Journal of Immunological Methods xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Research paper

Statistical approaches for analyzing immunologic data of repeated observations: A practical guide

Alexander Panda [a,1], Shu Chen [b,1], Albert C. Shaw [c], Heather G. Allore [b,*]

[a] Section of Pulmonary, Critical Care and Sleep Medicine, Yale University School of Medicine, New Haven, CT 06520, USA
[b] Section of Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06520, USA
[c] Section of Infectious Diseases, Yale University School of Medicine, New Haven, CT 06520, USA

ARTICLE INFO

Article history:
Received 3 June 2013
Received in revised form 16 August 2013
Accepted 2 September 2013
Available online xxxx

Keywords:
Heterogeneity
Mixed model
Repeated measurement
Multiple comparisons

ABSTRACT

Translational research not only encompasses transitioning from animal to human models but also must address the greater heterogeneity of humans when designing and analyzing experiments. Appropriate study designs can address heterogeneity through a priori data collection, and taking repeated measures can improve the power and efficiency of a study to detect clinically meaningful differences. Although common in other areas of biomedical research, modern statistical methods using repeated measurements on the same subject and accounting for their potential correlations are not widely utilized in immunologic studies. To highlight these analytic issues, we present a practical guide to understanding and applying analytic methods from commonly used T-tests without adjusting for multiple comparisons to mixed models with subject-specific adjustments for correlations using our data on Toll-like receptor-induced cytokine production in monocytes from young and older adults.

1. Introduction

The rapidly evolving and changing field of immunology has led to the discovery of hundreds of cell surface and intracellular proteins critical for mediating host defense against pathogens. Frequently, genetically homogeneous mice of a single sex are utilized so that there is little variation within a genotype and a resultant focus on genotypic differences. For example, in a particular knockout mouse line the assumption is that the expression of selected immunologic proteins at baseline should be comparable. There usually are no covariates, such as age, weight, and sex of the mice, as these are selected to be similar on these factors. Hence, because of the reduced variability within a group, the sample size needed to test between-group differences is lower than in studies of humans. However, a common analytic mistake is that Student's T-test is used to make many comparisons between small groups. Although a robust test, the T-test assumes normality of the measure — this often may not be known and is difficult to determine for sample sizes of less than 30 subjects (Freedman et al., 2007). In this regard, when analyzing experimental data with small sample sizes, non-parametric tests which do not make assumptions on the underlying distributions of the measurements should be used (e.g. the Wilcoxon rank sum test).

In translational research mechanistic findings from animal models are studied in human subjects. Consequently, analytic methods are required that can account for heterogeneity among human subjects. Here, we will use data from previously published studies on the effects of aging on human Toll-like receptor (TLR)-induced cytokine production (Van Duin et al., 2007a,b) and provide important statistical concepts and the analytic steps laboratory personnel may follow. We will first discuss the importance of characterizing the distribution of outcome observations. This fundamental statistical concept determines if certain statistical tests can be applied and whether data need to be transformed to satisfy

* Corresponding author at: Department of Internal Medicine, Yale University School of Medicine, 300 George Street/Suite 775, New Haven, CT 06520-8031, USA. Tel.: +1 203 737 1892; fax: +1 203 785 4823.
E-mail address: heather.allore@yale.edu (H.G. Allore).
[1] These authors made equal contributions to this work.

0022-1759/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jim.2013.09.004

an assumption of normality. We will then address challenges in analyzing repeated cross-sectional measurements of immunologic parameters.

2. Study setting and background information

We previously reported an age-associated decrease in TLR1/2 function in human monocytes (Van Duin et al., 2007a, b). In this study, participants were recruited at influenza vaccination clinics organized by the Yale University Health Services. In brief, heparinized blood from 159 healthy volunteers was obtained with informed consent under a protocol approved by the Human Investigation Committee of the Yale University School of Medicine. Older (age ≥ 65 years) or young (21–30 years) participants with no history of immunologic disease or acute illness in the 2 weeks prior to enrollment were evaluated. Blood was again drawn 6 to 7 weeks after vaccination to assess antibody response to the inactivated trivalent vaccine, as measured by a hemagglutination inhibition assay. Blood was processed as previously described (Van Duin et al., 2007a). We used flow cytometry and intracellular cytokine staining of monocytes and observed a substantial, highly significant defect in TLR 1/2-induced TNF-α ($P = 0.0003$) and IL-6 ($P < 0.0001$) production, in older, compared to young adults.

These differences in TLR-induced cytokine production were highly significant after adjustment for heterogeneity between young and older groups (e.g., gender, race, body mass index, number of comorbid medical conditions) using mixed-effects statistical modeling.

