisfaction. The evaluation of the intervention concentrated on measures of outcome that took the same form on each questionnaire. These were:

1. a measure of clinical outcome - whether the study condition had cleared up and whether the child was now better;

2. a measure of parental anxiety - whether parents were still anxious about their child's condition.

3. A measure of parental satisfaction - how satisfied parents were with the care that their child received during the consultation.

The presentation of results in this chapter follows a similar format to those presented in earlier chapters. For each of the three measures described above, preliminary (univariable) results are given followed by the results of the generalised linear modelling.

11.2 Problems with administration of postal questionnaires

The timetable for research activities within each practice was described in Table 1.2. For most aspects of data collection, the timetable was strictly adhered to. The prevalence survey was sent out at the end of week 4, interviews were administered during week 10 and enhancement of medical records began as scheduled. Unfortunately administrative

reasons prevented the mailing of postal questionnaires in accordance with the timetable. The number of questionnaires sent out, by calendar month, for each of the study conditions is given in Table 11.1.

Table 11.1 Number of questionnaires mailed by calendar month, by study condition by subphase

Condition	Sub-phase	Calendar month that questionnaire was sent													Total
		Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	
Acute cough	1A			67	46		139	95	42	71	58	24	23		565
	2A									11	31		56	9	107
Acute vomiting	1A			50	27	7	74	41	10	29	16	18	102	5	379
	2A									32	26		16	3	77
Itchy rash	1A							865	126	299	234	187	49		1760
	1B						379	116	173	150		457	20		1295
	2A								1004	295	318		582	63	2262
	2B							144	207	478	614	178	26		1647
Wheezy chest	1A										977	853			1867
	1B						533	151	163	150		427			1442
	2A	63	123	151	138	148	127	167	450	318	215	27			1996
	2B							101	194	397	497	161			1367

For the two acute conditions, no questionnaires were sent out until November in subphase 1A and May in subphase 2A. For these two conditions, it was felt that no questionnaires should be sent out more than 6 weeks after the identification of the child by the doctor. The conditions were self limiting and it was felt that results would be difficult to interpret if this deadline were to be extended. As a result, the number of questionnaires sent out in subphase 2A was very small. The power of the study to detect

any changes in these conditions due to standard setting was clearly severely compromised.

Questionnaires for itchy rash and wheezy chest were not sent out until March and June respectively in subphase 1A. In subphase 2A, questionnaires for wheezy chest were sent out on schedule from the beginning of the period of data collection but questionnaires for itchy rash were not sent out until April. For these two conditions it was decided that it was appropriate to send out the questionnaires even after a considerable period of time had elapsed since the return of the prevalence survey. Large backlogs of questionnaires were sent out in March and June in 1985, February 1986, and April 1987. Because questionnaires for recurrent wheezy chest were not sent out until June in subphase 1A, the length of time between receipt of the two questionnaires was much shorter for most parents in phase 1 than in phase 2.

The problems that affected the administration of the questionnaires clearly had to be kept in mind when undertaking the analysis.

11.3 Response rates

Of the 14 764 questionnaires mailed, 11 376 were returned—a response rate overall of 77%. A breakdown of the number of questionnaires returned by study condition and by calendar month in which they were sent is given in Table 11.2.

Table 11.2 Number of questionnaires returned by calendar month in which they were sent by condition by subphase

Condition	Sub-phase	Calendar month that questionnaire was sent													Total
		Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	
Acute cough	1A			48	40		112	70	32	54	41	17	18		432
	2A									8	15		38	6	67
Acute vomiting	1A			40	22	5	46	30	5	25	13	15	65	5	271
	2A									17	20		6	1	44
Itchy rash	1A							703	101	232	186	148	40		1410
	1B						297	92	140	120		349	10		1008
	2A								742	227	258		417	54	1698
	2B							110	150	348	479	150	22		1259
Wheezy chest	1A										757	683			1472
	1B						419	120	127	116		317			1109
	2A	58	94	99	106	126	107	142	339	255	177	20			1567
	2B							82	147	294	372	127			1039

11.4 Clinical outcome

11.4.1 Descriptive statistics

The first question on each questionnaire asked about the current state of the condition. The general form of the condition was, "How is the condition now?" A successful outcome was defined as being the event that the condition was not causing a problem at the time the questionnaire was completed. The proportion of successful responses by study condition and subphase is given in Table 11.3. This table includes only those children with a chronic condition for whom there were valid responses in both subphase A and subphase B. Thus the denominators in this table, representing matched responses, are smaller than the number of questionnaire returned (Table 11.2).

Table 11.3 Clinical outcome: proportion of successful responses by study condition and subphase

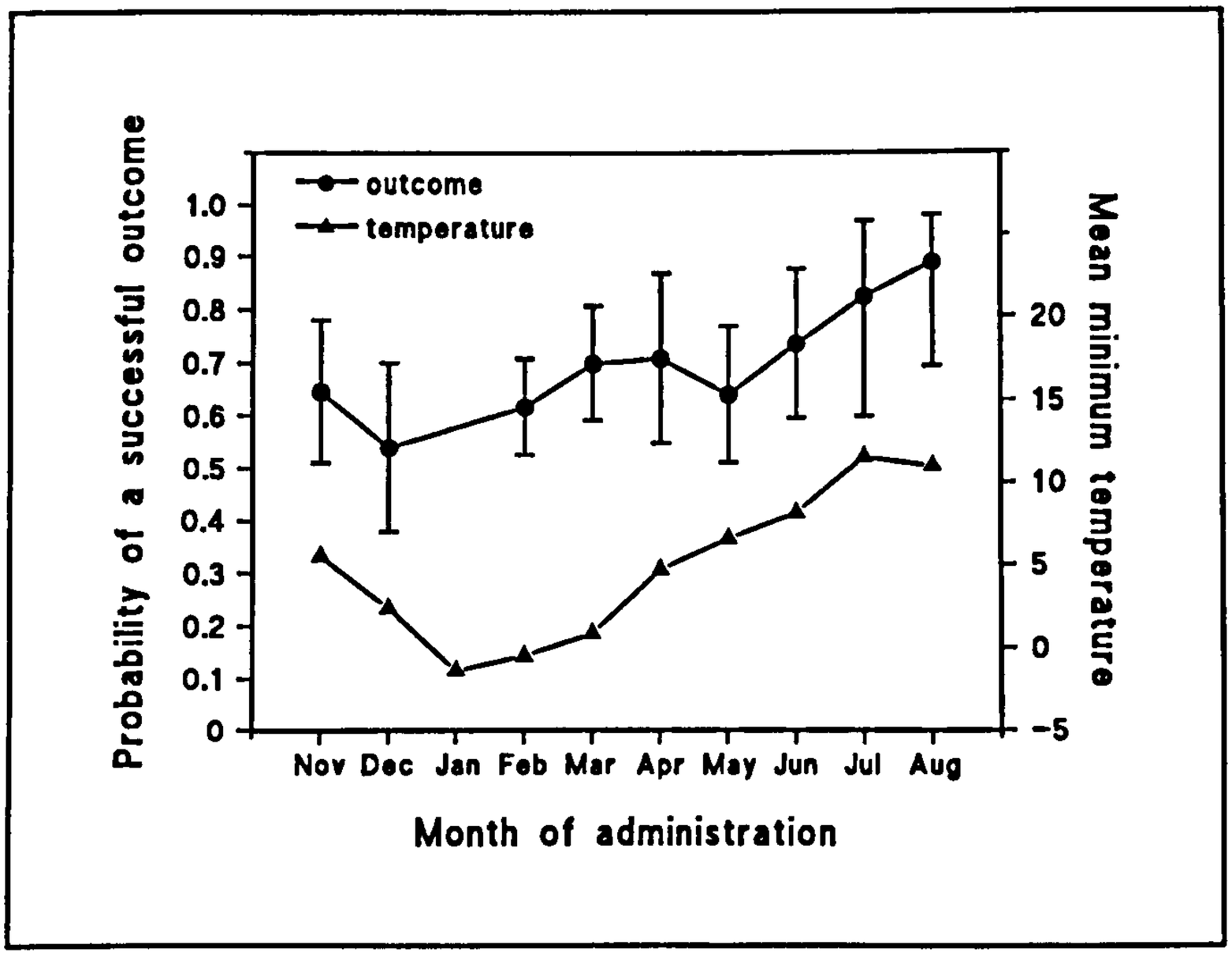
	Period of data collection			
Study condition	**Subphase 1A**	**Subphase 1B**	**Subphase 2A**	**Subphase 2B**
Acute cough	282/422 (66·8%)		50/63 (79·4%)	
Acute vomiting	217/262 (82·8%)		39/44 (88·6%)	
Itchy rash	487/938 (51·9%)	507/938 (54·1%)	565/1130 (50·0%)	614/1130 (54·3%)
Wheezy chest	614/1054 (58·3%)	618/1054 (58·6%)	520/1015 (51·2%)	606/1015 (59·7%)

For the two acute conditions the probability of a successful outcome appeared to be greater in subphase 2A than in subphase 1A. For acute cough, a simple chi-squared test indicated that the difference was almost significant at the five percent level. One possible explanation for this difference, might be that successful outcome is related to the time of year at which the questionnaires were sent out. In subphase 2A, no questionnaires were sent out during the Winter. It is possible that this conditions are less likely to clear up if the weather is cold.

This supposition was investigated by plotting the probability of a successful outcome by calendar month for cough questionnaires that were returned in subphase 1A (Figure 11.1). The mean minimum temperature (based on data from three sites spread across the North of England) for each calendar month was also plotted. The error bars represent 95% confidence intervals for the probability of a successful outcome. These error bars are fairly large and so some caution must be exercised when comparing the two lines. Even so there does seem to be some evidence of an association between

outcome and temperature—the higher the temperature the greater the likelihood of a successful outcome.

Figure 11.1 Probability of a successful clinical outcome for acute cough (based on questionnaires returned in subphase 1A) and mean minimum temperature by calendar month

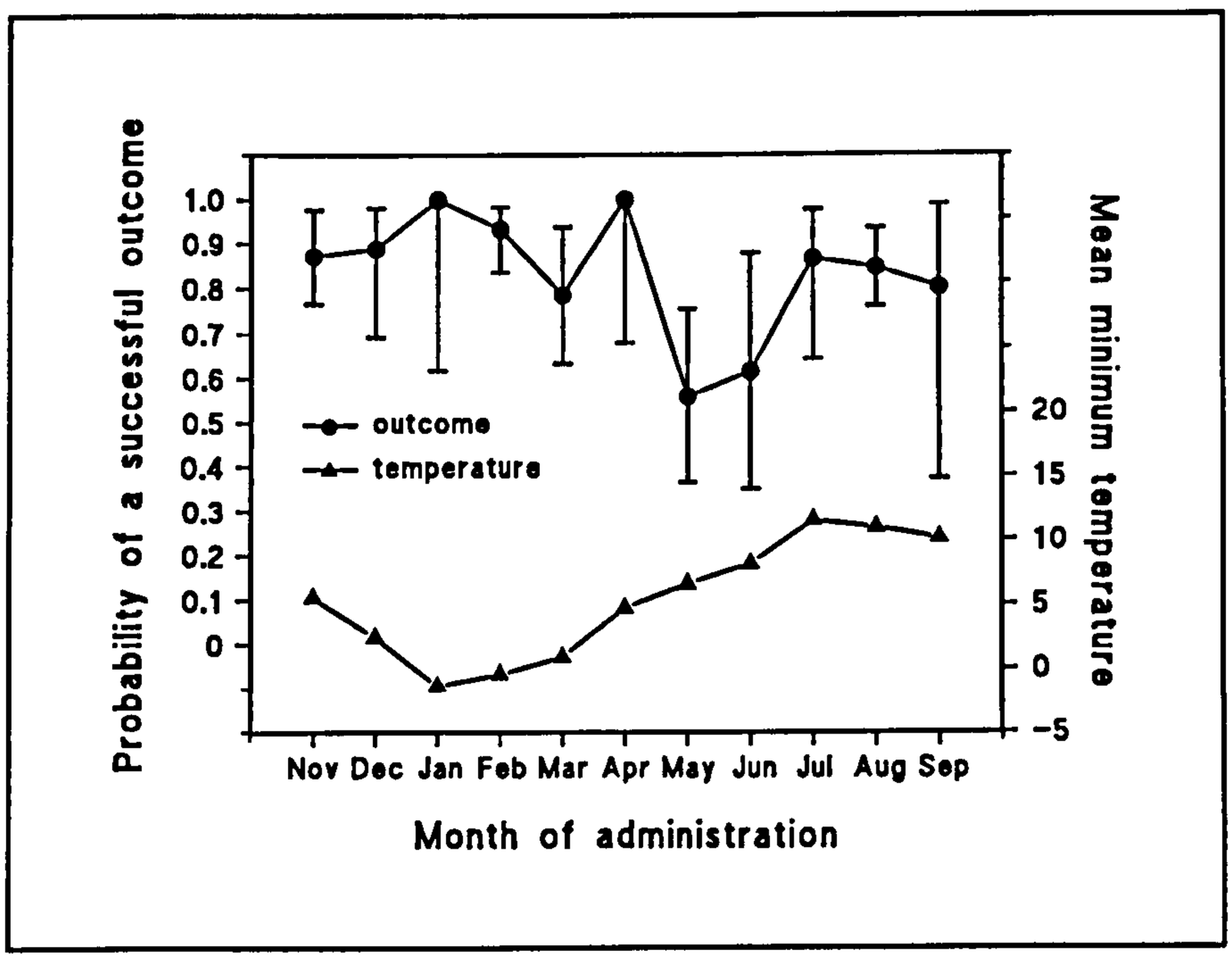


The difference in outcome for acute vomiting observed between phase 1 and phase 2 was not significant ($\chi_1^2 = 0.93$; $p = 0{\cdot}33$). A plot of clinical outcome by calendar month (Figure 11.2) does not indicate that there is any association between the likelihood of a successful response and temperature.

