Improved Methods for Interrupted Time Series Analysis Useful When Outcomes are Aggregated: Accounting for heterogeneity across patients and healthcare settings

IMPROVED METHODS FOR INTERRUPTED TIME SERIES ANALYSIS USEFUL WHEN OUTCOMES ARE AGGREGATED: ACCOUNTING FOR HETEROGENEITY ACROSS PATIENTS AND HEALTHCARE SETTINGS

BY

JOYCELYNE EFUA EWUSIE, B.Sc., M.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF

HEALTH RESEARCH METHODS, EVIDENCE, AND IMPACT

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2018) McMaster University

(Health Research Methods, Evidence, and Impact) Hamilton, Ontario, Canada

TITLE: Improved methods for interrupted time series analysis useful when outcomes are aggregated: accounting for heterogeneity across patients and healthcare settings

AUTHOR: Joycelyne Efua Ewusie

B.Sc. (Statistics & Computer Science)

University of Ghana, Legon, Ghana

M.Sc. (Statistics)

McMaster University

SUPERVISOR: Dr. Jemila S. Hamid

NUMBER OF PAGES: xvi, 135

This thesis is dedicated to the memory of my dear brother, Nana Kojo Blankson, who always gave me a reason to work harder.

Abstract

In an interrupted time series (ITS) design, data are collected at multiple time points before and after the implementation of an intervention or program to investigate the effect of the intervention on an outcome of interest. ITS design is often implemented in healthcare settings and is considered the strongest quasi-experimental design in terms of internal and external validity as well as its ability to establish causal relationships. There are several statistical methods that can be used to analyze data from ITS studies. Nevertheless, limitations exist in practical applications, where researchers inappropriately apply the methods, and frequently ignore the assumptions and factors that may influence the optimality of the statistical analysis. Moreover, there is little to no guidance available regarding the application of the various methods, and a standardized framework for analysis of ITS studies does not exist. As such, there is a need to identify and compare existing ITS methods in terms of their strengths and limitations. Their methodological challenges also need to be investigated to inform and direct future research. In light of this, this PhD thesis addresses two main objectives: 1) to conduct a scoping review of the methods that have been employed in the analysis of ITS studies, and 2) to develop improved methods that address a major limitation of the statistical methods frequently used in ITS data analysis. These objectives are addressed in three projects.

For the first project, a scoping review of the methods that have been used in analyzing ITS data was conducted, with the focus on ITS applications in health research. The review was based on the Arksey and O'Malley framework and the Joanna Briggs Handbook for scoping reviews. A total of 1389 studies were included in our scoping

review. The articles were grouped into methods papers and applications papers based on the focus of the article. For the methods papers, we narratively described the identified methods and discussed their strengths and limitations. The application papers were summarized using frequencies and percentages. We identified some limitations of current methods and provided some recommendations useful in health research.

In the second project, we developed and presented an improved method for ITS analysis when the data at each time point are aggregated across several participants, which is the most common case in ITS studies in healthcare settings. We considered the segmented linear regression approach, which our scoping review identified as the most frequently used method in ITS studies. When data are aggregated, heterogeneity is introduced due to variability in the patient population within sites (e.g. healthcare facilities) and this is ignored in the segmented linear regression method. Moreover, statistical uncertainty (imprecision) is introduced in the data because of the sample size (number of participants from whom data are aggregated). Ignoring this variability and uncertainty will likely lead to invalid estimates and loss of statistical power, which in turn leads to erroneous conclusions. Our proposed method incorporates patient variability and sample size as weights in a weighted segmented regression model. We performed extensive simulations and assessed the performance of our method using established performance criteria, such as bias, mean squared error, level and statistical power. We also compared our method with the segmented linear regression approach. The results indicated that the weighted segmented regression was uniformly more precise, less biased and more powerful than the segmented linear regression method.

In the third project, we extended the weighted method to multisite ITS studies, where data are aggregated at two levels: across several participants within sites as well as across multiple sites. The extended method incorporates the two levels of heterogeneity using weights, where the weights are defined using patient variability, sample size, number of sites as well as site-to-site variability. This extended weighted regression model, which follows the weighted least squares approach is employed to estimate parameters and perform significance testing. We conducted extensive empirical evaluations using various scenarios generated from a multi-site ITS study and compared the performance of our method with that of the segmented linear regression model as well as a pooled analysis method previously developed for multisite studies. We observed that for most scenarios considered, our method produced estimates with narrower 95% confidence intervals and smaller p-values, indicating that our method is more precise and is associated with more statistical power. In some scenarios, where we considered low levels of heterogeneity, our method and the previously proposed method showed comparable results.

In conclusion, this PhD thesis facilitates future ITS research by laying the groundwork for developing standard guidelines for the design and analysis of ITS studies. The proposed improved method for ITS analysis, which is the weighted segmented regression, contributes to the advancement of ITS research and will enable researchers to optimize their analysis, leading to more precise and powerful results.

Publications Related to Thesis

1. Ewusie JE, Blondal E, Soobiah C, Beyene J, Thabane L, Straus S, Hamid JS. (2017). "Methods, Applications, Interpretations and Challenges of Interrupted Time Series (ITS) Data: Protocol for a Scoping Review." *BMJ Open*, 7(6), e016018.

2. Ewusie JE, Blondal E, Soobiah C, Beyene J, Thabane L, Straus S, Hamid JS. (2018). "Methods, Applications, Interpretations and Challenges of Interrupted Time Series (ITS) Data: a scoping review", *BMJ Open*, revision submitted

3. Ewusie JE, Beyene J, Thabane L, Straus S, Hamid JS. "An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient Heterogeneity Using Weighted Analysis", *BMC Medical Research Methodology*, under review.

4. Ewusie JE, Beyene J, Thabane L, Straus S, Hamid JS. "Multi-center Interrupted Time Series Analysis: Incorporating Within and Between Center Heterogeneity", *Statistical Methods in Medical Research*, under review.

Acknowledgements

I would like to thank the Almighty God for His guidance and help during my stay at McMaster University.

I would like to express my profound gratitude to my supervisor Dr. Jemila Hamid for her outstanding supervision, mentorship, support, patience, and incessant encouragement. I will be externally grateful for her great ideas, friendly discussions, cooperation and enthusiasm that contributed greatly to this thesis and my overall academic growth. I am immensely grateful to members of my thesis committee: Dr. Sharon Straus, Dr. Joseph Beyene and Dr. Lehana Thabane for their availability and support, as well as for providing consistent guidance and opportunities during my PhD journey. I would also like to thank all my co-authors for their contributions.

I extend my warmest gratitude to the friends, colleagues and staff in the Department of Health Research Methods, Evidence and Impact for providing a motivating and friendly learning environment. I would also like to give special thanks to friends, colleagues and collaborators at the Li Ka Shing Knowledge Institute. Many thanks to Caitlin, Thuva, Regina, Sayantee, Zelalem, Ashley, Lawrence, Anderson, the Keowns, Odames, Masons, Bowens, Hardings and Boots for their time, motivation and support in various ways.

My deepest appreciation goes to my husband Eric, my mum Dorcas, and my uncle Abeiku (Ben) for being my support system. I am forever thankful for their prayers, encouragement and for always believing in me. Finally, I would like to say a big thank you to my family and friends who have been with me throughout my academic journey.

Table of Contents

Abstract	iv
Publications Related to Thesis	vii
Acknowledgements	viii
Table of contents	ix
List of Tables	xii
List of Figures	xiii
List of Abbreviations	xv
Preface	xvi
Chapter One: Introduction	1
1.1. Background	1
1.1.1. Quasi-experimental Designs	1
1.1.2. Interrupted Time Series (ITS) Designs	3
1.1.3. Implementation and Methodological Challenges in ITS Studies	4
1.1.4. Statistical Methods Used in the Analysis of ITS Data	7
1.2. Rationale and Objectives	13
1.3. Organization and Scope of the Thesis	15
References	18
Chapter Two: Interrupted Time Series Methods and Applications in Health Research	**22**
Paper 1: Abstract	23
2.1. Introduction	24
2.2. Methods	26
2.2.1. Search Strategy	27
2.2.2. Eligibility Criteria	27
2.2.3. Study selection and data collection	28
2.2.4. Data extraction and synthesis	28
2.3. Results	29
2.3.1. Study Characteristics	29
2.3.2. Review of Statistical Methods for ITS Analysis	32
2.3.2.1. Papers on Novel ITS Methods	33
2.3.2.2. Papers on Improved or Adapted Methods	38

2.3.2.3. Papers on Methodological Comparisons 40
2.3.3. Review of the Application Studies involving ITS Design and Analysis 42
2.4. Discussion 45
References 52
Chapter Three: Weighted Segmented Regression for Analyzing Interrupted Time Series Data **57**
Paper 2: Abstract 58
3.1. Background 59
3.2. Methods 62
3.2.1. Weighted Segmented Regression 64
3.2.2. The Weights 65
3.3. Simulations 66
3.3.1. Description of Simulations 66
3.3.2. Simulation Results 68
3.3.2.1. Bias and MSE 68
3.3.2.2. Level of significance and statistical power 72
3.4. Real Data Example 80
3.5. Discussion 82
References 85
Chapter Four: Extension of the Weighted Segmented Regression to Account for Variability in Healthcare Settings **89**
Paper 3: Abstract 90
4.1. Background 91
4.2. Methods 94
4.2.1 Empirical Data 94
4.2.2. Weighted ITS Analysis 95
4.2.3. Empirical Evaluation 99
4.3. Results 100
4.4. Discussion 111
References 113
Chapter Five: Conclusions **116**
5.1. Summary of Findings 116
5.2. Future Directions 122

5.3. Concluding Remarks 124
References 127
Appendix **129**

List of Tables

Chapter Two

Table 2.1: Description of studies included in the review with respect to methods used in the analysis of ITS data. 31

Chapter Three

Table 3.1: Range of parameters considered in the simulation study. 66

Table 3.2: Bias for segmented linear regression and for weighted segmented regression with small, moderate and large variance heterogeneity. 71

Table 3.3: Mean squared error for segmented linear regression and for weighted segmented regression with small, moderate and large variance heterogeneity. 72

Table 3.4: Estimates of power for segmented linear regression and weighted segmented regression for different sample sizes. 78

Table 3.5: Change in level and slope with corresponding 95% confidence interval (CI) and p-values of segmented linear regression and weighted segmented regression for mobility. 81

Table 3.6: MSE, AIC and p-value of segmented linear regression and weighted segmented regression for mobility. 81

Chapter Four

Table 4.1: Estimates for intervention effects, 95% CI and p values obtained using segmented linear regression (SLR), pooled analysis (PA) and weighted segmented regression (*w*SR). 101

Table 4.2: Subgroup estimates for intervention effects, 95% CI and p values obtained using segmented linear regression (SLR), pooled analysis (PA) and weighted segmented regression (wSR). 105

List of Figures

Chapter Two

Figure 2.1: Flow chart outlining the search and review process, the records identified, included and excluded as well as the reasons for exclusion. 30

Figure 2.2: Trend of interrupted time series application papers over the last two decades. 43

Chapter Three

Figure 3.1: Empirical bias for change in level using weighted (a) and unweighted (b) methods where data was generated for increasing values of within patient variance and four different sample sizes. 69

Figure 3.2: Average mean squared error (MSE) for change in level (left panel) and trend (right panel) estimates, for large (a, d), moderate (b, e) and small (c, f) differences in of variability across different sample sizes per time point. 70

Figure 3.3: Empirical level for change in level across different sample sizes with low variance across time points. 73

Figure 3.4: Empirical level for change in trend across different sample sizes with low variance across time points. 73

Figure 3.5: Error rate for change in level using weighted (a, b) and unweighted (c, d) methods where data was generated for large (a, c) and moderate (b, d) levels of within patient variability across different sample sizes. 74

Figure 3.6: Error rate for change in trend across different sample sizes and levels of variance heterogeneity for unweighted (left panel) and weighted (right panel) methods. 76

Figure 3.7: Power curve for change in level (left panel) and trend (right panel) across different sample size and large (a, d), moderate (b, e) and small (c, f) between patient variability. 77

Figure 3.8: Power curve for change in level across different sample sizes when differences in error variance is large for different values of β. 79

Chapter Four

Figure 4.1: Weekly percentage of patients who were mobile at least 98

once a day for all sites and overall.

Figure 4.2: Forest plot for change in level of mobilization during the implementation of the intervention. 102

Figure 4.3: Forest plot for change in trend of mobilization during the implementation of the intervention. 103

Figure 4.4: Forest plot for change in level of mobilization after the implementation of the intervention. 104

Figure 4.5: Forest plot for change in level of mobilization post intervention for high between site heterogeneity. 106

Figure 4.6: Forest plot for change in level of mobilization during intervention for moderate between site heterogeneity. 107

Figure 4.7: Forest plot for change in trend of mobilization during intervention for moderate between and within site heterogeneity. 107

Figure 4.8: Forest plot for change in trend of mobilization post intervention for low between site heterogeneity. 108

Figure 4.9: Forest plot for change in level of mobilization post intervention for low within and between site heterogeneity. 109

Figure 4.10: Forest plot for change in level of mobilization post intervention for moderate within and between site heterogeneity. 110

Figure 4.11: Forest plot for change in trend of mobilization post intervention for high within site variability. 110

List of Abbreviations

AIC	Akaike Information Criterion
ARIMA	Auto Regressive Integrated Moving Average
CI	Confidence Interval
ITS	Interrupted Time Series
LOS	Length of stay
MLE	Maximum Likelihood Estimator
MOVE-ON	Mobility of Vulnerable Elders in Ontario
MSE	Mean Squared Error
OLS	Ordinary Least Squares
PA	Pooled Analysis
QED	Quasi-experimental designs
QI	Quality Improvement
SR	Segmented Regression
SLR	Segmented Linear Regression
wSR	Weighted Segmented Regression

Preface

This thesis is prepared as a "sandwich thesis" consisting of five chapters. Chapter 1 introduces the topic of the research, lays out the rationale and objectives of the studies and provides the scope of the thesis. The subsequent three chapters (2-4) consist of three manuscripts which form the basis for this thesis and are currently under review for publication in peer-reviewed journals. The summary of the thesis, direction for future research, as well as the contributions of this thesis to health research are provided in Chapter 5. Joycelyne Ewusie's contributions to all the manuscripts include: developing the research ideas and questions, writing the protocol, collecting the data, conducting data organization and management, performing the statistical analysis, interpreting the results, writing the manuscripts, submitting the manuscripts as well as responding to reviewers' comments. All work presented in the chapters was conducted between the fall of 2015 and summer of 2018.

Chapter 1

Introduction

1.1. Background

1.1.1. Quasi-experimental Designs

Quasi-experimental designs (QEDs) include a wide range of non-randomized designs that are used to evaluate the effect of interventions. In several scenarios, QEDs involve the use of multiple groups or different ways of measurement (Lucasey, 2002). They are usually used in social sciences, particularly in community interventional research, (Shadish et al., 2002) and in recent times, QEDs have gained increasing popularity in health research (Eliopoulos et al., 2005; Jandoc et al., 2015; Ewusie et al., 2017; Rockers et al., 2017). QEDs have been used in several studies to assess the impact of primary care policy reforms and programs on clinical practice or health system outcomes using administrative routine data (Wagenaar et al., 1988; O'Malley and Wagenaar, 1991; Melhuish et al., 2008; Sommers et al., 2014; Anchah et al., 2017; Preval et al., 2017).

In utilizing QEDs, several interventions can be compared or evaluated concurrently (Reichardt, 2009). For example, one can assess the impact of three different drug policies implemented at different times over the period of ten years on the utilization and costs of drugs of uninsured patients using QEDs. QEDs can also be used to assess the impact of both planned and unplanned interventions or events on an outcome of interest (Campbell

and Riecken, 1968). For instance, QEDs could be used to examine the effect of an earthquake on trauma-related deaths. They are the optimal practical approaches for assessing system shocks due to unexpected events as well as evaluating retrospectively the effect of health policies.

The purpose of QEDs is to test causality between an outcome and an intervention (Harris et al., 2006). Another aim is to use pretest measures or control groups to provide counterfactual inference about what would have happened in the absence of the intervention (Cook et al., 2002). In other words, they mainly help to answer the question of what would have been observed if the subjects had not been exposed to the intervention. Arguably, QEDs may be the best alternative approach for evaluating intervention effects when random assignment is not possible due to factors such as ethical considerations where random withholding of an intervention with well-established efficacy will not be allowed or may raise ethical issues, lack of power because of small sample size, or feasibility where it is impractical to randomize an intervention to individual patients or individual units due to the possibility of contamination (Campbell and Stanley, 1963; Campbell and Riecken, 1968; Black, 1996).

Despite being the best approach to evaluating intervention effects in these scenarios, the lack of randomization is a major weakness and poses a threat to the internal and external validity of QEDs (Cook et al., 1979; Cook et al., 2002). Internal validity refers to the degree to which the changes observed in the outcome can be correctly attributed to the exposure or intervention; that is, did the intervention really make a difference in the experiment? External validity refers to the degree of generalizability of the results. That is,

to what population, setting, treatment variables or measurement variables can the observed effect be generalized (Campbell and Riecken, 1968)? The main threats to internal validity are *history, maturation, testing, instrumentation, statistical regression/ regression to the mean, selection bias/confounding, experimental mortality/attrition* and *ambiguous temporal precedence*. The main threats to external validity are *reactive/interaction effect of testing, reactive effects of experimental arrangement* and *multiple treatment interference* (Campbell and Stanley, 1963; Cook et al., 1979; Shadish et al., 2002). To establish causality, the main aim of any researcher is to select a study design, which is strong in both types of validity (Campbell and Stanley, 1963).

Several forms of study designs are classified under QEDs. According to Shadish et al., (2002) who identified different types of QEDs, there are 17 possible QEDs. These designs were reviewed and grouped by Harris et al., (2006) into four main categories: 1) Quasi-experimental designs without control group 2) Quasi-experimental designs with control groups but no pretest, 3) Quasi-experimental designs with control groups and pretest and 4) Interrupted time series designs (Shadish et al., 2002; Harris et al., 2006). These categories are listed in increasing hierarchical order in terms of their internal/external validity and consequently their ability to establish causality.

1.1.2. Interrupted time series (ITS) Designs

The strongest and frequently utilized QED in health research is the interrupted time series (ITS) study design (Gillings et al., 1981; Shadish et al., 2002; Penfold and Zhang, 2013). For ITS designs, observations are made at multiple instances over time before and after the implementation of an intervention or program (e.g. mass media campaign on breast cancer

awareness). With ITS designs, investigators can examine whether the program had a significantly higher effect on an outcome of interest than any underlying secular trend (Ramsay et al., 2003). ITS designs also ensure that natural variation in the repeated data is not mistaken for a real change in the outcome of interest due to intervention.

ITS designs have a wide range of applications in health research. The value of information provided by high quality ITS studies have received increasing recognition in implementation science, especially to assess the effects of evidence-based complex interventions, in pharmaceutical research, and generally in public health evaluation (Ansari et al., 2003; Moreno-Torres et al., 2011; Kirkland et al., 2012; Kastner et al., 2014; Mead et al., 2016). ITS designs have been instrumental in assessing the impact of clinical practice guidelines in changing clinical practice and improving patient care (Grimshaw et al., 2000; Mol et al., 2005; Lu et al., 2014). Furthermore, they are included in Cochrane systematic reviews conducted by the Cochrane Effective Practice and Organization of Care Group (EPOC) (Bero et al., 2002).

1.1.3. Implementation and Methodological Challenges in ITS Studies

ITS designs have several strengths compared to traditional before and after studies. Having multiple observations of an outcome prior to and then after the introduction of an intervention, as stated by Wagner et al., (2002), provide investigators with a tool that is more sensitive to potential confounding factors compared to using a simple pretest – posttest study designs. By making multiple observations, ITS designs can address major threats to internal validity such as history and maturation. A major strength of an ITS design is that it allows the statistical investigation of potential biases in the estimate of the effect

of an intervention (Ramsay et al., 2003). The biases that are usually investigated include secular trend, cyclical or seasonal effects, duration of intervention and autocorrelation.

The second notable strength of ITS design is the ability to model the data at the population level rather than at the individual level (Penfold and Zhang, 2013). When the population level data for instance show a clear linear trend, it is advisable to model the data using population level summary estimates. Moreover, using population level data help to control individual level confounding factors, which are likely to introduce bias in the results, unless the factor occurs simultaneously with intervention. Population shifts over time can also be adjusted for using standardization (Briesacher et al., 2011; Penfold and Zhang, 2013). Further, researchers using ITS designs can obtain overall effects of intervention as well as assess the impact of intervention on subgroups of the population using stratified analysis (Wagner et al., 2002; Penfold and Zhang, 2013).

A third strength of ITS design is that it can be applied when interventions cannot be randomized due to feasibility or ethical considerations (Wagner et al., 2002; Ramsay et al., 2003; Taljaard et al., 2014). In the field of implementation science for instance, the aim as mentioned previously is usually to evaluate the effects of the scale up of a program (e.g. quality improvement program) or to assess post-hoc effects of a policy with regional or national coverage. Randomization is hence not practical and ITS becomes the best alternative to assess intervention. Moreover, in analyzing administrative data such as routine health information systems data, which are used for health evaluations especially in low and middle-income countries, ITS designs are considered the best approach for accessing the effect of interventions or policies (Wagenaar et al., 2015).

ITS designs also enable researchers to assess unplanned consequences of an intervention or program. Researchers using ITS designs can also assess the unintended effect of an intervention on outcomes other than the primary outcome (Chace et al., 2012; Penfold and Zhang, 2013). Finally, ITS analysis provides graphical results that are easy to read/interpret. Apart from providing estimates of intervention effect, readers and knowledge users can easily understand the trend of the data prior to the implementation of the intervention and any changes that have occurred thereafter.

Despite these various strengths, it is important to also acknowledge the inherent limitations of ITS designs, even when they are adequately implemented. These limitations are mostly related to their internal validity and the potential generalizability of their findings. One of the major limitations of ITS designs is the threat of time-varying confounders. These include factors or events that occur around the same time as the implementation of the intervention and that can potentially affect the outcome of interest. They might also include other interventions or natural events occurring at the time of implementing the study intervention and which target the same outcome. There are several ways of determining if these confounders provide plausible explanations of the observed effect. These include using a control group or control outcome unaffected by the intervention; a multiple baseline design where the intervention is introduced at different locations and times; or to analyze the series with reconstructed units of measurements (e.g. using monthly data instead of yearly data) to specify the exact time when the intervention occurred and identify if any event that might affect the outcome was happening concurrently (Shadish et al., 2002; Bernal et al., 2017).

Seasonal variability is another major challenge of an ITS design (Bernal et al., 2017). One instance where seasonality can cause a problem is when there is uneven distribution of months pre- and post- intervention, that is more winter months in one period for an outcome more frequent in the winter (e.g. flu rates). When designing ITS studies, it is imperative that researchers anticipate any seasonal or cyclical patterns which may affect the outcome and attempts should be made to balance them between the pre and post intervention periods.

Another important limitation in the application of ITS designs is selection bias, where the composition of the intervention group is likely to change at the same time as the intervention (Penfold and Zhang, 2013). For instance, a large dropout rate at the start of the intervention due to the intervention requirements (e.g. diet change). Instrumentation is another threat to the internal validity and consequently a major limitation of ITS designs. Instrumentation causes bias in the results when the nature of the intervention causes a change in the way the outcome is measured or observed (Hacker et al., 2012).

1.1.4. Statistical Methods Used in the Analysis of ITS Data

One of the main features that can be used to determine a well implemented ITS study is the use of appropriate statistical methods to obtain adequate results. The methods used in ITS studies seek to evaluate the changes in the time course of an outcome of interest as a function of the inception of an intervention such as a policy or program. Several statistical methods are currently available for the analysis of ITS data (Ramsay et al., 2003; Shardell et al., 2007; Jandoc et al., 2015; Ewusie et al., 2017). These methods include: non-regression models such as t-tests and ANOVA which do not account for the time

component of the ITS design or any dependence in the observed time series data thus may lead to spurious results (Crosbie, 1993; Biglan et al., 2000); non-segmented regression models such as linear regression and generalized linear models which account for the underlying trend as well as potential confounders but do not assess the change in trend after implementation of an intervention (Donnelly, 2005; Shardell et al., 2007); and segmented regression models which account for the underlying trend in the data as well as the change in trend after the intervention is implemented (Wagner et al., 2002; Taljaard et al., 2014).

Segmented Regression

Segmented regression (SR) also known as piecewise regression involves fitting separate regression models (usually linear regression models, although other statistical modelling approaches can be used) to each segment of the time series data or to each intervention period. This way, each segment can exhibit different levels and trends, and the outcome of interest can also evolve differently over time before and after the intervention (Wang et al., 2013). The estimates of slope and intercept obtained for the pre- and post-intervention are then compared to examine the impact of the intervention. A change in level of the outcome post-intervention implies an immediate intervention effect while a change in trend of the outcome constitutes an intervention effect that was realized over time. The change in trend thus allows evaluators to measure the sustainability of the intervention effect (Shardell et al., 2007; Wang et al., 2013). In medical research, segmented linear regression models (SLR) are currently the most common statistical method employed in the analysis of ITS data (Zhang et al., 2009; Taljaard et al., 2014; Jandoc et al., 2015).

Despite the frequency in use and the various advantages, SLR as it is currently used has been shown to have several limitations (Wagner et al., 2002; Taljaard et al., 2014). Some of the major challenges that are often encountered in the use of SLR are serial correlation (autocorrelation), seasonality and non-constant error variances. These challenges are discussed further in detail below.

Serial Correlation

A critical assumption of regression modelling is that the errors are independent. With time series data however, observations are taken at successive time intervals or lags thus the errors ($\varepsilon_i's$) tend to be correlated. This is known as serial correlation or autocorrelation. Autocorrelation can be either positive or negative in magnitude. Both forms of autocorrelation violate the assumption of independence since the magnitude of the correlation between errors over time is different from zero. This violation causes standard error estimates produced by the regression model to be biased leading to erroneous significance or lack thereof of intervention effects (Sen and Srivastava, 2012).

There are two ways of assessing the presence of autocorrelation in a time series data. However, before assessing the presence of autocorrelation the data must first be made stationary. A stationary series is a series which is transformed such that the series does not increase or decrease with time, that is, has a constant mean and variance (Chatfield, 2016). One approach that can be used to achieve stationarity is differencing (Biglan et al., 2000). Differencing can be performed easily using packages in most statistical software such as SPSS, R or SAS.

The presence of autocorrelation in a time series data can then be assessed by either visual inspection or statistical testing. One approach for performing visual inspection of the data by use of the correlograms (Metcalfe and Cowpertwait, 2009). A couple of statistical tests can also be used to determine the presence of autocorrelation. The frequently used test is the Durbin-Watson (D-W) test (Sen and Srivastava, 2012). Another well-known statistical test that can be used is the Q statistics. These tests are described in various statistical textbooks (Anderson, 2011). All statistical tests for autocorrelation can be performed with ease using common statistical software packages in R, SAS or SPSS.

When autocorrelation is detected in a time series data, they can be accounted for using SLR with autoregressive error models or the segmented autoregressive integrated moving average (ARIMA) model. The SLR with autoregressive error models approach involve first fitting the SLR model as described by Wagner et al., 2002. The residuals from the model are then examined to test presence of autocorrelation. If autocorrelation is not present, then the basic SLR model would suffice, otherwise, the SLR model incorporating autoregressive process is employed. The SLR model with autocorrelated errors has two components: the structural component and the autoregressive component. The models are generally expressed as:

$$y_i = E(y_i) + R_i,$$

where

$y_i =$ outcome of interest (e.g. number of patients screened) at time i,

$E(y_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ (the structural component i.e. segmented linear regression model),

$R_i = \varphi_1 R_{i-1} + \cdots + \varphi_p R_{p-1} + \varepsilon_i$ (the autoregressive component), and φ_p= the autoregressive parameter at lag p.

The ARIMA model on the other hand comprises three components: the autoregressive (AR) component which refers to autocorrelation between errors at different time points; the moving average (MA) component which refers to each observation as being a function of the average of the errors; and the integrated (I) component which refers to the drift or trend that may be in the series (Biglan et al., 2000). In ARIMA modelling, the assumption is that the time series is stationary, that is does not increase or decrease with time. Thus, any trends and random drifts need to be transformed to meet this assumption. The transformation is usually done using the iterative process of differencing to remove all the trends. For ITS analysis, the ARIMA model is initially built using the pre-intervention data. Once a feasible model for the baseline data is identified, it is then applied to the complete ITS data.

A major limitation of ARIMA is the complexity of application. As described by Box & Jenkins (1979), who first presented the ARIMA approach for fitting statistical models, the ARIMA approach involves three steps: 1) Identification, 2) Estimation and 3) Diagnosis. The details of the approach are beyond the scope of this thesis but can be found in their widely accepted text on time series models (Box et al., 2015). A second major limitation is that to obtain adequate estimates approximately 50 time points or observations are required to perform the analysis (Biglan et al., 2000).

The complexity of application coupled with the large data requirement makes the ARIMA modelling a less favorable choice especially in health research where in some cases it is not possible to obtain that many observations due to resource and time constraints (Zhang et al., 2009). This makes the SLR incorporating autocorrelated error models more appealing since they tolerate fewer time points or observations, a total of 8 to 10 observations per period will be enough to adequately estimate effect sizes (Ramsay et al., 2003; Penfold and Zhang, 2013). Additionally, they are relatively less complex.

It is important to note here that the SLR with autoregressive error models do not account for moving average processes. However, the AR (1) process is what is mostly encountered in practice and in most cases little or no evidence of autocorrelation exists in a time series data after adjusting for autocorrelation using AR (1) process (Sen and Srivastava, 2012; Bernal et al., 2017). The ARIMA and SLR models can both be implemented using among others, the ARIMA package in R or PROCAUTOREG in SAS.

Seasonality

Autocorrelation can also be a consequence of seasonality. Both ARIMA and SLR models can account for seasonality by incorporating autoregressive parameters at a given seasonal lag, for instance incorporating AR parameters at lags 1 and 4 for quarterly data. To account for seasonality, having enough datapoints is a necessary requirement. For example, at least 24 monthly data points is required to account for seasonal patterns in monthly data (Bernal et al., 2017).

Homogeneity of error variances

Another major assumption that is likely to be violated in utilizing SLR is the homogeneity of error variances. This is related to the assumption that the observations, and consequently the error terms, have a constant variance. That is, $var(\varepsilon_i) = var\ (y_i) = \sigma^2$. The violation of the homogeneity assumption is often referred to as heteroscedasticity and this violation causes the variances of the parameter estimates to be large, leading to invalid conclusions about intervention effects. Heteroskedasticity can be caused by a time series with high seasonal patterns, whose variability increases with time. In such scenarios, de-seasonalizing the series can concurrently correct the heteroscedasticity (Anderson, 2011).

