When studying antimicrobial resistance it is clear that individuals do not exist in isolation and are often clustered into groups. Data within groups are generally not independent, but standard statistical approaches assume independence of observations. When data are clustered (e.g. students in schools, patients in general practices, etc.) multilevel analysis can be used. The overall idea of multilevel analysis is that the clustering is taken into account in the analysis and provides additional information on the interactions between individuals and groups. The lowest level is often the individual and additional levels are formed by clustering in groups (the higher levels). This article introduces the principles behind multilevel modelling. The approach is to provide readers with sufficient information to understand outcomes in which this statistical technique is used, without expecting the reader to be able to perform such an analysis. As multilevel modelling can be seen as an extension of linear regression analysis, this is the starting point of the article. Other concepts and terms are introduced throughout, resulting in the explanation of the accompanying article on antimicrobial prescribing and resistance in Irish general practice (Vellinga A, Tansey S, Hanahoe B et al. J Antimicrob Chemother 2012; 67: 2523–30).

Keywords: applications, statistical methods, hierarchical models, clustered data

Introduction

Antimicrobial agents are unique therapeutics because they treat more than just the individual; antimicrobials also affect the microbial population of the host (including the pathogen population) and thereby society.1 Whereas the clinician focuses primarily on the most effective cure for a specific individual, an epidemiologist focuses on the effect of cumulative prescribing on resistance at the population level.2 This can cause conflict, as optimizing immediate treatment success for the individual could have less favourable effects at a population level.3 Although this is widely understood at a qualitative level, a more quantitative approach is needed to weigh the risks and benefits of antimicrobial use.

Individuals have many traits that could explain a given outcome, but to find out which traits are most likely to explain the outcome we have to compare groups. Grouping many individuals with and without the outcome will show which individual traits are more often present in those with the outcome compared with individuals without the outcome. However, when studying antimicrobial resistance (AMR), or infectious disease in general, it is clear that individuals do not exist in isolation and outcomes in different subjects are not independent.5 The dependency between data and the additional information provided by differences between individuals is ignored when data are aggregated. An example of aggregated data is when antimicrobial consumption and surveillance data on AMR are aggregated at the country level and the correlation between prescribing and resistance is compared between countries.

When there is a structure in the data and a hierarchy can be assigned (e.g. patients from different practices, practices within different countries), the structure itself often contains information that can be of value in understanding associations. For instance, the chance that an individual in that area will have an infection with an antimicrobial-resistant organism is influenced by the overall AMR level within that area.

Multilevel modelling is a statistical method in which both individual- and group-level information is included to understand their separate and combined influence on the outcome.5 Its application to studies of AMR can help provide a more quantitative perspective on the impact of antimicrobial consumption on AMR.

The present article introduces the principles behind multilevel modelling, but is written as a stand-alone introduction to this method and accompanies our article ‘Trimethoprim and ciprofloxacin resistance and prescribing in urinary tract infection associated with Escherichia coli: a multilevel model’.6 The online courses, textbooks7 and software (MLWin)8,9 of the Centre for Multilevel Modelling (CMM) at Bristol University10 were helpful in...
developing the present article, as were the articles on multilevel modelling by Merlo et al.11–13

**Linear regression analysis**

Multilevel modelling is based on regression analysis, a method in which information on a dependent or outcome variable $y$ is explained by one or more independent variables $x_1, x_2, x_3, \ldots$. Regression analysis describes $y$ as a function of the $x$s, or in other words, aims to predict $y$ from the $x$s.14,15 The simplest expression of a regression analysis is presented in Figure 1 in which the variable $y$ is plotted against the variable $x$ for each individual (each dot is an individual observation) and a line with slope $b$ can be drawn. The line intersects the $x$-axis at the intercept $a$. This regression line for $y$ on $x$ is expressed in the linear equation $y = a + bx$. Once a regression line is calculated, it can be used to predict an outcome $y$ by replacing $x$ with the actual value.

In reality, not every point lies neatly on the diagonal line. The individual data points are more scattered and form a cloud (Figure 2). To determine a line that best fits the pattern, the least square method is usually applied. This method starts with an overall line and calculates the distance from the line to each point. The distance from the line to the data point is the residual (Figure 2). This value may be positive (point above the line) or negative (point below the line). To remove negative values, these residuals are squared. All these squared residuals of the cloud around the line are summed (sum of squares).

