How much effect of different antihypertensive medications on cardiovascular outcomes is attributable to their effects on blood pressure?


The debate over whether certain antihypertensive medications have benefits beyond what would be expected from their blood pressure lowering spurred the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial, which randomized 42,418 participants to chlorthalidone (15,255), amlodipine (9048), lisinopril (9054), or doxazosin (9061). We compared chlorthalidone, the active control, with each of the other three agents with respect to the primary outcome, fatal coronary heart disease or nonfatal myocardial infarction, and several other clinical endpoints. The arms were similar with respect to the primary endpoint, although some differences were found for other endpoints, most notably heart failure. Although the desire was to achieve similar blood pressure reductions in the four arms, we found some systolic blood pressure and diastolic blood pressure differences. A natural question is to what degree can observed treatment group differences in cardiovascular outcomes be attributed to these blood pressure differences. The purpose of this paper was to delineate the problems inherent in attempting to answer this question, and to present analyses intended to overcome these problems. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: blood pressure; measurement error; meta-analysis; regression calibration; regression dilution bias; time-varying covariate analysis

1. Introduction

It is well recognized that hypertension increases the risk of cardiovascular events and that antihypertensive treatment ameliorates this risk. What was not known until recently is whether different
antihypertensive medications confer similar cardiovascular benefits. This is an important issue not only for potential differences in health outcomes but also because of drug costs, both to the individual and to society. The cost of some newer antihypertensive drug classes exceeds the cost of generic thiazide diuretics. From the individual perspective, members of some patient groups experience disproportionate effects from drug costs. For example, elderly, Hispanic, or African American minorities are more vulnerable to the adverse clinical consequences of hypertension and are more likely to restrict antihypertensive use because of cost [1]. From the societal perspective, antihypertensive drug selection has enormous effects on healthcare expenditures. A shift from current antihypertensive regimens to diuretic-based regimens recommended in national guidelines has been estimated to reduce prescription costs by approximately $1.2bn per year in the USA [2].

With this background, the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) [3] was designed to compare three newer antihypertensive agents with a diuretic—the established drug in previous major hypertension trials with cardiovascular disease outcomes—to see whether the impact on cardiovascular events of lowering blood pressure (BP) depends on how it is lowered. The primary endpoint was fatal coronary heart disease (CHD) or nonfatal myocardial infarction (MI). Other important endpoints were total mortality, stroke, and heart failure (HF).

The ALLHAT results have been presented [4, 5]. No differences were found for the primary endpoint of fatal CHD or nonfatal MI, but the chlorthalidone arm had statistically significantly fewer cardiovascular events, chiefly HF. But chlorthalidone also lowered the mean systolic blood pressure (SBP) more than the other arms, 0.8 mmHg lower than amlodipine and 2.0 mmHg lower than lisinopril over 5 years of follow-up.

There are at least two major problems in trying to determine how much of the difference in events can be attributed to differences in BP:

Problem 1: Participants achieving a given level of BP in one treatment arm are not necessarily comparable with those achieving the same BP in another treatment arm.

Problem 2: BP measurement errors and transient BP changes cause regression dilution bias.

Problem 1 arises because a natural way to see if event differences are due to differential BP effects is to compare participants who achieve a given BP on two different drugs. This can be performed formally using a proportional hazards model with follow-up BP as a time-varying covariate [5] or less formally by simply computing relative risks within categories of follow-up BP. If the relative risks differ substantially from 1 for participants achieving similar follow-up BP, then we are tempted to conclude that drug effects are not mediated entirely through BP. But consider a hypothetical example in which drug A is so effective that everyone reaches goal BP, whereas only the healthiest 10% of participants on drug B achieve goal BP. Comparing those who reach goal BP in the two arms means comparing all participants in arm A with the healthiest 10% of participants in arm B. We might find fewer cardiovascular events among those reaching goal BP in arm B and erroneously conclude that there must be a non-BP beneficial effect of drug B. This problem is not entirely resolvable because the characteristics defining health level may be measured imprecisely or not at all.

We now elaborate on problem 2 by using stroke as a concrete example. If we could measure BP without error, then we might see a linear relationship between BP (horizontal axis) and the log odds of stroke (vertical axis), as shown in the hypothetical example in Figure 1(a). Measurement errors tend to ‘spread out’ the horizontal axis, as shown in Figure 1(b). The result of fitting a positively sloped regression line over a more spread out horizontal axis is that the slope of the line decreases. Imagine sticking a thumbtack loosely into the average horizontal and vertical points on the regression line, slightly rotating the line clockwise, and then pushing the thumbtack all the way in. That is essentially what measurement error does to the regression line. Suppose we could eliminate measurement error completely, and doing so resulted in the dotted line of Figure 2, which shows log odds of stroke in two different arms versus BP. Drug A produces lower BP and has fewer strokes than drug B, but the stroke difference is entirely explained by lower BP in arm A. Now add measurement error and transient BP differences. This is like sticking thumbtacks into each arm’s average horizontal and vertical points and rotating the lines, as shown by the solid lines in Figure 2. These lines now have different intercepts. In other words, one would now erroneously conclude that the difference in log odds of stroke is greater than that expected from the BP lowering (note that this is not an indictment of adjusting for baseline covariates measured with error; baseline covariates tend to have centers almost on top of each other, so regression dilution bias does not cause a problem [6]).
Figure 1. (a) A perfect world in which blood pressure (BP) measurement error and transient deviations have been eliminated and in which the relationship between the log odds of an event (say, stroke) and BP is perfectly linear. The large dark circle represents the ‘average’ horizontal and vertical points. Measurement error and transient deviations cause the horizontal positions of the points to change, as indicated by the horizontal arrows. (b) The result of adding the horizontal deviations caused by measurement error and transient deviations. The central point indicated by the large dark circle is essentially unchanged because errors in the two different directions cancel each other out. Because the heights stay the same but the points spread out horizontally, the regression line relating log odds to true BP (dotted line) is essentially rotated clockwise about its center, resulting in the less steep solid line.