3. Preliminary data analysis: visualizing distribution characteristics

Visually inspecting the distribution of the raw data is a critical step, as data that are not normally distributed may need to undergo transformation procedures before they can be analyzed correctly. Parametric tests assume that the data are drawn from a normal distribution, which can be visually determined with histograms or by applying tests of normality. Many commonly used parametric statistical tests, such as the T-test, and ordinary linear regression assume that the data or model's error term is normally distributed. Using parametric tests for data which are not normally distributed may be influenced by outliers, causing biased results. In this regard, it is important to stress that biological factors (e.g. amount of cytokines produced, expression of co-stimulatory molecules) rarely follow a normal distribution. Because departures from normality are not uncommon, several data transformation methods are available for non-normally distributed data (Hollander and Wolfe, 1973).

4. Testing for normality

To determine if such data transformation is required, a histogram of the outcome distribution can be approximated to the normal distribution (the classical bell-shaped Gaussian distribution, with a single peak at the mean and 95% of observations falling between 2 standard deviations of the mean; see superimposed plot in Fig. 1). Simple descriptive statistics can provide some insights; within this context,

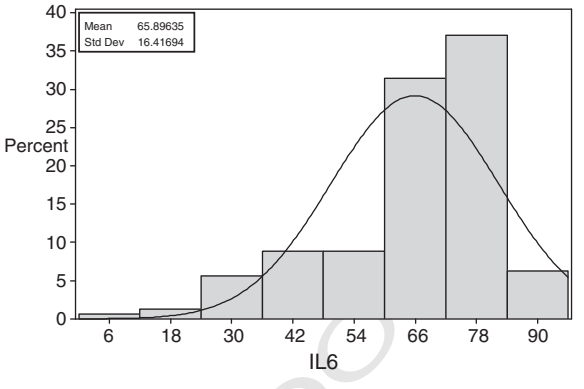

Fig. 1. Histogram plot for IL6 with superimposed normal distribution with mean and standard deviation.

visually assessing the shape of the distribution can be very useful. A superimposed normal plot on the histogram allows one to not only see if the data are approximately normally distributed, but also show where there may be departures from normality. For example, if the skewness (which measures the deviation of the distribution from symmetry) clearly differs from zero, the distribution is asymmetric, while normal distributions are perfectly symmetric. If the kurtosis (which measures "peakedness" of the distribution) clearly differs from zero, the distribution is either flatter or more peaked than normal; the kurtosis of a normal distribution is zero. Fig. 1 displays an example for the distribution of the cytokine interleukin 6 (IL6) generated after stimulation of monocytes with different TLR ligands. The superimposed normal plots show that the underlying distribution is asymmetrical.

There are several tests for normality, such as the Shapiro–Wilk, Kolmogorov–Smirnov, Cramer–von Mises and Anderson–Darling tests (D'Agostino et al., 1990). As noted, most linear regression techniques assume that errors are normally distributed; though minor departures from normality of the outcome may not always be serious. To determine whether these departures from normality may be serious, one analyzes the residuals (the difference between the observed values of a dependent variable and the values predicted by the regression line) of the regression model and inspects whether they are normally distributed in a plot depicting the residuals on the vertical axis and the independent variable on the horizontal axis (Fig. 2A). Additionally, if the points in a residual plot are randomly dispersed around zero across the predicted range, a linear regression model is appropriate for the data (Fig. 2B). These plots can be easily inspected for outliers (residuals ± 2 standard deviations). If serious departures from normality are found a non-linear model may be appropriate or transformations may be needed.

Another graphical method of assessing the extent of deviation from a normal distribution is the comparison of two probability distributions in a Q–Q plot ("Q" stands for quantile). A Q–Q plot compares a sample of data on the vertical axis to a statistical population/hypothetical population with a normal Gaussian distribution (Fig. 2C). For example, inspection of the intracellular production of the cytokine IL6 in monocytes (as shown in Fig. 2C) reveals that the tails deviate from normality; the departure of the distribution from the expected trend along the diagonal line is due to the presence