A sum of squares can be calculated for various such lines and the line with the smallest sum of squares is the best-fitting line. The sum of squares is a measure of the variation between individuals.

Although the accompanying article6 concerns antimicrobial susceptibility/resistance in E. coli [dichotomous (yes/no) variable] the concepts are more easily illustrated with continuous variables. As an example, the relation between (systolic) blood pressure (continuous variable) and different individual and group characteristics will be used as an example throughout this article.

In a regression analysis, the blood pressure ($x$) of patients from different general practices can be plotted against the patient’s age ($y$). The expectation is that, in general, the older the patient, the higher their blood pressure, resulting in a positive correlation between age and blood pressure. However, other variables might also influence the outcome, for instance, the smoking status of the patient. When smoking status is also included in the regression analysis it is possible to calculate discrete regression lines for smokers and for non-smokers (Figure 3).

This graph can be interpreted within this context as ‘the older a patient, the higher their blood pressure’ and ‘smokers generally have higher blood pressure compared with non-smokers'; the original association is the same for smokers and non-smokers, but the value of the intercept for smokers is greater. The effect of the variable smoking is said to be fixed, which means that the variable has a similar influence on the outcome of all patients.

---

**Figure 1.** Graph of outcome variable $y$ explained by one independent variable $x$. The second graph shows the simplest expression of a regression line $y = a + bx$, with $a$ corresponding to the intercept and $b$ corresponding to the slope.

**Figure 2.** Regression line drawn through a cloud of points. Each point is an individual observation. The second graph shows detail of the residual, the distance of each point from the regression line.
A variable is a fixed effect variable when the intercepts are different and the slopes run parallel. This concept (inclusion of smoking status) can be expanded to include more variables—the calculations become more complex, but the principles remain the same.

Interaction

Blood pressure is often managed in clinical practice with antihypertensive medication. Not including the information on medication in the overall analysis will distort the results. For instance, if there is no increase or much less increase in blood pressure (the outcome) with age in patients on antihypertensive medication, the slope of the regression line would be less. The effect of higher blood pressure with older age in patients not on medication may be partially or entirely ‘cancelled out’ by a lesser or opposite effect in patients on blood pressure lowering medication. An extreme case would be if patients on medication had lower blood pressure with increasing age and patients without medication had higher blood pressure with increasing age (Figure 4). The difference between those on medication and those without medication is not only the difference in the intercept, but also a difference in slope. This phenomenon is called interaction, or effect modification of the variable medication on the outcome blood pressure.

Interaction means that the effect of the variable is random. Whereas a variable has a fixed effect when regression lines are parallel, but intercepts are different (smoking status), a variable has a random effect when the regression lines have a different slope (medication) and cross at some point.

In the context of the accompanying article,6 regression lines were calculated for the association between the probability of infection with an AMR E. coli for the patient and the number of antimicrobial prescriptions in the previous 12 months. The model showed that an increased number of antimicrobial prescriptions to a patient in the previous year increased the chance that infection would be with an AMR E. coli (OR increased for each additional prescription compared with no prescriptions). This variable has the same effect for each individual and is therefore a fixed effect. If the number of prescriptions in the previous year had a different effect on the chance that infection would be with an AMR E. coli depending on the patient, the effect would be random.

Group-level studies (ecological studies)

Individuals belong to social groups or networks of individuals that may share characteristics and interactions that influence the outcome of interest. For instance, patients attending a particular general practice can be considered a group. Groups can differ in terms of the area where they are living (rural or urban), socioeconomic status or ethnicity. In a database where the patients are considered at the individual level, the practice attended represents a higher group level, which would result in a two-level model. Practices can in turn be clustered into areas to make a third level, and so on.

To revert to the blood pressure example, it may be of interest to determine how blood pressure management differs between practices. For instance, does the time allocated to a consultation in a practice have an influence on the patients’ blood pressure? This question can be answered in different ways, depending on what data are available and the way these variables are included in the analysis.

An ecological study is the analysis of data that are aggregated at a higher level. In the blood pressure example, this is done by calculating a mean practice blood pressure from the individual blood pressure measurements of all patients in a practice and comparing this with the mean time a general practitioner spends with patients in that practice. A graph plotting these means for a number of practices would result in a graph as seen in Figure 2, and a correlation (regression line) can be calculated for this association. However, a correlation is not indicative of a causal relationship. For example, the consultation time may reflect differences in patient population or in some other aspects of the practices’ approach to patient care.

