Statistical methods for testing carryover effects: A mixed effects model approach

S. Gwynn Sturdevant a,*, Thomas Lumley b

a Laboratory for Innovation Science at Harvard, 175 N. Harvard Street Suite 1350, Boston, MA, 02134, USA
b Department of Statistics, University of Auckland, Private Bag 92019, Auckland, 1142, New Zealand

ABSTRACT

Carryover, or the effects of treatment after it ceases, has been largely ignored in statistical literature except as a nuisance parameter. When testing for carryover, comparing cumulative incidence rates is biased when diagnosis is based on a noisy measurement crossing a threshold (such as in blood pressure) then followed by open-label treatment. This issue was raised in the context of preventing hypertension by the TROPHY trial. We show that modelling the noisy measurement itself using linear mixed effect models, then computing the expected proportion over the threshold, gives valid tests and consistent estimates. The key insight is that the data made unavailable by open-label treatment after diagnosis are missing at random. We demonstrate the analysis in simulations based on a large set of blood pressure measurements from a New Zealand healthcare organisation and show that properly specified random effects models accurately estimate carryover effects even in the presence of data censored at diagnosis.

1. Introduction

One goal of healthcare is prevention; “prevention is better than treatment” [1]. As such, some researchers want to investigate treatments that delay the onset of diseases, this is equivalent to measuring the carryover effects. Unfortunately, statistical methodology has not yet been developed that enables investigators to test a carryover hypothesis. We use Stephen Senn’s Definition of carryover effects as the “residual effects” of treatment after it has ceased [2]. He also laments “that little has appeared... on the subject of modelling for carry-over [sic] that is more grounded in clinical and pharmacological reality” [2]. This article begins to fill this gap. Recently, three studies have attempted to measure carryover effects of pharmacological treatments [3–5] but all used faulty methodology. For the purpose of concreteness we focus on one, the TROPHY trial.

Rates of diabetes and hypertension are on the rise and responsible for increased death and disease [6]. To explore a prevention strategy to reducing rates of hypertension, AstraZeneca tested Candesartan for carryover effects. Since parallel-group trials with blinding are the gold standard for determining causality [7], TROPHY trial investigators believed it could effectively test a carryover hypothesis. They randomised 809 participants with blood pressure (BP) close to the 140/90 mm Hg threshold for hypertension to the treatment arm (Candesartan for 2 years followed by 2 years of monitoring) or placebo for 4 years. Fig. 1 plots a possible BP trajectory of a participant in the treatment arm of the TROPHY trial which involved BP measurements every 3 months for the duration of the study. Investigators diagnosed participants with hypertension when any 3 measurements were above the 140/90 mm Hg threshold. Cumulative incidence in the treatment arm was 53.2% and 63.0% in the placebo arm. The investigators concluded that “the effect of active treatment on delaying the onset of hypertension can extend up to 2 years after the discontinuation of treatment” [5].

Other authors have criticised TROPHY’s design. Meltzer (2006) said an “idiosyncratic primary endpoint seriously impairs external applicability.” Persell and Baker (2006) noted that cumulative diagnosis rates would differ even with identical underlying BP.

A person is diagnosed with hypertension when their long-term average BP exceeds a threshold. However, BP measurements vary for multiple reasons [8–13]. Investigators must carefully consider this variation when analysing data as the instantaneous BP, not the long-term average, is used in most analyses, and was used in TROPHY. The long-term average BP requires multiple measurements over multiple days and is.

Not practical to implement. Noisy measurements complicates
diagnosis of hypertension which occurs when long-term average BP crosses a threshold; localising when the threshold is crossed is challenging.

We can visualize the problem with the TROPHY design by simulating systolic BP ($Y_{it}$ for individual $i$ at time $t_j$) as normally distributed around an individual-specific long-term average trend

$$Y_{it} = a_i + b_j t_j + c_i X_{it} + d_i Z_{it} + \varepsilon_{it}$$

where $Y_{it}$ is the BP measurement of person $i$ at time $t_j$, $a_i \sim \text{Unif}(125, 140)$, $b_i \sim N(0, \Sigma)$, $c_i$ estimates the treatment effects, $d_i$ estimates the carryover, $X_{it}$ is 1 if person $i$ is on treatment at time $t$ and 0 otherwise, and $Z_{it}$ starts at 1 when someone stops treatment and decreases linearly to 0 over the predefined carryover period. TROPHY’s inclusion criteria allowed for randomisation of participants with baseline BP close to the threshold of 140/90 mm Hg. To account for this, we modeled a uniform distribution on the individual random intercepts.