A striking feature of Table 11.3 is the drop in successful responses for recurrent wheezy chest in subphase 2A. A simple chi-squared test of the difference in outcome between subphase 1A and subphase 2A indicated that the difference was significant at the one percent level ($\chi_1^2 = 10.0$; $p < 0{\cdot}01$). A potential source of this difference in outcome is clearly the difference in the way in which the questionnaires for recurrent wheezy chest

were administered during each of the periods of data collection. More questionnaires were administered during the Winter months in subphase 2A than in subphase 1A.

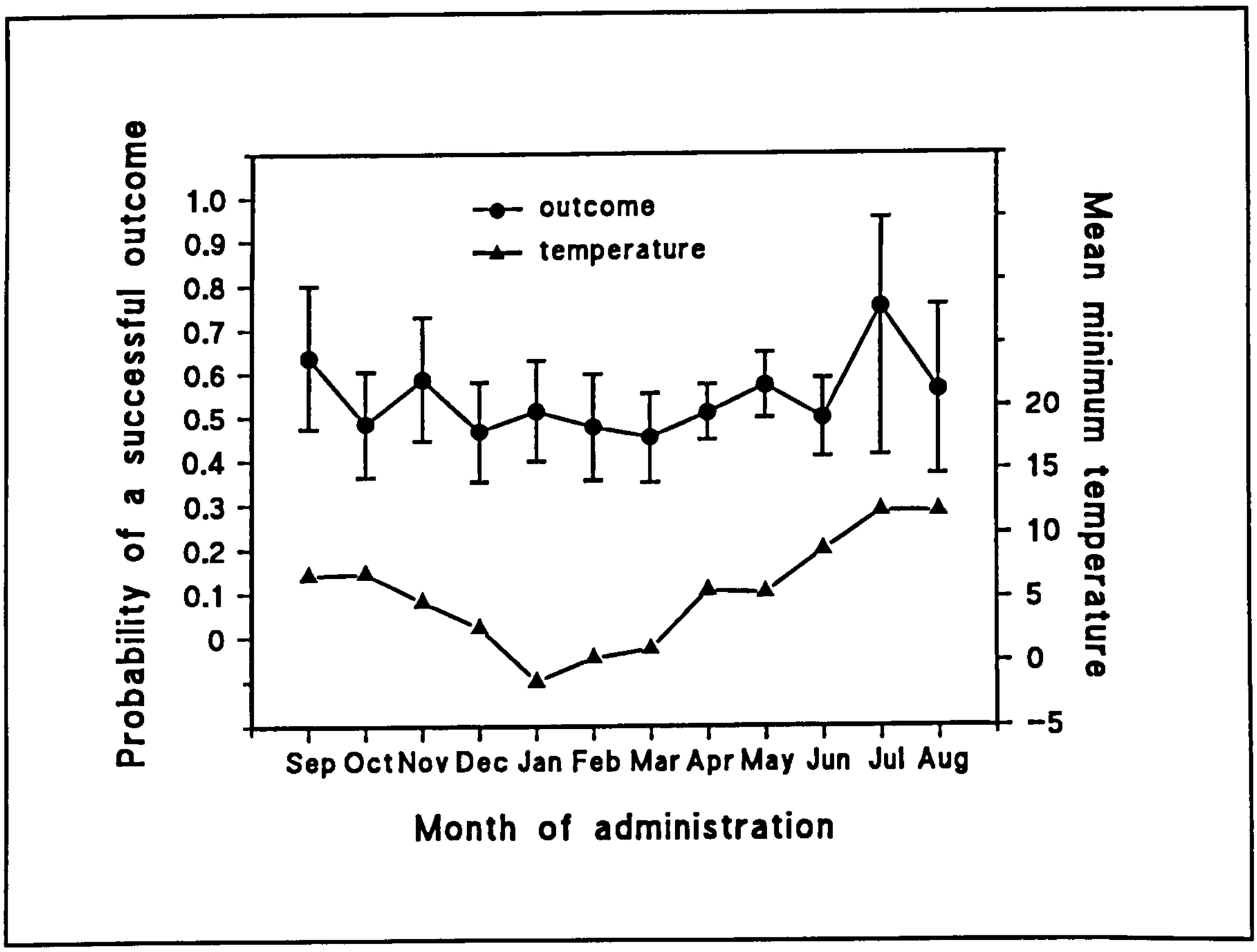
Figure 11.2 Probability of a successful clinical outcome for acute vomiting (based on questionnaires returned in subphase 1A) and mean minimum temperature by calendar month



A graph of the probability of a successful outcome by calendar month for wheeze questionnaires returned in subphase 2A is plotted in Figure 11.3. It is not clear from this figure, the extent to which the probability of a successful outcome is affected by seasonal influences. There does not appear to be a clear relationship between successful outcome and temperature.

An alternative reason that might explain the difference between outcome in subphase 1A and subphase 2A is the extended delay between administration of the prevalence survey and administration of the postal questionnaire in subphase 1A. This was caused by a delay in the production of the questionnaires. Evidence from interviews suggested that,

Figure 11.3 Probability of a successful clinical outcome for recurrent wheezy chest (based on questionnaires returned in subphase 2A) and mean minimum temperature by calendar month



for some children, their recurrent wheeze did clear up after a period of time. It is possible that it is not the probability of a successful outcome which was particularly low in subphase 2A but that the corresponding probability in subphase 1A was particularly high because of the delay.

11.4.2 Estimating the effects of standard setting on clinical outcome

In view of the problems explained above, it was decided to analyse responses from final questionnaires only. For recurrent wheezy chest, the difference in the way in which questionnaires were administered in subphase 1B and subphase 2B was much smaller than the difference between subphase 1A and subphase 2A. For itchy rash the way in

which questionnaires were administered in subphases 1B was roughly comparable with the way in which they were administered in subphase 2B—questionnaires were sent out between February and August in subphase 1B and between March and August in subphase 2B. For the two acute conditions, parents were only sent one questionnaire (in subphase 1A before standard setting and in subphase 2A after standard setting). Responses from these were included in the analysis. For children with acute cough, temperature was included as a covariate to allow for the differences in administration between the two data collection phases.

Details of the analysis are given in Table 11.4. After fitting the Grand Mean (model 1) the ratio of the residual deviance (6652) to the residual degrees of freedom (4927) was 1·35 suggesting that there may be some overdispersion in the data. In addition to comparing changes in deviance with the appropriate chi-squared distribution, the deviance ratio test described in Chapter 2 was used to assess improvements in fit when terms were added to the model.

Differences between conditions were expected and so this term was included in the model at an early stage (model 2). The improvement in fit was significant at the 0·1% level. There was also a difference between children who were reported to have asthma and other children with recurrent wheezy chest (model 3). When variation between practices (model 4) was incorporated into the model, the improvement in fit was not significant.

After dropping variation between practices, the reduction in deviance upon fitting a phase effect (model 5) indicated that there was some evidence of a difference in outcome between phase 1 and phase 2 (the improvement in fit was significant at the five percent level). But if we first fit a term that allows for an association between outcome for cough and temperature ('CTMP'—model 6) the difference between phases is no longer

Table 11.4 Clinical outcome at time of final questionnaire: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p*
Number	Specification					
1	GM	6652·0	4927			
2	GM + COND	6527·8	4924	124·2	3	<0·001
3	GM + COND + ASTH	6318·4	4923	209·4	1	<0·001
4	GM + COND + ASTH + PRAC	6230·0	4863	88·4	60	0·20
5	GM + COND + ASTH + PHAS	6312·2	4922	5·2	1	0·04
6	GM + COND + ASTH + CTMP	6309·3	4922	9·1	1	0·007
7	GM + COND + ASTH + CTMP + PHAS	6305·4	4921	3·9	1	0·08
8	GM + COND + ASTH + CTMP + PSTD	6309·2	4921	0·1	1	0·78
9	GM + COND + ASTH + CTMP + CPST	6306·3	4918	3·0	4	0·67
4	GM + COND + ASTH + PRAC	6230·0	4863			
10	GM + COND + ASTH + PRAC + PHAS	6223·9	4862	6·1	1	0·03
11	GM + COND + ASTH + PRAC + CTMP	6223·5	4862	6·5	1	0·02
12	GM + COND + ASTH + PRAC + CTMP + PHAS	6218·5	4861	5·0	1	0·05
13	GM + COND + ASTH + PRAC + CTMP + PHAS + COND·PHAS	6215·6	4858	2·9	3	0·52
14	GM + COND + ASTH + PRAC + CTMP + PHAS + PSTD	6218·3	4860	0·2	1	0·69
15	GM + COND + ASTH + PRAC + CTMP + PHAS + CPST	6215·9	4857	2·6	4	0·73
16	GM + COND + ASTH + PRAC + PSTD	6229·5	4862	0·5	1	0·53
17	GM + COND + ASTH + PRAC + CPST	6226·0	4859	3·9	4	0·55

* p values are based on the deviance ratio test described in Chapter 2 (section 2.4.2).

significant (model 7). This would indicate that the difference between phases noted above might be explained by a difference in outcome for children with cough. More questionnaires were sent out during the Winter months in phase 1, perhaps resulting in a poorer overall level of clinical outcome. There was no overall effect of standard setting (model 8) and no condition specific effects of standard setting (model 9).

The modelling strategy was then repeated but with variation between practices included in the model. This approach was not completely consistent with the analytic strategy set out in Chapter 2 as variation between practices (model 4) was not significant. This particular data set was highly unbalanced for the reasons mentioned earlier relating to problems with the administration of the questionnaire. The level of confounding of effects was potentially, therefore, very high. It is possible that an effect of standard setting might be masked by difference between practices. It was felt appropriate to refit the models after allowing for differences between practices. (A second reason why this approach was considered was that a test that did not take into account the overdispersion in the data would indicate that variation between practices was significant at the one percent level.)

The results are fairly similar. Comparing model 10 with model 4 there is some evidence of a difference between phase 1 and phase 2. If we include a temperature effect for children with cough (model 11) the improvement is significant but only at the 5% level. This is different to when this term was fitted before inclusion of variation between practices when the improvement was significant at the 1% level. This difference arises because the temperature effect is confounded with the term representing variation between practices. In phase 1, data on cough outcome was collected on children from 52 practices. In phase 2, the data came from only 25 practices—a difference of 27. Data collection was scheduled for the Winter months for many of the practices from which no data was actually collected. Further the difference between phases can no longer be explained by the temperature effect (model 12)—'CTMP' and 'PHAS' are each significant at the 5% level regardless of the order in which they are fitted. In models 14 and 15 these terms are retained while possible effects of standard setting are investigated. In models 16 and 17 both effects are omitted. Regardless of whether these effects were included in the model, there was no general effect of standard setting

(model 14 and model 16) and there were no condition specific effects of standard setting (models 15 and 17).

There is no obvious choice of "best model". There is little to choose between models 4, 6 and 12. Of these the most parsimonious is 'GM + COND + ASTH + CTMP'. If we fit this model, the regression coefficient of the temperature term is (on the logistic scale; with temperature measured in degrees Centigrade) equal to 0·07 with 95% confidence interval from 0·02 to 0·12. A breakdown of the raw data corresponding to the remaining terms in the model is given in Table 11.5.

