Testing for Carryover Effects After Cessation of Treatments: A Parallel Design Approach Does Not Work

S. Gwynn Sturdevant Thomas Lumley

July 1, 2013

Abstract

In interventions it may be important to determine whether the benefits extend beyond the active treatment period. This is clearly of interest for intensive lifestyle interventions, and there are also examples in the pharmaceutical literature. We consider estimation of carryover effects on time-to-event outcomes such as incident hypertension or incident diabetes. These are defined by a noisy measurement exceeding a diagnostic threshold, and diagnosis is followed by interventions that make subsequent measurements useless for treatment comparison.

We present the results of a systematic simulation study to determine the ability of a parallel-group trial design to detect carryover. None of the designs we examined had acceptable Type I error rate; most also had low power. When a treatment is effective during the intervention period, reliable testing for a carryover effect is difficult. Parallel-group designs do not appear to be a feasible approach.

1 Introduction

Hypertension and diabetes are responsible for significant mortality, morbidity, and cost in both developed and developing countries [6]. Rather than intervening after these high-risk conditions develop, it would be preferable to intervene to prevent incident hypertension and diabetes. In this paper we discuss trial design for evaluating interventions that prevent hypertension or diabetes, in particular the problem of estimating the duration of response to treatment. Put simply, how long does the effect continue after the active intervention ceases?

The question of carryover effects is obviously of interest for intensive, short-duration lifestyle interventions. For example, members of our department are currently taking part in a six-week fitness and nutrition program, which includes blood pressure reduction as a target. It is not feasible to prolong these interventions indefinitely, but they will not be useful unless they have long-term effects. The statistical issues in estimating carryover effects were in fact first studied in the context of a pharmacologic intervention. The Trial Of Preventing Hypertension [4, TROPHY] was conducted to determine if two years of treatment with an established antihypertensive drug, candesartan, in people with borderline high blood pressure (prehypertension) reduced the incidence of hypertension over the two years after treatment ceased. Analogous carryover analyses have also been carried out in diabetes prevention trials[5, 14]. These analyses have used naive comparisons of cumulative incidence at the end of the post-treatment follow up period. The TROPHY study design has been criticised on a number of grounds, including a simulation study that suggested it would have had a Type I error rate of close to 80%, far from the nominal 5% [7, 10, 13, 15].

Reliably estimating carryover effects requires that incident hypertension is diagnosed without differential bias by treatment group, and rapidly enough to distinguish zero, short, and long-term carryover. This paper is a systematic simulation study where we attempt to find more robust methods with which to test a carryover hypothesis by focussing on altering various parameters and analysing: Type I error rates and power. For concreteness, we describe the study designs in terms of systolic blood pressure and hypertension, but the results transfer to other incident events defined by thresholds in similar ways.

Section 1.1 presents the complexities in measuring blood pressure and their relation to measuring carryover. Section 2 describes the simulations conducted with a particular focus upon defining hypertension through differing rules. Section 3 attempts to correct TROPHY design by using these rules for diagnosis and adjusting inclusion criteria to decrease the likelihood of false positives and increase power. Lastly, in section 4 we summarize our results and suggest topics for further development.

1.1 Difficulties in Measuring Carryover

Blood pressure varies throughout the day, over the year, and with a range of outside influences and non-negligible measurement error[1, 3, 8, 11, 12, 17]. Diagnosing hypertension based on a single measurement will introduce unacceptable levels of noise, but averaging many measurements taken over a long interval makes it impossible to localise incident hypertension accurately in time. A practical study design must compromise and define hypertension in terms of a small number of measurements[9] taken at relatively infrequent intervals. A further complication is that, for ethical reasons, an individual who crosses the diagnostic threshold for hypertension must receive treatment that will change all future blood pressure measurements and make their subsequent data effectively useless.

Taking all of the above into account, the TROPHY [4] investigators designed a 4 year trial in which 809 subjects with systolic BP measurements between 130 – 139 mm Hg or diastolic BP 85 – 89 mm Hg were randomised to either treatment or placebo for 2 years. BP measurements were taken every 3 months, and diagnosis of hypertension occured when any 3 systolic or diastolic measurements were above the threshold of 140/90 mm Hg. Cumulative incidence in the placebo arm was 63.0%, and in the treatment arm 53.2%. 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. “

In this paper we consider a univariate measurement, systolic blood pressure, rather than the bivariate measurement (systolic, diastolic) used in TROPHY and in the prior simulation studies. Generalising to a bivariate measurement will make the design perform worse.