1.2. Rationale and Objectives

ITS studies in health research typically involve modelling data at the population level instead of the individual level. Consequently, the outcomes measured at a given time point are estimated or aggregated from individuals within a setting, such as a healthcare facility. For instance, consider a study assessing the weekly number of patients who had a mammography screening following the commencement of the mass media campaign. Thus, the outcome at each time point (y_i), is the weekly total number of patients ordered a screening. That is, $y_i = \sum_{p=1}^{k} x_{pi}$, where x represents daily number of patients, p represents the number of days per week and $k = 7$. Since participants' characteristics and total sample size are likely to vary within each time point, the variance is no longer constant. That is, $var\ (y_i) = var(\varepsilon_i) = \sigma_i^2$. Hence, aggregated data are a common cause of heteroskedasticity in ITS studies. Moreover, most ITS studies are conducted across

multiple sites and as a result, heterogeneity can be introduced because of differences in settings across the sites involved in the study.

Several methods have been developed and established as optimal methods to deal with some of the major challenges/limitations encountered when using SLR. These include ARIMA and SLR with autoregressive error models which can be used when there is serial correlation or seasonality. However, although the negative effects of the heterogeneity associated with aggregated data on results obtained from analysis of ITS data has been highlighted in several papers, little attention is directed towards overcoming this challenge (Wagner et al., 2002; Shardell et al., 2007; Taljaard et al., 2014; Jandoc et al., 2015). Additionally, to our knowledge, there has not been any method specifically developed to optimally account for the inherent heterogeneity in ITS data due to patient variability. Nonetheless, for accounting site to site variability, Gebski et al., (2012) introduced the pooled analysis method where the intercept and slopes obtained from individual site analysis are pooled, using a meta-analytic approach, to provide the overall estimates. This method of pooling summary estimates, instead of using all the available data, despite providing improved analysis, still leads to loss of information and hence is associated with imprecision and loss of statistical power. Moreover, their method does not account for the variability across participants within the same site.

The motivation for this thesis is therefore based on the fact that 1) most ITS data in health research are aggregated, 2) heterogeneity is introduced due to aggregation of varying patient population and different healthcare settings, 3) the SLR, which is the most commonly used method for ITS analysis, does not account for this heterogeneity and, 4)

the pooled analysis aimed at accounting between site variability can be improved. Additionally, this thesis is also motivated by real-life applications, where we encountered the above methodological gaps in a multi-site ITS study involving an evidenced-based intervention (Liu et al., 2013; Liu et al., 2017). The intervention was implemented across various hospitals in Ontario, Canada and the aim of the study was to examine its impact on the mobility of vulnerable elders admitted to acute care hospitals.

Considering the gaps in current methods, the overall objective of this thesis is to provide improved methods for analyzing ITS data and provide a methodological framework that will allow variability in the data to be incorporated. The specific objectives are to 1) conduct a scoping review to identify existing statistical methods for ITS analysis, describe the methods, elucidate methodological differences and identify challenges and gaps in usage of the methods, 2) develop and evaluate a novel method useful for aggregated data, where the patient variability is incorporated as weights and, 3) develop and evaluate a method that incorporates both participant variability and site to site variability in ITS studies conducted across multiple sites.

1.3. Organization and Scope of the Thesis

This thesis is organized as a sandwich thesis, where the three specific objectives are presented as papers in Chapters 2, 3 and 4. In Chapter 1, background material related to ITS designs, their strengths and limitations was presented. This chapter also provided a summary of the SLR approach, which is the commonly applied method in the analysis of

ITS data and is also the focus of this thesis. Rationale and objectives of the thesis has also been presented in Chapter 1.

In Chapter 2, the results of the scoping review of statistical methods that have been developed for or utilized in the analysis of ITS data in health research are presented. The scoping review was conducted as an initial step towards exploring the methodological gaps in segmented linear regression. We identified and described the statistical methods used in ITS analysis and examined their applications in health research. The study protocol for this review has been published (Ewusie et al., 2017).

In Chapter 3, an improved method for analyzing ITS studies is presented and evaluated. The method accounts for heterogeneity in aggregated data caused by variations in the patient population by applying weights that reflect the sample size and variability at each of the time points. This proposed method known as weighted segmented linear regression (wSR) was used with the aim of reducing bias and improving precision and statistical power. Extensive simulations were conducted to evaluate performance of our proposed method under many different scenarios, mimicking real data sets. Comparative analysis using extensive simulations were also done to compare performance of our method with that of the traditional SLR. Established performance criteria such as bias, mean squared error (MSE), level, and power are used in our comparison. We also illustrated the usefulness of our proposed method by using real world data.

The method presented in Chapter 3 accounts for patient level heterogeneity, that is the variability across patients within sites at each time point. In Chapter 4, we present an

extension of this method to multisite or multiple baseline ITS studies. Here, the objective was to account for the variability introduced due to aggregation of data across multiple sites, in addition to the variability across patients (within site) which was considered in Chapter 3. Extensive comparative evaluation of our method with traditional SLR and the pooled method by Gebski et al. (2012) was performed empirically using a multisite ITS study involving 14540 patients across 14 hospitals in Ontario, Canada.

Chapter 5 provides a summary of the thesis, discusses potential areas for future research and some concluding remarks regarding the implications and contribution of this thesis to health research, in particular the contribution of the thesis in light of the recently observed increasing trend of ITS use in evaluating the impact of health policy, quality improvement initiatives and in assessing benefits and impacts of clinical practice guidelines.

References

Anchah, L., M. A. Hassali, M. S. H. Lim, M. I. M. Ibrahim, K. H. Sim and T. K. Ong (2017). "Health related quality of life assessment in acute coronary syndrome patients: the effectiveness of early phase I cardiac rehabilitation." Health and quality of life outcomes **15**(1): 10.

Anderson, T. W. (2011). The statistical analysis of time series, John Wiley & Sons.

Ansari, F., K. Gray, D. Nathwani, G. Phillips, S. Ogston, C. Ramsay and P. Davey (2003). "Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis." Journal of Antimicrobial Chemotherapy **52**(5): 842-848.

Bernal, J. L., S. Cummins and A. Gasparrini (2017). "Interrupted time series regression for the evaluation of public health interventions: a tutorial." International journal of epidemiology **46**(1): 348-355.

Bero, L., R. Grilli, J. Grimshaw, G. Mowatt, A. Oxman and M. Zwarenstein (2002). "Cochrane effective professional and organisation of care group." Cochrane Collaboration. The Cochrane Library(1).

Biglan, A., D. Ary and A. C. Wagenaar (2000). "The value of interrupted time-series experiments for community intervention research." Prevention Science **1**(1): 31-49.

Black, N. (1996). "Why we need observational studies to evaluate the effectiveness of health care." BMJ: British Medical Journal **312**(7040): 1215.

Box, G. E., G. M. Jenkins, G. C. Reinsel and G. M. Ljung (2015). Time series analysis: forecasting and control, John Wiley & Sons.

Briesacher, B. A., Y. Zhao, J. M. Madden, F. Zhang, A. S. Adams, J. Tjia, D. Ross-Degnan, J. H. Gurwitz and S. B. Soumerai (2011). "Medicare Part D and changes in prescription drug use and cost burden: national estimates for the Medicare population, 2000–2007." Medical care **49**(9): 834.

Campbell, D. T. and H. Riecken (1968). "Quasi-experimental design." International encyclopedia of the social sciences **5**: 259-263.

Campbell, D. T. and J. C. Stanley (1963). "Experimental and quasi-experimental designs for research." Handbook of research on teaching. Chicago, IL: Rand McNally.

Chace, M. J., F. Zhang, C. A. Fullerton, H. A. Huskamp, D. Gilden and S. B. Soumerai (2012). "Intended and unintended consequences of the gabapentin off-label marketing lawsuit among patients with bipolar disorder." The Journal of clinical psychiatry **73**(11): 1388-1394.

Chatfield, C. (2016). The analysis of time series: an introduction, CRC press.

Cook, T. D., D. T. Campbell and A. Day (1979). Quasi-experimentation: Design & analysis issues for field settings, Houghton Mifflin Boston.

Cook, T. D., D. T. Campbell and W. Shadish (2002). Experimental and quasi-experimental designs for generalized causal inference, Houghton Mifflin Boston.

Crosbie, J. (1993). "Interrupted time-series analysis with brief single-subject data." Journal of Consulting and Clinical Psychology **61**(6): 966.

Donnelly, N. J. (2005). The use of interrupted time series analysis to evaluate the impact of pharmaceutical benefits scheme policies on drug utilisation in Australia, University of New South Wales.

Eliopoulos, G. M., A. D. Harris, E. Lautenbach and E. Perencevich (2005). "A systematic review of quasi-experimental study designs in the fields of infection control and antibiotic resistance." Clinical Infectious Diseases **41**(1): 77-82.

Ewusie, J. E., E. Blondal, C. Soobiah, J. Beyene, L. Thabane, S. E. Straus and J. S. Hamid (2017). "Methods, applications, interpretations and challenges of interrupted time series (ITS) data: protocol for a scoping review." BMJ Open **7**(6): e016018.

Gillings, D., D. Makuc and E. Siegel (1981). "Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care." American journal of public health **71**(1): 38-46.

Grimshaw, J., M. Campbell, M. Eccles and N. Steen (2000). "Experimental and quasi-experimental designs for evaluating guideline implementation strategies." Family practice **17**(suppl_1): S11-S16.

Hacker, K., R. Penfold, F. Zhang and S. B. Soumerai (2012). "Impact of electronic health record transition on behavioral health screening in a large pediatric practice." Psychiatric Services **63**(3): 256-261.

Harris, A. D., J. C. McGregor, E. N. Perencevich, J. P. Furuno, J. Zhu, D. E. Peterson and J. Finkelstein (2006). "The use and interpretation of quasi-experimental studies in medical informatics." Journal of the American Medical Informatics Association **13**(1): 16-23.

Jandoc, R., A. M. Burden, M. Mamdani, L. E. Lévesque and S. M. Cadarette (2015). "Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations." Journal of clinical epidemiology **68**(8): 950-956.

Kastner, M., A. M. Sawka, J. Hamid, M. Chen, K. Thorpe, M. Chignell, J. Ewusie, C. Marquez, D. Newton and S. E. Straus (2014). "A knowledge translation tool improved osteoporosis disease management in primary care: an interrupted time series analysis." Implementation Science **9**(1): 109.

Kirkland, K. B., K. A. Homa, R. A. Lasky, J. A. Ptak, E. A. Taylor and M. E. Splaine (2012). "Impact of a hospital-wide hand hygiene initiative on healthcare-associated infections: results of an interrupted time series." BMJ Qual Saf: qhc-2012-000800.

Lu, C. Y., F. Zhang, M. D. Lakoma, J. M. Madden, D. Rusinak, R. B. Penfold, G. Simon, B. K. Ahmedani, G. Clarke and E. M. Hunkeler (2014). "Changes in antidepressant use

by young people and suicidal behavior after FDA warnings and media coverage: quasi-experimental study." Bmj **348**: g3596.

Lucasey, B. (2002). "Quasi-experimental design." Orthopaedic Nursing **21**(1): 56-57.

Mead, E. L., R. Cruz-Cano, D. Bernat, L. Whitsel, J. Huang, C. Sherwin and R. M. Robertson (2016). "Association between Florida's smoke-free policy and acute myocardial infarction by race: A time series analysis, 2000–2013." Preventive medicine **92**: 169-175.

Melhuish, E., J. Belsky, A. H. Leyland, J. Barnes and N. E. o. S. S. R. Team (2008). "Effects of fully-established Sure Start Local Programmes on 3-year-old children and their families living in England: a quasi-experimental observational study." The Lancet **372**(9650): 1641-1647.

Metcalfe, A. V. and P. S. Cowpertwait (2009). Introductory time series with R.

Mol, P. G., J. E. Wieringa, P. V. NannanPanday, R. O. Gans, J. E. Degener, M. Laseur and F. M. Haaijer-Ruskamp (2005). "Improving compliance with hospital antibiotic guidelines: a time-series intervention analysis." Journal of Antimicrobial Chemotherapy **55**(4): 550-557.

Moreno-Torres, I., J. Puig-Junoy and J. M. Raya (2011). "The impact of repeated cost containment policies on pharmaceutical expenditure: experience in Spain." The European Journal of Health Economics **12**(6): 563-573.

O'Malley, P. M. and A. C. Wagenaar (1991). "Effects of minimum drinking age laws on alcohol use, related behaviors and traffic crash involvement among American youth: 1976-1987." Journal of studies on Alcohol **52**(5): 478-491.

Penfold, R. B. and F. Zhang (2013). "Use of interrupted time series analysis in evaluating health care quality improvements." Academic pediatrics **13**(6): S38-S44.

Preval, N., M. Keall, L. Telfar-Barnard, A. Grimes and P. Howden-Chapman (2017). "Impact of improved insulation and heating on mortality risk of older cohort members with prior cardiovascular or respiratory hospitalisations." BMJ open **7**(11): e018079.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies." International journal of technology assessment in health care **19**(4): 613-623.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies." Int J Technol Assess Health Care **19**(4): 613-623.

Reichardt, C. S. (2009). "Quasi-experimental design." The SAGE handbook of quantitative methods in psychology **46**: 71.

Rockers, P. C., P. Tugwell, J.-A. Røttingen and T. Bärnighausen (2017). "Quasi-experimental study designs series—paper 13: realizing the full potential of quasi-experiments for health research." Journal of clinical epidemiology **89**: 106-110.

Sen, A. and M. Srivastava (2012). Regression analysis: theory, methods, and applications, Springer Science & Business Media.

Shadish, W. R., T. D. Cook and D. T. Campbell (2002). Experimental and quasi-experimental designs for generalized causal inference, Wadsworth Cengage learning.

Shardell, M., A. D. Harris, S. S. El-Kamary, J. P. Furuno, R. R. Miller and E. N. Perencevich (2007). "Statistical analysis and application of quasi experiments to antimicrobial resistance intervention studies." Clin Infect Dis **45**(7): 901-907.

Sommers, B. D., S. K. Long and K. Baicker (2014). "Changes in mortality after Massachusetts health care reform: a quasi-experimental study." Annals of internal medicine **160**(9): 585-593.

Taljaard, M., J. E. McKenzie, C. R. Ramsay and J. M. Grimshaw (2014). "The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care." Implementation Science **9**(1): 77.

Wagenaar, A. C., R. G. Maybee and K. P. Sullivan (1988). "Mandatory seat belt laws in eight states: A time-series evaluation." Journal of Safety Research **19**(2): 51-70.

Wagenaar, B. H., K. Sherr, Q. Fernandes and A. C. Wagenaar (2015). "Using routine health information systems for well-designed health evaluations in low-and middle-income countries." Health policy and planning **31**(1): 129-135.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." Journal of clinical pharmacy and therapeutics **27**(4): 299-309.

Wang, J. J., W. Scott, G. Raphael and O. Jake (2013). A comparison of statistical methods in interrupted time series analysis to estimate an intervention effect. Australasian Road Safety Research, Policing and Education Conference.

Zhang, F., A. K. Wagner, S. B. Soumerai and D. Ross-Degnan (2009). "Methods for estimating confidence intervals in interrupted time series analyses of health interventions." J Clin Epidemiol **62**(2): 143-148.

Chapter 2

ITS Methods and Applications in Health Research

Summary

In this chapter, we present the Paper 1 of the thesis. There are numerous statistical methods that can be used to analyze data from ITS designs. However, there has not been any study to identify which methods are available and appropriate for different data types. In this chapter, we present an article, which is currently under review at "BMJ Open", where we conducted a scoping review of methods for analyzing ITS data in health research. This review was done as an initial step to explore the statistical methods for ITS analysis, examine the application of the methods, discuss the methodological differences and identify the gaps in existing methods. We classified the studies included in our review into methods and application papers based on the focus of the article. For the methods papers, we summarized the methods identified and narratively described their strengths and limitations. For the application papers, we summarized the data using summary statistics, mainly frequencies and percentages. We also identified the gaps and methodological deficiencies in existing methods.

Citation: Ewusie JE, Blondal E, Soobiah C, Beyene J, Thabane L, Straus S, Hamid JS. (2017). "Methods, Applications, Interpretations and Challenges of Interrupted Time Series (ITS) Data: Protocol for a Scoping Review." *BMJ Open*, 7(6), e016018.

Ewusie JE, Blondal E, Soobiah C, Beyene J, Thabane L, Straus S, Hamid JS. (2017). "Methods, Applications, Interpretations and Challenges of Interrupted Time Series (ITS) Data: a scoping review", *BMJ Open*, revision submitted.

Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review

Joycelyne E. Ewusie,[1] Charlene Soobiah,[2,3] Erik Blondal,[2,3] Joseph Beyene,[1] Lehana Thabane,[1,4] Sharon Straus, [2,5] Jemila S. Hamid[1,2,3]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
[2]Li Ka Shing Knowledge Institute of St Michael's Hospital, Toronto, Ontario, Canada
[3]Institute of Health Policy Management and Evaluation (IHPME), University of Toronto, Ontario, Canada
[4]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare, Hamilton, Ontario, Canada
[5]Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

Abstract

Objectives: Interrupted time series (ITS) designs are robust quasi-experimental designs commonly used to evaluate the impact of interventions and programs implemented in healthcare settings. This scoping review aims to 1) identify and summarize existing methods used in the analysis of ITS studies conducted in health research, 2) elucidate their strengths and limitations, 3) describe their applications in health research and 4) identify any methodological gaps and challenges. **Design:** Scoping review. **Data sources**: Searches were conducted in MEDLINE, JSTOR, PUBMED, EMBASE, CINAHL, Web of Science and the Cochrane Library from inception until September 2017. **Study selection:** Studies in health research involving ITS methods or reporting on the application of ITS designs. **Data extraction:** Screening of studies was completed independently and in duplicate by two reviewers. One reviewer extracted the data from relevant studies in consultations with a second reviewer. Results of the review were presented with respect to methodological and application areas, and data were summarized using descriptive statistics. **Results**: A total of 1389 articles were included, of which 98.27% (N=1365) were application papers. Segmented linear regression was the most commonly used method (26%, N=360). A small percentage (1.73%, N=24) were methods papers, of which 11 described either the development of novel methods or improvement of existing methods, 7 adapted methods from other areas of statistics, while 6 provided comparative assessment of conventional ITS methods. **Conclusion**: A significantly increasing trend in ITS use over time is observed, where its application in health research almost tripled within the last decade. Several statistical methods are available for analyzing ITS data. Researchers should consider the types of data and validate the required assumptions for the various methods. There is a significant methodological gap in ITS analysis involving aggregated data, where analyses involving such data did not account for heterogeneity across patients and hospital settings.

Keywords: Interrupted time series, segmented linear regression, ARIMA, limitations, methods, scoping review

2.1. Introduction

Quasi-experimental designs (QEDs) refer to non-randomized designs that are used to evaluate the effect of interventions and programs (Shadish et al., 2002). Interrupted Time Series (ITS) design is considered the strongest among QEDs and is a powerful tool used for evaluating the impact of interventions and programs implemented in healthcare settings (Wagner et al., 2002; Penfold and Fang, 2013). With this design, outcomes are measured at different time points before and after implementing an intervention, allowing the change in level and trend of outcomes to be compared, to evaluate intervention effects.

ITS designs are applied in a wide range of applications and healthcare settings. ITS designs are commonly used to evaluate quality improvement initiatives and infection control programs in hospitals (Wagner et al., 2002; Harris et al., 2006; Gebski et al., 2012; Kastner et al., 2014; Taljaard et al., 2014; Liu et al., 2017). A recently published systematic review also shows that ITS designs are being increasingly used in drug utilization research (Jandoc et al., 2015). With the increasing research focus on knowledge translation (KT) and evidence-based medicine (EBM), and the growing importance of the uptake of best evidence into clinical practice, ITS analysis has become a common tool used in assessing the effect of clinical practice guidelines and recommendations (Matowe et al., 2002; Cortoos et al., 2011; Dowell et al., 2012; Dayer et al., 2015; Etchepare et al., 2017). For instance, ITS design was used to look at de-adoption of tight glycemic control in critically ill patients across 113 intensive care units (ICUs) in the USA following publication of a randomized trial showing that tight glycemic control can increase mortality (Niven et al., 2015). In another study, ITS design was used in evaluating the impact of Otolaryngology

Head and Neck Surgery (OHNS) guidelines on perioperative care process and patient outcomes in children undergoing tonsillectomy (Mahant et al., 2015). In recent years ITS studies have also been included in Cochrane systematic reviews, done by the Effective Practice and Organization of Care Review group (Bero et al., 2002), indicating their usefulness in studies of organizational and practice change interventions.

ITS designs are generally highly regarded for their rigor (compared to the traditional before and after studies) and are arguably considered the optimal approach in evaluating the impact of hospital-wide interventions and new policies implemented nationwide. ITS designs allow us to statistically test potential biases such as, autocorrelation, seasonality, stationarity, heteroskedasticity, history, maturation and random fluctuations (Ramsay et al., 2003). Moreover, to increase the internal validity of a study, ITS designs can be modified to include a nonequivalent dependent variable or a control outcome unaffected by the intervention to control for possible concurrent events (Bernal et al., 2017).

There are several statistical methods that can be used in analyzing data from an ITS design. The decision on what method to use is often based on several factors including type of outcome (e.g. continuous, binary or count), distribution (e.g. Gaussian, Skewed) (Shardell et al., 2007), assumptions such as autocorrelation or seasonality (Nelson, 1998; Carroll, 2006), the number of groups/sites included in the ITS design (single site vs multi-site analysis) (Linden and Adams, 2011), or the inclusion of a control group (Boel et al., 2016). Despite availability of various statistical methods for ITS analysis, limitations exist in practical applications, where researchers frequently ignore checking the assumptions

and the various factors influencing the optimality of the different methods (Ewusie et al., 2017). As such, some ITS data are currently being analyzed using inadequate methods, and hence leading to inaccurate results, erroneous or misleading conclusions, and potentially affecting patient care.

Therefore, in order to reduce bias, increase precision and enhance statistical power it is imperative that researchers use appropriate methods for their analysis (Ramsay et al., 2003). There is also a need for the different statistical methods for ITS analysis to be identified and compared to inform researchers of the available methods and inform future research regarding the strengths and limitations of the methods as well as identify methodological gaps which will pave the way for improvements in ITS design and analysis.

In this study, we conducted a scoping review with the aim of 1) identifying and systematically summarizing available methods used in the analysis of ITS studies, 2) elucidating the strengths and limitations of existing methods, 3) identifying potential methodological gaps and 4) providing an extensive review of the applications of ITS designs and analysis in health research.

2.2. Methods

We performed the scoping review using the methods outlined by Arksey and O'Malley and the Joanna Briggs Handbook for conducting systematic scoping reviews (Arksey and O'Malley, 2005; Peters et al., 2015). Scoping review methods were used, since our aim was to identify the methods that have been used to analyze ITS data, provide an overview of their strengths and limitations and evaluate the frequency of their application without an in

depth assessment of the application papers with respect to their clinical content, risk of bias as well as quality of the design and analysis (Arksey and O'Malley, 2005). We developed a protocol for this scoping review based on the PRISMA-P guidelines (Moher et al., 2015). The protocol is published in a peer-reviewed journal (Ewusie et al., 2017) and a brief description of our scoping review methods is outlined here.

2.2.1. Search Strategy

We searched electronic databases, MEDLINE, JSTOR, PUBMED, EMBASE, CINAHL, Web of Science and the Cochrane Library from inception until September 2017 for relevant articles and conference abstracts. An experienced information specialist worked with JEE and JSH to further refine the search strategy presented in the study protocol. We also contacted methodological experts in the field of ITS for any difficult-to-locate and unpublished materials. Additionally, we scanned the references of included articles for other potentially relevant articles. We restricted our search to studies that were reported in English.

2.2.2. Eligibility Criteria

Our eligibility criteria were based on the recommendations by the Effective Practice and Organization of Care (EPOC) Cochrane Group (Bero et al., 2002). All health-related studies that reported on the development, adaptation, comparison or application of ITS design or methods were included. Studies with at least 3 time points before and after the intervention and had a clearly defined time point or period within which the intervention was implemented were included. Furthermore, studies were included if the outcome was

objectively measured. We excluded ITS studies that had less than 3 time points per period to be consistent with the recommendations found in literature (Bero et al., 2002).

2.2.3. Study selection and data collection

The search results and potentially relevant full-text articles were screened independently by two reviewers and in duplicate (JEE, CS and EB). To ensure reliability, a calibration exercise was performed prior to screening of titles and abstracts as well as full-text articles. Data extraction was completed by JEE in consultations with JSH. All conflicts that arose were resolved by discussion or consultation through a third reviewer. Since this is a scoping review, we did not access the risk of bias of included studies (Arksey and O'Malley, 2005; Peters et al., 2015).

2.2.4. Data extraction and synthesis

Data were extracted from all included studies. The studies were classified as methodological papers if they reported on the development, adaptation, important additions to common statistical methods, description or comparison of statistical methods for ITS design. Studies were classified as application papers if they used an ITS design or analysis for assessing intervention effect. Data extracted for methodological papers included; article characteristics (e.g. title, author, year of publication), type of method, description of method, type of outcome the method is developed for, assumptions involved as well as strengths and limitations of the method as described by the authors. For application papers, the data extracted include; article characteristics, field of application (e.g. clinical, pharmaceutical, etc.), setting (e.g. single site, multi-site), statistical method used as

reported by the authors, number of time points, and the assumptions (e.g. presence of autocorrelation) checked or tested.

Results from the review were summarized in two categories: methodological articles and articles on applications of ITS analysis. For methodological papers, we used narrative review to describe the methodological processes or analysis strategies. We also discussed the methods with respect to their strengths and weaknesses, the type of outcome assessed, the assumptions made, the type of software used and the limitations, if stated. Data from application papers were summarized using frequencies and percentages and we also provided tables and figures to depict the distribution.

2.3. Results

A total of 9278 articles were returned from the database search, of which 6890 were identified as unique articles. After the initial screening of titles and abstracts, 2262 articles were found to be relevant. After the full text screening of these articles, 1389 articles were included for the review. A detailed illustration of the study flow is provided in Figure 2.1.

2.3.1. Study Characteristics

Of the 1389 papers, about 2% (N=24) were classified as methods papers and the majority (98%, N=1365) were classified as application papers. An overall description of papers included in the scoping review, using broad categories, is provided in Table 2.1. As shown in the table, the methods papers were presented in three categories, with most of the articles presenting development of novel methods (46%, N=11). For application papers, ITS design has been used most frequently in clinical research (N=621, 46%) and in population and

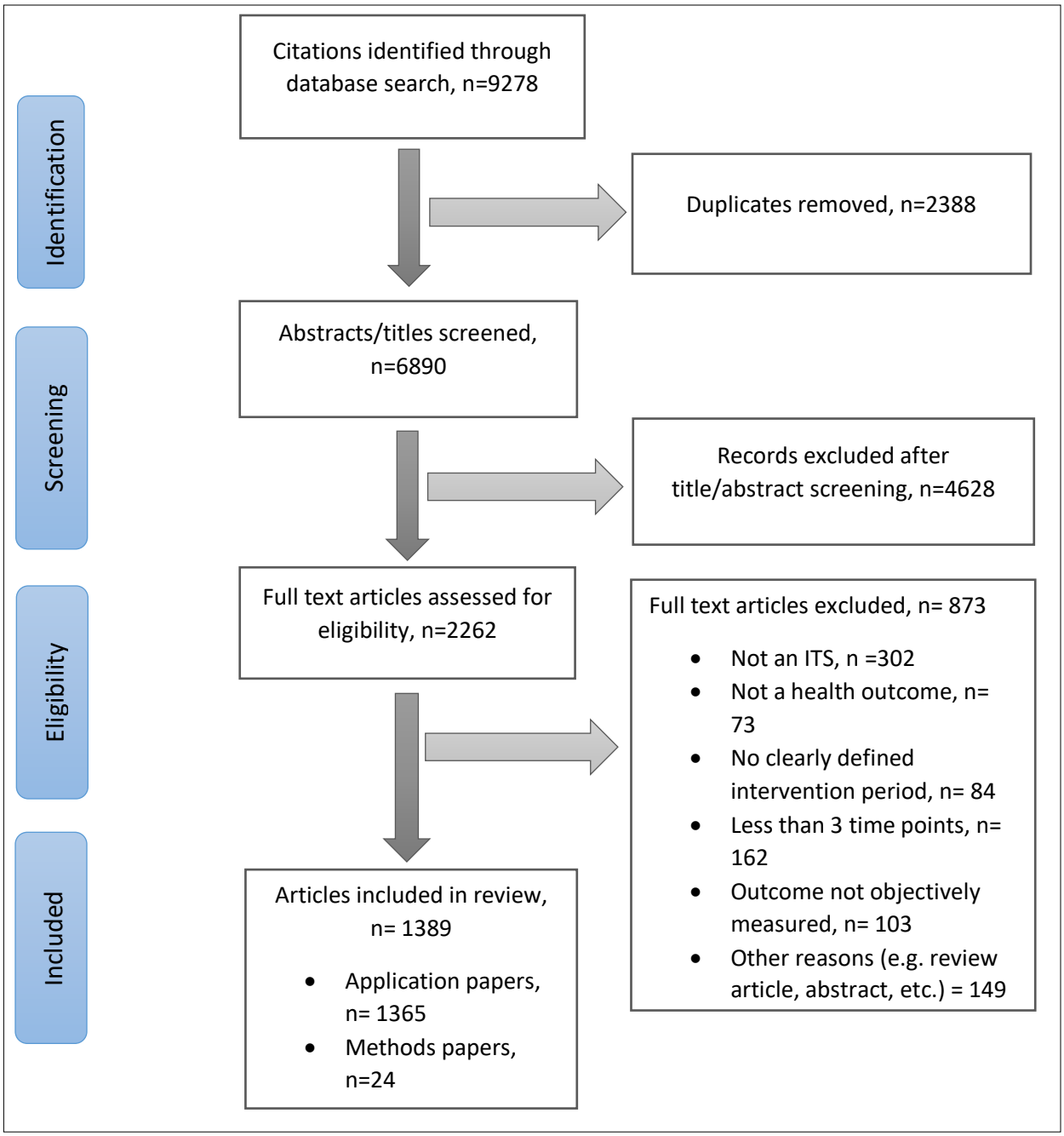

Figure 2.1: Flow chart outlining the search and review process, the records identified, included and excluded as well as the reasons for exclusion.

Table 2.1*:* Description of studies included in the review with respect to methods used in the analysis of ITS data

Characteristic	Number of Studies included in the review, N, (%) * N= 1389
Methods Papers	24 (1.73)
Novel Methods	11 (45.8)
Method adaptation and important Contribution	7 (29.2)
Method comparison	6 (25.0)
Application Papers	1365 (98.27)
Field of application	
Clinical	621 (45.5)
Pharmaceutical	238 (17.4)
Guideline Implementation	69 (5.1)
Public Health/policy	437 (32.0)
Setting/Design	
Single site	353 (25.9)
Multiple Baseline/multi-site	392 (28.7)
Controlled ITS	237 (17.4)
National (Population study)	383 (28.1)
Statistical methods used	
Segmented regression	
Segmented regression using linear models	360 (26.4)
Segmented regression using GLM, GEE or GAM**	261 (19.1)
Segmented regression using ARIMA	268 (19.6)
Non-segmented regression	110 (19.6)
Non-regression methods e.g. t-test	82 (6.0)
Difference in Differences	17 (1.2)
Unspecified	267 (19.6)
Type of outcome	
Continuous	131 (9.6)
Count	1029 (75.4)
Binary	205 (15.0)
Number of time points	
Less than 16 (or < 8 per period)	141 (10.3)
At least 16 (or ≥ 8 per period)	634 (46.5)
At least 50	590 (43.2)

Autocorrelation checked	
Yes	812 (59.5)
Other biases checked	
Yes	607 (44.5)
Specific biases* **	
Seasonality	407 (67.1)
Stationarity	290 (47.8)
Heteroskedasticity	123 (20.3)
Confounding	203 (33.4)
Clustering	68 (11.2)
Presentation of ITS results	
Figures	414 (30.3)
Tables	105 (7.7)
Both	804 (58.9)
None	42 (3.1)

*All percentages are out of the total number of corresponding papers
**ITS- interrupted time series; GLM- Generalized Linear Models; GAM- Generalized Additive Models; GEE – Generalized Estimating Equation
***The frequencies and percentages are out of the total of 607.

public health research (N=437, 32%). It is also applied frequently in multiple baseline or multiple site studies (N=392, 29%).