By aggregating data (at the country level, practice level or any other group level), interesting details will be lost and a more detailed analysis of the specific factors that can explain the correlation is not possible. For instance, both age and gender are associated with the resistance of a uropathogenic E. coli, but by calculating a mean, this detail cannot emerge from this type of study.16

In studies of the relationship between antimicrobial consumption and prevalence of AMR, ecological studies are widely used. For example, the European Surveillance of Antimicrobial Consumption (ESAC) group (Figure 5)17 plotted antimicrobial prescribing of each European country against the prevalence of AMR.

Figure 3. Regression line of outcome y (blood pressure) explained by independent variable x for smokers and non-smokers in which smokers overall have higher blood pressure compared with non-smokers.

Figure 4. Interaction: the association between x and y is not the same for patients on or not on medication. In the extreme example given in the text, x is age and y is blood pressure.
in that country. This graph shows a positive correlation; the higher the penicillin use in a country, the higher the occurrence of penicillin non-susceptible *Streptococcus pneumoniae*.

Ecological studies are useful monitoring tools for public health surveillance, as they often use routinely collected data that can be used for hypotheses generation. Ecological studies are not suitable to test causal hypotheses or to identify factors that influence associations between, in our case, prescribing of antimicrobials and emergence of AMR or blood pressure and consultation time.

### Group and individual levels

Levels in a multilevel analysis are the point at which data are aggregated.

To analyse data with a multilevel structure, the database has to be set up in such a way that individuals are assigned to hierarchical levels. This is also known as clustering; patients are clustered within practices. Because of this clustering, levels are assigned to the data, with the lowest level assigned to the individual. The presented example of patients within practices (and the accompanying article) has a two-level structure.

A group-level practice effect means that an imaginary patient visiting different practices will be managed differently depending on the practice that patient visits. For instance, if the difference between practices was the time allocated for a consultation, comparable patients going to one practice where the consultation time was longer might have lower blood pressure overall compared with patients going to a practice where very limited time was available for each consultation. The patients’ blood pressure is influenced by an overall practice effect. This is called dependency between patients going to a practice, or, in general, dependency of data. Dependency of data results in a difference between the sum of the individual effects compared with the overall effect.

When analysing studies including both the information on the individual and the group, a multilevel analysis is performed. A multilevel analysis retains and models all the detailed information, so the correlation between blood pressure and age is estimated in the context of other factors, like differences in management between practices or differences between individual patients. In a multilevel model a distinction can be made between the variation explained by the differences between patients (individual level) and the variation explained by the differences between practices (group level). The variation in blood pressure is ‘partitioned’ into a within-practice component (the variation between individuals) and a between-practice component (the variation between practices). The patient residuals or the individual effect can be explained by individual characteristics (e.g. the age, weight, ethnicity, etc.). The practice residuals or group effects represent practice characteristics that affect patient outcomes, for instance, the time allocated to a consultation or the availability of a practice nurse.

### Empty model

The analysis is stepwise, which means that new variables are introduced in steps, starting off with an ‘empty model’. An empty model does not include any explanatory variables: only the outcome (blood pressure measurements from all patients) and the allocation of each patient to a practice are included in the calculation of the empty model. The blood pressure measurements show variation, and this variation is allocated partly to differences between patients and differences between practices. With the least square method and by multiple calculations the computer comes up with a division of the variation in such a way that the individual and practice residuals are as small as possible. This is known as partitioning of (the unexplained) variance and gives an idea of how much of the variation in outcome (blood pressure) is due to variation between individuals (patients) and between groups (practices). To quantify the amount of variation, a variance partition coefficient (VPC) is calculated, which is a measure of the percentage of variance explained by each level (and should add up to 100%).

The variation between the patients is modelled so that the sum of squares for the practice will be lowest, resulting in a regression line for each practice (Figure 6) and an overall regression line combining these regression lines of each practice (Figure 6, red/thick line is the combined effect).

Subsequent further modelling will introduce individual variables to explain the individual variation and, for instance, the introduction of smoking status into the model will show that part of the difference in blood pressure is due to smoking. Similarly, practice variables are introduced to explain the variation between practices.