Fig. 2 visualises variation and its impact upon diagnosis using the threshold of 140 mm Hg. The plot on the left includes long-term average values for both treatment and control arms. The plot on the right includes both long-term and instantaneous BP for the treatment arm. In the plot on the right, two points show the impact of variation on instantaneous BP. At time 0 there are two instantaneous measurements: one slightly before treatment starts, the other after. At the first measurement, the participant has instantaneous BP in excess of the threshold despite the long-term average BP being below it. For the second measurement, because the long-term average BP is much lower, the instantaneous BP is not likely to be in excess of the threshold. Medicated participants have lower long-term average BP, so are less likely to have instantaneous BP in excess of the threshold.

In the TROPHY trial, hypertension was diagnosed when 3 measurements were above the threshold and the treatment arm had lower BP for the first two years. Random variation likely resulted in measurements in excess of the threshold in the control arm regardless of carryover and when the long-term average measurement was in fact below the threshold. Furthermore, simulations suggest an 80% Type I error rate for the null hypothesis of no carryover in the underlying data [14].

Another complication of the TROPHY trial is missing data. After
diagnosis occurs, participants must be treated for hypertension and subsequent measurements cannot be used to determine the long-term average BP. This further impairs our ability to test a carryover hypothesis.

A more thorough discussion of the faults in TROPHY and carryover effects can be found in Sturdevant and Lumley (2016) [15].

1.1. Solutions

We explored several methods to test a carryover hypothesis: a simple design, a rerandomisation design, methodology similar to the Prostate Cancer Prevention Trial (PCPT), and using a mixed effect model to effectively impute missing data.

In a simple design, comparing cumulative incidences in a parallel or a crossover design is ineffective at testing a carryover hypothesis [15]. In the parallel design, noisy measurements combined with treatment result in diagnosis rates which are consistently lower in the treatment arm regardless of carryover. In the crossover design, a nominal trend results in more diagnoses in the second period which skews estimates. In addition, a three arm trial with both parallel, crossover and control arms are also ineffective. Sturdevant and Lumley (2016) [15] discuss this further.

In a rerandomisation design, participants are randomised to treatment or control at the beginning of the study, after the first period the remaining participants are rerandomised to treatment or control [16]. Applying this method to TROPHY would require investigators rerandomise remaining participants after treatment has ceased in the treatment arm, then analysing data. Nason suggests two methods for analysis: mixture models and the Maentel—Haenszel estimator. Our data violates one assumption for the mixture model, that there is no “period effect” [16]. In BP measurements, there is a non-negligible trend which results in more diagnosis in the second period [14,17]; this also caused the crossover design to fail [15]. The second approach uses the Maentel—Haenszel estimator. This estimator compares risks of participants with the same treatment in the past period but different treatments in the current period [18]. Investigators would compare risks of participants who both had treatment in the first period, but are currently on treatment or control. Unfortunately, our question differs: for people with different treatments in the past period but the same (control) treatment in the current period, how does the risk differ? It may be possible to modify the estimator to look at carryover, but it does not currently. Currently, rerandomisation is not a feasible solution to test a carryover hypothesis.

The PCPT was designed to determine if treatment with finasteride was effective at preventing prostate cancer [19]. The investigators randomised 18,882 men to treatment or placebo for 7 years. Finasteride impacts upon PSA levels which increases rates of diagnosis in the treatment arm. The investigators addressed this in two ways: all participants were required to have biopsies at the end of the study, and the investigators used different thresholds of PSA levels in the two arms of the study to determine if a biopsy was necessary; high levels were sent to a statistical centre for analysis [19].

Although extremely burdensome, PCPT methodologies could be implemented as a solution. Ambulatory BP (participants wear a cuff for an extended time which measures BP every 15–30 min) could be used a few years after the end of the trial to diagnose hypertension in all participants [20]. A wash-out period is necessary to remove any carryover effects.

Although extremely dubious, thresholds in the treatment and control arms could have been adjusted to allow similar diagnosis rates in both arms. An initial threshold in both arms of 140–90 mm Hg could be used. The first half of the trial would require more participants diagnosed in the treatment group, as their BP is artificially lowered. The second half requires more diagnosis in the control group. A further complication is variation in the impact of treatment on BP [21]. Diagnosing people with the highest BP with treatment may not equate to diagnosing people with the highest underlying BP. We believe that simplier methodologies exist.

As a simple design, a rerandomisation design, and methodology similar to the PCPT are either not feasible nor practical for this problem, the rest of this article will discuss a mixed effect model.