We model systolic blood pressure (Y_{it} for individual i at time t) as normally distributed around an individual-specific long-term average trend

$$Y_{it} = a_i + b_i t + c_i X_{it} + d_i Z_{it} + \epsilon_{it},$$

where Y_{it} is the BP measurement, $a_i \sim Unif(125, 140)$, $b_i \sim N(0, \Sigma)$, 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 carryover period. The reason for the uniform distribution on the individual random intercepts is that entry into the study is based on blood pressure thresholds.

Figure 1 shows possible models for the delay. As we can see from the graph, BP is lowered during treatment with medication. At the end of treatment BP returns either quickly to a normal trend, or more gradually, depending upon if carryover does exist and the length, in regards to our model the Z_{it} are altered.

Figure 2 illustrates the simulation process: the long-term blood pressure trend for each individual is simulated, then exchangeable measurement error is added, then measurements over a threshold, 140 mm Hg, are counted.

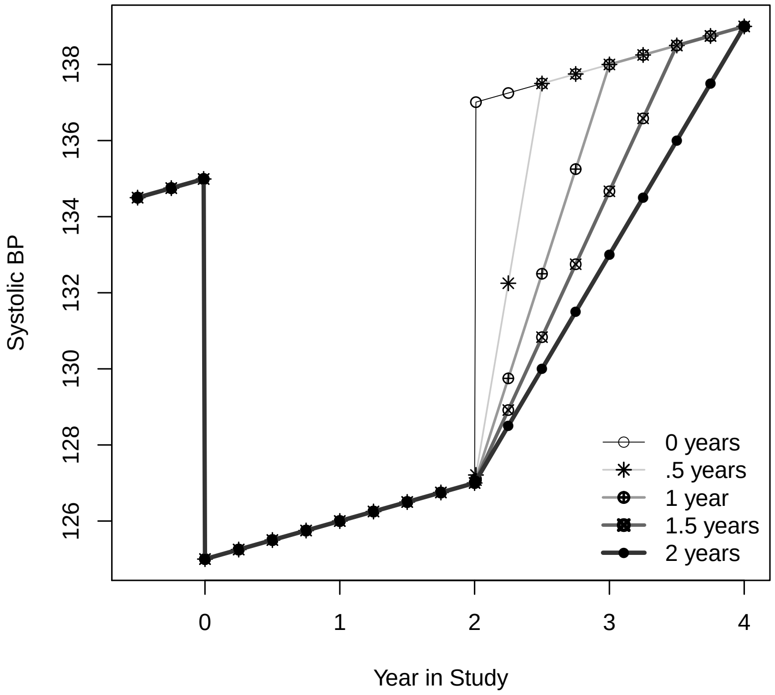

Figure 1: Systolic BP simulation with and without carryover. There are 5 different lengths of carryover: 0, 0.5, 1, 1.5, and 2 years.