Table 11.5 Proportion of successful responses by study condition and by whether asthma was a cause of the child's recurrent wheezy chest

				Recurrent wheezy chest	
	Acute cough	Acute vomiting	Itchy rash	Non-asthmatics	Asthmatics
Proportion of successful responses	$\frac{332}{485}$ (68·5%)	$\frac{256}{306}$ (83·7%)	$\frac{1121}{2068}$ (54·2%)	$\frac{869}{1200}$ (72·4%)	$\frac{355}{869}$ (40·9%)
95% confidence interval	(64·3 to 72·6)	(79·5 to 87·8)	(52·0 to 56·4)	(69·9 to 74·9)	(37·6 to 44·1)

As was found in the analysis of data from parent interviews, there was a big difference between responses from those parents who thought that asthma was the cause of their child's chest problem and responses from other parents of children with wheezy chest. Only 41% of parents of children who were believed to be asthmatic reported that their child had not had a chest problem during the month preceding the final interview. The corresponding figure for parents of other children with wheezy chest was just over 72%.

11.5 Parental anxiety

On each questionnaire, there was an item that asked whether there was anything about their child's condition that was still causing the parent anxiety or concern. An answer of "no" was regarded as a successful outcome; responses of "yes" or "not sure" were classified as unsuccessful outcomes. Anxiety was analysed as a binary variable. Responses to this question broken down by phase and study condition are given in Table 11.6.

Table 11.6 Parental anxiety: proportion of successful responses broken down by study condition and subphase

	Period of data collection			
Study condition	**Subphase 1A**	**Subphase 1B**	**Subphase 2A**	**Subphase 2B**
Acute cough	359/416 (86·3%)		53/62 (85·5%)	
Acute vomiting	216/237 (91·1%)		39/39 (100%)	
Itchy rash	741/934 (79·3%)	787/934 (84·3%)	853/1080 (79·0%)	933/1080 (86·4%)
Wheezy chest	815/1053 (77·4%)	862/1053 (81·9%)	768/1004 (76·5%)	830/1004 (82·7%)

For the reasons given in the section describing the analysis of clinical outcome, only responses from final interviews were included in the analysis. The analysis is reported in Table 11.7. After fitting the Grand Mean (model 1) the ratio of residual scaled deviance to the residual degrees of freedom is 0·86. When, as in this case, this ratio is less than unity, comparison of changes in deviance with the percentage points of the appropriate chi-squared distribution is the more conservative of the two proposed tests of model improvement. The significance levels reported in Table 11.7 are those associated with this more conservative test.

Table 11.7 Parental anxiety—final questionnaires: model selection

Model		Residual deviance	Residual degrees of freedom	Change in deviance	Change in degrees of freedom	p
Number	Specification					
1	GM	4155·5	4824			
2	GM + COND	4129·9	4821	25·6	3	<0·001
3	GM + COND + ASTH	4046·5	4820	83·4	1	<0·001
4	GM + COND + ASTH + CTMP	4046·1	4819	0·4	1	0·53
5	GM + COND + ASTH + PSTD	4044·2	4819	2·3	1	0·13
6	GM + COND + ASTH + CPST	4039·3	4816	7·2	4	0·13
7	GM + COND + ASTH + PHAS	4041·2	4819	5·3	1	0·02
8	GM + COND + ASTH + PHAS + PSTD	4040·6	4818	0·6	1	0·44
9	GM + COND + ASTH + PHAS + CPST	4035·5	4815	5·8	4	0·21
10	GM + COND + ASTH + PRAC	3966·2	4760	80·4	60	0·04
11	GM + COND + ASTH + PRAC + PSTD	3962·5	4759	3·7	1	0·05
12	GM + COND + ASTH + PRAC + CPST	3957·6	4756	8·5	4	0·07
13	GM + COND + ASTH + PRAC + PHAS	3960·0	4759	6·1	1	0·013
14	GM + COND + ASTH + PRAC + PHAS + PSTD	3958·8	4758	1·2	1	0·27
15	GM + COND + ASTH + PRAC + PHAS + CPST	3953·9	4755	6·1	4	0·19

There was significant variation between conditions (model 2) and a big difference between children for whom asthma was given as a cause of their chest trouble and other children with recurrent wheeze (model 3). There was no evidence of a temperature effect for children suffering from acute cough (model 4). There was some evidence of a difference in outcome between phases 1 and 2 (models 7, and 13) and variation between practices (model 10). Both effects were significant at the 5% level. Because of the unbalanced nature of the experimental design, it is likely that these effects are confounded with the intervention effect. General and condition specific effects of standard setting were therefore fitted with each possible combination of phase and practice effects (models 5 and 6, 7 and 8, 11 and 12 and 14 and 15). The improvement in fit was greatest for models 11 and 12 when variation between practices was included

in the model but no allowance was made for a general change between phases 1 and 2. But the significance level was much greater than the criteria of one percent specified in the analytic strategy. The final conclusion therefore is that there was no evidence that standard setting had any effect on parental anxiety.

Again the best choice of final model is not clear. The two models that would seem reasonable are model 3 and model 13. Both models allow for a difference between conditions and a difference between children with asthma and other children with recurrent wheezy chest. Model 13 also allows for variation between practices and a difference between the two periods of data collection.

The proportion of cases with a successful outcome broken down by study condition and phase was presented in Table 11.6. Understandably parents of children suffering from a chronic condition are more anxious or concerned than parents of children presenting with an acute condition. The data for children with recurrent wheezy chest (final interviews only) are broken down by whether the chest trouble was caused by asthma in Table 11.8. Parents of children with chest trouble caused by asthma were less likely to provide a successful response (indicating that they were more anxious or concerned about their child's condition) than parents of other children.

Table 11.8 Proportion of parents with children suffering from recurrent wheezy chest who were not anxious or concerned about their child's condition by cause of chest trouble and period of data collection

Period of data collection	Cause of chest trouble: Asthma		Cause of chest trouble: Other causes	
Subphase 1B	273/379	(72·0%)	589/674	(87·4%)
Subphase 2B	360/486	(74·1%)	470/518	(90·7%)

In both Tables 10.6 and 10.8 there appears to be a slight increase in the proportion of successful response from phase 1 to phase 2 across all groups of children. The increase is small and was significant only at the five percent level. The possible causes of such an increase are not known but there was no evidence to suggest that they were caused by the study interventions.

11.6 Satisfaction with the consultation

The questions used to assess satisfaction of care on the interview schedule (Chapter 10) were also included on the postal questionnaires. The five items that asked parents about their satisfaction with the care provided during the consultation were summed to form a single index of satisfaction. The internal reliability of this scale was 0·80. Once again, for the reasons given earlier, only responses to the final questionnaires were analysed.

The method of analysis was similar to that used in the analysis of data from parent interviews (Chapter 10). Mean satisfaction scores were calculated for responses relating to children with a particular condition for each practice for each phase of data collection. A distinction was made between children whose chest trouble was thought to be caused by asthma and other children with recurrent wheezy chest. Practice means were then

analysed using a weighted least squares procedure—the weights used were the number of children corresponding to each particular cell. The results of this analysis are given in Table 11.9.

Differences between conditions were not significant (model 2) but there was a difference between children with chest trouble caused by asthma and all other children (model 3). When variation between practices was included (model 4) there was a very large reduction (from 3123 to 1529) in the residual sum of squares for the loss of 60 degrees of freedom - an improvement significant at the 0·1% level. There was no difference

Table 11.9 Satisfaction with the consultation - final questionnaires: model selection

Model		Residual sum of squares	Residual degrees of freedom	Change in sum of squares	Change in degrees of freedom	Mean square	p
Number	Specification						
1	GM	3172·4	490				
2	GM + COND	3168·5	487	3·9	3	1·3	0·90
3	GM + ASTH	3123·3	486	45·2	1	45·2	0·008
4	GM + ASTH + PRAC	1528·5	429	1599·2	60	26·7	<0·001
5	GM + ASTH + PRAC + PHAS	1522·0	428	6·5	1	6·5	0·18
6	GM + ASTH + PRAC + PSTD	1514·2	428	14·3	1	14·3	0·045
7	GM + ASTH + PRAC + COND	1509·5	426	4·7	3	1·6	0·73
8	GM + ASTH + PRAC + COND + CPST	1477·2	422	32·3	4	8·1	0·06
9	GM + ASTH + PRAC + STND	1511·5	428	17·0	1	17·0	0·03
10	GM + ASTH + PRAC + STND + PSTD	1508·4	427	3·0	1	3·0	0·36
11	GM + ASTH + PRAC + STND + COND	1490·0	425	21·4	3	7·1	0·11
12	GM + ASTH + PRAC + STND + COND + CPST	1468·8	421	21·2	4	5·3	0·20

between phase 1 and phase 2 (model 5). There was some evidence of a standard setting effect but the improvement in fit of the model was significant only at the 5% level (model 6). Examination of the parameter estimates for this model suggested that satisfaction was slightly less in phase 2 for consultations with doctors who set standards than with other consultations (a difference of 0·30 with 95% confidence interval 0·08 to 0·52).

Variation between conditions was fitted (model 7) so that a separate effect of standard setting could then be fitted for each condition (model 8). The improvement in the fit of the model was not significant. Any effect of standard setting was consistent across all four conditions. The apparent effect of standard setting was further investigated by removing the term 'PSTD' from model 6 and fitting the effect 'STND' (model 9). This term fits a difference between doctors who set standards and all other doctors. This difference is assumed to be the same in phase 1 and phase 2 and can therefore not be attributed to standard setting. This improvement generated by adding this term was also significant at the five percent and the reduction in the residual sum of squares was even larger than when the term 'PSTD' was fitted. When the term 'PSTD' is now added (model 10) the improvement in fit is no longer significant. This would suggest that although there is some evidence of a difference between doctors who set standards and other doctors, this difference is the same before and after standard setting. That is parents of children with cough who consulted with doctors who set a standard for cough were less satisfied than other parents whose children consulted for cough (and likewise for the other four conditions) but this difference was the same in phase 1 and phase 2. There is no obvious explanation for why there should be such a difference in phase 1 because that was before the group to which the trainer belong had been randomised to a condition. That is, in phase 1 the trainer had no prior knowledge that he or she would be setting a standard for a particular condition. This result was consistent across all five conditions. Such a difference is very difficult to interpret and it clearly cannot be

attributed to an effect of the intervention. It is significant only at the five percent level it is probably best omitted from the model (the only rational explanation is that it is a chance finding).

The model that best represents the data is model 4, "GM + ASTH + PRAC". Examination of the parameter estimates of this model indicates that parents of children with chest trouble caused by asthma are more satisfied than other parents. The difference was 0·32 with 95% confidence interval from 0·18 to 0·47. The reason for this difference is not clear. It is likely that children with a diagnosis of asthma are amongst those in the study population who are affected most severely by their condition. It is possible that doctors give a higher priority to the treatment of these children. A difference of the same order of magnitude was found for parents interviewed - parents of children who were believed to have asthma had a higher mean satisfaction score than parents of other children with recurrent wheeze - but because the sample size was very much smaller the difference was not significant in the analysis reported in Chapter 10.

11.7 Effects of standard setting

Ninety five percent confidence intervals for the effects of standard setting are given in Table 11.10. For each outcome variable analysed, there was no single model that best fitted the data. For each variable, estimates of the effects of standard setting are therefore based on two alternative sets of models. There are two models that correspond to each of the rows in Table 11.10. The first model was used to generate an estimate of an overall effect of standard setting; the second model was used to generate estimates of condition specific effects of standard setting.

Clinical outcome and parental anxiety were binary variables; the effects of standard setting are given in the form of odds ratios. Satisfaction with the consultation was

analysed as a continuous variable with a normal error structure; direct estimates of the effects of standard setting are therefore given.

Table 11.10 Ninety five percent confidence intervals for the effects of standard setting on outcome variables

Outcome	Table and models		Overall	Condition specific effects of standard setting			
				Acute cough	Acute vomit	Itchy rash	Wheezy chest
Clinical outcome	11.4	14 and 15	(0·75 to 1·21)	(0·43 to 3·17)	(0·14 to 816)	(0·62 to 1·18)	(0·70 to 1·56)
	11.4	16 and 17	(0·86 to 1·35)	(0·70 to 4·63)	(0·15 to 1120)	(0·70 to 1·29)	(0·78 to 1·71)
Parental anxiety	11.7	8 and 9	(0·83 to 1·53)	(0·22 to 1·70)	(0·01 to 2×10^{4})	(0·62 to 1·37)	(0·99 to 2·63)
	11.7	12 and 13	(0·86 to 1·68)	(0·22 to 1·84)	(0·00 to 5×10^{6})	(0·62 to 1·52)	(1·04 to 3·16)
Satisfaction with care	11.9	6 and 8	(-0·41 to -0·00)	(-0·92 to +0·60)	(-0·87 to +2·48)	(-0·32 to +0·25)	(-0·83 to -0·15)
	11.9	10 and 12	(-0·36 to 0·13)	(-0·81 to +0·72)	(-0·74 to +2·62)	(-0·24 to +0·39)	(-0·75 to -0·02)

In general the confidence intervals for a condition specific effect of setting a standard for acute vomiting are much wider than confidence intervals for the effects of setting standards for the other three conditions. The estimates for this condition are based on very few observations.