2.3.2. Review of Statistical Methods for ITS Analysis

Among the 11 methods papers proposing novel statistical methods for ITS analysis, a considerable percentage (55%, N=6) were developed for either controlled ITS designs (Linden and Adams, 2011; Fretheim et al., 2015; Pechlivanoglou et al., 2015) or multiple baseline ITS designs (Velicer, 1994; Gebski et al., 2012; Huitema et al., 2014), where multiple baseline indicates that the intervention is introduced in different localities (e.g. sites, units) at different times. A significant percentage (25%, N=6) of the methods papers provided a comparative analysis of existing ITS methods where empirical analysis and/or simulated data were used to compare performance of the different methods (Shardell et al.,

2007; Nunes et al., 2011; Harrington and Velicer, 2015; Kontopantelis et al., 2015; Andersson Hagiwara et al., 2016; Burke et al., 2016).

2.3.2.1. Papers on Novel ITS Methods

A detailed summary of the 11 articles with respect to the descriptions of methods' process, assumptions made, statistical software used, strengths and limitations have been provided in the Appendix (Table A1). The methods include, the extended time series model introduced by Sun et al. (2012) which allows the identification of immediate and gradual changes in an outcome of interest. This time series model comprises a stochastic component and a structural or intervention component. The stochastic component is modelled using ARIMA and the intervention component is divided into 4 subcomponents to detect: 1) pre-intervention marginal change in outcome, 2) post intervention marginal change in outcome, 3) short-term change in outcome and 4) additional impact of intervention over the observed period (Sun et al., 2012). Their method also allows the assessment of different post intervention time points. However, it is not possible to assess the effect of concomitant interventions that may affect the outcome.

Duncan & Duncan (2004) applied the latent growth curve modelling approach to pooled interrupted time series data. In their approach, they added a growth curve model that captures both the intercept and slope differences over the baseline and intervention periods. The model gives linear and average level change for both periods. This model can be applied to designs with multiple units and different time points, time spacing and growth functions (Duncan and Duncan, 2004). It is indicated that their method can be used for a short time series (>2 time points per period) and allows the evaluation of treatment effects

in the presence of different covariates. However, they also mentioned that their method is not optimum when the time series is lengthened since the linear growth form may not adequately define the series.

The robust interrupted time series method, using a two-stage approach, was proposed by Cruz et al. (2017). The first stage involves identifying the change point and the second stage involves performing formal tests for differences in the correlation structure and variability between pre- and post- intervention (Cruz et al., 2017). For this method, the time of intervention is assumed to not necessarily be the same as the time at which the effect of intervention initiate (i.e. the change point). The authors demonstrated that their method performed better than traditional segmented regression with regards to mean squared error.

Kong et al. (2012) developed an extended logistic regression method for count and proportion data. The authors showed how the incidence rate or proportion of an outcome of interest can be estimated while accounting for seasonality and serial correlation. The method incorporates harmonic functions to account for seasonality and a first-order autoregressive model to account for serial autocorrelation (Kong et al., 2012). The authors also described the different estimation procedures for estimating the parameters and their associated variances. The authors indicated that their method performed better than the conventional segmented linear regression model.

For studies involving multiple baseline ITS designs, Gbeski et al. (2012) introduced the pooled and stacked ITS method of data analysis, where the data are from several units

within one site (e.g. hospital) or across multiple sites. The data are thus aggregated across the units (or sites) to obtain the overall intervention effect (Gebski et al., 2012). For the pooled analysis, they fitted a separate segmented regression model for each unit and then calculated the weighted average of the individual estimates of the parameters where the weights were the inverse variances of the estimates obtained from the individual fitted models. The stacked analysis approach involved fitting a single segmented regression model but accounting for the unit effect by incorporating a variable to represent the units using one unit as reference. Although their methods accounts for unit effect and heterogeneity among units, having large number of units may lead to over dispersion or an increase in Type 1 error due to the increase in number of required parameters.

Huitema et al. (2014) implemented an extended time series regression model which involved a within unit and between unit analysis. For the within unit analysis which is performed for designs with multiple units and interventions that are rolled out over time and at different time points per unit, they incorporated a continuous variable which they called a penetration variable to evaluate the extent of penetration of the intervention. The penetration variable was calculated as the ratio of number of units with intervention to overall number of units, and it ranged from 0 to 1 (Huitema et al., 2014). Their method is appropriate for interventions that are not introduced in full after baseline. Moreover, the authors indicated that if an effect is identified, the function estimated from the regression can be used to predict the outcome from the degree of intervention penetration. This method, however, does not account for unit effect and heterogeneity across units.

Similarly, Velicer (1994) introduced the pooled time series analysis approach to analyze multiple baseline time series data, where data are combined from different units and comparison between the units is allowed. The method involved using a patterned transformation matrix and a design matrix. The transformation matrix transforms serially dependent variables to independent variables. The design matrix is defined based on the parameters of interest (level and trend) and differences between units. (Velicer, 1994). The author stated that the pooled time series method has the advantage of being easily implemented in existing computer programs with little modification; it can also be adapted to other modelling approaches. For instance, instead of the general transformation matrix used in the proposed method, the ARIMA (1,0,0) transformation matrix can be used for most cases, and thus the method can be implemented in R or SAS. This method, however, requires substantial understanding of transformation, which is a barrier to uptake of the proposed method. The method does not also account for heterogeneity across units.

The propensity score-based weighted interrupted time series analysis was introduced by Linden & Adams (2011) for controlled series. The method allows the pre-intervention characteristics (level and slope) for the treatment and control groups to be comparable. This method uses standard regression techniques, and hence easy to implement (Linden and Adams, 2011). However, the control groups must have significant overlap with treatment group, in terms of basic characteristics, for the weighting to be effective. Fretheim et al. (2005) presented a method which is similar to the difference-in-differences approach. As with all controlled ITS studies, this analysis method allows for the detection of anomalous effects and co-interventions (Fretheim et al., 2015). The authors

indicated that this method requires more than 6 time points per period for reliable estimation of the regression coefficients.

The state-space method was introduced by Pechlivanoglou et al. (2015) for controlled interrupted time series. The method comprises an observation equation and a state equation. The state equation contains the level and trend parameters that are affected by the intervention and allowed to vary over time (Pechlivanoglou et al., 2015). The main advantage of this method is the ability to capture effects of co-interventions by the addition of other variables. This method is however unable to test the assumption of comparability of the control group and thus overall results might be suboptimal if groups are not comparable. Park (2012) implemented the intervened time series central mean subspace, which is a non-parametric approach to analyzing ITS data. The method is an extension of the central mean subspace in time series to a nonparametric intervention analysis. The dynamic reduction technique is used to analyze the time series using interventions as a covariate (Park, 2012). The authors showed that the method is a viable alternative to ARIMA and does not require model specification. The authors stated that a large number of observations is required to ensure model accuracy, however, they did not explicitly specify how large this is required to be.

2.3.2.2. Papers on Improved or Adapted Methods

A total of 7 methods papers (29%) presented a description of existing statistical methods, used in other areas, that are adapted to ITS analysis (Gillings et al., 1981; Wagner et al., 2002; Yau et al., 2004; Taljaard et al., 2014) or presented important contributions to the adapted methods (Huitema and McKean, 2007; Zhang et al., 2009; Zhang et al., 2011). These articles include a publication by Gillings et al. (1981) who described the implementation of ITS design and the application of segmented linear regression (SLR) methodology in health services research to assess mortality trends, following the implementation of a regionalized perinatal care program (Gillings et al., 1981). The authors fitted both SLR and a single non-linear regression to the data; and they showed that SLR explains a greater proportion of variation in the data compared to fitting a single non-liner regression. Moreover, it is much easier to interpret. They also concluded that the SLR is relevant for analyzing ITS data, when the errors are independent.

Wagner et al. (2002) presented the SLR approach, which is also known in statistical literature as piece-wise regression. The authors described how this method can be applied to evaluate the effect of a policy or educational intervention implemented to improve quality of medication use (Wagner et al., 2002). The authors highlighted various factors that should be considered, when using SLR. These include testing for autocorrelation, seasonality as well as identifying outliers and their effects on the results. The authors indicated that SLR model is a robust modelling technique that allows the estimation of dynamic changes in outcomes following interventions intended to change the use of medication.

Yau et al. (2004) described the extension of the zero inflated Poisson (ZIP) regression model to handle time series of count data with excess zeros in application to occupational health (Yau et al., 2004). In their paper, the authors discussed how the method enables the evaluation of an occupational intervention using population level aggregated count data containing extra zeros. The authors stated the assumptions under which the ZIP model can be used, for instance, the existence of perfect and imperfect states of the underlying process. They also proposed the ZIP mixed autoregression model which is an extension of the ZIP model to account for the dependency between serially observed counts.

Studies presenting important contributions to existing methods include Zhang et al. (2009). The authors considered SLR estimators for absolute and relative changes in outcomes and provided methods for obtaining confidence intervals (CIs) of the estimators (Zhang et al., 2009). Similarly, Zhang et al. (2011) considered design aspects of SLR and provided simulation-based power calculations. They indicated that SLR models with 24 or more time points have more than 80% power to detect an effect size of 1 or greater, depending on the degree of autocorrelation and number of parameters to be estimated (Zhang et al., 2011). Huitema & McKean (2007) proposed a new method for dealing with autocorrelation in ITS analysis, in the context of SLR. Their method is a form of portmanteau test that identifies autocorrelated errors generated from processes of higher-order autoregressive models. Unlike the conventional portmanteau tests, their method provides more satisfactory small sample properties and better inferential properties by

accounting for the biases associated with the autocorrelation estimator and the error variance estimators (Huitema and McKean, 2007).

2.3.2.3. Papers on Methodological Comparisons

Six of the studies focused on comparative evaluation of statistical methods used in the analysis of ITS studies (Shardell et al., 2007; Nunes et al., 2011; Harrington and Velicer, 2015; Kontopantelis et al., 2015; Andersson Hagiwara et al., 2016; Burke et al., 2016). Shardell et al. (2007) performed empirical evaluation to compare the 2 group tests, SLR analysis and time series analysis in terms of the characteristics, assumptions, strengths and limitations of the methods. The authors used data from a study conducted with the objective of evaluating the impact of a hospital-based intervention to reduce antimicrobial infection rates and overall length of stay. The conclusions drawn from this study indicated that all three methods can be effectively used for ITS data analysis (Shardell et al., 2007). For statistical validity, however, researchers must consider the research question, the data requirements and modeling assumptions to obtain high quality and unbiased results. The authors also provided guidance on examining the requirements for each of the methods.

Nunes et al. (2011) compared two statistical methods, seasonal autoregressive integrated moving average (ARIMA) and cyclical regression models, used to analyze ITS data on mortality. Based on their results, the seasonal ARIMA was found to produce non-autocorrelated residuals in the weeks where there were excess deaths due to influenza epidemic. The seasonal ARIMA model had a lower residual mean square than the cyclical regression model. These results led to a lower 95% confidence limit for the estimates obtained from the seasonal ARIMA model, making it more efficient (Nunes et al., 2011).

Hagiwara et al. (2016) compared SLR analysis and statistical process control (SPC) for evaluating the longitudinal effects of quality improvement interventions. They discussed the implementation processes of both methods, their differences, strengths and limitations. Based on their analysis of the empirical data, the authors concluded that SLR analysis was more statistically robust than SPC analysis in comparing the effectiveness of different interventions, while the SPC was more appropriate for controlling a process since it is relatively easy to conduct and interpret (Andersson Hagiwara et al., 2016). Kontopantelis et al. (2015) compared different regression modeling approaches for ITS analysis, based on the level of complexity in their implementation, which they labeled as basic, advanced and expert levels. The authors provided a description of the method processes, the assumptions made, and the technical details of each modeling technique using secondary data obtained from a quality and outcomes framework performance research (Kontopantelis et al., 2015). In the same way, Burke et al. (2016) compared three methods; SLR, ARIMA, and standardized incidence ratio (SIR) and compared the methods' ability to determine whether Amerithrax influenced patient utilization. The results obtained based on the three methods were discussed and their differences were reported. The authors, however, stated that, they were not able to successfully evaluate the impact of the intervention using the three methods due to the limited level of detail in the timeframes (Burke et al., 2016). They, therefore, concluded that granularity of timeframes is as important as number of data points in a time series analysis.

2.3.3. Review of the Application Studies involving ITS Design and Analysis

A large percentage (98%, N=1365) of the studies in our review involve application papers, where studies reported use of ITS design and/or analysis. The most common application area identified was clinical research (46%, N=621) followed by public health or health policy applications (32%, N=437) and pharmaceutical research (17%, N= 238) (Table 2.1). In clinical research, ITS designs are mostly utilized in studies involving interventions to reduce anti-microbial resistance and in other quality improvement studies (40%, N=246). Common public health and policy research areas, where ITS designs are used, include assessing the effect of mass media campaign programs and national or regional regulations such as smoking ban and traffic laws (27%, N=120). Pharmaceutical research areas, where ITS designs, are frequently used include studies on drug dose evaluation and drug regulatory approval.

Our review reveals that there is an overall increasing trend in ITS use (Figure 2.2). Most of the application papers using ITS are published in recent years, where 197 studies using ITS analysis were published in 2016 alone compared to only 69 in 2011 (approximately 185 percent point increase in the last 5 years); and just 27 in 2006 indicating more than 600 percent point increase in publications within the last decade.

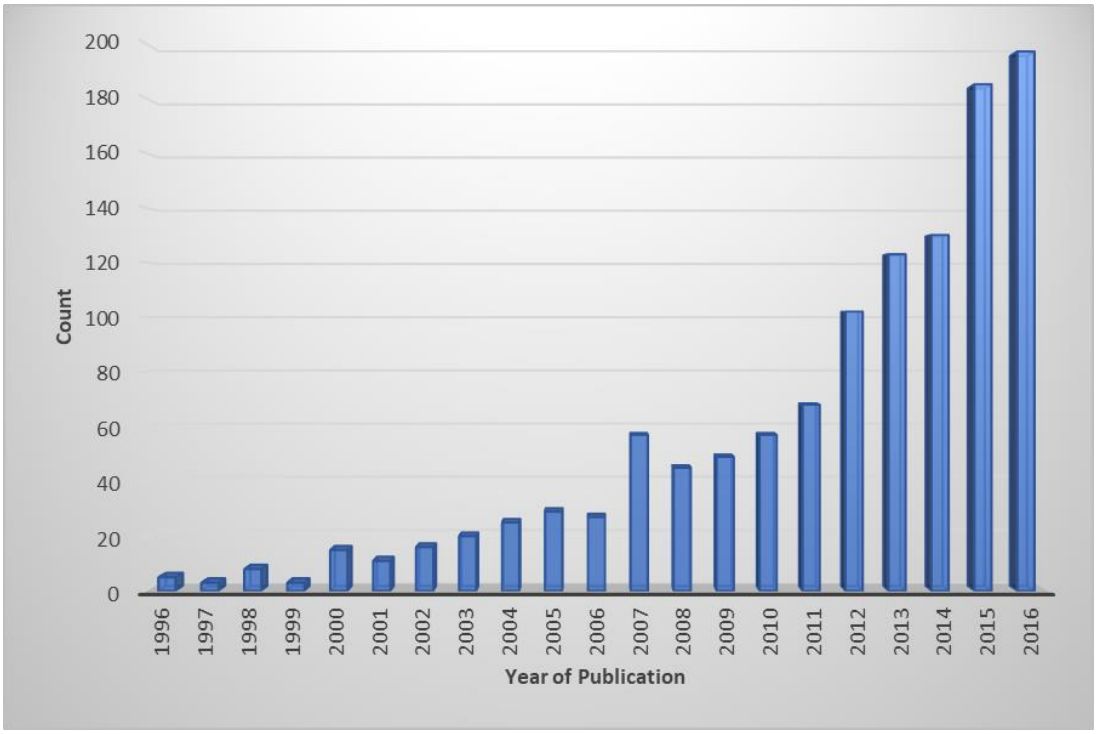

Figure 2.2: Trend of interrupted time series application papers over the last two decades.

Evaluation of the application papers with respect to the methods utilized in the statistical analysis shows that most of the studies (65%, N=889) reported segmented regression (SR) techniques. Specifically, SR using ordinary linear regression modelling (SLR) was the most commonly applied method (26%, N=360), followed by ARIMA modeling (20%, N=268).

SR using other modelling approaches such as generalized linear model (e.g. Poisson and Binomial models) was implemented in 19% (N=261) of the studies. Although, most of the studies used segmented regression approaches (e.g. SLR) to analyze data, there was considerable shortfall in the application of these methods. For instance, among the studies

that used ARIMA models, approximately 23% (N= 62) did not meet its at least 50 time points requirement for adequate analysis. Similarly, 10% (N=65) of the studies that used SLR had less than 8 time points per period. Of the studies that reported using SR modelling approaches, other than ARIMA, 66% (N=407) stated that they tested and where necessary accounted for the presence of autocorrelation in their data.

Overall, about 60% (N=812) of the included studies checked or accounted for autocorrelation while approximately 45% (N=607) checked or accounted for other biases, such as stationarity (48%, N=290), heteroskedasticity (20%, N=123), confounding variables such as age and sex (30%, N=203), clustering effect (11%, N=68) and seasonality (67%, N=407) when suitable. Furthermore, a considerable percentage (21%, N=132) of the studies inappropriately used SLR or ARIMA to analyze ITS data with count outcome. This was often done without checking for linearity or validating assumptions such as the normality assumption. For considerable percentage of the application studies (48%, N=658), the outcome of interest was summarized as rate.

A sizeable number of studies (6%, N=82), that implemented an ITS design, inappropriately used statistical methods that do not account for time. The methods used include ANOVA, t-tests and Chi-square tests. Approximately 20% (N=267) of the studies included in our review reported using ITS analysis without specifying the modelling approach they implemented in performing analysis.

Substantial number of the studies included in our review implemented a multiple baseline (multi-site) design (29%, N=392) and a considerable number (17%, N=237)

implemented controlled ITS designs. For studies that implemented controlled ITS designs, a very small percentage (7%, N=16) reported using methods such as difference in differences in analyzing their data, which is a method arguably considered as non-ITS. For studies that reported the statistical software used (70%, N=958), a considerable number performed their analysis using either SAS (37%, N=353) or STATA (36%, N=348) statistical software. PROC AUTOREG in SAS is the most commonly reported software package (15%, N=52) among the SAS users.

2.4. Discussion

Our review demonstrated the increasing trend in ITS use in health research, where publications using ITS designs and analysis have almost tripled within the last decade. This increase might be attributed to advancements in the field of implementation science in recent years. Our findings also show that one of the factors contributing to this increasing trend is the use of routinely collected administrative data to answer important healthcare related questions. These administrative data provide relatively inexpensive options compared to using resource-intensive prospective data. ITS designs are also being increasingly used in evidence synthesis and uptake, where ITS studies are used to evaluate the impact of evidence-based recommendations and policies, clinical practice guidelines (CPGs) as well as publication of other important evidence. With the growing knowledge and awareness of health ethics combined with other advances in patient care delivery, feasibility and ethical considerations might be contributing factors to the increasing use of ITS designs.

Our results show that ITS designs are being implemented mostly in clinical research and public health and policy applications. The observed increase in ITS applications is consistent with findings in a recent systematic review focused on ITS use in drug utilization research (Jandoc et al., 2015). Our review also revealed that most of the studies used either a multiple baseline design or a controlled ITS design. A few of the studies also used a control outcome or non-equivalent dependent variable. These design adaptations have been recommended in literature to increase the internal validity of ITS designs by controlling for threats such as time-varying confounders (Wagner et al., 2002; Bernal et al., 2017). Moreover, even without a control group, ITS designs can address important biases such as history and maturation by having multiple observations before and after the intervention (Wagner et al., 2002). We identified several methods that have been utilized in ITS data analysis. The most common methods were SR analysis and ARIMA. We also identified some new and improved methods that have been developed in recent years such as the robust ITS, the extended ITS and the pooled ITS analysis.

This scoping review highlights some major limitations in the implementation of ITS designs and their analysis. First, a considerable number of studies analyzed their data inappropriately. For instance, despite having sufficient time points per phase (study period), some studies used methods such as ANOVA and t-tests when performing analyses. These methods do not account for any underlying secular trend or autocorrelations often present in the data, thus making such methods suboptimal (Ramsay et al., 2003). Similarly, although the recommendation by the Effective Practice and Organization of Care (EPOC) Cochrane Group (Bero et al., 2002) suggest that ITS designs should have at least 3 time

points per period for inclusion in their reviews, methodological literature examined in this scoping review show that a minimum of 8 time points per period (or 50 time points for ARIMA models) is required to gain sufficient power in estimating the regression coefficients (Penfold and Fang, 2013). Nevertheless, our findings revealed that some of the studies, despite fewer time points, used time series regression techniques to analyze their data thus making their results underpowered. Second, for studies that used appropriate time series regression approaches, over 40% of them did not test or account for potential biases that might be present in the data due to, for instance, autocorrelation, seasonality, or heteroskedasticity. The effects of correlated errors or serial dependency in time series data have been highlighted in literature (Hartmann et al., 1980; Nelson, 1998; Ramsay et al., 2003; Carroll, 2006). Testing or accounting for such potential biases in time series data is imperative to ensure that the standard errors are not biased, and that the significance of the intervention is not overestimated.

There are some methodological gaps in current and frequently used ITS methods that were identified in our review. Although a few of the gaps have been acknowledged in previous studies (Ramsay et al., 2003; Zhang et al., 2009; Gebski et al., 2012), the limited methodological advances in the area remain a problem. One major issue, for instance, is the use of SR to analyze aggregated data (Wagner et al., 2002). For almost all identified ITS studies, data at a given time point are often aggregated or summarized across different patients or participants. Further, with the multiple baseline (or multi-site) designs, the final data are pooled across the sites or units, resulting in another level of data aggregation. Thus, for both single and multi-site designs, data at each time point are estimated and hence

associated with an imprecision due to variability of patient outcomes or variability across sites at a given time point. Nevertheless, our results show that a substantial number of these studies used SLR or ARIMA models, which do not account for this imprecision and hence may lead to biased (aggregation bias) or suboptimal results. The method proposed by Gebski and colleagues, (Gebski et al., 2012) accounts for the heterogeneity across sites using meta-analytic approaches. However, the method does not account for imprecision introduced due to aggregation across patients within the same site. Moreover, pooling the intercept and slope estimates leads to loss of information (and hence loss of power), since summarized data rather than individual patient data are used. Similarly, other methods that have been developed for multi-site studies do not account for the heterogeneity across the patients or across the different sites/units (Velicer, 1994; Huitema et al., 2014). Hence, there is a need for methods that account for the variability at both the patient and site levels.

Another issue highlighted in our review is the lack of guidance in design aspect of ITS studies and lack of clarity on the adequate sample size per time point. Although a sample size of 100 at each time point is stated as desirable for acceptable level of variability of estimates (Wagner et al., 2002), we found no information in literature concerning the adequate sample size required at each time point to provide optimal results in analyzing ITS data. Since sample size calculation is essential in estimating effect sizes, it will be imperative for future research to consider studies that will help determine the minimum sample size required at each time point for adequate evaluation of intervention effects.

Yet another issue is the slow uptake and implementation of the methods that have been developed in recent years. While only a few recently developed methods were

identified in our review (N=11), we noticed that none of the recently developed methods have been applied by other researchers in the analysis of ITS data, even if they had similar data structures. This issue is, however, not unique to ITS methods; the problem of dissemination and uptake of new statistical methods into health research has been discussed extensively in literature (Pullenayegum et al., 2016). The reasons for the limited uptake of new methods include: 1) the lack of statistical expertise to implement the method and understand the findings especially for methods such as extended time series model (Huitema et al., 2014) and pooled time series model (Velicer, 1994) that require substantial mathematical or statistical knowledge 2) lack of software packages to implement the method and 3) lack of awareness of the existence of the method. This knowledge translation gap emphasizes the objective of this scoping review of available methods used in ITS data analysis.

Based on these observed shortfalls in the analysis of ITS data and the gaps present in current methods used in ITS analysis, we make the following recommendations: 1) future researchers and reviewers of ITS studies must carefully consider the assumptions and requirements of the various statistical methods to ensure that conclusions about intervention effects are not spurious, 2) future research comparing available methods should be based on simulation studies to assess the bias, mean square error, and power. This will enable readers to confidently use methods based on their preferences, 3) ITS method developers should provide user friendly software packages as was done by Cruz et al. (2017), to ensure effective and efficient uptake of their new methods in biomedical research, and 4) there is the need for the provision of some form of guideline to help

researchers and novice users decide on the appropriate use of ITS methods with respect to both design and analysis. Currently, two articles have provided a checklist for researchers and reviewers of ITS designs (Ramsay et al., 2003; Jandoc et al., 2015). In addition, reporting guidelines will also be helpful in ensuring consistency in ITS articles. These guidelines will have the potential to reduce heterogeneity as well as standardize analysis, interpretation and reporting of ITS studies. This standardization will not only enhance the methodological rigor, but also facilitate more appropriate assessment of study quality, evidence gathering and evidence synthesis through meta-analyses of ITS studies.

In our study, we considered specific databases and studies published in English to limit the scope of our research, thus we may not have retrieved all studies particularly those on the application of ITS designs. We acknowledge that this may be a possible limitation of our study, however, we believe that our sample is representative of ITS methods utilized in biomedical research since we have captured all the methods papers that meet our eligibility criteria. Furthermore, the application papers were needed to describe the overall trend in ITS use, thus the number of papers included in this study is sufficient to provide a reliable estimate and achieve our aim. Moreover, we expect the magnitude of increase in ITS to be higher than what is reported in this paper if non-English studies and applications outside health research are included in our review. This further underscore the importance of ITS methods. Finally, we acknowledge the lack of detail in the presentation of the application papers as another limitation of this review. However, we believe that the level of detail presented is enough to answer our objective. Nonetheless, we highlight that the

studies found in our literature search can be used as bases for future systematic review with an in-depth evaluation of application articles.

The findings from this study accentuate the need for improved methods for design and analysis of ITS studies such as when the data are aggregated per time point. This review serves as a first step towards developing standard guidelines for ITS studies and for filling the methodological gaps identified in current literature.

References

Andersson Hagiwara, M., B. Andersson Gare and M. Elg (2016). "Interrupted Time Series Versus Statistical Process Control in Quality Improvement Projects." J Nurs Care Qual **31**(1): E1-8.

Arksey, H. and L. O'Malley (2005). "Scoping studies: towards a methodological framework." International journal of social research methodology **8**(1): 19-32.

Bernal, J. L., S. Cummins and A. Gasparrini (2017). "Interrupted time series regression for the evaluation of public health interventions: a tutorial." International journal of epidemiology **46**(1): 348-355.

Bero, L., R. Grilli, J. Grimshaw, G. Mowatt, A. Oxman and M. Zwarenstein (2002). "Cochrane effective professional and organisation of care group." Cochrane Collaboration. The Cochrane Library(1).

Boel, J., V. Andreasen, J. O. Jarlov, C. Ostergaard, I. Gjorup, N. Boggild and M. Arpi (2016). "Impact of antibiotic restriction on resistance levels of Escherichia coli: a controlled interrupted time series study of a hospital-wide antibiotic stewardship programme." J Antimicrob Chemother **71**(7): 2047-2051.

Burke, L. K., C. P. Brown and T. M. Johnson (2016). "Historical Data Analysis of Hospital Discharges Related to the Amerithrax Attack in Florida Historical Data Analysis of Hospital Discharges Related to the Amerithrax Attack in Florida." Perspectives in Health Information Management: 1-16.

Carroll, N. (2006). Application of segmented regression analysis to the Kaiser Permanente Colorado critical drug interaction program. Proceedings of the Fifthteenth Annual Western Users of SAS Software Conference.

Cortoos, P.-J., C. Gilissen, P. G. Mol, F. Van den Bossche, S. Simoens, L. Willems, H. Leenaers, L. Vandorpe, W. E. Peetermans and G. Laekeman (2011). "Empirical management of community-acquired pneumonia: impact of concurrent A/H1N1 influenza pandemic on guideline implementation." Journal of antimicrobial chemotherapy **66**(12): 2864-2871.

Cruz, M., M. Bender and H. Ombao (2017). "A robust interrupted time series model for analyzing complex health care intervention data." Statistics in Medicine **29**: 29.

Dayer, M. J., S. Jones, B. Prendergast, L. M. Baddour, P. B. Lockhart and M. H. Thornhill (2015). "Incidence of infective endocarditis in England, 2000–13: a secular trend, interrupted time-series analysis." The Lancet **385**(9974): 1219-1228.

Dowell, D., L. H. Tian, J. A. Stover, J. A. Donnelly, S. Martins, E. J. Erbelding, R. Pino, H. Weinstock and L. M. Newman (2012). "Changes in fluoroquinolone use for gonorrhea following publication of revised treatment guidelines." American journal of public health **102**(1): 148-155.

Duncan, T. E. and S. C. Duncan (2004). "A latent growth curve modeling approach to pooled interrupted time series analyses." Journal of Psychopathology and Behavioral Assessment **26**(4): 271-278.

Etchepare, F., E. Pambrun, H. Verdoux and M. Tournier (2017). "Trends in patterns of antidepressant use in older general population between 2006 and 2012 following publication of practice guidelines." International journal of geriatric psychiatry **32**(8): 849-859.

Ewusie, J. E., E. Blondal, C. Soobiah, J. Beyene, L. Thabane, S. E. Straus and J. S. Hamid (2017). "Methods, applications, interpretations and challenges of interrupted time series (ITS) data: protocol for a scoping review." BMJ Open **7**(6): e016018.

Fretheim, A., F. Zhang, D. Ross-Degnan, A. D. Oxman, H. Cheyne, R. Foy, S. Goodacre, J. Herrin, N. Kerse, R. J. McKinlay, A. Wright and S. B. Soumerai (2015). "A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation." J Clin Epidemiol **68**(3): 324-333.

Gebski, V., K. Ellingson, J. Edwards, J. Jernigan and D. Kleinbaum (2012). "Modelling interrupted time series to evaluate prevention and control of infection in healthcare." Epidemiology and Infection **140**(12): 2131-2141.

Gillings, D., D. Makuc and E. Siegel (1981). "Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care." American journal of public health **71**(1): 38-46.

