STATISTICS IN MEDICINE
Statist. Med. **17**, 909–944 (1998)

TUTORIAL IN BIOSTATISTICS

DEVELOPMENT OF A CLINICAL PREDICTION MODEL FOR AN ORDINAL OUTCOME:
The World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants

FRANK E. HARRELL, Jr.[1]*, PETER A. MARGOLIS[2], SANDY GOVE[3], KAREN E. MASON[3], E. KIM MULHOLLAND[4], DEBORAH LEHMANN[5], LULU MUHE[6], SALVACION GATCHALIAN[7] AND HEINZ F. EICHENWALD[8]
and the
WHO/ARI YOUNG INFANT MULTICENTRE STUDY GROUP

[1] *Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, University of Virginia, Charlottesville, U.S.A.*
[2] *The Division of Community Pediatrics, University of North Carolina, Chapel Hill, U.S.A.*
[3] *Programme for the Control of Acute Respiratory Infection (ARI) of the World Health Organization, Geneva, Switzerland*
[4] *MRC, The Gambia*
[5] *Papua New Guinea Institute of Medical Research, Goroka*
[6] *Department of Paediatrics and Child Health, Addis Ababa University, Ethiopia*
[7] *Research Institute for Tropical Medicine, Alabang, Philippines*
[8] *Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, U.S.A.*

SUMMARY

This paper describes the methodologies used to develop a prediction model to assist health workers in developing countries in facing one of the most difficult health problems in all parts of the world: the presentation of an acutely ill young infant. Statistical approaches for developing the clinical prediction model faced at least two major difficulties. First, the number of predictor variables, especially clinical signs and symptoms, is very large, necessitating the use of data reduction techniques that are blinded to the outcome. Second, there is no uniquely accepted continuous outcome measure or final binary diagnostic criterion. For example, the diagnosis of neonatal sepsis is ill-defined. Clinical decision makers must identify infants likely to have positive cultures as well as to grade the severity of illness. In the WHO/ARI Young Infant Multicentre Study we have found an ordinal outcome scale made up of a mixture of laboratory and diagnostic markers to have several clinical advantages as well as to increase the power of tests for risk factors. Such a mixed ordinal scale does present statistical challenges because it may violate constant slope

* Correspondence to: Frank E. Harrell Jr, PhD, Box 600, Health Sciences Center, University of Virginia, Charlottesville VA 22908, U.S.A. E-mail: fharrell@virginia.edu

Contract grant sponsor: Agency for Health Care Policy and Research; contract grant number: HS-06830, HS-07137

CCC 0277–6715/98/080909–36$17.50

assumptions of ordinal regression models. In this paper we develop and validate an ordinal predictive model after choosing a data reduction technique. We show how ordinality of the outcome is checked against each predictor. We describe new but simple techniques for graphically examining residuals from ordinal logistic models to detect problems with variable transformations as well as to detect non-proportional odds and other lack of fit. We examine an alternative type of ordinal logistic model, the continuation ratio model, to determine if it provides a better fit. We find that it does not but that this model is easily modified to allow the regression coefficients to vary with cut-offs of the response variable. Complex terms in this extended model are penalized to allow only as much complexity as the data will support. We approximate the extended continuation ratio model with a model with fewer terms to allow us to draw a nomogram for obtaining various predictions. The model is validated for calibration and discrimination using the bootstrap. We apply much of the modelling strategy described in Harrell, Lee and Mark (*Statist. Med.* **15**, 361–387 (1998)) for survival analysis, adapting it to ordinal logistic regression and further emphasizing penalized maximum likelihood estimation and data reduction.

1. INTRODUCTION

The presentation of an acutely ill young infant presents health workers in all parts of the world with one of their most difficult problems. Serious infections are the main cause of morbidity and mortality in infants under 3 months of age in developing countries. Diagnosis is difficult – meningitis or pneumonia might appear as clear clinical syndromes, but more often the picture is mixed and the infant is labelled as 'sepsis'. Even in industrialized countries, treatment is usually based on clinical impressions supported by laboratory data, which by itself is often inconclusive. In developing countries, clinical signs are the only tools available in most places. The ability to detect serious bacterial infection early in young infants is important in defining appropriate prevention and treatment strategies. A better clinical prediction rule to be used by peripheral health workers might result in more appropriate referral to hospital as well as less antibiotic use in very low-risk infants.

We set out to determine which combination of clinical signs most accurately predict the group of infants who have meningitis, sepsis or pneumonia. There is no single 'gold standard' against which to correlate these signs. It is tempting to use as an endpoint the physician's expert clinical diagnosis. This would induce a circularity which would inflate the predictive discrimination of the prediction model, because clinical signs are major determinants of the overall clinical impression. Death is the only endpoint that can be ascertained for every infant, but truly ill infants who were successfully treated early with antibiotics may not die.

Even without a gold standard, however, there are a number of generally agreed laboratory tests which could be used to construct a reasonable outcome scale, including cerebro-spinal fluid (CSF) culture from a lumbar puncture (LP), blood culture (BC), arterial oxygen saturation (SaO_2, a measure of lung function), and chest X-ray (CXR). In this study, 249 infants died, and death could be placed at the top of an ordinal outcome scale. Many of the deaths were related to starting treatment too late so they were not preventable with antibiotics. For this paper, we choose to ignore death (but not to exclude patients who died) when constructing the scale as the main goal was to predict treatable disease. Ignoring death resulted in some of the clinical signs being stronger predictors. Two-thirds of the deaths were redistributed to other positive outcome categories.

We model an ordinal outcome scale using the proportional odds (PO) form of an ordinal logistic model[1] and the forward continuation ratio (CR) ordinal logistic model.[2] (see references 3–18 for some excellent background references, applications, and extensions to the ordinal

models.) We predicted this ordinal outcome using clinical symptoms, signs, and basic variables such as age, weight, temperature and respiratory rate. Nine major statistical problems had to be addressed in these analyses:

1. How does one avoid estimating a separate coefficient for the large number of clinical signs? (Overfitting and poor model validation would result if all signs were treated as separate candidate variables in the model.)[19]
2. Can expert clinicians assign weights for signs *a priori* that adequately predict the outcomes?
3. Given that the clinical signs can be combined in meaningful ways, how should a cluster of such variables be quantified? Should weights for multiple signs be summed or should the cluster be scored using the weight associated with the most severe sign present? Is the union of all signs (that is, presence of any sign) within a cluster an adequate summary of that cluster?
4. How can continuous predictors such as respiratory rate or temperature be modelled flexibly without assuming linearity?
5. Since the response variable is a hierarchical assignment made up of disparate measurements is the proportional odds assumption likely to be violated?
6. Will another type of ordinal logistic model provide a better fit?
7. How can the constant slopes assumption be relaxed without causing overfitting?
8. How does one diagram the final ordinal model so that field health workers can quickly obtain predicted risks of various severities of outcome?
9. How does one validate an ordinal regression model without sacrificing sample size?

Section 2 gives a brief overview of the World Health Organization/Acute Respiratory Infection (WHO/ARI) Multicentre Study design. Section 3 provides the definition of the ordinal outcome scale. In Section 4 we discuss how clinical signs were scored (individually) and clustered. Section 5 tests the adequacy of weights specified by subject-matter specialists and depicts the utility of various scoring schemes using a tentative ordinal logistic model. Section 6 depicts a simple way to assess the assumption of ordinality of the response with respect to each predictor, and to examine the PO and CR assumptions separately for each predictor. In section 7 we derive a tentative proportional odds model using cluster scores and using regression splines to allow other predictors to be flexibly related to the log odds of an outcome. Section 8 shows how residuals from binary logistic models can be adapted to the ordinal case, and uses smoothed residual plots to assess the proportional odds assumption with respect to each predictor. Section 9 examines the fit of a continuation ratio model. Section 10 shows how the CR model can easily be extended (and fitted using standard software) to allow some or all of the regression coefficients to vary with cut-offs of the response level as well as to provide formal tests of constant slopes. Section 11 shows how penalized maximum likelihood estimation is used to improve future predictive accuracy. In Section 12 the full model is approximated by a sub-model, and a nomogram is constructed for the approximate model. Section 13 demonstrates how the ordinal model is validated using the bootstrap. Many of the methods discussed here were discussed in Harrell *et al.*[20] where the focus was on survival analysis. This modelling strategy used here generally follows that paper, with additional stress on penalized estimation.

Cole *et al.*[17] also presented a case study in developing a PO ordinal logistic model for diagnosing illness in infants under 6 months of age. In their study, which was based on patients who were less severely ill than those in the present study, the ordinal outcome was physicians' subjective assessment of the severity of illness. The analysis was based on stepwise variable selection of individual clinical signs which would be expected to prevent the model calibration to

be accurate for very low and very high risk infants. That paper included a nice example of optimally rounding regression coefficients so that a simple severity score could be derived.

For almost all steps of the analysis, computer code is shown, both to make the steps more concrete as well as to show their feasibility. All analyses were done using S-plus version 3.2[21] on UNIX using Sun Sparcstation 2 and 10 computers in conjunction with the `Design` library of UNIX and Microsoft Windows S-plus functions.[22] For binary and PO logistic models `Design` has a general penalized maximum likelihood estimation facility in the `lrm` function. It also has a function `cr.setup` which allows the CR model to be fitted in an extremely flexible way using a binary logistic model on a modified input data set. `Design` is available at `http://www.med.virginia.edu/medicine/clinical/hes/biostat.htm`. `varclus`, `transcan`, `impute`, and `scatld` are separate functions in the `Hmisc` library in `statlib`, also written by the first author.

2. STUDY DESIGN

The WHO/ARI Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants was undertaken in Ethiopia, The Gambia, Papua New Guinea, and the Phillippines to collect data that would allow better screening criteria to be derived for finding infants at high risk of serious infection.[23] Standardized laboratory and clinical evaluations (clinical history, risk factors such as low birth weight, CXR, SaO_2, BC, LP etc.) were done. Infants brought for primary care were enrolled if *any* of the following were present: cough; difficult, fast or noise breathing; fever or hypothermia; not feeding well (less than half of normal intake); abnormally sleepy or difficult to wake; convulsions; rectal temperature $\geqslant 37{\cdot}5^{\circ}C$ or $\leqslant 35{\cdot}5^{\circ}C$; or mother volunteered that the baby was very sick, irritable, or has stopped breathing or turned blue/black. Infants were excluded if the illness began in hospital (except for delivery), the clinic visit was for trauma, burn, or routine care such as immunization, weight < 1500 g during the first 48 hours of life, there was a documented episode of previous pneumonia, sepsis, or meningitis within the last 3 weeks, if an obvious congenital malformation was present, or if the infant had previously been enrolled in the study. 8418 infants were screened and the 4552 having a positive symptom without an exclusion were enrolled.

To be a candidate for BC and CXR, an infant had to have a clinical indication for one of the three diseases, according to prespecified criteria in the study protocol ($n = 2398$). Blood work-up (but not LP) and CXR was also done on a random sample intended to be 10 per cent of infants having no signs or symptoms suggestive of infection ($n = 175$).* Infants with signs suggestive of meningitis had LP. All 4552 infants received a full physical exam and standardized pulse oximetry to measure SaO_2. The vast majority of infants getting CXR had the X-rays interpreted by three independent radiologists.

For determining outcomes, BC was considered positive if the blood culture was definite or probable for bacterial infection.[23] CSF was positive if the CSF culture was positive or there were more than 10 white cells in the CSF, after a correction was made for a bloody tap. CXR is positive if all radiologists who read the CXR and ruled it interpretable classified the infant's X-ray as definitely or probably abnormal.

* Some mothers refused having blood drawn for their children, adding a slight bias to the remaining sampled group which made them slightly more sick. At one site, the sample was systematic (every fifth patient) but sampling was done more often when mothers refused to participate

Table I. Ordinal outcome scale

Outcome level Y	Definition	n	Fraction in outcome level		
			BC, CXR indicated ($n = 2398$)	Not indicated ($n = 1979$)	Random sample* ($n = 175$)
		4552			
0	None of the above	3551	0·63	0·96	0·91
1	$90\% \leqslant SaO_2 < 95\%$ or CXR +	490	0·17	0·04†	0·05
2	BC + or CSF + or $SaO_2 < 90\%$	511	0·21	0·00‡	0·03

* A separate sample of patients not indicated for laboratory work-up but having it anyway
† SaO_2 was measured but CXR was not done
‡ *Assumed* zero since neither BC nor LP were done

The analyses which follow are not corrected for verification bias[24] with respect to BC, LP and CXR, but Section 3 has some data describing the extent of the problem.

3. ORDINAL OUTCOME SCALE

Rationale and details of the outcome scale construction are found in a background paper.[25] As discussed by Follman,[26] it is useful to derive new outcome variables (risk scores) by observing how non-fatal events predict death. We followed this scheme but extended it with a two-stage strategy. First, we found that the more important non-fatal response measures, BC+ , CSF+ , and severe hypoxemia ($SaO_2 < 90$ per cent, altitude adjusted), had roughly equal weight in predicting death* so the union of these findings was used as the top outcome level. Next, CXR and moderate hypoxemia ($90 < SaO_2 < 95$ per cent) were examined for their association with the probability that the patient had BC+ or CSF+ . These two markers were found to have the same associations with this worse outcome (probabilities of BC+ $\cup$ CSF+ were equal for CXR+ and for $SaO_2 \in$ [90 per cent, 95 per cent), and were lower but equal for CXR− and $SaO_2 \geqslant 95$ per cent).

Patients were then assigned to the worst qualifying outcome category. Table I shows the definition of the ordinal outcome variable Y and shows the distribution of Y by the laboratory work-up strategy.

The effect of verification bias is a false negative fraction of 0·03 for $Y = 2$, from comparing the detection fraction of zero for $Y = 2$ in the 'not indicated' group with the observed positive fraction of 0·03 in the random sample that was fully worked up. The extent of verification bias in $Y = 1$ is $0{\cdot}05 - 0{\cdot}04 = 0{\cdot}01$. In what follows, these biases will be ignored.