Estimates for the effects of standard setting on clinical outcome are unremarkable. The models in the first row (Table 11.4, models 14 and 15) include terms that represent a general change between phases 1 and 2 and an effect of temperature on outcome for children suffering from acute cough. When these terms are removed (Table 11.4, models 16 and 17) the estimates of the effects of standard setting are increased very slightly.

Interval estimates for the effects of standard setting on parental anxiety are also fairly unremarkable. In the second row, the models on which the estimates are based (models 12 and 13 in Table 11.7) include a term representing variation between

practices. These intervals are slightly wider than those given in the first row. The estimates in the final column indicate that setting a standard for recurrent wheezy chest may have caused an improvement (a decrease) in parental anxiety. There is however no justification for regarding this result as being significant—when estimating multiple 95% confidence intervals the probability that at least one will indicate a significant effect purely by chance is considerably larger than 5%.

Estimates of the effect of standard setting on parents' satisfaction with the care provided during the consultation are given in the final two rows of Table 11.10. In the bottom row the additional term 'STND' is included in the models. This was fitted because, in consultations with a trainer who had set a standard for the condition, satisfaction with care provided seemed to be lower in both phases of the study. It is extremely difficult to suggest a logical explanation for why this should be so. It is most likely that this is just a chance occurrence. Examining the interval estimates in both rows, indicates nothing unusual for three of the conditions. Setting a standard for wheezy chest, however, would appear to cause a reduction in satisfaction in consultations for that condition. In previous chapters it was found that

1. doctors who set a standard for recurrent wheezy chest, changed their prescribing behaviour for that condition
2. setting a standard for wheezy chest had a beneficial effect on clinical outcome for children consulting with that condition.

While a reduction in parental anxiety might be regarded as being consistent with these findings it is difficult to see why there should be a reduction in parents' satisfaction.

Once again the most likely explanation is that is just a chance occurrence.

Chapter 12

Discussion

12.1 Introduction

The evaluative component of a major study into the effects of performance review (medical audit) on the behaviour of doctors and on the resulting outcome of care for their patients has been described in this thesis. A replicated Latin square design was used (Table 1.1). Data were collected before and after the educational intervention. Although the underlying experimental design was fairly simple, there were a number of difficulties and issues in its implementation that caused problems when the analysis was undertaken.

The analyses in this thesis have been presented in a logical order. The prevalence survey was used to help identify children suffering from the five study conditions. It was possible to use the results of this survey to make some assessment of whether there were any global changes in perceived prevalence rates or reported consultation rates for the entire study population. Any changes, particularly if they appeared to be due to the conduct of the study, would need to be taken into account in subsequent analysis of the other data sets. This analysis was reported first. The analysis of the process data sets was reported next. It is natural to investigate whether the intervention caused doctors to change their behaviour. If there was no behaviour change, then there was no good

reason why there should be any change in outcome. Analysis of the two outcome data sets was therefore reported last.

12.2 Prevalence survey

Data relating to reported prevalence and consultation rates were available from every practice for each condition for both trainers and partners. Although data were collected from individual practices, the unit of randomisation in the study design was the trainer group. In order to make the trainer group the unit of analysis, data from individual practices were combined to yield a summary statistic for the trainer group. This was done for each of the ten trainer groups, for each of the five conditions for both trainers and partners giving a total of 100 observations ($10 \times 5 \times 2$) corresponding to the 100 cells of the four replicates of the 5×5 Latin square described in Chapter 2. With no missing data, it was natural to use classical methods of analysis in which the model defined by equation (3.1) was fitted to the data.

As the main effects were all orthogonal, it was possible to partition the total sums of squares into the components for the main effects and interaction terms as described in section 3.5. There was no confounding; results were unambiguous. The main findings were a general reduction in reported prevalence of the study conditions between the two surveys and a difference in consultation behaviour between patients attending trainers and patients attending their partners. No effects of standard setting or performance review were noted.

The design of the study and method of analysis proved powerful enough to permit the detection of two fairly obscure changes. First, an apparent change in the prevalence of acute cough within just one of the 64 participating practices was detected. This turned out to be the spurious by-product of the failure of the practice computer. Secondly, a

change in children's consultation patterns within five other practices—all located near Sellafield, a nuclear fuel reprocessing plant in West Cumbria—was identified. This change probably resulted from public concern arising from the apparent increase in the incidence of leukaemia near the plant. At the time it was argued that, as the study was powerful enough to detect the two obscure changes described in section 3.8, any changes arising due to the interventions must be small in comparison (Ho et al 1990). The results were consistent with expectations (no changes had really been expected) and were accepted.

It was felt that the analysis of the prevalence survey would serve as a blueprint for the analysis of the process and outcome data sets but, as can be seen from the analysis presented in subsequent chapters, this assessment was incorrect. The main reason for this was the presence of additional (and unwelcome) features in the process and outcome data sets that necessitated the use of a very different analytic strategy. However, a second possible reason is that perhaps the analysis of the prevalence survey was not as comprehensive as it might have been. In particular, insufficient consideration was given to the issue that a number of the effects of critical interest did not coincide exactly with specific main effects or specific interactions fitted in the model defined by equation 3.1.

When the analysis was undertaken, no evidence had been collected about the relative effectiveness of the different levels of the intervention. The only treatment effect investigated was one with four degrees of freedom corresponding to differences between all five levels of medical audit. Evidence from interviews with trainers now suggests that one of the interventions—setting a clinical standard for a condition—was likely to have a much greater impact on their behaviour than the other types of audit. Indeed, due to the poor quality of presentation, many trainers reported that they were unable to understand the feedback that they received. It is likely therefore that a number of the interventions did not produce any discernible effect. In analyses of subsequent data sets a contrast

with one degree of freedom was used to investigate whether there was a specific effect of standard setting; the other four levels of intervention were grouped together. A similar term could be fitted to the prevalence data.

In the case of the prevalence data, there is a further complication. Data corresponding to trainers' partners were also collected. In practice, the extent to which partners were involved in the study, was left to the discretion of individual trainers. Trainers were free to discuss the study with their partners but the partners themselves did not attend small group meetings and did not take an active role in the setting of clinical standards. It is likely that many of the partners had no active involvement with any aspect of the study at all. Consequently any intervention effects were likely to be substantially greater for trainers than their partners. But the estimate of the effect of the intervention (main effect C in Tables 3.1 and 3.4) is based on data corresponding to all doctors. Thus any effect due to a change in behaviour of the trainers may have been diluted because of no change in behaviour among their partners. There is an interaction term ($\gamma\omega$) in equation 3.1 corresponding to different intervention effects for trainers and partners but this term has four degrees of freedom - corresponding to the case where partners behave differently from trainers for each of the five levels of medical audit. If we combine the evidence that setting a standard was the only intervention likely to affect clinical practice and that only trainers took part in standard setting, it would be appropriate to fit a contrast with just one degree of freedom comparing prevalence and consultation rates for trainers who set a standard for the relevant condition with all other doctors. Using GLIM this procedure involves the computation of a new indicator variable. While this technique was used extensively in the analysis of the process and outcome data sets, it was not used in the analysis of the prevalence survey. Further analysis would be appropriate. Unfortunately due to the corruption of a key data file which occurred when the university changed its mainframe computer system this is now not possible.

12.3 Process of care

The first data set analysed was that relating to the process of care—the information abstracted by fieldworkers from the children's medical records. The first major problem that was identified was the loss of balance in the design that resulted from two causes. The prevalence of bedwetting was lower than the estimate obtained from the pilot study (Russell et al, 1986). As a result, many doctors did not identify their quota of children with this condition. The second cause was, apparently, a diminishing enthusiasm of doctors for the onerous task of enhancing clinical records—fewer children had their records enhanced in phase 2 than phase 1 (Table 4.2). In fact, for some conditions, a number of doctors identified no cases at all during phase 2 (Table 6.3). Careful consideration was given as to how this lack of balance might affect the usual Latin square analysis where orthogonality is assumed.

The usual method of analysis for this study design would be to (i) calculate, for each combination of doctor and study condition, a summary statistic (such as a group mean) for each of the phases of data collection (one before and one after standard setting); (ii) calculate their difference; (iii) aggregate these differences to obtain a summary statistic for each combination of trainer group and study condition (corresponding to the cells of the Latin square given in Table 2.1); and (iv) then fit a statistical model such as that defined by equation 2.1 to the aggregated data.

The shortfall of cases identified in phase 2 was not consistent across all doctors. In the extreme case, some doctors failed to identify any children for some of the conditions in phase 2 and one of the summary statistics in step (i) could not be computed. Data from these doctors would therefore need to be excluded from the remaining steps in the classical analysis. In other cases, the variability between doctors in the shortfall between phase 1 and 2 meant that potentially there might be considerable variability in the standard error associated with the difference calculated in step (ii) and consequently in

the standard errors associated with observations in each cell calculated in step (iii). As a result, assumptions relating to homogeneity of variance made in step (iv) could not be justified.

This lack of balance also necessitated careful consideration of possible covariates. The shortfall in identification of cases by some doctors resulted in the possibility that demographic characteristics of the sample were not the same in both phases of data collection. The proportion of children from rural areas say or from specific socio-economic backgrounds was not necessarily the same in each phase. Although analyses based on a number of the steps described above were sometimes carried out, alternative methods were also considered.

To try and overcome some of the problems described above, one of the solutions that was considered was to make observations corresponding to individual children the unit of analysis. This allowed some covariates such as age and sex to be fitted directly. The main problem with this approach was that it was then necessary to allow for correlated responses between children consulting with a particular doctor. This was done by modelling variation between doctors as a fixed effect. There were two reasons for this approach. Firstly, at the time that this analysis was undertaken (1989/1990) there were no readily available packages for fitting mixed effects models to binary data (although there were some procedures available for normally distributed data). Secondly, trainers were not randomly allocated to trainer groups. It was felt that the best way to control for any systematic effects caused by this method of allocation was by incorporating variation between doctors as a fixed effect. Arguably, fitting doctor variation as a fixed effect rather than as a random effect reduces the generalisability of these findings to other doctors. But in the case of this study, the problem of generalisability would exist anyway. The doctors who participated in this study were 80% of the general practitioner trainers in the Northern Region who worked in practices with at least one other partner.

It is not clear that results obtained in a study involving perhaps the more motivated and more experienced doctors will be the same as those obtained from a similar study involving "ordinary" general practitioners. As doctors need to demonstrate some merit before being granted the status of trainer it is likely that differences exist between trainers and other general practitioners.

A useful piece of future research might be to review this decision to model variation between doctors as a fixed effect. In a slightly different area of application, that of meta analysis, Aitkin (1997) has argued that there are considerable benefits to modelling differences between studies as a random effect even if the studies have been selected systematically. It would be interesting to compare the parameter estimates for the effects of standard setting obtained from a mixed effects model with those reported here. The multilevel modelling package MLn (Rasbach and Woodhouse, 1995) now has a built in command to help set up a model in which there are cross-classified fixed and random effects. [A *crossed effects model* would be necessary in the case of the North of England study because each doctor (doctors would be treated as a random effect) provided data for each condition (conditions would still be modelled as a fixed effect). The effects are therefore crossed rather than nested. Cross-classified models are discussed in some detail by Goldstein (1995).]

For some variables (such as the number of items of current medical history), in addition to analysing individual observations, alternative analyses based on steps (i) and (ii) in the classical analysis described above were also carried out. Means were calculated for groups of children consulting with a particular doctor for a particular condition and then included as the dependent variable in the generalised linear modelling. Assumptions relating to the distribution of the residual errors were checked carefully at each stage. In these analyses too, for the reasons given above, variation between doctors was included as a fixed effect.

For most of the analyses of the process variables, variation between doctors was highly significant perhaps indicating substantial differences between doctors. There may be a number of causes of this variation. One source of variation may be the demographic characteristics of the population for whom the doctor provides care. Fitting variation between doctors takes into account the lack of randomisation of patients to doctors. Secondly it is likely that there are real differences between doctors that are quite sizeable. For some variables, such as decisions about whether or not to discharge a child, these differences may be due to differences in recording style. For other variables, such as the prescription of antibiotic drugs, there was evidence to suggest that doctor's actual behaviour concurred closely with their recorded behaviour and that differences reflect variation in management of the condition. With an unbalanced design it would clearly not be appropriate to fit a standard setting effect without first allowing for such variation.