Harrington, M. and W. F. Velicer (2015). "Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies." Multivariate Behav Res **50**(2): 162-183.

Harris, A. D., J. C. McGregor, E. N. Perencevich, J. P. Furuno, J. Zhu, D. E. Peterson and J. Finkelstein (2006). "The use and interpretation of quasi-experimental studies in medical informatics." Journal of the American Medical Informatics Association **13**(1): 16-23.

Hartmann, D. P., J. M. Gottman, R. R. Jones, W. Gardner, A. E. Kazdin and R. S. Vaught (1980). "Interrupted time-series analysis and its application to behavioral data." Journal of Applied Behavior Analysis **13**(4): 543-559.

Huitema, B. E. and J. W. McKean (2007). "An improved portmanteau test for autocorrelated errors in interrupted time-series regression models." Behavior Research Methods **39**(3): 343-349.

Huitema, B. E., R. Van Houten and H. Manal (2014). "Time-series intervention analysis of pedestrian countdown timer effects." Accid Anal Prev **72**: 23-31.

Jandoc, R., A. M. Burden, M. Mamdani, L. E. Lévesque and S. M. Cadarette (2015). "Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations." Journal of clinical epidemiology **68**(8): 950-956.

Kastner, M., A. M. Sawka, J. Hamid, M. Chen, K. Thorpe, M. Chignell, J. Ewusie, C. Marquez, D. Newton and S. E. Straus (2014). "A knowledge translation tool improved

osteoporosis disease management in primary care: an interrupted time series analysis." Implement Sci **9**: 109.

Kong, M., A. Cambon and M. J. Smith (2012). "Extended Logistic Regression Model for Studies with Interrupted Events, Seasonal Trend, and Serial Correlation." Communications in Statistics-Theory and Methods **41**(19): 3528-3543.

Kontopantelis, E., T. Doran, D. A. Springate, I. Buchan and D. Reeves (2015). "Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis." bmj **350**: h2750.

Linden, A. and J. L. Adams (2011). "Applying a propensity score-based weighting model to interrupted time series data: Improving causal inference in programme evaluation." Journal of Evaluation in Clinical Practice **17**(6): 1231-1238.

Liu, B., J. E. Moore, U. Almaawiy, W.-H. Chan, S. Khan, J. Ewusie, J. S. Hamid, S. E. Straus and M. O. Collaboration (2017). "Outcomes of Mobilisation of Vulnerable Elders in Ontario (MOVE ON): a multisite interrupted time series evaluation of an implementation intervention to increase patient mobilisation." Age and ageing **47**(1): 112-119.

Mahant, S., M. Hall, S. L. Ishman, R. Morse, V. Mittal, G. M. Mussman, J. Gold, A. Montalbano, R. Srivastava and K. M. Wilson (2015). "Association of national guidelines with tonsillectomy perioperative care and outcomes." Pediatrics **136**(1): 53-60.

Matowe, L., C. R. Ramsay, J. Grimshaw, F. Gilbert, M.-J. Macleod and G. Needham (2002). "Effects of mailed dissemination of the Royal College of Radiologists' guidelines on general practitioner referrals for radiography: a time series analysis." Clinical radiology **57**(7): 575-578.

Moher, D., L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle and L. A. Stewart (2015). "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement." Systematic reviews **4**(1): 1.

Nelson, B. K. (1998). "Statistical methodology: V. Time series analysis using autoregressive integrated moving average (ARIMA) models." Acad Emerg Med **5**(7): 739-744.

Niven, D. J., G. D. Rubenfeld, A. A. Kramer and H. T. Stelfox (2015). "Effect of published scientific evidence on glycemic control in adult intensive care units." JAMA internal medicine **175**(5): 801-809.

Nunes, B., I. Natario and M. L. Carvalho (2011). "Time series methods for obtaining excess mortality attributable to influenza epidemics." Stat Methods Med Res **20**(4): 331-345.

Park, J. H. (2012). "Nonparametric approach to intervention time series modeling." Journal of Applied Statistics **39**(7): 1397-1408.

Pechlivanoglou, P., J. E. Wieringa, T. de Jager and M. J. Postma (2015). "The effect of financial and educational incentives on rational prescribing. A state-space approach." Health Econ **24**(4): 439-453.

Penfold, R. B. and Z. Fang (2013). "Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements." Academic Pediatrics: S38-44.

Peters, M. D., C. M. Godfrey, H. Khalil, P. McInerney, D. Parker and C. B. Soares (2015). "Guidance for conducting systematic scoping reviews." Int J Evid Based Healthc **13**(3): 141-146.

Pullenayegum, E. M., R. W. Platt, M. Barwick, B. M. Feldman, M. Offringa and L. Thabane (2016). "Knowledge translation in biostatistics: a survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods." Statistics in medicine **35**(6): 805-818.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies." Int J Technol Assess Health Care **19**(4): 613-623.

Shadish, W. R., T. D. Cook and D. T. Campbell (2002). Experimental and quasi-experimental designs for generalized causal inference, Wadsworth Cengage learning.

Shardell, M., A. D. Harris, S. S. El-Kamary, J. P. Furuno, R. R. Miller and E. N. Perencevich (2007). "Statistical analysis and application of quasi experiments to antimicrobial resistance intervention studies." Clin Infect Dis **45**(7): 901-907.

Sun, P., J. Chang, J. Zhang and K. H. Kahler (2012). "Evolutionary cost analysis of valsartan initiation among patients with hypertension: a time series approach." J Med Econ **15**(1): 8-18.

Taljaard, M., J. E. McKenzie, C. R. Ramsay and J. M. Grimshaw (2014). "The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care." Implement Sci **9**: 77.

Velicer, W. F. (1994). "Time series models of individual substance abusers." NIDA Res Monogr **142**: 264-301.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." J Clin Pharm Ther **27**(4): 299-309.

Yau, K. K., A. H. Lee and P. J. Carrivick (2004). "Modeling zero-inflated count series with application to occupational health." Comput Methods Programs Biomed **74**(1): 47-52.

Zhang, F., A. K. Wagner and D. Ross-Degnan (2011). "Simulation-based power calculation for designing interrupted time series analyses of health policy interventions." Journal of clinical epidemiology **64**(11): 1252-1261.

Zhang, F., A. K. Wagner, S. B. Soumerai and D. Ross-Degnan (2009). "Methods for estimating confidence intervals in interrupted time series analyses of health interventions." J Clin Epidemiol **62**(2): 143-148.

Chapter 3

Weighted Segmented Regression for Analyzing Interrupted Time Series Data

Summary

Segmented linear regression, (SLR) as it is currently used, is prone to bias when applied to aggregated data from interrupted time series (ITS) studies. This is because the variability introduced by aggregation of the data, for instance across participants per time point, is not accounted for, which leads to imprecision and loss of power to detect clinically meaningful differences. This chapter includes an article currently under review in the "BMC Medical Research Methodology" journal. In the article, we developed and presented a novel weighted segmented regression (wSR) method where the variability associated with the aggregated data is included as weights in the SLR model. We performed extensive simulations to evaluate the performance of our method and compared the results with the segmented linear regression (SLR) method. We also illustrated the application of our method using real world data.

Citation: Ewusie JE, Beyene J, Thabane L, Straus S, Hamid JS. "An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient Heterogeneity Using Weighted Analysis", *BMC Medical Research Methodology*, under review.

An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient Heterogeneity Using Weighted Analysis

Joycelyne Ewusie[1], Joseph Beyene[1], Lehana Thabane[1,2], Sharon E Straus[3,4], Jemila S. Hamid[1,3,*]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
[2]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare, Hamilton, Ontario, Canada
[3]Li Ka Shing Knowledge Institute of St Michael's Hospital, Toronto, Ontario, Canada
[4]Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

Abstract

Background: Interrupted time series (ITS) design is commonly used to evaluate the impact of interventions in healthcare settings, where data across multiple time points before and after intervention is collected to compare changes in level and trend of important outcomes. Segmented linear regression (SLR) is the most commonly used statistical method and has been shown to be useful in practical applications involving ITS designs. Nevertheless, the method suffers from some limitations when applied to aggregated data. SLR is prone to aggregation bias, which leads to imprecision and loss of power to detect clinically meaningful differences. The main objective of this article is to present a novel weighted segmented regression (wSR) method, where variability across patients within the healthcare facility and across time points is incorporated through weights in ITS analysis. **Methods:** We present the methodological framework for the proposed wSR, provide optimal weights associated with data at each time point and discuss relevant statistical inference. We conduct an extensive simulation to evaluate performance of our method and provide comparative analysis with the segmented linear regression (SLR). We use established performance criteria such as bias, mean square error (MSE), level of significance and statistical power. Illustrations using real data set is also provided. **Results:** In most simulation scenarios considered, the wSR method produced estimators that are uniformly more precise and relatively less biased compared to the traditional SLR. The wSR approach is also associated with higher statistical power in the scenarios considered. The performance difference is much larger for data with high variability (heterogeneity) across patients within healthcare facilities. **Conclusion:** ITS is the most useful in evaluating the impact of programs in healthcare settings, where data are often aggregated across patients. The weighted method proposed here allows us to account for the heterogeneity in the patient population, and hence leading to increased accuracy and power across all scenarios. Nevertheless, in studies involving highly variable data and limited sample size, both methods lack adequate power. We, therefore, recommend researchers to carefully design their studies and determine their sample size by incorporating heterogeneity in the patient population.

Keywords: Weighted segmented regression, Simulation study, Heteroskedasticity, Method comparison, Interrupted time series.

3.1. Background

Interrupted time series (ITS) designs are regularly used to examine the effect of population and hospital level interventions and evaluate the impact of programs implemented in various healthcare settings (Albu et al., 2017; Aregawi et al., 2017; Berrevoets et al., 2017; Bond et al., 2017; Gutacker et al., 2017; Rosich Marti et al., 2017; Taylor et al., 2017). ITS designs are arguably the best approach for evaluating post hoc effects of policies or programs (e.g. quality improvement programs) using administrative or other routinely collected data (Bobo et al., 2014; Pow et al., 2015; Balkhi et al., 2016; Yang et al., 2017). In ITS designs, outcomes are measured pre- and post-interventions and changes in levels and trends are compared over time between the different periods to evaluate intervention impacts (Ramsay et al., 2003). ITS designs have become increasingly more popular in implementation science due to their relatively higher rigor and optimality in terms of its statistical performance, compared to before and after analysis. ITS designs are used in several applications in implementation science; including in assessing the effects of quality improvement initiatives (e.g. hand hygiene) and prevention programs in healthcare settings (Hanson et al., 1015; Michielutte et al., 2000; Ansari et al., 2003; Gebski et al., 2012; Kastner et al., 2014; Taljaard et al., 2014; Langford et al., 2016; Liu et al., 2017).

A recent scoping review showed that application of ITS in health research has significantly increased in recent years, where 197 studies using ITS analysis were published in 2016 compared to 57 in 2010, and only 27 in 2006 (Ewusie et al., 2018). The review also revealed that ITS analysis was mostly applied in clinical and public health research as well as in drug utilization and drug policy studies. Another, more focused

systematic review showed that ITS is being increasingly used in drug utilization research, where an average of 15 articles were published per year since the year 2000 compared to only a total of 17 articles published between 1984 and 2000 (Jandoc et al., 2015).

Several factors may be attributed to this increase in use of ITS designs. One of the potential contributing factors of this increasing trend in the use of ITS designs relates to the use of routinely collected big administrative data to answer important healthcare-related questions, as these data sets provide relatively inexpensive options for using available data rather than using resource-intensive prospective data (Dayer et al., 2015; Wagenaar et al., 2015; Carter et al., 2016; Graves et al., 2016). For instance, Walley et al. (2013) evaluated the impact of state supported overdose education and nasal naloxone distribution (OEND) programs on rates of opioid related death from overdose and acute care utilization in Massachusetts. They used routine data from the Massachusetts Registry of Vital Records and Statistics and from the Massachusetts Division of Health Care Finance and Policy respectively (Walley et al., 2013). A second contributing factor for the recent increase in ITS use is associated with its application in evidence synthesis and uptake, where ITS designs are being used to evaluate the impact of evidence dissemination and implementation such as that from clinical practice guidelines (Bussieres et al., 1501; Buyle et al., 2010; Judge et al., 2015). For instance, ITS design was employed to examine longitudinal trends in screening mammography utilization and the presence of any changes in utilization following the release of the 2009 U.S Preventive Services Task Force updated guideline (Jiang et al., 2015). Mackie et al. (2016) also assessed the rate of infective

endocarditis hospitalizations in Canada before and after the publication of the 2007 American Heart Association guidelines (Mackie et al., 2016).

Several statistical methods are currently available for analyzing data from ITS studies. The two most commonly used methods are segmented linear regression (SLR) and autoregressive integrated moving average (ARIMA), with SLR being the most popular (Zhang et al., 2009; Jandoc et al., 2015; Ewusie et al., 2018). Despite their frequent application, these methods are associated with several limitations. ITS studies are often conducted in healthcare settings, where outcomes are aggregated across patients within the healthcare facilities. Data are, therefore, associated with heterogeneity due to differences among the patient populations (Gillings et al., 1981; Taljaard et al., 2014). Moreover, when evaluating healthcare related programs and initiatives, studies are usually conducted across multiple sites. This leads to further aggregation of outcomes across sites, introducing yet another level of heterogeneity due to, for instance, clinical practice variation. Current ITS methods, including the most popular method SLR, do not account for the heterogeneity among patients and across sites. The methods, therefore, suffer from aggregation bias, imprecision and loss of power (Ramsay et al., 2003; Shardell et al., 2007; Gebski et al., 2012; Harrington and Velicer, 2015; Ewusie et al., 2018).

In this paper, we proposed a weighted segmented regression (wSR) method to account for variability among patients within healthcare facilities. We hypothesized that our method produces less biased and more precise estimators and higher statistical power, compared to traditional unweighted approaches. Since the method assigns less weights to

data with higher variability, we further hypothesized that our method will be robust in the presence of outliers.

3.2. Methods

Consider an ITS study aimed at evaluating the impact of an intervention in improving patient outcomes (e.g. mental health services utilization rates, incidence of a major cardiac event or the cost of hospitalization). Suppose an outcome *x* is measured at several time points pre- and post-interventions. Without loss of generality, and for clarity of presentation, we assume that the outcome is continuous. We would like to highlight that the approach is valid for other outcomes (e.g. count and binary), although the summary measures used might be different (e.g. proportion of patients for binary outcomes). We also assume, without loss of generality, that the outcome can be represented as a linear function of time. When the distribution of the outcome is not normal, generalized linear regression (e.g. logistic, Poisson or multinomial) can be used or outcomes can be transformed using appropriate transformations.

Let x_{ij} denote the value of the outcome for patient j at time i. The outcome at each time point i, is first summarized by calculating the average across all the patients at time i, i.e.,

$$y_i = \frac{\Sigma_j x_{ij}}{N}, \tag{3.1}$$

where N is the total number of patients at time i. The SLR model can be described as;

$$y_i = \beta_0 + \beta_1 * t + \beta_2 * Int + \beta_3 * t * Int + \varepsilon_i, \tag{3.2}$$

where t represents the time at which the outcome is measured or estimated, Int is a dichotomous variable representing the intervention period (e.g. 0 for pre- and 1 for post-intervention periods) and $t * Int$ is an interaction variable represents the interaction of time and intervention. The parameters β_0 and β_1 , respectively, estimate the baseline intercept (level) and slope (trend), while β_2 and β_3 represent the changes in level and trend between the two intervention periods. The model in (3.2) can be re-written in a matrix presentation as;

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{3.3}$$

where Y denotes a vector of outcomes $(y_1, y_2, y_3, \ldots y_k)$, measured at times 1, 2, …, k; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_p)$ denotes a vector representing the regression coefficients (parameters) associated with time, the intervention periods, interactions as well as other potential time varying covariates that might influence the outcome ; $\boldsymbol{X}$ is a p x k matrix of variables associated with the regression coefficients; and $\boldsymbol{\varepsilon}$ is the error vector.

Using ordinary of least squares (OLS) approach, the estimators of $\boldsymbol{\beta}$ can be derived by minimizing the error sum of squares, that is:

$$\sum \varepsilon_i^2 = (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}), \tag{3.4}$$

which provides the well-established estimator,

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y. \tag{3.5}$$

The parameter estimates for SLR were, therefore, calculated as such. Standard errors, 95% confidence intervals as well as statistics for testing statistical significance can be found in

standard statistical literature and are implemented in most statistical software (Sen and Srivastava, 2012; Draper and Smith, 2014).

3.2.1. Weighted Segmented Regression

In minimizing the squared error, OLS method is an optimal approach, when the variance of the error terms is constant (also known as homoscedastic); that is $var(\varepsilon_i) = \sigma^2$ (Draper and Smith, 2014). Specifically, the variance associated with the data at each time point is assumed to be constant and is due to random error alone; $r(y_i) = var(\varepsilon_i) = \sigma^2$. For studies involving ITS in healthcare settings, however, data at a given time point is aggregated hence not observed (e.g. average prescriptions per week across all patients). Consequently, data at each time point is associated with some level of statistical uncertainty (γ_i^2), which is a function of heterogeneity in the patient population and the sample size used in the aggregation (i.e., number of patients). The variability in the data is therefore not constant across time, rather, $var(y_i) = var(\varepsilon_i) = \sigma_i^2; and\ \sigma_i^2 = \sigma^2 * \gamma_i^2$. Additionally, since the data are aggregated across different sets of patients and different sample sizes, the associated variance at each time point is different, that is $\gamma_i^2 \neq \gamma_j^2$. As a result, the method of least squares will not be optimal both from the perspective of estimation (does not lead to minimum variance unbiased estimators, resulting in imprecise estimates and wider confidence intervals) as well as from the perspective of hypothesis testing (tends to have inflated Type I error and decreased power).

To address this limitation of the OLS method, we propose the weighted SR (wSR) model and use weighted least square (WLS) method to perform statistical inference. The

WLS approach is a modification of the OLS, where data are weighted to account for differences in variance. Recall the generalized regression model described in equation (3.3), we now minimize the weighted sum of squares to derive the estimators in a wSR frame work:

$$\begin{aligned} \sum w_i \varepsilon_i^2 &= (\boldsymbol{Y} - \boldsymbol{X\beta})' \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X\beta}), \\ &= \boldsymbol{Y'WY} - \boldsymbol{Y'WX\beta} - \boldsymbol{\beta' X'WY} + \boldsymbol{\beta' X'WX\beta}, \end{aligned} \tag{3.6}$$

leading to estimators,

$$\widehat{\boldsymbol{\beta}}_w = (\boldsymbol{X'WX})^{-1} \boldsymbol{X'WY}, \tag{3.7}$$

where $\boldsymbol{W}$ represents a diagonal matrix with vector, $\boldsymbol{w} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_k})$. We hypothesize that the weighted estimators will lead to smaller bias and mean squared error and the test statistic calculated using the weighted estimates will have smaller type one error and higher power. The standard error and the 95% confidence interval (CI) are calculated accordingly based on these weighted estimates and the formulas are similar to what is found in literature (Sen and Srivastava, 2012).

3.2.2. The Weights

There are several approaches to determining the weights associated with each of the observations at different time points. The optimality of the method depends on the choice of the weights. Several strategies are proposed in literature. For instance, the weight can be based on experience or prior information using some theoretical model or using residuals of the model (Sen and Srivastava, 2012). If unknown, one approach for estimating weights is using the squared residual regression approach, which is a two-stage estimation

procedure (Gilstein and Leamer, 1983; Tsai and Wu, 1990; Sen and Srivastava, 2012). In this study, we follow approaches used in evidence synthesis and meta-analysis; and use between patient variability as weights. Thus, our weights are calculated as the inverse of the sum of the between patient variability (within time) and the random error (variability across time). That is, if $\gamma_i{}^2$ represents the variability across patients at time point i; and σ^2 represents the variance associated with the random error as shown in model (3.3) then the weight is therefore $w_i = \frac{1}{\sigma^2+\gamma_i{}^2}$.

3.3. Simulations

3.3.1. Description of Simulations

We performed extensive simulations, where we considered a total of 350 scenarios with various parameter values, variabilities and sample sizes. The range of the parameters used are provided in Table 3.1.

Table 3.1: Range of parameters considered in the simulation study

Settings	**Values**
Sample Size (n)	10 to 100
Variance (γ^2) across patients per time point	Ranged from 2 to 12 Low: < 5, Medium: 5-8, High: >8
Change in Level (β_2) Change in Slope (β_3)	Ranged from 0.1 to 2.1 Ranged from 0.1 to 1.2
Number of time points	50 (25 per each period)

For all the scenarios, data at each time point were generated independently from the normal distribution, $N(\mu, \sigma^2)$, where the mean (μ) was modeled as a function of time and

intervention in a SR setup as $\mu_i = \beta_0 + \beta_1 * t_1 + \beta_2 * Int + \beta_3 * t_2$, where $t_1 \in \{1, \dots, 50\}$, $Int \in \{0,1\}$ for pre and post intervention respectively and

$$t_2 = \begin{cases} 0 & for\ pre - intervention \\ 1\ to\ 25 & for\ post - intervetion \end{cases}.$$

The simulation parameters associated with the slopes and intercepts ($\beta_i s$) for pre- and post-intervention were obtained from a real data set to mimic real life scenarios (Liu et al., 2017). Data for some of the scenarios were generated by assuming a constant error variance across the different time points, that is $var(\varepsilon_i) = \sigma^2, for\ all\ \ i\ in\ \{1, \dots, 50\}$. To assess the effect of heterogeneity of the error variances, we also generated data from the normal distribution with different levels of variance, $var(y_i) = var(\varepsilon_i) = {\sigma_i}^2$, where ${\sigma_i}^2 = \sigma^2 + \gamma_i^2\ and\ \gamma_i^2 \neq \gamma_j^2$. The variance we considered ranged from 2 to 12 in increments of 0.5 and variances were generated from a uniform distribution: $\sigma_i^2 \sim U(1, k), k \in \{2, \dots, 12\}$. For the constant variance assumption, $k = 3$ is used. The variance values were chosen to reflect expected differences in variance between patients at a time point. As indicated in Table 3.1, different sample sizes per time point were considered to investigate the performance of the methods for small, moderate and large sample sizes. Without loss of generality, we considered a total of 50 time points for all simulation scenarios. We would like to highlight that the sample sizes and the number of time points used were chosen to be representative of what is common seen in literature (Ewusie et al., 2018) and account for any underlying or seasonal trends.

At each time point, we generated independent data with different seeds. Data were then summarized (aggregated) as a mean to represent the nature of ITS data in healthcare

settings. The traditional SLR and the wSR proposed in this paper were then applied to the aggregated data and statistical inference performed using the two methods. We calculated empirical bias, empirical square error (MSE), Type I error rate and power based on 5000 simulations. The number of simulations were chosen to ensure that the estimate produced is within 5% accuracy of the true parameter value for change in level, with a 0.05 level of significance. To account for the sampling variability, the simulations were also repeated 50 times to generate 50 estimates of the empirical bias and level. We performed the simulations using R version 3.4.3 software.

3.3.2. Simulation Results

3.3.2.1. Bias and MSE

Our simulation results show that both the wSR and unweighted SR produced estimates with minimal to no bias, when the within patient variance was small compared to data with large variability across patients (Figure 3.1). Although, the bias values did not show any clear pattern between the two methods for the different levels of variance, the wSR method (weighted ITS) overall produced estimates with relatively less bias compared to the SLR (unweighted ITS) approach (Figure 3.1; Table 3.2). Further, our results show that both methods produced estimates with large bias (inaccurate estimates) when the sample size was small ($n<30$) compared to when the sample size was large ($n\geq50$).

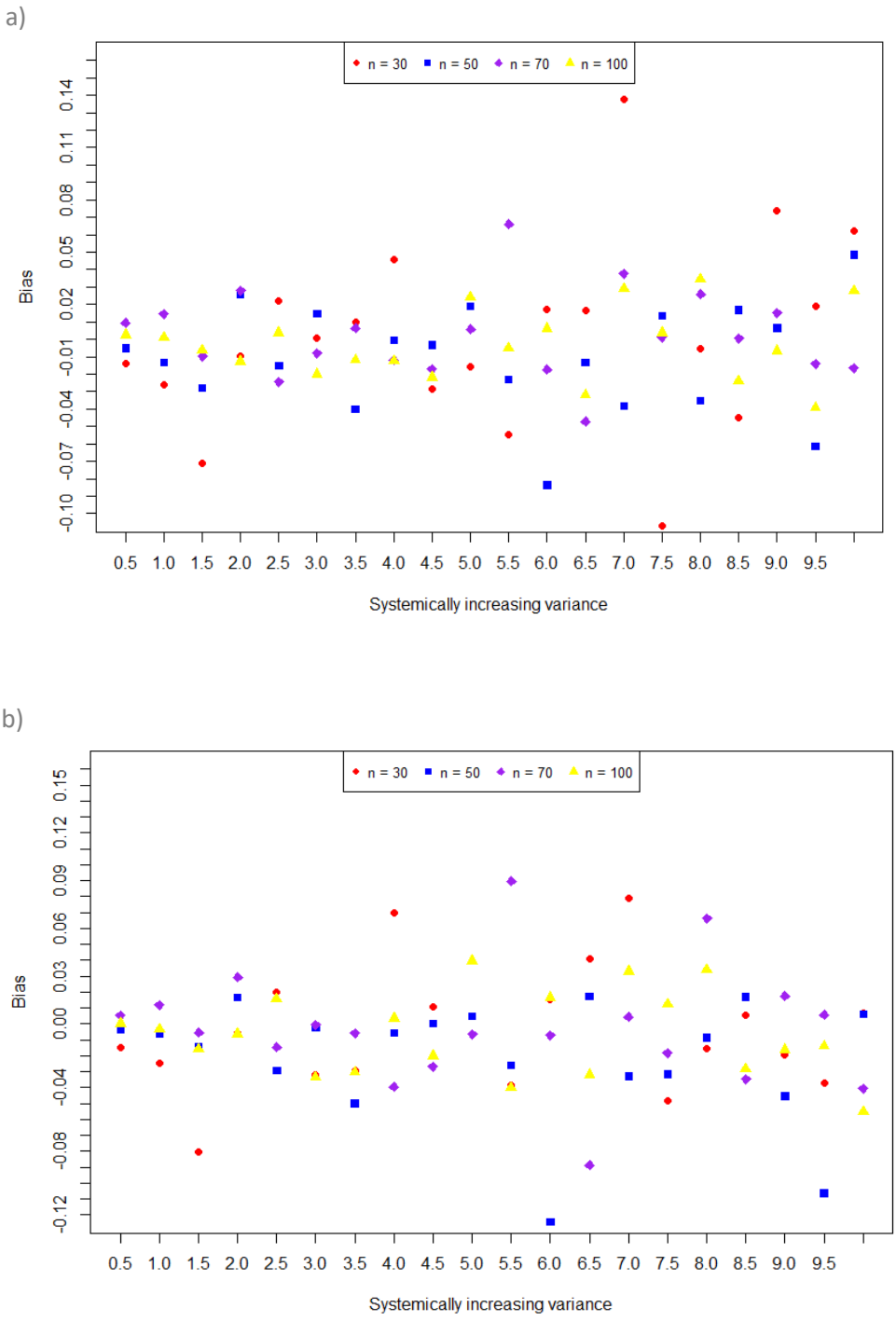

Figure 3.1: Empirical bias for change in level using weighted (a) and unweighted (b) methods where data was generated for increasing values of within patient variance and four different sample sizes.

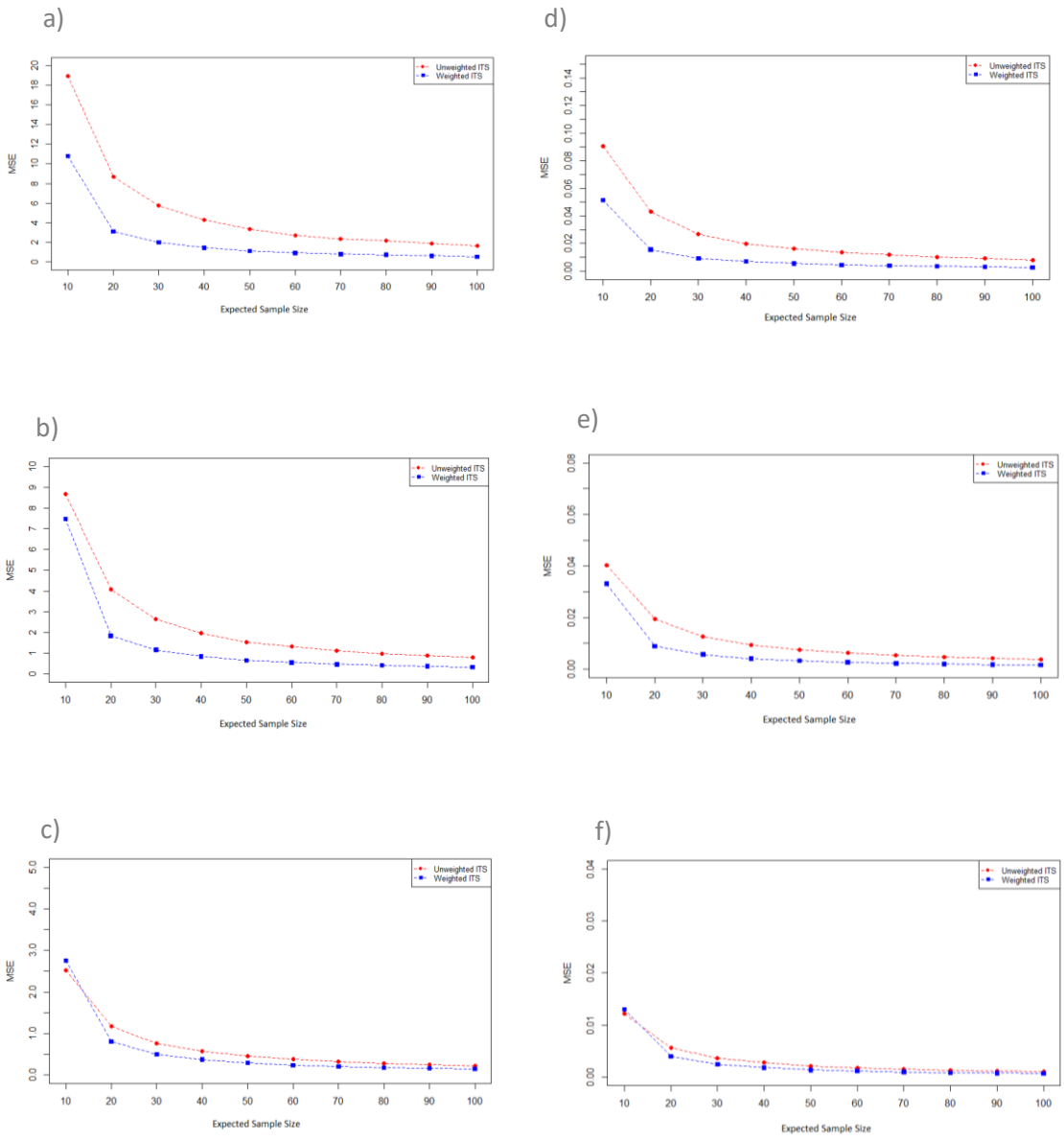

Figure 3.2: Average mean squared error (MSE) for change in level (left panel) and trend (right panel) estimates, for large (a, d), moderate (b, e) and small (c, f) differences in of variability across different sample sizes per time point.

