Multiple Imputation of Accelerometer Data for Physical Activity Measurement: A Comparison of Methods with an Application to the Active Play Study

By:

Lauren Paul

Supervisor:

Dr. Michael McIsaac

Biostatistics Practicum Report

August 2016

Department of Public Health Sciences

Queen's University

99 University Avenue

Kingston, Ontario

Canada K7L 3N6

Acknowledgements

I would primarily like to thank my supervisor, Dr. Michael McIsaac, for his continuous encouragement, guidance, and patience throughout this process. I would also like to thank my lab group, as well as the faculty and staff at the Department of Public Health Sciences, for all the help and support they provided along the way. Finally, I would like to thank Nicole, Jill, Katherine, and the first-year crew for keeping me sane over the last year; I couldn't have done it without you guys.

Funding and Support

The Active Play Study was funded by the Heart and Stroke Foundation of Canada. The principal investigator of the Active Play Study is Dr. Ian Janssen at Queen's University, Canada. Lauren Paul was additionally supported by a Graduate Research Assistant Fellowship with Dr. Ian Janssen, and the QEII-GSST Award.

Abstract

Physical activity level is a key exposure that is studied in a wide range of health fields. Measurement of physical activity can be a challenge as traditional self-report measures are subject to social desirability bias and recall error. Accelerometers have gained popularity as an objective measurement tool with the ability to provide more accurate and detailed measurements of physical activity level that avoid exposure bias. However, compliance with accelerometer wear is often imperfect, and the removal of accelerometers throughout the day leads to gaps in the data collection sequence. These gaps in data collection result in a need to modify subsequent analyses, as commonly employed analyses often rely on strong, implausible assumptions that may introduce bias.

Multiple imputation is a powerful, flexible statistical tool that can be used in a wide variety of settings with missing data. This technique can be applied to fill in missing accelerometer counts during periods of non-wear for more accurate physical activity level estimation. However, the nature of accelerometer data provides unique challenges for imputation. In particular, imputation methods should account for zero-inflation, autocorrelation, multilevel structure, and high dimensionality present in accelerometer data.

This investigation involved exploration of the theoretical and empirical properties of two zero-inflated Poisson imputation models, and a zero-inflated Poisson Log-normal imputation model, that were designed to address the unique challenges inherent in imputing accelerometer data. Mean imputation surprisingly outperformed all three imputation models in terms of imputation accuracy with the Active Play Study data, indicating that none of the models should be employed with the goal of obtaining accurate estimates of true physical activity patterns. However, the imputation models performed well in the estimation of amount of daily average time spent in moderate-to-vigorous intensity physical activity in the sample, which is a useful result for investigators who only wish to obtain summary measures of physical activity level.

List of Abbreviations and Notation

FCS	Fully conditional specification
JM	Joint modeling
MAD	Mean area distance
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MICE	Multiple imputation by chained equations
ML	Maximum likelihood
MNAR	Missing not at random
MVN	Multivariate normal
MVPA	Moderate-to-vigorous intensity physical activity
PA	Physical activity
PMM	Predictive mean matching
R	Missing data indicator
RMSE	Root mean squared error
X	An $n \times p$ matrix of completely observed covariates
Y	Outcome variable of interest
Y_{obs}	The subset of Y that is observed
Y_{miss}	The subset of Y that is missing
ZIP	Zero-inflated Poisson
ZIPLN	Zero-inflated Poisson Log-Normal
θ	A set of parameters of interest
ψ	A set of parameters from the missingness model

Contents

1 Introduction

1.1 Background and motivating problem

An active lifestyle allows children and youth to not only improve their health, but also improve their self-confidence, mood, and achievement as a student (CSEP, 2012). As part of an active lifestyle, daily moderate-to-vigorous intensity physical activity (MVPA) has been consistently shown to provide important health benefits (Poitras et al., 2016). Moderate-intensity activities allow individuals to burn roughly three to six times more energy per minute than sedentary behavior, whereas vigorous-intensity activities allow individuals to burn greater than six times more energy per minute than sedentary behavior (Harvard T.H. Chan School of Public Health, 2016). The Canadian Society for Exercise Physiology (CSEP) recommends that children and youth accumulate at least 60 minutes of MVPA each day (CSEP, 2012).

Accurate physical activity (PA) measurement is necessary to evaluate current PA levels in populations, identify any changes in PA levels over time, and assess the effectiveness of interventions designed to increase PA levels (Prince et al., 2008). Although Canadian childhood obesity has increased sharply since the 1980s, corresponding with the downfall of active play (Janssen, 2013, 2014), self-reported data indicates that the majority of Canadian children and youth are sufficiently active (Colley et al., 2011). Traditional self-report measures of PA, such as through questionnaires like the PAQ-C (Kowalski, Crocker, & Donen, 2004), have been commonly used due to their low cost and low participant burden (Prince et al., 2008). Unfortunately, the practicality of these measures is overshadowed by their subjectivity to social desirability bias and recall error (Dale W Esliger & Tremblay, 2007; LeBlanc & Janssen, 2010;

Prince et al., 2008). These issues create a need for more objective PA measurement tools that possess greater validity.

Presently, accelerometers are considered the gold standard for free-living PA measurement (Borghese et al., 2016; Esliger & Tremblay, 2007). Accelerometers can capture sporadic bouts of activity often performed by children and youth during active play time, which may be difficult to quantify with a questionnaire (LeBlanc & Janssen, 2010). They can also provide valuable insight into an individual's pattern of activity throughout the day, including sleep, sedentary behavior, and intensity or duration of any PA performed (Esliger & Tremblay, 2007). From accelerometer data, important population health indicators can be derived such as daily average time spent in MVPA (LeBlanc & Janssen, 2010), or other PA summary measures of interest.

Despite these advantages, accelerometers lose their accurate measurement ability when they are not consistently worn. In children and youth, the most common reasons for accelerometer removal are for watersports, if the device is not waterproof, or organized and contact sports where wearing the device is not permitted (D.W. Esliger, Copeland, Barnes, & Tremblay, 2005). The removal of accelerometers throughout the day leads to gaps in the data collection sequence, and if this problem not handled strategically, PA measurements will likely be inaccurate.

The Canadian Health Measures Survey (CHMS) collects important health data on a nationally representative sample of the population, including accelerometer data that makes it possible to assess the proportion of Canadian children and youth that are meeting the CSEP PA guidelines (Colley et al., 2011; Statistics Canada, 2014). Based on the Cycle 3 CHMS from 2013, only 9%

of Canadian children and youth were meeting the CSEP guidelines (Statistics Canada, 2015). If children and youth are removing their accelerometers for important reasons that create systematic differences between wear time and non-wear time, such as for organized sport, PA measurements arising from traditional methods of analysis will be biased. Remedying any identifiable issues in the way accelerometer data is being analyzed in practice may reduce the amount of bias introduced into accelerometer-based PA measurement. These improvements could reveal that a higher or lower proportion of Canadian children and youth are meeting the CSEP PA recommendations than originally estimated, which would have important implications for Canadian public health policy and program development.

1.2 Accelerometer data

Accelerometers are small devices, worn either on the wrist or the hip, that have the ability to track PA by a count value measured over a pre-specified epoch, such as every 15 seconds. They measure body movement in terms of acceleration, with greater accelerations producing greater counts, and count values are used to estimate PA intensity and duration. The activity counts over time are stored in the device, and the data can later be retrieved for analysis (Esliger et al., 2005). An example of accelerometer data from the Active Play Study conducted at Queen's University is seen in Figure 1.1.

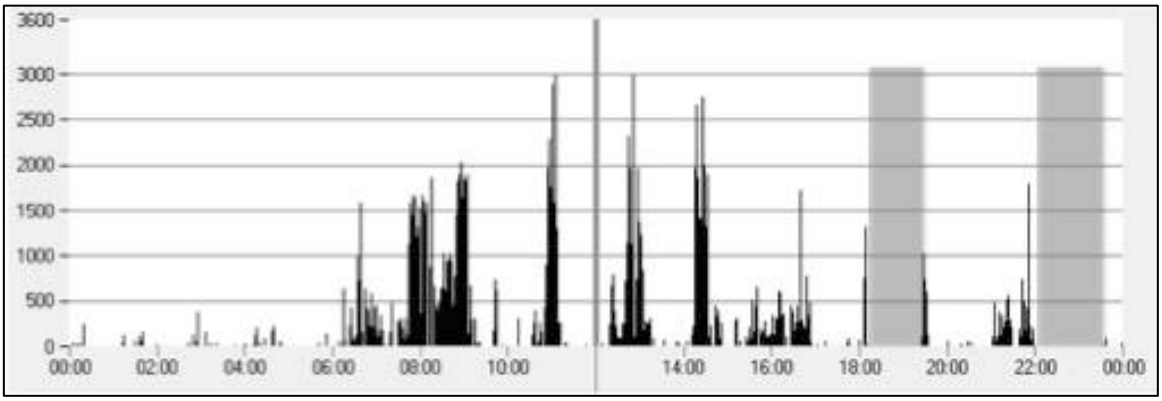

Figure 1.1: Example of data collected by an accelerometer over a single day (12:00am – 12:00am). The x-axis represents the time of day and the y-axis represents the intensity of activity, measured in terms of count value. Each black bar represents the amount of movement done in a single epoch, and the more black bars grouped together, the longer the duration of activity. The two shaded areas in grey highlight extended periods of zero counts.

1.3 Incomplete accelerometer data

The two shaded grey areas in Figure 1.1 represent extended periods of zero counts. These extended periods of zero counts mean that no movement was recorded by the accelerometer. Unless additional information was collected from the wearer, such as through a log sheet, the analyst will not know definitively whether the accelerometer was in fact removed, or if the wearer was simply sedentary during this period. Usually a cutoff of 20, 30, or 60 minutes of consecutive zeros is chosen to indicate a period of device removal; if the period of zeros is longer than this cutoff, it is unlikely that the wearer had remained perfectly still for this long. In Figure 1.1 the cutoff was set at 60 minutes. It was thus assumed that the accelerometer was removed for two periods during the day displayed in Figure 1.1, and there was likely some level of activity that was not recorded. These extended periods of zero counts are referred to as missing data intervals or non-wear time (Catellier et al., 2005; J. A. Lee & Gill, 2016; P. H. Lee, 2013).

In the past, a common approach to handling missing accelerometer data was to simply average over missing intervals during analysis, excluding non-wear periods from both the numerator and denominator. However, this approach can lead to a number of issues. It creates an underlying assumption that the wear time and non-wear time are not systematically different, which is often false and can lead to an overestimation or underestimation of PA level. Additionally, some analysis methods apply data reduction techniques that drop individuals or days without sufficient wear time to obtain more reliable estimates of habitual PA. Sufficient wear time can be described as an individual having at least a certain number of valid wearing days during the period the accelerometer was worn (e.g. at least 3 days). A valid wearing day can be considered a day where the accelerometer was worn for at least a certain number of hours (e.g. a 10 hour minimum of wearing time out of 24 hours) (Alhassan, Sirard, Spencer, Varady, & Robinson, 2008; Mâsse et al., 2005; Penpraze et al., 2006). However, dropping individuals or days from analysis may similarly introduce bias if these individuals or days without sufficient wear time are systematically different from those with sufficient wear time. Finally, data reduction techniques result in a smaller sample size, a loss of information, and a decrease in study power.

Instead of dropping individuals or days, or averaging over missing intervals, we can try and fill in non-wear intervals with plausible count values. We do this by making use of information we have from the observed data, such as demographic information, time of day during the non-wear period, or average activity level. We can create a statistical model based on the observed data, and draw plausible values from the posterior predictive distribution of this model to fill in the missing values. This process is called imputation, and it can reduce the amount of bias arising in the analysis of accelerometer-based PA measures.

1.4 Report structure

The remainder of this report is structured as follows. Section 2 gives a general overview of the concept of missing data and common statistical methods for handling this issue, both theoretically and in the context of accelerometer data. Section 3 focuses on multiple imputation for accelerometer data in more detail, including the challenges, a review of previously proposed methods, and the shortcomings of these methods that need to be addressed and improved upon. The prediction performance of three multiple imputation methods is assessed through an application to the Active Play Study data in Section 4. Finally, in Section 5, the observed properties of multiple imputation methods are examined using a pseudosimulated data set based on the Active Play Study data. Overall conclusions and future directions are discussed in Section 6.

2 Missing Data Overview

In statistical analyses, interest often lies in making inferences on an unknown parameter, or set of unknown parameters, θ, based on the distribution of a set of complete data, Y_{com}, sampled from a population. The process for making inferences on θ is complicated when some parts of Y_{com} are missing (Schafer & Graham, 2002), as standard statistical methods are often developed for use with complete, rectangular data sets (Little & Rubin, 2002).

