Journal of the American Medical Informatics Association, 25(12), 2018, 1669–1674
doi: 10.1093/jamia/ocy121
Advance Access Publication Date: 28 September 2018
Review

Review

Beyond smartphones and sensors: choosing appropriate statistical methods for the analysis of longitudinal data

Ian Barnett,[1] John Torous,[2,3] Patrick Staples,[4] Matcheri Keshavan,[2] and Jukka-Pekka Onnela[4]

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [2]Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA, [3]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA, and [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Corresponding Author: Ian Barnett, PhD, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA (ibarnett@pennmedicine.upenn.edu)

Received 7 March 2018; Revised 19 June 2018; Editorial Decision 16 August 2018; Accepted 23 August 2018

ABSTRACT

Objectives: As smartphones and sensors become more prominently used in mobile health, the methods used to analyze the resulting data must also be carefully considered. The advantages of smartphone-based studies, including large quantities of temporally dense longitudinally captured data, must be matched with the appropriate statistical methods in order draw valid conclusions. In this paper, we review and provide recommendations in 3 critical domains of analysis for these types of temporally dense longitudinal data and highlight how misleading results can arise from improper use of these methods.
Target Audience: Clinicians, biostatisticians, and data analysts who have digital phenotyping data or are interested in performing a digital phenotyping study or any other type of longitudinal study with frequent measurements taken over an extended period of time.
Scope: We cover the following topics: 1) statistical models using longitudinal repeated measures, 2) multiple comparisons of correlated tests, and 3) dimension reduction for correlated behavioral covariates. While these 3 classes of methods are frequently used in digital phenotyping data analysis, we demonstrate via actual clinical studies data that they may sometimes not perform as expected when applied to novel digital data.

Key words: digital phenotyping, mHealth, longitudinal data

INTRODUCTION

Smartphones and sensors are increasingly common tools for behavioral measurement in psychiatric research studies. Smartphones are owned and used as personal digital devices by a broad range of populations, including those with mental illnesses, making them uniquely suited as measurement devices.[1] They represent unique data collection tools, providing non-intrusive, high-frequency, real-time, multivariate sensor data and event logs such as GPS, accelerometer, light, temperature, gyroscope, screen state, Wi-Fi, and microphone. These passive measures may also be paired with ecological momentary assessments, or surveys, deployed on the smartphone. Translating this information into clinical insights requires inferring meaningful behavioral features from sensor data and relating these behaviors to clinical outcomes. Potential real-time behavioral patterns and clinically relevant measures include mobility, location, sociability, activity, routine, sleep, and behavioral anomalies. Collection of such behavioral data is difficult or even impossible with older actigraphy methods, but today is feasible with a smartphone app. This concept of *digital phenotyping*, defined to be the "moment-by-moment quantification of

the individual-level human phenotype *in situ* using data from personal digital devices,"[2] has already proven feasible in diverse psychiatric conditions, including schizophrenia, depression, and bipolar disorder.[3–9]

But data collection is only the first step in digital phenotyping research and must be followed by thorough and statistically thoughtful analysis. New data from digital phenotyping brings both novel opportunities as well as challenges. These smartphone-based tools rapidly generate large quantities of data—up to one million data points per day in studies that collect raw sensor data. These data meet the common definition of big data, being of high velocity, volume, and variety. Further complexities arise from the longitudinal nature of digital phenotyping data that introduce a temporal dimension to all resulting measurements. While other fields of research, such as genetics and neuroimaging, have each developed standard and rigorous methods to avoid spurious results when faced with complex data, such approaches have not become standardized in digital phenotyping. As methods and standards are developed for smartphone-based work, it becomes increasingly important for clinicians and researchers to understand foundational principles of digital phenotyping analysis. In this paper, we address some commonly used statistical methods for digital phenotyping analysis of longitudinal studies and demonstrate potential issues and advantages to a few common approaches. While our application of interest is digital phenotyping, the same issues are present in any longitudinal study with long follow-up periods with many repeated measures, and the same conclusions are therefore applicable in those settings as well.