The final sum of the (squared) residuals gives the overall unexplained and explained variance. The magnitude of the unexplained variance and its corresponding standard error are combined in a log likelihood (−2LL in the output of a linear regression model). The smaller the −2LL, the lower the unexplained variance and the better the linear model fits the data.
The absolute value of this log likelihood is not very informative, but it is used for the comparison of different models. A $-2\text{LL}$ is calculated every time a new variable is added to the model. As not every new variable explains a significant amount of the variation, the $-2\text{LL}$ of the model with the new variable is compared with the previous $-2\text{LL}$. The difference between these $-2\text{LL}$s gives an indication of the relative importance of the added variable (which can then be tested for its significance). In the final model, only variables that explain a significant amount of variation in the outcome are retained.

**Intra-class correlation (ICC)**

The calculation of a (practice) regression line is based on minimizing the overall sum of the individual residuals. These residuals are an indication of how important the practice is in explaining the overall variation, an ICC is calculated. The ICC is a measure of clustering and gives an indication of the relative importance of the cluster (practice) on the outcome (blood pressure measurements of the patients). The ICC is calculated by dividing the between-practice variance (from the practice residuals) by the sum of the within-(between individuals) and between-practice variances. The value of the ICC lies between 0 and 1, and a higher ICC indicates more clustering. A high ICC means that most of the variation in the outcome (blood pressure) is due to practice-related factors, while a low ICC indicates that it is mostly individual differences that influence the outcome blood pressure. The ICC is particularly interesting for interventions, as it can give an indication of whether an individual- or a practice-based intervention would be most (cost) effective.

**Logistic regression analysis**

The previous explanation refers to linear regression, i.e. if the outcome variable is a continuous variable with a normal distribution. In logistic regression the outcome variable is dichotomous, i.e. the presence or absence of a condition—for instance, susceptibility or resistance to an antimicrobial is the absence or presence of resistance. This outcome is not normally distributed (not a bell-shaped curve), but after a number of transformations, the probability of having this outcome can be analysed with the linear regression techniques as described before.

The probability of having a particular dichotomous outcome (susceptibility or resistance to trimethoprim) from a single test lies in the range of 0% - 100%. Consider a probability of detection of resistance on each individual test is 30%. If we perform 100 independents tests, the number of times we find a resistant result will not be 30 every time. Sometimes it will be 28 or 32 and less often it may be 20 or 40. The probability distribution for this process can be calculated and is called a binomial distribution. However, to perform a linear regression a normal distribution (bell-shaped curve) is required. To obtain a normal distribution from a binomial distribution the relative chance to have the outcome is calculated, also known as the odds of having a certain outcome. A 50% chance has odds of 0 and a 100% chance has odds of infinity. More information on how this transformation is performed is given in the step-by-step instructions (see Supplementary data available at JAC Online). After performing the regression analysis, the coefficients of the variables are transformed back for interpretation and result in the presentation of ORs of having a certain outcome.

Coefficient estimation in linear regression is done with the method of least squares. In logistic regression, because we are estimating probabilities, the least squares estimation is not possible and in its place maximum likelihood estimation is used. This means that instead of estimating the actual values of the model coefficients, parameters are estimated that produce a distribution that give the observed data the greatest probability. While this is a more complex method of estimation, the principles remain the same.

For a more in-depth introduction to logistic regression, a good reference is the work of Hosmer and Lemeshow.18
change in the outcome resulting from a unit change in the predictor variable \(x\). In logistic regression, due to the dichotomy, the probability \(P\) of the outcome is calculated, i.e. the probability of \(y\) occurring given known values of \(x\). The same equation is applicable, but since a number of transformations happened to calculate a linear regression, this \(b\) value (named \(\beta\) in a logistic regression) has to be transformed back to predict the change of the dichotomous outcome. This back-transformation is done by calculating the exponent of \(\beta\) (\(e^\beta\)) and results in the odds of \(y\) occurring. This means that the value \(\beta\) represents the odds of the outcome variable associated with a one-unit change in the predictor variable.

**Median OR (mOR)**

In linear multilevel modelling (the blood pressure example), the ICC is calculated from the individual-level variance (variance between patients) and the higher aggregated-level variance (cluster, i.e. practice). The individual-level and practice-level variance are not dependent.