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004

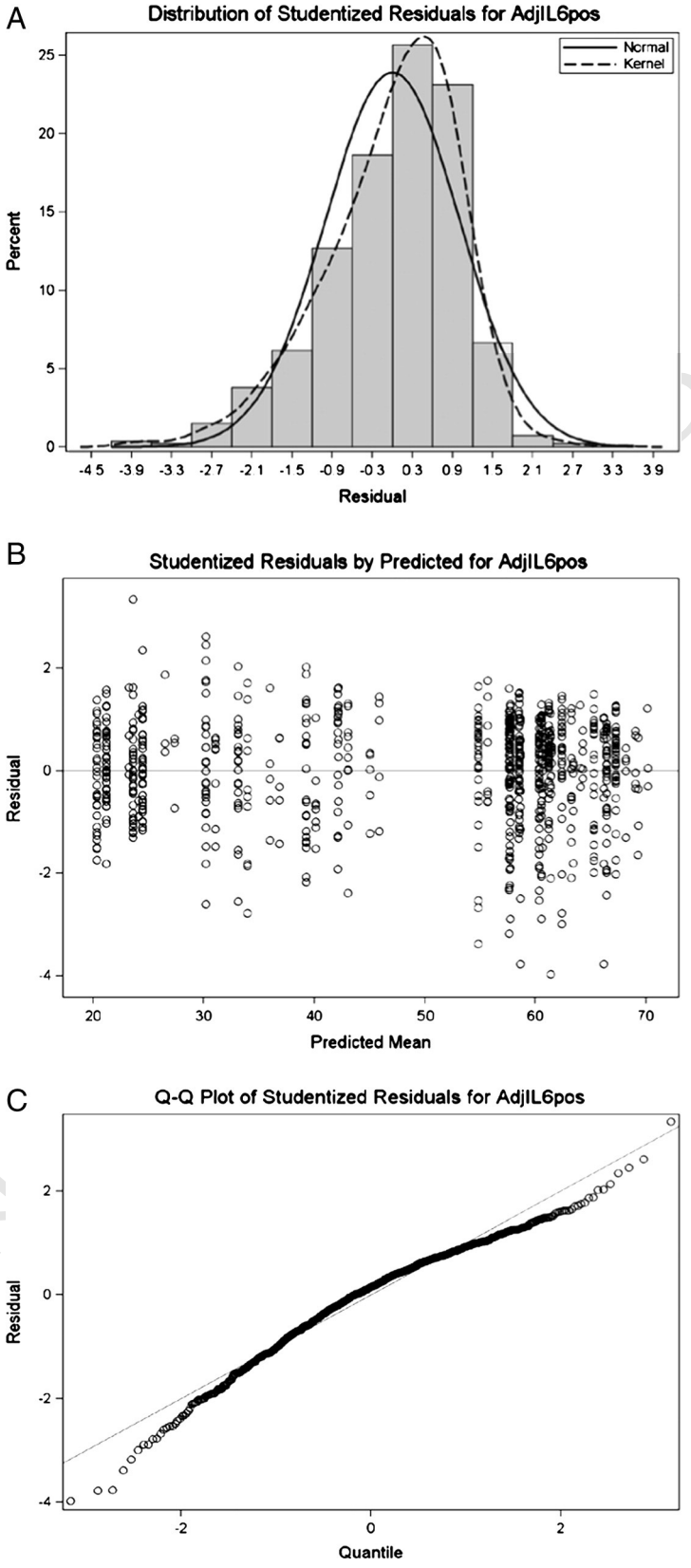

Fig. 2. Studentized residuals for IL6. Panel A: Distribution of residuals with normal distribution superimposed. Panel B: Spread of residuals over the predicted values of IL6. Panel C: Normal quantile–quantile plot.

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004

of substantially larger test statistic values than would be expected if all data were normally distributed. Overall, however, the distribution follows the expected trend along the diagonal line. From Fig. 2 all diagnostics show that most of the residuals fall within ±2 standard deviations with only a few outliers, thus suggesting that there are no serious violations of normally distributed errors.

5. Applying covariate adjustments to address heterogeneity

If the researcher fails to control for variables that are associated with the dependent variable this might also result in biased or imprecise effect estimation. Improved translational human immunologic study designs can record important covariates and multivariable statistical methods are routinely used in medical research.

In our study of the association of TLR-induced cytokine responses with the effect of age group (21–30 versus ≥65 years old) we accounted for the covariates of gender, race and influenza vaccination status from the preceding year. As it is evident from Table 1 (*independent with covariates* versus *independent without covariates*, see Subsection 5 below) the inclusion of these covariates into the model leads to different effect estimates, standard errors and thus, different P-values. This suggests that the outcome (production of IL6) is partially associated to the covariates, as well as ligand-specific age effects.

6. Similarities and differences of correlation and covariance

Correlation is a measurement of linearity; it defines the direction and strength of a linear relationship between two quantitative variables. While there are several measures of correlation, here we restrict the discussion to the Pearson's correlation that assumes two variables as measured on a comparable continuous scale and summarizes the extent that two variables are linearly related to each other. Correlations range from −1 to 1, with a value near or equal to zero implying little or no linear relationship between IL6 produced by a particular pair of ligands, and values approaching −1 or 1 indicating a strong linear relationship. A common first step in data analyses is to create a correlation matrix of all variables to assess the degree of their independence.