1.2. Mixed effect model

As we discussed in the introduction, analysis of data is complicated due to missing data. However, all data subsequent to diagnosis is missing deterministically based upon previous measurements. As missing data is based on previous measurements, the missing data is missing at random and a linear mixed model should give unbiased estimates of carryover parameters.

Mathematically, to discuss diagnosis times we define $D_t$ to be the time person $i$ is diagnosed as hypertensive. We let $Y_{it}$ be the systolic BP of person $i$ at time $t$ and $Y_{it}$ be the counterfactorial trajectory of systolic BP of person $i$ at time $t$ without treatment after diagnosis. To discuss how data relate to diagnosis we use rule $f_{it}$ laid out in trial protocol that defines when person $i$ is diagnosed as hypertensive using $Y_{it}$. For example, the TROPHY trial diagnosed people with hypertension when 3 measurements were above the threshold. We have that $Y_{it} = Y_{it} > D_{it}$ (until diagnosis occurs). After $D_{it}$ they differ due to $f_{it}$: if person $i$ is in the control group after diagnosis they will receive treatment per trial protocol, if not they will receive new treatment or more treatment after diagnosis. $f_{it}$ results in measurements $Y_{it}, t > D_{it}$ not adhering to Eq (2).

The $D_{it}$ can be seen as a stopping time for the stochastic process.

$$
Y_{it} = a_i + \beta Y_{it} + \delta Z_{it} + \epsilon_{it}
$$

where $Y_{it}$ is the systolic BP measurement, $a_i$ estimates the intercept for person $i$, $\beta$ is the average trend, $\delta$ is the average treatment effect, $Z_{it}$ is 1 if person $i$ is on treatment at time $t$ and 0 otherwise, and $\epsilon_{it}$ is 1 at the beginning of period 2, when treatment ceases, and decreases linearly to 0 over the predefined carryover period. Unfortunately, we cannot access all $Y_{it}$ as diagnosis occurs and subsequent data is missing due to ensuing treatment beginning when $f_{it} = 1$.

To estimate parameters we analyse $Y_{it}$ where data is available only until diagnosis. The unavailability of all $Y_{it}$ results in biased parameter estimates if analysis is based on a linear model.

As our missing data is due to diagnosis based upon past measurements, and none of the parameters have any impact upon diagnosis, our data is missing at random [22]. We formalise in Theorem 1.1.

**Definition.** $D_t = d$ if $f_{it} = 1$ and $f_{it} = 0$ for all $t < d$.

**Theorem 1.1.** If we set $Y_{it}$ to missing for $t > D_{it}$ it is missing at random.

**Proof.** Diagnosing person $i$ is based on $Y_{it}$ in a deterministic manner defined by $f_{it}$, $D_t = d$ is completely determined by $Y_{it}, t < d$, and therefore is independent of $Y_{it}, t > D_t$ conditional on $Y_{it}, t < d$. If $Y_{it}, t > D_t$ are treated as missing, they are missing at random.

If we coerced $Y_{it}$ then missing data after diagnosis would be missing at random only if diagnosis was deterministic given the coerced data. For example, if $A_{it}$ was the indicator that $Y_{it} \geq 140$ missing data after diagnosis would be missing at random if $f_{it}$ defined hypertension by counting the number of measurements above the 140 mm Hg. It would not be missing at random if $f_{it}$ diagnosed hypertension by averaging measurements as the likelihood ceases to factor into independent products with only one containing parameters. For the rest of this article we refer to $Y_{it}, Y_{it} > D_t$ as $Y_{it}$.

If data missing after diagnosis are missing at random, the likelihood of the parameters factors into the data distribution and a function of the parameters for the probability of missingness. These two sets of parameters differ and are independent. The restricted maximum likelihood of the parameters is dependent on only the likelihood for the parame-
ters; the probability of missingness is irrelevant. This likelihood for person \( i \) is:

\[
l_i = \sum_{j=1}^{M} \left( -\frac{1}{2} \log|\Sigma| - \frac{1}{2} \left( Y_{ij} - (a_i + t_{ij} + X_{ij} + Z_{ij})\right)\Sigma\left( Y_{ij} - (a_i + t_{ij} + X_{ij} + Z_{ij})\right) \right)
\]

with measurement times \( t_j = t_1, t_2, ..., t_M \). The likelihood for the whole data set is \( L = \prod_{i=1}^{N} l_i \) where there are \( i = 1, 2, ..., N \) participants.