4. VARIABLE CLUSTERING

Expert clinical judgement was used to enumerate a list of clinical variables to collect, including 47 clinical signs. The list reflects the content of an expert paediatric examination. As a first step in

* Proportion of deaths for BC+ , BC− , CSF+ , CSF− , $SaO_2 < 90$ per cent, $SaO_2 \geqslant 90$ per cent were, respectively, 0·30, 0·08, 0·29, 0·05, 0·25, 0·04

coding the predictor variables, all questionnaire items that were connected (for example, using skip rules such as 'if condition was present, what was its severity?') were scored as a single variable using equally spaced codes, with 0–3 representing, for example, sign not present, mild, moderate, severe. The resulting list of clinical signs with their abbreviations is given in Table II. The signs are organized into clusters as will be discussed below. Here, `hx` stands for history, `ausc` for auscultation, and `hxprob` for history of problems. Two signs (`qcr`, `hcm`) were listed twice because they were later placed into two clusters each.

When there are many candidate predictors, several authors[20,27–29] have demonstrated that because variable clustering reduces the number of regression coefficients to test or estimate, it results in better validating models than either stepwise modelling or fitting a full model with at least one regression coefficient per candidate predictor. A commonly used variable clustering technique is a rotation of principal components (see Section 3.2 of D'Agostino *et al.*,[29] Chapters 5, 6, 8, 9, 12, 14 of Cureton and D'Agostino[30] and Sarle[31]) by which variables are separated into groups so that the first principal component of that group of variables explains the majority (for example 0·8) of the variance for that group of variables and so that the correlation of individual variables in different groups is low. Instead of using specialized variable clustering procedures, there are advantages to using traditional cluster analysis to cluster variables. Traditional cluster analysis uses a distance matrix to cluster subjects, but cluster analysis can also cluster variables by using a 'similarity matrix' as input. The advantages of this clustering approach are: (i) a multitude of similarity measures are available, including ones that allow for non-monotonic relationships,[32] (ii) there are many clustering techniques available (a concise summary may be found in Venables and Ripley[33], p. 311–315), including some that allow overlap between clusters. Here we used the matrix of squared Spearman rank correlation coefficients in the similarity matrix. The `varclus` function in the Hmisc library, which uses the S-plus hierarchical clustering function `hclust`, was used as follows:

```
vclust ← varclus(~ illd + hlt + slpm + slpl + wake + convul + hfa +
                   hfb + hfe  + hap  + hcl  + hcm + hcs  + hdi + fde +
                   chi + twb  + ldy  + apn  + lcw + nfl  + str + gru +
                   coh + ccy  + jau  + omph + csd + csa  + aro + qcr +
                   con + att  + mvm  + afe  + absu + stu  + deh + dcp +
                   crs + abb  + abk  + whz  + hdb + smi2 + abd + conj +
                   oto + puskin,   sim = "spearman")
plot(vclust)
```

The output appears in Figure 1.

Overall, the statistical clusterings made clinical sense, for example, `stu` (skin turgor) and `deh` (dehydrated) are closely related, and these two are somewhat less related to `abk` (sunken fontanelle) than to `hdi` (history of diarrhoea). In many cases, the clusters suggested by such output can be used directly in data reduction and summary scale construction. More often, though, the output serves as a starting point for clinicians to use in constructing more meaningful clinical clusters. That was the case here, and the clusters in Table II were the consensus of the clinicians who were the investigators in the WHO/ARI study. Prior subject matter knowledge plays a key role at this stage in the analysis.

See Sections 3 and 4 of D'Agostino *et al.*[29] for a full description of a process for eliciting clusters from subject-matter experts and then using statistical clustering techniques to check that each cluster represents only one central concept.

Table II. Clinical signs

Cluster name	Sign abbreviation	Name of sign	Values
bul.conv	abb	bulging fontanelle	0–1
	convul	hx convulsion	0–1
hydration	abk	sunken fontanelle	0–1
	hdi	hx diarrhoea	0–1
	deh	dehydrated	0–2
	stu	skin turgor	0–2
	dcp	digital capillary refill	0–2
drowsy	hcl	less activity	0–1
	qcr	quality of crying	0–2
	csd	drowsy state	0–2
	slpm	sleeping more	0–1
	wake	wakes less easy	0–1
	aro	arousal	0–2
	mvm	amount of movement	0–2
agitated	hcm	crying more	0–1
	slpl	sleeping less	0–1
	con	consolability	0–2
	csa	agitated state	0–1
crying	hcm	crying more	0–1
	hcs	crying less	0–1
	qcr	quality of crying	0–2
	smi2	smiling ability $\times$ age $>$ 42 days	0–2
reffort	nfl	nasal flaring	0–3
	lcw	lower chest in-drawing	0–3
	gru	grunting	0–2
	ccy	central cyanosis	0–1
stop.breath	hap	hx stop breathing	0–1
	apn	apnoea	0–1
ausc	whz	wheezing	0–1
	coh	cough heard	0–1
	crs	crepitation	0–2
hxprob	hfb	fast breathing	0–1
	hdb	difficulty breathing	0–1
	hlt	mother report respiratory problems	none, chest, other
feeding	hfa	hx abnormal feeding	0–3
	absu	sucking ability	0–2
	afe	drinking ability	0–2
labor	chi	previous child died	0–1
	fde	fever at delivery	0–1
	ldy	days in labour	1–9
	twb	water broke	0–1
abdominal	abd	abdominal distension	0–4
	jau	jaundice	0–1
	omph	omphalitis	0–1
fever.ill	illd	ge-adjusted number days ill	
	hfe	hx fever	0–1
pustular	conj	conjuctivitis	0–1
	oto	otoscopy impression	0–2
	puskin	pustular skin rash	0–1

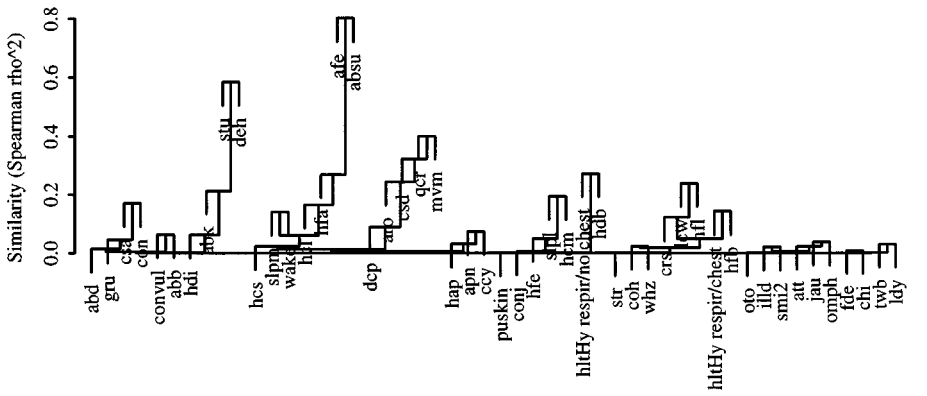

Figure 1. Hierarchical variable clustering using Spearman ρ^2 as a similarity measure for all pairs of variables. Note that because the `hlt` variable was nominal, it is represented by two dummy variables here

5. THE PROPORTIONAL ODDS MODEL AND DEVELOPING CLUSTER SUMMARY SCORES

The clusters listed in Table II were first scored by the first principal component – the linear combination of signs (using the codes in Table II) that explains the maximum variance of all signs in a cluster explainable by a single dimension (a single linear combination).[29,34,35] We denote the first principal component by PC_1. Knowing that the resulting weights may be too complex for clinical use, the primary reasons for analysing the principal components was to see if some of the clusters could be removed from consideration so that the clinicians would not spend time developing scoring rules for them. We decided to 'peak' at Y to assist in scoring clusters at this point, but to do so in a very structured way that did not involve the examination of a large number of individual coefficients.

We did not actually compute PC_1s on the raw signs but rather used a psychometric scaling technique similar to that of Kuhfeld[36] which is implemented in the S-plus `transcan` function. Here, for any cluster which contains a sign with more than two levels, the levels are automatically re-scored so as to increase the variance explained by the PC_1. This could be called a non-linear principal components analysis.

To judge any cluster scoring scheme, we had to pick a tentative outcome model. For this purpose we chose the most commonly used ordinal logistic model, which was described in Walker and Duncan[1] and later called the *proportional odds* (*PO*) *model* by McCullagh.[14] The PO model is stated as follows:

$$\Pr[Y \geqslant j \mid X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]} \tag{1}$$

where there is an implicit assumption that the regression coefficients (β) are independent of j, the cut-off level for Y. Note that the PO model makes no assumption whatever about the magnitude of spacings between levels of Y. By using the 14 PC_1s corresponding to the 14 clusters, the fitted PO model had a likelihood ratio (LR) χ^2 of 1155 with 14 d.f., and the predictive discrimination of

Table III. Clinician combinations, rankings and scorings of signs

Cluster	Combined/ranked signs in order of severity	Weights
bul.conv	abb∪convul	0–1
drowsy	hcl, qcr > 0, csd > 0∪slpm∪wake, aro > 0, mvm > 0	0–5
agitated	hcm, slpl, con = 1, csa, con = 2	0, 1, 2, 7, 8, 10
refort	nfl > 0, lcw > 1, gru = 1, gru = 2, ccy	0–5
ausc	whz, coh, crs > 0	0–3
feeding	hfa = 1, hfa = 2, hfa = 3, absu = 1∪afe = 1, absu = 2∪afe = 2	0–5
abdominal	jau∪abd > 0∪omph	0–1

the clusters was quantified by a Somers' D_{xy} rank correlation between $X\hat{\beta}$ and Y of 0·596.* The following clusters were not statistically important predictors and we assumed that the lack of importance of the PC_1s in predicting Y (adjusted for the other PC_1s) justified a conclusion that no sign within that cluster was clinically important in predicting Y: `hydration`, `hxprob`, `pustular`, `crying`, `fever.ill`, `stop.breath`, `labor`. This list was identified using a backward step-down procedure on the full model. The total Wald χ^2 for these 7 PC_1s was 22·4 ($P = 0{\cdot}002$). The reduced model had LR $\chi^2 = 1133$ with 7 d.f., $D_{xy} = 0{\cdot}591$. The bootstrap validation in Section 13 penalized for fitting the 7 predictors.

During a meeting of the study group, the clinicians were asked to rank the clinical severity of signs within each potentially important cluster. During this step, the clinicians also ranked severity levels of some of the component signs, and some cluster scores were simplified, especially when the signs within a cluster occurred infrequently. The clinicians also assessed whether the severity points or weights should be equally spaced, assigning unequally spaced weights for one cluster (`agitated`). The resulting rankings and sign combinations are shown in Table III. The signs or sign combinations separated by a comma are treated as separate categories, whereas some signs were unioned ('or'-ed) when the clinicians deemed them equally important. As an example, if an additive cluster score was to be used for `drowsy`, the scorings would be 0 = none present, 1 = `hcl`, 2 = `qcr` > 0, 3 = `csd` > 0 or `slpm` or `wake`, 4 = `aro` > 0, 5 = `mvm` > 0 and the scores would be added.

This table reflects some data reduction already (unioning some signs and selection of levels of ordinal signs) but more reduction is needed. Even after signs are ranked within a cluster, there are various ways of assigning the cluster scores. We investigated six methods. We started with the purely statistical approach of using PC_1 to summarize each cluster. Second, all sign combinations within a cluster were unioned to represent 0–1 cluster score. Third, only sign combinations thought by the clinicians to be severe were unioned, resulting in `drowsy`=`aro`>0 or `mvm`>0, `agitated`=`csa` or `con`=2, `reffort`=`lcw`>1 or `gru`>0 or `ccy`, `ausc`=`crs`>0, and `feeding`=`absu`>0 or `afe`>0. For clusters that are not scored 0–1 in Table III, the fourth summarization method was a hierarchical one which used the weight of the worst applicable category as the cluster score. For example, if `aro`=1 but `mvm`=0, `drowsy` would be scored as 4. The fifth method counted the number of positive signs in the cluster. The sixth method summed the weights of all signs or sign

* See reference 20 for details; $D_{xy} = 2(c - \frac{1}{2})$ where c is the probability of concordance between pairs of $X\hat{\beta}$ and Y values, which is a generalization of a receiver operating characteristic curve area

Table IV. Predictive information of various cluster scoring strategies

Scoring method	LR χ^2	d.f.	AIC
PC_1 of each cluster	1133	7	1119
Union of all signs	1045	7	1031
Union of higher categories	1123	7	1109
Hierarchical (worst sign)	1194	7	1180
Additive, equal weights	1155	7	1141
Additive using clinician weights	1183	7	1169
Hierarchical, data-driven weights	1227	25	1177

combinations present. Finally, the worst sign combination present was again used as in the second method, but the points assigned to the category were data driven ones obtained by using extra dummy variables. This provides an assessment of the adequacy of the clinician-specified weights. By comparing rows 4 and 7 in Table IV we see that response data-driven sign weights have a slightly worse Akaike information criterion (AIC) or LR $\chi^2 - 2 \times$ d.f. (which penalizes the model for complexity[37]), indicating that the number of extra β parameters estimated was not justified by the improvement in χ^2. The hierarchical method, using the clinicians' weights, performed quite well. The only cluster with inadequate clinician weights was `ausc` – see following. The PC_1 method, without any guidance, performed well, as in Harrell *et al.*[19] The only reasons not to use it are that it requires a coefficient for every sign in the cluster and coefficients are not translatable into simple scores such as $0, 1, \ldots$.