Table 3.2: Bias for segmented linear regression and for weighted segmented regression with small, moderate and large variance heterogeneity.

	Small variance		Moderate variance		Large variance	
Sample size*	**SLR**	***w*SR**	**SLR**	***w*SR**	**SLR**	***w*SR**
Change in Level, with order of magnitude 10^{-3}						
10	0.71	1.37	1.39	1.89	0.34	3.61
30	-0.19	-0.11	-0.38	-0.08	-1.88	-0.14
50	-0.38	-0.24	-0.68	-0.27	-0.74	-0.53
70	-0.56	-0.46	-0.99	-0.67	-2.09	-0.79
100	-0.05	-0.04	-0.10	-0.04	-0.66	-0.19
Change in Trend, with order of magnitude 10^{-3}						
10	-0.01	0.06	-0.02	0.10	0.12	0.28
30	-0.01	-0.01	-0.02	-0.01	-0.02	-0.05
50	-0.02	-0.01	-0.03	-0.001	-0.06	0.04
70	-0.01	-0.01	0.02	0.01	0.06	0.01
100	-0.01	-0.01	0.01	0.01	0.02	0.02

*sample size refers to the expected sample size per time point
SLR: Segmented linear regression, wSR: Weighted Segmented Regression

In terms of precision, results from our simulation study show that the wSR produced estimates with uniformly lower mean squared error (MSE) compared to the estimates from SLR (Figure 3.2; Table 3.3). In the large between patient variance scenario, the MSE values obtained for the wSR analysis were consistently smaller than those from the SLR. Similarly, in the moderate variance scenario, the MSE values were consistently lower for the wSR method. For the small variance scenario, the MSE values were only smaller for SLR when the average sample size per time point was less than 20.

Table 3.3: Mean squared error for segmented linear regression and for weighted segmented regression with small, moderate and large variance heterogeneity.

	Small variance		Moderate variance		Large variance	
Sample size*	**SLR**	***w*SR**	**SLR**	***w*SR**	**SLR**	***w*SR**
Change in Level, with order of magnitude 10^{-2}						
10	15.12	18.93	66.69	50.08	155.93	89.58
30	4.77	3.75	20.86	10.05	48.63	18.04
50	2.69	2.06	11.81	5.54	27.58	9.95
70	2.00	1.49	8.74	3.91	20.38	6.94
100	1.37	1.01	6.02	2.69	14.07	4.81
Change in Trend, with order of magnitude 10^{-3}						
10	0.75	0.92	3.29	2.46	7.68	4.43
30	0.22	0.18	0.99	0.47	2.30	0.83
50	0.14	0.10	0.61	0.28	1.43	0.51
70	0.09	0.07	0.41	0.19	0.95	0.34
100	0.07	0.05	0.29	0.13	0.67	0.24

*sample size refers to the expected sample size per time point
SR: Segmented linear regression, wSR: Weighted Segmented Regression

3.3.2.2. Level of significance and statistical power

For the initial analysis, we maintained a small difference in between patient variance and estimated the Type I error rates for change in level and trend with different sample sizes. As seen in Figure 3.3, the average error rates of the two models for change in level (β_2) were similar for most sample sizes apart from the scenario where we considered an average sample size of 10 per time point. In this instance, we observed an average error rate of 0.1 for the wSR. As we increased the size of the sample per time point, the error rates obtained from the wSR became comparable to those obtained from the SLR.

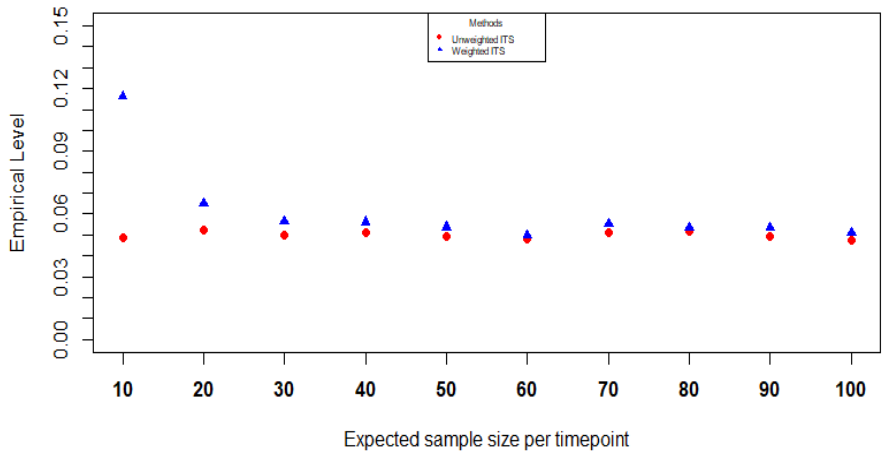

Figure 3.3: Empirical level for change in level across different sample sizes with low variance across time points.

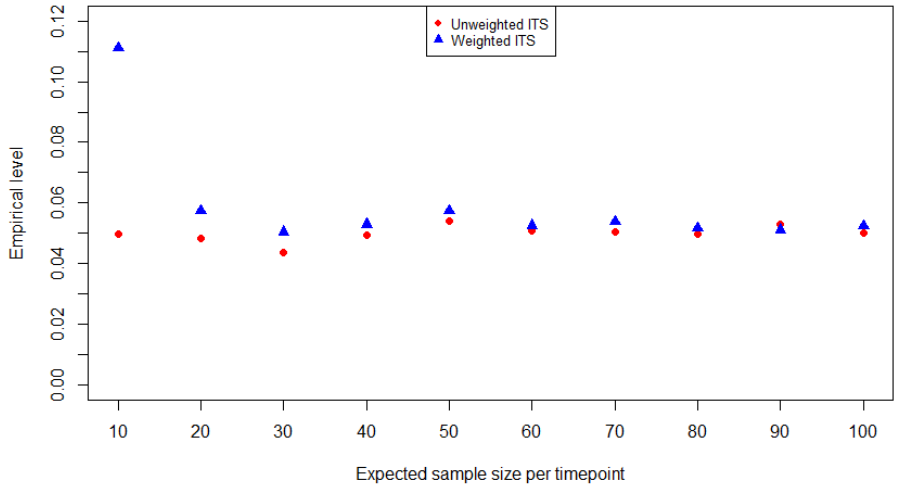

Figure 3.4: Empirical level for change in trend across different sample sizes with low variance across time points.

Similar results were observed when we assessed the Type I error rate for change in trend (β_3). (Figure 3.4).

When the differences in variance within time (between patients) was varied between moderate to large, the Type I error rates obtained for change in level (β_2) across the different sample sizes considered were slightly more conservative under the SLR

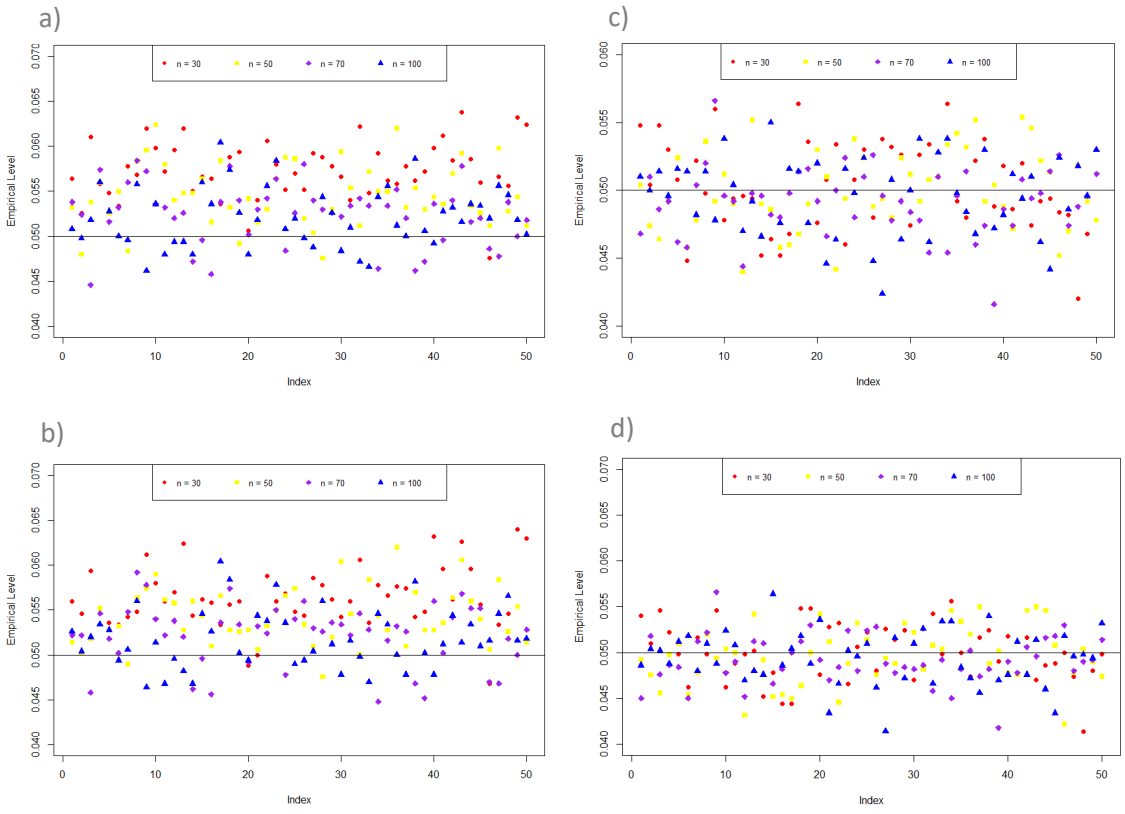

Figure 3.5: Error rate for change in level using weighted (a, b) and unweighted (c, d) methods where data was generated for large (a, c) and moderate (b, d) levels of within patient variability across different sample sizes.

method compared to the wSR. The error rates were distributed around the 0.05 threshold for the SLR (Figure 3.5) and around 0.055 for the wSR. Nonetheless, like what was obtained for the small differences in within patient variance, the error rate calculated under the wSR model remained significantly higher than 0.05 when the average sample size was less than 30 per time point. The same result pattern was observed for the change in trend (β_3) (Figure 3.6).

The power curves for change in level and change in trend across different sample sizes and different variances (small, moderate, and large) is shown in Figure 3.7 and Table 3.4. The wSR model had a consistently higher power to detect at least a 0.5 change in level across all sample sizes. For moderate to large differences in between patient variances, the weighted method had considerably higher power to detect a 0.5 change in level compared to the unweighted method (Figure 3.7). Even for small samples where both methods performed poorly in terms of power, the wSR had considerably higher power to detect a 0.5 change in level (Table 3.4). A similar performance was observed when we assessed the two methods in terms of power to detect a unit change in level (i.e. a change in level of 1) across all sample sizes. Overall, for large between patient variance, the wSR method performed considerably better in terms of power to detect varying estimates for change in level (Figure 3.8).

Similarly, for large between patient variance, the wSR model had considerably higher power to detect a 0.1 change in trend compared to the SLR method across all sample sizes (Figure 3.7).

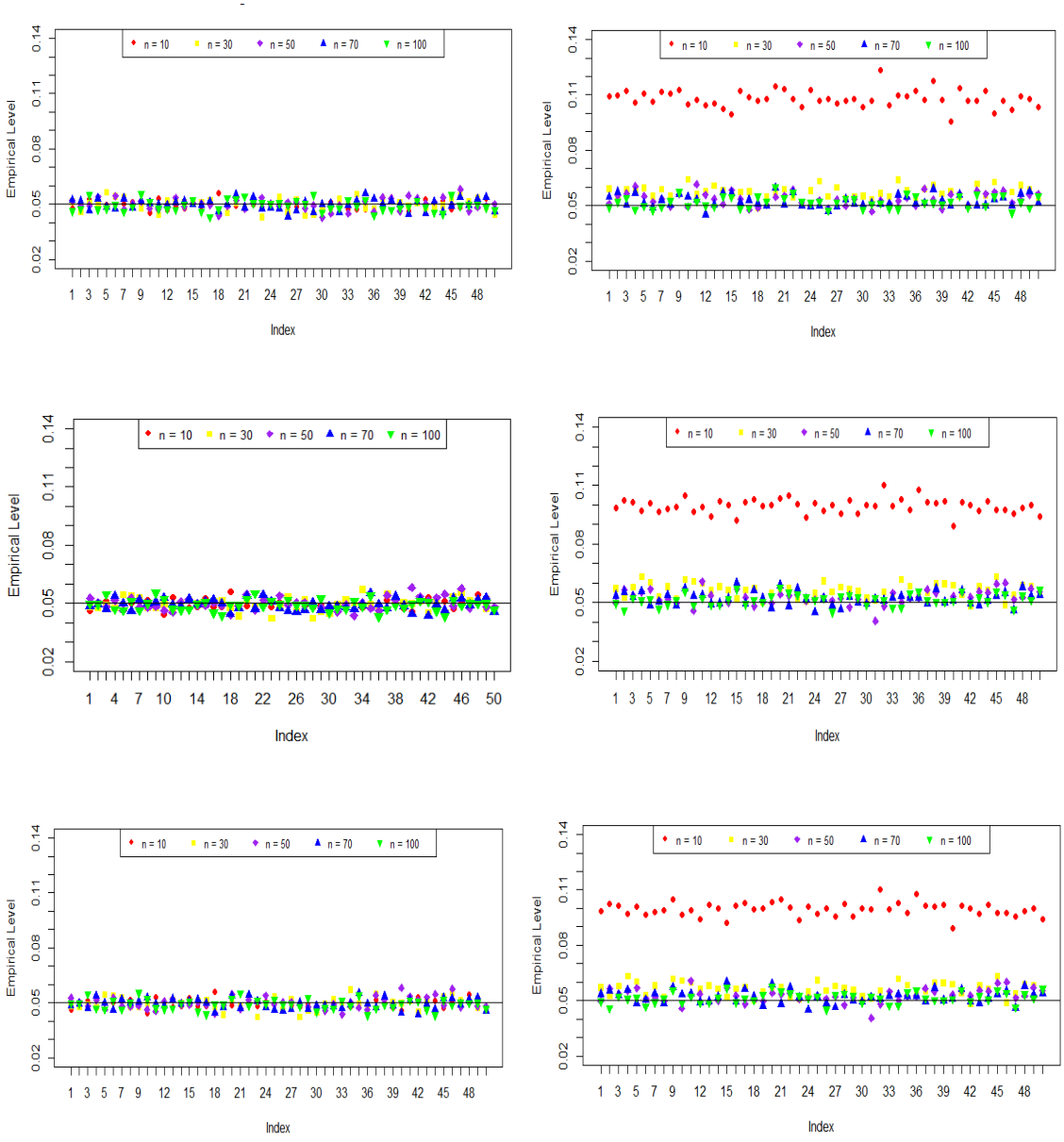

Figure 3.6: Error rate for change in trend across different sample sizes and levels of variance heterogeneity for unweighted (left panel) and weighted (right panel) methods.

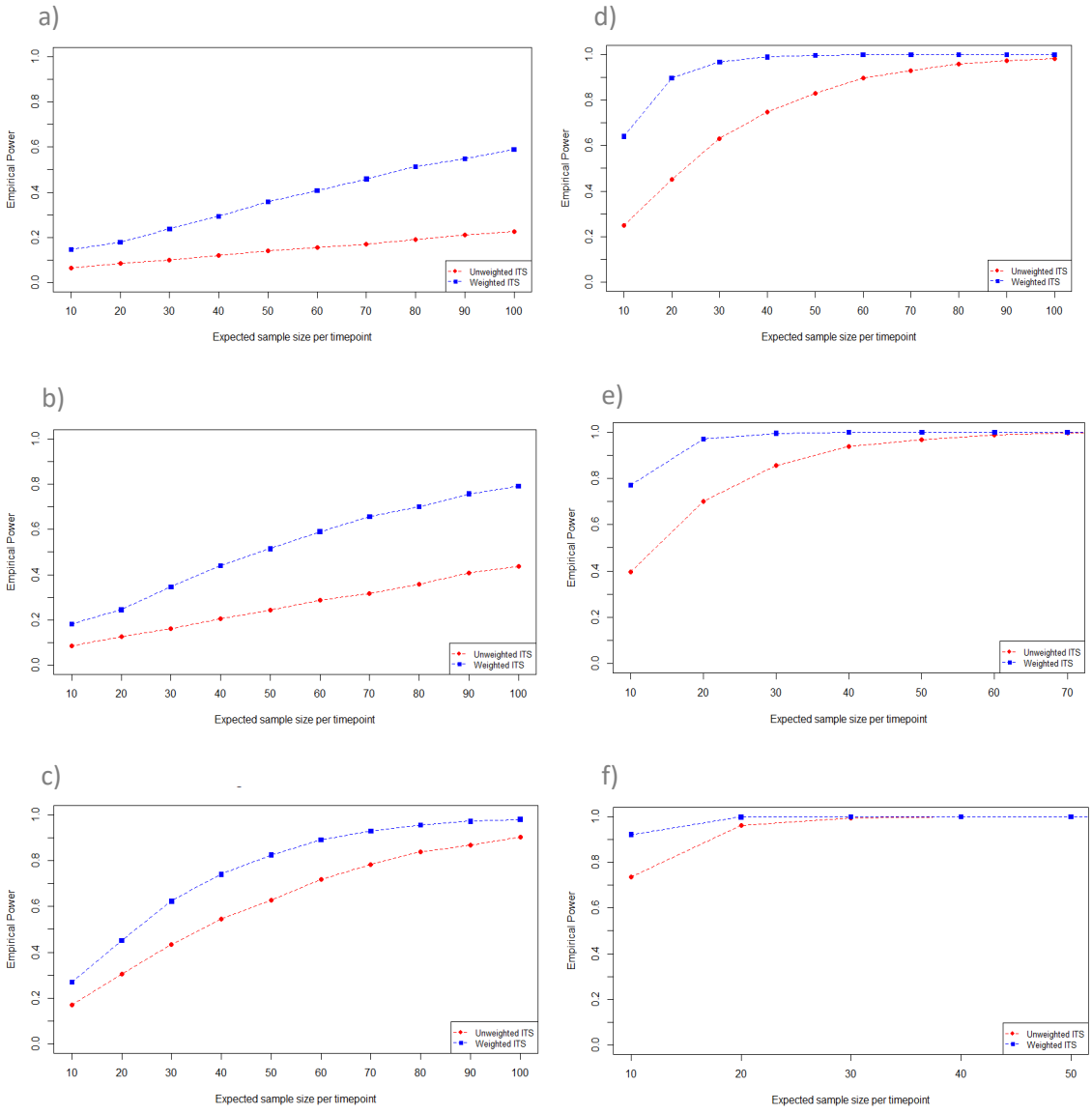

Figure 3.7: Power curve for change in level (left panel) and trend (right panel) across different sample size and large (a, d), moderate (b, e) and small (c, f) between patient variability.

The weighted model also performed much better for moderate differences in variances when the sample size was less than 30 per time point. Overall, for large sample sizes (>70 per time point) with moderate differences in variance, the results for both methods were comparable in terms of power to detect a change of 0.1 in trend (Figure 3.7).

A similar performance was observed when we wanted to detect a change in trend of 0.2. Finally, when the between patient variance was kept low across all time points, the SLR and the wSR models produced comparable results for change in trend (β_3).

Table 3.4: Estimates of power for segmented linear regression and weighted segmented regression for different sample sizes

	Change in Level					
	Low variance		Medium variance		High variance	
Sample size*	**SLR**	**wSR**	**SLR**	**wSR**	**SLR**	**wSR**
10	0.179	0.286	0.091	0.192	0.065	0.158
30	0.410	0.618	0.153	0.376	0.106	0.255
50	0.623	0.812	0.217	0.524	0.145	0.372
70	0.781	0.920	0.315	0.688	0.183	0.461
100	0.905	0.998	0.435	0.808	0.219	0.584
	Change in trend					
10	0.728	0.923	0.405	0.792	0.255	0.633
30	0.999	1.000	0.859	0.997	0.621	0.951
50	1.000	1.000	0.973	1.000	0.835	0.998
70	1.000	1.000	1.000	1.000	0.957	1.000
100	1.000	1.000	1.000	1.000	0.999	1.000

*sample size refers to the expected sample size per time point
SLR: Segmented linear regression, wSR: Weighted Segmented Regression

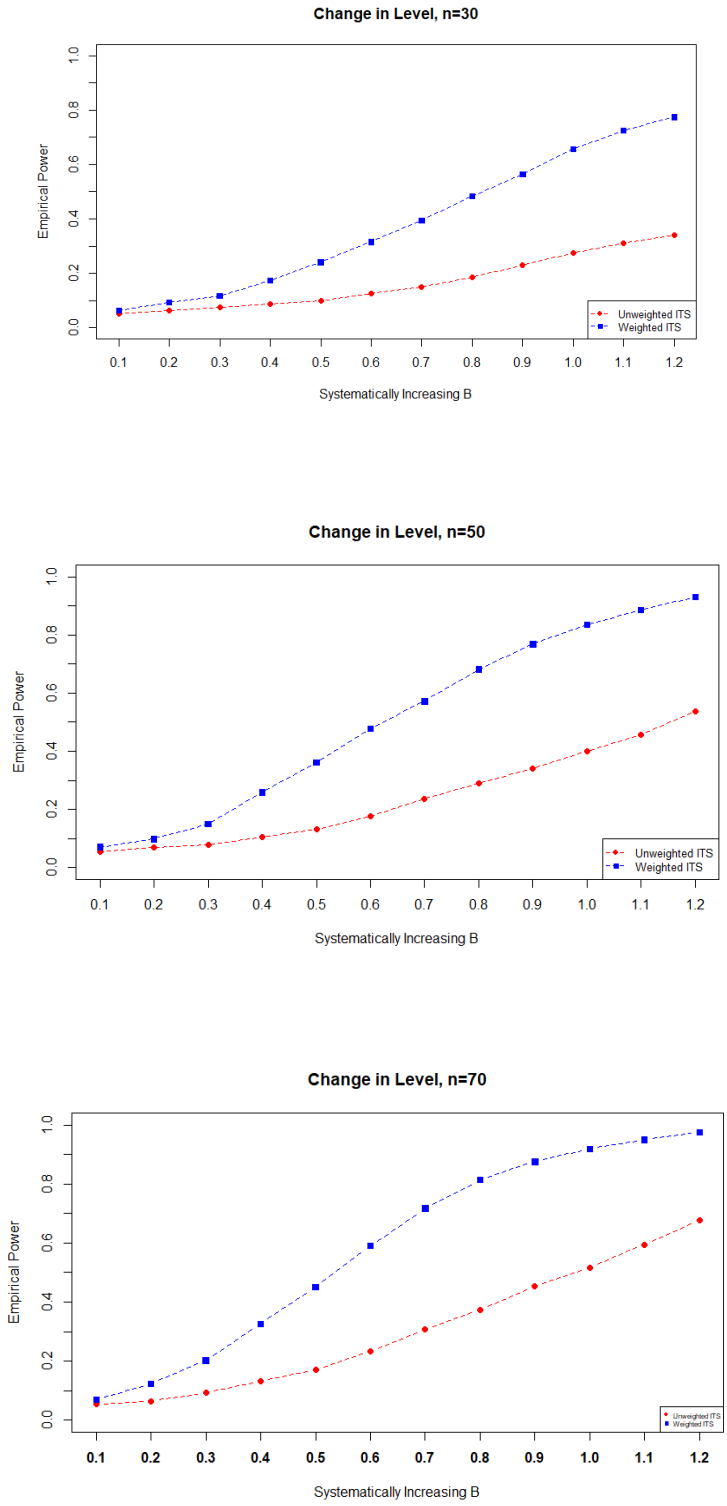

Figure 3.8: Power curve for change in level across different sample sizes when differences in error variance is large for different values of β.

3.4. Real Data Example

We illustrate application of our proposed weighted approach, using data from an ITS study involving an evidence-based intervention aimed at improving mobility of vulnerable elders admitted to Ontario hospitals (Liu et al., 2013; Liu et al., 2017). The main outcomes considered in this study were patient mobility observed weekly through audits and length of stay (LOS) obtained from hospital decision support data. For illustration purposes, we use the mobility data from two of the hospitals, with varying levels of heterogeneity and sample sizes; average sample size of 39. In each hospital, data was collected twice a week for a period of 10 weeks before intervention, 8 weeks during intervention and 20 weeks post-intervention.

The weekly mobility was first calculated as the proportion (p) of patients, who were out of bed at least once a day, as it was originally analyzed in the study (Liu et al., 2017). Thus, the total number of time points was 38. The variance of the proportion at each time point was estimated as $var\ (\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$ and weights were calculated as the inverse of these estimates. We fitted the both the traditional SR and wSR models to the data and compared the estimates obtained from both models. We also compared the mean square error (MSE) and the Alkaike information criteria (AIC) obtained from the two models to assess the performance of the methods and model fit respectively.

The estimates of the changes level (intercept) and trend (slope) were obtained by fitting both the SLR and the wSR to the data and the results for both are presented in Table 3.5. As can be seen from the results, the two methods resulted in different estimates. The p-values of the wSR estimates were generally lower than the p-values of the unweighted

model (i.e. SLR). Furthermore, the 95% confidence intervals for the wSR estimates are also narrower than that of the SLR. Results from comparative model fit analysis, with respect to mean squared error (MSE), Akaike information criteria (AIC) and p-value of model fit, are provided in Table 3.6. For the two sites, the wSR has smaller MSE and AIC compared to SLR. The results observed corresponds to the results of the simulation study where the wSR is shown to be more powerful compared to the SLR. The results also showed the wSR provides more precise estimate and is a better fit when there are moderate to high differences in within patient variability (Tables 3.5 & 3.6).

Table 3.5: Change in level and slope with corresponding 95% confidence interval (CI) and p-values of segmented linear regression (SLR) and weighted segmented regression (wSR) for mobility

	Change in Level		**Change in Trend**	
Site	Segmented Linear Regression	Weighted Segmented Regression	Segmented Linear Regression	Weighted Segmented Regression
Site 1	-14.58 (-28.85, -0.31) *p-value=0.05*	-13.59 (-25.79, -1.40) *p-value=0.03**	1.43 (-1.34, 4.21) *p-value=0.30*	1.70 (-0.81,4.22) *p-value=0.18*
Site 2	3.81 (-7.79, 15.41) *p-value=0.51*	3.69 (-5.34, 12.71) *p-value=0.41*	-0.21 (-2.47, 2.04) *p-value= 0.85*	-0.35 (-1.97, 1.26) *p-value = 0.66*

The first row within each site corresponds to the estimate, the second to the 95% CI and the third to the p-value. The asterisk, *, denotes significance at α=0.05.

Table 3.6: MSE, AIC and p-value of segmented linear regression (SLR) and weighted segmented regression (wSR) for mobility

	MSE		**AIC**		*p-value**	
Site	Segmented Linear Regression	Weighted Segmented Regression	Segmented Linear Regression	Weighted Segmented Regression	Segmented Linear Regression	Weighted Segmented Regression
Site 1	51.52	47.78	265.1	263.7	0.2419	0.1539
Site 2	33.77	11.10	229.9	215.4	0.6008	0.1328

*p-value <0.05 implies model is a good fit. MSE: Mean square error, AIC: Akaike information criteria

3.5. Discussion

Interrupted time series (ITS) designs are arguably the best approach in implementation science for evaluating hospital-wide intervention effects and post hoc effects of policies especially in healthcare settings. Segmented linear regression (SLR) is the most popular statistical method that has been used to analyze data from interrupted time series designs (Ewusie et al., 2018). Despite its known advantages, SLR has several limitations. One such limitation is when applied to aggregated data, since such data are usually associated with some level of variability, which is often ignored under SLR modeling (Ramsay et al., 2003; Taljaard et al., 2014; Harrington and Velicer, 2015).

In this study, we presented the wSR model, which takes into account the variability associated with such aggregated data. At each time point, we calculated estimate of interest (e.g. mean) and the associated variance of the estimate. This variance was then used to estimate the weights used in our proposed method. The weights were calculated as the inverse of the sum of the within and across time variance. To assess the performance of the model, we simulated data similar to real world examples and evaluated the bias, MSE, empirical power and Type I error rate of the method and compared it to the traditional SLR method.

Our simulation study provides insight into the extent to which the magnitude of differences in within time variance would yield biased effect estimates versus estimates with decreased precision. The wSR model was statistically superior to the traditional SLR model when the differences in variance within time was moderate to large. We observed that the wSR method was comparable to the SLR method when the sample size per time

point was small (<30). However, for small sample sizes, the wSR model may yield biased results. For small samples, when the differences in variance was also small, the wSR model had a higher power at the expense of the type one error rate. This means that although the wSR method has a higher likelihood of detecting an intervention effect, it is also more likely to provide a false positive finding when the sample size per time point is less than 20.

Considering the results based on scenarios where the differences in variance lies between moderate to large, the precision-bias trade-off will be in favor of the wSR. Therefore, it seems reasonable to conclude that the wSR method presented in our study, is preferable in situations where there is moderate to large differences in variance within time. Even in situations where the difference is small, we found out that the wSR method does not perform any worse than the conventional SLR method in terms of the error-power tradeoff.

Our simulation study also gives insight into the average sample size to have per time point to detect a significant effect of intervention even when the within time variance is low. We would recommend based on our results that to detect significant intervention effect, even for situations where there are low differences in the variance, we would need a sample size greater than 30 per time point.

Our study has some limitations. First, our simulations may be limited with respect to generalizability to non-normal distributions. All scenarios were considered under the normal distribution, thus for ITS data following a skewed distribution, conclusions drawn

based on our simulations may not apply. Second, we limited our simulations to single site studies, thus we did not consider multi-site studies where there are two levels of aggregation and variability introduced; that is, at the patient level and the site level. Despite the limitations, our study has several strengths. Since we used parameters based on real ITS examples, we assume that our method will be applicable in ITS designs where data are normally distributed, which is a common case in most ITS studies. Moreover, with most of us being researchers in ITS and ardent readers of ITS articles, we believe the scenarios considered here represent a considerable proportion of published ITS studies.

In summary, the wSR model is superior to the conventional SLR where there is moderate to large differences in the variance within time (between patients). The weighted SR method however does not perform well when the sample size per time point is small. Therefore, we recommend that future research should aim to develop robust ITS methods that can be used in scenarios where the sample size per time point is less than 30, since such situations sometimes occur in some ITS designs.

References

Albu, J. B., N. Sohler, L. Rui, L. Xuan, E. Young, E. W. Gregg, D. Ross-Degnan, R. Li and X. Li (2017). "An Interrupted Time Series Analysis to Determine the Effect of an Electronic Health Record-Based Intervention on Appropriate Screening for Type 2 Diabetes in Urban Primary Care Clinics in New York City." Diabetes Care **40**(8): 1058-1064.

Ansari, F., K. Gray, D. Nathwani, G. Phillips, S. Ogston, C. Ramsay and P. Davey (2003). "Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis." Journal of Antimicrobial Chemotherapy **52**(5): 842-848.