We review methods related to 3 important areas: mixed effects models, multiple comparisons of correlated tests, and dimension reduction of behavioral covariates. We use examples from actual clinical studies to underscore the power of correctly used methods as well as the dangers of improperly applied ones.

METHODS OF ANALYSIS

Association studies

The goal of association studies is to identify significant correlations between behavioral features and clinical outcomes. Digital phenotyping is currently used in association studies to answer questions such as whether geolocation as determined by a smartphone's GPS sensor correlates with a patient's mood symptoms, or whether social data obtained from smartphone call logs correlate with negative symptoms in schizophrenia. Using a smartphone to continually sample or survey with repeated measurements from the same person tends to create correlated data, so statistical models need to take this within-person correlation into account. The most popular methods for association studies for the analysis of longitudinal data that can account for such correlations are generalized estimating equations (GEEs)[10] and generalized linear mixed models (GLMMs).[11] While GLMMs can be powerful when the model and the distribution of the outcome are correctly specified, the GEE approach is semiparametric and therefore more robust to model misspecification. The use of sandwich estimators, where a robust empirical estimate of the variance is "sandwiched" between 2 parametric estimators, enables consistent standard error estimation for regression coefficients in mis-specified semiparametric models.[12–14] Because of this, in practice when the distribution of the outcome is uncertain, many choose to model their data with GEE. For example, one digital phenotyping study that followed patients with schizophrenia after being discharged from the hospital used GEE to test for bivariate relationships between self-report of symptoms and behavior.[5]

Robustness to model misspecification makes GEE a popular choice. However, GEE may not be appropriate for many digital phenotyping studies, particularly when there is a large number of observations obtained for each subject. In contrast, GLMM might be a better choice in terms of overall accuracy when model misspecification is not severe, as might be the case in digital phenotyping studies. To demonstrate the performance of GEE vs GLMM for the analysis of digital phenotyping data, we evaluated 17 patients with schizophrenia for up to 3 months through their smartphone use.[2] For each day of data collection, 16 summary features of mobility were estimated from GPS. For example, distance traveled, fraction of time spent at home, and the fraction of time spent not moving are several of the features used to summarize daily patient mobility. For patients with Android smartphones, text and call logs were used to calculate 15 aggregate summary features of sociability, such as the number of texts/calls sent and number of texts/calls received. In addition to these daily measures of mobility and sociability, phone surveys were given to the patients, measuring their anxiety, depression, sleeping habits, psychosis, and warning signs of psychosis. This study, which combined smartphone-based surveys to record self-report of clinical symptoms alongside passively collected behavioral features based on smartphone sensors, represented a design typical of most current studies using smartphone-based digital phenotyping in populations with psychiatric disorders.

Based on this data, we simulated 10 000 datasets under the null hypothesis of no association between anxiety self-report and a behavioral digital phenotype. The *P*-values from this simulation are displayed in Figure 1, where the linear mixed model has an accurate Type I error due to anxiety being approximately normally distributed, while GEE had a greatly inflated Type I error. An inflated Type I error from GEE would lead to a large number of bivariate associations being falsely reported as significant. Despite being theoretically robust to model misspecification, GEE fails on account of the finite data structure properties inherent to the study design of many digital phenotyping studies. Specifically, the sandwich estimator of the covariance matrix might be too volatile when each patient in the study contributes a large number of observations throughout the period of followup. The number of elements in the covariance matrix grows quadratically with the number of longitudinal (repeated) measurements, making empirical estimation of these elements difficult. Traditionally, sandwich estimators were designed for the analysis of large samples with few repeated measurements per patient, but current digital phenotyping studies tend to be the opposite, with relatively few data-rich patients being followed intensively for long periods of time. Therefore, in digital phenotyping studies, we recommend avoiding sandwich covariance estimators and instead modeling within-patient correlation parametrically with GLMM or with regularized sandwich estimators for GEE.[15] Also, when sample size is relatively small, we recommend avoiding overly complicated random effect structures in a GLMM, as there may not be sufficient data to support the estimation of the large number of parameters required for elaborate random effects. While large sample sizes can alleviate these issues, different computational problems related to handling the enormity of the resulting datasets can emerge.