Representation of clusters by a simple union of selected signs or of all signs is inadequate, but otherwise the choice of methods is not very important in terms of explaining variation in Y. We chose the fourth method, a hierarchical severity point assignment (using weights which were prespecified by the clinicians), for its ease of use and of handling missing component variables (in most cases) and potential for speeding up the clinical exam (examining to detect more important signs first). Because of what was learned regarding the relationship between `ausc` and Y, we modified the `ausc` cluster score by redefining it as `ausc=crs>0` (crepitations present). Note that neither the 'tweaking' of `ausc` nor the examination of the seven scoring methods displayed in Table IV will be taken into account in the model validation.

One attractive alternative approach that we did not try was the battery reduction strategy described in Chapter 12 of Cureton and D'Agostino[30] in which one finds a subset of the variables in each cluster whose PC_1 adequately represents the whole cluster's PC_1.

6. ASSESSING ORDINALITY OF Y FOR EACH X, AND UNADJUSTED CHECKING OF PO AND CR ASSUMPTIONS

A basic assumption of all commonly used ordinal regression models is that the response variable behaves in an ordinal fashion with respect to each predictor. Assuming that a predictor X is linearly related to the log odds of some appropriate event, a simple way to check for ordinality is to plot the mean of X stratified by levels of Y (denote these by $\hat{E}(X|Y = y)$). These means should be in a consistent order. If for many of the Xs, two adjacent categories of Y do not distinguish the means, that is evidence that those levels of Y should be pooled.

One can also estimate the mean or expected value of $X|Y=j$ given that the ordinal model assumption hold. This is a useful tool for examining those assumptions, at least in an unadjusted fashion. For simplicity, assume that X is discrete, and let $P_{jx} = \Pr(Y=j|X=x, \text{model})$ be the probability that $Y=j$ given $X=x$ that is dictated from the model being fitted, with X being the only predictor in the model. Then

$$\Pr(X=x|X=j, \text{model}) = \Pr(Y=j|X=x, \text{model})\Pr(X=x)/\Pr(Y=j)$$
$$E(X|Y=j, \text{model}) = \sum_x xP_{jx}\Pr(X=x)/\Pr(Y=j) \tag{2}$$

and the expectation can be estimated by

$$\hat{E}(X|Y=j, \text{model}) = \sum_x x\hat{P}_{jx}f_x/g_j \tag{3}$$

where $\hat{P}_{jx}$ denotes the estimate of P_{jx} from the fitted 1-predictor model*, f_x is the frequency of $X=x$ in the sample of size n, and g_j is the frequency of $Y=j$ in the sample. This estimate can be computed conveniently without grouping the data by X. For n subjects let the n values of X be $x_1, x_2, \ldots, x_n$. Then

$$\hat{E}(X|Y=j) = \sum_{i=1}^{n} x_i\hat{P}_{jx_i}/g_j. \tag{4}$$

Figure 2 was produced by the S-plus function `plot.xmean.ordinaly` in the `Design` library, which plots simple Y-stratified means overlaid with $\hat{E}(X|Y=j, \text{model})$, with j on the x-axis. Here we expect strongly non-linear effects for `temp`, `rr` and `hrat`, so for those predictors we plot the mean absolute differences from suitable 'normal' values as an approximate solution:

```
par(mfrow=c(3,4)) #3 x 4 matrix of plots
plot.xmean.ordinaly(Y ~ age + abs(temp-37) + abs(rr-60) + abs(hrat-125) +
                    waz + bul.conv + drowsy + agitated + reffort +
                    ausc + feeding + abdominal, cr=T)
```

The plot is shown in Figure 2. Y does not seem to operate in an ordinal fashion with respect to `age`, `|rr-60|` or `ausc`. For the other variables, ordinality holds, and PO holds reasonably well except for `bul.conv`, `drowsy` and `abdominal`. For heart rate, the PO assumption appears to be satisfied perfectly. CR model assumptions appear to be no less tenuous than PO assumptions, at least when fitting one variable at a time.

There is a relationship between score residuals defined later in equation (6) and a slightly different comparison between stratified means of X and expected values under the model. Suppose that simple means and expected values were computed under the condition that $Y \geqslant j$ instead of the condition $Y=j$. Then $\hat{E}(X|Y \geqslant j) - \hat{E}(X|Y \geqslant j, \text{model})$ is proportional to the mean of the score residuals for the PO model.

7. A TENTATIVE FULL PROPORTIONAL ODDS MODEL

Using summary cluster scores that were developed in Section 5, the original list of 14 clusters with 47 signs was reduced to 7 predictors as listed in Table III: two unions of signs (`bul.conv`,

* For inner values of Y in the PO models, these probabilities are differences between terms given by equation (1)

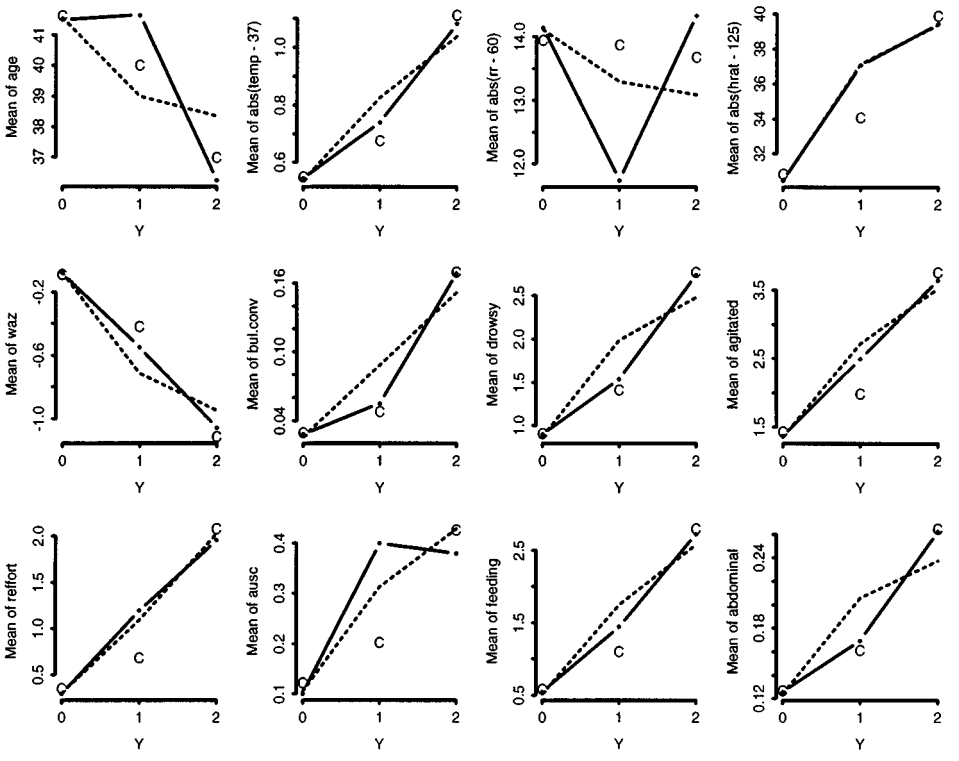

Figure 2. Examination of the ordinality of Y for each predictor by assessing if differing Y distinguish the mean X and if the the trend is monotonic. Solid lines connect the simple stratified means, and dashed lines connect the estimated expected value of $X|Y = j$ given that PO holds. Estimated expected values from the CR model are marked with c's

abdominal); one single sign (ausc), and four 'worst category' point assignments (drowsy, agitated, reffort, feeding). Seven clusters were dropped because of weak associations with Y*. Such a limited use of variable selection avoids most of the severe problems inherent with that technique such as the lack of replication of the list of 'significant' variables if one sampled repeatedly from the same population.[20]

At this point in model development we added to the model age and vital signs: temp (temperature); rr (respiratory rate); hrat (heart rate); and waz, weight-for-age Z-score. Since age was expected to modify the interpretation of temp, rr and hrat, and interactions between continuous variables would be difficult to use in the field, we categorized age into three intervals: 0–6 days ($n = 302$); 7–59 days ($n = 3042$); and 60–90 days ($n = 1208$).† This was done with the S-plus cut2 function from the Hmisc library in statlib:

```
ageg ← cut2(age, c(7,60))
```

* These clusters were reinstated as candidate predictors in the final model validation, to penalize for having tested them for association with Y

† These age intervals were also found to adequately capture more of the interaction effects

The new variables temp, rr, hrat, waz were missing in, respectively, $n = 13$, 11, 147 and 20 infants. Because the three vital sign variables are somewhat correlated with each other, customized imputation models were developed to impute all the missing values without assuming linearity or even monotonicity of any of the regressions. The S-plus transcan and impute functions from Hmisc were used to impute vital signs as follows:

```
vsign.trans ← transcan(~ temp + hrat + rr, imputed=T)
temp ← impute(vsign.trans, temp)
hrat ← impute(vsign.trans, hrat)
rr   ← impute(vsign.trans, rr)
```

After transcan estimated optimal restricted cubic spline transformations, temp could be predicted with adjusted $R^2 = 0{\cdot}17$ from hrat and rr, hrat could be predicted with adjusted $R^2 = 0{\cdot}14$ from temp and rr, and rr could be predicted with adjusted R^2 of only 0·06. The first two R^2, while not large, mean that customized imputations are more efficient than imputing with constants. Imputations on rr were closer to the median rr of 48/minute as compared with the other two vital signs whose imputed values have more variation. In a similar manner, waz was imputed using age, birth weight, head circumference, body length, and prematurity (adjusted R^2 for predicting waz from the others was 0.74).

The continuous predictors temp, hrat, rr were not assumed to linearly relate to the log odds that $Y \geqslant j$. Flexible piecewise cubic polynomials (restricted cubic spline functions[38–41] with 5 knots or join points for temp, rr and 4 knots for hrat,waz*) were used to model the effects of these variables, using the rcs function with the binary and PO logistic regression function lrm in the Design library:

```
f1 ← lrm(Y ~ ageg * (rcs(temp,5) + rcs(rr,5) + rcs(hrat,4)) + rcs(waz,4) +
         bul.conv + drowsy + agitated + reffort + ausc +
         feeding + abdominal, x=T, y=T) #x=T, y=T used by resid( ) below
```

Here the asterisk in the formula indicates that main effects and interactions are to be fitted. This model has LR χ^2 of 1393 with 45 d.f. and $D_{xy} = 0{\cdot}653$. Wald tests of non-linearity and interaction are obtained using the statement anova(f1), whose output is shown in Table V.†

The bottom four lines of the table are the most important. First, there is strong evidence that some associations with Y exist (45 d.f. test) and very strong evidence of non-linearity in one of the vital signs or in waz (26 d.f. test). There is moderately strong evidence for an interaction effect somewhere in the model (22 d.f. test). The anova output does not contain Wald χ^2 statistics for main effects alone as these are meaningless; main effect parameters are pooled with interaction ('higher order') parameters to yield meaningful overall tests for predictors. We see that the grouped age variable ageg is predictive of Y, but mainly as an effect modifier for rr. temp is extremely non-linear and rr is moderately so. hrat, a difficult variable to measure reliably in young infants, is perhaps not important enough ($\chi^2 = 19{\cdot}0$, 9 d.f.) to keep in the final model.

* Four knots were used for hrat because it was thought to be less important *a priori*, and fewer were used for waz because it was thought to operate almost linearly

† Actually, the statement which produced this output is latex(anova(f1)), which typeset the output using the LaTeX document processing language[42]

Table V. Wald statistics for Y in the proportional odds model

	LR χ^2	d.f.	P
ageg (factor + higher order factors)	41·49	24	0·0147
All interactions	40·48	22	0·0095
temp (factor + higher order factors)	37·08	12	0·0002
All interactions	6·77	8	0·5617
Non-linear (factor + higher order factors)	31·08	9	0·0003
rr (factor + higher order factors)	81·16	12	< 0·0001
All interactions	27·37	8	0·0006
Non-linear (factor + higher order factors)	27·36	9	0·0012
hrat (factor + higher order factors)	19·00	9	0·0252
All interactions	8·83	6	0·1836
Non-linear (factor + higher order factors)	7·35	6	0·2901
waz	35·82	3	< 0·0001
Non-linear	13·21	2	0·0014
bul.conv	12·16	1	0·0005
drowsy	17.79	1	< 0·0001
agitated	8·25	1	0·0041
reffort	63·39	1	< 0·0001
ausc	105·82	1	< 0·0001
feeding	30·38	1	< 0·0001
abdominal	0·74	1	0·3895
ageg × temp (factor + higher order factors)	6·77	8	0·5617
Non-linear	6·40	6	0·3801
Non-linear interaction: $f(A, B)$ versus AB	6·40	6	0·3801
ageg × rr (factor + higher order factors)	27·37	8	0·0006
Non-linear	14·85	6	0·0214
Non-linear interaction: $f(A, B)$ versus AB	14.85	6	0·0214
ageg × hrat (factor + higher order factors)	8·83	6	0·1836
Non-linear	2·42	4	0·6587
Non-linear interaction: $f(A, B)$ versus AB	2·42	4	0·6587
Total non-linear	78·20	26	< 0·0001
Total interaction	40·48	22	0·0095
Total non-linear + interaction	96·31	32	< 0·0001
Total	1073·78	45	< 0·0001

The clinicians did not presuppose that any of the clinical signs had special importance in combination with other signs or with vital signs. Therefore interactions involving the clinical signs, which would have been great in number, were not examined.