We found the restricted maximum likelihood (REML) estimates by maximising \( \int L \theta \) where \( \theta = (\alpha, \beta, \gamma, \delta)^T \). This gave consistent estimates

\[
\begin{pmatrix}
\hat{a}_i
\hat{b}_i
\hat{c}_i
\hat{d}_i
\end{pmatrix}
\]

of distribution \( N \left[ \begin{pmatrix}
\alpha
\beta
\gamma
\delta
\end{pmatrix}, \Sigma \right] \)

where \( \Sigma \) is an unconstrained covariance matrix not assumed to be diagonal. We optimised the likelihood using the lmer function from the lme4 package in R.

With this model, we used a parametric bootstrap to find the risk ratio (\( rr \)) of hypertension. We define the \( rr \) as "a measure of the risk of a certain event happening in one group compared to the risk of the same event happening in another group \([23]\)." Mathematically, we have

\[
rr = \frac{TE/(TE + TN)}{CE/(CE + CN)}
\]

where \( TE \) is the number of people with hypertension in the treatment arm, \( TN \) the number of non-hypertensives in the treatment arm, \( CE \) is the number of people with hypertension in the control arm, and \( CN \) the number of non-hypertensives in the control arm. We bootstrapped the risk ratio for diagnosis of hypertension at the first time point after treatment ceases. In Fig. 1 we find the \( rr \) 2 years and 3 months after the start of the trial, when systolic BP with no carryover and all other lengths of carryover differs the most.

To find the true risk ratio of hypertension from the mixed model, Eq (2), we simulated 10^4 participants with no measurement error (\( \epsilon_{ij} = 0 \)). When carryover is 0, at the first measurement after treatment ceased, the risk ratios were centred around 1. Risk ratios centred on 1 suggests that diagnosis of hypertension is equiprobable in the two arms: because there is no carryover, the effects of treatment cease immediately, with no residual benefits of treatment, diagnosis rates in the two arms should be similar.

To test a carryover hypothesis fit a mixed model using lmer; conduct a parametric bootstrap of the risk ratio for the first measurement after treatment ceases using bootMer and check to see if 1 is in the confidence interval. If it is, based upon our simulations there is a 99% probability that no carryover was in the model used to simulate the data. After some preliminary data analysis to ensure the assumptions of our models, the details follow in Section 2.

1.3. Exploring the correlation structure

Assumptions for a linear mixed model are: random effects in the data, normality of random effects, and errors are normal \( \epsilon_{ij} \sim N(\mu = 0, \sigma^2) \), and uncorrelated \([24]\). The first two assumptions show that when using linear mixed models, correct specification of the random effects is important. To ensure a random effects model was realistic, we analysed a large dataset of longitudinal BP.

Repeated BP measurements were found in the chronic care management programme in Counties Manukau. Counties Manukau has high Pacific.

- **Fig. 3.** Q-Q Plots for both systolic and diastolic blood pressures. Besides terminal digit preference, there is not evidence against normality.

- **Fig. 4.** The relative flatness of the variogram shows no autocorrelation. The red line indicates a 4 year lag, which is the length of our study.
Table 1
Summary of fixed and random effects trends for systolic and diastolic blood pressures. Random effects represent the data better.

|                | Less than 50 | More than 50 |
|----------------|--------------|--------------|
| **Systolic**   |              |              |
| Fixed trend    | 70%          | 64%          |
| Random trend   | 97%          | 92%          |
| **Diastolic**  |              |              |
| Fixed trend    | 85%          | 72%          |
| Random trend   | 97%          | 96%          |

Islander and Maaori population [25]; 34% of the population lives in “very deprived” conditions [25]. The District Health Board analysed increasing rates of acute adult medical admissions and found that they could be reduced by up to 30% “with more timely primary care intervention.” The aim of the study was “to develop an effective and efficient process for the seamless delivery of care for targeted patients with specific chronic diseases.” One of the critical objectives was “to achieve best-practice management of ... blood pressure control” [25]. We obtained 103,098 repeated BP measurements from 9,043 people from the chronic care management programme in Counties Manukau [25].

Our purpose was to explore the correlation structure of BP for people who were prehypertensive. To have data that reflected this population, we deleted rows with missing systolic measurements or systolic measurements below 50 mm Hg or above 160 mm Hg. The elevated upper bound of 160 mm Hg was due to the uncertainty in measurement; systolic BP can vary up to 20 mm Hg in any given day [26]. We deleted diastolic measurements less than 30 mm Hg which were likely the result of a recording error. Participants with only 1 measurement were also deleted.