Table 1
Model comparison for analyzing IL6.

Outcome IL6			Ligand Flagellin		
			Mean	SE	P-value
T-test		Young	64.61	2.06	0.33
		Old	67.14	1.62	
Adjusted T-test		Hochberg	–	–	0.63
		Bonferroni	–	–	1.00
Mixed models*	Independent without covariates	Young	64.61	1.76	0.31
		Old	67.14	1.73	
	Independent with covariates	Young	66.10	1.80	0.17
		Old	69.60	1.97	
	Compound symmetry	Young	66.10	1.91	0.20
		Old	69.60	2.43	
	Unstructured	Young	65.74	1.98	0.16
		Old	69.76	2.38	

Variation describes the spread or scatter of the data, while covariance measures the strength of relationship between two or more variables. If variables are truly independent (i.e. the correlation is zero), the covariance is zero; however, as for correlation, a nonlinear relationship would also result in a zero covariance but non-independence. The primary concern when analyzing outcomes that are correlated is that the variances of the correlated factors are inflated which may be measured by the variance inflation factor in a linear regression. This will also result in biased test statistics and P-values (Bozdogan, 1987).

7. Repeated observations on a single subject

Repeated observations on subjects are commonly used in immunological studies without making use of statistical methods to analyze them (Table 1 T-test). For example, when blood is drawn from a cohort of humans or animals, and the same immunological outcome is measured under different stimuli, this represents a cross-sectional repeated observations design. Clearly, the advantage of doing so is that for any given cohort, outcome observations on a variety of markers can be measured, reducing either human recruitment costs or animal care fees. In our experiments on human TLR function, by stimulating monocytes with a 4 different TLR ligands and measuring the production of the same cytokine, one obtains 4 repeated outcome observations per subject.

Repeated observations of the same outcome, such as IL6, on the same subject are likely to be correlated. In our study, we repeatedly measured the production of IL6 (among other cytokines) on the same individual after stimulation with different TLR ligands. The inset of values of Fig. 3 shows the Pearson's correlation of responses for IL6 production associated with the indicated pair of ligands. As an example, the Pearson's correlation of r = 0.91 for the relationship of IL6 responses following stimulation with the ligands Flagellin and lipopolysaccharide (LPS) in older adults indicates that knowing the response to stimulation to LPS is strongly related to the response to stimulation to Flagellin and vice versa. This may reflect underlying relationships among TLR signaling pathways; thus, for analytic purposes, these variables cannot be considered to be independent, and must be analyzed using a method that accounts for such correlation.

8. Understanding multiple comparisons and applying multiple testing corrections

Estimates derived from the same model share the same error distribution, so post-hoc adjustments are not routinely undertaken. This is fundamentally different when estimates are derived from separate models when different hypotheses are tested. When many hypotheses are tested, the classical dilemma of multiple comparisons arises, in which the chance of one or more incorrect significant findings among all these tests (also referred as "familywise error rate") increases with the number of tests performed (Wolfinger, 1993, 1996; Shaffer, 1995). Setting the Type I error to 0.05 will result in 5 of 100 coefficients appearing significant by chance. Thus, when multiple tests are performed the original Type I error is not maintained. It is well known that the probability of false