In addition to small sample sizes and data-rich subjects, pervasive missing data pose an analysis problem in many digital phenotyping studies. There is an extensive literature on accounting for dropout in longitudinal studies for both GEE[16,17] and GLMM.[18–20] However, these methods rely heavily on the specific pattern of missingness that corresponds to dropout. Unfortunately, in digital

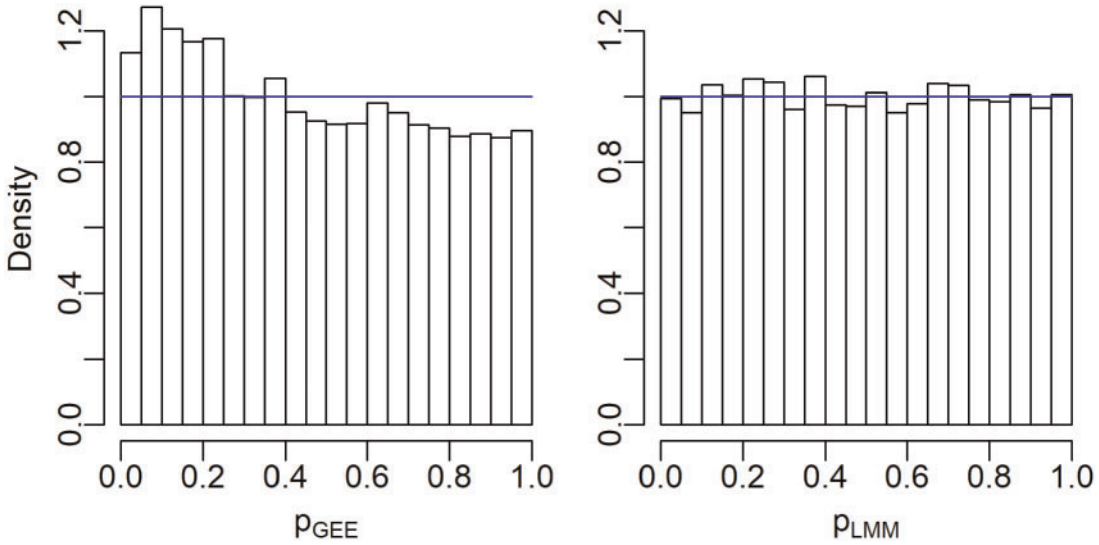

Figure 1. P-values from Type I error simulations of GEE and LMM on a longitudinal schizophrenia cohort. A linear mixed effects random intercept model is compared with the GEE model assuming exchangeable correlation within patient. 10 000 simulations under the null marginal model between outcome (self-report of anxiety) and a single digital biomarker are used to generate *P*-values under the null hypothesis of no association. The GEE model uses the Wald test with sandwich standard errors. The LMM model uses model-based standard errors. For accurate Type I error, the histograms should be approximately uniformly distributed from 0 to 1. While the LMM is accurate, GEE is anti-conservative, which would lead to an inflation in the number of false positives.

phenotyping studies, many of the data are missing sporadically due to many reasons, such as a phone being turned off intermittently or data collection being suspended temporarily by the phone's operating system to conserve resources, and dropout-specific methods cannot be applied in this situation. The most reasonable alternative in our view is using data imputation prior to using complete-data analysis methods. For example, a multiple imputation approach for intermittent missing GPS data from smartphones has been proposed.[21] Some digital phenotyping data can be represented as a time series, in which case, versions of the bootstrap designed for serially correlated data, such as the sieve bootstrap, can be used to impute missing data.[22]