1.4. Normality

Fig. 3 demonstrates a well-documented phenomenon: terminal digit preference in BP measurements. Research indicates the terminal digit is 0 in 45% of systolic BP measurements and 48% of diastolic BP. Even numbers are more likely to occur than odd, with the most common odd digit being 5 [27]. Besides terminal digit preference the quantiles of the data appear to follow a normal distribution.

1.5. Correlation structure

We used a variogram to determine how time impacted upon systolic BP. For each person, we found all combinations of measurement times and the corresponding BP measurements. We computed $y_{ijk} = \frac{1}{2}BP_{ij} - BP_{ij}^2$ and $x_{ijk} = t \_i - t \_k$ where $t \_j$ and $BP_{ij}$ refer to time of measurement and BP measurement, respectively, for person $i$ at time $t$. For the first 4 years, we partitioned them into 90 day intervals, after 4 years, we increased the interval lengths to 180 days for smoothness; 4 years is also the duration of our study. Fig. 4 shows the mean of each partition.

The stability of Fig. 4 suggests no evidence of autocorrelation in this data. Fluctuations are likely due to variation among the data.

1.6. Random effects

We also used these data to identify if fixed or random effects in the gradient most accurately represented the measurements within people. We separated the data into participants above and below age 50 and randomly drew 10,000 samples of size 400 from both subpopulations. We fit 2 mixed effect models to both systolic and diastolic BP to all 20,000 samples. One model was $Y_{ij} = a + b t \_i + \epsilon$, the other $Y_{ij} = a + b t \_i + \epsilon + \delta$; the first model has a fixed trend, the second a random trend. We fit fixed and random trend effects for systolic and diastolic BP to the 23,786 measurements for people less than age 50, and the 79,312 above to find the population trend. The percent of confidence intervals for $β$ and $b_i$ which contained the population trend are found in Table 1. A random effect model is more likely to contain the true trend in its confidence interval, so was used for both trend and treatment effect in our third set of simulations.

2. Methods

We conducted three systematic simulation studies as outlined in Fig. 5.

The first set demonstrated that restricted maximum likelihood gave unbiased estimates of the carryover parameter using the lmer function in R [28,29]. The second set demonstrated the same using both systolic and diastolic BP. The third involved using the bootMer function to find confidence intervals for the risk ratio.

Our first simulations removed some of the complications by only simulating systolic BP. We started simply to explore bias in our estimates prior to proceeding. We began by randomly generating a number from a uniform distribution between 125 and 140 mm Hg; TROPHY trial participants had systolic BP close to the 140 mm Hg threshold [5]. Trends of 0, 1, and 2 mm Hg per year were used as systolic BP increases over time [30,31]. The treatment arm was given treatments of either –5 mm Hg or −10 mm Hg for either 1, 1.5, 2, 2.5, or 3 years [14]. Measurements were taken every 3 months, 6 months, or yearly, and carryovers of length 0, 0.5, 1, 1.5, and 2 years were included. BP varies due to both measurement error and intraindividual variability which we combined and assumed normally distributed standard deviations of 3, 5, and 7 mm Hg [14,17]. We analysed the data by fitting a model similar to Eq (2) using both lm and lmer in R; lm fits a linear model, while lmer fits a mixed effect model with random and fixed effects outlined in Fig. 5.

The second set of simulations involved both systolic and diastolic measurements. Although more extreme, the simulations were qualitatively similar and had the goal of exploring bias in estimates. A description of the systolic BP simulation is above. Baseline diastolic BP was randomly generated between 80 and 90 mm Hg and a quadratic
effect with age meant trends of −1, 0, and 1 mm Hg per year were used [32]. All combinations of systolic and diastolic trends were used in the simulation. To reflect day-to-day variation and measurement error, we added normally-distributed random error with standard deviation 3, 5, and 7 mm Hg for systolic BP which corresponded to 2, 3, and 4 mm Hg for diastolic BP [32]. Active treatment was assumed to lower systolic pressure by 5 or 10 mm Hg which corresponded to diastolic decreasing by 3 and 6 mm Hg, respectively [32]. We used the same lengths of study, durations of treatments, and carryover lengths as our first set.

We fit the mixed effects model $Y_{iktj} = a_{ik} + \beta_k t_{itj} + \gamma_k X_{itj} + \delta_k Z_{itj} + \epsilon_{iktj}$ using lm and lmer in R. Here, $Y_{iktj}$ is systolic BP when $k$ is 1, and diastolic BP otherwise. The lm estimates used a sum of squares maximisation, while lmer used restricted maximum likelihood [28, 29].