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004

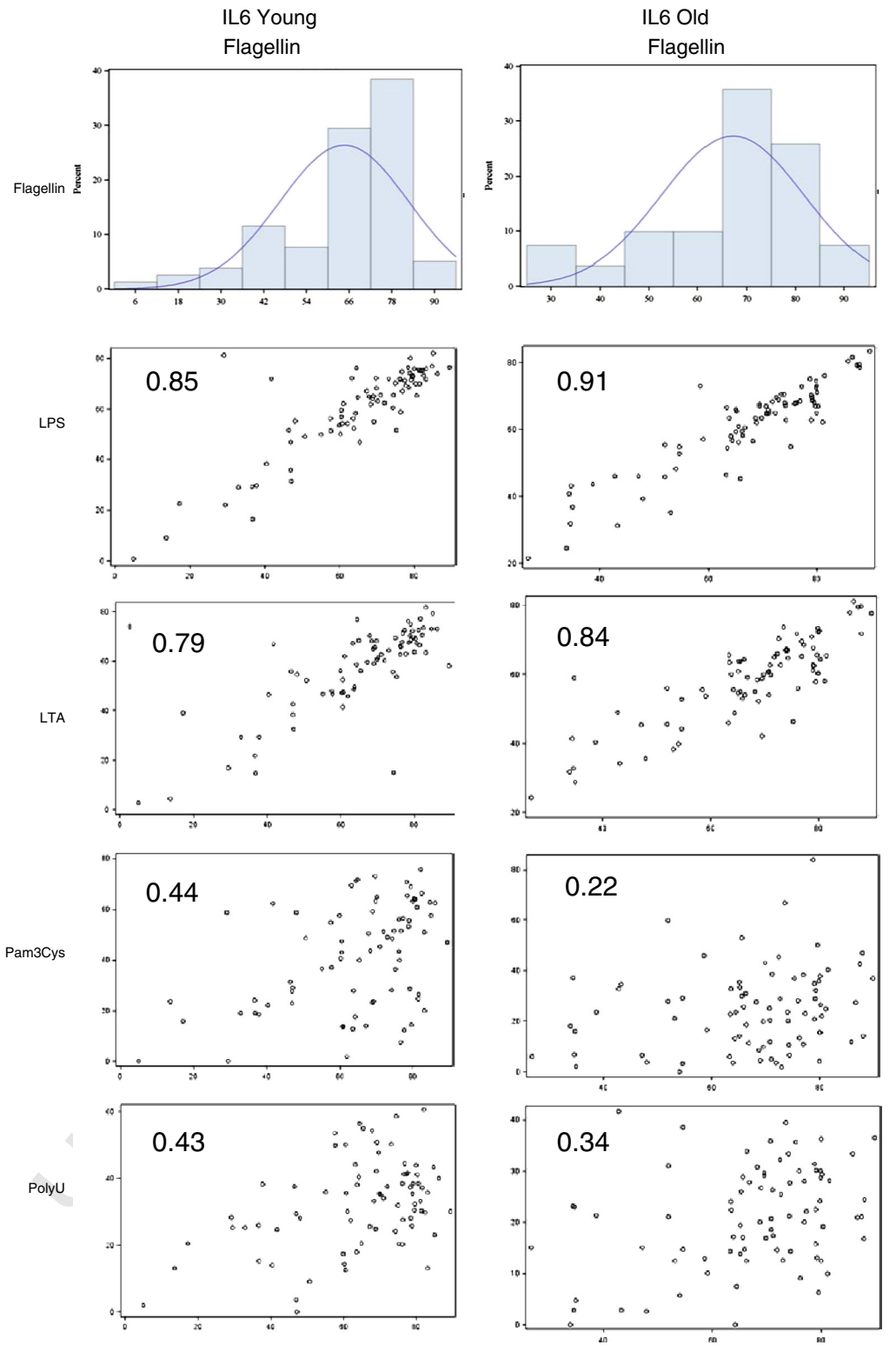

Fig. 3. Histogram plot for IL6 for Flagellin young and old groups. Scatter plot for IL6 for Flagellin and other ligands for young and old groups. Numbers in the plot are Pearson's correlation coefficients.

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004

positive results (or Type I error) increases with multiple tests or multiplicity.

Within a given study, many separate models are usually tested; thus, multiple tests or comparisons are common. While problems with study design, measurement error and inappropriate analytic methods may result in erroneous conclusions, researchers should be aware that multiplicity, especially in exploratory studies, may become a problem.

Multiple comparison procedures are used to control for the familywise error rate. For example, in our report on age-related differences in TLR induced cytokine production in human monocytes, we compared the production of certain cytokines after stimulation with five TLR ligands (TLR 1, 2, 3, 4 and 8). We compared whether the older and young adult group least-squared (LS) means for IL6 (among others), differed after stimulation in separate observations of each of the five ligands. If we wanted to carry out all pairwise comparisons of the group LS means that for IL6 there would be 10 comparisons (i.e. TLR1 compared with TLR2, TLR4, TLR5 and TLR8; TLR2 with TLR4, TLR5 and TLR8; TLR4 with TLR5 and TLR8; and TLR 5 with TLR8). We refer to such sets of comparisons as "family". Each of these 10 comparisons will yield 10 different P-values for a certain Type I error designated as α (at which we would reject the null hypothesis of equality of means). However, the null hypothesis tests at the pre-specified α for each of the comparisons, but not for the whole set of comparisons. The probability of incorrectly rejecting some of the null hypotheses would be larger than α. This might result in the incorrect rejection of the null hypothesis (Type I error). Hence, multiple comparisons procedures are useful to control the familywise error rate.

Multiple comparisons methods can be divided into two types: single-step methods and sequentially rejective (stepwise) methods. With single-step methods where a single critical value is calculated against which tests are compared, directionality of the mean difference and confidence intervals can be constructed; however, they generally have lower power. Most sequentially rejective (stepwise) methods, test mean equality, not directionality, and these methods test null hypotheses for certain significance levels. These methods are less conservative than the corresponding single-step procedures.