Let Y_{com} denote a matrix with *n* rows and *p* columns, where *n* is the number of individuals in the data set and *p* is the number of outcome variables. In practice, some entries in Y_{com} could be missing, so Y_{com} can further be broken down into observed values, Y_{obs}, and missing values, Y_{miss} (i.e. $Y_{com} = (Y_{obs}, Y_{miss})$). Let the *n* x *p* matrix R be a matrix of missing data indicators,

where if an entry in Y_{com} is observed, the corresponding entry in R takes on a 1, and if an entry Y_{com} is missing, the corresponding entry in R takes on a 0. Additionally, we let X be an n x q matrix of completely observed covariates, and we suppose that interest lies in making inferences on θ based on the observed data.

The possible values of θ given the observed data are summarized by the posterior distribution $P(\theta|Y_{obs}) = \int P(\theta|Y_{obs}, Y_{miss})P(Y_{miss}|Y_{obs})dY_{miss}$, where $P(\theta|Y_{obs}, Y_{miss})$ is the posterior distribution of θ given the hypothetical complete data, and $P(Y_{miss}|Y_{obs})$ is the posterior distribution of the missing data given the observed data (van Buuren, 2012). This posterior distribution effectively averages over the distribution of the missing data (Schafer & Olsen, 1998).

2.1 Missing data mechanisms

The work done by Rubin (1976) helped develop three classifications of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). These three classifications help describe the relationship between the missingness and the variables in the data set (Little & Rubin, 2002).

MCAR is the strictest classification. It states that the probability of a value being missing does not depend on any other covariates, including both those observed in the data set and those not observed the data set. In equation form, this assumption can be written as $P(R|Y_{obs}, Y_{miss}, X, \psi) = P(R|\psi)$, where ψ is a set of parameters from the missingness model (Little & Rubin, 2002).

The MAR classification relaxes the MCAR requirements. Here, missingness is permitted to depend upon any of the observed covariates in the data set; however, it cannot additionally depend upon any covariates not observed the data set. The equation form of the MAR assumption can be written as $P(R|Y_{obs}, Y_{miss}, X, \psi) = P(R|Y_{obs}, X, \psi)$ (Little & Rubin, 2002). This assumption is the minimum requirement for many of the statistical methods that handle missing data in order to achieve unbiased and efficient estimates of the parameters of interest (Schafer & Graham, 2002). Previous studies have shown that it is often a plausible assumption to make, and departures from the MAR mechanism may only have minimal impact on estimates and standard errors (Collins, Schafer, & Kam, 2001).

If the missing data mechanism is not MCAR or MAR, it is MNAR. This missingness mechanism is the most difficult to work with as the missingness depends on some unobserved covariates. One does not have information about these covariates to help explain the missingness in the data, and so the missingness can only be partly explained by the observed data. The distribution of missingness is simply written as $P(R|Y_{obs}, Y_{miss}, X, \psi)$, and cannot be simplified any further. A joint probability model can be defined for both the observed data and the missingness indicator, R, to yield valid inferences. This joint probability model can be written as $P(Y_{obs}, R|\theta, \psi) = \int P(Y_{obs}, Y_{miss}|\theta)\, P(R|Y_{obs}, Y_{miss}, \psi)\, dY_{miss}$, and the inferences for θ now depend on the model for R, $P(R|Y_{obs}, Y_{miss}, \psi)$ (Schafer & Graham, 2002). Inferences may vary greatly depending on the choice of the R model, and choosing the right model usually requires detailed knowledge about the data under study (Allison, 2002).

Unfortunately, there are no objective tests to identify under which mechanism the missingness in a data set has arisen, and so it is rare to assume the mechanism with confidence. The MAR assumption cannot be verified as one does not have any information about the unobserved covariates on which the missingness may depend (Little & Rubin, 2002). For this reason, potential violations of the MAR assumption should always be considered, and the impact of departures from the MAR mechanism on results should be investigated through sensitivity analyses (Allison, 2002; Schafer & Graham, 2002).

2.2 Complete-case analysis

In complete-case analysis, also known as listwise deletion, individuals are removed from the sample if they have any missing data on any of the variables in the data set. This yields a "completely observed" sample to which the usual statistical analyses can be applied.

This method effectively assumes that the individuals with complete data are representative of the population, which is often false. Additionally, not making use of the partially observed information from individuals with incomplete data renders this method inefficient. This is especially problematic in the case of data sets with many variables, any of which may contain missing values (Little & Rubin, 2002; Schafer & Graham, 2002).

Approaches to handling missing data that do not discard available information are usually more desirable. Two methods for parameter estimation with incomplete data are considered further – likelihood-based methods and imputation methods.

2.3 Maximum likelihood

Maximum likelihood estimates the parameters of interest directly from the observed data (Enders, Mistler, & Keller, 2016). An underlying model is specified based on the observed data, and the likelihood of the parameters of interest is maximized based on the corresponding likelihood function (Allison, 2002).

The likelihood function of the complete data for the parameters of interest can be written as $L(\theta|Y_{obs}, Y_{miss})$. In the presence of missing data, the full likelihood given the observed data, (Y_{obs}, R), is obtained by integrating the complete data likelihood over all possible values of the missing data. Assuming that the two sets of parameters are independent, the full likelihood can be written as $L(\theta, \psi|Y_{obs}, R) = \int L(\theta|Y_{obs}, Y_{miss})L(\psi|Y_{obs}, Y_{miss}, R)dY_{miss}$. Inferences for θ must be based on the full likelihood when the data are MNAR. However, if the data are MAR, inferences for θ can be based on the partial likelihood function that ignores the missing data mechanism $L(\theta|Y_{obs}) \propto \int L(\theta|Y_{obs}, Y_{miss})dY_{miss}$. This is because under the MAR assumption $L(\psi|Y_{obs}, Y_{miss}, \mathrm{R}) = \mathrm{L}(\psi|Y_{obs}, \mathrm{R})$, and so this term can be move out of the integral (Little & Rubin, 2002). The observed data likelihood effectively averages over the distribution of the missing data, and the ML estimate of θ is the value of θ for which the likelihood function is the greatest (van Buuren, 2011).

2.4 Imputation

Imputation is the process of filling in missing data with plausible values based on the information provided by the observed data (Little & Rubin, 2002). There are two general types of imputation: single imputation and multiple imputation.

2.4.1 Single imputation

Single imputation produces a single plausible completed data set, filling in each missing data value with an educated guess based on information available from the observed data. Methods of single imputation include mean imputation, conditional mean imputation, conditional distribution imputation, and unconditional distribution imputation. These methods can only be considered ad hoc solutions to missing data problems, and must be used carefully as they have the potential to distort variable distributions and inter-variable relationships (Schafer & Graham, 2002). One critical shortcoming is that they all fail to reflect the added variability stemming from the uncertainty surrounding the values being imputed (Little & Rubin, 2002). This leads to an underestimation of the total variance, making parameter estimates falsely appear more efficient, p-values artificially low, and rates of Type I error higher than nominal levels (Schafer & Olsen, 1998). This issue can be addressed by expanding to multiple imputation.

2.4.2 Multiple imputation

Multiple imputation (MI), developed by Rubin (1987), is the gold standard of imputation methods and arises from a Bayesian perspective. Missing values and unknown parameters are treated as random, and all of the data's evidence about the missing values and parameters of interest is summarized using probability distributions (Little & Rubin, 2002).

The likelihood function of the parameters given the observed data can be multiplied by a prior distribution, $P(\theta)$, which gives the updated posterior distribution of the parameters, $P(\theta|Y_{obs}) \propto P(\theta)L(\theta|Y_{obs})$ (Little & Rubin, 2002). There are several choices of prior distribution that can be made. One can choose to incorporate informative priors if some additional external knowledge about the parameters of interest exists, or non-informative priors if there is no additional

knowledge (Schafer & Graham, 2002). One common method of estimating the posterior distribution using non-informative priors involves obtaining ML estimates of the parameters of interest from the observed data, and assuming the parameters follow a normal posterior distribution with mean and variance defined by the ML estimates (Schafer & Olsen, 1999). Whatever the choice of prior distribution, its influence in the overall posterior distribution diminishes as the sample size increases. As MI relies on large-sample approximations for the complete data distribution, the choice of prior distribution rarely has a significant impact on results (Schafer & Graham, 2002).

MI creates $m \geq 2$ versions of the complete, filled-in data set with various sets of plausible values based on repeated random draws from the predictive distribution of an imputation model. This is the posterior predictive distribution of the missing values under a particular specified model for missingness, $P(Y_{miss}|Y_{obs}, \dot{\theta})$, given the observed data. The parameters $\dot{\theta}$ are drawn from the posterior distribution calculated from the observed data, $P(\theta|Y_{obs})$. Then, standard complete-data analysis methods can be run on each imputed data set, and the results combined by Rubin's rules (D. B. Rubin, 1987) to obtain overall parameter inferences (Little & Rubin, 2002).

MI methods can be further broken down into several approaches. Our focus is on explicit imputation, rather than implicit imputation where imputed values are borrowed from the observed values of other similar individuals.

Joint modeling

Joint modelling (JM) is a parametric imputation approach. It assumes that the incomplete variables follow a common multivariate distribution (Enders et al., 2016), and leads to imputation procedures whose statistical properties are known under a correctly specified joint imputation model (van Buuren, 2007). However, it can be difficult to define a single joint model for non-normal data sets (Yucel, 2008), data sets with mixed data types, or for high-dimensional data sets with many variables (R. He, 2012). In any of these situations, an alternative imputation approach often must be considered.

Fully conditional specification

Fully conditional specification (FCS), also known as multiple imputation by chained equations (MICE) or sequential imputation, is a popular imputation method for handling multivariate missing data as it does not require that a joint model be defined (Liao et al., 2014; van Buuren, 2007). FCS factorizes the joint distribution of the data as a sequence of unique conditional imputation models for each incomplete variable, and draws missing values to impute in an iterative fashion. These univariate imputation models can be of any form, and tailored to the distribution of the incomplete variables (Liao et al., 2014; van Buuren, 2007). FCS does not require the user to specify a covariance structure among the variables, and it has a lower sample size requirement than the JM approach (Lloyd, Obradović, Carpiano, & Motti-Stefanidi, 2013). This method is particularly useful when no suitable identifiable joint distribution of the incomplete variables exists (van Buuren, 2012), when there are many incomplete variables (Kalaycioglu, Copas, King, & Omar, 2016), when incomplete data is of mixed types (R. He, 2012), or when it is necessary to preserve unique features in the data and maintain constraints

between different variables (van Buuren, 2007). FCS imputation is often the only way to conduct MI with unconventional, non-normal data sets (Yucel, 2008), and simulation studies have shown that FCS generally yields unbiased estimates that have appropriate coverage (van Buuren, 2007).

In general, imputation performance is affected by four factors: the number of covariates included in the imputation model(s), the covariate correlations with the incomplete variable(s), the amount of missingness in the data, and the missing data mechanism at work (Catellier et al., 2005). When applying MI methods, all relevant available information should be included to the fullest extent possible, as this helps to reduce any systematic differences between completely observed and partially observed individuals (R. He & Belin, 2014).

MI is often considered more attractive than ML methods because ML methods are problem specific, whereas MI is more flexible and can be implemented similarly for a variety of analyses (Zhao & Yucel, 2009). Analysts who do not have the statistical expertise required for ML missing data methods can simply analyze the imputed, complete data sets with their usual complete-data analysis models and software (Y. He, Yucel, & Raghunathan, 2011). Additionally, if multiple people are to analyze the data, then the creation of an imputed data set prior to any analyses ensures comparability of results across analyses (Schafer & Graham, 2002). A final key feature of MI is that the imputation model does not necessarily have to be congenial with the analysis models to yield correct inferences (Y. He et al., 2011; Little & Rubin, 2002).

3 Multiple Imputation for Accelerometer Data

3.1 Challenges

There are three particular aspects of accelerometer data that make imputation challenging: zero-inflation, high dimensionality, and a multilevel structure to the data. Typically with accelerometer data, it is desirable to avoid specifying the complex covariance structure of the data, and a multivariate JM strategy is not feasible. For this reason, it is best to focus on FCS methods that have a lower sample size requirement and bypass the need for specifying a single joint model.

Further details on each of these three identified challenges are provided in the following subsections.