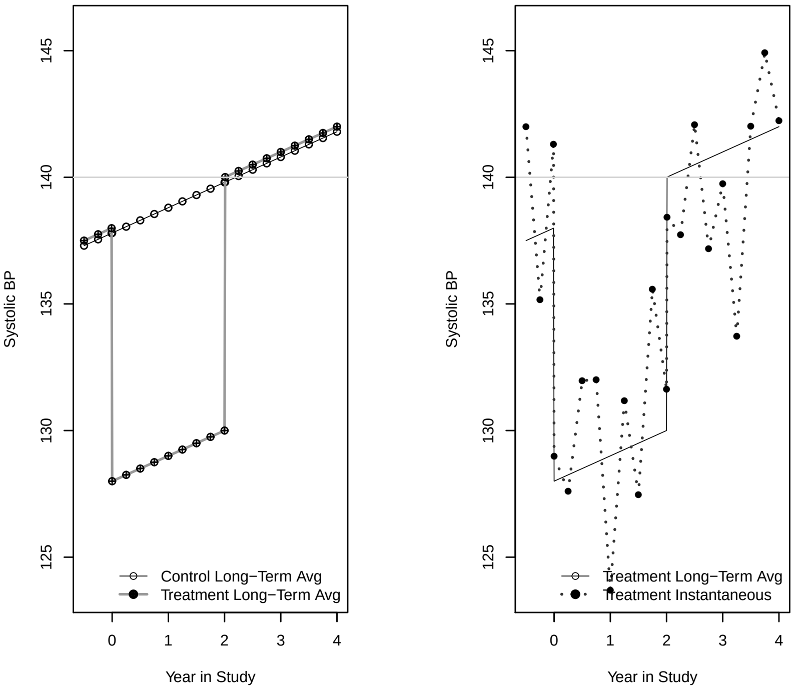

Figure 2: The graph on the left simulates systolic BP with no measurement error, the right, the treatment arm both with and without error.

2 Simulation design

To find tests to measure carryover, two sets of simulations were completed — one involving altering inclusion criteria, the other, altering other parameters — then analysis of false positives and power were done in each to determine the effectiveness of the design. Code for these simulations, and a complete display of results can be found at www.stat.auckland.ac.nz/∼sstu011/.

2.1 Simulation parameters

Our first simulation began by randomly generating baseline long-term-average systolic blood pressure in the range from a uniform distribution with range 125–140 mm Hg [4]. We then added an increasing trend with age [2, 16]; we assumed increases of 0, 1, or 2 mm Hg per year. To reflect day-to-day variation and measurement error, we added Normally-distributed random error with standard deviation 3, 5, and 7 mm Hg to account for this [7] . Active treatment was assumed to lower systolic pressure by 5 or 10 mm Hg [7], and control treatment to have no effect. The length of the study was assumed to be 4 [4] years with measurement times either every 3 months, 6 months, or yearly [7]. The duration of the treatment was either 1, 1.5, 2, 2.5, or 3 years [7]. Carryovers of lengths 0, 0.5, 1, 1.5, and 2 years were used [4].

Our second simulations looked at varying the inclusion criteria, by sampling from a uniform distribution with the baseline BP from 110–140, 120–140, or 130–140 mm Hg. We fixed the design at 2 years of treatments [4], 1 mm Hg per year trend in BP [7], and a standard deviation of 5 mm Hg [7]. The carryover duration was 0, .5, 1, 1.5, or 2 years [4], measurements 3 monthly, 6 monthly, or yearly [7].

2.2 Rules for Diagnosis

Since blood pressure is known to be variable, diagnosis is typically not made on a single measurement. The performance of the designs will depend on the diagnosis rule: using more measurements will lead to fewer errors, but to a longer delay in diagnosis.

We analysed five feasible criteria for diagnosing hypertension using a threshold of 140 mm Hg: if one measurement was above, if two consecutive measurements were above, if the average of two consecutive measurements were above, if three measurements were above, and if the average of three consecutive measurements were above. To illustrate the importance of measurement error we also considered a rule that diagnosed when both the instantaneous BP and the underlying long-term systolic BP were above