Correcting for performing multiple statistical tests

Consider an association study in which bivariate associations are sought between each self-reported outcome and a behavioral marker. In the setting of digital phenotyping, the number of possible tests that may be performed can be large. Using our schizophrenia cohort as an example, we measured clinical outcomes with 6 symptom scores and 31 daily behavioral measures relating to mobility and sociability, corresponding to $6*31=186$ potential statistical tests. As digital phenotyping studies become more advanced, the number of outcomes and predictors also increases. With so many simultaneous statistical tests, many non-related outcome/predictor pairs will be found to be statistically significant by chance at, say, the 0.05 significance level. These false positives will be abundant unless a more strict significance threshold is used. If allowing any false positives can be harmful or problematic, it is important to control the family-wise error rate (FWER), or the probability of having one or more false positives. The drawback of controlling for FWER is that this can lead to a high number of false negatives and low specificity. A more lenient alternative is to try to limit how large is the fraction of significant tests that are false positives, known as the false discovery rate (FDR). For exploratory analyses where false negatives are more harmful than false positives, we recommend controlling for FDR instead of FWER. If a validation dataset is available,

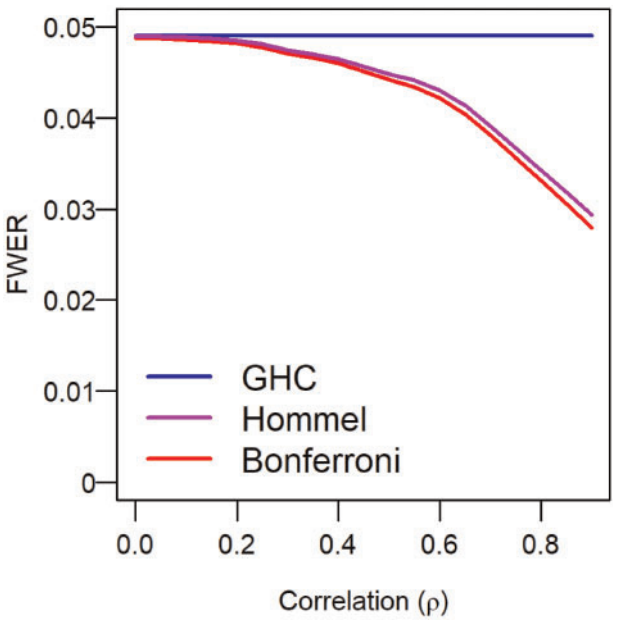

Figure 2. Control of the family-wise error rate for different levels of correlation between tests. Test statistics are generated 1000 times under the null distribution with autocorrelation of ρ, varied from 0 to 0.9 by increments of 0.05. For each ρ, the fraction of the 1000 simulations where there is a false positive at the 0.05 significance level, the estimated FWER, is shown with a smoothed curve.

which is not yet common in this field, one can select a significance level to match a desired sensitivity and specificity through an ROC curve. Given the early state of the field, we focus here on what approaches are useful in the currently more typical and difficult case of discovery-phase studies.

Much of the methodology designed to control both FDR and FWER unfortunately relies on the tests being independent[23–27] or