Bonferroni and Sidak tests are single-step methods which use adjustment for the number of tests compared (Table 1 Adjusted T-test). The Bonferroni–Holm and Sidak–Holm methods modify these to become "step down" methods—that is, one adjusts the smallest P-value, then the second smallest, and so on to the largest. The term "step down" refers to the fact that one starts with the most significant and "steps down" to the least significant. The Hochberg procedure is a "step up" procedure and works in the reverse direction: one starts by adjusting the P-value for the least significant test and "steps up" to the most significant (Table 1 Adjusted T-test). Step down methods maintain strong control over the familywise error rate when there are strong positive dependencies among the tests.

We would be wrong to suggest that all multiple testing inference issues are resolved by selecting an appropriate multiple comparison procedure, as there are pre-planned and post-hoc analyses. As with any statistical inference method, there is never only one correct method for the analysis of data. However, with multiple comparison procedures there can be meaningful differences between the P-values obtained before and after multiplicity adjustments.

9. Introducing mixed effect models for repeated observations and controlling heterogeneity: putting it all together in a unified model

A mixed model is a statistical model containing both fixed and random effects. We believe that these models are particularly applicable to translational immunologic research. Typically, experiments observe fixed effects (e.g. genotype, age groups or disease status). When working with heterogeneous human populations, researchers should include covariates, but these will not address unmeasured sources of variation. Thus, random effects may be useful to account for the unmeasured latent factors each subject may have in relation to the outcome. For example, the genetic makeup of individual members of our cohort of older and young people arises from some unknown distribution which may contribute to variation in TLR-induced cytokine production.

Over the last few decades most forms of regression models have been enhanced to include random effects. They are particularly useful in settings where repeated measurements on the same subject are performed. For translational immunologic research, the accommodation of repeated measurement analysis (as discussed in Section 7) and random effects makes mixed effects models attractive. We have employed a mixed effects model to estimate the effects of age on TLR function (Van Duin et al., 2007a,b; Panda et al., 2010) and have found Q11 Q12 that they improve the overall model fit.

The most critical factor in analyzing data is to select the model that is most appropriate to address the hypothesis given the collected data. Models should address the hypothesis in question: thus, in-depth knowledge of the question and the cohort under study is critically important in model creation.

10. Introducing common covariance structures: what makes biologic sense?

For any analysis to be valid, the covariances among repeated observations must be modeled properly. It is helpful to consider different covariance structures (Little et al., 2000). Q13 Here, we will explore three common covariance structures:

Independent (often referred to as *variance components*) — assuming independence, the covariance between repeated IL6 outcome after stimulation with different TLR ligands is zero. This model estimates each ligand-stimulated TLR response simultaneously; thus, the standard errors are adjusted, but the mean estimates remain the same. Because there is more information on the IL6 outcome, this approach is more powerful than separate models. The model can also include covariates that partially account for the heterogeneity of study populations. Table 1 shows that the estimates of the means, their standard errors and subsequent P-values change in the independent models when covariates are added. In some cases, such analyses will reveal significant relationships — while in other situations they will account for confounding, thereby preventing the

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004

overestimation of statistical significance. As seen in Table 1, the mean estimate for the generation of IL6 after stimulation of cells with TLR5 ligand Flagellin in the young participants is 64.61 with a standard error of 1.76 for the independent model without covariates and changes to 66.10 with a standard error of 1.8 for the independent model with covariates. Thus, the P-value changes from 0.31 for the independent model without covariates to 0.17 for the independent model with covariates (≈55% reduction).

Compound symmetry: This structure assumes that the relationship between variables is not independent and that the covariances among all pairs of observations are the same. Repeated measures ANOVA assumes this covariance structure. This assumption is appropriate when there are only 2 repeated observations, or the repeated observations arise from the same underlying immunologic mechanism; however, it may not hold when observations are repeated over time or the underlying mechanisms differ. Typically, measurements that are relatively closely spaced (i.e. consecutive measurements, for example at baseline and 6 weeks later) will be more highly correlated than measurements made farther apart (for example at baseline and 12 months later). As for the example from our study (Van Duin et al., 2007a,b) TLRs recognize different molecular patterns conserved in pathogens such as bacteria, viruses and fungi and TLR dysfunction might be the consequence of the same or different defects in the underlying signaling cascade. Table 1 shows that the mean estimates are the same as the independent model, but the standard errors change as do the subsequent P-values.