For each process variable, the actual analytic strategy that was adopted comprised a number of steps. Frequency distributions and summary statistics were used to inform the decision as to the most appropriate method of analysis. Generalised linear modelling was then carried out using an appropriate error structure and link function. In general, effects relating to the rows and columns of the Latin square design were the first to be incorporated into the model—variation between study conditions and variation between trainer groups or variation between doctors. Differences between data collection phases were included as a fixed effect if significant. Terms representing the effects of standard setting were incorporated in the model with and without various covariates. Due to the confounding of many of the factors several series of nested models were considered before the model (or models) that best represented the data was (or were) selected.

The initial strategy was to look for global changes in the pattern of care provided. For example, did the setting of standards lead to greater recording of family histories? Were

these changes uniform across all five conditions? A number of variables took the form of a binary variable representing the presence or absence of information of a particular type. In such cases it was probably reasonable to expect that setting standards would lead to an increase in the recording of such information for each of the five conditions. But for other variables, for example those relating to the management of the condition, it did not seem reasonable to expect changes to be the same for each condition. Different effects of standard setting were therefore included for these variables by fitting an interaction between the standard setting effect and the effect of study conditions.

It was mentioned above that, for some variables, alternative methods of analysis were possible. For example, for continuous variables such as the mean number of diagnoses, it was possible either to treat observations from each phase as repeated measures or to analyse the difference between corresponding observations taken in phase 1 and phase 2. When alternative analyses were undertaken the results were examined carefully to check for inconsistencies. It was reassuring that very few were found.

Few effects of standard setting on the process of care were detected. The area where standard setting seems to have had the most effect is on the prescription of drugs. There is evidence that data on drug management is more reliable than other data; it was the one area where there was little difference in the amount of information recorded on enhanced records and statutory records. Changes consistent with advice given in the standards were found for the prescription of antibiotics and there appeared to be some changes in the prescription of other therapeutic drugs (particularly for doctors who set standards for bedwetting) although the evidence was not entirely conclusive. Outside of drug management there was some evidence to suggest that setting standards had affected doctors' willingness to discharge patients but again the evidence was not totally conclusive. It is possible that the lack of positive findings in areas of care other than

drug management may be partially attributable to poor reliability of data rather than the ineffectiveness of the intervention but there is no way of testing this hypothesis.

12.4 Outcome: evidence from interviews with parents

12.4.1 Clinical outcome

Data were collected by interview for only three of the study conditions. Clinical outcome was first assessed using a binary measure for each of the conditions. Parents were asked whether the condition was still affecting their child's health at the time of the interview; responses were classed either as successes or failures. There were big differences between conditions. For children with acute cough, the length of time between the consultation for the cough and the interview explained a lot of the variation in outcome. For children with recurrent wheezy chest, there was a substantial difference in outcome between those children who were believed to be asthmatic and those whose chest trouble was believed to be brought about by other causes. Children reported as suffering from asthma tended to have a poorer clinical outcome than other children. Using a binary measure of clinical outcome, there was no evidence of any effects of standard setting.

For the two chronic conditions, a graded measure of clinical outcome was available. For bedwetting this variable took the form of the square root of the number of nights on which the child wet the bed in the month preceding the interview. For recurrent wheeze, a composite index was formed—based on three separate symptoms. The difference in scores between initial and final interviews was calculated for each condition. For each condition there were spikes in the frequency distributions (Figures 8.2 and 8.7) which corresponded to children for whom no symptoms were reported at either interview. There was a concern that if these children were retained in the analysis the assumptions underlying the F-tests for comparing nested models might be violated but, that if they

were omitted the estimates of the effects of standard setting might be biased. The analysis was therefore done twice—initially all children were included but then those children for whom no symptoms were reported in either phase were omitted. Reassuringly, the two analyses gave results that were almost identical (Table 8.13). In contrast with the analysis of the binary outcome variable, there was evidence to suggest that, for children with recurrent wheezy chest, outcome was better for those who consulted with doctors who had set a standard for that condition than for those who consulted other doctors.

12.4.2 Parents' satisfaction with care

Parents' satisfaction was assessed using a 12 item scale developed in the United States. The psychometric properties of the scale were examined. It was found that responses were highly skewed; most patients were very satisfied. It is possible that the individual items lacked a sufficient number of response categories. Fitzpatrick (1992) describes the application of general principles of attitude measurement to surveys of patient satisfaction. He suggests that items typically should be measured using a five point Likert scale. In general the reliability of items increases as the number of response alternatives increases (Nunnally, 1978) although there is evidence (Lissitz and Green, 1975) that there is little to be gained by going beyond five points. In this study the satisfaction items had either two or three response categories only.

Psychometric analysis has shown that summed scales are more reliable than individual items (Oppenheim, 1992, page 165). The next step in the analysis was to investigate which sets of items could be summed to form scales that had both face validity and good internal reliability. Results from a principal components analysis (Table 9.3) suggested that the most suitable grouping of items to measure various aspects of satisfaction was slightly different from that proposed by the original authors. It was decided that the scale relating to satisfaction with the consultation was the one that was most relevant to

there might still be a problem with ceiling effects. Mean satisfaction scores for parents of children with a particular condition registered with a given practice were analysed using a weighted least squares estimation procedure. The scale was able to detect large differences between practices. There was no evidence that standard setting had had any effect on satisfaction with the consultation.

12.5 Outcome: evidence from postal outcome questionnaires

The final data set to be analysed was outcome assessed by means of a self completion postal questionnaire. The administration of the questionnaires had been beset with a number of problems. The net result of these was that the timetable for mailing out questionnaires was not adhered to (Table 11.1). An analysis of clinical outcome, parental anxiety and parents' satisfaction were undertaken making allowance for this. No effects of standard setting were observed but it is probable that the power of the study design to detect such effects was compromised by the problems affecting the administration of the questionnaires. Because of the lack of balance many effects were confounded. When one variable was entered into the model, it explained some of the variation that may have been due to another. It is likely that fitting variation between practices, variation between conditions and making allowance for temperature effects would explain some of the variation due to an effect of standard setting (if it had existed). It would be interesting to do a simulation study based on pattern of questionnaires returned in Table 11.2. Using estimates of the main effects obtained from the final model of, say, clinical outcome an additional effect of standard setting could be simulated. The power of the design to detect an effect of standard setting (allowing the magnitude of the effect to vary) using the method of analysis described could be investigated.

12.6 Measuring health outcomes

This study did not provide a great deal of evidence that outcome for children was influenced by the setting of clinical standards. There may be a number of reasons for that. One possible explanation may be the nature of the outcome measures used. Many of the measures were fairly crude. It is interesting that evidence of an improvement in clinical outcome for children with recurrent wheezy chest came from a graded measure of clinical outcome. Using a binary measure of outcome, there was no evidence of any change. It may be that the binary measure is less sensitive than the graded measure.

Since 1982 when this study was designed, there has been a considerable amount of work done on developing health outcome measures (see for example, Bowling 1991 and 1995; Wilkin, Hallam and Doggett, 1992; and Jenkinson, 1994). Many of the outcome measures developed take the form of scales made up of multiple items. These are summed to provide an index score that measures a particular aspect of health status (McDowell and Newell, 1987). In addition to being more reliable, multi-item scales are likely to be more sensitive in detecting change than single item measures.

A number of authors have suggested that is advantageous to use a combination of generic and condition specific measures in order to assess outcome for patients. Bowling (1995, pages 14-16) discusses the relative merits of the two types of scale. In the context of a Latin square design, generic measures have the advantage that they can be used across all of the conditions in the study; in each cell there is an equivalent measure of outcome. The disadvantage of generic measures is that they may not be able to detect sometimes small but clinically significant changes in health status and levels of disease severity. Bowling cites a number of studies where this was found to be the case (Dhillon et al, 1982; and Morris, 1990) and one example of a study (Kantz, 1992) where this was not the case. There is a problem with the use of generic measures that is specific to the research design employed for this study. It will be addressed in the

discussion of the study design in the next section. The main problem with condition specific measures, used within a Latin square design, is that the outcome variable does not take exactly the same form in each cell, creating problems at the analysis stage.

The condition specific measures in this study related to presence and frequency of symptoms associated with each conditions. The graded measure took the form of the number of days per month on which symptoms occurred. Some transformations of the data were necessary to get the data in a form that could be analysed without violating assumptions relating to the distribution of the residual errors. A square root transformation improved the fit to a Normal distribution. A linear combination of the transformed wheeze symptoms (the first principal component) gave a single index score for that condition. The transformed bedwetting symptom score was then scaled so that it had the same variance as that of the wheeze index thereby ensuring approximate homogeneity of variance across each of the conditions.

12.7 Evaluation of the research design

12.7.1 Choice of the study conditions

One important aspect of the study design was that the conditions were symptomatic; they were not defined by specific medical diagnoses. Children were recruited if they were reported as suffering from particular symptoms. One of the main reasons behind this strategy was to get round the variability in the diagnosis of specific conditions. There was a belief that asthma for example was under-diagnosed. It was felt that standard setting might lead to an improvement in the diagnosis of this condition. If a diagnosis of asthma had been used as the criteria for entry into the study it is possible that the characteristics of the children identified in the two phases might differ due to additional children being identified after standard setting.

Opting for symptomatic conditions however created other problems. It is likely that the resulting groups of children are less homogeneous than if diagnostic criteria had been used to identify them. It is clear from reviewing the clinical standards that doctors who set standards for recurrent wheezy chest were concerned almost exclusively with the condition of asthma. Indeed one group explicitly renamed their standard as a protocol for the treatment of children with asthma. By using symptoms to identify children it is likely that the children recruited included a large group who were not suffering from asthma and for whom the standards were not relevant. The mostly likely effect of including children to whom the standards do not apply is to reduce the power of the study to detect changes that are bought about by standard setting. For this particular condition, an effort was made, during the analysis, to make some allowance for this by fitting a difference between children diagnosed (or were thought to be suffering from) asthma and other children with recurrent wheezy chest.

The symptomatic condition itchy rash encompassed a wide range of clinical conditions - one of the standards identified 16 common conditions and 12 uncommon or rare conditions. The course of action recommended in all the standards for itchy rash depended critically upon whether or not a diagnosis could be made and, if one was made, the clinical condition from which the child suffered. This made it difficult to predict with any certainty the likely effects of standard setting on the group of children (suffering from itchy rash) as a whole.

For all the conditions, for any given child, only a part of the standard was relevant. Perhaps the most critical measure of process would have been, for each child, to determine whether or not the doctor had followed the course of action recommended in the standard given the information available. In practice this proved to be impossible due to a lack of resources and lack of necessary information.

Due to the nature of the intervention, decisions about which data items should be collected had to be made before the clinical standards were set. Consequently a large amount of data was collected in the expectation that it would be relevant to an evaluation of the effects of standard setting. In the event only a small amount of it was actually used for that purpose (although many data items were used for other purposes such as economic assessments of the costs of looking after children with the study conditions). Prior knowledge of the content of the standards would have made it easier to collect data to enable an assessment of the extent to which doctors adhered to standards.

Another reason why not all the data could be used for evaluative purposes within the Latin square design was the diverse nature of the study conditions. In order to take advantage of the Latin square design it was necessary to derive a variable that took approximately the same form for each of the study conditions. The derivation of comparable measures of clinical outcome, parental anxiety and satisfaction with care for each study condition has been described. For a number of other variables that were potentially of interest, it was not possible to derive similar measures for each condition. A considerable amount of data relating to additional financial costs arising from the child's condition was available for bedwetting for example but much less data relating to economic outcome were collected for the other conditions. Similarly it was felt that standard setting might well have an influence on the extent to which parents complied with advice and treatment prescribed by doctors but, in practice, it proved impossible to derive an index of compliance for any condition except recurrent wheezy chest. For such variables it was only possible to carry out a univariate analysis (usually multiple regression) outwith the Latin square.

A logistic regression analysis of compliance with medication reported by children with recurrent wheezy chest produced the only instance where there was any evidence of an

effect of standard setting. The results of this analysis have been published elsewhere (North of England Study, 1992b). There was some evidence of an improvement in compliance during the twelve months immediately after doctors had set standards, but that the improvement was not sustained during the following year.

The five conditions were very diverse in nature. This was a deliberate choice; it was felt that most aspects of the management of care of children in primary care would be involved in one or other of these conditions. Two of the conditions chosen were therefore acute. The problem with these conditions is that they are self limiting—children tend to get better by themselves. This makes the assessment of outcome a little problematic. The assumption was that improved care would lead to these children getting better more quickly. Thus at the time they were surveyed it was expected that more of the children who consulted doctors who set standards would have got completely better than those who consulted with other doctors. It is possible that different estimates of the effects of the intervention would be obtained depending upon the follow up period selected. In this case, because of the way in which the children were recruited, the time between consultation and survey was variable. An attempt was made to take this into account by fitting time between consultation and survey as a covariate.