In the third set of simulations we generated data using random effects for both treatment effect and trend. When both systolic and diastolic were included, and diagnosis happened early in a simulated trial, the model fit by lmer was unidentifiable or barely identifiable. For that reason, we included on systolic BP. In addition, these simulations included bootstrapping the risk ratio at the time point after treatment ceased.

We generated baseline systolic measurements as above. Our simulated trend for each person was sampled from a random normal distribution with mean 0, 1 or 2 mm Hg per year and standard deviation 1.5 mm Hg, which our data indicated. Random treatment effects were found by sampling from a normal distribution with mean −5, −10 or 0 mm Hg and standard deviation 2.5 mg Hg [21]. We fit a mixed effects model with random effects for intercept, trend, and treatment effects. We used the measurement schedules and lengths of carryover outlined above.

\[
Y_{itj} = a_i + h_{itj} + c_{itj} X_{itj} + d_{itj} Z_{itj} + \epsilon_{itj} \quad (4)
\]

For each combination of parameters we simulated 400 trials, fit a mixed effects model using lmer, and used the bootMer function to simulate 100 parametric bootstraps of the risk ratio found in Eq (3) [28, 29]. A schematic of these simulations can be found in Fig. 9.

In a parametric bootstrap we assume a distribution based on the estimates obtained from fitting a model to the data, then sample from this distribution [33]. In our case, we fit a mixed effects model and sampled from the model to find bootstrap estimates of the risk ratio.

We explored differing definitions of hypertension that investigators could define in trial protocol and used them for all 3 sets of simulations. There were 6 rules, some of which involved averaging multiple measurements, and others that counted the number of measurements above the threshold. Our rules diagnosed hypertension when:

---

Fig. 6. Boxplots of mean bias in estimates of carryover parameter for the mixed effects model (left panel) and the linear model (right panel) from our primary simulations. Bias in the mixed effects model is consistently smaller.
3. Results and discussion

3.1. Primary simulation results

Our first simulation demonstrated that the lmer function in R gave negligibly biased estimates for the carryover parameter. Fig. 6 shows bias in estimates for carryover found using the linear model and the mixed effects model fit to 100 simulated trials. The relevant parameters are: treatment effect –10 mm Hg, carryover 2 years, and diagnosis defined by 3 measurements above the 140 mm Hg threshold. The horizontal axis separates lengths of treatment 1, 1.5, 2, 2.5, or 3 years and the bottom two graphs have standard deviation 3 mm Hg, the top 7 mm Hg. Trials with standard deviation 5 mm Hg were also simulated with results similar to those in Fig. 6.

The baseline BP for the simulated data came from a uniform distribution as described in Sec 2 but our model assumed a normal distribution for baseline BP. As we can see from Fig. 6 our model appears to be robust to the normality assumption at baseline.

3.2. Bivariate simulation

Fig. 7 shows the mean biases in our estimates of the carryover parameter from all 7016 combinations of parameters discussed in Sec 2. For each set of parameters, we conducted 100 bivariate simulated trials,
Fig. 9. Schematic overview of bootstrapping for relative risks. This method allows us to accurately test a carryover hypothesis. Adapted from [34].

Fig. 10. Proportion of bootstrap confidence intervals that contain 1 partitioned by length of carryover. When carryover is 0, 1 is the true value and is likely to be in the confidence intervals.
then found the mean of the 100 estimates of the carryover parameters. All the biases for feasible models for both systolic and diastolic measurements were computed and are included here.

The bivariate linear mixed model also appears robust to the normality assumption at baseline as the simulated data came from a uniform distribution.

### 3.3. Risk ratio

Simulations that used a parametric bootstrap to find a 95% confidence interval of the risk ratio for the time after treatment ceased are helpful in testing a carryover hypothesis.

Mathematically, we let \( r_{k_j}(f_j^*) \) be the \( k \)th simulated bootstrap risk ratio from the \( j \)th simulated trial at time \( t \). We define \( f_j^* \) as the mixed effect model found using lmer to maximise the likelihood from the \( j \)th simulated trial. For one combination of parameters in our simulations, \( l = 1, 2, \ldots, 400 \) and \( j = 1, 2, \ldots, M \). We found the 97.5% and 2.5% quantiles of \( r_{k_j}(f_j^*) \) where \( k = 1, 2, \ldots, 100 \) and \( t \) was the first time point after treatment ceased. We tested these confidence intervals for 1.