8. RESIDUALS FOR CHECKING THE PROPORTIONAL ODDS ASSUMPTION

Peterson and Harrell[15] developed score and likelihood ratio tests for testing the PO assumption. The score test is used in the SAS LOGISTIC procedure,[43] but it yields P-values that are far too small in many cases.[15] Other techniques, especially graphical ones, are needed for verifying PO. Schoenfeld residuals[44] are very effective[45] in checking the proportional hazards assumption in the Cox[46] survival model. For the PO model one could analogously compute each subject's contribution to the first derivative of the log-likelihood function with respect to β_m, average them

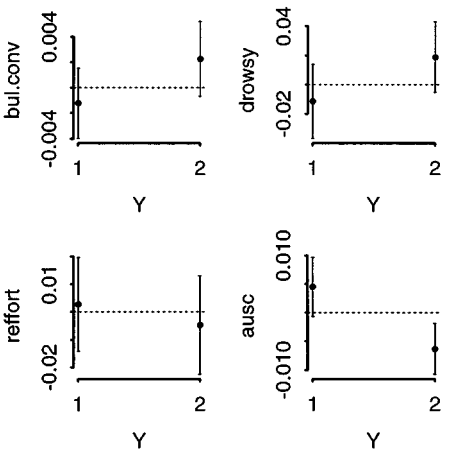

Figure 3. Binary logistic model score residuals for binary events derived from two cut-offs of the ordinal response Y. Note that the mean residuals, marked with closed circles, correspond closely with differences between solid and dashed lines at $Y = 1, 2$ in Figure 2. Bars are 0·95 confidence limits. Score residual assessments for spline-expanded variables such as `rr` would have required one plot per d.f.

separately by levels of Y, and examine trends in the residual plots. A few examples have shown that such plots are usually hard to interpret. Easily interpreted score residual plots for the PO model can be constructed however by using the fitted PO model to predict a series of binary events $Y \geqslant j, j = 1, 2, \ldots, k$, using the corresponding predicted probabilities

$$\hat{P}_{ij} = \frac{1}{1 + \exp[-(\hat{\alpha}_j + X_i\hat{\beta})]} \tag{5}$$

where X_i stands for a vector of predictors for subject i. Then, after forming an indicator variable for the event currently being predicted ($[Y_i \geqslant j]$), one computes the score (first derivative) components U_{im} from an ordinary binary logistic model:

$$U_{im} = X_{im}([Y_i \geqslant j] - \hat{P}_{ij}) \tag{6}$$

for the subject i and predictor m. Then, for each column of U, plot the mean $\bar{U}_{\cdot m}$ and confidence limits, with Y (that is, j) on the x-axis. For each predictor the trend against j should be flat if PO holds.*

For the tentative PO model, score residuals for four of the variables were plotted using

```
par(mfow=c(2,2))
resid(f1, 'score.binary', pl=T, which=c(17,18,20,21))
```

The result is shown in Figure 3. We see strong evidence of non-PO for `ausc` and moderate evidence for `drowsy` and `bul.conv`, in agreement with Figure 2.

* If $\hat{\beta}$ were derived from separate binary fits, all $\bar{U}_{\cdot m} \equiv 0$

In binary logistic regression, *partial residuals* are very useful because after the analyst fits linear effects for all predictors, computes partial residuals, and smooths the relationship between each predictor and its partial residuals, the resulting trend is an estimate of the true relationship between each predictor and the log odds. The partial residual is defined as follows, for the ith subject and mth predictor variable:[47,48]

$$r_{im} = \hat{\beta}_m X_{im} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)} \tag{7}$$

where

$$\hat{P}_i = \frac{1}{1 + \exp[-(\hat{\alpha} + X_i\hat{\beta})]}. \tag{8}$$

A smoothed plot (for example, using the moving linear regression algorithm in `lowess`[49]) of X_{im} versus r_{im} provides a non-parametric estimate of how X_m relates to the log relative odds that $Y = 1 | X_m$.

For ordinal Y, we just need to compute binary model partial residuals for all cut-offs j:

$$r_{im} = \hat{\beta}_m X_{im} + \frac{[Y_i \geqslant j] - \hat{P}_{ij}}{\hat{P}_{ij}(1 - \hat{P}_{ij})} \tag{9}$$

then to make a plot for each m showing smoothed partial residual curves for all j, looking for similar shapes and slopes for a given predictor for all j. Each curve provides an estimate of how X_m relates to the relative log odds that $Y \geqslant j$. Since partial residuals allow examination of predictor transformation (linearity) while simultaneously allowing examination of PO (parallelism), partial residual plots are generally preferred over score residual plots for ordinal models.

In Figure 4, smoothed partial residual plots were obtained for all predictors, after first fitting a simple model in which every predictor was assumed to operate linearly. Interactions were temporarily ignored and `age` was used as a continuous variable:

```
f2 ← lrm(Y ~ age + temp + rr + hrat + waz +
         bul.conv + drowsy + agitated + reffort + ausc +
         feeding + abdominal, x=T, y=T)
par(mfrow=c(3,4))

resid(f2, 'partial', pl=T)   # pl=T : plot
```

The degree of non-parallelism generally agreed with the degree of non-flatness in Figure 3 and with the other score residual plots which were not shown. The partial residuals show that `temp` is highly non-linear and that it is much more useful in predicting $Y = 2$. For the cluster scores, the linearity assumption appears reasonable, except possibly for `drowsy`. Other non-linear effects will be taken into account using splines as before (except for `age`, which will be categorized).

A model can have significant lack of fit with respect to some of the predictors and still yield quite accurate predictions. To see if the case for this PO model, we computed predicted probabilities of $Y = 2$ for all infants from the model and compared these with predictions from a binary logistic model derived specifically to predict $\Pr(Y = 2)$ (that is, a model with no assumptions connecting different levels of Y). The mean absolute difference in predicted probabilities between the two models is only 0·02, but the 0·90 quantile of that difference is 0·059. For

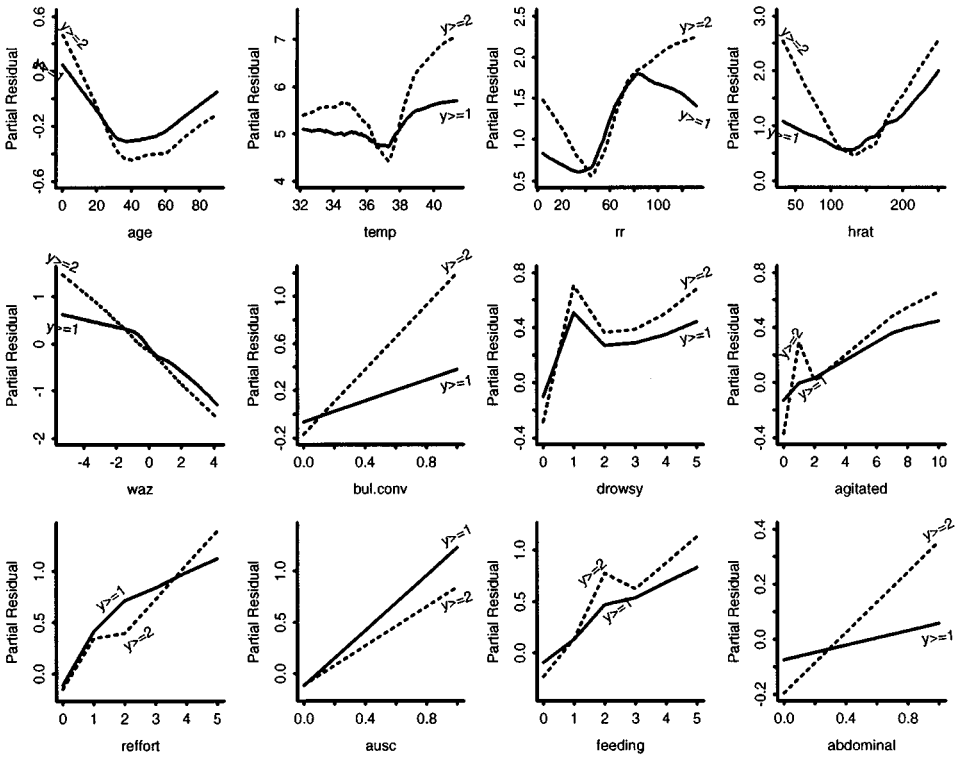

Figure 4. Smoothed partial residuals corresponding to two cut-offs of Y, from a model in which all predictors were assumed to operate linearly and additively. The smoothed curves estimate the actual predictor transformations needed.

high-risk infants, discrepancies of 0·2 were common. Therefore we elected to consider a different model.

9. CONTINUATION RATIO ORDINAL LOGISTIC MODEL

Unlike the PO model, which is based on cumulative probabilities, the continuation ratio (CR) model is based on conditional probabilities. The (forward) CR model[2,6,8] is stated as follows for $Y = 0, \ldots, k$ (here $k = 2$):

$$\begin{aligned} \Pr(Y = j \mid Y \geqslant j, X) &= \frac{1}{1 + \exp[-(\theta_j + X\gamma)]} \\ \text{logit}(Y = 0 \mid Y \geqslant 0, X) &= \text{logit}(Y = 0 \mid X) \\ &= \theta_0 + X\gamma \\ \text{logit}(Y = 1 \mid Y \geqslant 1, X) &= \theta_1 + X\gamma. \end{aligned} \tag{10}$$

The CR model has been said to be likely to fit ordinal responses when subjects have to 'pass through' one category to get to the next (which is the case here with respect to SaO_2). The CR model is a discrete version of the Cox proportional hazards model.

To check CR model assumptions, binary logistic model partial residuals are again valuable. We fit a sequence of binary logistic models using a series of binary events and the corresponding applicable (increasingly small) subsets of subjects, and plot smoothed partial residuals against X for all of the binary events. In S-plus we now fit the sequence of binary fits and then use the plot.lrm.partial function, which assembles partial residuals for a sequence of fits and constructs one graph per predictor:

```
cr0 ← lrm(Y==0 ~ age + temp + rr + hrat + waz +
          bul.conv + drowsy + agitated + reffort + ausc +
          feeding + abdominal, x=T, y=T)
# Use the update function to save repeating model right hand side
# An indicator variable for Y=1 is the response variable below
cr1 ← update(cr0, Y==1 ~ ., subset=Y>=1)
plot.lrm.partial(cr0, cr1, center=T)
```

The output is in Figure 5. There is not much more parallelism here than in Figure 4. For the two most important predictors, ausc and rr, there are strongly differing effects for the differing events being predicted (for example, $Y = 0$ vs. $Y = 1 | Y \geqslant 1$). As is often the case, there is no one constant β model that satisfies assumptions with respect to all predictors simultaneously, especially when there is evidence for non-ordinality for ausc in Figure 2. The CR model will need to be generalized to adequately fit this data set.

10. EXTENDED CONTINUATION RATIO MODEL

By comparing Figures 4 and 5 it is seen that the CR model in its ordinary form has no advantage over the PO model for this data set. The PO model has been extended by Peterson and Harrell[15] to allow for unequal slopes for some or all of the X's for some or all levels of Y. This partial PO model requires specialized software, and with the demise of SAS Version 5 PROC LOGIST, software is not currently available. Armstrong and Sloan[6] and Berridge and Whitehead[8] showed how the CR model can be fitted using ordinary binary logistic model software, after certain rows of the X matrix are duplicated and a new binary Y vector is constructed. For each subject, one constructs separate records by considering successive conditions $Y \geqslant 0$, $Y \geqslant 1$, ... , $Y \geqslant k-1$ for a response variable with values $0, 1, \ldots, k$. The binary response for each applicable condition or 'cohort' is set to 1 if the subject failed at the current 'cohort' or 'risk set', that is, if $Y = j$ where the cohort being considered is $Y \geqslant j$. The constructed cohort variable is carried along with the new X and Y. This variable is considered to be categorical and its coefficients are fitted by adding $k-1$ dummy variables to the binary logistic model. The CR model is restated as follows:

$$\Pr(Y = j | Y \geqslant j, X) = \frac{1}{1 + \exp[-(\alpha + \theta_j + X\gamma)]}. \tag{11}$$

Here α is an overall intercept, $\theta_0 \equiv 0$, and $\theta_1, \ldots, \theta_{k-1}$ are increments from α. In S-plus notation, the model is

$$\texttt{y} \sim \texttt{cohort} + \texttt{X1} + \texttt{X2} + \texttt{X3} + \ldots,$$

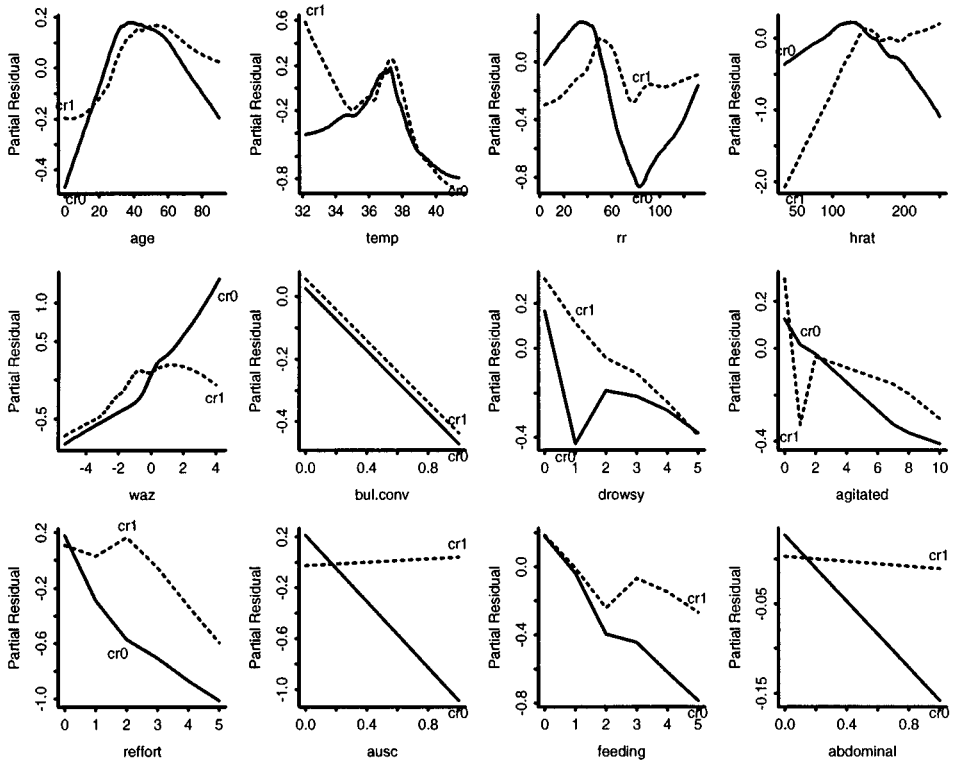

Figure 5. lowess smoothed partial residual plots for binary models which are components of an ordinal continuation ratio model

with cohort denoting a polytomous variable and the columns of X denoted by X1, X2 ... , etc. The CR model can be extended to allow for some or all of the γ's to change with the cohort or Y-cut-off.[6] Suppose that non-constant slope is allowed for X1 and X2. The S-plus notation for the extended model would be

$$\texttt{y} \sim \texttt{cohort} * (\texttt{X1} + \texttt{X2}) + \texttt{X3}$$

The interaction notation ($*$) implies that lower-order effects are also included in the model. The extended CR model is a discrete version of the Cox survival model with time-dependent covariables.