Figure 2. The regression dilution bias when two arms achieve different blood pressures (BPs). The horizontal separation of the two large dark circles is the difference in average BP in arm A versus arm B. The vertical separation is the difference in log odds of events in arm A versus arm B. Suppose that the relationship between true BP and log odds of event is the same in both arms, indicated by the dotted line. The regression dilution bias causes a similar rotation about each arm’s center point. The resulting (solid) lines now have different intercepts, creating the illusion that the difference in events is not explained entirely by the BP difference.

Problem 2 has been addressed in both observational studies and clinical trials. One approach [7] in observational studies has been to first classify participants into subgroups determined by baseline BP, recognizing that measurement error and transient deviations cause the lowest and highest subgroups to be more extreme than they should be, as in Figure 1(b). A follow-up BP measurement is used to estimate the usual BP for participants in a given baseline BP subgroup. For example, if the baseline SBP subgroup is 140–149 mmHg and the follow-up average SBP for such participants is 142 mmHg, then participants in that subgroup are treated as having a usual SBP of 142 mmHg. One then plots event incidence against usual SBP. Usual SBP is much less prone than individual measurements to measurement error or transient deviations because it is an average over all participants in a given baseline SBP category.
The aforementioned approach is problematic in a clinical trial because the follow-up BP reflects drug effects rather than the usual BP of someone in a given baseline category. Instead, one could use observational studies or previous trials to compare the treatment effect observed in the current trial with that predicted from its average treatment minus control SBP [4, 5]. Again the key idea is that averaging SBP across many participants greatly reduces measurement error or variability due to transient fluctuations.

Meta-analysis of clinical trials provides another opportunity to reduce measurement error and transient fluctuations by averaging over many participants. Meta-regression analysis plots the treatment-to-control log odds ratios from different trials against their between-arm difference in average SBP. Weighted regression is then used to determine the effect on clinical events of a given reduction in BP [8]. Meta-regression was used to analyze results of the lipid-lowering component of the ALLHAT. A regression line relating treatment/control log odds ratios for mortality (and separately for CHD events) to cholesterol differences among trials other than the ALLHAT was plotted; the ALLHAT’s log odds ratio and cholesterol difference was superimposed to see if it seemed consistent with the line [9]. Meta-regression can be applied within the ALLHAT by treating collections of clinics as though they were separate trials.

There are advantages and disadvantages of applying meta-regression to clinics within the ALLHAT. In addition to dealing with problem 2, meta-regression circumvents problem 1. We no longer compare the nonrandom subset of participants achieving similar SBP reductions on different agents in a given trial; rather, we compare only the randomly assigned treatment groups in a given trial. The major disadvantage of meta-regression is that the unit of analysis is the clinic rather than the individual participants, effectively reducing sample size and power dramatically.

A method of dealing with problem 2 while maintaining the individual participant as the unit of analysis is regression calibration [10]. Each person is assumed to have some true BP at a given time, but measurement error and transient fluctuations prevent us from observing it. Regression calibration replaces someone’s observed value by the best estimate of his or her true value, given the observed value. That best estimate is a weighted combination, \( wY + (1 - w) \text{(average } Y \text{)}, \) of the person’s observed value \( Y \) and the average BP of all participants. The weight \( w \) is an estimate of the ratio of the variability across people to the total variability. The effect of regression calibration is to dampen extreme values by bringing them closer to the mean of all participants. This essentially reverses the horizontal dispersion seen in Figure 1.

The objective of this paper was to present exploratory analyses of the ALLHAT data to determine the degree to which observed treatment group differences in cardiovascular outcomes are attributable to BP differences.

2. Methods

The design of the ALLHAT has been presented [3]. Briefly, participants were randomized to amlodipine (9048), lisinopril (9054), doxazosin (9061), or the active control, chlorthalidone (15,255). The ratio of the sample sizes in the chlorthalidone arm and each other arm was approximately 1.7 to increase power for comparisons with chlorthalidone. As the desire was to achieve comparable BPs in the different arms, failure to reach the common BP goal of less than 140/90 prompted dose escalation up to a maximum dose of the initial drug, at which time open label therapy with atenolol, reserpine, clonidine, or hydralazine was added in any of the four arms if the participant still did not reach goal. Other drugs could be added for BP control at the clinician’s discretion, including low doses of the randomized drug classes. About 13% of chlorthalidone patients were also taking a calcium channel blocker or an ACE inhibitor, whereas 9% were taking a calcium channel blocker or ACE inhibitor alone. Similar rates of crossing over to other randomized drugs were seen in the other arms.