3.2.1 Zero inflation

Accelerometer data are zero inflated, where the activity count data contain more zero values than typically predicted with a standard count model, such as a Poisson model. Ignoring the zero-inflated nature of a variable in the imputation process can severely distort the variable's marginal distribution, or its relationship with other variables (Schafer & Olsen, 1999). Imputation methods tailored to zero-inflated count data have been shown to produce unbiased parameter estimates and measures of uncertainty compared to standard count data imputation methods (Kleinke & Reinecke, 2013). As we are interested in the distribution of activity counts for the accelerometer sample, it is natural to assume that the counts follow a Poisson distribution, including additional zeros from sedentary time. The process contributing to the zero inflation needs to be modeled separately, such as through a zero-inflated model like the zero-inflated Poisson (ZIP) model

(Lambert, 1992). This report will focus on ZIP models as they are simple to fit using R software and the *pscl* package, and readily available to apply as an imputation model using the *mice* package and its extensions.

The ZIP model is a mixture of two data-generating processes – the zero-inflation process and the activity count process. Zero inflation occurs with probability π, and activity counts with probability $1-\pi$, giving the probability distribution

$P(Y=y|x) = \begin{cases} \pi + (1-\pi)P(0|x) \text{ if } y = 0 \\ (1-\pi)P(y|x) \text{ if } y > 0 \end{cases}$, where $P(y|x) = \frac{e^{-\lambda}\lambda^{y}}{y!}$ is the Poisson probability mass function. The zero inflation probability, π, can additionally be specified as a function of covariates $\pi(w^T\gamma)$, where w is a set of covariates to help explain the probability of zero inflation, and γ is a corresponding vector of regression coefficients. Common model choices for $\pi(w^T\gamma)$ are the logit and probit models, and the covariates w can be different from the covariates that define the activity count process (Kleinke & Reinecke, 2013).

3.2.2 High dimensionality

Accelerometer data are also high-dimensional. Data sets typically consist of hundreds of individuals, each with thousands of PA measurements in addition to baseline and demographic covariates such as age and sex. The Active Play Study data, for example, consists of 332 individuals, each with a set of measurements that were taken every 15 seconds during awake time for 7 days. Having information on a large number of variables may make the MAR assumption more plausible (R. He, 2012), but it creates a number of barriers for using JM imputation methods. Simpler traditional models, such as the multivariate normal (MVN) model with a general covariance matrix that allows each variable to be correlated with all other

variables, will usually be over-parameterized for the sample size (R. He, 2012). This is especially an issue in data sets with a large variable-to-individual ratio. Additionally, in large data sets variables are often of mixed data types (e.g. count, categorical, continuous), which increases the complexity of trying to model the joint distribution of all variables (Liao et al., 2014).

Liao et al. (2014) compared seven different imputation methods in both simulated and real high-dimensional phenomic data containing mixed data types. These included mean imputation, MICE, a random forest-based method (missForest in R), and various K-nearest neighbor imputation methods. In their simulation study, they found no method to universally outperform other methods, although mean imputation was consistently the worst method. They concluded that the choice between MICE, missForest and K-nearest neighbor method depends on the data types being imputed and the data structure. He and Belin (2014) proposed a MVN JM approach to handle high-dimensional incomplete data with both continuous and binary variables. Normal latent variables were introduced for binary variables so that MVN JM could be used to impute data, and results from their simulations indicated that this method could adapt to different MCAR and MAR missingness mechanisms, as well as various covariance structures. In He's corresponding PhD dissertation (2012), he referenced a method developed by Dunson (2005), in which Poisson latent variables were used in a latent variable model for mixed count, binary, and ordinal data. He suggested that the Poisson latent variable model could be extended to include continuous and nominal variables as well. However, both He (2012), and He and Belin (2014), recognized that if a data set includes count or semicontinuous variables, incorporating these variables into a JM imputation method is challenging, and that an FCS imputation approach would likely still be required.

3.2.3 Multilevel structure and Autocorrelation

Finally, accelerometer data have a multilevel structure, where days of activity measurements are nested within individuals. In addition to time-invariant effects, such as age, sex, and BMI, there are also time-varying effects, including day effects, hour effects, minute effects, and second effects. This multilevel structure must be considered when selecting an imputation method, as basic single-level imputation methods applied to multilevel data can introduce substantial bias and lead to underestimation of standard errors, even under the MAR assumption (Enders et al., 2016; van Buuren, 2011). The imputation model should aim to preserve the correlations that arise from the multilevel structure (Yucel, 2008). In the context of accelerometer data imputation, failing to account for similar activity levels across days from the same individual will lead to underestimation of longitudinal correlations among activity counts, and may not properly recover an individual's true pattern of physical activity during the period the accelerometer was worn (Allison, 2002). In addition to dependence across days, there is a time-associated effect within days, where consecutive accelerometer measurements are expected to be correlated. This can additionally be thought of as a temporal or longitudinal effect within days, and if these within-day effects are ignored in the imputation model, efficiency is lost and bias may similarly be introduced (Junger & Ponce de Leon, 2015). As a result, additional structure needs to be incorporated into imputation models in order to yield more efficient and unbiased parameter estimates for accelerometer data.

A general approach to accounting for dependence across days is to include a random effect for individual. Much of the work on multilevel mixed-effects MI has focused on JM for continuous data with MVN assumptions (Schafer & Yucel, 2002; Yucel, 2008, 2011). Zhao and Yucel

(2009) branched off to explore FCS imputation for one continuous and one binary variable using generalized linear mixed-effects models in multilevel settings. Although the FCS approach was shown to perform well in the imputation of continuous variables, and it outperformed PAN (an R package for imputation using multivariate generalizations of linear mixed-effects models) in the imputation of binary variables, there were a number of convergence issues and poor parameter inference for the binary variable in the presence of higher intracluster correlation coefficients. Enders et al. (2016) recognized that to date there has been limited success in combining generalized linear mixed models with FCS.

For within-day effects, Cano and Andreu (2010) showed that by incorporating appropriate data structure, such as through simple lag effects, MI could be applied successfully to time series data and yield plausible values. Kalaycioglu et al. (2016) found that when normally distributed repeated measures data displayed a first-order auto-regressive correlation structure between measurements, FCS imputation with a moving time window of lag and lead covariates performed well in terms of bias and efficiency.

Methods to capture both across-days and within-day correlations simultaneously have been assessed by Nevalainen et al. (2009) and Welch et al. (2014). Nevalainen et al. proposed a doubly iterative FCS imputation approach, also called two-fold FCS, for repeated measures data with multiple variables that was computationally intensive, but yielded valid parameter inferences. This was achieved through univariate models for each variable at each time point that were iterated both "within-time" including one lag and one lead covariate, and "among-time". Welch et al. applied the doubly iterative FCS approach in high-dimensional longitudinal settings

with larger proportions of missingness. It was shown that two-fold FCS still created unbiased and efficient parameter estimates in these settings, especially in the case of imputing time-dependent covariates with greater longitudinal correlations.

3.2 Previously proposed methods

Two studies by Catellier et al. (2005) and Lee (2013) contributed key findings to the advancement of imputation for accelerometer data. Catellier et al. examined an EM algorithm and JM MI approach that both made use of information from valid days to help impute invalid days. They showed that these two imputation methods outperformed complete-case analysis in terms of bias and precision for the imputation of the mean and standard deviation of intensity-weighted minutes of MVPA per day. Lee found that incorporating information from both valid and invalid days in a two-step MI approach further improved performance. Compared to methods that did not incorporate information available from invalid days, the two-step approach yielded unbiased estimates of average counts per minute with lower estimation error.

However, the methods proposed by Catellier et al. and Lee used imputation to only obtain summary measures from incomplete accelerometer data, such as intensity-weighted minutes of MVPA per day or average counts per minute. Methods that only impute summary measures are simpler to execute because considerations do not have to be made for zero inflation, autocorrelation, or multilevel structure among activity counts across time. These types of imputations are not as useful when we are interested in results beyond summary measures, such as an individual's pattern of PA throughout the day. This caveat created a need for methods of imputation that filled in the actual accelerometer counts over all epochs during non-wear time.

Lee and Gill (2016) defined a minute-level FCS imputation approach based on a mixture of zero-inflated Poisson and Log-normal distributions that was inspired by the accelerometer data structure. Also called the ZIPLN model, this imputation model was shown to effectively handle the zero-inflated and autocorrelated nature of multivariate accelerometer count data (see the Appendix for additional detail on the ZIPLN model).

Lee and Gill demonstrated the importance of including lag and lead variables in the imputation model for better performance. Through correlation heat maps, they displayed the autoregressive correlation structure present in the data, which confirmed the need for only a finite number of lag and lead effects to properly capture the longitudinal relationships among activity counts. By examining residual correlation heat maps, as well as through comparisons of a test statistic developed by Srivastava and Yanagihara (2010), Lee and Gill found that the ZIPLN model with 3 lag and 3 leads effects was most effective in comparison to other ZIPLN models with different numbers of lags and leads, as well as two ZIP models and a zero-inflated Negative Binomial model that included some lags and leads. Lee and Gill additionally assessed prediction accuracy for wear time, and imputation accuracy for non-wear time through comparisons of root mean squared error and mean area difference. The ZIPLN model including both lags and leads outperformed all other models considered.

Lee and Gill created the package *accelmissing* in R to help facilitate easier implementation of the ZIPLN imputation model for researchers with their own accelerometer data. Computational details for the ZIPLN imputation model are found in the Appendix.

3.3 Issues and extensions

The Active Play accelerometer data set has observations from 7 days nested within individuals, which can be considered multilevel or longitudinal data. Failing to account for similarity in activity level across days from the same individual during imputation will lead to underestimation of longitudinal correlations across activity counts, and may not properly recover an individual's true pattern of physical activity (Allison, 2002). Lee and Gill (2016) effectively treated all days independently, as if each day of daily activity counts in the data set came from a unique individual, which made the data appear as if it was seven times the true sample size. The lag and lead effects included in their models incorporated dependency among activity counts within days, but not dependency among activity counts across days from the same individual, likely leading to an underestimation of variance. There are a few ways to try and incorporate dependence across days from the same individual. One option is to include a random effect corresponding to individual in the imputation model. However, there are no easily implemented packages in R for imputation using zero-inflated Poisson mixed models that we could find, especially for FCS imputation. Another option is to use a two-fold, doubly iterative, FCS imputation method with lag and leads effects both within and across days, similar to the method proposed by Nevalainen et al. (2009). Given the size of our data set, the number of epochs, and the number of missing data scenarios we wanted to consider, this approach would have likely been highly computationally intensive and time consuming. We decided to stick with a singly iterative FCS method, and to instead try including a simple time-varying composite measure in the imputation model that incorporated some information about an individual's unique activity habits across the 7 days. A summary measure of average activity level on weekdays and weekends, broken down into 4 time periods over the day, was chosen. This created 8 different

average activity levels for each individual, corresponding to average morning activity level (9:00am – 9:30am), average early afternoon activity level (9:31am – 2:59pm), average late afternoon activity level (3:00pm – 5:58pm), and average evening activity level (6:00pm – 9:00pm), both on weekdays and weekends. The time intervals were chosen based on time-of-day cut points from the Active Play Study, and our choice to be consistent with Lee and Gill and only impute from 9:00am – 9:00pm. Time-varying covariates were not originally permitted in the imputation function from the *accelmissing* package, so we adjusted their function to allow for these covariates, as well as to allow for imputation every 15 seconds instead of every minute, and to include both lag and lead effects in the zero inflation portion of the model opposed to a single lag effect (see the Appendix for details on the ZIPLN imputation procedure developed by Lee and Gill). We also decided to focus on ZIP and ZIPLN models to account for the zero inflation in the data, as zero-inflated Negative Binomial models often require substantial sample size (Kleinke & Reinecke, 2013), and we ran into estimation problems when trying to fit them to our 332×2880 data set. Finally, we sought to assess a more comprehensive set of imputation methods and missingness scenarios compared to what was shown by Lee and Gill.

4 Data Application

The Active Play study is an ongoing cross-sectional study that is being conducted at Dr. Ian Janssen's Physical Activity Epidemiology Lab here at Queen's University. The main purpose of the study is to assess and describe the characteristics of children's active play. 444 children and youth (222 boys and 222 girls) aged 10-13 from Kingston were sampled, and each study participant wore an Actical accelerometer for 7 days to collect data on their physical activity with the goal of measuring active play. All study participants were instructed to wear the

accelerometer for 24 hours each day for the 7 days following their preliminary lab visit. Each participant additionally filled out a computer-based questionnaire, had physical measurements taken, and was given a log sheet to record supplementary information on their daily activity such as wakeup time, bedtime, and any additional periods of accelerometer removal. Parents of participants also completed a computer-based questionnaire. Table 4.1 provides a summary of the variables from the Active Play Study. Some of the data collected from this study was used to evaluate the performance of three imputation models on a real accelerometer data set. In this report we focused on five variables: activity count, sex, age, race, and BMI.