The cr.setup function in Design returns a list of vectors useful in constructing a data set used to 'trick' a binary logistic function into fitting CR models. The subs vector in this list contains observation numbers in the original data, some of which are repeated so that subscripting on subs will cause the subject's row of predictor variable values to be replicated the desired number of times ($\min(Y + 1, k)$ times if $Y = 0, 1, \ldots, k$). Each replication corresponds to the conditioning event or risk set (cohort) $Y \geqslant j$, and the new dummy response variable y indicates whether or not $Y = j$.


```
u ← cr.setup(Y)            # Y is original ordinal response vector
attach(mydata[u$subs,]) # my data is the original dataset
                           # mydata[i,] subscripts the input data,
                           # using duplicate values of i for repeats
y       ← u$y              # constructed binary response
cohort  ← u$cohort         # cohort or risk set categories
```

Here the cohort variable has values 'all', 'Y >= 1' corresponding to the conditioning events in equation (10). After the attach command runs, vectors such as age are lengthened (to 5553 records) by duplicating the correct observations according to the magnitude of a subject's Y value. Now we fit a fully extended CR model which makes no equal slopes assumptions, that is, the model *has* to fit Y assuming the covariables are linear and additive. At this point, we omit hrat but add back all variables which were deleted by examining their association with Y. Recall that most of these 7 cluster scores were summarized using PC_1. Adding back 'insignificant' variables will allow us to validate the model fairly using the bootstrap, as well as to obtain confidence intervals which are not falsely narrow.[50]

```
full ← lrm(y ~ cohort*(ageg*(rcs(temp,5) + rcs(rr,5)) + rcs(waz,4) +
                       bul.conv + drowsy + agitated + reffort +
                       ausc + feeding + abdominal +
                       hydration + hxprob + pustular + crying +
                       fever.ill + stop.breath + labor), x=T, y=T)
# x=T, y=T is for pentrace, validate, calibrate below
latex(anova(full))
```

This model has LR $\chi^2 = 1824$ with 87 d.f. Wald statistics produced by anova(full) are in Table VI. For brevity, tests of non-linear effects and many tests with $P > 0{\cdot}1$ are not shown.

The global test of the constant slopes assumption in the CR model (test of all interactions involving cohort) has $\chi^2 = 172$ with 43 d.f., $P < 0{\cdot}0001$. Consistent with Figure 5, the formal tests indicate that ausc is the biggest violator, followed by waz and rr.

At this point we select the CR model for this problem because of its flexibility, both in testing the equal slopes assumption and in parameterizing non-equal-slopes extensions.

11. PENALIZED ESTIMATION

The traditional estimation technique used for logistic models, maximum likelihood estimation (MLE), is optimal (that is, has lowest variance) among techniques which yield unbiased estimates for large samples. The bias of an estimator is not its most important attribute, however. The probability that a parameter estimate $\hat{\beta}_i$ is close to the true population value is a function of the mean squared error of that estimate, which is the variance plus the square of the bias. In many cases, especially for small samples, one can sacrifice the bias and lower the variance by a sufficient amount so that the mean squared error of the estimate is lower than that of the MLE. As an example, if one were using patients' sex to predict mortality and there were two males in the sample, the probability of death for males would be more reliably estimated by 'shrinking' it toward the probability of death for the overall sample, which is dominated by females. Shrinkage can be much more general, for example, shrinking a non-linear regression effect toward a linear effect if the evidence for non-linearity is weak.

Table VI. Wald statistics for y in the extended CR model

	χ^2	d.f.	P
cohort	199·47	44	< 0·0001
All interactions	172·12	43	< 0·0001
ageg	48·89	36	0·0742
temp	59·37	24	0·0001
rr	93·77	24	< 0·0001
waz	39·69	6	< 0·0001
bul.conv	10·80	2	0·0045
drowsy	15·19	2	0·0005
agitated	13·55	2	0·0011
reffort	51·85	2	< 0·0001
ausc	109·80	2	< 0·0001
feeding	27·47	2	< 0·0001
hxprob	6·62	2	0·0364
stop.breath	5.34	2	0·0693
labor	5·35	2	0·0690
ageg × temp	8·18	16	0·9432
ageg × rr	38·11	16	0·0015
cohort × rr	19·67	12	0·0736
cohort × waz	9·04	3	0·0288
cohort × ausc	38·11	1	< 0·0001
cohort × fever.ill	3·17	1	0·0749
cohort × stop.breath	2·99	1	0·0839
cohort × ageg × temp	2·22	8	0·9736
cohort × ageg × rr	10·22	8	0·2500
Total non-linear	93·36	40	< 0·0001
Total interaction	203·10	59	< 0·0001
Total non-linear + interaction	257·70	67	< 0·0001
Total	1211·73	87	< 0·0001

Penalized MLE (PMLE)[51–53] is a general technique for shrinking (stabilizing) regression fits. Instead of maximizing the log-likelihood, PMLE maximizes a penalized log-likelihood which is the sum of the ordinal model log-likelihood and a penalty, resulting in

$$\log L - \frac{1}{2}\lambda \sum_{i=1}^{p} (s_i \beta_i)^2. \tag{12}$$

Here $s_1, s_2, \ldots, s_p$ are scale factors chosen to make $s_i\beta_i$ unitless. Most authors standardize the data first and do not have scale factors in the equation,[51] but equation (12) has the advantage of allowing estimation of β on the original scale of the data. The usual methods (for example, Newton–Raphson) are used to maximize equation (12). The usual default values for s are sample standard deviations of columns of the design matrix, but special consideration has to be given to dummy variables,[52] which gives rise to a more general form of the penalized log-likelihood

$$\log L - \frac{1}{2}\lambda \beta' P \beta \tag{13}$$

where P is a penalty matrix. Rows and columns of P can easily be set to zero for parameters for which no shrinkage is desired.[52,53]

The main problem in using PMLE is the choice of λ. Many authors use cross-validation to solve for the λ which optimizes an unbiased estimate of predictive accuracy, but it is easy to show that one must use a huge number of data splits to get a precise estimate of the optimum λ. A faster and usually more reliable strategy, based on findings from a small number of simulation studies, is to choose the λ which maximizes the 'effective' AIC. Gray (Eq. 2·9)[53] and others show how to compute the 'effective d.f.' in this situation (that is, higher λ causes more shrinkage which lowers the effective d.f.). The effective AIC is

$$\text{LR } \chi^2 - 2 \times \text{effective d.f.} \tag{14}$$

where LR χ^2 is the likelihood ratio χ^2 for the penalized model, but ignoring the penalty function.

The `lrm` function will do PMLE, and a separate function called `pentrace` searches for the optimum λ based on effective AIC once the analyst specifies a vector of λs to try. `pentrace` can also allow for differing λ for different types of terms in the model. Here we want to do a grid search to determine the optimum penalty for simple main effect (non-interaction) terms and the penalty for interaction terms, most of which are terms interacting with `cohort` to allow for unequal slopes. The following code uses `pentrace` on the full extended CR model fit to find the optimum penalty factors. All combinations of `simple` and `interaction` λ's for which the interaction penalty $\geqslant$ the penalty for the simple parameters are examined. The range of penalty factors to try for each type of parameter was found by computing effective AIC in a trial and error process.

```
pentrace(full, list(simple=c(0,.025,.05,.075,.1),
                    interaction=c(0,10,50,100,125,150)))
Best penalty:

  simple interaction       df      aic
    0.05         125    49.75 1672.6

  simple  interaction      df      aic      bic    aic.c
   0.000            0 87.000 1650.3 1074.2 1647.5
   0.000           10 60.628 1670.8 1269.4 1669.5
   0.025           10 60.110 1671.6 1273.5 1670.2
   0.050           10 59.797 1671.6 1275.6 1670.3
   0.075           10 59.581 1671.5 1276.9 1670.2
   0.100           10 59.421 1671.3 1277.8 1670.0
   0.000           50 54.640 1671.3 1309.5 1670.2
   0.025           50 54.135 1672.0 1313.5 1670.9
   0.050           50 53.829 1672.0 1315.5 1670.9
   0.075           50 53.619 1671.8 1316.8 1670.8
   0.100           50 53.463 1671.6 1317.6 1670.5
   0.000          100 51.613 1671.9 1330.1 1670.9
   0.025          100 51.113 1672.5 1334.1 1671.6
   0.050          100 50.809 1672.6 1336.1 1671.6
   0.075          100 50.600 1672.4 1337.3 1671.4
   0.100          100 50.445 1672.1 1338.1 1671.2
   0.000          125 50.553 1671.9 1337.2 1671.0
```



```
0.025      125 50.054 1672.6 1341.1 1671.7
0.050      125 49.750 1672.6 1343.2 1671.7
0.075      125 49.542 1672.4 1344.4 1671.5
0.100      125 49.387 1672.1 1345.1 1671.3
0.000      150 49.653 1671.8 1343.0 1670.9
0.025      150 49.155 1672.5 1347.0 1671.6
0.050      150 48.852 1672.5 1349.0 1671.6
0.075      150 48.643 1672.3 1350.2 1671.5
0.100      150 48.489 1672.1 1351.0 1671.2
```

We see that shrinkage from 87 d.f. down to 49·8 effective d.f. results in an increase in AIC of 22·3. The optimum penalty factors were 0·05 for simple terms and 125 for interaction terms.*

We now store a penalized version of the full fit, determine the kind of model terms for which the effective d.f. were reduced, and compute χ^2 for each factor in the model. We take the effective d.f. for a collection of model parameters to be the sum of the diagonals of the matrix product defined underneath Gray's Eq. 2·9[53] that correspond to those parameters:

```
full.pen ← update(full, penalty=list(simple=.05, interaction=125))
effective.df(full.pen)
Original and Effective Degrees of Freedom
                                Original    Penalized
                     All            87        49.75
             Simple Terms           20        19.98
  Interaction or Nonlinear          67        29.77
               Nonlinear            40        16.82
             Interaction            59        22.57
  Nonlinear Interaction             32         9.62
plot(anova(full.pen))
somers2(predict(full.pen)[cohort=='all'], y[cohort=='all'])
     C     Dxy       n   Missing
 0.836   0.672    4554         0
```

This will be the final model except for the model used in Section 12. The model has LR $\chi^2 = 1772$. The output of effective.df shows that non-interaction terms have barely been penalized, and coefficients of interaction terms have been shrunken from 59 d.f. to effectively 22·6 d.f. Predictive discrimination was assessed by computing the Somers' D_{xy} rank correlation between $X\hat{\beta}$ and whether or not $Y = 0$, in the subset of records for which $Y = 0$ is what was being predicted. Here $D_{xy} = 0{\cdot}672$ and the ROC area is 0·836 (the unpenalized model had an apparent $D_{xy} = 0{\cdot}676$ for the training sample). To summarize in another way, the effectiveness of this model in screening

* See Hurvich and Tsai[54] for the definition of aic.c, the 'corrected AIC'. Regarding the Bayesian information criterion (bic) of Schwarz,[55] several simulations have shown that models selected by BIC had too much shrinkage and hence in validation samples predicted less well than ones selected using AIC

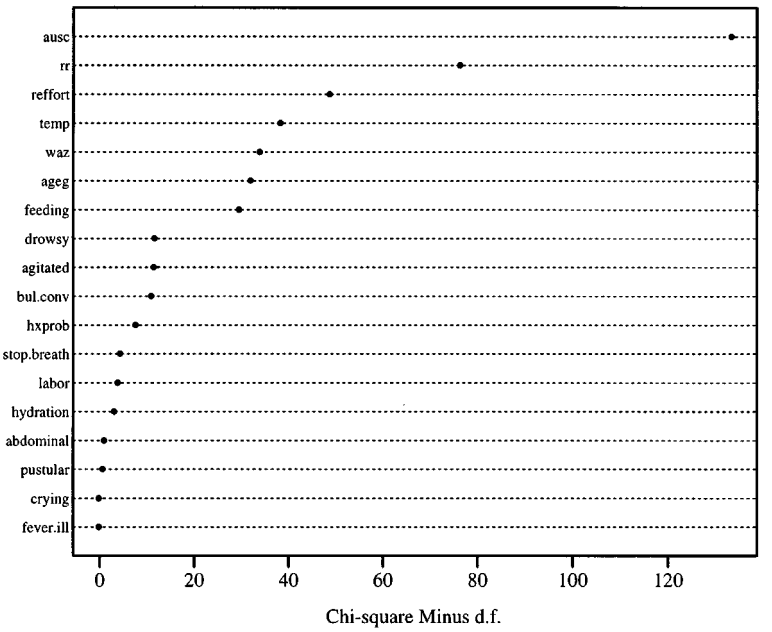

Figure 6. Importance of predictors in full penalized model, as judged by partial Wald χ^2 minus the predictor d.f. The Wald χ^2 values for each line in the dot plot include contributions from all higher-order effects. For example, the `cohort` effect includes all `cohort` interaction effects

infants for risks of any abnormality, the fraction of infants with predicted probabilities that $Y > 0$ being $< 0{\cdot}05$, $> 0{\cdot}25$, and $> 0{\cdot}5$ are, respectively, 0·10, 0·28 and 0·14. `anova` output is plotted in Figure 6 to give a snapshot of the importance of the various predictors. The Wald statistics used here are computed on a variance-covariance matrix which is adjusted for penalization (Gray Eq. $2{\cdot}6^{53}$).