However, in logistic regression analysis, the chance of a given outcome (resistance to trimethoprim) in an individual instance depends on the prevalence of this outcome at the aggregated level. The interpretation of the ICC in a logistic regression, i.e. the relative importance of the difference between clusters, depends on the prevalence. As a consequence, the ICCs of different studies cannot be compared, as the prevalence of the outcome might be different. An alternative to the ICC for logistic regression is to calculate an mOR. The mOR is defined as the median value of the ORs obtained when comparing two individuals with identical characteristics from two different randomly chosen clusters. The mOR can be conceptualized as the increased risk that an individual would encounter when moving from one cluster to another cluster with a higher risk. The advantage of an mOR is that it can be compared between studies.

**Applying multilevel modelling to the study of trimethoprim and ciprofloxacin resistance in *E. coli* associated with urinary tract infection (UTI) in general practice**

In the companion article the multilevel analysis of prospectively collected data on antimicrobial prescribing to the individual and AMR of *E. coli* is presented. All individuals with an *E. coli* UTI diagnosed during a 9-month period in one of the 22 participating practices were enrolled in the study. The analysis was performed using a multilevel model, with individual prescribing data as well as individual results of the susceptibility of the *E. coli* from UTI. The individuals were grouped within general practices and a number of variables related to the practice were included, including prescription rates, size of the practice and practice resistance.

The association between previous prescribing of trimethoprim or ciprofloxacin and subsequent resistance of *E. coli* to these antimicrobial agents was quantified correcting for other confounding factors. These confounding factors included age, gender, nursing home residence and other morbidities at the individual level, and practice-related factors at the higher group level. The results show that there is an association between antimicrobial prescribing and subsequent resistance of the *E. coli*; individuals who had trimethoprim/ciprofloxacin prescribed in the year prior to the UTI have higher odds of this *E. coli* being resistant to trimethoprim/ciprofloxacin compared with individuals who did not have trimethoprim/ciprofloxacin prescribed. Additionally, it was shown that this association between prescribing and resistance of *E. coli* was incremental, which suggests a causal link.

These associations were more extreme for ciprofloxacin. The difference is most likely due to the difference in prevalence of resistance, which is much higher and more established for trimethoprim.

When looking at the influence of the general practice on this association between prescribing and resistance, the mOR showed a difference between practices in the odds of having an antimicrobial-resistant *E. coli* detected. This can be represented as the difference in the chance of being diagnosed with a resistant *E. coli* in the imaginary event of a patient moving from a practice with low to a practice with high resistance. For trimethoprim this OR was 1.17 and for ciprofloxacin 1.33, which would mean changing practice could increase the chance of having a resistant *E. coli* by as much as 17% and 33%, respectively. The variation between practices is also illustrated in the caterpillar plot, which shows the variation in resistance between practices and is plotted according to increasing rank of the practice variation. This caterpillar plot also illustrates the bigger differences between practices for ciprofloxacin.

Even though the mOR shows a significant 95% credible interval (CrI), which is based on the distribution of the mOR, the variation between the practices was not significant in the overall multilevel model. This might be due to a lack of power or because the practice level is not the correct boundary for looking at differences in AMR of *E. coli* UTI.

The application of a multilevel model in the analysis of antimicrobial prescribing and resistance showed the importance of individual- and group-level factors in the occurrence of the outcome. Multilevel analysis allows quantification of the relative influence of individual- and group-level factors.

**Conclusions**

Multilevel modelling is used when there is structure or hierarchy in the data and individual data are clustered within groups. If clustering is high (mOR and ICC are relatively large), the clusters explain an important part of the variation in outcome between individuals. On the other hand, if clustering is not significant, when the mOR/ICC is small, the outcome can be mainly explained by individual factors.

One of the challenges in communication with the public and with policymakers regarding AMR is that unwanted effects (AMR) may be perceived as irrelevant in comparison with the immediate individual benefit. Multilevel modelling is valuable in studies of AMR because it quantifies both the risks of antimicrobial consumption for the individual and for the community.

**Funding**

This work was supported by a project grant from the Health Research Board (HRB) of Ireland.
Transparency declarations
None to declare.

Supplementary data
Instructions on transforming data from dichotomous to continuous are available as Supplementary data at JAC Online (http://jac.oxfordjournals.org/).

References