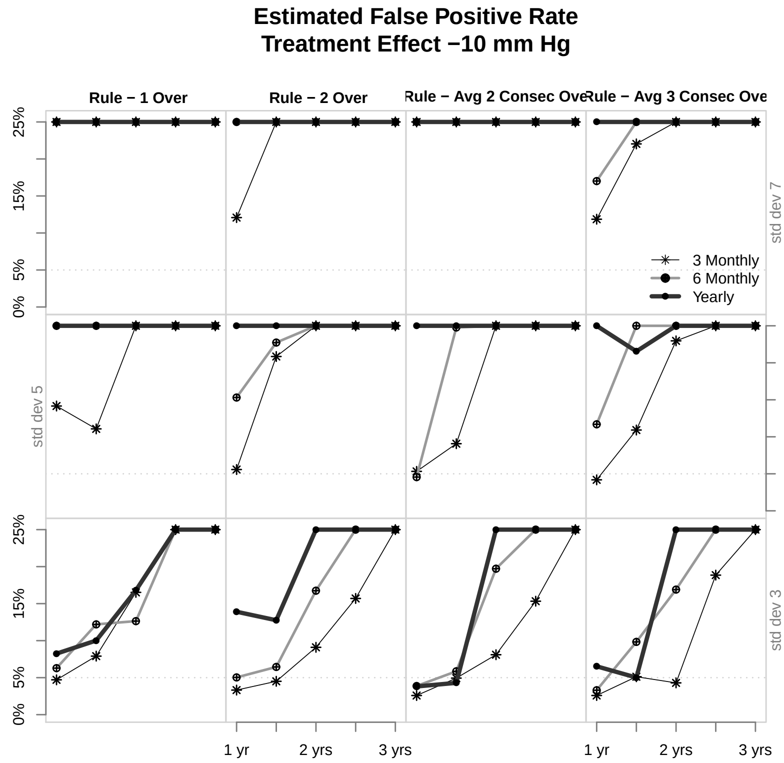

Figure 3: This graph shows estimated rates of false positives for the rules mentioned, most are far from the normal 5%.

threshold. This rule is not of practical use, although it could potentially be operationalized by averaging a large number of measurements over a period of days for anyone who had a single measurement over 140 mm Hg.

3 Results

3.1 False Positives

Figure 3 shows the rates of false positives across four rules studied, which are significantly higher than the accepted rate of 5%. The x axis of each graph tells us the length of time participants received treatment, with differing measurement standard deviations found in rows and columns signifying varying rules. The line types distinguish the frequency of measurements, as indicated in the key. All the results have trend of 1 mm Hg per year. Type I

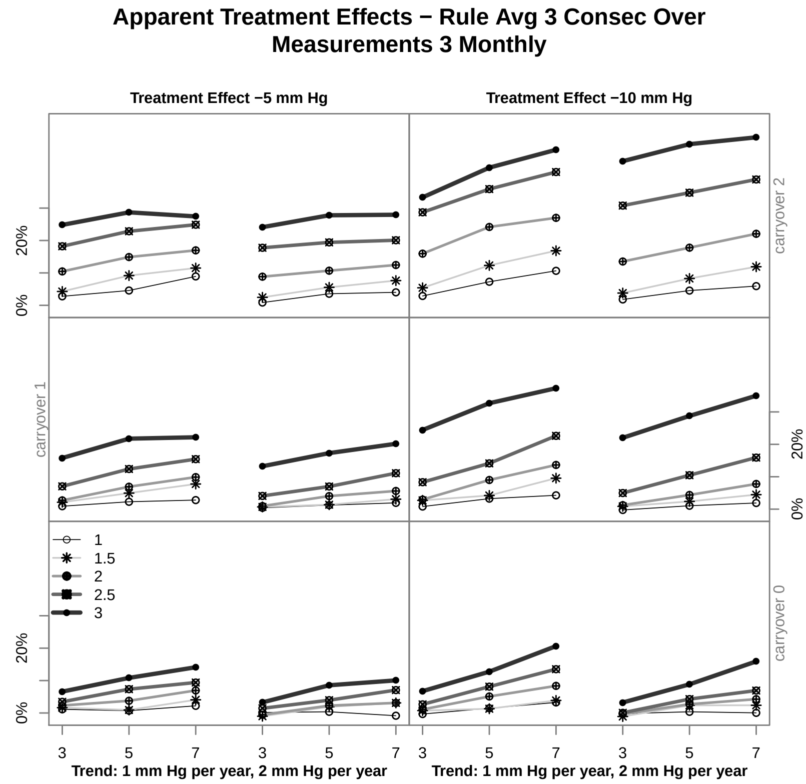

Figure 4: This graph shows the differences in cumulative incidences of diagnosis between treatment and control.

error rate is inflated except for the smallest measurement error and shortest period of active treatment.

Two rules are omitted from the graph. One, where people are diagnosed when 3 measurements are above the threshold, has been studied previously [10, 13, 15], The other is the infeasible rule that uses the true long-term-average blood pressure. This rule is the only one that does achieve close to nominal Type I error rate.

Figure 4 again demonstrates why testing for carryover using naive differences in cumulative incidence is ineffective. Each line type shows differing lengths of treatment, as indicated by the key. The graphs on the left side of the two columns include runs with trend 1 mm Hg per year, the others 2. Standard deviations are along the x axis. When carryover is 1 year, in the middle row, differences in cumulative incidence are substantial only when the duration of active treatment is long (2.5 or 3 years), the setting where the

Type I error inflation is largest.