The full equation for the fitted model is obtained using the S-plus statement `latex(full.pen)`, which typeset the output below using LaTeX.* Restricted cubic spline functions which were fit using `rcs` have been automatically written in more simple unrestricted form (Herndon and Harrell, Eq. 1^{41}). Only the part of the equation used for predicting $\Pr(Y = 0)$ is shown:

$$\Pr\{Y = 0\} = \frac{1}{1 + \exp(-X\beta)}$$

where

$$X\hat{\beta} = -5{\cdot}543 + 0{\cdot}1075\{\text{ageg} \in [7,60)\} + 0{\cdot}1971\{\text{ageg} \in [60,90]\}$$
$$+ 0{\cdot}1979\,\text{temp} + 0{\cdot}1092(\text{temp} - 36{\cdot}2)_+^3 - 2{\cdot}833(\text{temp} - 37)_+^3 + 5{\cdot}071(\text{temp} - 37{\cdot}3)_+^3$$

* The fitted equation could be written in the S-plus or SAS language using `Design`'s `Function` function had third-order interactions not been present

$$
\begin{aligned}
&- 2{\cdot}508(\text{temp} - 37{\cdot}7)_+^3 + 0{\cdot}1606(\text{temp} - 39)_+^3 \\
&+ 0{\cdot}02091\,\text{rr} - 6{\cdot}337\times10^{-5}(\text{rr} - 32)_+^3 + 8{\cdot}405\times10^{-5}(\text{rr} - 42)_+^3 + 6{\cdot}152\times10^{-5}(\text{rr} - 49)_+^3 \\
&- 0{\cdot}0001018(\text{rr} - 59)_+^3 + 1{\cdot}96\times10^{-5}(\text{rr} - 76)_+^3 \\
&- 0{\cdot}0759\,\text{waz} + 0{\cdot}02509(\text{waz} + 2{\cdot}9)_+^3 - 0{\cdot}1185(\text{waz} + 0{\cdot}75)_+^3 + 0{\cdot}1226(\text{waz} - 0{\cdot}28)_+^3 \\
&- 0{\cdot}02916(\text{waz} - 1{\cdot}73)_+^3 - 0{\cdot}4418\,\text{bul.conv} - 0{\cdot}08185\,\text{drowsy} - 0{\cdot}05327\,\text{agitated} \\
&- 0{\cdot}2304\,\text{reffort} - 1{\cdot}159\,\text{ausc} - 0{\cdot}16\,\text{feeding} - 0{\cdot}1609\,\text{abdominal} \\
&- 0{\cdot}0541\,\text{hydration} + 0{\cdot}08086\,\text{hxprob} + 0{\cdot}00752\,\text{pustular} + 0{\cdot}04712\,\text{crying} \\
&+ 0{\cdot}004299\,\text{fever.ill} - 0{\cdot}3519\,\text{stop.breath} + 0{\cdot}06864\,\text{labor} \\
&+ \{\text{ageg}\in[7{,}60)\}[6{\cdot}5\times10^{-5}\,\text{temp} - 0{\cdot}0028(\text{temp} - 36{\cdot}2)_+^3 - 0{\cdot}008691(\text{temp} - 37)_+^3 \\
&- 0{\cdot}004988(\text{temp} - 37{\cdot}3)_+^3 + 0{\cdot}02592(\text{temp} - 37{\cdot}7)_+^3 - 0{\cdot}009445(\text{temp} - 39)_+^3] \\
&+ \{\text{ageg}\in[60{,}90]\}[0{\cdot}000132\,\text{temp} - 0{\cdot}001826(\text{temp} - 36{\cdot}2)_+^3 - 0{\cdot}0164(\text{temp} - 37)_+^3 \\
&- 0{\cdot}0476(\text{temp} - 37{\cdot}3)_+^3 + 0{\cdot}09142(\text{temp} - 37{\cdot}7)_+^3 - 0{\cdot}02559(\text{temp} - 39)_+^3] \\
&+ \{\text{ageg}\in[7{,}60)\}[-0{\cdot}0009438\,\text{rr} - 1{\cdot}045\times10^{-6}(\text{rr} - 32)_+^3 - 1{\cdot}67\times10^{-6}(\text{rr} - 42)_+^3 \\
&- 5{\cdot}189\times10^{-6}[\text{rr} - 49)_+^3 + 1{\cdot}429\times10^{-5}(\text{rr} - 59)_+^3 - 6{\cdot}382\times10^{-6}(\text{rr} - 76)_+^3] \\
&+ \{\text{ageg}\in[60{,}90]\}[-0{\cdot}001921\,\text{rr} - 5{\cdot}521\times10^{-6}(\text{rr} - 32)_+^3 - 8{\cdot}628\times10^{-6}(\text{rr} - 42)_+^3 \\
&- 4{\cdot}147\times10^{-6}(\text{rr} - 49)_+^3 + 3{\cdot}813\times10^{-5}(\text{rr} - 59)_+^3 - 1{\cdot}984\times10^{-5}(\text{rr} - 76)_+^3]
\end{aligned}
$$

and $\{c\} = 1$ if subject is in group c, otherwise, $(x)_+ = x$ if $x > 0$, 0 otherwise.

To show the shapes of effects of the predictors we use the following code. For the continuous variables temp and rr which interact with age group, we show the effects for all three age groups, separately for each Y cut-off. All effects have been centred so that the log odds at the median predictor value is zero when cohort='all', so these plots actually show log odds relative to the reference values. The patterns in Figure 7 are in agreement with those in Figure 5:

```
par(mfrow=c(4,4))
yl←c(-2.5, 1)  # put all plots on common y-axis scale

# Plot predictors which interact with another predictor
# Vary ageg over all age groups, then vary temp over its
# default range (10th smallest to 10th largest values in data)
# Make a separate plot for each 'cohort'
# ref.zero centers effects using median x

for(co in levels(cohort)) {
  plot(full.pen, temp=NA, ageg=NA, cohort=co, ref.zero=T, ylim=yl, conf.int=F)
  text(37.5, 1.5, co)        # add title showing current cohort
}
```


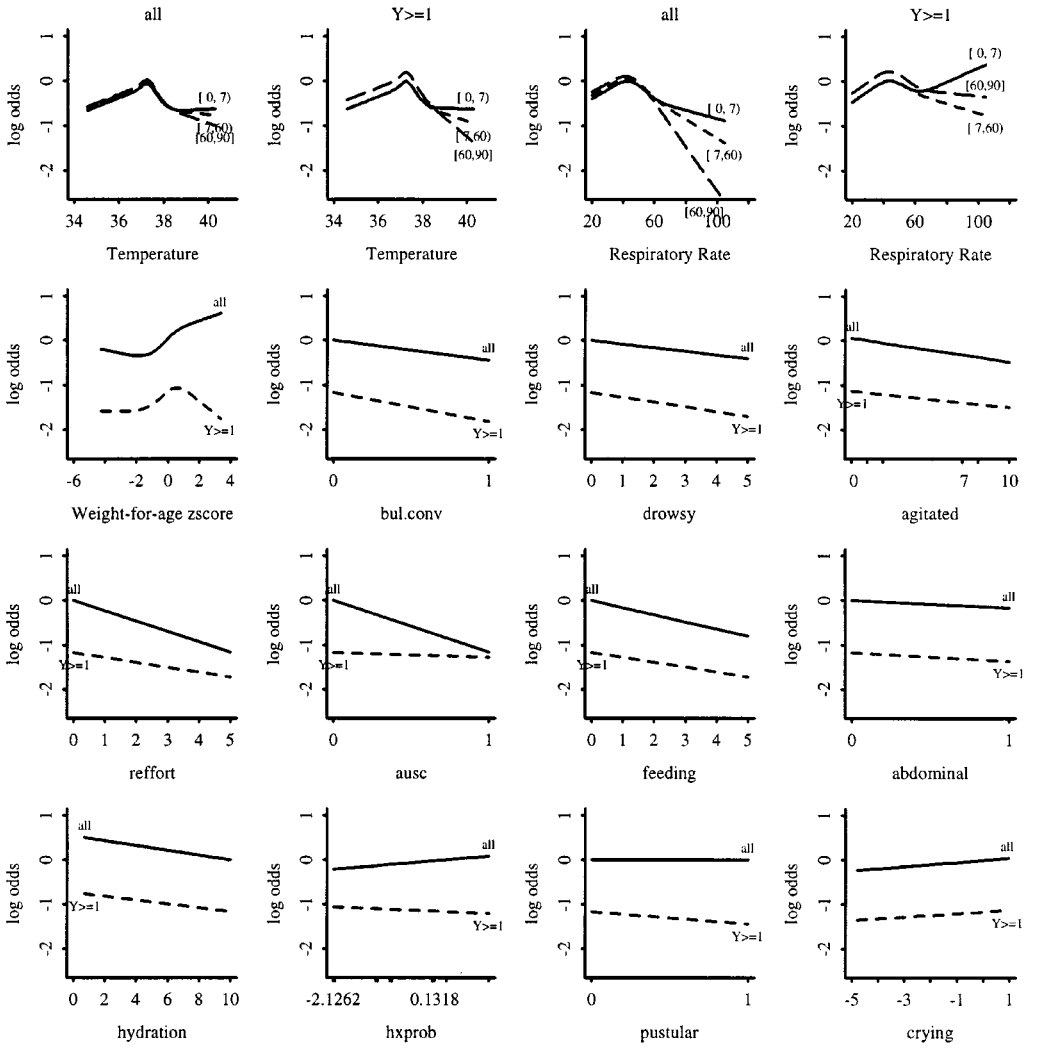

Figure 7. Centred effects of predictors on the log odds. The first four plots show interaction effects with the age intervals noted. For others, interaction with cohort are shown. For predictors having fewer than 10 unique values, *x*-axis tick marks appear only for values which occurred in the data. No plot was made for the fever.ill, stop.breath, or labor cluster scores. The title all refers to the prediction of $Y = 0 | Y \geqslant 0$, that is, $Y = 0$

```
for(co in levels(cohort)) {
  plot(full.pen, rr=NA, ageg=NA, cohort=co, ref.zero=T, ylim=yl, conf.int=F)
  text(70, 1.5, co)
}

# For each predictor which only interacts with cohort, show the
# Differing effects of the predictor for predicting Pr(Y=0) and
```



```
# Pr(Y=1 | Y > 0) on the same graph

plot(full.pen, waz      = NA, cohort=NA, ref.zero=T, ylim=yl, conf.int=F)
plot(full.pen, bul.conv = NA, cohort=NA, ref.zero=T, ylim=yl, conf.int=F)
. . . .
plot(full.pen, crying   = NA, cohort=NA, ref.zero=T, ylim=yl, conf.int=F)
```

12. USING APPROXIMATIONS TO SIMPLIFY THE MODEL

It is tempting to use *P*-values and stepwise methods to develop a parsimonious prediction model. Besides invalidating confidence limits and causing measures of predictive accuracy such as adjusted R^2 to be optimistic, there are many other reasons not to rely on stepwise techniques (see Harrell *et al.*[20] for citations). We follow Spiegelhalter's advice to use full model fits in conjunction with shrinkage.[56]

Parsimonious models can be developed, however, by approximating predictions from the model to any desired level of accuracy. Let $\hat{L} = X\beta$ denote the predicted log odds from the full penalized ordinal model, including multiple records for subjects with $Y > 0$. Then we can use a variety of techniques to approximate $\hat{L}$ from a subset of the predictors (in their raw form). With this approach one can immediately see what is lost over the full model by computing, for example, the mean absolute error in predicting $\hat{L}$. Another advantage to full model approximations is that shrinkage used in computing $\hat{L}$ is inherited by any model that predicts $\hat{L}$. In contrast, the usual stepwise methods result in $\hat{\beta}$ that are too large since the final coefficients are estimated as if the model structure was prespecified.*

Even though CART (classification and regression trees[58]) when used on X and Y often finds prediction rules that validate poorly because of the extremely large number of models searched,[27] CART can be very useful as an approximator for a complex model. For the current problem, CART would be particularly useful as it would result in a prediction tree that would be easy for health workers to use. Unfortunately, a 50-node CART was required to predict $\hat{L}$ with an $R^2 \geqslant 0{\cdot}9$, and the mean absolute error in the predicted logit was still 0·4. This will happen when the model contains many important continuous variables.

We chose to approximate the full model using its important components, by using a stepdown technique predicting $\hat{L}$ from all of the component variables using ordinary least squares. In using stepdown with the least squares function `ols` in `Design` there is a problem with infinite F statistics when the initial $R^2 = 1{\cdot}0$, so we will specify $\sigma = 1$ to `ols`. Because `cohort` interacts with the predictors, separate approximations can be developed for each level of Y. For this example we approximate the log odds that $Y = 0$ using the cohort of patients used for determining $Y = 0$, that is, $Y \geqslant 0$ or `cohort='all'`:

```
plogit ← predict(full.pen)

f ← ols(plogit ~ ageg*(rcs(temp,5) + rcs(rr,5)) + rcs(waz,4) +
        bul.conv + drowsy + agitated + reffort + ausc + feeding +
```

* The *lasso* method of Tibshirani[57] addresses this problem


```
        abdominal + hydration + hxprob + pustular + crying +
        fever.ill + stop.breath + labor, sigma=1,
        subset=cohort='all')

# Do fast backward stepdown
fastbw(f, aics=1e10)   # 1e10 causes all variables to eventually be
                       # deleted so can see most important ones in order

# Fit an approximation to the full penalized model using most
# important variables
full.approx ← ols(plogit ~ rcs(temp,5) + ageg*rcs(rr,5) + rcs(waz,4) +
                  bul.conv + drowsy + reffort + ausc + feeding,
                  subset=cohort=='all')
```

The approximate model had R^2 against the full penalized model of 0·972, and the mean absolute error in predicting $\hat{L}$ was 0·17. The D_{xy} rank correlation between the approximate model's predicted logit and the binary event $Y = 0$ is 0·665 as compared with the full model's $D_{xy} = 0{\cdot}672$.