Table 4.1: Summary of the variables from the Active Play Study data (N=332 with 7 days of data per individual). Bolded variables were considered in this report.

Variable		Mean (SD) or n (%)
Activity count (awake wear time)		83.27 (249.40)
Sex	Male	167 (50.3%)
	Female	165 (49.7%)
Age		11.85 (1.16)
Race	White	283 (85.2%)
	Other	49 (14.8%)
BMI z-score		0.017 (1.00)
SBP z-score		0.012 (1.00)
Health	No health condition	301 (90.7%)
	Health condition	31 (9.3%)
Family status	Double parent	277 (83.4%)
	Single parent	55 (16.6%)
Parent education	High school education or less	29 (8.7%)
	2-year college diploma	100 (30.1%)
	4-year college diploma or	203 (61.1%)
Fast food	Rarely or never	113 (34.0%)
	1-2 times per month	169 (50.9%)
	1+ times per week	50 (15.1%)
Weekly snacking frequency		4.23 (3.92)
Maturity		186.06 (30.30)
Annual family income	No answer	49 (14.8%)
	< $50,000	60 (18.1%)
	$50,000 - $100,000	84 (25.3%)
	> $100,000	139 (41.9%)
Season	Winter	121 (36.4%)
	Spring	83 (25.0%)
	Summer	52 (15.7%)
	Fall	76 (22.9%)

As is typical of accelerometer data, there were a number of periods of non-wear time in the study. Non-wear time was identified as periods of zero counts longer than 60 minutes (240 epochs), and any periods of accelerometer removal recorded by participants in their log sheets. Table 4.2 provides a breakdown of the proportion of missingness during awake time by day of the week, as well as the overall proportion of missingness in the entire data set (daily awake time was specified in participant log sheets). Table 4.3 gives the missing rate across different levels of the four variables considered.

Table 4.2: Missing rate by day (awake time).

Day of the week	Missing rate (%)
Sunday	8.94
Monday	7.30
Tuesday	6.41
Wednesday	6.14
Thursday	5.86
Friday	8.30
Saturday	9.43
Overall	7.48 (SD = 0.014)

Table 4.3: Missing rate by variable (awake time).

Variable		Missing rate (%)
Sex	Male	7.80
	Female	7.14
Age	[10, 11)	5.54
	[11, 12)	7.26
	[12, 13)	7.12
	[13, 14]	10.72
Race	White	6.79
	Other	11.50
BMI z-score	< 0	6.82
	> 0	8.48

The proportion of missingness was similar across days, with an average of 7.48% and a standard deviation of 0.014%. The missing rate among males and females was similar, with a slightly higher missing rate among males. Non-wear also tended to be greater among those individuals who were older, of non-white race, and of higher than average BMI.

It is interesting to explore the relationship between non-wear and time of day. Figure 4.1 displays the proportion of individuals who were wearing their accelerometer between 12:00am – 12:00am across all days, as well as for weekends only, and weekdays only. The wearing proportion in the sample always remained high (above 85%).

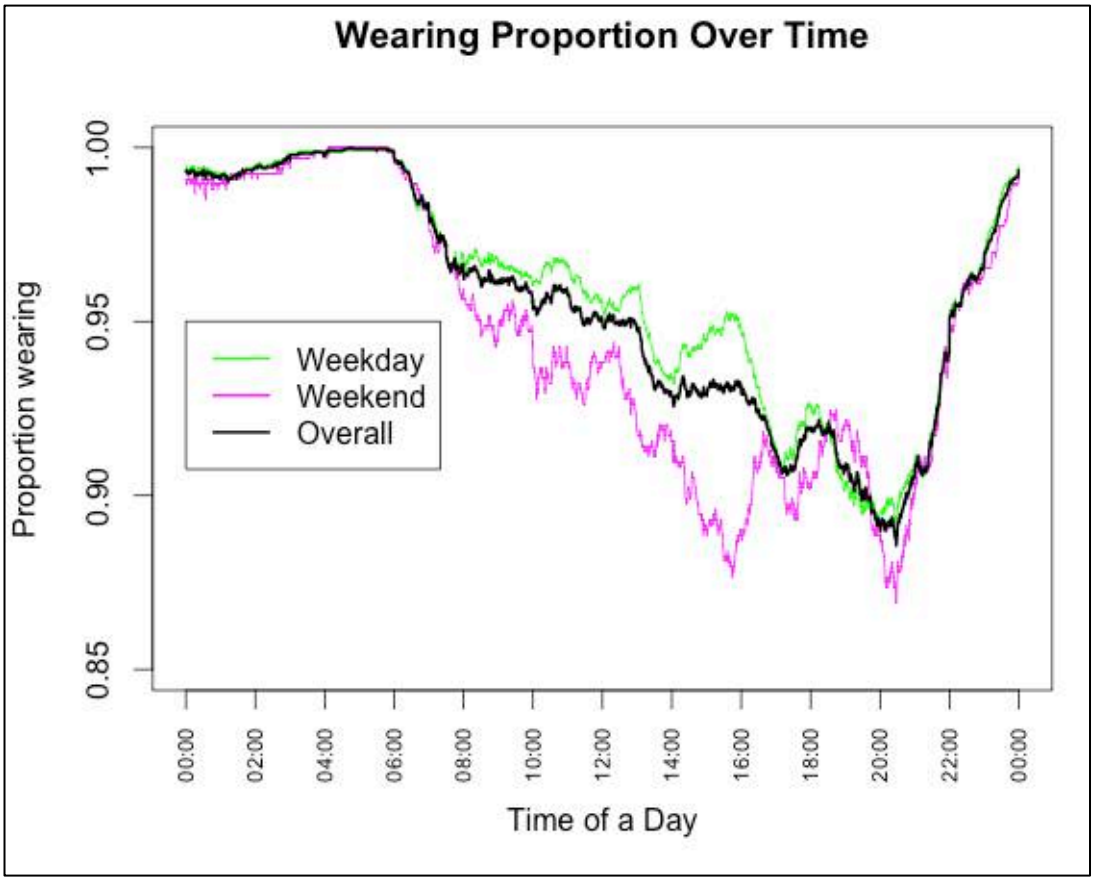

Figure 4.1: Proportion of individuals who were wearing their accelerometer over time (12:00am – 12:00am) across all days (black), and stratified by weekdays (green) and weekend (pink).

Next, the level of autocorrelation in the observed data was examined. The autocorrelation can be conceptualized in two ways: the correlation between activity counts from epochs close in time within days, and the correlation between activity counts across days. Figure 4.2 displays correlation heat maps corresponding to these two conceptualizations at specific times.

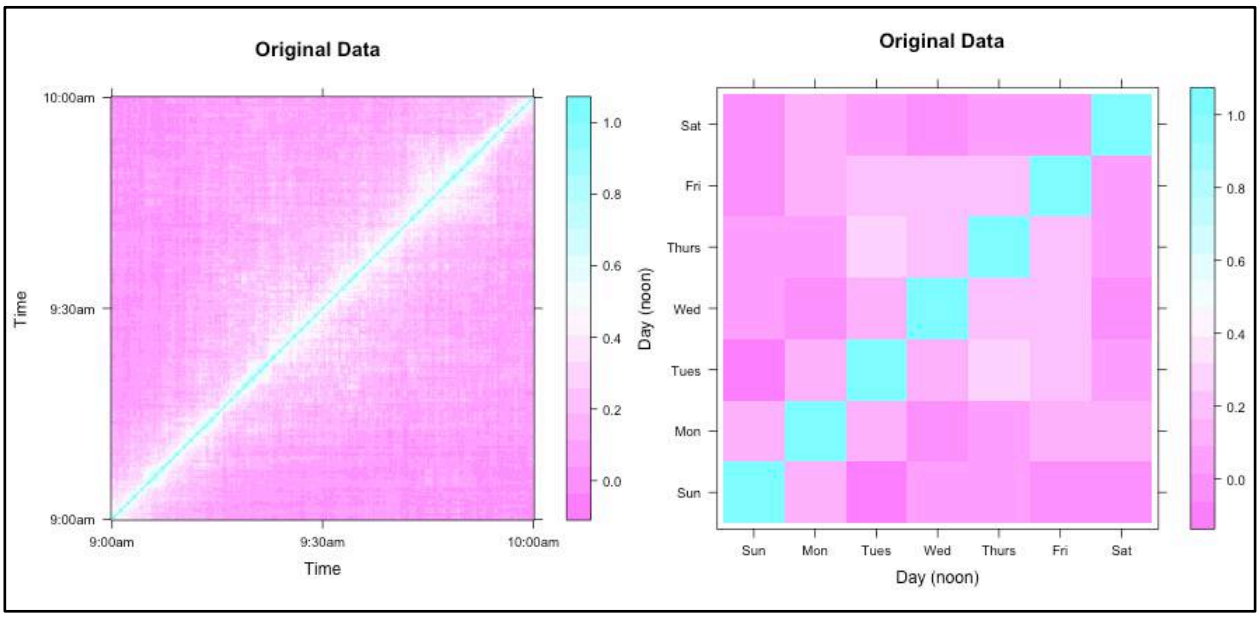

Figure 4.2: Correlation heat maps of activity counts within days from 9:00am – 10:00am across all individuals (left plot), and of activity counts across days at 12:00pm across all individuals (right plot).

The plot on the left of Figure 4.2 displays the correlations across epochs from 9:00am – 10:00am within days. The diagonal blue line represented the perfect correlation of an epoch with itself. There was a clear banding correlation structure on either side of the diagonal, suggesting that the strongest correlations were between the epochs close together in time (shades of blue, white, and light pink in the graph), and that correlations weakened between epochs as they became further apart in time (darker shades of pink). This was indicative of an autoregressive correlation structure, and was similarly shown by Lee and Gill (2016).

The plot on the right of Figure 4.2 displays the correlations across days at a single epoch, 12:00pm (noon). Again, the diagonal blue squares represented the perfect correlation of activity counts on a single day at 12:00pm with itself. The off-diagonal squares represented the strength of activity count correlations across days, with shades of lighter pink indicating stronger correlations, and shades of darker pink indicating weaker correlations. By the lighter shaded 4×4 square in the center of the plot, it appeared that Tuesday – Friday had the strongest activity count correlations, suggesting that activity levels during weekdays were more similar. Monday also notably had moderate correlations with Tuesday, Friday, Saturday, and Sunday, and weekdays tended to be only weakly correlated with weekends.

Next, ZIP and ZIPLN models using sex, age, race, BMI z-score, and a weekend/weekday indicator were fit to the data to assess model fit and prediction accuracy. The ZIP model was considered both with and without the time-varying (T-V) average activity level covariate, and K = 1, 2, 3, 4, 5, 10, and 20 lag and lead effects were added to the zero-inflation and activity count portions of each model. We additionally included two types of mean imputation for comparison – imputation using the overall mean activity count from wear time (the grand mean method), and imputation using epoch-specific mean activity counts from wear time (the epoch mean method). Model fit was assessed in two ways based on residual correlation – graphically and using a test statistic. Prediction accuracy was assessed based on wear time, as the true activity counts during non-wear time were unknown.