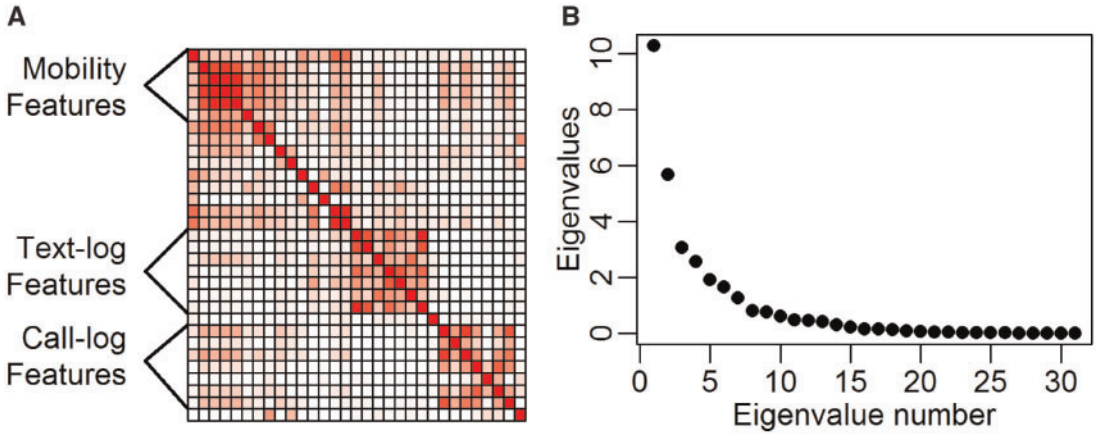

Figure 3. Correlation structure of digital phenotypes. The correlation heatmap of 31 digital phenotypes collected daily for a patient with schizophrenia shows blocks of highly related digital phenotyping features (A). The scree plot shows that there are only a few important factors underlying the full feature set that can be well represented by 4 to 7 latent factors (B). These latent factors correspond to individual blocks of correlated digital phenotypes.

having a specific dependence structure.[28–32] In digital phenotyping, many outcomes and behavioral features can be highly correlated with unknown correlation structure, so if independence is assumed when controlling the FDR and FWER, a decrease in sensitivity is likely. Two popular approaches, proposed by Bonferroni[23] and Hommel,[24] control for the FWER but tend to be particularly conservative when tests are correlated. The generalized higher criticism (GHC) is an alternative approach that was designed to accommodate this correlation and can also be used for controlling the FWER.[33,34] The conservative nature of the approaches that incorrectly assume independence is demonstrated in Figure 2. Alternatively, one could avoid attempting to model the correlation and instead rely on an empirical construction of the null distribution.[35–37]

To investigate methods that attempt to control the FDR, we employ a similar simulation structure that we used for FWER method evaluations. Based on these simulations and using a desired FDR rate of 10%, the Benjamini-Hochberg method[25] led to an FDR of 8%, while the Benjamini-Hochberg-Yekutieli method[28] led to a FDR of 1.6%, implying overly conservative corrections. Surprisingly, both of these rates hardly change as correlation increases between tests. Despite this, both methods, and the Benjamini-Hochberg-Yekutieli method in particular, are overly conservative relative to the desired false discovery rate. When using these approaches in practice, choosing a more generous target FDR, such as 20% or 30%, may be used to offset the decrease in sensitivity caused by these conservative approaches. Another alternative to consider, the principal factor approximation (PFA) method, is capable of accommodating arbitrary correlation between tests while controlling for the FDR.[38] Ultimately, choosing a stringent target for the false discovery rate while employing a conservative multiple testing correction procedure should be avoided, unless specificity is considered more important than sensitivity.

Dimension reduction

Redundant predictors can greatly hamper a wide variety of statistical analyses by increasing variability in estimation, increasing degrees of freedom, and obfuscating model interpretations. Dimension reduction methods aim to find and eliminate these redundancies prior to analysis. The correlation of 31 digital phenotypes for a patient with schizophrenia is displayed in Figure 3. There is a clear clustering of digital phenotyping features based on the behavior type they represent, such as mobility and social communication. This pattern of strong correlation blocks indicates latent factors that could be used to reduce data dimensionality. Including all features from the same correlation block can introduce redundancies and will hamper downstream analyses. For example, both distance traveled and maximum travel diameter are highly correlated, and therefore only one may be necessary to include in the analysis.