Next, turn to diagramming this model approximation so that all predicted values can be computed without the use of a computer. We draw a type of nomogram which converts each effect in the model to a 0–100 scale which is just proportional to the log odds. These points are added across predictors to derive the 'total points', which are converted to $\hat{L}$ and then to predicted probabilities. For the interaction between rr and ageg, Design's nomogram function automatically constructs 3 rr axes – only one is added into the total point score for a given subject. Here we draw a nomogram for predicting the probability that $Y > 0$, which is $1 - \Pr(Y = 0)$. This probability is derived by negating $\hat{\beta}$ and $X\hat{\beta}$ in the model derived to predict $\Pr(Y = 0)$.

```
f ← full.approx
f$coefficients        ← f$coefficients
f$linear-predictors ← f$linear.predictors

nomogram(f,
         temp=32:41, rr=seq(20,120,by=10), waz=seq(-1.5,2,by=.5),
         fun=plogis, funlabel='Pr(Y > 0)',
         fun.at=c(.02,.05,seq(.1,.9,by=.1),.95,.98))
# plogis is S-PLUS's builtin 1/(1+exp(-x)) function
```

The nomogram is shown in Figure 8. As an example in using the nomogram, a 6-day old infant gets approximately 9 points for having a respiratory rate of 30/min, 19 points for having a temperature of 39°C, 11 points for waz=0, 14 points for drowsy=5, and 15 points for reffort=2. Assuming that bull.conv=ausc=feeding=0, that infant gets 68 total points. This corresponds to $X\hat{\beta} = -0{\cdot}6$ and a probability of 0·35. Values computed directly from the full.approx formula were $X\hat{\beta} = -0{\cdot}68$ and a probability of 0·34.*

* To see how this compares with prediction using the full model, the extra clinical signs in that model that are not in the approximate model were predicted individually on the basis of $X\hat{\beta}$ from the reduced model along with the signs that are in that model, using ordinary linear regression. The signs not specified when evaluating the approximate model were then set to predicted values based on the values given for the 6-day old infant above. The resulting $X\hat{\beta}$ for the full model is $-0{\cdot}81$ and the predicted probability is 0·31, as compared with $-0{\cdot}68$ and 0·34 quoted above

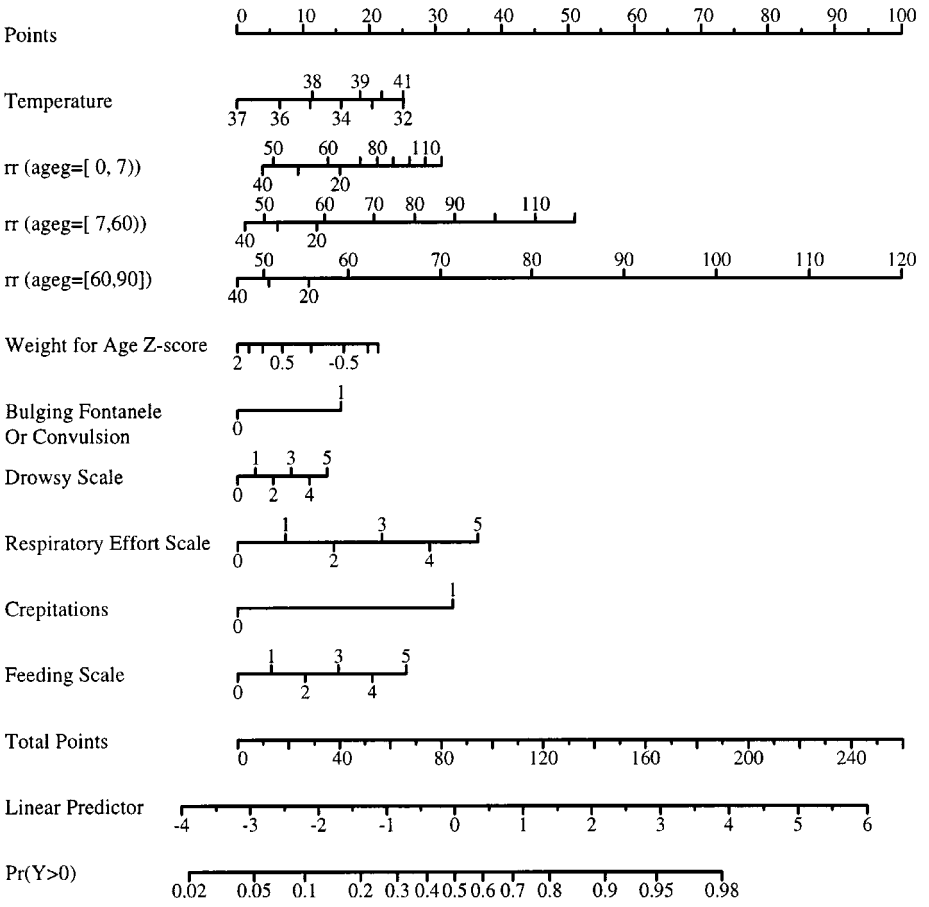

Figure 8. Nomogram for predicting $\Pr(Y > 0)$ from the penalized extended CR model, using an approximate model fitted using ordinary least squares ($R^2 = 0{\cdot}972$ against the full model's predicted logits)

For some applications a further simplification of the model is required. Assuming that maximum information is to be derived from the important continuous variables, one way to simplify the model is to dichotomize each clinical sign into a present/absent coding. To find the simplest approximation of the model that adequately discriminated between low- and high-risk infants, we again predicted the 'good standard' predicted log odds of outcomes, but used as candidate variables the 3-interval age variable, the vital signs, vital sign by age interactions, and all of the individual clinical signs and clinical history variables. To enable the procedure to automatically find the best cutpoints for multi-level signs, those signs were represented using a series of binary variables. For example, `hfa` (history of feeding) is a 0–4 variable, and the following candidate variables were used to represent `hfa`: mildly reduced or worse; severely reduced or unable to feed, and unable to feed. All variables were fitted in an ordinary multiple regression model, and variables were deleted in increasing order of explained variation. A model

which retained 7 individual signs resulting in $D_{xy} = 0{\cdot}664$ and R^2 against the optimal model's predicted logit of 0·954.

13. VALIDATING THE MODEL

Most analysts validate a fitted model using held-back data, but this method has severe drawbacks.[20] The bootstrap technique[59] allows the analyst to derive bias (overfitting) – corrected estimates of predictive accuracy without holding back valuable data during the model development phase. The steps required for using the bootstrap to bias-correct indexes such as D_{xy} and calibration error was summarized in Harrell *et al.*[20] For the full CR model which was fitted using PMLE, we used 150 bootstrap replications to estimate and then to correct for optimism in various statistical indexes: D_{xy}; generalized R^2;[60] intercept and slope of a linear recalibration equation for $X\hat{\beta}$ (related to Section 7 of van Houwelingen and le Cessie;[61] see also Phillips *et al.*[62]); the maximum calibration error for $\Pr(Y = 0)$ based on the linear-logistic recalibration (`Emax`), and the Brier quadratic probability score `B`.[63] PMLE is used at each of the 150 resamples. During the bootstrap simulations, we sample with replacement from the *patients* and not from the 5553 expanded *records*, hence the specification `cluster=u$subs`, where `u$subs` is the vector of sequential patient numbers computed from `cr.setup` above. To be able to measure the predictive accuracy of the predicted probability of a single event, the `subset` parameter is specified so that $\Pr(Y = 0)$ is being assessed even though 5553 observations are used to develop each of the 150 models. The output and the S-plus statement used to obtain the output are shown below:

```
validate(full.pen, B=150, cluster=u$subs, subset=cohort=='all')
```

	index.orig	training	test	optimism	index.corrected	n
Dxy	0.672	0.675	0.666	0.009	0.662	150
R2	0.376	0.383	0.370	0.013	0.363	150
Intercept	−0.031	−0.033	0.001	−0.034	0.003	150
Slope	1.029	1.031	1.002	0.029	1.000	150
Emax	0.000	0.000	0.001	0.001	0.001	150
B	0.120	0.119	0.121	−0.002	0.122	150

We see that for the apparent $D_{xy} = 0{\cdot}672$ the optimism from overfitting was estimated to be 0·009 for the PMLE model, so the bias-corrected estimate of predictive discrimination is 0·662. The intercept and slope needed to recalibrate $X\hat{\beta}$ to a 45° line are very near (0, 1). The estimate of the maximum calibration error in predicting $\Pr(Y = 0)$ is 0·001 which is quite satisfactory. The corrected Brier score is 0·122.

The simple calibration statistics just listed do not address the issue of whether predicted values from the model are miscalibrated in a non-linear way. Steps for estimating bias-corrected calibration curves for survival time models and for non-parametrically estimating a smooth calibration curve for a binary logistic model on a separate validation sample were given previously.[20] Putting these two techniques together we arrive at the following plan for estimating a calibration curve using the bootstrap, with the only assumption being the smoothness of the curve. Choose a single binary event for which to check the calibration of the estimated

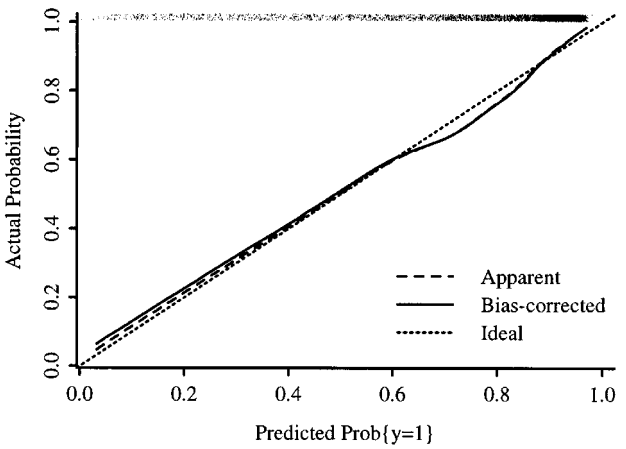

Figure 9. Bootstrap calibration curve for the full penalized extended CR model. 150 bootstrap repetitions were used in conjunction with the lowess smoother.[49] Also shown is a 'rug plot' to demonstrate how effetive this model is in discriminating patients into low and high risk groups for $\Pr(Y = 0)$ (which corresponds with the derived variable value $y = 1$ when cohort = 'all')

probabilities. The actual occurrences of binary responses are smoothed using lowess (with the 'no iteration' option) to estimate probabilities. Then choose a grid of predicted values, for example, 0·01, 0·03, 0·05, ... ,0·99. Fit lowess to the predicted probabilities derived from the final model and the actual binary outcomes from the final model and the actual binary outcomes from the original sample. Then evaluate the smoothed estimates at the grid. Differences between the lowess estimates and the 45° line are the estimates of apparent calibration accuracy. Then for each bootstrap resample, the ordinal model is fitted using PMLE from a sample with replacement from the *patients*, and the coefficients from this model are used to predict probabilities for the original sample. The discrepancies from the 45° line are compared with the discrepancies present when the bootstrap model was evaluated on the bootstrap sample. The difference in the discrepancies is the estimate of optimism. After averaging over 150 replications, separately for each probability level in the uniform grid, the estimates of optimism in the original, apparent, calibration errors are added to those errors. Then the bootstrap-corrected calibration curve is plotted.

All these steps are done using the following Design functions:

```
cal ← calibrate(full.pen, B=150, cluster=u$subs, subset=cohort=='all')
plot(cal)
```

The results are shown in Figure 9. One can see a slightly non-linear calibration function estimate, but the overfitting-corrected calibration is excellent everywhere, being only slightly worse than the apparent calibration. The estimated maximum calibration error is 0·043. The excellent validation for both predictive discrimination and calibration are a result of the large sample size, frequency distribution of Y, initial data reduction and PMLE.

14. SUMMARY

Clinically-guided variable clustering and item weighting, done with very limited use of the outcome variable, resulted in a great reduction in the number of candidate predictor degrees of freedom and hence increased the true predictive accuracy of the model. Sources summarizing clusters of clinical signs, along with the temperature, respiratory rate, and weight-for-age after suitable non-linear transformation and allowance for interactions with age, are powerful predictors of the ordinal response. Graphical methods are effective for detecting lack of fit in the PO and CR models and for diagramming the final model. Model approximation is a better approach than stepwise methods (that use Y) to develop parsimonious clinical prediction tools. Approximate models inherit the shrinkage from the full model. For the ordinal model developed here, substantial shrinkage of the full model was needed.

The bootstrap, as in a wide variety of other situations, is an effective tool for validating an ordinal logistic model with respect to discrimination and calibration without having the need to hold back data during model development. The final CR ordinal logistic model accurately predicted severity of diagnosis/outcome (as summarized by several disparate outcome variables) in infants screened for pneumonia, sepsis, and meningitis in developing countries. There was nothihg about the continuation ratio model that made it fit the data set better than other ordinal models (which we have found to be the case in one other large data set), and in fact there is some evidence that the equal-slopes CR model fits the data more poorly than the equal-slopes PO model. The real benefit of the CR model is that using standard binary logistic model software one can flexibly specify how the equal-slopes assumption can be relaxed.