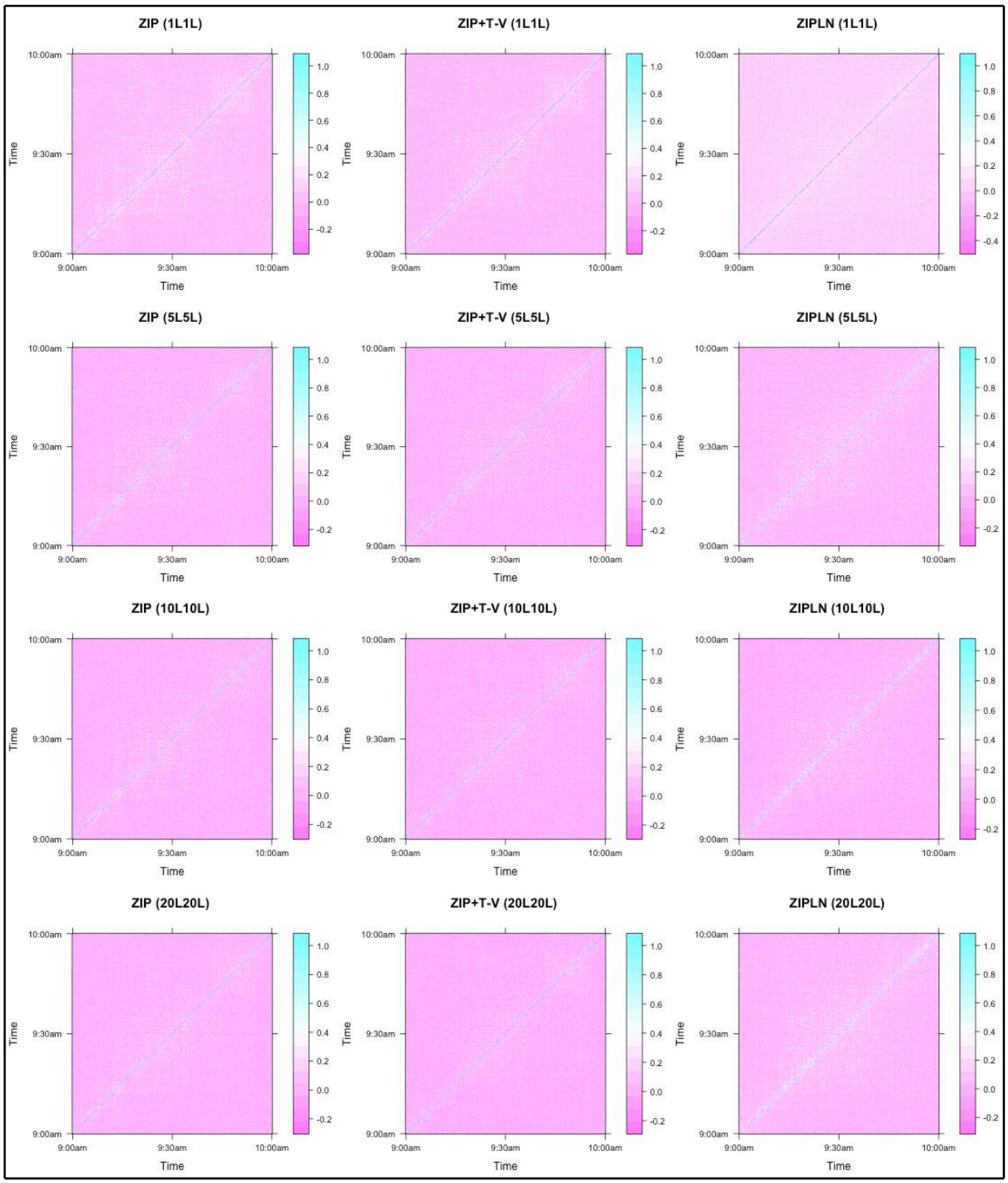

Figure 4.3: Correlation heat maps of model residuals within days from 9:00am – 10:00am.

If the model fits the data well, one would expect any autocorrelation to be removed from the model residuals, and the correlation matrix to resemble an identity matrix (J. A. Lee & Gill, 2016). Figure 4.3 displays the residual correlation heat maps within days from 9:00am – 10:00am. K = 1, 5, 10, and 20 lags and leads for the ZIP model without the T-V covariate, the ZIP model with the T-V covariate, and the ZIPLN model are shown. Visually, there is no clear banding correlation structure in any of the plots, and all models seemed to have effectively removed most of the autocorrelation from the residuals. There may have been some block-diagonal positive correlation structure in the ZIP model with K = 1, and weak overall correlations in the ZIPLN model with K = 1, but these correlations seemed to disappear as more lag and lead effects were added. Residual correlation could be evaluated further with a simple test statistic tailored to high-dimensional data. Schott (2005) proposed a test for the complete independence of a high-dimensional sample correlation matrix. Assuming that the model residuals were asymptotically normal, we could apply this test to determine if they were independent, which is equivalent to the correlation matrix being diagonal. Following the notation of Schott (2005) if ρ_{ij} is the $(i,j)th$ element of the correlation matrix, the null hypothesis can be written as $H_o: \rho_{ij} = 0\ (i > j)$. The statistic for this test under the null hypothesis with mean 0 is given by $t_{nm} = \sum_{i=2}^{m}\sum_{j=1}^{i-1} r_{ij}^2 - \frac{m(m-1)}{2n}$, with variance equal to $\sigma_{t_{nm}}^2 = var(t_{nm}) = \frac{m(m-1)(n-1)}{n^2(n+2)}$. The value r_{ij} is the $(i,j)th$ element of the sample correlation matrix, n is the sample size $N - 1$, and m is the number of epochs being considered. In the Active Play Study, looking at the correlations within days from 9:00am – 9:00pm, $n = 332(7) - 1 = 2323$ and $m = 2880$ (the number of 15 second epochs from 9:00am – 9:00pm). Across days, $n = 332 - 1 = 331$ and $m = 2880(7) = 20160$. If the complete-

independence hypothesis holds, then the z-score $\frac{t_{nm}}{\sigma_{t_{nm}}}$ will have mean 0 and variance 1. Table 4.4 gives the test results and associated p-values for all the models considered under the two conceptualizations of correlation.

Table 4.4: Test to assess the complete independence of the PA count residuals resulting from the different modeling procedures. The null hypothesis is zero correlation in all entries of the residual correlation matrix off the diagonal. A smaller test statistic indicates that the complete-independence hypothesis holds, meaning that the model more effectively removes the autocorrelation from the residuals. The smallest test statistic of each model considered is bolded.

		9:00am – 9:00pm Within days		9:00am – 9:00pm Across days	
Model	K	z-score	p-value	z-score	p-value
Grand mean	–	**27892.180**	< 0.001	**1215510**	< 0.001
Epoch mean	–	**27892.180**	< 0.001	**1215510**	< 0.001
ZIP	1	2141.259	< 0.001	4847.299	< 0.001
	2	1957.836	< 0.001	4781.673	< 0.001
	3	1857.861	< 0.001	4738.793	< 0.001
	4	1805.728	< 0.001	4714.434	< 0.001
	5	1770.050	< 0.001	4693.206	< 0.001
	10	1650.318	< 0.001	4618.344	< 0.001
	20	**1466.208**	< 0.001	**4471.558**	< 0.001
ZIP + T-V covariate	1	1820.483	< 0.001	4667.431	< 0.001
	2	1701.057	< 0.001	4624.222	< 0.001
	3	1636.284	< 0.001	4593.284	< 0.001
	4	1601.203	< 0.001	4574.797	< 0.001
	5	1576.382	< 0.001	4556.944	< 0.001
	10	1486.963	< 0.001	4494.266	< 0.001
	20	**1341.843**	< 0.001	**4367.147**	< 0.001
ZIPLN	1	1964.185	< 0.001	4643.918	< 0.001
	2	**1940.152**	< 0.001	**4608.705**	< 0.001
	3	2008.378	< 0.001	4608.992	< 0.001
	4	2071.747	< 0.001	4615.309	< 0.001
	5	2117.021	< 0.001	4624.866	< 0.001
	10	2256.939	< 0.001	4708.269	< 0.001
	20	2291.218	< 0.001	4856.871	< 0.001

Although the p-values were all very small, which indicated that the null hypothesis $H_o: \rho_{ij} = 0$ did not hold, we were more interested in comparing the test statistics relatively to assess which

model leads to the least amount of remaining residual correlation. As expected due to poor fit to the true data, both mean imputation methods had very large test statistic values compared to those from the ZIP and ZIPLN models, indicating that neither method effectively reduced the amount of autocorrelation among the residuals. For the ZIP models considered, the test statistic results indicated that adding more lag and leads effects to the model corresponded to a smaller amount of remaining residual correlation. This makes sense intuitively, as incorporating more information about the longitudinal correlations among activity counts within days would be expected to yield more accurate predictions of the true activity level at each epoch, and therefore reduce the amount of autocorrelation among the residuals. Surprisingly, the ZIPLN model did not display the same relationship of smaller test statistics with more lags and leads. For both conceptualizations of correlation, the ZIPLN model with $K = 2$ lags and leads yielded the smallest test statistic. Lee and Gill similarly found a point at which additional lags and leads did not help in removing autocorrelation from the residuals, however, their results indicated that the ZIPLN model with $K = 3$ lags and leads was optimal. Within days, the ZIP+T-V covariate model produced the smallest test statistic values for a given number of lags and leads, and this was believed to be attributable to the incorporation of extra information about activity level habits across days. Across days, the ZIPLN model had the smallest test statistic values out of all three models for $K = 1$ and 2 lags and leads, although they were similar to those from the ZIP+T-V covariate model. With $K = 3$ or more lags and leads, the ZIP+T-V covariate model again produced the smallest test statistic values.

Next we assessed the prediction accuracy for wear time. This was done based on the idea that if a model predicts well for wear time, then it should also predict well for non-wear time under the

MAR assumption (J. A. Lee & Gill, 2016). The prediction accuracy was assessed using root mean squared error (RMSE) and mean area difference (MAD). RMSE was calculated as $RMSE = \left[\sum_{t\in T}\sum_{i\in N}\frac{(Y_{it}-\hat{Y}_{it})^2}{|T||N|}\right]^{\frac{1}{2}}$, where Y_{it} was the true activity count for all epochs t classified as wear time, $\hat{Y}_{it}$ was the corresponding predicted activity count from each model, T was the number of wear time epochs, and N was the number of days from all individuals. MAD was calculated as $MAD = \sum_{t\in T}\sum_{i\in N}\frac{|X_i(t)-\hat{X}_i(t)|}{|T||N|}$, where $X_i(t)$ was the predicted value from a B-spline fit to the true data for all epochs t classified as wear time, and $\hat{X}_i(t)$ was the corresponding predicted value from a B-spline fit to the predicted data from each model. Table 4.5 gives the RMSEs and MADs calculated for wear time for each model considered. Smaller values of both measures indicated better prediction accuracy.

Table 4.5: Comparison of prediction accuracy for wear time. A smaller RMSE and MAD indicate better prediction accuracy. The smallest RMSE/MAD of each model considered is bolded.

Model	K	RMSE	MAD
Grand mean	–	**260.4249**	**105.7426**
Epoch mean	–	**259.5225**	**104.1045**
ZIP	1	164.1544	21.35827
	2	163.1579	21.11089
	3	161.9195	21.20378
	4	161.1037	21.26250
	5	160.3945	21.27973
	10	157.2554	21.08272
	20	**151.2810**	**20.33611**
ZIP + T-V covariate	1	162.3844	20.88467
	2	161.5267	20.63484
	3	160.4230	20.69978
	4	159.6658	20.73526
	5	158.9924	20.73831
	10	155.9378	20.51768
	20	**150.0567**	**19.79549**
ZIPLN	1	**163.3765**	**27.05473**
	2	166.7019	30.84502
	3	168.5127	33.32196
	4	169.6644	34.64452
	5	170.4528	35.50286
	10	172.5439	37.40369
	20	174.0313	38.00704

Results in Table 4.5 were similar to those seen in Table 4.4. Again, prediction accuracy was poor for the two mean imputation methods. For both ZIP models, including more lag and leads effects produced smaller RMSEs and MADs, indicating better prediction accuracy. The ZIP+T-V covariate model also produced the smallest values of RMSE and MAD out of all three models for a given number of lags and leads. Reasons for these results are expected to be similar to those hypothesized for the Table 4.4 results. Out of the ZIPLN models considered, the model with K = 1 lag and lead had the smallest RMSE and MAD, and this result was congenial with what was seen from Lee and Gill.

Figure 4.4 displays the Wednesday activity counts from 9:00am – 9:00pm for individual 5 in the Active Play Study. Three prediction curves with numbers of lags and leads that produced the smallest RMSE and MAD for the ZIP and ZIPLN model are shown. The prediction curve from the grand mean and a curve fit to the true data were also included for comparison.