Unstructured: Sometimes no standard covariance structure fits well. In our study, allowing each individual to have its own covariance structure instead of a shared common covariance structure may best model the biology (Van Duin et al., 2007a,b).

An unstructured covariance structure permitted each participant to have a unique covariance structure, such that IL6 responses for each TLR ligand have their variances and covariances estimated. This addresses the problems of a heterogeneous cohort and the cross-sectional repeated measurement of cytokine outcomes shown in Fig. 3. How did this affect significance levels as compared to the T-test? The P-value for the generation of IL6 after stimulation of monocytes with the TLR5 ligand Flagellin changed from 0.33 for the T-test not corrected for multiple comparisons versus 0.63 when the P-value was adjusted for five multiple comparisons (Table 1). Subsequently, however, this P-value changed to 0.16 when the model with covariates and an unstructured covariance structure was applied. Comparing the other estimates from the T-test not corrected for multiple comparisons and the models with an unstructured covariance, one can see that the mean estimate for the generation of IL6 after stimulation of cells with TLR5 ligand Flagellin in the older adults is 67.14 for the T-test not corrected for multiple comparisons and changes to 69.76 for the model with an unstructured covariance. A larger proportional change is the standard error change from 1.62 to 2.38 in older adults; thus, including covariates and accounting for correlation among TLR ligands are critical to minimize biases in the analysis of this data. Although the use of these methods in our case did not ultimately result in a change of statistical significance, it is easy to imagine circumstances in which such substantial magnitudes in P-value variation could profoundly affect data near the threshold of significance.

10.1. Choosing among covariance structures: letting model fit help select

Final model selection is based on measures of "goodness of fit". One such "goodness of fit" test is the Akaike Information Criterion corrected for sample size (AICC) [Bozdogan, 1987]. Such tests tend to be composed of two parts, one that reflects the accuracy of the fit and another that penalizes for increased numbers of parameters estimated in the model. Thus, one fits the data using different covariance structures and chooses the one with the smallest AICC. The AICC provides a means for comparison among models — a tool for model selection. In our case, the model that appeared to fit best was the model with an unstructured covariance. It is important to remember that the results of diagnostic analysis depend on the model. For example, an observation can be highly influential and/or an outlier because the model is not correct. The appropriate action is to change the model by transforming the data, distribution of the outcome or covariance structure, not to remove the data point. Outliers can be the most important and noteworthy data points, since they can point to a model misspecification. The task is to develop a model that fits the data, not to develop a set of data that fits a particular model.

11. Illustrative example the step by step approach

Our illustrative example focuses on only one out of five ligands (Flagellin, which engages TLR5); however the results are from a model described in Section 2 which includes all five repeated observations of IL6 by after stimulation with five different ligands. Detailed instructions on the data structure, SAS programming code, interpretation and citations of articles using this methodology are available at: http://grasp.med.yale.edu.

11.1. Step 1: visualizing distributional characteristics and model fit

As the superimposed normal plot on the histogram for IL6 production shows, the underlying distribution of IL6 is somewhat skewed to the right, showing departures from the assumption of a symmetric normal distribution (Fig. 1). We determined whether these departures from normality might be serious, analyzing the Studentized residuals of the regression model (Fig. 2B) and inspecting whether they are normally distributed by plotting them in relationship to a standard normal curve (Fig. 2A). For the cytokine IL6, the quantile–quantile plot of the Studentized residuals mostly follows the expected trend along the diagonal line with departures at the tails (Fig. 2C). It is also evident that most of these residuals fall within ±2 standard deviations, though there is a trend toward lower residuals with increased predicted mean IL6. Taken together, the residual results suggest that there are no serious violations of normally distributed errors.

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004

11.2. Step 2: visualizing and estimating correlations of repeated observations

The Pearson correlation coefficients for the generation of IL6 between the TLR ligands Flagellin and Pam3CSK4, LTA, LPS and PolyU, respectively, range from 0.43 to 0.85 for young and from 0.22 to 0.91 for older adults (Fig. 3). Thus, these findings do not indicate that IL6 production by different ligands is independent. Furthermore, the correlations among all pairs measured are not similar. Thus, compound symmetry, which assumes the same correlation among all pairs, would not be the appropriate correlation structure. Notably, the correlation coefficients are not the same within ligands in different age groups. For example, correlation coefficients for IL6 between Flagellin and Pam3CSK4 are halved in older adults compared to younger adults, suggesting that an unstructured covariance structure might fit best.

11.3. Step 3: adding covariates to control for confounding and heterogeneity

Race and influenza vaccination in the previous year were associated with age group ($P < 0.0001$, $P = 0.0002$, respectively) and the outcome of IL6 level ($P = 0.01$, $P = 0.02$, respectively); thus, they are confounders and were included in the model of IL6 production. These variables were associated with age group due to sampling imbalances.