There are a variety of methods capable of removing redundancies and reducing the dimension of the predictor set in a principled fashion. When there are no missing data present, classical principal component analyses (PCA) or factor analyses can be used. However, it is much more likely that some digital phenotypes will be missing. Least squares PCA methods can work when there are few missing values.[39–41] More robust alternatives exist, such as probabilistic approaches[42] or estimates of the correlation matrix,[43–45] which are iterative procedures that perform data imputation or use an EM procedure[46] to account for missing variables. The primary drawback shared by most dimension reduction methods is that the interpretability of the resulting factors or transformed predictor variables can be poor, but this danger does not apply to block correlation structures, which are common in digital phenotyping studies. We highly recommend using 1 of these methods to condense clusters of related digital phenotypes into single consensus factors prior to data analyses for most digital phenotyping studies. When examining relationships between digital phenotype data and other independent large datasets, such as imaging and genetics, one can use the parallel ICA approach.[47]

CONCLUSIONS

Digital phenotyping offers unique scientific advantages, such as scalability and data-rich, high-frequency followup for each subject. This approach suggests strong potential for allowing researchers to conduct remote, large-scale, longitudinal studies that capture not only self-reported outcomes, but actual measures of functional and behavioral outcomes. However, as we demonstrate, the use of popular statistical methods and corrections may not perform as expected in this new setting and at times can lead to incorrect conclusions.

The 3 examples covered in this paper offer important lessons for avoiding common pitfalls with digital phenotyping analysis. First, it is important to avoid empirical estimation of within-person

correlations, and instead one should model the associations using GLMMs or GEEs using model-based non-robust inference. Second, when performing exploratory bivariate association analyses, controlling for multiple testing is essential, and the high correlation and redundancy among many digital phenotypes necessitate the use of methods that do not assume independence, such as GHC for FWER control or PFA for FDR control. Third, steps should be taken to remove redundancies in the set of digital phenotyping predictors through dimension reduction methods. Given these methodological challenges, as digital phenotyping research evolves, it is critical that researchers and data scientists collaborate closely to ensure that analyses are conducted properly and correct scientific and clinical conclusions are reached.

CONTRIBUTORS

All authors made substantial contributions to editing and drafting of the manuscript. IB was responsible for the conceptual development and data analysis. JPO developed the smartphone platform, and JT conducted the data collection.

Conflict of interest statement. The authors have no competing interests to declare.

FUNDING

IB, PS, and JPO are supported by NIH/NIMH 1DP2MH103909 (PI: JPO) and the Harvard McLennan Dean's Challenge Program (PI: JPO). JT, LS, and MK are supported by the Natalia Mental Health Foundation. JT is also supported by a Dupont-Warren Fellowship from the Harvard Medical School Department of Psychiatry as well as a Young Investigator Grant form the Brain and Behavior Research Foundation.