Aregawi, M., K. L. Malm, M. Wahjib, O. Kofi, N. K. Allotey, P. N. Yaw, W. Abba-Baffoe, S. Segbaya, F. Owusu-Antwi, A. T. Kharchi, R. O. Williams, M. Saalfeld, N. Workneh, E. B. Shargie, A. M. Noor and C. Bart-Plange (2017). "Effect of anti-malarial interventions on trends of malaria cases, hospital admissions and deaths, 2005-2015, Ghana." Malar J **16**(1): 177.

Balkhi, B., E. Seoane-Vazquez and R. Rodriguez-Monguio (2016). "Osteoporosis Drugs Marketed in the United States: Generic Competition, Pricing Structure, and Dispersion among Payers." International Journal of Technology Assessment in Health Care **32**(6): 385-392.

Berrevoets, M. A. H., J. H. L. W. Pot, A. E. Houterman, A. T. S. M. Dofferhoff, M. H. Nabuurs-Franssen, H. W. H. A. Fleuren, B. J. Kullberg, J. A. Schouten and T. Sprong (2017). "An electronic trigger tool to optimise intravenous to oral antibiotic switch: A controlled, interrupted time series study." Antimicrobial Resistance and Infection Control **6**(1).

Bobo, W. V., R. A. Epstein, Jr., R. M. Hayes, R. C. Shelton, T. V. Hartert, E. Mitchel, J. Horner and P. Wu. (2014). "The effect of regulatory advisories on maternal antidepressant prescribing, 1995-2007: An interrupted time series study of 228,876 pregnancies." 1. Retrieved 't Jong, G. W., Einarson, T., Koren, G., & Einarson, A. (2012) Antidepressant use in pregnancy and persistent pulmonary hypertension of the newborn (PPHN): a systematic review. Reprod Toxicol 34:293-97., 17, from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc11&NEWS=N&AN=2013-35318-001.

Bond, S. E., A. J. Chubaty, S. Adhikari, S. Miyakis, C. S. Boutlis, W. W. Yeo, M. J. Batterham, C. Dickson, B. J. McMullan, M. Mostaghim, S. Li-Yan Hui, K. R. Clezy and P. Konecny (2017). "Outcomes of multisite antimicrobial stewardship programme implementation with a shared clinical decision support system." Journal of Antimicrobial Chemotherapy **14**: 14.

Bussieres, A. E., A. E. Sales, T. Ramsay, S. M. Hilles and J. M. Grimshaw (1501). "Impact of imaging guidelines on X-ray use among American provider network chiropractors: Interrupted time series analysis." Spine Journal **14**(8): 1501-1509.

Buyle, F., D. Vogelaers, R. Peleman, G. Van Maele and H. Robays (2010). "Implementation of guidelines for sequential therapy with fluoroquinolones in a Belgian hospital." Pharm World Sci **32**(3): 404-410.

Carter, R., A. Quesnel-Vallee, C. Plante, P. Gamache and J. F. Levesque (2016). "Effect of family medicine groups on visits to the emergency department among diabetic patients in Quebec between 2000 and 2011: a population-based segmented regression analysis." BMC Fam Pract **17**: 23.

Dayer, M. J., S. Jones, B. Prendergast, L. M. Baddour, P. B. Lockhart and M. H. Thornhill (2015). "Incidence of infective endocarditis in England, 2000-13: a secular trend, interrupted time-series analysis." Lancet **385**(9974): 1219-1228.

Draper, N. R. and H. Smith (2014). Applied regression analysis, John Wiley & Sons.

Ewusie, J., C. Soobiah, E. Blondal, J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review." Revisions Submitted.

Ewusie, J., C. Soobiah, E. Blondal, J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review." BMJ Open **(Revisions Submitted)**.

Gebski, V., K. Ellingson, J. Edwards, J. Jernigan and D. Kleinbaum (2012). "Modelling interrupted time series to evaluate prevention and control of infection in healthcare." Epidemiology and Infection **140**(12): 2131-2141.

Gillings, D., D. Makuc and E. Siegel (1981). "Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care." American journal of public health **71**(1): 38-46.

Gilstein, C. Z. and E. E. Leamer (1983). "The set of weighted regression estimates." Journal of the American Statistical Association **78**(384): 942-948.

Graves, A. J., K. B. Kozhimannil, K. P. Kleinman and J. F. Wharam (2016). "The Association between High-Deductible Health Plan Transition and Contraception and Birth Rates." Health Serv Res **51**(1): 187-204.

Gutacker, N., K. Bloor, R. Cookson, C. P. Gale, A. Maynard, D. Pagano, J. Pomar, E. Bernal-Delgado and J. Pomar (2017). "Hospital Surgical Volumes and Mortality after Coronary Artery Bypass Grafting: Using International Comparisons to Determine a Safe Threshold." Health Services Research **52**(2): 863-878.

Hanson, C. C., G. D. Randolph, J. A. Erickson, C. M. Mayer, J. T. Bruckel, B. D. Harris and T. S. Willis (1015). "A reduction in cardiac arrests and duration of clinical instability after implementation of a paediatric rapid response system." Postgraduate Medical Journal **86**(1015): 314-318.

Harrington, M. and W. F. Velicer (2015). "Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies." Multivariate Behav Res **50**(2): 162-183.

Jandoc, R., A. M. Burden, M. Mamdani, L. E. Lévesque and S. M. Cadarette (2015). "Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations." Journal of clinical epidemiology **68**(8): 950-956.

Jiang, M., D. R. Hughes and R. Duszak (2015). "Screening mammography rates in the Medicare population before and after the 2009 US Preventive Services Task Force guideline change: an interrupted time series analysis." Women's Health Issues **25**(3): 239-245.

Judge, A., G. Wallace, D. Prieto-Alhambra, N. K. Arden and C. J. Edwards (2015). "Can the publication of guidelines change the management of early rheumatoid arthritis? An interrupted time series analysis from the United Kingdom." Rheumatology (Oxford) **54**(12): 2244-2248.

Kastner, M., A. M. Sawka, J. Hamid, M. Chen, K. Thorpe, M. Chignell, J. Ewusie, C. Marquez, D. Newton and S. E. Straus (2014). "A knowledge translation tool improved osteoporosis disease management in primary care: an interrupted time series analysis." Implement Sci **9**: 109.

Langford, B. J., J. Seah, A. Chan, M. Downing, J. Johnstone and L. M. Matukas (2016). "Antimicrobial Stewardship in the Microbiology Lab: Impact of Selective Susceptibility Reporting on Ciprofloxacin Utilization and Gram-Negative Susceptibility in a Hospital Setting." J Clin Microbiol.

Liu, B., U. Almaawiy, J. E. Moore, W.-H. Chan and S. E. Straus (2013). "Evaluation of a multisite educational intervention to improve mobilization of older patients in hospital: protocol for mobilization of vulnerable elders in Ontario (MOVE ON)." Implementation Science **8**(1): 76.

Liu, B., J. E. Moore, U. Almaawiy, W.-H. Chan, S. Khan, J. Ewusie, J. S. Hamid, S. E. Straus and M. O. Collaboration (2017). "Outcomes of Mobilisation of Vulnerable Elders in Ontario (MOVE ON): a multisite interrupted time series evaluation of an implementation intervention to increase patient mobilisation." Age and ageing **47**(1): 112-119.

Mackie, A. S., W. Liu, A. Savu, A. J. Marelli and P. Kaul (2016). "Infective endocarditis hospitalizations before and after the 2007 American Heart Association prophylaxis guidelines." Canadian Journal of Cardiology **32**(8): 942-948.

Michielutte, R., B. Shelton, E. D. Paskett, C. M. Tatum and R. Velez (2000). "Use of an interrupted time-series design to evaluate a cancer screening program." Health Educ Res **15**(5): 615-623.

Pow, J. L., A. A. Baumeister, M. F. Hawkins, A. S. Cohen and J. C. Garand. (2015). "Deinstitutionalization of American public hospitals for the mentally ill before and after the introduction of antipsychotic medications." 3. Retrieved American Medico-Psychological Association, Committee on Statistics; National Committee for Mental Hygiene, Bureau of Statistics. (1918). The statistical manual for the use of institutions for the insane. New York: National Committee for Mental Hygiene, 1918., 23, from

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc12&NEWS=N&AN=2015-20600-002.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies." Int J Technol Assess Health Care **19**(4): 613-623.

Rosich Marti, I., A. Allepuz, G. Rodriguez Palomar, F. Ortin Font and M. Soler Cera (2017). "Impact of an intervention on the prescription of aliskiren after new evidence on safety reported." Pharmacoepidemiology & Drug Safety **26**(1): 91-96.

Sen, A. and M. Srivastava (2012). Regression analysis: theory, methods, and applications, Springer Science & Business Media.

Shardell, M., A. D. Harris, S. S. El-Kamary, J. P. Furuno, R. R. Miller and E. N. Perencevich (2007). "Statistical analysis and application of quasi experiments to antimicrobial resistance intervention studies." Clin Infect Dis **45**(7): 901-907.

Taljaard, M., J. E. McKenzie, C. R. Ramsay and J. M. Grimshaw (2014). "The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care." Implement Sci **9**: 77.

Taylor, J. E., S. J. McDonald, A. Earnest, J. Buttery, B. Fusinato, S. Hovenden, A. Wallace and K. Tan (2017). "A quality improvement initiative to reduce central line infection in neonates using checklists." European Journal of Pediatrics **176**(5): 639-646.

Tsai, C.-L. and X. Wu (1990). "Diagnostics in transformation and weighted regression." Technometrics **32**(3): 315-322.

Wagenaar, B. H., K. Sherr, Q. Fernandes and A. C. Wagenaar (2015). "Using routine health information systems for well-designed health evaluations in low-and middle-income countries." Health policy and planning **31**(1): 129-135.

Walley, A. Y., Z. Xuan, H. H. Hackman, E. Quinn, M. Doe-Simkins, A. Sorensen-Alawad, S. Ruiz and A. Ozonoff (2013). "Opioid overdose rates and implementation of overdose education and nasal naloxone distribution in Massachusetts: interrupted time series analysis." Bmj **346**: f174.

Yang, C., Q. Shen, W. Cai, W. Zhu, Z. Li, L. Wu and Y. Fang (2017). "Impact of the zero-markup drug policy on hospitalisation expenditure in western rural China: an interrupted time series analysis." Tropical Medicine & International Health **22**(2): 180-186.

Zhang, F., A. K. Wagner, S. B. Soumerai and D. Ross-Degnan (2009). "Methods for estimating confidence intervals in interrupted time series analyses of health interventions." J Clin Epidemiol **62**(2): 143-148.

Chapter 4

Extension of the Weighted Segmented Regression to Account for Variability in Healthcare Settings

Summary

In the previous chapter, we presented a novel weighted segmented regression (wSR) method, which incorporates the variability introduced when data are aggregated across participants per time point in ITS studies. For multisite or multi-center studies, another level of variability is introduced when the data are aggregated across the various sites involved in the overall analysis. This chapter includes an article submitted to "Statistical Methods in Medical Research" journal, where we extended the wSR to multisite studies. The extended method allows us to incorporate two levels of variability: the between participant (within site) variability, and the site to site (between site) variability. Empirical data analysis was completed using data from a multisite ITS data. We compared the precision and accuracy of estimates produced by our method with those from the traditional SLR method as well as those from a previously developed pooled analysis method.

Citation: Ewusie JE, Beyene J, Thabane L, Straus S, Hamid JS. "Multi-center Interrupted Time Series Analysis: Incorporating Within and Between Center Heterogeneity", Statistical Methods in Medical Research, under Review.

Multi-center Interrupted Time Series Analysis: Incorporating Within and Between Center Heterogeneity

Joycelyne Ewusie[1], Lehana Thabane[1], Joseph Beyene[1], Sharon E Straus[2], Jemila S. Hamid[1,3*]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
[2]Li Ka Shing Knowledge Institute of St Michael's Hospital, Toronto, Ontario, Canada
[3] Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada

Abstract

Background: Segmented linear regression (SLR) is the most common statistical method used in the analysis of interrupted time series (ITS) data. The model allows one to examine immediate effects of an intervention as well as the effects over time. However, this modeling strategy is indicated to produce spurious results when applied to aggregated data. For multisite ITS studies, data at a given time point is often aggregated across participants and across different sites, thus conventional SLR analysis may not be an optimal approach. Our objective is to provide a robust method for analysis of ITS data, while accounting of two sources of heterogeneity – between participants and across sites. **Methods:** We provide a methodological framework within the segmented regression (SR) modeling strategy, where we introduced weights to account for between participant variation as well as the differences in settings across multiple sites. We empirically compared the weighted segmented regression (wSR) method proposed in this study with the conventional SLR method as well as with a previously published pooled analysis method. Data from the Mobility of Vulnerable Elders in Ontario (MOVE-ON) project, which was a multisite ITS study, was used for the empirical evaluation. **Results:** The wSR method produced the most precise estimates, with the narrowest 95% confidence interval (CI), for the overall intercepts and slopes, while the SLR method resulted in the least precise estimates. Our method also led to smaller p-values, indicating increased power. The pooled analysis method produced results similar to the wSR, when the number of sites included in the overall analysis was ≤ 4 sites, and when there was moderate to high between-site variability as measured by the I^2 statistic. **Conclusions:** Incorporating the participant-level and site-level variability led to estimates that were more precise and accurate in determining the magnitude of the effect of an intervention and led to increased statistical power. This result underscores the importance of accounting for the inherent variability in aggregated data. Extensive simulations are required to further evaluate the methodological framework presented in this paper and compare the methods in a wide-range of scenarios and outcome types.

Keywords: Aggregated data, Weighted segmented regression, Pooled analysis, Interrupted time series, Multisite studies.

4.1. Background

Interrupted time series (ITS) design involves repeated measurements of an outcome at several time points before and after implementation of interventions or programs and can be designed to investigate impact of interventions and programs in healthcare settings. Segmented linear regression (SLR) is the most common statistical method used in the analysis of ITS data (Wagner et al., 2002; Zhang et al., 2009; Penfold and Fang, 2013; Ewusie et al., 2018). A recent scoping review, aimed at identifying methods for analyzing ITS data showed that SLR models were the most utilized methods, having been used in approximately 26% of the included studies, followed by autoregressive integrated moving average (ARIMA), which was used in approximately 20% of the studies to analyze their ITS data (Ewusie et al., 2018). These methods have been applied in various health research areas (e.g. clinical research, public health, and health services) to assess the impact of interventions (on patient important outcomes or clinical practice) such as quality improvement initiatives, drug subsidization policies, educational programs as well as dissemination and implementation of clinical practice guidelines (Ansari et al., 2003; Smith et al., 2006; Fowler et al., 2007; Hartung et al., 2008; Katikireddi et al., 2016; Naimer et al., 2017; Bedard et al., 2018).

Segmented linear regression (SLR), also known as piecewise regression, is a special case of multiple linear regression with an indicator variable representing the intervention periods, a continuous variable representing time at which observations are taken and an interaction variable (Kong et al., 2012). For ITS analysis employing SLR, a linear regression line is fitted to each segment of the time series, for instance, the pre- and post-

intervention periods (Wang et al., 2013). This approach assesses the magnitude of the effect of intervention on the outcome of interest. Using this approach, both the immediate effect of the intervention as well as the effect of intervention over time can be evaluated. SLR analysis controls for potential confounders (such as age and gender), for example, by including covariates in the model to account for the confounders (Bernal et al., 2017). The model also enables us to evaluate if other plausible or rival explanations are available for the observed change in the outcome after the onset of intervention.

Compared to other methods employed in ITS analysis, SLR models have other advantages. The model can be used to examine underlying secular trends in the data. For example, the length of stay (LOS) of patients in a hospital may be decreasing over time prior to a quality improvement (QI) initiative to improve patient care. If this trend in LOS prior to intervention is not accounted for, any observed decrease could be attributed to the QI initiative. The SLR model can also be used to investigate seasonal or cyclical patterns that occur over time, which may confound with the intervention effect. For instance, flu rate may be higher in the winter season hence a study examining the effect of a health policy on hospital admission may lead to biased results if the preintervention period contains several winter months. Further, SLR models can be used to assess the presence of serial correlation, also known as autocorrelation, in the ITS data and account for any observed autocorrelation. This is usually achieved by including autoregressive parameters in the SLR model. Segmented linear regression with autoregressive error models are usually preferred to autoregressive integrated moving average (ARIMA) models because they require fewer time points to adequately estimate parameter values (Biglan et al., 2000;

Zhang et al., 2009). Finally, SLR approaches can be used to investigate intervention effects that may be temporal by disintegrating the time points to identify the exact point or period the intervention was most effective. For instance, an intervention that is effective only in the first 3 months can be identified by dividing yearly data into monthly data. Decomposing time points in this way, however, reduces the sample size per time point; hence there is a trade-off between having large number of time points versus large sample size per time (Biglan et al., 2000; Ramsay et al., 2003).

Despite these advantages, SLR, like most traditional regression approaches, has several limitations. One major limitation is when it is applied to aggregated data (Gillings et al., 1981; Wagner et al., 2002; Taljaard et al., 2014). Aggregated data refers to summary statistics (e.g. mean, percentage, median) calculated across study participants. In health research involving ITS, aggregated data are very common since interventions are often implemented to evaluate the effect at healthcare facility level and data at a given time point are often summarized across participants within the healthcare facility. For multisite ITS studies, further aggregation occurs, where the data are summarized across several sites (e.g. hospital units) to evaluate overall impact of interventions. Additionally, administrative routine data collected from different subjects across different regions are increasingly being used in ITS studies to perform post-hoc evaluation of nationwide policies and programs (Wagenaar et al., 2015; Ewusie et al., 2018). Such data are usually only available in aggregated forms.

Aggregated data are a major cause of heteroskedasticity, which is a violation of the assumption of constant error variances required during inference involving regression

models (Box et al., 2015). When data are aggregated, variability is introduced due to factors such as the differences in participant characteristics, sample size and settings. Hence, the variance across time, and consequently the error variance is no longer constant. To account for the variability introduced by aggregating data, Gebski et al. (2012) proposed the pooled time series analysis method, where the intercepts and slopes are pooled from individual site analysis, using meta-analytic approaches, to calculate the overall effect of the intervention. Their method may not be optimal for all types of aggregated data, since the variability across participants within the same site was not accounted for. Further, the use of summary estimates (intercepts and slopes) instead of using all the available information from the data (individual participant data) may lead to loss of power, which will in turn lead to suboptimal results (Garrett, 2003; Lyman and Kuderer, 2005).

In this study, we proposed the extended weighted SR (wSR)method, with the aim of incorporating both participant level and site level variability. The precision, accuracy and statistical significance of our method was compared to the SLR method as well as the pooled analysis (PA) method proposed by Gebski and colleagues using empirical data.

4.2. Methods

4.2.1. Empirical Data

We use data from the Mobility of Vulnerable Elders in Ontario (MOVE-ON) project for our empirical evaluation (Liu et al., 2013; Liu et al., 2017). The MOVE-ON study involved implementation and evaluation of the impact of an evidence-based intervention, which aimed to promote early mobilization and prevent functional decline among older patients

admitted to acute care academic hospitals across Ontario. An interrupted time series design was used to evaluate the impact of the intervention. The study was conducted across 14 hospitals consisting of 30 units that provided care to inpatients including those aged 65 years and older. Data were collected over a period of 38 weeks with a 10-week pre-implementation period followed by an 8-week implementation period, where the intervention was rolled out, and a 20-week post-implementation period. The primary outcome of the study was the mobilization status of patients, who were assessed on twice-weekly visual audits, which occurred three times per day (in the morning, at lunch and in the afternoon). The secondary outcomes included hospital length of stay (LOS), functional status at admission and discharge, number of falls and discharge destination. Data on patient characteristics such as age, gender and place of residence prior to admission were also collected. Further details on the study can be found elsewhere (Liu et al., 2013; Liu et al., 2017). In this paper, we considered the primary outcome and presented the proposed method using this data as an illustration. We also compared our method with previous methods empirically.

4.2.2. Weighted ITS Analysis

Consider the primary outcome, patient mobility, in the MOVE-ON project. Within each site, and at each time point, patient mobility was summarized as the percentage of patients who were observed (during audits) out of bed at least once in a day. That is, if we let x_{ij} represent the mobility status for patient j at time point i, then the percentage of patients who were observed out of bed at a time point i is calculated as;

$$y_i = \frac{\sum_j I_{(x_{ij} \geq 3)}}{N}, \quad (4.1)$$

where $I_{(x_{ij} \geq 3)}$ is an indicator function dichotomizing the primary outcome into two categories; whether a patient is out of bed or not out of bed. N represents the total number of patients observed at each time point. To examine the impact of the intervention on patient mobility, a SLR of the form described in equation (4.2) below was fitted and comparisons in level and trend of mobility were made among the pre-, during, and post-implementation periods (Wagner et al., 2002).

$$y_i = \beta_0 + \beta_1 * t + \beta_2 * Int1 + \beta_3 * t * Int1 + \delta_1 * Int2 + \delta_1 * t * Int2 + \varepsilon_i , \quad (4.2)$$

where t represents the time at which the outcome is measured, $Int1$ and $Int2$ are dichotomous variables representing the intervention periods. Interactions between time and intervention periods are also included in the model. The parameter β_0, estimates the baseline percentage of patients who were out of bed; β_1 represents the slope (trend) of mobilization prior to the intervention; β_2 and β_3 represent the changes in intercepts and slopes between the pre and during intervention periods; while δ_1 and δ_2 represent the changes in intercepts and slopes between the pre and post intervention. The model can be presented in matrix format as;

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad (4.3)$$

where Y denotes a vector representing the outcomes $(y_1, y_2, y_3, \ldots y_k)$, measured at times 1, 2, …, k; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_p)$ denotes a vector representing the regression coefficients (parameters) associated with time, the intervention periods and interactions (as listed in

equation (4.2)) as well as other potential covariates that might influence the outcome; $\boldsymbol{X}$ is a p x k matrix of independent variables and $\boldsymbol{\varepsilon}$ is the error vector.

Recall that the data at each time point is aggregated across patients, which also involves summary statistics based on different sample sizes, thus introducing one level of variability. Further, being a multisite study, data are aggregated not only at the site level but also across sites. Hence, there are two levels of heterogeneity introduced; within site (that is across participants within hospitals and units) and across sites (across hospitals). Figure 4.1 provides a depiction of the two levels of heterogeneity in the MOVE-ON data sets, where the overall percentage mobility at each time point (represented using red triangular shape points) is averaged across the 14 participating Ontario hospitals. Each overall estimate is associated with heterogeneity introduced by variations across the sites; the different magnitudes of variability is shown using elliptical shapes. Moreover, the site level estimates themselves (represented using circular points) are summarized across different patients within the respective hospitals; and summarized using different sample sizes ranging from 30 to 146; thus, associated with various levels of imprecision. For instance, the 4 site level summaries circled in red, are associated with percentage mobility estimates based on sample sizes of 32, 44, 104 and 75 respectively.

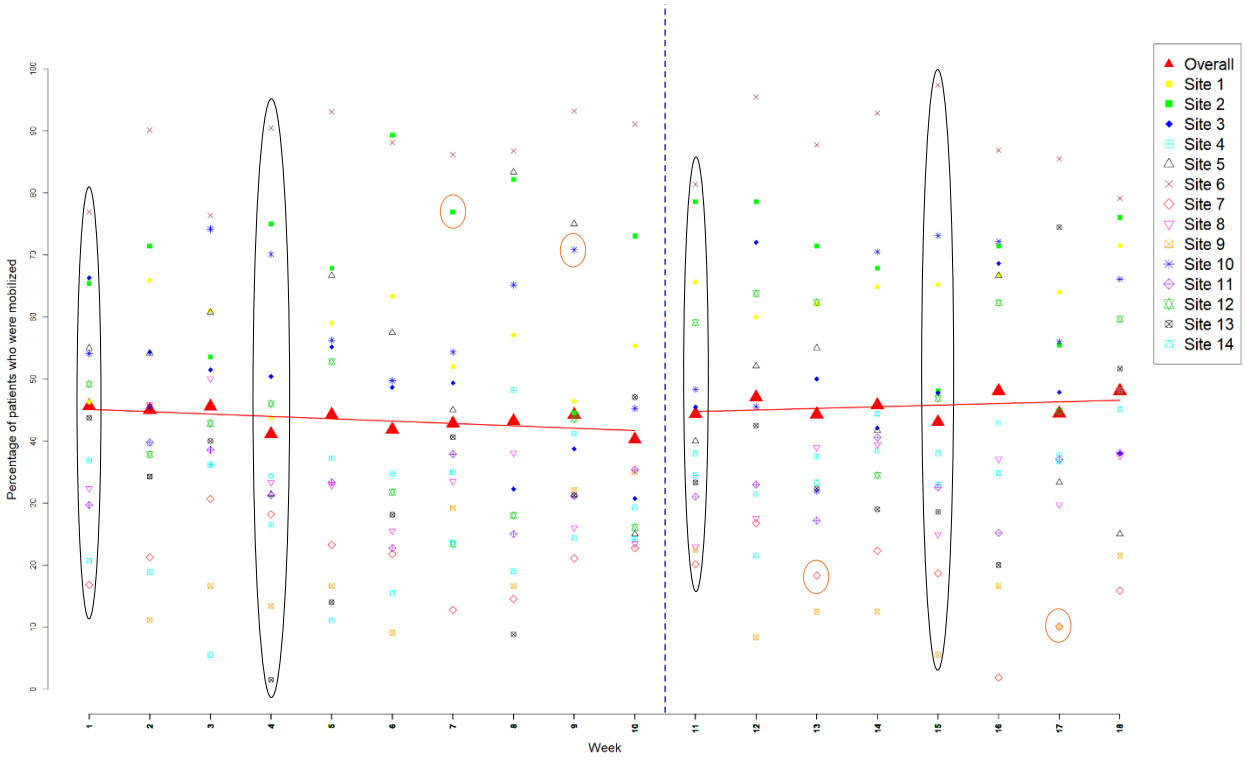

Figure 4.1: Weekly percentage of patients who were mobile at least once a day for all sites and overall.

Due to the unaccounted heterogeneity presented above, the SLR approach will be associated with increased biased and decreased precision when used on aggregated data. It has also been previously established that the violation of the homoscedasticity assumption (due to the difference in levels of variability) leads to inflated type one error and decreased power (Wagner et al., 2002; Sen and Srivastava, 2012; Ewusie et al., 2018). To overcome this limitation, we propose the extended wSR approach, where the weighted least squares method is used to estimate the parameters of the model in equation (4.2). The estimators are given by;

$$\widehat{\boldsymbol{\beta}}_w = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}, \tag{4.4}$$

where Y and X are as defined in (4.3) and $\boldsymbol{W}$ represents a diagonal matrix which consists of the vector, $\boldsymbol{w} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_k})$ as its diagonal elements. The weights are calculated as the inverse of the sum of the within site and between site variances. That is, if ${\gamma_i}^2$ represents the variance within a site (across patients) at time point i; ${\tau_i}^2$ represents the variance between sites at time point i; and σ^2 represents the variance associated with the random error in model (4.4) then the weights are given as;

$$w_i = \frac{1}{\sigma^2 + {\gamma_i}^2 + {\tau_i}^2}. \tag{4.5}$$

The variance of the weighted regression estimates is calculated as;

$$Var\left(\hat{\beta}_w\right) = \sigma^2 (X'WX)^{-1}. \tag{4.6}$$

The standard errors and 95% confidence intervals (CI) can therefore be constructed accordingly, using established formulas found in literature (Sen and Srivastava, 2012).

4.2.3. Empirical Evaluation

Our empirical evaluation consists of comparative analysis involving our extended wSR approach, the SLR approach as well as the pooled analysis (PA) proposed by Gebski et al., 2012. To examine the performance of the proposed method to changes in number of sites and various levels of within and between site heterogeneity, we performed several analyses for different scenarios, where we varied the number of sites included, the amount of within site variability and the magnitude of between site heterogeneity (low, moderate and high) as measured by the I^2 statistic used in meta-analysis literature to quantify the amount of between study heterogeneity (Higgins and Thompson, 2002; Higgins et al., 2003). The

heterogeneity is defined as low when the $I^2 \sim 25\%$, moderate when $I^2 \sim 50\%$, and high when $I^2 \sim 75\%$. For all scenarios considered, the results (estimates of slopes and intercepts) and the associated 95% confidence intervals (95% CIs) as well as the p-values obtained from the three methods were compared to assess precision and accuracy of the estimates respectively. For time series data, the Durbin-Watson statistic is the most common measure used to evaluate the presence of autocorrelation in the data (Sen and Srivastava, 2012). However, the data used in this study is from a previous project where the authors reported there was no significant autocorrelation in the data. Hence, we did not perform any further assessment of autocorrelation. All analysis was performed using the R version 3.4.3 statistical software.

4.3. Results

During the 38-week study, a total of 125,025 observations were made from a total of 14,540 patients across the 14 sites, of which 3,943 patients were audited pre-intervention; 3,216 during intervention and; 7,381 patients post intervention. The overall mean age of patients was 80.0 ± 8.36 years; the average age was similar across intervention periods and 54% (N=7,805) of the patients were female and private home/apartment was the most common place of residence prior to admission, 33.8% (N=4917).

For analysis involving overall patient mobility across the 14 hospitals, the estimates of the regression coefficients obtained using each of the three methods are presented in Table 4.1. Here, we present the comparisons between pre-intervention, during intervention and post intervention. The wSR estimates have the narrowest confidence interval for all estimates,

Table 4.1: Estimates for intervention effects, 95% CI and p values obtained using segmented linear regression (SLR), pooled analysis (PA) and weighted segmented regression (*w*SR)

Variable	**SLR* Estimate** (95% CI)	**Pooled Estimate** (95% CI)	***w*SR** Estimate** (95% CI)
β_0 (Baseline level)	45.47	44.16	45.49
	(37.78, 53.17)	(32.54, 55.77)	(41.21, 49.77)
	p value < 0.0001	*p value < 0.0001*	*p value < 0.0001*
β_1 (Week)	-0.38	-0.21	-0.38
	(-1.61, 0.86)	(-1.02, 0.58)	(-1.07, 0.31)
	p value = 0.55	*p value = 0.60*	*p value = 0.27*
β_2 (Difference in level pre to during intervention)	2.80	2.51	2.79
	(-8.07, 13.66)	(-4.18, 9.20)	(-3.33, 8.91)
	p value = 0.61	*p value = 0.46*	*p value = 0.36*
β_3 (Difference in trend pre to during intervention)	0.64	0.39	0.64
	(-1.47, 2.76)	(-0.91, 1.69)	(-0.55, 1.83)
	p value = 0.55	*p value = 0.56*	*p value = 0.28*
δ_1 (Difference in level pre to post intervention)	13.03	9.45	13.33
	(-3.64, 29.70)	(-0.79, 19.68)	(4.02, 22.66)
	p value = 0.13	*p val1ue = 0.07*	*p value = 0.01*
δ_2 (Difference in trend pre to post intervention)	0.09	0.09	0.05
	(-1.22, 1.41)	(-0.72, 0.90)	(-0.68, 0.79)
	p value = 0.89	*p value = 0.82*	*p value = 0.88*

*SLR: Segmented linear regression; ***w*SR: Weighted segmented regression
p-value <0.05 implies estimate is significant

while the SLR has the widest confidence interval for most estimates. The wSR approach produced smaller p-values in most cases as well, indicating increased statistical power.