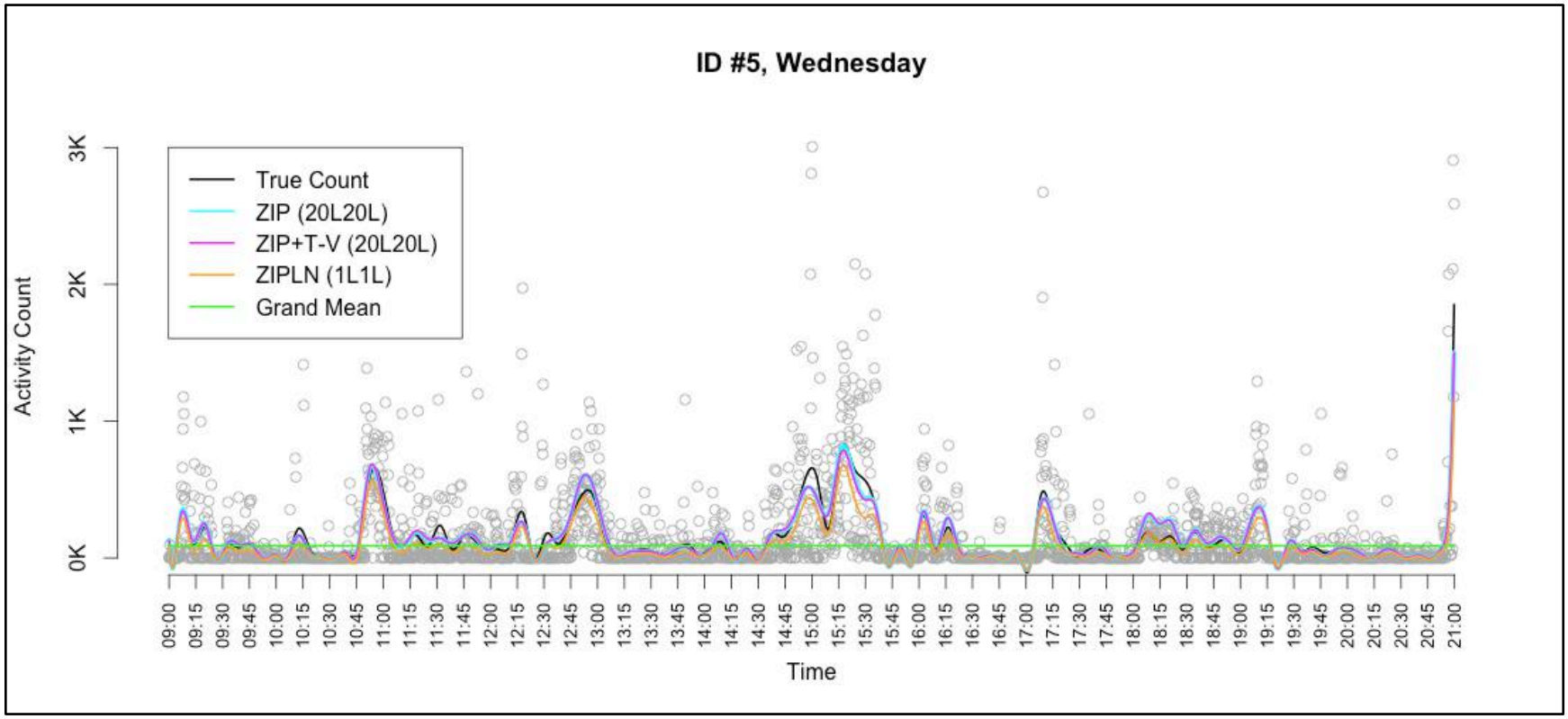

Figure 4.4: Activity count plot from 9:00am – 9:00pm on Wednesday for individual 5. Overlaid splines correspond to the true activity counts, the grand mean predicted activity counts, and the predicted activity counts from the ZIP and ZIPLN models with lags and leads that produced the smallest RMSE and MAD from Table 4.5.

The prediction curves for the ZIP and ZIP+T-V covariate models overlapped with the true count curve for the majority of the epochs, however, there were some intervals where both models overestimated the activity count. This was notably seen between 6:00pm – 6:30pm. The ZIPLN model prediction curve also followed the true count curve reasonably well, but tended to be lower than the two ZIP model curves and occasionally underestimated peaks of activity counts. These underestimations and overestimations were similarly seen across all individuals and days in the data set.

5 Observed Properties of Multiple Imputation Methods

A pseudosimulated complete data set was created in order to assess the imputation accuracy for non-wear time of the three models considered. The goal of this analysis was to compare how well the different models recovered the activity counts during non-wear time, and if they lead to accurate estimation of overall daily average time spent in MVPA (a common indicator used to assess population health (Statistics Canada, 2014)). As previously mentioned, imputation performance is affected by four factors: the number of covariates included in the imputation model(s), the covariate correlations with the incomplete variable(s), the amount of missingness in the data, and the missing data mechanism at work (Catellier et al., 2005). Holding the first two factors constant, we compared the imputation performance of our models under six missing data scenarios corresponding to combinations of the three missing data mechanisms (MCAR, MAR and MNAR) and two proportions of non-wear time in the data (10% and 20%).

As it would be very challenging to accurately capture variable relationships in simulating high-dimensional accelerometer data from scratch, we instead made use of the completely observed

days from the individuals in the Active Play Study. To create a pseudosimulated complete data set, we subsetted the Active Play Study data to completely observed days from individuals who had at least one completely observed weekday and one completely observed weekend day. After subsetting the data, we were left with a total of 803 days from 182 individuals, with varying numbers of days per individual, which constituted approximately 35% of the original sample. Table 5.1 displays the summaries of the variables considered for this reduced sample. Variable distributions were similar to those seen in the original Active Play Study sample (Table 4.1).

Table 5.1: Summary of the variables from the complete data (N=182 with varying days per individual).

Variable		Mean (SD) or n (%)
Sex	Male	89 (48.9%)
	Female	93 (51.1%)
Age		11.62 (1.09)
Race	White	160 (87.9%)
	Other	22 (12.1%)
BMI z-score		-0.016 (0.89)

In order to create pseudosimulated incomplete data sets with missingness similar to that observed in the Active Play Study, we selected the lengths of the simulated non-wear periods based on random draws from the lengths of the non-wear periods that arose in the Active Play Study. The placement of the non-wear periods in time was selected using random draws from a Binomial distribution. If a 1 was drawn at a particular epoch, this indicated the start of a non-wear period, and subsequent epochs were also set as missing for the associated randomly drawn non-wear period length. Under the MCAR assumption, Binomial probabilities were set to yield approximately 10% and 20% missingness. Under the MAR assumption, Binomial probabilities

were modeled using logistic regressions with covariates sex, age, BMI z-score dichotomized at 0, and a weekend/weekday indicator. Regression parameters were selected to reflect the missingness trends in the real data. From Tables 4.2 and 4.3, males, older individuals, individuals with an above average BMI, and weekends tended to have more missingness. Based on these observations, parameters were set as $\beta_1 = -0.25$, $\beta_2 = 0.1$, $\beta_3 = 0.5$, $\beta_4 = 0.5$, and β_0 was varied to yield approximately 10% and 20% missingness. Under the MNAR assumption, an indicator covariate for activity count exceeding 375 was additionally included in the logistic regression, as this was the MVPA cutoff used in the Active Play Study. We wanted to make it so that greater activity counts were associated with a higher probability of missingness, so $\beta_5 = 1$. Even though individuals of non-white race also tended to have more missingness in the original sample, race was not included in the logistic regression models as the pseudosimulated complete data set only contained 22 individuals of non-white race. Because of this small representation, we ran into issues trying to fit the imputation models when individuals of non-white race were set to have a higher probability of missingness. These issues occurred when there were not enough individuals of non-white race without missingness to inform the models at each time point.

Using the pseudosimulated complete data, imputation accuracy could be assessed for non-wear time. K = 1 and K = 2 lags and leads were of interest from the results seen in Tables 4.4 and 4.5, and for the sake of computational time, K = 5 was selected to show the effects of adding more lags and leads. Similar to Table 4.5, Table 5.2 gives the RMSEs and MADs calculated for non-wear time from each model considered. m = 5 imputed data sets were created per ZIP and

ZIPLN imputation, and RMSE and MAD were calculated for each imputed data set and averaged to yield final measures. Smaller values of both measures indicated better imputation accuracy.

Table 5.2: Comparison of imputation accuracy for non-wear time. A smaller RMSE and MAD indicated better imputation accuracy. The smallest RMSE/MAD of each model considered is bolded.

			10% non-wear		20% non-wear	
Missing data mechanism	Model	*K*	RMSE	MAD	RMSE	MAD
MCAR	Grand mean	–	**195.3844**	**100.6665**	**205.3645**	**103.8683**
	Epoch mean	–	**186.6050**	**99.51552**	**197.4076**	**102.3259**
	ZIP	1	222.4692	113.8935	231.9786	119.5379
		2	222.4536	114.3544	233.7072	117.2954
		5	**219.6213**	**113.0595**	**231.0787**	**115.8108**
	ZIP + T-V covariate	1	217.5453	109.9382	229.0052	113.1865
		2	217.7598	110.6880	229.2803	113.5857
		5	**216.3616**	**109.7137**	**227.8501**	**112.7657**
	ZIPLN	1	213.2742	101.6670	224.4635	104.9852
		2	209.9684	100.3232	221.5449	103.8871
		5	**204.7788**	**97.7128**	**217.1748**	**101.5615**
MAR	Grand mean	–	**210.2115**	**106.0366**	**207.1119**	**106.5767**
	Epoch mean	–	**202.8043**	**104.0816**	**199.1993**	**104.7210**
	ZIP	1	233.7599	121.4655	233.3182	122.8718
		2	233.7598	122.3791	232.8965	122.7449
		5	**230.6686**	**120.3285**	**230.6200**	**121.4650**
	ZIP + T-V covariate	1	228.1848	117.4008	228.0959	**118.4106**
		2	228.3489	118.3094	228.5449	119.2088
		5	**226.9460**	**117.1736**	**227.3393**	118.4228
	ZIPLN	1	224.2858	108.1475	223.6355	109.6924
		2	220.8827	106.6043	220.7249	108.5830
		5	**215.5849**	**103.9496**	**216.4299**	**106.4039**
MNAR	Grand mean	–	**210.2532**	**107.6144**	**214.5714**	**107.7390**
	Epoch mean	–	**202.3379**	**105.5409**	**207.4516**	**105.9753**
	ZIP	1	238.1671	123.9644	239.7490	119.5595
		2	238.1594	124.5626	239.4154	119.9372
		5	**235.3630**	**122.7930**	**237.3780**	**118.8039**
	ZIP + T-V covariate	1	233.4081	**120.1388**	235.0584	**116.1131**
		2	233.9562	120.9621	235.7049	116.8597
		5	**232.4803**	120.3882	**234.7905**	116.2046
	ZIPLN	1	229.4246	111.2659	230.5101	108.5054
		2	226.4174	110.1162	227.6755	107.3741
		5	**221.5479**	**107.5819**	**223.8308**	**105.5891**

Across all missingness scenarios considered, mean imputation surprisingly outperformed all ZIP and ZIPLN imputation models in terms of both RMSE and MAD. This was expected to be indicative of an issue with the ZIP and ZIPLN models' ability to recover true activity patterns in the data. Focusing on the ZIP and ZIPLN models, under the MCAR and MAR mechanisms, adding more lags and leads lead to better imputation accuracy. For the ZIP+T-V covariate model, more lags and leads lead to improved imputation accuracy under the MCAR mechanism and MAR mechanism with 10% missingness. However, the RMSE and MAD did not agree on the best model under the MAR mechanism with 20% missingness, although the differences in these measures across the different numbers of lags and leads could be considered negligible. Under the MNAR mechanism, adding more lags and leads improved imputation accuracy for the ZIP and ZIPLN models. The RMSE and MAD for the ZIP+T-V covariate model did not agree on the best number of lags and leads, although again, the differences in RMSE and MAD across the different numbers of lags and leads for either amount of missingness could be considered negligible. Overall, the ZIPLN model had the smallest RMSE and MAD for a given number of lags and leads in each missingness scenario compared to the two ZIP models, indicating the best imputation accuracy. This contrasted the results in Table 4.5 for prediction accuracy where the ZIP+T-V covariate model was shown to produce the smallest RMSE and MAD. The ZIP model always had the largest RMSE and MAD in all missingness scenarios, indicating that it had the worst imputation accuracy out of all the methods considered.

Finally, we wanted to assess how well the methods recovered a common summary measure of PA level, namely daily average time spent in MVPA. Daily average time spent in MVPA was calculated as the number of epochs in the data set (both wear time and non-wear time) with

activity counts exceeding 375, divided by the total number days. We additionally divided by 4 to give the measure in terms of minutes, instead of 15-second epochs. This measure was calculated for each of the $m = 5$ imputed data sets from each ZIP and ZIPLN imputation, averaged, and compared with the amount of daily average time spent in MVPA calculated from the pseudosimulated complete data set. We additionally included daily average time spent in MVPA calculated from the wear time only under each missingness scenario, effectively ignoring the non-wear periods. This method was equivalent to the "averaging" approach discussed in Section 1.3. Table 5.3 gives the results of these calculations, as well as the difference between the measure derived from the imputed data and the "true" measurement. A smaller amount of error indicated better estimation accuracy.

Table 5.3: Comparison of estimation accuracy for daily average time spent in MVPA. A smaller amount of error indicated better estimation accuracy. The smallest errors from each model considered, and the associated MVPA estimate, are bolded.