For these reasons, for acute conditions it is difficult to see how it might be possible to produce a condition specific outcome measure that is valid, reliable and responsive to change. Further there are obvious problems with using a generic measure of outcome for these conditions if many of the patients are no longer suffering from the condition by the time the survey arrives. At the very least, there must implications for the sample size required if you expect to see an effect in only a fraction of the sample.

12.7.2 Choice of study design

There is very little published literature that deals specifically with the issues of study design within health services research. Most of the books that deal with the design, conduct and interpretation of health services research (see for example Wortman, 1981 and Crombie and Davies, 1997) are fairly general in nature. The reason for this may be because the nature of the research questions that arise within this discipline are so enormously diverse. Examples of such questions include: the relative effectiveness of different therapies; the causes and or prevalence of particular conditions; the effects of changes in health care policy; the effectiveness of alternative modes of delivering care; effectiveness of screening programs and consideration of how best to change the behaviour of health care professionals. Usually a decision about how to provide optimum care for patients involves consideration of a complex range of issues. It would now be unusual, for example, for a piece of health services research not to include an economic component.

The nature of the research question will directly affect the research design employed. For comparing alternative therapies a simple two armed randomised controlled trial would normally be the most appropriate design. Such designs are particularly powerful when it is possible to randomise individual subjects to treatment groups and keep that allocation blind to both the subjects themselves and the health care professionals involved. There is a very large literature concerning the design of such trials. Books include Pocock (1983), Friedman et al (1995) and Piantadosi (1997) and there have been a number of special issues of the journal *Statistics in Medicine* devoted to specific aspects of clinical trials (Ashby, 1993; Geller, Freedman, Lee and Der Simonian, 1996; and Souhami and Whitehead, 1994). There have also been a number of papers pointing out the problems that arise when subjects are not properly randomised to treatment groups and when blinding is not undertaken rigorously (Schulz et al 1994, 1995 and 1996).

For research questions relating to the causes of disease an epidemiological study might be appropriate. Again there is an extensive literature dealing with these types of design. Breslow and Day, for example describe the design and analysis of case control studies (1980) and cohort studies (1987). The review *Statistical Methods in Medical Research* has published issues devoted to epidemiological studies (Everitt, Dunn and Holford 1995a) and screening studies (Everitt, Dunn and Holford 1995b). These epidemiological studies (specifically case-control and cohort studies) might be considered as particular examples of a class of studies where it is not possible to randomise experimental subjects to groups. These studies are sometimes referred to as quasi-experimental designs (Cook and Campbell, 1979) or observational studies (Rosenbaum, 1995). Both sets of authors also discuss the use of this type of design to evaluate interventions. The main problem with these studies is that it is extremely difficult to attribute observed outcomes of alternative interventions to the interventions themselves rather than to concomitant variation in any of a wide range of other factors which affect these outcomes. In such studies it is important to try and assess the extent of any bias arising from the lack of randomisation.

The study described in this thesis does not fall neatly into any of these categories. There were opportunities for randomisation (trainer groups were randomised to treatments and conditions) but there was also systematic allocation (doctors to trainer groups, and patients to doctors or practices). Also due to the lack of opportunities for blinding there were problems such as a possible reactive study effect that are common to many observational studies.

There is often a considerable gap between original research which demonstrates the effectiveness of a particular therapy and the implementation of that therapy in practice. Antman and colleagues (1992) demonstrated that 13 years elapsed between the publication of papers that showed the effectiveness of thrombolysis and routine

recommendation of its use in even half of text books or review articles. It is recognised that the value of the original research is limited unless the findings are actually implemented in practice. There is thus considerable interest in methods of changing the behaviour of health care professionals to take account of the new evidence. Studies designed to evaluate these methods are often termed behaviour change studies and are becoming more common (Cochrane Collaboration on Effective Practice, 1995; Oxman et al 1995).

Koescoff et al (1987) have demonstrated that publishing national consensus statements in professional journals is ineffective in changing behaviour. Oxman and colleagues (1995) concluded that dissemination only strategies such as conferences or the mailing of unsolicited materials demonstrated little or no change in health professional behaviour or health outcome when used alone and that more active implementation strategies need to be considered.

The goal of the North of England Study was to improve the quality of care provided in general practice. The intervention, small group work and setting clinical standards, was intended to change the way doctors provided care and might be thought of as an active implementation strategy. Today, the clinical standards developed during the study would be referred to as clinical guidelines. Since the end of the study there has been much interest in the development and implementation of guidelines (Grimshaw and Russell, 1993 & 1994; Grol, 1992 & 1993; NHS Executive, 1994; Mittman et al, 1992). Many of these authors identify the need for rigorous evaluations of the effectiveness of implementation strategies although relatively little has been written about suitable research designs for such evaluations.

Grimshaw and colleagues (1995), in a paper reviewing 91 studies in which such an evaluation has been undertaken, briefly mention a number of designs which they

regarded as suitable for the purpose. Although the aim of these authors was to assess retrospectively the evidence arising from behaviour change studies that had already been undertaken, and not to recommend designs for future studies, their comments relating to study design are still pertinent. They suggest a three tier hierarchy of study design. They argue that randomised controlled trials generally provide the best evidence of the effectiveness of implementation but recognise that in behavioural research such trials may be susceptible to a greater range of bias than in other types of research. They suggest that one of the most reliable trial designs for these types of interventions is one in which each participating doctor simultaneously experiences both guidelines for some conditions and status quo for others in a balanced incomplete block design. The highest tier in their hierarchy comprised: randomised controlled trials in which doctors were randomised either individually or in groups; randomised crossover trials; and trials incorporating the balanced incomplete block design. Evidence coming from such trials was regarded as being the least susceptible to bias and categorised as grade I.

The second tier comprised: before and after studies with non-randomised controls which compare changes in the targeted behaviour with a control group of activities performed by the same doctors but not targeted by the guidelines; and simple randomised controlled trials in which patients are randomised. Evidence coming from such trials was considered to be grade II. The third tier comprised before and after studies controlled by data from other sites where non-randomised controls are selected in the belief that they may experience changes similar to those of the study population provided the baseline characteristics and performance in control and study sites was similar and data collection was contemporaneous in both sets of sites throughout the whole of the study. They felt that simple uncontrolled before and after studies were not suitable for this type of evaluation as it was impossible to attribute any observed changes to the intervention.

The study design used in the North of England Study, a Latin square, is an example of a balanced incomplete block design. It was therefore considered by the above authors to provide grade I evidence of the effectiveness of guideline implementation through the use of small group work. It is interesting to consider the advantages and disadvantages arising from the use of this particular design in this particular study.

12.7.3 Advantages and disadvantages of the Latin square design

In this study very few of the data sets were complete and so the full benefits of the Latin square design were not realised during the analysis, but considerable use was made of many of the features of the design. In addition, the problems that arose during data collection would have had serious implications for the analysis of data arising from any alternative design employed. Some specific issues are considered below.

Reactive study effect

In order to carry out a study it is necessary to make special arrangements that may result in subtle changes in the way care is provided. At the simplest level, just being aware that they are involved in a study concerning a particular condition may cause doctors to change their behaviour. Often there are more obtrusive influences such as the requirement to provide information. (The North of England Study involved doctors in identification of cases and additional data recording; there were visits by fieldworkers to practices; and data were collected from patients.) It is possible that these special arrangements might influence the results of the study. This effect is often referred to as the *Hawthorne effect* following its initial description in the important Hawthorne studies (Roethlisberger and Dickson, 1939).

This effect is a problem if it is different for different experimental groups. This might arise in a simple randomised controlled trial where one of the experimental groups

receives an intervention but the other does not. In the case of the North of England Study if a "control group" of doctors had not taken part in setting clinical standards it is possible that they would have had a very different attitude towards the additional burdens of data recording in comparison with doctors who received the intervention. Clearly this sort of interaction between the study effect and the treatment will lead to biased estimates of the effect of the treatment. This is a particular problem in behaviour change studies when it is almost impossible to keep health care professionals blind to the allocation to experimental groups.

A major advantage of the balanced incomplete block design used is that all doctors received the stimulus of taking part in setting clinical standards. Provided that the reactive effect is equal in each of the experimental groups, the amount of bias in the estimates of any treatment effect should be minimised. This requirement has implications for the choice of clinical conditions in such an evaluation. If doctors perceive one condition to be much more significant clinically than another then the reactive study effect might not be completely equal in each of the experimental groups. Unfortunately it is very difficult to quantify such effects.

Doctors acting as their own controls

The intervention was aimed at groups of doctors. Allocation to groups was not random. In designing the study it was important to recognise that this might lead to systematic differences between groups. To allow for these potential effects in the analysis differences between doctors was modelled as a fixed effect. This was facilitated by doctors acting as their own controls. The estimates of the condition specific effects of standard setting for acute cough for example were based on simultaneous comparisons of the recorded behaviour, in consultations for acute cough in phase 2 of doctors who set clinical standards for that condition with:-

1. recorded behaviour of other doctors in consultations for acute cough in phases 1 and 2;
2. their own recorded behaviour for consultations for acute cough in phase 1;
3. their own recorded behaviour for consultations for other conditions in phases 1 and 2.

In the second and third of these comparisons the doctors are acting as their own controls. There are two features of the design that enables this to happen. The first is the before and after component which permits measurement of change in behaviour over time. Taking observations before and after an intervention is not a feature that is particular to the Latin square design but something that can be undertaken within any research design. The second feature that generates repeated measures is the use of more than one clinical condition. This is a particular feature of the balanced incomplete block design. Had there been no missing data we would have had, corresponding to each individual doctor, five observations in phase 1 (one for each condition) and five observations in phase 2. The estimate of the fixed effect for each doctor would therefore be based on ten observations (assuming that we combine data across different patients to form a single observation). In a simple before and after design with just one clinical condition, if the same analytic approach were adopted, the fixed effect for each doctor would be based on just two observations and would be highly confounded with any treatment effect. Within the Latin square design, because of the level of missing data, there was still confounding between the fixed doctor effect and the treatment effect but the additional number of repeated measures went some way to alleviating this problem.

The analysis of the quantitative measure of outcome reported in Chapter 9 involved only two of the conditions—bedwetting and recurrent wheezy chest. The analysis indicated that setting a standard for wheezy chest improved outcome for children with that condition. Part of the evidence for this effect was that outcome for those children with bedwetting who consulted with doctors who set standards for wheezy chest had not

improved. This suggested that it was unlikely that there was some phenomenon causing a general improvement in outcome among all children consulting with those particular doctors.

Five conditions

There was interest in whether the interventions would be effective in a range of clinical conditions rather than just one. The only way that to address this issue was to include a range of conditions. In practice, on the basis of the results reported in this thesis it is difficult to draw any conclusions about the about the relative merits of setting standards for different types of conditions. In part this is because there were so few positive findings with respect to the effects of the intervention.

There is another issue related to the use of multiple conditions within a study such as this. A proportion of children will suffer from more than one condition. This can cause a problem if a generic measure is used to assess outcome. As an example consider a child with itchy rash who consults with a trainer who set a standard for acute cough. The child is identified as suffering from itchy rash and the satisfaction score in the outcome survey will pertain to that condition. It is then possible that the child may also consult with the same doctor for acute cough. If there is an effect of standard setting on patient satisfaction the satisfaction score in the outcome survey may be influenced by this later consultation. If the consultation occurs outside the six week period of prospective identification for acute conditions, that child will not be identified as also suffering from acute cough. The net effect is that the satisfaction scores of some of the controls may be enhanced by this mechanism. At the very least this must influence the expected effect size that one might observe and thus influence any sample size considerations. The choice of two acute conditions may have exacerbated this problem. (Any child may develop an acute condition but generally either a child suffers from a chronic condition or does not.)

Five treatments

The study was designed to evaluate a number of interventions. It was perhaps this aspect of the design that gave most cause for concern. To analyse a Latin square design which has just one observation per cell it is necessary to make the assumption that there is no interaction between the main effect of treatments and the main effects represented by the rows and columns of the design. When the study was originally planned the intention was to undertake the usual Latin square analysis in which there was a single treatment effect with five degrees of freedom that was orthogonal to the main effects of rows and columns. Such a model is only applicable if each intervention has an effect of a particular magnitude that is uniform across all conditions. Subsequent investigation has shown that it was very unwise to make this assumption. Each condition has a very different impact on the health of a child; it is very questionable that any intervention could produce the same effect on each condition.

By taking the consultation (process data) or the child (outcome data) as the unit of analysis we had multiple observations per cell. It was therefore possible to investigate interactions between treatment and study effects. In particular condition specific effects of standard setting have been considered in a number of chapters. The main problem is that the study had much less power to detect these condition specific effects than it had to detect the main effects. The confidence intervals for condition specific effects of standard setting given in the results chapters tend to be much larger than those give for a single uniform effect (for example, compare the confidence intervals in Table 5.5 with those in Table 5.6).