Nonetheless, the parameter estimates are relatively comparable for the SLR and wSR methods in most cases.

All the models produced a negative coefficient for the Week (β_1) variable, which implies that there was a decline in patient mobilization prior to the implementation of the educational intervention. The positive coefficient of the intervention (β_2) variable implies that there was an immediate increase in percentage of patient mobilization during the implementation of the intervention (Figure 4.2, Table 4.1). Similarly, there was a substantial increase in the weekly percentage of patients who were mobile during the

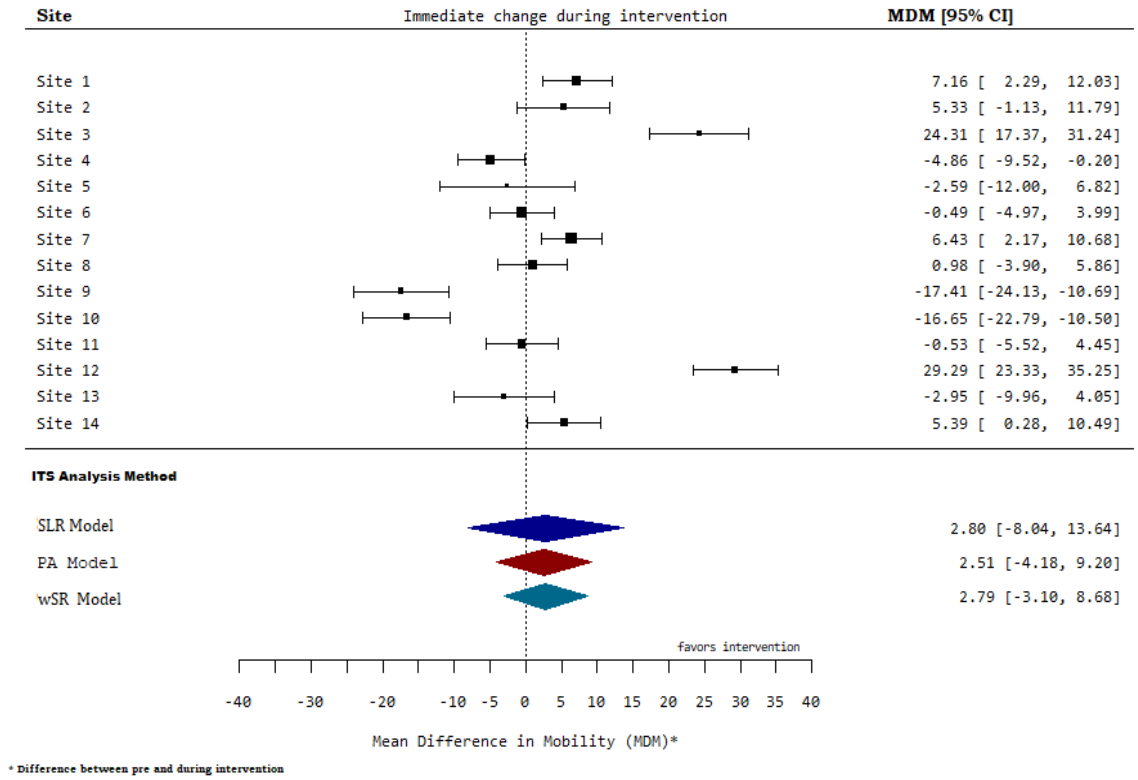

Figure 4.2: Forest plot for change in level of mobilization during the implementation of the intervention.

intervention period, β_3= 0.64 (95% CI: -0.55, 1.83; *p-value = 0.28*); from the wSR estimate. We again observed that although all three models showed an increase in trend of patient mobilization during the intervention period, this increase was not statistically significant (Figure 4.3, Table 4.1).

Considering the change in level from the pre-intervention phase to post-intervention, there was a significant increase in mobility, where 13% more patients who were mobile in the post-intervention period compared to pre-intervention (Figure 4.4, Table 4.1). This increase was shown to be statistically significant using our proposed wSR method (p-value = 0.01). However, this difference was not statistically significant using both the SLR and PA methods. Once more indicating that our method has

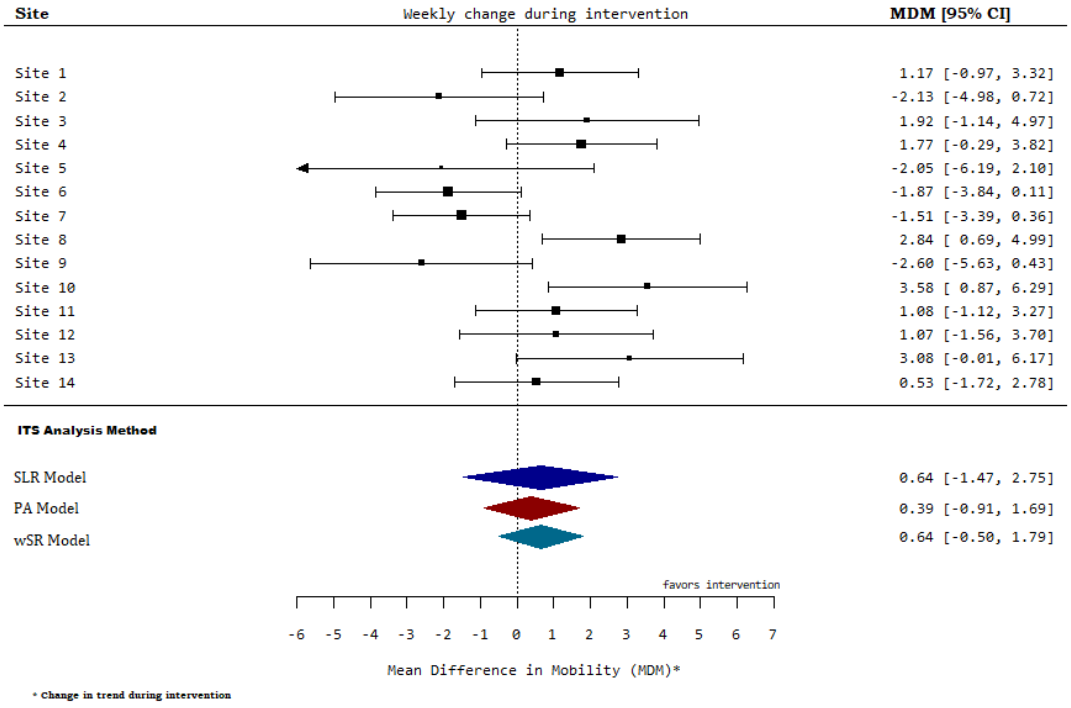

Figure 4.3: Forest plot for change in trend of mobilization during the implementation of the intervention.

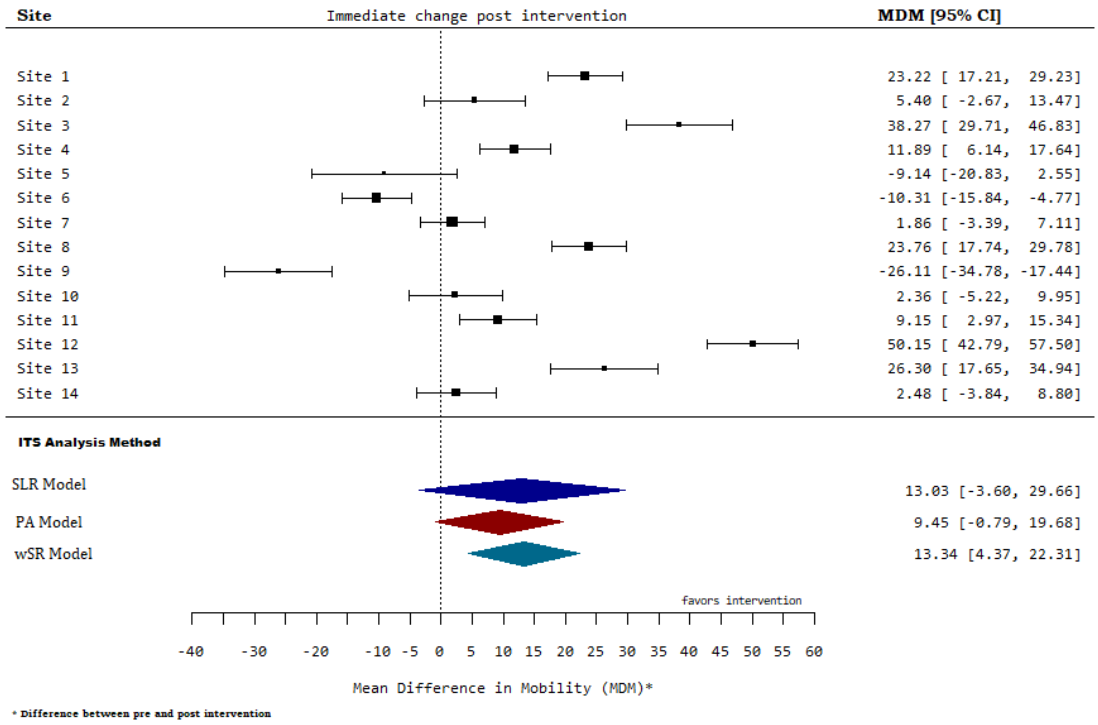

Figure 4.4: Forest plot for change in level of mobilization after the implementation of the intervention.

increased statistical power compared to the other two methods considered in this paper. For all the methods, the weekly percentage of patients who were mobile was shown to have increased by approximately 0.1% per week post intervention, albeit not significant (Table 4.1).

From the forest plots in Figures 4.2-4.4, it can be seen that there is substantial uncertainty within sites and heterogeneity across sites. To perform more extensive comparative analysis with varying levels of heterogeneity, we performed further analyses where we considered different scenarios. The results for these analyses are presented in Figures 4.5-4.11 as well as in Table 4.2.

Table 4.2: Subgroup estimates for intervention effects, 95% CI and p values obtained using segmented linear regression (SLR), pooled analysis and weighted segmented regression (wSR).

Variable	**SLR* Estimate** (95% CI)	**Pooled Estimate** (95% CI)	***w*SR** Estimate** (95% CI)
Moderate to high Heterogeneity ($I^2 \geq 30\%$) – 4 sites			
β_2 (Difference in level pre to during intervention)	12.07 (-5.61, 29.74) *p value = 0.18*	8.95 (-0.76, 18.66) *p value = 0.07*	12.07 (5.26, 18.87) *p value = 0.001*
β_3 (Difference in trend pre to during intervention)	-0.11 (-5.08, 4.87) *p value = 0.97*	-0.33 (-2.11, 1.46) *p value = 0.72*	-0.11 (-1.53, 1.31) *p value = 0.87*
δ_1 (Difference in level pre to post intervention)	19.49 (-7.46, 46.43) *p value = 0.15*	14.18 (-3.44, 31.80) *p val1ue = 0.11*	19.58 (9.21, 29.95) *p value = 0.0005*
δ_2 (Difference in trend pre to post intervention)	-0.38 (-3.46, 2.69) *p value = 0.80*	-0.19 (-1.14, 0.76) *p value = 0.70*	-0.44 (-1.31, 0.44) *p value = 0.88*
Moderate to high Heterogeneity ($I^2 \geq 50\%$) – 8 sites			
β_2 (Difference in level pre to during intervention)	2.56 (-9.27, 20.11) *p value = 0.68*	0.84 (-12.14, 13.82) *p value = 0.89*	2.62 (-7.70, 12.96) *p value = 0.61*
β_3 (Difference in trend pre to during intervention)	0.95 (-2.22, 3.49) *p value = 0.52*	0.56 (-0.82, 1.94) *p value = 0.42*	0.95 (-0.12, 2.02) *p value = 0.08*
δ_1 (Difference in level pre to post intervention)	16.84 (-8.45, 36.38) *p value = 0.09*	15.37 (-4.60, 35.34) *p val1ue = 0.13*	17.57 (1.83, 33.30) *p value = 0.03*
δ_2 (Difference in trend pre to post intervention)	0.02 (-1.60, 1.94) *p value = 0.98*	0.05 (-0.69, 0.79) *p value = 0.90*	-0.002 (-0.66, 0.66) *p value = 0.99*

*SLR: Segmented linear regression; ** *w*SR: Weighted segmented regression
p-value <0.05 implies estimate is significant

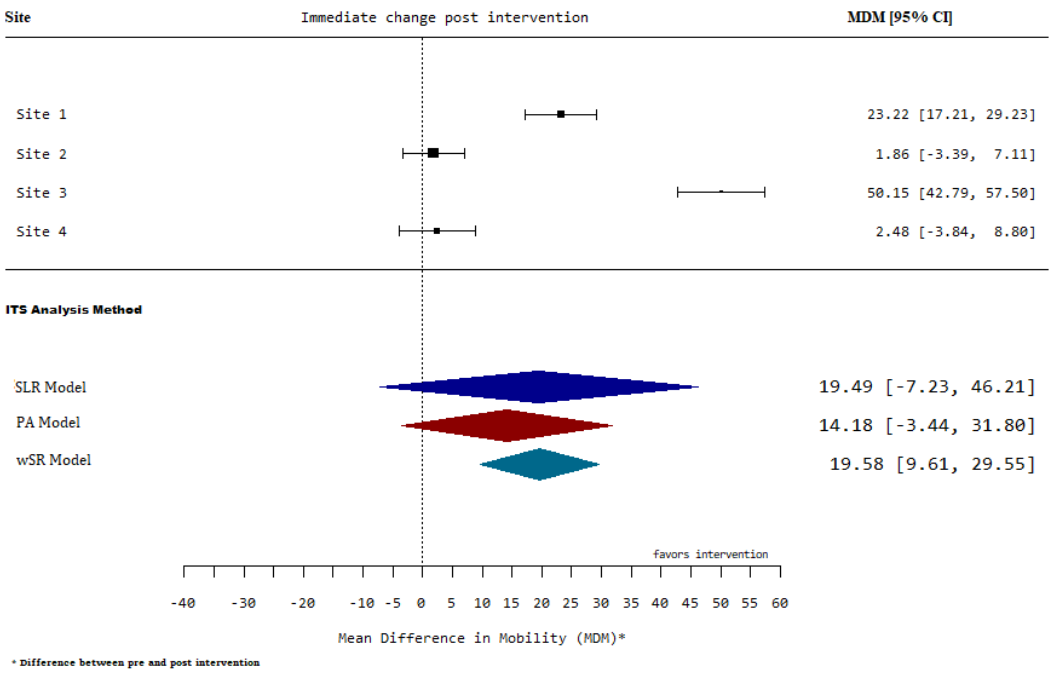

Figure 4.5: Forest plot for change in level of mobilization post intervention for high between site heterogeneity.

For analysis involving 4 sites, when the between site heterogeneity was very high, $I^2 \geq$ 75%, the wSR method produced the most precise estimate compared to the PA method and the SLR method (Figure 4.5), where the length of the confidence intervals for our estimates are significantly smaller compared to estimates based on the SLR and the PA

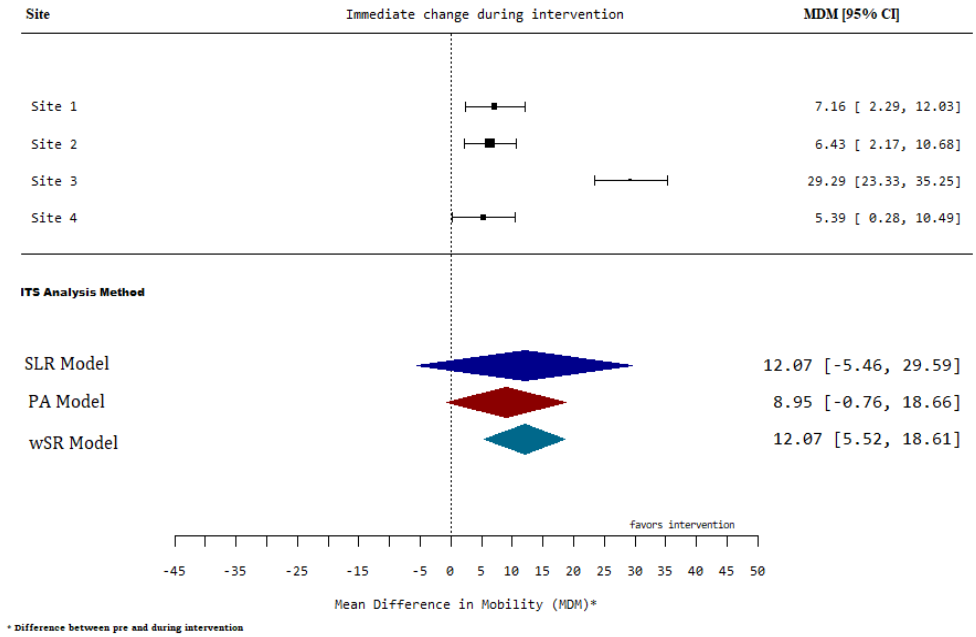

Figure 4.6: Forest plot for change in level of mobilization during intervention for moderate between site heterogeneity.

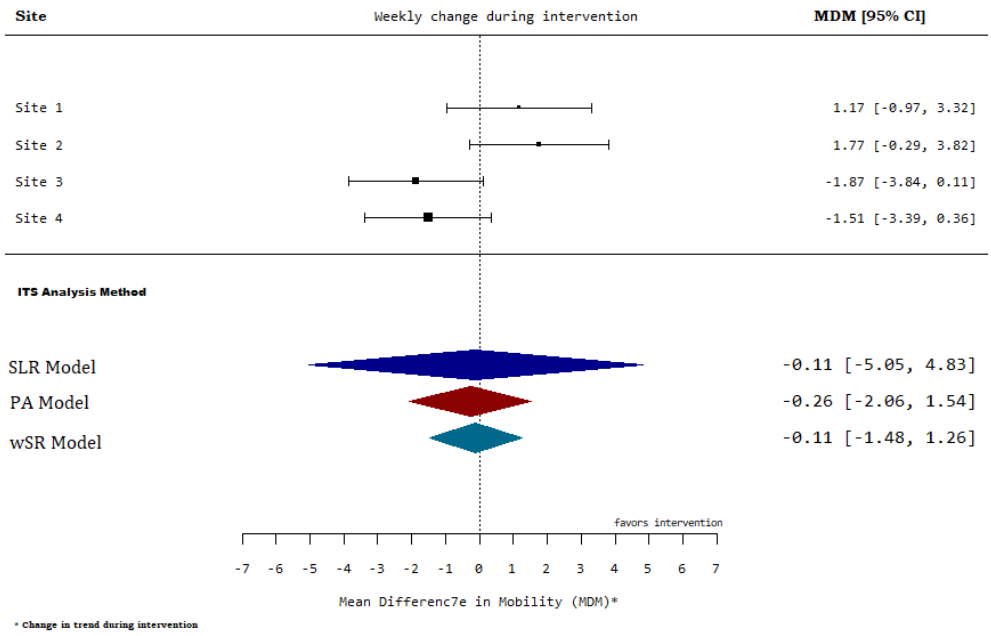

Figure 4.7: Forest plot for change in trend of mobilization during intervention for low between and moderate within site heterogeneity.

methods. Our method also led to increased statistical power demonstrated by significantly lower p-values associated with our method (Table 4.2). The same conclusions were drawn when there was moderate between site variability $I^2 \sim 50\%$ (Figure 4.6). When there was substantial within site variability and moderate between-site variability, the wSR and the PA method gave results that were similar in precision whereas the SLR still had the worst performance (Figure 4.7). When there is low between site variability, $I^2 \leq 25\%$, but substantial within site variability, the wSR method provided the most precise estimates while the PA and the SLR methods produced similar results (Figure 4.8). Interestingly,

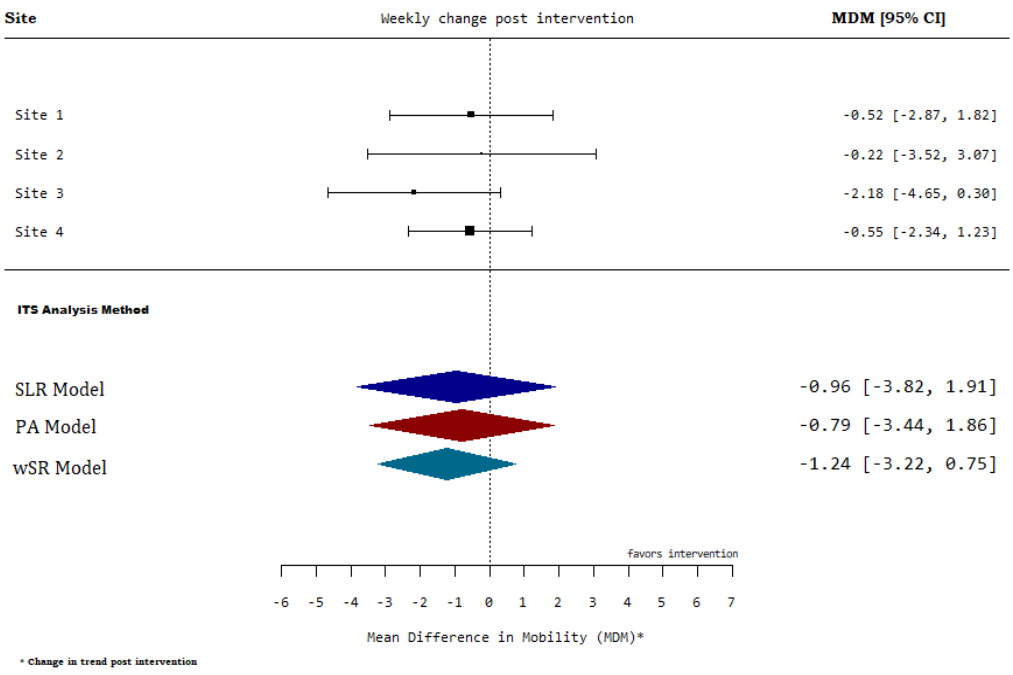

Figure 4.8: Forest plot for change in trend of mobilization post intervention for low between site heterogeneity.

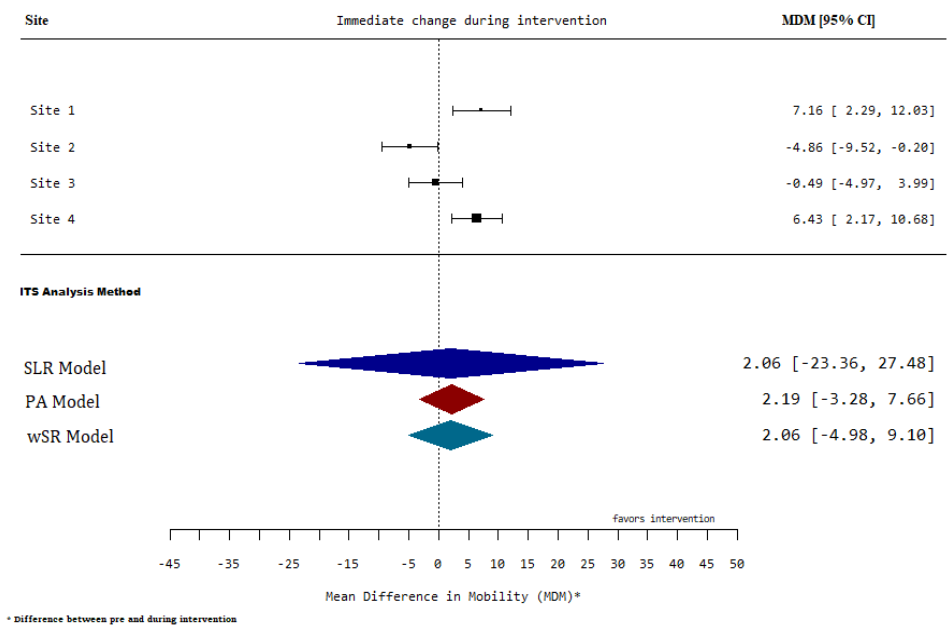

Figure 4.9: Forest plot for change in level of mobilization post intervention for low within and moderate between site heterogeneity.

even in some instances when the within and between site variability was low, the pooled method produced the most precise estimates, while the SLR performed significantly worst (Figure 4.9).

Similarly, when the number of included sites was increased to 8, the wSR method produced estimates with the narrowest confidence intervals and the smallest p-values in most cases (Table 4.2). When the between site variability was at least moderate, $I^2 \geq 50\%$, the weighted segmented regression gave the most precise estimates compared to the other two methods (Figure 4.10). Interestingly, we observed that in some cases, where there was substantial within site variability and moderate to high between site variability, the PA method produced estimates associated with the widest confidence intervals (Figure 4.11).

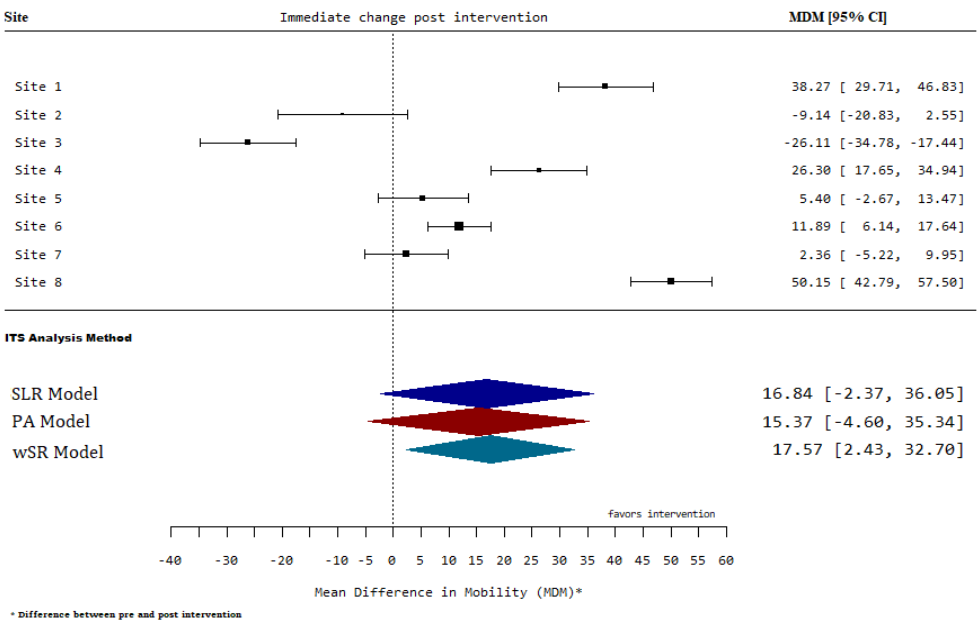

Figure 4.10: Forest plot for change in level of mobilization post intervention for high within and between site heterogeneity.

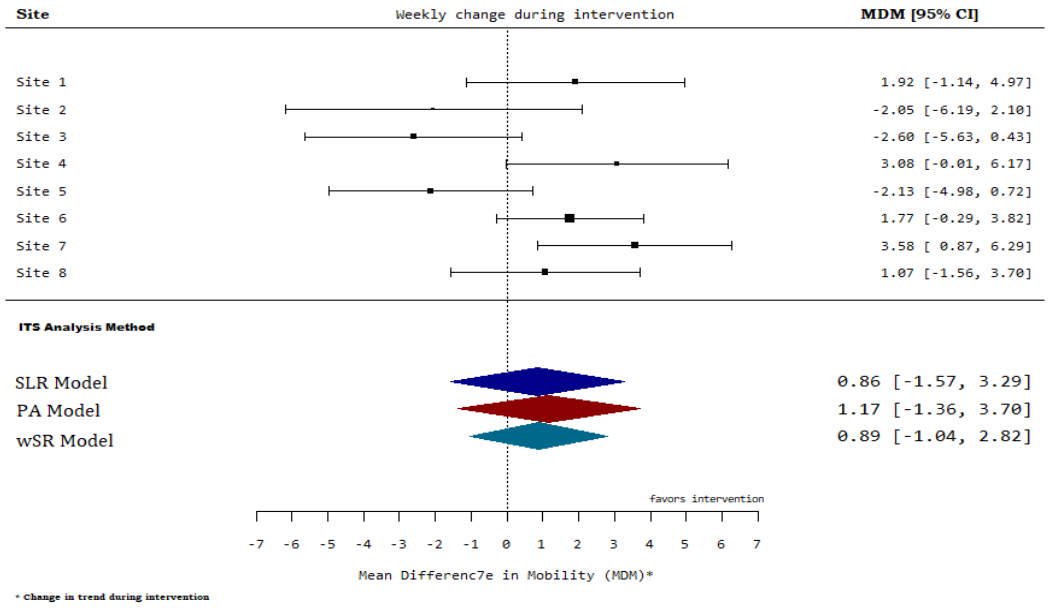

Figure 4.11: Forest plot for change in trend of mobilization post intervention for high within site variability and low between site variability.

4.4. Discussion

In this study, our aim was to improve the SLR approach, which is most commonly used in applications involving ITS studies and provide a methodological framework that allows incorporation of within site (between participants) and across site (between healthcare facilities) variability. We performed extensive empirical evaluations by creating various scenarios using a multi-site ITS study and compared our proposed method with that of the SLR and the PA method previously published.

Overall, our proposed wSR method produced estimates with the narrowest 95% confidence intervals for most of the scenarios considered, indicating that our method led to increased precision by incorporating two levels of heterogeneity in the data. The PA approach produced comparable estimates with that of the wSR in some scenarios, while the conventional SLR method had the widest 95% confidence intervals compared to the other two methods. The observed wider confidence interval for the pooled method in situations with substantial within site variability and relatively large number of sites agrees with the reported limitation of the pooled analysis method (Gebski et al., 2012). However, further investigation, preferably using extensive simulations, is required to understand performance limitations for the pooled method in comparison with the weighted analysis.

For almost all the scenarios considered, our proposed weighted method produced estimates with narrower confidence intervals leading to significantly smaller p-values, and hence allowing meaningful differences to be detected. This finding indicates that our method led to increased statistical power. Finally, we would like to highlight that the extensive empirical evaluations performed in this paper lay the ground work for further

study involving simulations to establish more extensive performance characteristics of our method and compare performance with the other methods in terms of well-established performance measures such as bias, mean square error (MSE), type I error rate and statistical power.

In conclusion, the findings in this study showed that accounting for participant variability and differences across healthcare facilities in analysis of ITS data leads to increased precision and statistical power. The results also showed that simply pooling slopes and intercepts from site level analysis has performance limitations. Taking into consideration that most ITS studies are conducted in healthcare settings and involve aggregated data, we believe this study provides findings that encourage researchers to consider differences in the participant populations and across healthcare settings. Our study also highlights the need for researchers who collect administrative routine data, which are usually presented in aggregated forms, to also report the variability associated with such data for meaningful statistical analysis to be performed.

References

Ansari, F., K. Gray, D. Nathwani, G. Phillips, S. Ogston, C. Ramsay and P. Davey (2003). "Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis." Journal of Antimicrobial Chemotherapy **52**(5): 842-848.

Bedard, N. A., D. E. DeMik, N. A. Glass, R. A. Burnett, K. J. Bozic and J. J. Callaghan (2018). "Impact of Clinical Practice Guidelines on Use of Intra-Articular Hyaluronic Acid and Corticosteroid Injections for Knee Osteoarthritis." JBJS **100**(10): 827-834.