Faraway[64] has demonstrated how all data-driven steps of the modelling process increase the real variance in 'final' parameter estimates, when one estimates variances without assuming that the final model was prespecified. For ordinal regression modelling, the most important modelling steps are (i) choice of predictor variables; (ii) selecting or modelling predictor transformations; and (iii) allowance for unequal slopes across Y-cut-offs (that is, non-PO or non-CR). Regarding steps (ii) and (iii) one is tempted to rely on graphical methods such as residual plots to make detours in the strategy, but it is very difficult to estimate variances or to properly penalize assessments of predictive accuracy for subjective modelling decisions. Regarding (i), shrinkage has been proven to work better than stepwise variable selection when one is attempting to build a main-effects model.[56] Choosing a shrinkage factor is a well-defined, smooth, and often a unique process as opposed to binary decisions on whether variables are 'in' or 'out' of the model. Likewise, instead of using arbitrary subjective (residual plots) or objective (χ^2 due to `cohort` $\times$ covariable interactions, that is, non-constant covariable effects) assessments, shrinkage can systematically allow model enhancements in so far as the information content in the data will support, through the use of differential penalization. Shrinkage is a solution to the dilemma faced when the analyst attempts to choose between a parsimonious model and a more complex one that fits the data. Penalization does not require the analyst to make a binary decision, and it is a process that can be validated using the bootstrap.

APPENDIX

WHO/ARI Young Infant Multicentre Study Group*

Study sites	Addis Ababa, Ethiopia	Fajara, Gambia	Goroka, Papua New Guinea	Manila, Philippines
Principal investigator	Lulu Muhe	Kim Mulholland	Deborah Lehmann	Salvacion Gatchalian
Co-investigators, Study coordinators		Olayinka Ogunlesi Martin Weber	Gerard Saleu	Beatriz Quiambao
Other investigators, clinicians	Meaza Tilahun Sileshi Lulseged Senait Kebede	Mark Manary Ayo Palmer	Alphonse Rongap Mexy Kakazo Pioto Namuigi Sebeya Lupiwa Rebecca sehuko	Ana Marie Moreles Leticia Abraham
Bacteriologists	Afeworti Yohanes† Bahrie Belete Signe Ringertz	Richard Adegbola Osman Secka	Alison Clegg Audrey Michael Tony Lupiwa Matthew Omena Mark Mens	Lydia Sombrero Ma. Victoria Abraham
Radiologists	Tsegaye Desta			
Data management	Kidanemariam Wlyesus	Joseph Bangali	Don Lewis	Elinor S. Sunico Teresita C. Cedulla
Institution/ hospitals	Department of Paediatrics and Child Heath, Ethio-Swedish Children's Hospital, Addis Ababa University	Medical Research Council Hospital and Royal Victoria Hospital	Papua New Guinea Institute of Medical Research Goroka Base Hospital	Research Institute of Tropical Tropical Medicine, Phillippines General Hospital, Quezon City General Hospital
Institution Directors	Nebiat Tafari	Brian Greenwood	Michael P. Alpers	

* Not inclusive
† Deceased

Study Co-ordination

Scientific co-ordinator: Dr Sandy Gove, WHO/ARI
Data management: Dr Peter Byass, University of Nottingham Medical School
Data analysis: Dr Frank Harrell, University of Virginia, Charlottesville Mrs Karen Mason, WHO/ARI
Management, logistic, supplies: Mrs Frances McCaul, WHO/ARI
Mrs Sue Parker, WHO/ARI
Study advisors: Dr Claire Broome, Centers for Disease Control (CDC), Atlanta
Dr H. F. Eichenwald, University of Texas Southwestern Medical Center, Dallas
Mr Mike Gratten, Queensland Institute of Medical Research, Brisbane
Dr P. Margolis, University of North Carolina at Chapel Hill
Dr R. Facklam, CDC, Atlanta
Radiology working group: Dr H. Tschappeler, Universitat Bern
Dr A. Lamont, The Leicester Royal Infirmary
Dr G. M. A. Hendry, Royal Hospital for Sick Children, Edinburgh
Professor Philip E. S. Palmer, University of California, Davis

ACKNOWLEDGEMENTS

This work was supported by The World Health Organization ARI Programme, and for F. Harrell, Research Grants HS-06830 and HS-07137 from the Agency for Health Care Policy and Research, Rockville, Maryland, U.S.A., and grants from the Robert Wood Johnson Foundation,

Princeton, NJ, U.S.A. F. Harrell wishes to dedicate his work on this project to the memory of his dear colleague L. Richard Smith whose critical reading of this paper resulted in significant improvements.

REFERENCES

1. Walker, S. H. and Duncan, D. B. 'Estimation of the probability of an event as a function of several independent variables', *Biometrika*, **54**, 167–178 (1967).
2. Fienberg, S. E. *The Analysis of Cross-Classified Data*, 2nd edn, MIT Press, Cambridge, MA, 1980.
3. Agresti, A. 'A survey of models for repeated ordered categorical response data', *Statistics in Medicine*, **8**, 1209–1224 (1989).
4. Anderson, J. A. and Philips, P. R. 'Regression, discrimination and measurement models for ordered categorical variables', *Applied Statistics*, **30**, 22–31 (1981).
5. Anderson, J. A. 'Regression and ordered categorical variables', *Journal of the Royal Statistical Society*, Series B, **46**, 1–30 (1984).
6. Armstrong, B. G. and Sloan, M. 'Ordinal regression models for epidemiologic data', *American Journal of Epidemiology*, **129**, 191–204 (1989).
7. Ashby, D., West, C. R. and Ames, D. 'The ordered logistic regression model in psychiatry: Rising prevalence of dementia in old people's homes', *Statistics in Medicine*, **8**, 1317–1326 (1989).
8. Berridge, D. M. and Whitehead, J. 'Analysis of failure time data with ordinal categories of response', *Statistics in Medicine*, **10**, 1703–1710 (1991).
9. Brazer, S. R., Pancotto, F. S., Long III, T. T., Harrell, F. E., Lee, K. L., Tyor, M. P. and Pryor, D. B. 'Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia', *Journal of Clinical Epidemiology*, **44**, 1263–1270 (1991).
10. Cox, C. 'Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach', *Statistics in Medicine*, **14**, 1191–1203 (1995).
11. Greenland, S. 'Alternative models for ordinal logistic regression', *Statistics in Medicine*, **13**, 1665–1677 (1994).
12. Hastie, T. J., Botha, J. L. and Schnitzler, C. M. 'Regression with an ordered categorical response', *Statistics in Medicine*, **8**, 785–794 (1989).
13. Koch, G. G., Amara, I. A. and Singer, J. M. 'A two-stage procedure for the analysis of ordinal categorical data', in Sen, P. K. (ed.), *BIOSTATISTICS*: *Statistics in Biomedical*, *Public Health and Environmental Sciences*, Elsevier Science Publishers B.V., North-Holland, 1985.
14. McCullagh, P. 'Regression models for ordinal data', *Journal of the Royal Statistical Society*, *Series B*, **42**, 109–142 (1980).
15. Peterson, B. and Harrell, F. E. 'Partial proportional odds models for ordinal response variables', *Applied Statistics*, **39**, 205–217 (1990).
16. Whitehead, J. 'Simple size calculations for ordered categorical data', *Statistics in Medicine*, **12**, 2257–2271 (1993).
17. Cole, T. J., Morley, C. J., Thornton, A. J., Fowler, M. A. and Hewson, P. H. 'A scoring system to quantify illness in babies under 6 months of age', *Journal of Royal Statistical Society*, *Series A*, **154**, 287–304 (1991).
18. Yee, T. W. and Wild, C. J. 'Vector generalized additive models', *Journal of the Royal Statistical Society*, *Series B*, **58**, 481–493 (1996).
19. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
20. Harrell, F. E., Lee, K. L. and Mark, D. B. 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, **15**, 361–387 (1996).
21. MathSoft. *S-Plus User's Manual*, *Version 2.3*, MathSoft, Inc., Seattle WA, 1995.
22. Harrell, F. E. 'Design: S-Plus functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. UNIX and Microsoft Windows versions available from http://www.med.virginia.edu/medicine/clinical/hes/biostat.htm' 1997.

23. WHO/ARI Study Group on the Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants, 'Study methods', in preparation (1997).
24. Zhou, X. 'Effect of verification bias on positive and negative predictive values', *Statistics in Medicine*, **13**, 1737–1745 (1994).
25. WHO/ARI Study Group on the Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants, 'Ordinal outcome scale', in preparation (1997).
26. Follmann, D. 'Multivariate tests for multiple endpoints in clinical trials', *Statistics in Medicine*, **14**, 1163–1175 (1995).
27. Harrell, F. E., Lee, K. L., Matchar, D. B. and Reichert, T. A. 'Regression models for prognostic prediction: Advantages, problems, and suggested solutions', *Cancer Treatment Reports*, **69**, 1071–1077 (1985).
28. Marshall, G., Grover, F. L., Henderson, W. G. and Hammermeister, K. E. 'Assessment of predictive models for binary outcomes: An empirical approach using operative death from cardiac surgery', *Statistics in Medicine*, **13**, 1501–1511 (1994).
29. D'Agostino, R. B., Belanger, A. J., Markson, E. W., Kelly-Hayes, M. and Wolf, P. A. 'Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model', *Statistics in Medicine*, **14**, 1757–1770 (1995).
30. Cureton, E. E. and D'Agostino, R. B. *Factor Analysis, An Applied Approach*, Erlbaum Publishers, New Jersey, 1983.
31. Sarle, W. S. 'The VARCLUS procedure', in *SAS/STAT User's Guide*, vol. 2, 4th edn, SAS Institute, Inc., Cary NC, 1990, Chapter 43, pp. 1641–1659.
32. Hoeffding, W. 'A non-parametric test of independence', *Annals of Mathematical Statistics*, **19**, 546–557 (1948).
33. Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York, 1994.
34. Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, New York, 1986.
35. Jackson, J. E. *A User's Guide to Principal Components*, Wiley, New York, 1991.
36. Kuhfeld, W. F., 'The PRINQUAL procedure', in *SAS/STAT User's Guide*, vol. 2, 4th edn, SAS Institute, Inc., Cary NC, 1990, Chapter 34, pp. 1265–1323.
37. Atkinson, A. C. 'A note on the generalized information criterion for choice of a model', *Biometrika*, **67**, 413–418 (1980).
38. Stone, C. J. and Koo, C. Y. 'Additive splines in statistics', *Proceedings of the Statistical Computing Section ASA*, 45–48 (1995).
39. Devlin, T. F. and Weeks, B. J. 'Spline functions for logistic regression modeling', in *Proceedings of the Eleventh Annual SAS Users Group International Conference*, SAS Institute, Inc., Cary NC, 1986, pp. 646–651.
40. Harrell, F. E., Lee, K. L. and Pollock, B. G. 'Regression models in clinical studies: Determining relationships between predictors and response', *Journal of the National Cancer Institute*, **80**, 1198–1202 (1988).
41. Herndon, J. E. and Harrell, F. E. 'The restricted cubic spline hazard model', *Communications in Statistics – Theory and Methods*, **19**, 639–663 (1990).
42. Lamport, L. *LaTeX: A Document Preparation System*, 2nd edn, Addison-Wesley, Reading, MA, 1994.
43. SAS Institute, Inc. *SAS/STAT User's Guide*, vol. 2, 4th edn, SAS Institute, Inc., Cary NC, 1990.
44. Schoenfeld, D. 'Partial residuals for the proportional hazards regression model', *Biometrika*, **69**, 239–241 (1982).
45. Grambsch, P. and Therneau, T. 'Proportional hazards tests and diagnostics based on weighted residuals', *Biometrika*, **81**, 515–526 (1994). Amendment and corrections in **82**, 668 (1995).
46. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society Series B*, **34**, 187–220 (1972).
47. Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. 'Graphical methods for assessing logistic regression models (with discussion)', *Journal of the American Statistical Association*, **79**, 61–83 (1984).
48. Collett, D. *Modelling Binary Data*, Chapman and Hall, London 1991.
49. Cleveland, W. S. 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**, 829–836 (1979).

50. Altman, D. G. and Andersen, P. K. 'Bootstrap investigation of the stability of a Cox regression model', *Statistics in Medicine*, **8**, 771–783 (1989).
51. le Cessie, S. and van Houwelingen, J. C. 'Ridge estimators in logistic regression', *Applied Statistics*, **41**, 191–201 (1992).
52. Verweij, P. and van Houiwelingen, H. C. 'Penalized likelihood in Cox regression', *Statistics in Medicine*, **13**, 2427–2436 (1994).
53. Gray, R. J. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
54. Hurvich, C. M. and Tsai, C. 'Regression and time series model selection in small samples', *Biometrika*, **76**, 297–307 (1989).
55. Schwarz, G. 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464 (1978).
56. Spiegelhalter, D. J. 'Probabilistic prediction in patient management', *Statistics in Medicine*, **5**, 421–433 (1986).
57. Tibshirani, R. 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B*, **58**, 267–288 (1996).
58. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., *Classification and Regression Trees*, Wadsworth and Brooks/Cole Pacific Grove, CA, 1984.
59. Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
60. Nagelkerke, N. J. D. 'A note on a general definition of the coefficient of determination', *Biometrika*, **78**, 691–692 (1991).
61. van Houwelingen, J. C. and le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **8**, 1303–1325 (1990).
62. Phillips, A. N., Thompson, S. G. and Pocock, S. J. 'Prognostic scores for detecting a high risk group: Estimating the sensitivity when applied to new data', *Statistics in Medicine*, **9**, 1189–1198 (1990).
63. Brier, G. W. 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, **75**, 1–3 (1950).
64. Faraway, J. J. 'The cost of data analysis', *Journal of Computational and Graphical Statistics*, **1**, 213–229 (1992).