Power at a nominal 5% Type I error rate typically did not greatly exceed the actual Type I error rate, so that power at honest test size appears to be poor. This is of secondary importance, as we are unable to control the rate of false positives and so the designs are not useable in practice.

3.2 Inclusion criteria

Measurement error can lead to false positive diagnosis only when true blood pressure is relatively close to the threshold, so varying inclusion criteria for baseline long-term-average blood pressure were considered. Figure 5 shows difference in cumulative incidence of diagnosis for three inclusion thresholds (110, 120, 130 mm Hg) in the presence and absence of carryover, for the same diagnosis rule in 4. Including participants with lower blood pressure reduces the estimated carryover effect under the null hypothesis, but also under the alternative hypothesis.

4 Discussion

Although carryover effects are potentially important, especially for intensive lifestyle interventions, they are difficult to assess reliably. In a wide range of parallel-group designs with data simulated we have shown that randomisation fails to preserve the Type I error rate even approximately.

The bias is smaller when the active treatment period is short and the followup is long (relative to the spacing of measurements), when the measurement error in systolic blood pressure is (unrealistically) small, and when the inclusion criteria are broad enough to allow participants who are far from the threshold, and thus not ethical to treat. A short active treatment period and broad inclusion criteria also reduce the estimated carryover effect when carryover is truly present, so they are not a solution to the problem. Rather than modifying the design so that carryover effects can be demonstrated by comparing cumulative incidence of diagnosis, it may be necessary to develop new statistical methodology to extract valid estimates from these designs.

The example of TROPHY shows that potential design bias for carryover effects can be hard to recognise, even for sophisticated researchers. We strongly recommend that simulation studies be performed to validate the planned design and analysis for any trial attempting to measure carryover effects.

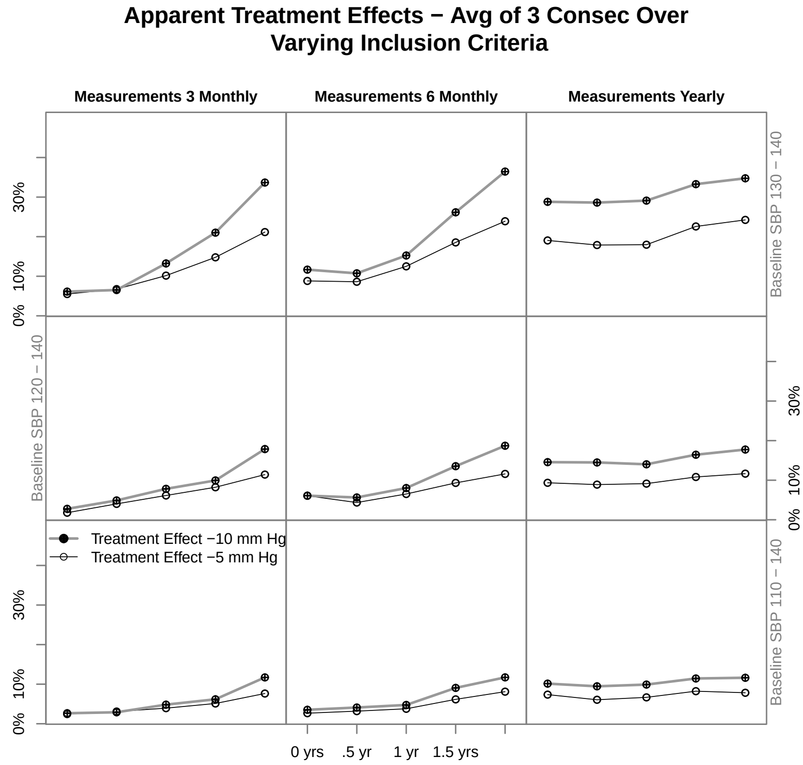

Figure 5: This graph demonstrates the impact of altering inclusion criteria upon differences in cumulative diagnosis rates. The x axis indicates the length of carryover.

References

[1] L. A. Clark, L. Denby, D. Pregibon, G. A. Harshfield, T. G. Pickering, S. Blank, and 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):671–81, 1987.

