Journal of Medical Genetics 1988, **25**, 118–121

A combinatorial method for grouping cases with multiple malformations

R M WINTER*, R D CLARK†, K ASHLEY‡, AND G GIBBS§

*From *the Division of Inherited Metabolic Disease, Clinical Research Centre, Northwick Park Hospital, and Kennedy Galton Centre, Harper Lane, Radlett, Herts WD7 9HQ; †the Institute of Child Health, 30 Guilford Street, London WC1N 1EH; ‡the University of London Computer Centre, Guilford Street, London WC1N 1EH; and §the Clinical Research Centre, Watford Road, Harrow, Middlesex HA1 3UJ.*

SUMMARY A combinatorial method is described for grouping cases with multiple malformations for the purpose of identifying previously undescribed syndromes. This method includes ways of carrying out 'tight' or 'loose' grouping, of allowing for variability of reporting of physical features by different observers, and of minimising the number of 'spurious' groups. Evaluation using a test data set of known dysmorphic syndromes showed that the method provides a feasible and useful means of grouping undiagnosed cases.

Over 50% of children with multiple malformations remain undiagnosed, even after many consultations with experts and recourse to computer databases.[1] Many of these children have undescribed syndromes of varying aetiologies but, because these syndromes are rare, similarly affected subjects may be widely separated in time and space, making matching of cases dependent on chance in many instances. One answer is a central database where the features of cases from many centres can be stored and compared on a regular basis.[2 3] However, such a repository is likely to accumulate thousands of cases and the matching process becomes a considerable analytical problem. Although much has been written about the use of numerical taxonomy in the classification of birth defects,[4] most workers have concentrated their efforts on analysing relatively small numbers of patients with restricted diagnoses and a limited set of characteristics. A multiple malformation register may involve thousands of patients with over a thousand possible characteristics and virtually an infinite number of possible similarity groups, depending on the degree of similarity required. Although registers exist, very little consideration has been given to the optimal techniques for grouping similar cases. This paper describes the methods developed for grouping patients in the London Dysmorphology Database,[2] in the hope that other workers will be stimulated to develop and publish improved methods or refinements.

Received for publication 25 June 1986.
Revised version accepted for publication 5 January 1987.

General considerations

It is appropriate to set out the general requirements of a system for grouping large numbers of undiagnosed cases.

(1) The recognition that several patients with a particular pattern of malformations exhibit a 'new' syndrome is highly subjective, at least in the initial stages. It is unlikely that statistical criteria could be devised to 'prove' that a particular group of patients must have a previously unrecognised 'new' syndrome. Therefore, the aim of any procedure for matching cases should be the identification of smaller subgroups of patients who share many features. Once grouped, cases can be further evaluated by analysis of photographs, clinical examination, and other subjective assessments.

(2) Any matching procedure must create a manageable number of small groups of patients with similar features from a large number of possible candidates. A balance must be struck between producing too many groups with patients having too few features in common, and discarding possible 'correct' matches because the grouping criteria are too strict.

(3) The grouping methods should be flexible enough to allow for different degrees of stringency in the definition of abnormal features, so that 'tight' or 'loose' searches can be carried out.

(4) 'Spurious' matches, based on features that have little clinical significance, should be kept to a minimum.

Methods

Three key steps in the analysis are the initial coding of patient features, the specification of equivalent features, and the grouping process itself.

CODING OF PATIENT FEATURES

The features of each case are coded using a master list of physical abnormalities, as described elsewhere.[2] Each abnormality is given a three level code. The first level represents some general region of the body (for example, head 03·00·00), the second a particular subdivision of that region (for example, scalp 03·06·00), and the third a specific abnormality (for example, scalp defect, 03·06·02). This system can be used to carry out 'loose' versus 'tight' matches by using level 1 or 2 codes instead of level 3 codes in some instances.

THE SPECIFICATION OF EQUIVALENT FEATURES

There are four reasons why specification of equivalent features may be indicated. First, because it is impossible to define unambiguously some clinical features, observers may classify a feature differently in the same case. This happens particularly where a feature cannot be measured objectively. For example, 'mid-face hypoplasia' may be classified as a 'flat face' by some observers or as a 'hypoplastic maxilla' by others. For the purpose of matching cases, it would be desirable to join all these features into an 'equivalence class', to compensate for observer bias. Cases with any of these 'equivalent' features would tend to be grouped together, because the features would be synonymous for the purpose of matching cases. Second, equivalence classes may comprise malformation sequences and field defects,[5] for example, anencephaly, posterior encephalocele, meningomyelocele, and spina bifida occulta could be an equivalence class. Third, equivalence classes may be used to overcome the lack of accurate information about particular defects. For example, 'abnormalities of the heart' (a level 2 code) could be used as an equivalence class, if many cases had unspecified heart defects, without more detailed investigations. Finally, equivalence classes can also be used to match cases from centres using different coding systems for clinical features. In this event the equivalence classes would function as translation tables allowing different sets of data to be combined, even if the feature codes used for each data set were not strictly comparable.

GROUPING METHODS

The general method

The computer programme can identify groups of cases with any specified minimum number of features in common. Specifying a minimum that is too low will give too many groups, making subsequent analysis impossible. Specifying a minimum that is too high will make finding a match less likely.

For any particular case, the number of matches will depend on the minimum number of common features specified and the total number of features of the case. Cases with a large number of features will generate many spurious matches if too low a minimum is specified. Conversely, cases with only a few features may not be matched at all if too high a minimum is specified. For example, if two cases have only three significant abnormal features (for example, syndactyly, anal atresia, and iris coloboma), it would be very important to retrieve them as a group. However, specifying a minimum number of three common features may produce too many groups if applied to the whole data set. For this reason, the programme can analyse only those cases with a specified maximum number of features. This reduces both the number of cases in the analysis and the number of groups.