11.4. Step 4: model selection and final model

The model utilizing the unstructured covariance structure had the lowest AICC: the independent score was 6561, compound symmetry was 6272 and unstructured was 6008. Thus, models utilizing the unstructured covariance structure provided are the best fit of the three covariance structures.

12. Summary

Translational studies of human immunology will require analytic models that account for heterogeneous samples, correlation among predictors and among outcomes, control of covariates, and repeated observations on the same subject. Statistical analysis may be invalid if the assumptions behind those tests are violated. In general, methods to visualize how data are distributed, and to account for multiple comparisons and repeated measurements (when applicable) should be applied. When interpreting a set of P-values not corrected for multiple comparisons, readers must consider the possibility of a Type I error and overestimation of statistical significance. It is often best to consider these multiple comparisons as exploratory, intended to generate hypothesis that can be tested with future, more focused experiments. We provided step by step illustrations of these concepts to show how LS mean estimates and standard errors differ, and in turn, influence the significance levels of observations depending upon whether these methods are applied. Many of the concepts discussed in this paper apply widely in translational and biomedical research, and should be applied at the design stage of a study, as they affect sample size calculations and analytic plans. It is our hope that this practical guide will allow analytic laboratory personnel to become familiar with these methods to improve the analysis of experimental data in immunology.

13. Uncited references Q16

Hannan and Quinn, 1979
Shibata, 1989

Acknowledgment

This work was supported in part by the Center of Excellence in Aging at Yale University, funded by the John A. Hartford Foundation, and by the Yale Claude D. Pepper Older Americans Independence Center (P30 AG021342). This work was also supported by the National Institutes of Allergy and Infectious Diseases (U19 AI089992, Contract N01 272201100019C-3-0-1, and K24 AG042489 to A.C.S.). A.P. was a Brookdale Leadership in Aging Fellow and is a Beeson Scholar (1K08AG042825-01).

References

Bozdogan, H., 1987. Model selection and Akaike's information criteria (AIC): the general theory and its analytical extensions. Psychometrika 52, 345.
D'Agostino, R.B., Belanger, A., D'Agostino Jr., R.B., 1990. A suggestion for using powerful and informative tests of normality. Am. Stat. 44, 316.
Freedman, D., Pisani, R., Purves, R., 2007. Statistics, 4th edition. W.W. Norton & Company.
Hannan, E.J., Quinn, A.G., 1979. The determination of the order of an autoregression. J. R. Stat. Soc. B41, 190.
Hollander, M., Wolfe, D.A., 1973. Nonparametric Statistical Methods. John Wiley & Sons, Inc., New York.
Little, R.C., Pendergast, J., Natarajan, R., 2000. Modelling covariance structure in the analysis of repeated measure data. Stat. Med. 19, 1793.
Panda, A., Quian, F., Mohanty, S., van Duin, D., Newman, F.K., Zhang, L., Chen, S., Towle, V., Belshe, R.B., Fikrig, E., Allore, H.G., Montogomery, R.R., Shaw, A.C., 2010. Age-associated decrease in TLR function in primary human dendritic cells predicts influenza vaccine response. J. Immunol. 184 (5), 2518.
Shaffer, J.B., 1995. Multiple hypothesis testing. Annu. Rev. Psychol. 46, 561.
Shibata, Ritei, 1989. From Data to Model. Springer Verlag, New York 215.
Van Duin, D., Allore, H.G., Mohanty, S., Ginter, D., Newman, F.K., Belshe, R.B., Medzhitov, R., Shaw, A.C., 2007a. Prevaccine determination of the expression of costimulatory B7 molecules in activated monocytes predicts influenza vaccine responses in young and older adults. J. Immunol. 195 (11), 1590.
Van Duin, D., Mohanty, S., Thomas, V., Ginter, S., Montgomory, R.R., Fikrig, E., Allore, H.G., Medzhitov, R., Shaw, A.C., 2007b. Age associated defect in Human TLR 1/2 function. J. Immunol. 178 (2), 970 (184).
Wolfinger, R.D., 1993. Covariance structure selection in general mixed models. Commun. Stat. Simul. Comput. 22, 1079.
Wolfinger, R.D., 1996. Heterogeneous variance — covariance structures for repeated measures. J. Agric. Biol. Environ. Stat. 1 (2), 205.

Please cite this article as: Panda, A., et al., Statistical approaches for analyzing immunologic data of repeated observations: A practical guide, J. Immunol. Methods (2013), http://dx.doi.org/10.1016/j.jim.2013.09.004