Intuitively, with no carryover in the model, equiprobable distribution in the treatment and control arms (risk ratio of 1) after treatment ceases seems predictable. As there is no extended impact of treatment, due to randomisation, BP in both arms is similar, and diagnosis rates should be similar. To support this we use data. The true risk ratio was found by using a sample size of 10\(^4\) and no error, all other parameters were the same as those discussed in Sec 2. As we have no error in the data, we diagnose hypertension when 1 measurement is above the threshold. We found the risk ratio of incidence hypertension at times when simulated carryover was present in the model (i.e. \( Z_k \neq 0 \) in Eq (4)). We plotted the true risk ratio at the first measurement after treatment ceases. Fig. 8 shows the stability around 1 when carryover is 0. With increasing magnitudes of carryover, no participants in the treatment arm are diagnosed which results in a risk ratio of 0. This complicates distinguishing magnitudes of carryover which requires further work.

We simulated 400 trials for each combination of parameters. For each trial we fit a mixed effect model and used the bootMer function in the lme4 package in R to find 100 parametric bootstraps of risk ratio at the first measurement after treatment ceased (the first \( t \) where \( Z_k \neq 0 \)). Using these, we computed a 95% confidence interval which we tested for 1. The coverage for the combination of parameters was the percentage of bootstrap confidence intervals that contained the value 1, Fig. 10. This coverage at the first measurement after treatment ceased has by length of carryover. If I is in the confidence interval our simulations suggest that it is 99% likely that the data came from a simulation that contained no carryover in the model. The process is outlined in Fig. 9.

### 4. Conclusion

To summarise, our simulations suggest that using a mixed effect model similar to (4) to estimate carryover parameters will give negligibly biased estimates in both systolic and diastolic models. Once a model is fit, a parametric bootstrap of the risk ratio can be found for the measurement directly after treatment ceases using the bootMer function. If I is in the bootstrap confidence interval it is highly likely there is no carryover in the model from which the data were simulated. This tests a carryover hypothesis. However, testing for duration of carryover will require more effort as insufficient participants are diagnosed in the treatment arm. Survival analysis may provide a viable solution in these situations.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### References