At the design stage each intervention was given equal weight. In practice the vast majority of available resources were dedicated to just one of the interventions—setting standards. Comparatively few resources were spent on the other interventions—developing effective forms of feedback. Clearly these other interventions were

considered to be of lesser importance by those responsible for implementing the interventions. When modifications to process data collection were necessary, it was decided to maintain the level of data abstraction for the condition for which the doctor set a standard but to reduce abstraction targets for the other four conditions (Chapter 4). That is, priority was given to detecting a significant effect of standard setting in preference to detecting effects of the other interventions.

In retrospect it would probably have been better to make this decision at the design stage. It would have been sensible to either drop interventions b, c and d described in Figure 1.2 (so that for conditions for which doctors did not set standards there was no intervention at all) or replace interventions b, c, d and e with a single intervention (for example the doctors could have been sent a clinical standard for the four conditions for which they did not set a standard themselves—intervention b). The argument for the latter approach is that giving the doctors an existing standard is very much cheaper than developing one for themselves. Thus if standard setting is to be adopted it should be demonstrably better than the cheaper alternative.

Such a decision would then have had implications for the sample of children selected for study. The estimates of condition specific effects of standard setting reported in the results chapters are typically based on a very small number of cases (children with the condition for which the trainer set a standard) and a much larger number of 'controls' (children with one of the other four conditions). Usually, in this type of study power is maximised when the numbers of cases and controls is approximately the same (see for example Kraemer and Thiemann, 1987, p.42). Thus it would have been desirable to sample approximately four times as many children with the condition for which the doctor set a standard than children with each of the conditions. For one of the conditions, bedwetting, such a sampling scheme might not have been possible because of the low prevalence of the disease. To maintain the desired power, it would then have

been necessary to either increase the number of controls or select another, more prevalent, condition.

Choice of subjects

The intervention was undertaken as part of the in-service training of general practitioner trainers who were not single handed and whose practices were in the old Northern Region. It has already been mentioned that general practitioners need to demonstrate special qualities before they are awarded training status and thus the study doctors could definitely not be considered as a sample that was representative of the UK population of primary care doctors. Besides lack of generalisability there is another issue that must be considered as a consequence of restricting the intervention to just this group of doctors. It is now widely accepted that care for most of the UK population is provided by a primary health care team rather than by individual doctors. One of the practicalities of this is that patients do not always see the same doctor when they visit the surgery. This may be particularly true for acute episodes of care (a patient may need to visit the practice fairly urgently and find that the doctor with which they are registered is not available on that occasion). Thus it is probably not sensible to try to associate uniquely a patient with a given doctor.

In addition there is a range of health care professionals involved in the care of a child. A child with asthma for example may be seen by a practice nurse in an asthma clinic. The actions of this nurse may have as much impact on the health of that child as those taken by the doctor. It would seem sensible to target the behaviour change at the whole primary health care team rather than any one individual within it.

12.8 Recommendations

Based on the previous discussions the following recommendations are made to assist future researches who wish to evaluate strategies for changing the behaviour of health care professionals:

1. Keep the number of alternative strategies to a manageable size. A decision about which behaviour change strategy to adopt in practice is best informed by a pragmatic trial. Before undertaking such a trial there should be either good theoretical evidence (see Lomas, 1994, for a review of some of the theoretical models influencing practical strategies) or good empirical evidence for the effectiveness of the various strategies. This evidence should inform the sample size calculations which should be based on the smallest difference that it is desired to detect between alternative strategies. In the North of England Study consideration was given to the power of the study to detect a difference between standard setting and the control patients but not to its power to detect potentially much smaller differences between the different levels of feedback.

2. If interactions between the alternative strategies are of interest, a factorial design might be considered. It was not the aim of the North of England Study to look at the question of whether one type of audit was more effective if done in conjunction with another and hence a factorial design was not adopted. However, there may be behaviour change studies in future in which interaction effects are of interest.

3. The issue of a study (or Hawthorne) effect should be addressed at the design stage.

    - For the purpose of evaluating a single strategy a randomised controlled trial might be considered. The main problem with this design is that it is almost always impossible to keep health care professionals blind to the allocation to treatment groups. It is likely that the study effect will be greater in one arm

of the trial than in the other. In particular those professionals who are randomised to receiving no intervention (the control group) may be less enthusiastic about the study than other doctors. It may be possible to ameliorate this problem to some extent by allocating all health care professionals to the treatment group but delaying the implementation strategy for one half of them. The main disadvantage of this is that the duration of the study is longer (assuming further data collection after the delayed intervention). It is also possible that the study effect may not be fully equalised in the two groups.

- The balanced incomplete block design may also be considered at this stage. It may be regarded as suitable if there are two or more interventions which are more or less independent of each other. In the context of implementing clinical guidelines the most likely scenario would be the evaluation of evidence based guidelines for two separate, unrelated clinical conditions. One arm of the trial would comprise the experimental group for the first condition and the control group for the second. The other arm would be the control group for the first condition and the intervention group for the second. (It would be possible to consider evaluating simultaneously more than two guidelines but, as the number of guidelines increases, so will the problems associated with patients consulting with more than one of the specified conditions.) In order for the study effect to be equalised in the two groups the conditions should be perceived to be of equal importance.

4. If a balanced incomplete block design is chosen:

    - The conditions should be clinically independent of each other. Guidelines for the management of one condition should have no effect on the management of the second. In the North of England Study there was a high degree of

overlap between the two respiratory conditions—acute cough and recurrent wheezy chest.

- Patients with both conditions should be excluded from the study.
- Sample size calculations should be based on condition specific effects of the guidelines

5. Clinical conditions should not be self limiting. In the North of England Study children with acute cough and acute vomiting tended to get better anyway. It may not be realistic to expect an intervention to have a marked effect on clinical outcome for these children.

6. Both generic and condition specific measures of outcome should be used. This study did not actually provide very much evidence about the relative effectiveness of generic and condition specific outcome measures. Within the balanced incomplete block design it is easier to analyse generic measures of outcome. They take exactly the same form for each condition; it is easier to analyse both conditions simultaneously. On the other hand condition specific measures may be more sensitive.

7. Where possible, existing measures of outcome that have been shown to be valid, reliable and responsive to change should be used. This makes it easier to compare results across trials. In addition, over a period of time, it is likely that there will be an increased understanding of the level of clinical significance associated with different effect sizes for standard outcome measures.

8. If new measures of outcome are to be developed, they must first be validated and be shown to reliable before they are used in an evaluation.

9. The additional burden of work placed on health care professionals should be minimised. The evidence from trainer interviews indicated that the burden of enhancing medical records in this study caused negative feelings about the study as a whole. It would be desirable that routine medical records be used. Greater computerisation of medical records may facilitate data collection.

10. The reliability of any process measures collected should be investigated. Evidence from this study suggested that routine medical records can yield a reliable account of drug management but that some other aspects of disease management were inadequately documented.

11. If it is desired to modify behaviour to better comply with national evidence based guidelines, the guidelines themselves should be used to inform the choice of process and outcome measures. In the North of England study the measures of process and outcome were determined before the clinical standards had been drawn up. It is possible that some of these measures related to aspects of care not included in the standards.

12. A design incorporating repeated measures is to be preferred. The evidence from this study suggests that there tends to be large differences between health care professionals but that individuals tend to be reasonably consistent in the way that they provide care. When the correlation between repeated measures is high (greater than 0·5) Kraemer and Thiemann (1987, p.p. 46-49) suggest that the repeated measures design is more efficient than the single endpoint design.

13. All those involved with the study should be kept blinded as much as possible. This recommendation does not follow directly from the analysis reported in this thesis but

the problems associated with lack of blinding are well documented elsewhere (e.g. Pocock, 1983).

14. In behaviour change studies, the unit of randomisation should be made as small as possible within the constraint that the allocation should not lead to contamination across experimental groups. In primary care the most suitable choice for the unit of randomisation will probably be the practice.

15. If the study is in the primary care setting with the practice as the unit of randomisation, the intervention should be aimed at the whole primary care team. This will reduce the chance of collecting outcome data from patients who were not seen by a specific member of that team who was subject to the intervention.

16. The analysis should take into account any clustering within the design. In the North of England Study it was possible to do this using fixed effects models but in future studies which involve a higher degree of randomisation the use of mixed effects models should be investigated.

17. Behaviour change studies are susceptible to a range of different biases. Careful consideration must be given to potential sources of bias at each stage of the research process. Careful choice of a research design can help to minimise the threat that is posed to the validity of the study. During the analysis sensitivity analysis to assess the extent of any bias should be undertaken wherever possible.

12.9 Effects of standard setting on the process and outcome of care

12.9.1 Estimation of the effects

When the study was originally designed, it was anticipated that one of the advantages of using a Latin square design was that standard methods of analysis would be available. To some extent this was true—the use of these methods to analyse the prevalence survey is reported in Chapter 3. But even here the analysis was not completely straightforward. As pointed out earlier in this chapter (Section 12.2), tests of some of the main hypotheses of interest did not coincide exactly with specific main effects or specific interaction terms in the usual statistical model and perhaps the composite terms (developed to overcome this problem) in the analysis of the other data-sets should have been used here too.

For the remaining data sets there were a number of additional features that made application of standard procedures even more problematic. During their analysis, it was necessary to allow for the confounding of effects arising because of the loss of othogonality which arose either by design (interviews for only three conditions; postal questionnaires for only four conditions) or through other causes (diminishing enthusiasm of doctors to enhance medical records; failure to administer postal questionnaires). Essentially this necessitated the consideration of a number of alternative models in which terms were fitted in different sequences.

The strategy for testing the hypotheses of interest was set out in Chapter 2. This included consideration of the order of fitting terms and a rationale for how each of the effects of standard setting were to be tested. In general, the strategy was to develop composite terms that matched exactly the effects of standard setting of particular interest. One of the areas over which there was some debate was how a condition specific effect of standard setting should be tested in the case where no general effect (across all conditions) was found. The choice was between retaining the non-significant

main, general effect (denoted by PSTD) and adding the condition specific effect (CPST) in which case the test was based on four degrees of freedom or first removing the general effect and then fitting the condition specific effect in which case the test was based on five degrees of freedom. After some discussion, it was the opinion of the study statistics team that the latter approach was most appropriate. It was felt that if the hypothesis was that standard setting produced a different effect for each condition we would be estimating six parameters - one corresponding to each of the five conditions after the intervention and a sixth corresponding to all other observations. It was felt that the test should therefore be based on five degrees of freedom and that there was no justification for retaining the non-significant general effect in the model before fitting the condition specific effect. This decision is perhaps open to review. In practical terms, going back through the analyses there were no occasions where this choice critically affected the interpretation of the results.

Overall, the modelling strategy set out in Chapter 2 seems to have been appropriate.

12.9.2 Magnitude of the effects

In this study comparatively few effects of standard setting were found. There was some evidence that standard setting had had an effect on drug management—particularly on the prescribing of antibiotics for all children and on the prescribing of other therapeutic drugs for children with bedwetting. There was also evidence to suggest that, for children with recurrent wheezy chest, standard setting led to an improvement in clinical outcome.

That there were so few observed effects may be because the intervention was not particularly effective. However there may be other reasons. Perhaps the measures of process and outcome used in the study had insufficiently reliability, validity or sensitivity. But perhaps of more serious concern was whether the study had sufficient power to

detect such changes given the problems that arose during the data collection phases. In each of the results chapters confidence intervals for the effects of standard setting have been given. In general these tend to be fairly wide. It is likely that a number of these interval estimates include effect sizes that might be regarded as clinically significant. This is essentially a matter of clinical judgement.

Although, for most of the data-sets, a standard Latin square analysis was not possible a number of features of the design proved beneficial (as discussed in section 2.7). The analyses reported in this thesis have demonstrated that it is possible to use modern statistical techniques to analyse data from this type of evaluation. The extent of bias caused by deviations from the study protocol can be assessed and a valid analysis carried out. These methods, however, cannot compensate for any loss of power that deviations from the research protocol might cause.

Bibliography

Aitkin, M. (1997) Meta-analysis by random-effect modelling. *Burning issues in medical statistics abstracts* p.35. De Montfort University, Leicester.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical modelling in GLIM*. Oxford: Oxford Science Publications.

Anscombe, F.J. (1961) Examination of residuals. *Proceedings of the Fourth Berkeley Symposium* **1** 1-36.

Antman, E.M., Lau J., Kupelnick, B., Mosteller, F. and Chalmers, T.C. (1992) A comparison of results of meta-analyses of randomized controlled trials and recommendations of clinical experts. *Journal of the American Medical Association* **268** (2) 240-248.

Armitage, P. and Berry, G. (1987) *Statistical methods in medical research.* 2nd edn. Oxford: Blackwell.