"True" daily average time spent in MVPA = 59.2092						
			10% non-wear		20% non-wear	
Missing data mechanism	Model	*K*	MVPA	Error	MVPA	Error
MCAR	Wear time	–	**59.4422**	**-0.2330**	**64.2104**	**-5.0012**
	Grand mean	–	**53.1055**	**6.1037**	**47.4486**	**11.7606**
	Epoch mean	–	**53.1055**	**6.1037**	**47.4486**	**11.7606**
	ZIP	1	60.1560	-0.9468	60.6768	-1.4676
		2	60.0184	-0.8092	59.9972	-0.7880
		5	**59.4452**	**-0.2360**	**58.8366**	**0.3727**
	ZIP + T-V covariate	1	59.7948	-0.5856	59.5928	-0.3836
		2	59.7176	-0.5084	**59.3920**	**-0.1828**
		5	**59.1980**	**0.0112**	58.4502	0.7590
	ZIPLN	1	**57.7120**	**1.4972**	**55.7173**	**3.4919**
		2	57.2939	1.9153	54.9969	4.2123
		5	56.6164	2.5928	53.9079	5.3014
MAR	Wear time	–	**59.2498**	**-0.0406**	**60.9936**	**-1.7844**
	Grand mean	–	**53.0688**	**6.1404**	**46.2983**	**12.9110**
	Epoch mean	–	**53.0688**	**6.1404**	**46.2983**	**12.9110**
	ZIP	1	59.9956	-0.7864	60.4181	-1.2089
		2	59.8808	-0.6715	60.0648	-0.8555
		5	**59.2316**	**-0.0224**	**58.8820**	**0.3272**
	ZIP + T-V covariate	1	59.6031	-0.3938	59.6140	-0.4047
		2	59.5146	-0.3054	**59.4517**	**-0.2425**
		5	**59.0100**	**0.1993**	58.4894	0.7198
	ZIPLN	1	**57.6096**	**1.5996**	**55.3929**	**3.8163**
		2	57.1809	2.0283	54.6846	4.5246
		5	56.54732	2.6619	53.60772	5.6015
MNAR	Wear time	–	**69.6227**	**-10.4135**	**67.3561**	**-8.1469**
	Grand mean	–	**52.7404**	**6.4689**	**46.0943**	**13.1149**
	Epoch mean	–	**52.7404**	**6.4689**	**46.0943**	**13.1149**
	ZIP	1	59.74689	-0.5377	**58.49377**	**0.7154**
		2	59.59215	-0.3829	58.21575	0.9935
		5	**58.98443**	**0.2248**	57.17341	2.0358
	ZIP + T-V covariate	1	59.39601	-0.1868	**57.90193**	**1.3073**
		2	**59.29981**	**-0.0906**	57.78051	1.4287
		5	58.83655	0.3727	56.88107	2.3281
	ZIPLN	1	**57.27709**	**1.9321**	**54.17933**	**5.0299**
		2	56.8929	2.3163	53.45268	5.7565
		5	56.23506	2.9742	52.43524	6.774

Mixed results were seen for which models most accurately estimated daily average time spent in MVPA. Mean imputation consistently performed poorly, producing large underestimations. This was to be expected as both mean imputation methods imputed values below the 375 MVPA cutoff for all non-wear epochs, which greatly decreased the amount of MVPA seen in the data. This was especially the case under the MNAR mechanism. The wear-time-only method had varying performance with good estimations produced under the MCAR and MAR mechanisms with 10% missingness. However, it did not do well under the MCAR and MNAR mechanisms with 20% missingness, and under the MNAR mechanism with 10% missingness it produced a poorer estimate than those from the mean imputation methods. From the ZIP and ZIPLN models, under the MCAR mechanism the ZIP and ZIP+T-V covariate models produced close results, with the ZIP+T-V model often leading to slightly more accurate estimation. The ZIPLN model under the MCAR mechanism produced noticeably worse estimates than the two ZIP models that underestimated the amount of daily average time spent in MVPA, especially when the amount of missingness was doubled, and adding more lags and leads did not improve the estimation accuracy. This was likely due to the ZIPLN model's tendency to underestimate peaks in activity level, as seen in Figure 4.4. Similar results were seen for the MAR and MNAR mechanisms with 10% missingness, and doubling the amount of missingness often lead to worse estimation accuracy for all three models. Additionally, under the MNAR mechanism with 20% missingness, the ZIP model surprisingly produced the smallest estimation errors even though it had the worst imputation accuracy from Table 5.2. The inconsistencies in results seen between Tables 5.2 and 5.3 indicated that imputation accuracy, measured in terms of RMSE and MAD, was not a good measure of ability to recover information about PA level summary statistics, such as daily average time spent in MVPA.

Figure 5.1 displays the Tuesday activity counts from 9:00am – 9:00pm for individual 3 in the Active Play Study. Three curves fit to the imputed data with numbers of lags and leads that produced the smallest RMSE and MAD for each ZIP and ZIPLN model under the MNAR mechanism with 20% missingness are shown. A curve fit to the grand mean imputed data under the MNAR mechanism with 20% missingness and a curve fit to the true data were also included for comparison.

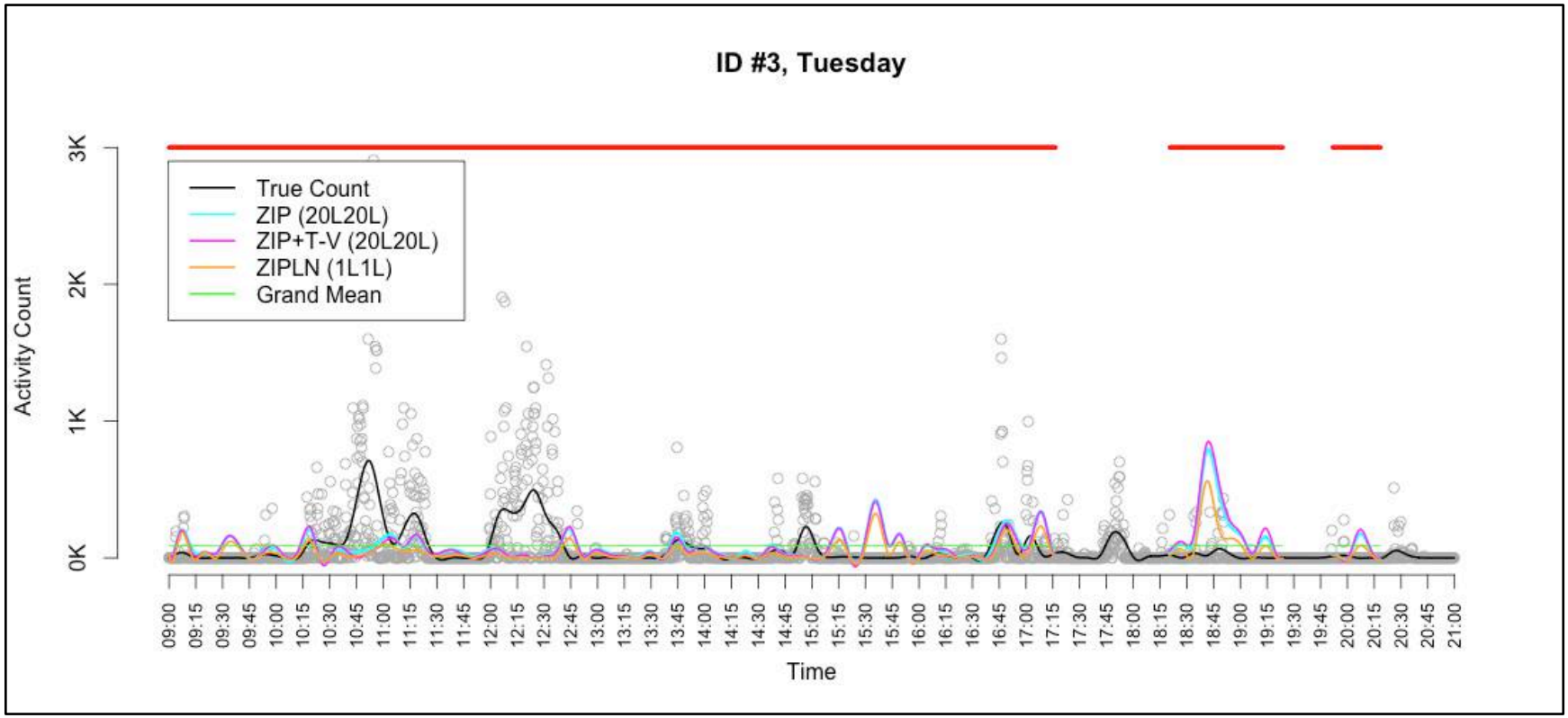

Figure 5.1: Activity count plot from 9:00am – 9:00pm on Tuesday for individual 3. Red bars at the top mark the periods of non-wear time. Overlaid splines correspond to the true activity counts, the imputed activity counts from the ZIP and ZIPLN models with lags and leads that produced the smallest RMSE and MAD from Table 5.2 under the MNAR mechanism with 20% missingness, and the grand mean imputed activity counts under the MNAR mechanism with 20% missingness.

The curves from the ZIP and ZIPLN models only poorly tracked the true count curve during non-wear time. There were many discrepancies between the peaks in the imputed data and the peaks in the true data, indicating that none of the ZIP or ZIPLN models were able to accurately recover the true pattern of activity, especially for longer periods of non-wear. These models may perform adequately in data sets with intermittent, shorter periods of non-wear, but this would be an unrealistic pattern of missingness in real-life accelerometer data. The grand mean imputation curve remained at approximately 88 count during non-wear time, which was also inaccurate. However, with a consistent imputed value of 88 during the non-wear periods, it better approximated the smaller activity counts in the true data (which were more frequent than large activity counts) and lead to smaller overall discrepancies between the mean imputation curve and the true data curve. These observations helped explain the results seen in Table 5.2 where mean imputation had the best overall imputation accuracy in terms of RMSE and MAD.

6 Conclusions

This report provided a useful contribution to research for the imputation of accelerometer data by considering a more comprehensive set of imputation methods and a wider range of missingness scenarios than previously investigated by other studies. Namely, the accelerometer data imputation methods considered were assessed for imputation accuracy under MCAR, MAR and MNAR missingness mechanisms with two levels of missingness, and time-varying covariates were considered to better account for individual activity habits over time. Additionally, this report provided a comparison of how well the imputation methods recovered a common population measure of PA level.

Overall, with the Active Play Study data, simple mean imputation more accurately imputed the values for non-wear time compared to the ZIP and ZIPLN models considered. This is an important result if interest lies in recovering an individual's true pattern of physical activity during the period the accelerometer was worn, as previously proposed imputation models from the literature were shown to perform poorly at this task. We therefore do not recommend any of the imputation methods considered here for this purpose. However, the ZIP and ZIPLN models had better performance recovering the amount of daily average time spent in MPVA for the sample compared to the mean imputation methods. Results indicated that the two ZIP models best served this purpose, as the ZIPLN model lead to substantial underestimations, likely due to its issues with underestimation of peaks in activity levels in the data. From these observations, we conclude that the goal of the accelerometer data analysis should guide the choice of which imputation method to employ.

Future research directions for the imputation of accelerometer data would include exploration of the use of mixed-effects FCS imputation models to account for dependency across days within individuals, and exploration of imputation outside of the 9:00am – 9:00pm time period that may require additional consideration for sleeping habits. Furthermore, in our study, both time-varying and time-independent covariates were completely observed, and so future work could explore more complex imputation in the situation where both activity count variables and model covariates could be missing.

Appendix

Following the notation of Lee and Gill (2016), the ZIPLN model is of dimension $d = 2K + 1$, and is a mixture of d independent Poisson models and a d-variate Log-normal model. It assumes that zero counts are observed with probability π_i, and that the activity counts can be modeled by a Poisson(λ_i) distribution with probability $1 - \pi_i$, where $\lambda_i = exp(x_i^T \beta + e_i)$, $e_i \sim N_d(0, \Sigma)$, and Σ denotes a $d\ x\ d$ variance-covariance matrix. In other words, a Log-normal distribution is placed on the mean parameter of the Poisson portion of the ZIP distribution (i.e. $\lambda_i \sim LN_d(x_i^T \beta, \Sigma)$).

The conditional expectation of the ZIPLN random variable is written as $E(Y_{i,t} | w_i, x_i) = (1 - logit^{-1}(w_i^T \gamma_t))\ exp(x_i^T \beta_t + e_{it})$. The first term of the conditional expectation is a logistic regression $logit(\pi_i) = w_i^T \gamma$, where w_i and γ are sets of covariates and regression coefficients, respectively, that help predict the probability of an activity count being from the zero inflation process, or from the Poisson Log-normal process. The covariates and regression coefficients x_i and β help describe the distribution of the activity counts that arise under the Poisson Log-normal model. The separation of the zero inflation portion and count portion of the ZIPLN model allows the analyst to separately specify which covariates cause zero inflation, and which cause activity.