REFERENCES

1. Smith A. *Record Shares of Americans Now Own Smartphones, Have Home Broadband*. Pew Research Center; 2017. http://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/ Accessed December 11, 2017.
2. Torous J, Kiang M, Lorme J, Onnela J. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Ment Health* 2016; 3 (2): e16.
3. Torous J, Onnela J, Keshavan M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Transl Psychiatry* 2017; 7 (3): e1053.
4. Onnela J, Rauch S. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacol* 2016; 41 (7): 1691–6.
5. Wang R, Aung M, Abdullah S, *et al*. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In: *Proceedingss of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*; 2016: 886–897.
6. Saeb S, Zhang M, Karr C, *et al*. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res* 2015; 17 (7): e175.
7. Bot B, Suver C, Neto E, *et al*. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data* 2016; 3 (3): 160011.
8. Garza-Rey J, Aguilo J. Remote assessment of disease and relapse (RADAR-CNS). *TMLAI* 2017; 5 (4): 565–571.
9. Spook J, Paulussen T, Kok G, Empelen PV. Monitoring dietary intake and physical activity electronically: feasibility, usability, and ecological validity of a mobile-based Ecological Momentary Assessment tool. *J Med Internet Res* 2013; 15 (9): e214.
10. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73 (1): 13–22.
11. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; 88 (421): 9–25.
12. Huber P. The behavior of maximum likelihood estimates under nonstandard conditions. *Proc Fifth Berkeley Symp Math Stat Probab* 1967; 1 (1): 221–33.
13. Eicker F. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann Math Statist* 1963; 34 (2): 447–56.
14. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; 48 (4): 817–38.
15. Warton D. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* 2011; 67 (1): 116–23.
16. Robins J, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; 89 (427): 846–66.
17. James R, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; 90 (429): 106–21.
18. Diggle P, Kenward M. Informative drop-out in longitudinal data analysis. *Appl Stat* 1994; 43 (1): 49–93.
19. Little R. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993; 88 (421): 125–34.
20. Little R. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; 81 (3): 471–83.
21. Barnett I, Onnela J. Inferring mobility measures from GPS traces with missing data. *arXiv preprint: arXiv: 1606*.06328, 2016.
22. Bühlmann P, Buhlmann P. Sieve bootstrap for time series. *Bernoulli* 1997; 3 (2): 123–48.
23. Dunnett C. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc* 1955; 50 (272): 1096–121.
24. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; 75 (2): 383–6.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995; 57 (1): 289–300.
26. Storey J. The positive false discovery rate: a bayesian interpretation oand the q-value. *Ann Statist* 2003; 31 (6): 2013–35.
27. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 2002; 23 (1): 70–86.
28. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Statist* 2001; 29 (4): 1165–88.
29. Sun W, Cai T. Large-scale multiple testing under dependency. *J R Stat Soc B* 2009; 71 (2): 393–424.
30. Storey J, Taylor J, Siegmund D. Strong control, conservative pioint estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc B* 2004; 66 (1): 187–205.
31. Leek J, Storey J. A general framework for multiple testing dependence. *Proc Natl Acad Sci USA* 2008; 105 (48): 18718–23.
32. Friguet C, Kloareg M, Causeur D. A factor model approach to multiple testing under dependence. *J Am Stat Assoc* 2009; 104 (488): 1406–15.
33. Barnett I, Mukherjee R, Lin X. The generalized higher criticism for testing SNP-set effects in genetic association studies. *J Am Stat Assoc* 2017; 112 (517): 64–76.
34. Donoho D, Jin J. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc Natl Acad Sci USA* 2008; 105 (39): 14790–5.
35. Efron B. Correlation and large-scale simultaneous significance testing. *J Am Stat Assoc* 2007; 102 (477): 93–103.
36. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Stat Assoc* 2004; 99 (465): 96–104.
37. Efron B. Size, power, and false discovery rates. *Ann Statist* 2007; 35 (4): 1351–77.

38. Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary covariance dependence. *J Am Stat Assoc* 2012; 107 (499): 1019–35.
39. Watanabe S, Pakvasa N. Subspace Method in Pattern Recognition. In: *Proceedings of the 1st IJCPR*; 1973: 25–31.
40. Diamantaras K, Kung S. *Principal Component Neural Networks: Theory and Applications*. New York, NY: John Wiley & Sons, Inc.; 1996.
41. Grung B, Manne R. Missing values in principal component analysis. *Chemometr Intell Lab Syst* 1998; 42 (1–2): 125–39.
42. Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. *J Mach Learn Res* 2010; 11: 1957–2000.
43. Ghahramani Z, Jordan M. *Learning from Incomplete Data*. Boston, MA: MIT AI Lab; 1995.
44. Boscardin J, Zhang X. Modeling the covariance and correlation matrix of repeated measures. In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*; West Sussex, England: John Wiley & Sons Ltd. 2004: 215–26.
45. Jolliffe I. *Principal Component Analysis and Factor Analysis*. New York, NY: Springer; 1986: 115–28.
46. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B Methodol* 1977; 39 (1): 1–38.
47. Pearlson G, Calhoun V, Liu J. An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways. *Front Genet* 2015; 6: 276.