Ashby, D. (1993) Papers from the conference on methodological and ethical issues in clinical trials [Special issue]. *Statistics in Medicine* **12** (15&16).

Baker, R.J. and Nelder, J.A. (1978) *The GLIM system release 3*. Oxford: Numerical Algorithms Group.

Bentler, P.M. (1989) *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

Black, D. (1984) *Investigation of the possible increase of cancer in West Cumbria.* London: Her Majesty's Stationary Office.

Bland, J.M. (1995) *An introduction to medical statistics*. 2nd edn. Oxford: Oxford University Press.

Bowling, A. (1991) *Measuring health: a review of quality of life measurement scales*. Milton Keynes: Open University Press.

Bowling, A. (1995) *Measuring disease: a review of disease-specific quality of life measurement scales*. Buckingham: Open University Press.

Chatfield, C. and Collins, A.J. (1980) *Introduction to multivariate analysis*. London: Chapman and Hall.

Cochran, W.G. and Cox, G.M. (1992) *Experimental designs*. 2nd edn. Chichester: Wiley.

Cochrane Collaboration on Effective Practice. (1995) Implementing findings of medical research: the Cochrane Collaboration on Effective Professional Practice. *Quality in Health Care* **4** (1) 45-47.

Cook, R.D. and Prescott, P. (1981) On the accuracy of Bonferroni significance levels for detecting outliers in linear models. *Technometrics* **23** 59-63.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-experimentation : design & analysis issues for field settings*. Chicago: Rand McNally College Pub.

Cox, D.R. and Snell, E. (1989) *Analysis of binary data*. 2nd edn, London: Chapman and Hall.

Crombie, I.K. and Davies, H.T.O. (1997) *Research in health care: design conduct and interpretation of health services research*. Chichester: Wiley.

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* **16** 297-334.

DeLury, D.B. (1946) The analysis of latin squares when some observations are missing. *Journal of the American Statistical Association* **41** 370-389.

Dhillon, R., Palmer, B., Pittam, B. and Shaw, H. (1982) Rehabilitation after major head and neck sugary: the patient's view. *Clin Otolaryngol* **7** 319-324.

Dobson, A.J. (1983) *An introduction to statistical modelling*. London: Chapman and Hall.

Everitt, B.S., Dunn, G. and Holford, T.R. (1995a) Design issues in epidemiological studies. *Statistical Methods in Medical Research* **4** (4).

Everitt, B.S., Dunn, G. and Holford, T.R. (1995b) Screening. *Statistical Methods in Medical Research* **4** (1).

Fitzpatrick, R. (1992) Surveys of patient satisfaction: II-designing a questionnaire and conducting a survey. In: Smith, R. (Ed.) *Audit in action*. London: British Medical Journal.

Friedman, L.M., Furburg, C.D. and DeMets, D.L. (1995) *Fundamentals of clinical trials*. 3rd edition. St Louis: Mosby-Year Book.

Geller, N., Freedman, L., Lee, Y.J. and DerSimonian, R. (Eds.). (1996) Conference on meta analysis in the design and monitoring of clinical trials [Special issue] *Statistics in Medicine,* **15** (12).

GLIM Working Party (1987) *The GLIM system release 3.77: manual.* 2nd edn. Oxford: Numerical Algorithms Group.

Goldstein, H. (1995) *Multilevel statistical models.* 2nd edn. London: Arnold.

Grimshaw, J. and Russell, I. (1993) Achieving health gain through clinical guidelines. II: ensuring guidelines change medical practice. *Quality in Health Care* **3** 45-52.

Grimshaw, J. and Russell, I.T. (1994) Achieving health gain through clinical guidelines. I: developing scientifically valid guidelines. *Quality in Health Care* **2** 243-248.

Grimshaw, J., Freemantle N., Wallace, S. Russell, I., Hurwitz, B., Watt I., Long A. and Sheldon, T. (1995) Developing and implementing clinical practice guidelines. *Quality in Health Care* **4** (1) 55-64.

Grol, R. (1992) Implementing guidelines in general practice care. *Quality in health care* **1** 184-191.

Grol, R. (1993) Development of guidelines for general practice care. *British Journal of General Practice* **43** 146-151.

Ho, M., Foy C., Avery P., Russell I. and Steen N (1990) Does quality assurance affect consultation rates in British Family Medicine? *Proceedings of the Social Statistics Section of the American Statistical Association* 270-275.

Irvine, D.H., Russell, I.T., Hutchinson, A., Foy, C.J.W., Addington-Hall, J.M., Barton, A.G., Donaldson, C., Haimes, E.V., Humphrey, R.D., Philips, P.R., Parkin, J.M. and Hewison, J. (1986a) *Northern regional study of standards and performance in general practice: preliminary report on phase 1.* Newcastle Upon Tyne: Health Care Research Unit (Report 28).

Irvine, D.H., Russell, I.T., Hutchinson, A., Foy, C.J.W., Addington-Hall, J.M., Barton, A.G., Donaldson, C., Haimes, E.V., Humphrey, R.D., Philips, P.R., Parkin, J.M. and Hewison, J. (1986b) Performance review in general practice: educational development and evaluative research in the Northern Region. In: Pendleton, D.A., Schofield, T.P.C. and Marinker, M.L. (Eds.) *In pursuit of quality.* London: Royal College of General Practitioners.

Jenkinson, C. (1994) *Measuring health and medical outcomes.* London: UCL.

Kantz, M.E., Harris, W.J., Levitsky, K., Ware, J.E., Jr. and Davies, A.R. (1992) Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. *Medical Care* **30** MS240-MS252.

Kim, J.O. and Mueller, C.W. (1994) Factor analysis: statistical methods and practical issues. In: Lewis-Beck, M.S. (Ed.) *Factor analysis and related techniques.* pp. 75-155. London: SAGE Publications Toppan Publishing.

Kosecoff, J., Kanouse, D.E., Rogers, W.H., McCloskey, L., Winslow, C.M. and Brook, R.H. (1987) Effects of the national institutes of health consensus development program on physician practice. *Journal of the American Medical Association* **258** (19) 2708-2713.

Kraemer, H,C. and Thiemann, S. (1987) *How many subjects? Statistical power analysis in research.* London: SAGE Publications.

Lindman, H.R. (1992) *Analysis of variance in experimental design.* New York: Springer-Verlag.

Lissitz, R.W. and Green, S.B. (1975) Effect of the number of scale points on reliability: a Monte Carlo approach. *Journal of Applied Psychology* **60** 10-13.

Lomas, J. (1994) Teaching old (and not so old) docs new tricks: effective ways to implement research findings. In: Dunn, E.V., Norton, P.G., Stewart, M., Tudiver, F.

and Bass, M.J. (Eds.) *Disseminating research/changing practice. Research methods for Primary Care Volume 6.* Thousand Oaks: Sage Publications.

McCullagh, P. and Nelder, J.A. (1989) *Generalized linear models.* 2nd edn, London: Chapman and Hall.

McDowell, I. and Newell, C. (1987) *Measuring health: a guide to rating scales and questionnaires.* Oxford: Oxford University Press.

Mendenhall, W., Wackerly, D.D. and Scheaffer, R.L. (1990) *Mathematical statistics with applications.* 4th edn. Boston: PWS-Kent.

Mittman, B.S., Tonesk, X. and Jacobson, P.D. (1992) Implementing clinical guidelines: social influence strategies and practitioner behaviour change. *Quality Review Bulletin* **18** 413-422.

Morris, J.N. (1990) *The quality of life of head and neck cancer patients: a review of the literature.* Discussion paper 72. York: Centre for Health Economics, Health Economics Consortium.

North of England Study of Standards and Performance in General Practice (1990a) *Final report: IA setting clinical standards within small groups - appendices.* Newcastle Upon Tyne: Health Care Research Unit (Report 40).

North of England Study of Standards and Performance in General Practice (1990b) *Final report: II methods for evaluating the setting and implementation of clinical standards.* Newcastle Upon Tyne: Health Care Research Unit (Report 41).

North of England Study of Standards and Performance in General Practice (1990c) *Final Report: III the effects of setting and implementing clinical standards.* Newcastle Upon Tyne: Health Care Research Unit (Report 42).

North of England Study of Standards and Performance in General Practice (1990d) *Final report: I setting clinical standards within small groups.* Newcastle Upon Tyne: Health Care Research Unit (Report 40).

North of England Study of Standards and Performance in General Practice (1991) *North of England study of standards and performance in general practice: an overview of the study.* Newcastle Upon Tyne: Centre for Health Services Research (Report No 50).

North of England Study of Standards and Performance in General Practice (1992a) Medical audit in general practice. I: effects on doctors' clinical behaviour for common childhood conditions. *British Medical Journal* **304** 1480-1484.

North of England Study of Standards and Performance in General Practice (1992b) Medical audit in general practice II: effects on health of patients with common childhood conditions. *British Medical Journal* **304** 1484-1488.

NHS Executive (1994) *Improving the effectiveness of the NHS.* Leeds: Department of Health.

Nunnally, J. (1978) *Psychometric theory.* 2nd edn. New York: McGraw Hill.

Oppenheim, A.N. (1992) *Questionnaire design, interviewing and attitude measurement.* New edn. London: Printer Publishers.

Oxman, A.D., Thomson, M.A., Davis, D.A. and Haynes, B. (1995) No magic bullets: a systematic review of 102 trials to improve professional practice. *Canadian Medical Association Journal* **153** (10) 1423-1431.

Piantadosi, S. (1997) *Clinical trials: a methodological perspective.* Chichester: Wiley.

Pocock, S.J. (1983) *Clinical trials: a practical approach.* Chichester: Wiley.

Rasbash, J. and Woodhouse, G. (1995) *MLn command reference.* Institute of Education, University of London: Multilevel Models Project`.

Roethlisberger, F.J. and Dickson, W.J. (1939) *Management and the worker: an account of a research program conducted by the Western Electric Company, Hawthrone Works, Chicago.* Cambridge, Massachusetts: Harvard University Press.

Roghmann, K.J., Hengst, A. and Zastowny, T.R. (1979) Satisfaction with medical care: its measurement and relation to utilization. *Medical Care* **17** 461-479.

Rosenbaum, P.R. (1995) *Observational studies.* New York: Springer-Verlag.

Russell, I.T. (1983) The evaluation of computerised tomography: a review of research methods. In: Culyer, A.J. and Horisberger, B. (Eds.) *Economic and medical evaluation of health care technologies.* Berlin: Springer-Verlag.

Russell, I.T., Foy, C.J.W., Garrett, A., Smyth, J.E., Parker, L., Addington-Hall, J.M., Barton, A.G., Haimes, E.V., Hewison, J., Humphrey, R.D., Hutchinson, A. and Philips,

P.R. (1986) *Northern regional study of standards and performance in general practice: report on pilot study in South Cumbria and North Lancashire.* Report 31, Newcastle Upon Tyne: Health Care Research Unit.

Schulz, K.F., Chalmers I, Grimes D.A. and Altman D.G. (1994) Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynaecology journals. *Journal of the American Medical Association* **272** (2) 125-128.

Schulz, K.F., Chalmers I, Hayes R.J. and Altman D.G. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* **273** (5) 408-412.

Schulz, K.F., Grimes D.A., Altman D.G. and Hayes, R.J. (1996) Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *British Medical Journal* **312** (7033) 742-744.

Snedecor, G.W. and Cochran, W.G. (1967) *Statistical methods.* 6th edn. Ames: Iowa State University Press.

Souhami, R.L. and Whitehead, J. (1994) Workshop on early stopping rules in cancer clinical trials, Robinson College, Cambridge, U.K. [Special issue] *Statistics in Medicine* **13** (13&14).

SPSS (1990) *SPSS reference guide.* Chicago, Illinois: SPSS Inc.

Stigler, S.M. (1986) *The History of Statistics.* Cambridge, Massachusetts: Belknap Press.

Streiner, D.L. and Norman, G.R. (1989) *Health measurement scales: a practical guide to their measurement and use.* Oxford: Oxford University Press.

Taylor, J. (1948) Errors of treatment comparisons when observations are missing. *Nature* **162** 262-263.

Wilkin, D., Hallam, L. and Doggett, M. (1992) *Measures of need and outcome for primary health care.* Oxford: Oxford University Press.

Wortman, P.M. (1981) *Methods for evaluating health services.* London: Sage Publications.

Yates, F. (1936) Incomplete latin squares. *Journal of Agricultural Science* **26** 301-315.

Yates, F. and Hale, R.W. (1939) The analysis of latin squares when two or more rows, columns or treatments are missing. *Journal of the Royal Statistical Society Supplement* **6** 67-79.

Zastowny, T.R., Roghmann, K.J. and Hengst, A. (1983) Satisfaction with medical care: replications and theoretic reevaluation. *Medical Care* **21** (3):294-322.