Reducing the number of 'spurious' groups

Some clinical features are important for the description of individual cases, but are so frequently found in the data set that they cause many spurious groups. In other words, their usefulness in discriminating between cases is low. Such features could be ignored for the purpose of grouping to minimise the number of non-specific groups. Some of these features are listed in table 1, with their frequency of use in the database of 923 published syndromes.

Assessment of similarity between cases

For each case in a group, the number of matching

TABLE 1 *Frequently occurring clinical features that could be ignored for the purpose of grouping.*

Clinical features	*No of syndromes*
Anteverted nares	58
Hypotonia	73
Epicanthic folds	74
Clinodactyly	76
High palate	82
Depressed/flat nasal bridge	84
Cryptorchid testes	86
Short stature, prenatal onset	91
Prominent forehead/frontal bossing	93
Low set ears	103
Strabismus	105
Seizures/abnormal EEG	114
Hypertelorism	138
Small mandible/micrognathia	163
Short stature, proportionate	267
Mental retardation	442

The numbers refer to the number of times each clinical feature has been used in a database of 923 published syndromes.

clinical features is expressed as a percentage of the total number of features for that case. This so-called 'goodness-of-fit' allows an average for each group to be calculated and for the groups to be sorted by average goodness-of-fit.

Thus, even using a simple combinatorial method, several different grouping strategies are possible.

Evaluation

TEST DATA

To evaluate the grouping programme, test data which consisted of coded features for 923 known malformation syndromes were used.[2] The features were coded using a master list of 1214 codes. Each syndrome had an average of 10 features (range three to 28) and 154 equivalence classes were created with an average of five features per class. The clinical features in table 1 could be ignored for the purpose of grouping.

TEST MATCHES

The numbers of groups of 'similar' syndromes generated in specific runs are shown in table 2. Simple code level 3 matching without equivalence classes was used. Initially, no features were ignored. There were more than 10 000 groups of syndromes with three or more features in common. This unmanageable number of groups cannot be evaluated by the comparison of individual cases. When only cases with a maximum of 10 features were analysed, the more manageable number of 393 groups was generated. However, 392 test cases were excluded from the analysis because they each had more than 10 features. With an unlimited number of features per case, a manageable number of groups was generated when five or six features were the minimum for grouping, giving 815 and 286 groups respectively. When the equivalence classes were used, grouping on five features produced 1125 groups and on six features 509 groups. In the latter case, ignoring the features listed in table 1 resulted in 310 groups and 173 groups when matching on five or six features respectively. Of the 310 groups generated with five features or more in common, 62% were deemed to be 'sensible' (for example, all cases in a group were acrocephalosyndactylies or mucopolysaccharidoses). The numbers of 'sensible' groups by average 'goodness-of-fit' are summarised in table 3.

TABLE 2 *Total number of groups of cases generated using different minimum number of features.*

Minimum common features	*Maximum features per case*	
	No limit	*10*
3	>10 000*	393
4	2436	82
5	815	12
6	286	0
7	110	
8	48	
9	24	
10	10	

*Computer run was terminated at this point.

Conclusions

Operation of the simple combinatorial programme described, using test data, has shown that this method provides a feasible way of grouping undiagnosed cases with multiple malformations. The equivalence class is a powerful and flexible device which compensates for the variability between different observers and phenotypic variation within the same syndromes.

Although it is difficult to quantify the usefulness of this method, the matches obtained with test data consisting of known syndromes did produce 'sensible' matches when equivalence classes were used and frequently occurring features ignored (table 3). It is hoped that this method will prove equally useful with undiagnosed cases. One unknown factor is the minimum number of undiagnosed cases needed in a database to have a reasonable chance of recognising 'new' entities. Our experience with a database of 500 undiagnosed cases suggests that this number is perhaps too small; the authors would welcome submission of further cases.

TABLE 3 *Number of groups, each comprising cases with five or more clinical features in common, by average 'goodness-of-fit'; 154 equivalence classes were specified and the features in table 1 were ignored.*

Average 'goodness-of-fit' of group	*No of groups*	*No of 'sensible' groups*	*Percentage of 'sensible' groups*
0–20%	0	0	0
21–40%	96	32	33
41–60%	158	106	67
61–80%	46	44	96
81–100%	10	10	100
Total	310	192	62

References

[1] Feingold M. Diagnosis of craniofacial anomalies: assessment of diagnostic data. *Birth Defects* 1980;**XVI**(5):75–82.

[2] Winter RM, Baraitser M, Douglas J. A computerised data base for the diagnosis of rare dysmorphic syndromes. *J Med Genet* 1984;**21**:121–3.

[3] Buyse M. Center for birth defects information services. *Birth Defects* 1980;**XVI**(5):83–92.

[4] Preus M. Diagnosis of craniofacial anomalies: the numerical versus intuitive approach to syndrome nosology. *Birth Defects* 1980;**XVI**(5):93–104.

[5] Spranger J, Benirshke K, Hall JG, *et al.* Errors of morphogenesis: concepts and terms. *J Pediatr* 1982;**100**:160–5.

Correspondence and requests for reprints to Dr R M Winter, Division of Inherited Metabolic Disease, Clinical Research Centre, Northwick Park Hospital, Watford Road, Harrow, Middlesex HA1 3UJ.