If $Z = (Z_{t-K}, \dots, Z_{t-1}, Z_{t+1}, \dots, Z_{t+K})^T$ represents a set of K lag and lead variables of Y_t, where $Z_t = log(Y_t) - x_i^T \beta_t$, then $Z \sim N_{2K}(0, \Sigma_{zz})$. The conditional expectation of the ZIPLN model that incorporates lag and leads effects can be updated as

$E\left(Y_{i,t}\middle|Z_i, w_i, x_i\right) = \left(1 - logit^{-1}\left(w_i^T\gamma_t + \delta_t H_{i,t-1}\right)\right) exp\left(x_i^T\beta_t + \Sigma_{yz}\Sigma_{zz}^{-1}Z\right)$, which follows from the normal conditional distribution property. Additionally, $H_{i,t-1} = log(Y_{i,t-1} + 1)$, $\Sigma_{zz} = cov\,(Z, Z)$, and $\Sigma_{yz} = cov\,(Z_t, Z)$.

Parametric Bayesian imputation with a ZIPLN imputation model is carried out in several steps. First, the entire data set is filled in using ZIP imputation with 1 lag and predictive mean matching (PMM). Imputed counts are drawn from the conditional expectation of the ZIP model, $E\left(Y_{i,t}\middle|Y_{i,t-1}, w_i, x_i\right) = (1 - logit^{-1}(w_i^T\gamma_t + \delta_t H_{i,t-1}))\exp(x_i^T\beta_t + \alpha_t H_{i,t-1})$ where $H_{i,t-1} = log(Y_{i,t-1} + 1)$. ZIP model parameters, $\hat{B} = (\hat{\gamma}, \hat{\delta}, \hat{\beta}, \hat{\alpha})$, are estimated from the observed data, Y_{obs}, using ML, and then it is assumed that the posterior distribution of the parameters is normal with mean and variance defined by the ML estimates. Draws from the posterior distribution to obtain updated parameter estimates are equivalent to $\dot{B} = \hat{B} + V^{-\frac{1}{2}}z$, where $V = cov(\hat{B})$ and $z \sim N(0, 1)$. Imputations for the non-wear time, Y_{miss}, are selected using PMM based on draws from the ZIP model defined by the updated zero-inflated probability $\dot{\pi}_j = logit^{-1}(w_j^T\dot{\gamma}_t + \dot{\delta}_t H_{j,t-1}))$, and mean parameter $\dot{\lambda}_j = \exp(x_j^T\dot{\beta}_t + \dot{\alpha}_t H_{j,t-1})$. Zero-inflated imputations are drawn based on the zero-inflated probability, where $\hat{Y}_j = 0$ if $\dot{\pi}_j > \upsilon_j$, $\upsilon_j \sim unif(0, 1)$. Count imputations are drawn from the Poisson model, where $\hat{Y}_j = \dot{\lambda}_j$ if $\dot{\pi}_j \leq \upsilon_j$. Then, PMM can be applied based on the smallest absolute differences between the draws from the ZIP model, $\hat{Y}_j$, and the observed data, Y_{obs}, which completes one iteration.

Next, ZIPLN imputation with *K* lags and leads is performed in a similar fashion for the remaining iterations. A set of lags for the zero-inflated portion of the model,

$H_{i,t-1} = log(Y_{i,t-1} + 1)$, is calculated based on the completed data from the first iteration. A ZIP model is again fit to the completed data to obtain model parameters to inform the posterior distribution, and to calculate the K lags and leads, $Z = (Z_{t-K}, \dots, Z_{t-1}, Z_{t+1}, \dots, Z_{t+K})^T$, needed to inform the Log-normal error term. Updated parameter draws are made using $\dot{B} = \hat{B} + V^{-\frac{1}{2}}z$, where $\hat{B} = (\hat{\gamma}, \hat{\delta}, \hat{\beta})$, and the updated zero-inflated probability and mean parameter are calculated as $\dot{\pi}_j = logit^{-1}(w_j^T \dot{\gamma}_t + \delta_t H_{i,t-1}))$ and $\dot{\lambda}_j = \exp(x_j^T \dot{\beta}_t)$.

Again, zero-inflated imputations are drawn based on $\dot{\pi}_j = P(\hat{Y}_j = 0)$, where $\hat{Y}_j = 0$ if $\dot{\pi}_j > \upsilon_j$. Count imputations are now updated with the Log-normal error term considering K lags and leads $\hat{Y}_j = \dot{\lambda}_j exp(\dot{e})$, where $\dot{e} = \hat{\Sigma}_{yz}\hat{\Sigma}_{zz}^{-1}Z$. This process is repeated for the remaining iterations to yield one complete, imputed data set.

References

Alhassan, S., Sirard, J. R., Spencer, T. R., Varady, A., & Robinson, T. N. (2008). Estimating physical activity from incomplete accelerometer data in field studies. *Journal of Physical Activity & Health*, *5 Suppl 1*, S112–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18364516

Allison, P. D. (2002). *Missing data*. Sage Publications.

Borghese, M. M., Tremblay, M. S., LeBlanc, A. G., Leduc, G., Boyer, C., & Chaput, J. P. (2016). Comparison of ActiGraph GT3X+ and Actical accelerometer data in 9-11-year-old Canadian children. *Journal of Sports Sciences*, 1–8. http://doi.org/10.1080/02640414.2016.1175653

Cano, S., Rovira, U., Virgili, I., & Andreu, J. (2010). Using Multiple Imputation to Simulate Time Series: A proposal to solve the distance effect.

Catellier, D. J., Hannan, P. J., Murray, D. M., Addy, C. L., Conway, T. L., Yang, S., & Rice, J. C. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S555–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16294118

Colley, R., Garriguet, D., Janssen, I., Craig, C., Clarke, J., & Tremblay, M. (2011). Physical activity of Canadian children and youth: Accelerometer results from the 2007 to 2009 Canadian Health Measures Survey. *Statistics Canada Catalogue No. 82-003-X*, *22*(1).

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11778676

CSEP. (2012). Canadian Physical Activity Guidelines. Retrieved May 23, 2016, from http://www.csep.ca/CMFiles/Guidelines/CSEP_Guidelines_Handbook.pdf

Dunson, D. B., & Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics (Oxford, England)*, *6*(1), 11–25. http://doi.org/10.1093/biostatistics/kxh025

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–40. http://doi.org/10.1037/met0000063

Esliger, D. W., Copeland, J. L., Barnes, J. D., & Tremblay, M. (2005). Standardizing and optimizing the use of accelerometer data for free-living physical activity monitoring.

Esliger, D. W., & Tremblay, M. S. (2007). Physical activity and inactivity profiling: the next generation. *Canadian Journal of Public Health = Revue Canadienne de Santé Publique*, *98 Suppl 2*, S195–207. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18213949

Harvard T.H. Chan School of Public Health. (2016). Examples of Moderate and Vigorous Physical Activity | Obesity Prevention Source. Retrieved from https://www.hsph.harvard.edu/obesity-prevention-source/moderate-and-vigorous-physical-activity/

He, R. (2012). Multiple Imputation of High-dimensional Mixed Incomplete Data. UCLA.

He, R., & Belin, T. (2014). Multiple imputation for high-dimensional mixed incomplete continuous and binary data. *Statistics in Medicine*, *33*(13), 2251–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24918245

He, Y., Yucel, R., & Raghunathan, T. E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine*, *30*(10), 1137–56.

http://doi.org/10.1002/sim.4201

Janssen, I. (2013). The Public Health Burden of Obesity in Canada. *Canadian Journal of Diabetes*, *37*(2), 90–96. http://doi.org/10.1016/j.jcjd.2013.02.059

Janssen, I. (2014). Active play: An important physical activity strategy in the fight against childhood obesity. *Canadian Journal of Public Health*, *105*(1), e22–7. http://doi.org/10.17269/CJPH.105.4154

Junger, W. L., & Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96–104. http://doi.org/10.1016/j.atmosenv.2014.11.049

Kalaycioglu, O., Copas, A., King, M., & Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(3), 683–706. http://doi.org/10.1111/rssa.12140

Kleinke, K., & Reinecke, J. (2013). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, *67*(3), 311–336. http://doi.org/10.1111/stan.12009

Kowalski, K., Crocker, P., & Donen, R. (2004). The Physical Activity Questionnaire for Older Children and Adolescents. Retrieved May 23, 2016, from http://www.hfsf.org/uploads/Physical Activity Questionnaire Manual.pdf

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34*(1), 1. http://doi.org/10.2307/1269547

LeBlanc, A. G. W., & Janssen, I. (2010). Difference between self-reported and accelerometer measured moderate-to-vigorous physical activity in youth. *Pediatric Exercise Science*, *22*(4), 523–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21242602

Lee, J. A., & Gill, J. (2016). Missing value imputation for physical activity data measured by accelerometer. *Statistical Methods in Medical Research*. http://doi.org/10.1177/0962280216633248

Lee, P. H. (2013). Data imputation for accelerometer-measured physical activity: the combined approach. *The American Journal of Clinical Nutrition*, *97*(5), 965–71. http://doi.org/10.3945/ajcn.112.052738

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., … Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*, *15*, 346. http://doi.org/10.1186/s12859-014-0346-6

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons Inc.

Lloyd, J., Obradović, J., Carpiano, R., & Motti-Stefanidi, F. (2013). JMASM 32: Multiple Imputation of Missing Multilevel, Longitudinal Data: A Case When Practical Considerations Trump Best Practices? *Journal of Modern Applied Statistical Methods*, *12*(1). Retrieved from http://digitalcommons.wayne.edu/jmasm/vol12/iss1/29

Mâsse, L. C., Fuemmeler, B. F., Anderson, C. B., Matthews, C. E., Trost, S. G., Catellier, D. J., & Treuth, M. (2005). Accelerometer data reduction: a comparison of four reduction algorithms on select outcome variables. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S544–54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16294117

Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Statistics in Medicine*, *28*(29), 3657–69. http://doi.org/10.1002/sim.3731

Penpraze, V., Reilly, J., MacLean, C., Montgomery, C., Kelly, L., Paton, J., … Grant, S. (2006).

Monitoring of physical activity in young children: How much is enough? *Pediatric Exercise Science*, *18*, 483–491.

Poitras, V. J., Gray, C. E., Borghese, M. M., Carson, V., Chaput, J.-P., Janssen, I., … Tremblay, M. S. (2016). Systematic review of the relationships between objectively measured physical activity and health indicators in school-aged children and youth. *Applied Physiology, Nutrition, and Metabolism = Physiologie Appliquée, Nutrition et Métabolisme*, *41*(6 Suppl 3), S197–239. http://doi.org/10.1139/apnm-2015-0663

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Connor Gorber, S., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *The International Journal of Behavioral Nutrition and Physical Activity*, *5*, 56. http://doi.org/10.1186/1479-5868-5-56

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. http://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. (D. B. Rubin, Ed.)*Harvard University*. Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470316696

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12090408

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, *33*(4), 545–71.

Schafer, J. L., & Olsen, M. K. (1999). Modeling and Imputation of Semicontinuous Survey Variables. *The Methodology Center, Penn State University, USA*.

Schafer, J. L., & Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *Journal of Computational and Graphical Statistics*, *11*(2), 437–457. http://doi.org/10.1198/106186002760180608

Schott, J. R. (2005). Testing for complete independence in high dimensions. *Biometrika*, *92*(4), 951–956. http://doi.org/10.1093/biomet/92.4.951

Srivastava, M. S., & Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, *101*(6), 1319–1329. http://doi.org/10.1016/j.jmva.2009.12.010

Statistics Canada. (2014). Canadian Health Measures Survey (CHMS). Retrieved from http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5071

Statistics Canada. (2015). Directly measured physical activity of children and youth, 2012 and 2013. Retrieved from http://www.statcan.gc.ca/pub/82-625-x/2015001/article/14136-eng.htm

Troiano, R. P. (2005). A timely meeting: objective measurement of physical activity. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S487–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16294111

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–42. http://doi.org/10.1177/0962280206074463

van Buuren, S. (2011). Multiple imputation of multilevel data. In *The Handbook of Advanced Multilevel Analysis* (pp. 173–196). Routledge, Milton Park, UK.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. CRC Press. http://doi.org/10.1201/b11826

Welch, C. A., Petersen, I., Bartlett, J. W., White, I. R., Marston, L., Morris, R. W., … Carpenter, J. (2014). Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, *33*(21), 3725–37. http://doi.org/10.1002/sim.6184

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *366*(1874), 2389–403. http://doi.org/10.1098/rsta.2008.0038

Yucel, R. M. (2011). Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling*, *11*(4), 351–370. http://doi.org/10.1177/1471082X1001100404

Zhao, E., & Yucel, R. M. (2009). Performance of Sequential Imputation Method in Multilevel Applications.