[1] Prevention & treatment. http://www.oxfordreference.com/view/10.1093
dictionaryauthority/0199343457/5. (Accessed 19 October 2017).
[2] S. Penn, Cross-over trials in statistics: the first 25 years. Stat. Med. 25 (2006) 3430–3442.
[3] S.G. Sturdevant, T. Lumley. Testing for carryover effects after cessation of intervention of metformin, N. Engl. J. Med. 344 (1) (2001) 3–10.
[4] W. Knowler, E. Barret-Conner, S. Fowler, R. Haman, L. Jlach, E. Walker, D. Nathan, Reduction in the incidence of type 2 diabetes: meta-analysis of Lifestyle intervention of metformin, E. Engl. J. Med. 346 (2002) 393–403.
[5] S. Julius, D.S. Nebbett, B.M. Egan, M.A. Weber, E.L. Michelson, N. Kaciroti, H. R. Black, R.H. Grimm, F.H. Messerli, S. Oparil, Feasibility of treating prehypertension with an angiotensin receptor blocker, N. Engl. J. Med. 354 (1) (2006) 1685–1697.
[6] C.M.M. Lawes, S.V. Hoon, M.R. Law, P. Elliott, S. MacMa- hon, A. Rodgers, High blood pressure, in: M. Ezzati, A.D. Lopez, A. Rodgers, C.J. Murray (Eds.), Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors, World Health Organization, 2004, pp. 281–389.
[7] M. Gibbald, M. Roland, Understanding controlled trials. Why are randomized controlled trials important? BMJ 316 (7126) (1998) 201.
[8] M.W. Miller-Craig, C.N. Bishop, E.B. Raferty, Circadian variation of blood-pressure, Lancet 311 (8068) (1978) 795–797.
[9] P.R. Woodhouse, K.-T. Khaw, M. Plummer, Seasonal variation of blood pressure and its relationship to ambient temperature in an elderly population, J. Hypertens. 11 (11) (1993) 1267–1274.
[10] G.D. James, I.S. Yee, T.G. Pickering, Winter–summer difference in the effects of environment, posture and place of measurement on blood pressure, Soc. Sci. Med. 31 (10) (1990) 1213–1217.
[11] L.A. Clark, L. Denby, D. Pregibon, G.A. Harshfield, T.G. Pickering, S. Blank, J. H. Laragh, A quantitative analysis of the effects of activity and time of day on the diurnal variations of blood pressure, J. Chron. Dis. 40 (7) (1987) 671–681.
[12] P. Lusardi, A. Zoppo, P. Preti, R.M. Pesce, E. Fizzaia, R. Fogari, Effects of insufficient sleep on blood pressure in hypertensive patients a 24-ha study. Am. J. Hypertens. 12 (1) (1999) 63–68.
[13] P.D. Neufeld, D.L. Johnson, Observer error in blood pressure mea- surement, Can. Med. Assoc. J. 135 (6) (1986) 633–637.
[14] T. Lumley, K.M. Rice, B.M. Psaty, Carryover effects after cessa- tion of drug treatment: trophies or dreams? Am. J. Hypertens. 21 (2008) 14–16.
[15] S.G. Sturdevant, T. Lumley, Testing for carryover effects after cessation of treatments: a design approach, BMC Med. Res. Methodol. 16 (1) (2016) 92.
[16] M. Nason, D. Pollman, Design and analysis of crossover trials for absorbing binary endpoints, Biometrics 66 (2010) 958–965.
[17] S.D. Persell, D.W. Baker, Studying interventions to prevent the progression from prehypertension to hypertension: does trophy win the prize? Am. J. Hypertens. 19 (11) (2006) 1095–1097.
[18] N. Mannel, W. Haa-nsel, Statistical aspects of the data analysis from retrospective studies of disease. J. Natl. Cancer Inst. 22 (4) (1959) 719–748.
[19] P.J. Goodman, I.M. Thompson, C.M. Tangen, J.J. Crowley, L.G. Ford, C. Obarznek, P.R. Conlin, E.R. M III, D.G. Simons- Morton, N. Karanja, P.-H. Lin, Effects on blood pressure of reduced dietary sodium and the dietary approaches to prevention atherosclerosis trials (DASH): final results of a multi-site randomized controlled trial. http://www.oxfordreference.com/view/10.1093
dictionaryauthority.20110803100344375 . (Accessed 19 October 2017).
[20] J. Wellingham, J. Tracey, H. Rea, B. Gribben, The development and its relationship to ambient temperature in an elderly population, J. Hypertens. 57 (2) (2011) 160–166.
[21] P.J. Bell, A. Hayen, P. Macaskill, J.C. Craig, B.C. Neal, K.M. Fox, W.J. Remme, F. Sturdevant and T. Lumley
[22] B. Sibbald, M. Roland, Understanding controlled trials. Why are randomized controlled trials important? BMJ 316 (7126) (1998) 201.
[23] G. Verbeke, G. Molenberghs, Linear Mixed Models for Longitudinal Data, Springer Series in Statistics, Springer, 2000.
[24] R. Black, R.H. Grimm, F.H. Messerli, S. Oparil, Feasibility of treating prehypertension with an angiotensin receptor blocker, N. Engl. J. Med. 354 (1) (2006) 1685–1697.
[25] P.J. Nietert, A.M. Wissell, C. Feifer, S.M. Ornstein, Effect of terminal digit dependency on blood pressure measurement and treatment in primary care, Am. J. Hypertens. 19 (2) (2006) 147–152, https://doi.org/10.1016/j. ajhype.2005.08.016.
[28] D. Bates, M. Maechler, B. Bolker, S. Walker, lme4: linear mixed-effects models using Eigen and S4, r package version 1.1-8. http://CRAN.R-project.org/package=lme4, 2015.

[29] D. Bates, M. Maechler, B.M. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, arXiv e-print; in press, J. Stat. Software 67 (1) (2015) 1–48, https://doi.org/10.18637/jss.v067.i01. http://arxiv.org/abs/1406.5823.

[30] I. Hajjar, T.A. Kotchen, Trends in prevalence, awareness, treatment, and control of hypertension in the United States, JAMA 290 (2003) 199–206.

[31] K. Wolf-Maier, R.S. Cooper, J.R. Banegas, S. Giampaoli, H.-W. Hense, M. Joffres, M. Kastarinen, N. Poulter, P. Primatesta, F. Rodriguez- Artalejo, B. Stegmayr, M. Thamm, J. Tuomilehto, D. Vancuzzo, F. Vescio, Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States, JAMA 289 (2003) 2363–2369.

[32] J.D. Wright, J.P. Hughes, Y. Ostchega, S.S. Yoon, T. Nwankwo, Mean systolic and diastolic blood pressure in adults aged 18 and over in the United States, 2001-2008, National Health Statistics Reports 35 (2011) 1–23.

[33] K.P. Murphy, Machine learning: a probabilistic perspective, in: Adaptive Computation and Machine Learning Series, MIT Press, 2012.

[34] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, 1994.