Bernal, J. L., S. Cummins and A. Gasparrini (2017). "Interrupted time series regression for the evaluation of public health interventions: a tutorial." International journal of epidemiology **46**(1): 348-355.

Biglan, A., D. Ary and A. C. Wagenaar (2000). "The value of interrupted time-series experiments for community intervention research." Prevention Science **1**(1): 31-49.

Box, G. E., G. M. Jenkins, G. C. Reinsel and G. M. Ljung (2015). Time series analysis: forecasting and control, John Wiley & Sons.

Ewusie, J., J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient Heterogeneity Using Weighted Analysis." BMC Medical Research Methodology **(Under Review)**.

Ewusie, J., C. Soobiah, E. Blondal, J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review." BMJ Open **(Revisions Submitted)**.

Ewusie, J., C. Soobiah, E. Blondal, J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review." Revisions Submitted.

Fowler, S., A. Webber, B. Cooper, A. Phimister, K. Price, Y. Carter, C. Kibbler, A. Simpson and S. Stone (2007). "Successful use of feedback to improve antibiotic prescribing and reduce Clostridium difficile infection: a controlled interrupted time series." Journal of Antimicrobial Chemotherapy **59**(5): 990-995.

Garrett, T. A. (2003). "Aggregated versus disaggregated data in regression analysis: implications for inference." Economics Letters **81**(1): 61-65.

Gebski, V., K. Ellingson, J. Edwards, J. Jernigan and D. Kleinbaum (2012). "Modelling interrupted time series to evaluate prevention and control of infection in healthcare." Epidemiology and Infection **140**(12): 2131-2141.

Gillings, D., D. Makuc and E. Siegel (1981). "Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care." American journal of public health **71**(1): 38-46.

Hartung, D. M., M. J. Carlson, D. F. Kraemer, D. G. Haxby, K. L. Ketchum and M. R. Greenlick (2008). "Impact of a Medicaid copayment policy on prescription drug and health services utilization in a fee-for-service Medicaid population." Medical care **46**(6): 565-572.

Higgins, J. P. and S. G. Thompson (2002). "Quantifying heterogeneity in a meta-analysis." Statistics in medicine **21**(11): 1539-1558.

Higgins, J. P., S. G. Thompson, J. J. Deeks and D. G. Altman (2003). "Measuring inconsistency in meta-analyses." BMJ: British Medical Journal **327**(7414): 557.

Katikireddi, S. V., G. Der, C. Roberts and S. Haw (2016). "Has childhood smoking reduced following smoke-free public places legislation? A segmented regression analysis of cross-sectional UK school-based surveys." Nicotine & Tobacco Research **18**(7): 1670-1674.

Kong, M., A. Cambon and M. J. Smith (2012). "Extended Logistic Regression Model for Studies with Interrupted Events, Seasonal Trend, and Serial Correlation." Communications in Statistics-Theory and Methods **41**(19): 3528-3543.

Liu, B., U. Almaawiy, J. E. Moore, W.-H. Chan and S. E. Straus (2013). "Evaluation of a multisite educational intervention to improve mobilization of older patients in hospital: protocol for mobilization of vulnerable elders in Ontario (MOVE ON)." Implementation Science **8**(1): 76.

Liu, B., J. E. Moore, U. Almaawiy, W.-H. Chan, S. Khan, J. Ewusie, J. S. Hamid, S. E. Straus and M. O. Collaboration (2017). "Outcomes of Mobilisation of Vulnerable Elders in Ontario (MOVE ON): a multisite interrupted time series evaluation of an implementation intervention to increase patient mobilisation." Age and ageing **47**(1): 112-119.

Lyman, G. H. and N. M. Kuderer (2005). "The strengths and limitations of meta-analyses based on aggregate data." BMC medical research methodology **5**(1): 14.

Naimer, M. S., J. C. Kwong, D. Bhatia, R. Moineddin, M. Whelan, M. A. Campitelli, L. Macdonald, A. Lofters, A. Tuite and T. Bogler (2017). "The effect of changes in cervical cancer screening guidelines on chlamydia testing." The Annals of Family Medicine **15**(4): 329-334.

Penfold, R. B. and Z. Fang (2013). "Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements." Academic Pediatrics: S38-44.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies." International journal of technology assessment in health care **19**(4): 613-623.

Sen, A. and M. Srivastava (2012). Regression analysis: theory, methods, and applications, Springer Science & Business Media.

Smith, D. H., N. Perrin, A. Feldstein, X. Yang, D. Kuang, S. R. Simon, D. F. Sittig, R. Platt and S. B. Soumerai (2006). "The impact of prescribing safety alerts for elderly persons in an electronic medical record: an interrupted time series evaluation." Archives of Internal Medicine **166**(10): 1098-1104.

Taljaard, M., J. E. McKenzie, C. R. Ramsay and J. M. Grimshaw (2014). "The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care." Implementation Science **9**(1): 77.

Wagenaar, B. H., K. Sherr, Q. Fernandes and A. C. Wagenaar (2015). "Using routine health information systems for well-designed health evaluations in low-and middle-income countries." Health policy and planning **31**(1): 129-135.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." J Clin Pharm Ther **27**(4): 299-309.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." Journal of clinical pharmacy and therapeutics **27**(4): 299-309.

Wang, J. J., W. Scott, G. Raphael and O. Jake (2013). A comparison of statistical methods in interrupted time series analysis to estimate an intervention effect. Australasian Road Safety Research, Policing and Education Conference.

Zhang, F., A. K. Wagner, S. B. Soumerai and D. Ross-Degnan (2009). "Methods for estimating confidence intervals in interrupted time series analyses of health interventions." J Clin Epidemiol **62**(2): 143-148.

Chapter 5

CONCLUSIONS

5.1. Summary of Findings

Interrupted time series (ITS) designs have been recognized as the most robust quasi-experimental designs (QEDs) used in health research to investigate the effect of interventions or programs implemented to improve patient outcomes (Wagner et al., 2002; Penfold and Fang, 2013; Taljaard et al., 2014). Although several statistical methods exist for analyzing data from ITS studies, several limitations have been identified with the frequently used ITS methods making them suboptimal in some scenarios (Gillings et al., 1981; Ramsay et al., 2003; Zhang et al., 2009; Gebski et al., 2012; Taljaard et al., 2014).

One major limitation occurs when ITS methods are applied to aggregated data, where data at each time point are not observed, but instead estimated (summarized), and hence associated with imprecision. This issue is particularly important since ITS studies are often conducted in healthcare facilities, where data are usually aggregated from various participants within facilities (e.g. hospitals) as well as across different settings.

The projects in this thesis address the limitation related to statistical methods used for analyzing aggregated ITS data, with particular focus on segmented linear regression (SLR), which has been identified as the most commonly used statistical method (Jandoc et al., 2015; Cruz et al., 2017; Ewusie et al., 2018). Additionally, there was a need for the

existing statistical methods for ITS analysis to be identified and compared to inform researchers of the available methods. It will also inform future applied research regarding the strengths and limitations of the methods and help identify relevant methodological gaps, which will pave the way for improvements in ITS design and analysis (Ewusie et al., 2017). In this section, we will briefly summarize the key findings from the thesis and discuss the potential implications of the findings on future studies.

Motivated by the methodological limitations we encountered in the analysis of multi-site ITS studies and as a first step towards addressing these limitations, we conducted a scoping review (Ewusie et al., 2017; Ewusie et al., 2018). The review, which is presented in Chapter 2, was conducted with the objective of reviewing and synthesizing all available methods employed in the analysis of ITS, examine their strengths and limitations as well as explore their applications in health research. Description of the methods from eligible articles were provided, where we classified the studies as methodological papers or application papers depending on the focus of the included studies. We further classified the methodological papers into papers presenting novel statistical methods, papers performing a comparative assessment of traditional ITS methods and papers describing method adaptations to different areas of health research or significant additions to traditional ITS methods. Our review revealed that most of the methods papers either proposed novel methods or presented improvements of traditional methods. This shows significant advancements have been made in terms of methods used in the analysis of ITS data, particularly over the last two decades (Jandoc et al., 2015; Cruz et al., 2017; Ewusie et al., 2018). Furthermore, a substantial amount of the methods papers performed comparative

evaluation of the common statistical methods used in ITS studies using empirical data and provided information on the strengths and limitations of the methods compared. Nonetheless, a comprehensive review of the existing statistical methods for ITS analysis has not been previously conducted. For the application papers, we synthesized data and provided quantitative results, using descriptive statistics to show the frequency at which ITS is applied in various fields of health research, the settings in which it is used, the type of statistical methods utilized, and whether the required assumptions were checked or tested. The results of the review show that use of ITS designs and analysis in health research has indeed increased significantly over the last decade, where the largest increase was seen within the last 10 years (Ewusie et al., 2018). This increase was attributed to various potential reasons, including 1) the advancements in the field of implementation science in recent years, 2) the increase in the use of administrative routine data to answer important healthcare and health outcome related questions, 3) the current research focus on knowledge translation and evidence synthesis, where ITS studies are, for instance, conducted to assess the effect of clinical practice guidelines, and 4) the growing awareness of health research ethics, which makes ITS designs more favorable in certain scenarios.

Segmented linear regression (SLR) was found to be the most commonly used method for analyzing ITS data (Ewusie et al., 2018). This finding is consistent with what has been reported in the literature (Wagner et al., 2002; Zhang et al., 2009; Jandoc et al., 2015). Despite the frequency of use and the advancements made in ITS studies, our scoping review showed that limitations do exist. One of the methodological gaps observed, which is the focus of this thesis, was in analysis of aggregated ITS data. Although this shortfall

has been acknowledged in many of the previous studies (Gillings et al., 1981; Wagner et al., 2002; Gebski et al., 2012; Taljaard et al., 2014), there is little to no evidence related to the extent at which the optimality of the estimates and statistical power is impacted. Moreover, the scoping review revealed that there has not been any method developed to specifically account for the bias and imprecision introduced due to aggregation across patients within sites. For ITS studies involving multiple sites, the pooled analysis (PA) method has previously been developed. This method uses a meta-analytic approach to pool slopes and intercepts from individual site analysis and provide an estimate of overall treatment impact. However, this involved using summary measures rather than the actual data across the sites, and hence improvements are needed.

To address the methodological gap specified above, we proceeded with two steps. The first step involved incorporating the imprecision introduced due to aggregation within site, which in turn is a function of variability across participants within a site and sample size (the number of individuals at a specific time point, across whom data is aggregated). As such, we developed and presented a novel method (Ewusie et al., 2018), presented in Chapter 3, where we introduced weights into the segmented linear regression model. The weights are functions of sample size and variance, where data with large sample size and small variance are weighted higher than those with small sample size and large variance. The weighted segmented regression (wSR) model, along with the ordinary least squares (OLS) approach, is then used to analyze the ITS data. Extensive simulations involving more than 300 scenarios were conducted to establish performance criteria and compare performance with traditional SLR.

In most of the scenarios, the wSR we proposed performed uniformly better than the SLR method, where our method led to higher statistical power to detect a change in intervention effects and decreased mean squared error in estimating the model parameters. However, both methods performed relatively similar in terms of bias, when the within participant heterogeneity was low. Our simulation results also show that when the sample size was small (approximately < 30), both methods produced estimates that were relatively more biased, highlighting importance of sample size considerations, even after incorporating variations across participants. In terms of type I error rate (level of the statistical tests), the wSR and the SLR methods gave comparable results, when the sample size per time point was at least around 30. However, when the sample size was small (<30), the SLR method had a smaller error rate, indicating that the test is more conservative. We illustrated application of the proposed method using real data and the results agreed with what was observed in our simulation study. The weighted method was also a better fit based on the observed Akaike Information Criterion (AIC) value.

The second step we considered in addressing the methodological gap focused on multi-site ITS studies, where there is an additional level of heterogeneity introduced by the site to site variability (e.g. due to difference in settings across healthcare facilities) in addition to the variability across participants discussed above. We, therefore, extended the weighting scheme to include both levels of heterogeneity, where data aggregated from samples with high participant variability, high site to site variability and small sample size is given a smaller weight compared to data aggregated from those with small participant variability, small site to site variability and large sample size. This work is presented in

Chapter 4 of the thesis (Ewusie et al., 2018). We used data from a multisite study (Liu et al., 2013; Liu et al., 2017) involving 14,540 patients and 14 hospitals to empirically compare our method to the SLR method and the pooled analysis (PA) method proposed by Gebski et al., (2012).

Overall, our findings depicted that the wSR method resulted in estimates with the narrowest 95% confidence intervals (CIs) and the smallest p-values. To further explore performance of our method under different scenarios, extensive empirical evaluations were conducted, where we generated many real-world scenarios by taking samples of the empirical data with varying number of sites and magnitude of the between and within site heterogeneity. The scenarios considered include those with low within and high between site variability; low within and moderate between site variability; moderate within and between site variability; moderate within and high between site variability; moderate within and low between site variability; low within and between site variability; and high within site variability.

The results from these extensive empirical evaluations showed that the wSR produced estimates with the narrowest 95% CIs in most scenarios with moderate to high between and within site variability. For all scenarios, where there was significant improvement in mobility following the educational intervention, the results produced by the wSR method were the most precise and most accurate, and accompanied by the smallest p-values, indicating increase in power. In few instances, where there was moderate to high within and between site variability, the wSR method produced estimates that showed significant improvement in mobility while the PA and SLR methods could not establish

significance. This was due to the variability accounted for by the wSR. Based on our findings, we concluded that the wSR method was the most optimal approach for evaluating the effect of an intervention, in situations where there is at least moderate heterogeneity between sites and substantial variability within the individual sites.

5.2. Future Directions

The wSR method accounts for the imprecision introduced by aggregating data, thus helping to provide more optimal results than the SLR method particularly when there is moderate to high variability associated with the aggregated data. The matrix formulation for the weighted regression introduced in Chapters 3 and 4 of this thesis can be modified to provide a unified and standardized framework that allows application of the method for different types of outcomes and various distributions. Moreover, the method allows incorporation of correlations across the different time points, and hence allows autoregressive and seasonal parameters to be included in the analysis where necessary. Since the approach falls under the well-established weighted linear model framework, we expect all the desirable properties of estimators and tests based on weighted least squares to be satisfied. Nevertheless, extensive simulations with various scenarios of within and between site variabilities as well as sample sizes are required to confirm performance properties and conduct further comparative evaluations. Weighting with respect to variability between participants (and sites) and sample size can also be used in other ITS methods and is expected to provide similar robust results, even in the presence of outliers.

From our simulation study in project 2 (Chapter 3), we observed that performance of our method (and that of the traditional SR method) was a function of sample size per time point. That is, the number of individuals from which the aggregated data are calculated from. Nevertheless, sample size calculations in ITS studies often focus on determining the number of time points (Zhang et al., 2011), while the number of participants involved at each time point are often ignored when designing ITS studies. This is perhaps because of limited knowledge (and guidance) on how to determine the required number of participants at a given time point. This is particularly important when the study involves implementation of an intervention in a prospective study, where for instance, the observations are measured from participants in healthcare facilities.

Our simulation algorithm and the power curves generated using various scenarios with different values of variability and sample size can be used to estimate the average number of participants at each time point. For instance, in one scenario a sample size of 30 was required to achieve an 80% power to detect a 1.2 change in level post intervention. Our work, therefore, can be used in future research to develop a more extensive simulation based computational algorithm, to estimate optimal sample size required to provide precise estimates and achieve adequate statistical power to detect both immediate intervention effect (change in level) and an effect realized over time (change in trend).

Finally, future research to identify ways of facilitating the uptake of ITS methods, including methods proposed in this thesis is required. One way this can be done is by developing a statistical package in R with an accompanying user manual, which will help

future researchers in the usage of the methods. We plan to consider and tackle some of these methodological and computational directions in future research.

5.3. Concluding Remarks

The research presented in this dissertation has important implications in health research. First, to our knowledge, the scoping review is the first of its kind, providing a comprehensive overview of ITS methods and their application in health research. The review can also serve as groundwork towards developing guidelines to standardize design and analysis of ITS studies. The guidelines can provide information such as the requirements needed for a study to be classified as a well implemented ITS study, the optimal methods to use for the different data types and various distributions as well as the expected consequences of using inappropriate methods. The guideline can also provide strategies on how to design ITS studies and advance the understanding of methods for ITS analysis. Developing this guideline will lead to the standardization of ITS design and analysis, and hence facilitate better implementation of ITS studies and consequently their inclusion in evidence synthesis.

Second, the analysis of administrative routine data has become increasingly common (Ewusie et al., 2018). Such administrative data are often in aggregated forms. As we have shown in Chapters 3 & 4, there is the need for the variability associated with aggregated data to be accounted for, to obtain optimal results. The manuscripts presented in Chapters 3 & 4 hence highlight the need for researchers involved in recording administrative data to include the variability associated with the data. This is common

when conducting evidence synthesis through meta-analysis and network meta-analysis studies, where aggregated data are often encountered. Nevertheless, since most of these aggregated data come with their associated variability, meaningful statistical analysis can still be performed, where the aggregated data are weighted in the estimation of pooled treatment effects.

Third, the manuscripts presented in Chapters 3&4 of this dissertation have each contributed to the advancement of ITS research by providing improved ITS methods namely through incorporating different sources of heterogeneity in the segmented linear regression model. This will enable researchers to optimize their analysis, leading to results which are more powerful, and precise. Further, the methodological framework we presented enables our methods to be extended to different outcome types and distributions hence showing the generalizability of the methods presented in this dissertation, and hence ensuring wide spread applications.

Fourth, the simulation study presented in Chapter 3 established the relationship between sample size (e.g. number of patients at each time point) and performance criteria for both the weighted segmented regression and the segmented linear regression methods. This finding has laid the groundwork for future research related to sample size considerations, which will further contribute significantly to the optimal analysis of ITS data.

Finally, the significant upsurge in the use of ITS design and analysis to assess the effect of interventions and programs particularly in the last decade cannot be

overemphasized (Jandoc et al., 2015; Ewusie et al., 2018). This increasing trend in the utilization of ITS can largely be attributed to the growing focus of implementation science, the use of administrative routine data in current research and the rising interest in evidence-based clinical practice. In light of this increasing trend, it is imperative that optimal design and analysis methods are employed to yield the best results that will help inform health decision making. The projects undertaken in this dissertation help to achieve this by advancing our understanding of existing methods for ITS analysis and subsequently improving the shortfalls of current practice.

References

Cruz, M., M. Bender and H. Ombao (2017). "A robust interrupted time series model for analyzing complex health care intervention data." Statistics in Medicine **29**: 29.

Ewusie, J., J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "Analysis of Multi-center Interrupted Time Series Data: Incorporating Within and Between Center Heterogeneity." Stat Methods and Applications **(Submitted)**.

Ewusie, J., J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient Heterogeneity Using Weighted Analysis." BMC Medical Research Methodology **(Under Review)**.

Ewusie, J., C. Soobiah, E. Blondal, J. Beyene, L. Thabane, E. Straus Sharon and J. Hamid (2018). "Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review." BMJ Open **(Revisions Submitted)**.

Ewusie, J. E., E. Blondal, C. Soobiah, J. Beyene, L. Thabane, S. E. Straus and J. S. Hamid (2017). "Methods, applications, interpretations and challenges of interrupted time series (ITS) data: protocol for a scoping review." BMJ Open **7**(6): e016018.

Gebski, V., K. Ellingson, J. Edwards, J. Jernigan and D. Kleinbaum (2012). "Modelling interrupted time series to evaluate prevention and control of infection in healthcare." Epidemiology and Infection **140**(12): 2131-2141.

Gillings, D., D. Makuc and E. Siegel (1981). "Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care." American journal of public health **71**(1): 38-46.

Jandoc, R., A. M. Burden, M. Mamdani, L. E. Lévesque and S. M. Cadarette (2015). "Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations." Journal of clinical epidemiology **68**(8): 950-956.

Liu, B., U. Almaawiy, J. E. Moore, W.-H. Chan and S. E. Straus (2013). "Evaluation of a multisite educational intervention to improve mobilization of older patients in hospital: protocol for mobilization of vulnerable elders in Ontario (MOVE ON)." Implementation Science **8**(1): 76.

Liu, B., J. E. Moore, U. Almaawiy, W.-H. Chan, S. Khan, J. Ewusie, J. S. Hamid, S. E. Straus and M. O. Collaboration (2017). "Outcomes of Mobilisation of Vulnerable Elders in Ontario (MOVE ON): a multisite interrupted time series evaluation of an implementation intervention to increase patient mobilisation." Age and ageing **47**(1): 112-119.

Penfold, R. B. and Z. Fang (2013). "Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements." Academic Pediatrics: S38-44.

Ramsay, C. R., L. Matowe, R. Grilli, J. M. Grimshaw and R. E. Thomas (2003). "Interrupted time series designs in health technology assessment: lessons from two

systematic reviews of behavior change strategies." Int J Technol Assess Health Care **19**(4): 613-623.

Taljaard, M., J. E. McKenzie, C. R. Ramsay and J. M. Grimshaw (2014). "The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care." Implement Sci **9**: 77.

Wagner, A. K., S. B. Soumerai, F. Zhang and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research." J Clin Pharm Ther **27**(4): 299-309.

Zhang, F., A. K. Wagner and D. Ross-Degnan (2011). "Simulation-based power calculation for designing interrupted time series analyses of health policy interventions." Journal of clinical epidemiology **64**(11): 1252-1261.

Zhang, F., A. K. Wagner, S. B. Soumerai and D. Ross-Degnan (2009). "Methods for estimating confidence intervals in interrupted time series analyses of health interventions." J Clin Epidemiol **62**(2): 143-148.

APPENDIX

Table A1: Summary of Methods Papers Included in the Scoping Review

Author (Year)	Method	Method Description	Number of time points	Type of outcome used in study	Software Used	Assumptions Made/tested	Advantages of method*	Limitations of method*
Cruz M., et al. (2017)	Robust interrupted time series model	- The method is developed in two stages. - In the first stage, a set of plausible change points are established based on the research question. Based on these change points, the mean parameters are estimated using ordinary least squares with the likelihood approach. - The change point whose parameter maximizes the likelihood is then chosen as the optimal change point. - In the second stage, residuals obtained in the first stage are used to examine and determine the structure of the stochastic process. - If the residuals act as white noise, it implies there is no correlation and hence the variances before and after the estimated change point	- min. of 5 per period plus the length of possible change points	Continuous	R Shiny toolbox	- Data is linear. - There is no seasonality. - Data is stationary.	- Allows the change point to be variable and different from the time of intervention - Provides a better model fit compared to traditional segmented regression with respect to mean squared error - Allows the easy assessment of the effect of intervention on the correlation structure - Ability to conduct variance comparisons in the absence of correlation allows for clear inference on the possible effects of intervention. - Method proposed has been developed into an interactive and user-friendly software"	only applicable for continuous outcomes

		are compared using the F test. - Otherwise, an ARIMA process is fit on the residuals in each phase separately. From this process, estimates of the correlation and variances are obtained using conditional likelihood methods.						
Duncan, T. E.et al. (2004) [39]	Latent Growth Curve Modeling	- The method involves a growth curve model - The growth model captures both the intercept and slope differences over baseline and treatment intervention periods in a single sample. - The added growth intercept and slope determines the intervention effect - The model gives linear and average level change for both phases of the study.	- at least 3 per period	Continuous, Likert scale	Mplus	- Constraints are placed on specific model parameters, e.g. means of intercepts and slopes for both phases are constrained to be equal. - Individual level correlation is assumed to be sufficiently homogeneous	- Any number of time points can be used once they can be modeled by the growth form. - Measurement errors in the model specification are incorporated in model. - Allows evaluation of treatment effects in the presence of different covariates. - Allows statistical tests of level and slope mean differences and prediction of changes in treatment outcomes and time-invariant and time-varying covariates.	- An increase in measured variables results in longer time series, which may not be defined adequately by a linear growth form. - Longer time series increases random error variance estimated by the model.
Park, J. H.(2012)	Intervened time series central mean subspace	- This method involves the dimension reduction technique. - The dimension reduction technique is used to analyze univariate time series with interventions as a covariate. - Specifically, a central mean subspace for a	Not specified	Continuous	Not stated	- The univariate time series and the conditional mean function are conditionally independent. - Their covariance is zero.	- Supplements and strengthens the conventional ARIMA. - Can identify intervention effects even before model is specified. - Simple and intuitive and does not require initial assumptions or pre-intervention modeling.	- Requires large sample size to increase accuracy of the model, at least 100 was used in the study.

		univariate time series that includes several change points is built, where the reduction in dimension is focused on the conditional mean function.				- The conditional mean is a function of the subspace.	- The method does not require specification of the model.	
Sun, P. et al. (2012)	Extended ARIMA	- Involves a stochastic component and a structural or intervention component. - The stochastic component is modeled by ARIMA. - The intervention component of the model is divided into 4 subcomponents: a) marginal change in outcome before intervention b) marginal change in outcome in the presence of intervention c) short-term change in outcome and d) additional impact of intervention initiation over the observed time.	- at least 20 per period	Continuous	Not stated	- Tested stationarity with the Dickey-Fuller (ADF) test. - Assessed residuals with Kolmogorov-Smimov and Ljung Box test. - Assessed homoscedasticity using Levene test	- Allows the identification of immediate and gradual changes in the outcome. - Allows the assessment of the number of post intervention time points before the offset point after which the outcome after intervention was at or lower than the pre-intervention phase.	- The threats to internal validity of ITS model, such as effect of concomitant interventions cannot be accounted for.
Gebski, V. et al. (2012)	Pooled and Stacked ITS analysis	- The pooled analysis fits a separate segmented regression for each unit. - The overall effect is then obtained by calculating the weighted average of the estimates of the parameters (level and slope) for each unit. - The weights are the inverse variances of the unit-level estimates.	Not specified	Count (rate)	Not stated	- Assessed autocorrelation using Dublin Watson approach	- Pooled analysis allows statistical adjustment for each unit. - Modelling individual units provides individual estimates, which allows for the evaluation of individual units. - Pooling overall effect using weighted analysis accounts for heterogeneity between units.	- Increase in units leads to increase in parameters required and may lead to over dispersion. - Increase in number of parameters for the stacked ITS analysis leads to an

		- The second method, stacked analysis, involves fitting a single model involving all the units. - The unit effect is accounted for by adding parameters to the model to represent the different units. - One unit is used as reference.					- Accounts for extra variability due to the inclusion of the unit effect in both analysis.	increase in Type 1 error.
Huitema, B. E. et al. (2014)	Extended time series regression model	- The method comprises a within and between unit analysis. - The within analysis is performed for ITS designs where intervention is rolled out over time across multiple units - The model includes a penetration variable used to evaluate the degree of intervention penetration present during any time point. - The penetration variable is a continuous variable indicating the extent of intervention penetration. - The variable is calculated as the ratio of the number of sites with intervention to overall number of sites and ranges from 0 to 1.	At least 20 per period	Count	Not stated	- Accounted for autocorrelation using autoregressive parameters included in the intervention model.	- If an effect is identified, the function estimated from the regression on penetration can be used to predict the outcome from the degree of intervention penetration. - Intervention need not be fully introduced right after baseline since the penetration variable accounts for this. - Allows the unknown external events occurring at the same time as the intervention (confounders) to be identified or accounted for.	- Time consuming and cumbersome. - Model does not account for heterogeneity among unit.
Kong, M. et al. (2012)	Extended Logistic	-The method is an extension of the logistic regression approach.	Not specified	Count and dichotomous	R	-Autocorrelation was assessed using Durbin Watson	-Suitable for ITS studies with count or binary outcomes.	-Does not account for the variability introduced due to

	Regression Model	- The method follows the quasi-likelihood approach. - It involves including variables in the logistic regression model to account for intervention effect. - Harmonic functions are also included to account for seasonality. -Autoregressive parameters are incorporated to account for serial correlation. -Estimation procedure for parameters is based on the fishers scoring method.					-Accounts for seasonality and serial correlation. -Provides different procedures for estimating parameters and associated variances.	summarized (aggregated) outcome.
Velicer, W. F. (1994)	Pooled time series analysis	- The method is an extension of a previous analysis involving a single unit to multiple units. - Method involves using a patterned transformation matrix which transforms the serially dependent variables to serially independent variables. - The choice of matrix and number of units varies by research interest. - The design matrix is defined based on the parameters of interest (level and trend) and differences between units.	Not specified	Not stated	GENTS and TSX	- All units are assumed to have a common transformation matrix	- Avoids the model identification step in ARIMA models. - It can be easily adapted to other modeling approaches. - It can be easily. implemented with slight modifications in existing computer programs.	- Requires substantial understanding of the transformation approach. - Units are assumed to have a common transformation matrix which may not always be the case.

Fretheim, A. et al. (2015)	Controlled segmented regression	- For this method, the differences in outcome between control and intervention groups are calculated for each time point. - The new set of values are then used to find the difference in slope and level changes between the treatment and control groups using segmented regression modeling.	At least 6 per period	Count (proportion or frequency)	STATA (REGRESS, PRAIS, ESTAT DWATSON , XTGEE)	- Assessed first order autocorrelation using Durbin Watson and adjusted using Prias-Winston method	- Allows the detection of anomalous effects and identification of co interventions. - Allows the detection of important dynamics that enables the understanding of effects over time and triggers further qualitative explorations.	- Less than 6 time points per period is too small for reliable ITS estimates.
Linden, A. et al. (2011)	Propensity score-based weighted ITS	- A weighted modelling technique of time series analysis following three steps; - First, the propensity score is estimated for the treatment group and all controls. - Second, weights are constructed based on these scores and the treatment assignment. - Third, the weights are used in a regression framework to provide the treatment effect estimates. - The ATT weighting approach was used to calculate the weights.	At least 15 per period	Count	STATA	- First order autocorrelation was assumed and tested using Durbin Watson and adjusted using Prias-Winston.	- It is technically less complicated to implement and rooted in regression techniques. - Can be implemented using any stats software without elaborate programming. - May accommodate any number of treatment units. - Allows for greater flexibility in the choice of treatment effect estimators (i.e. ATT or ATE or ATE on controls)"	- The assumption that all biases and confounding have been adjusted for in the model cannot be tested. - Control groups must have substantial overlap with treatment group for weighting to be effective.

Pechlivanoglou, P. et al. (2015)	State-space approach	- The model consists of an observation equation and a system or state equation. - The state equation contains the level and trend parameters that are affected by the intervention and can vary over time. - The model consists of a pulse variable which captures the instantaneous effects of specific events. - It also consists of the permanent and temporary level and trend variables to capture the level and trend effects of the intervention. - The model allows the addition of other variables to capture the effect of cointerventions.	Not specified	Count/ continuous	R	- Assessed Independence, heteroscedasticity and normality of residuals using Durbin and Koopman procedures.	- Handles non-stationarity issues in a straight forward way and do not require unit root testing procedure. - Suitable for studying multiple interventions. - Can deal with missing observations easier. - Allows large models to be estimated without disproportionate increase in computational times. - The specification of explanatory variables in either or both the observation and state equations allow for controlling for external parameters that might impact the studied variables. - Available in standard statistical software and hence convenient and easy to implement.	- Co-interventions are difficult to disentangle and might affect overall results. - Assumptions made about the comparability of the control group cannot be tested by the model. - Requires the control group to be as comparable as possible for optimum results.

*Includes advantages and limitations as reported in the articles by the authors.