[2] Ihab Hajjar and Theodore A. Kotchen. Trends in prevalence, awareness, treatment, and control of hypertension in the united states. *JAMA*, 290:199–206, 2003.

[3] Gary D. James, Lily S. Yee, and Thomas G. Pickering. Winter-summer difference in the effects of emotion, posture and place of measurement on blood pressure. *Social Science Medical*, 31(11):1213–7, 1990.

[4] Stevo Julius, Shawna D. Nesbitt, Brent M. Egan, Michael A. Weber, Eric L. Michelson, Niko Kaciroti, Henry R. Black, Richard H. Grimm, Franz H. Messerli, and Suzanne Oparil. Feasibility of treating prehypertension with an angiotensin-receptor blocker. *New England Journal of Medicine*, 354(16):1685–97, 2006.

[5] WC Knowler, E Barrett-Connor, SE Fowler, RF Hamman, JM Lachin, EA Walker, and DM Nathan. Reduction in the incidence of type 2 diabetes with lifestyle intervention of metformin. *New England Journal of Medicine*, 6(346):393–403, 2002.

[6] Carlene M. M. Lawes, Stephen Vander Hoorn, Malcolm R. Law, Paul Elliot, Stephen MacMahon, and Anthony Rodgers. High blood pressure. In Majid Ezzati, Alan D. Lopez, Anthony Rodgers, and Christopher J.L. Murray, editors, *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, pages 281–389. World Health Organization, 2004.

[7] Thomas Lumley, Kenneth M. Rice, and Bruce M. Psaty. Carryover effects after cessation of drug treatment: Trophies or dreams? *American Journal of Hypertension*, 21:14–16, 2008.

[8] Paola Lusardi, Annalisa Zoppi, Paola Preti, Rosa Maria Pesce, Elena Piazza, and Roberto Fogari. Effects of insufficient sleep on blood pressure in hypertensive patients a 24-h study. *American Journal of Hypertension*, 12(1):63–8, 1999.

[9] T Marshall. Blood pressure measurement: The problem and its solution. *Journal of Human Hypertension*, 18:757–9, 2004.

[10] Jay I. Meltzer. A specialist in clinical hypertension critiques the trophy trial. *American Journal of Hypertension*, 19(11), 2006.

[11] Michael W. Millar-Craig, Charles N. Bishop, and E. B. Raferty. Circadian variation of blood-pressure. *The Lancet*, 311(8068):795–7, 1978.

[12] Philip D. Neufeld and David L. Johnson. Observer error in blood pressure measurement. *Canadian Medical Association Journal*, 135(6):633–7, 1986.

[13] Stephen D. Persell and David W. Baker. Studying interventions to prevent the progression from prehypertension to hypertension: Does trophy win the prize? *American Journal of Hypertension*, 19(11):1095–7, 2006.

[14] Frank M. Sacks, Laura P. Svetkey, William M. Vollmer, Lawrence J. Appel, George A. Bray, David Harsha, Eva Obarznek, Paul R. Conlin, Edgar R. Miller, III, Denise G. Simons-Morton, Njeri Karanja, and Pao-Hwa Lin. Effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension. *New England Journal of Medicine*, 344(1):3–10, 2001.

[15] Heribert Schunkert. Pharmacotherapy for prehypertension-mission accomplished? *New England Journal of Medicine*, 354:1742–4, 2006.

[16] Katharina Wolf-Maier, Richard S. Cooper, José R. Banegas, Simona Giampaoli, Hans-Werner Hense, Michel Joffres, Mika Kastarinen, Neil Poulter, Paola Primatesta, Fernando Rodríguez-Artalejo, Birgitta Stegmayr, Michael Thamm, Jaakko Tuomilehto, Diego Vanuzzo, and Fenicia Vescio. Hypertension prevalence and blood pressure levels in 6 european countries, canada, and the united states. *JAMA*, 289:2363–69, 2003.

[17] Peter R. Woodhouse, Kay-Tee Khaw, and Martyn Plummer. Seasonal variation of blood pressure and its relationship to ambient temperature in an elderly population. *Journal of Hypertension*, 11(11):1267–74, 1993.