2.1. Participants

We present results for the ALLHAT participants randomized to chlorthalidone, amlodipine, or lisinopril. Doxazosin participants were excluded for two reasons. First, because doxazosin was terminated early, events in that arm should be compared with those of the chlorthalidone arm only up to the time of discontinuation of doxazosin. The fact that two comparisons would use all chlorthalidone events and the third would use only a subset would be confusing. Second, because the purpose of this paper was to illustrate the difficulty in trying to ascribe event differences to BP differences even in the best of circumstances, we felt it best to avoid the added complexity of including an arm that was stopped early.
2.2. Statistical methods

2.2.1. Comparability of participants with a given follow-up SBP. To identify baseline characteristics that differ between participants achieving a given BP on lisinopril and those achieving the same BP on chlorthalidone, we performed separate regressions or logistic regressions for each baseline characteristic, depending on whether that characteristic was continuous or binary. The baseline characteristic \( (Y) \) was the dependent variable, and the independent variables were year-1 SBP \( (x) \) and treatment \((z; z = 0 \) for chlorthalidone and \( z = 1 \) for lisinopril). Thus,

\[
\begin{align*}
E(Y) &= \beta_0 + \beta_1 x + \beta_2 z & \text{if } Y \text{ is continuous,} \\
P(Y = 1) &= \frac{\exp(\beta_0 + \beta_1 x + \beta_2 z)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 z)} & \text{if } Y \text{ is binary.}
\end{align*}
\]

For continuous baseline characteristics, the treatment coefficient \( \beta_2 \) is the between-arm difference in that baseline characteristic for participants achieving the same year-1 SBP. For example, if the baseline characteristic is age, then the treatment coefficient is the difference in baseline age between participants achieving a given SBP in the lisinopril arm and those achieving the same SBP in the chlorthalidone arm. If the baseline characteristic is binary, then the treatment coefficient may be interpreted as the logarithm of an odds ratio. For example, if \( Y \) is baseline gender, the treatment coefficient \( \beta_2 \) is the logarithm of the ratio of the odds of being male for participants achieving a given year-1 SBP in the lisinopril arm relative to the chlorthalidone arm. Similar analyses were conducted comparing amlodipine and chlorthalidone (treatment = 0 corresponds to chlorthalidone; treatment = 1 corresponds to amlodipine).

2.2.2. Correction for regression dilution 1: meta-regression. The ALLHAT clinics were first randomly ordered. For each clinic, the outcome having the fewest number of events was identified from among the ALLHAT outcomes: death, stroke, fatal CHD or nonfatal MI, and HF. The randomly ordered clinics were then combined by grouping successive clinics until the sum of the identified outcomes (i.e., the least occurring outcome for each respective clinic) equaled or exceeded 10. This guaranteed that a cluster had at least 10 events for any outcome. For each participant, the baseline SBP was subtracted from the year-1 SBP change was computed for each cluster. A Cox proportional hazards model was used to analyze each cluster to estimate the log hazard ratio for clinical events between a comparator and chlorthalidone, and its variance. Weighted regression was then used to relate the log hazard ratio (the dependent variable) \( Y_i \) and between-arm SBP difference \( x_i \) (the independent variable). The weights were inverses of the variances of the log hazard ratios.

\[
Y_i = \beta_0 + \beta_1 x_i + e_i, \quad Y_i = \text{clinic } i \text{ log hazard ratio, } x_i = \text{clinic } i \text{ SBP difference weights } w_i = 1/\text{var}(e_i).
\]

In this model, the intercept term \( \beta_0 \) indicates whether the treatment effect is entirely explained by the SBP difference; exponentiating the intercept yields the hazard ratio expected from equivalent BP lowering \((x = 0)\) in the two arms. Separate analyses were performed to compare lisinopril to chlorthalidone and amlodipine to chlorthalidone.

2.2.3. Correction for regression dilution 2: regression calibration. At any given time \( t \), the BP used is a weighted average of the patient’s observed BP (which is an average of two measurements) at time \( t \) and the average BP at time \( t \) of all patients still at risk. The weight assigned to the observed BP is the estimated ratio \( \sigma_{\text{true}}^2/\sigma_{\text{total}}^2 \), where \( \sigma_{\text{true}}^2 \) is the variance of the true BP of a given patient if we could eliminate measurement error and \( \sigma_{\text{total}}^2 \) is the total variability (includes variability of true BP and measurement error variability). Specifically, let \( \sigma_{\text{true}}^2 \) denote the within-patient variance of the two BP measurements at time \( t \), pooled across people. Let \( \sigma_{\text{total}}^2 \) be the sample variance of the per-patient averages of two measurements. Then we estimate \( \sigma_{\text{true}}^2 \) by \( \sigma_{\text{total}}^2 - (1/2)\sigma_{\text{true}}^2 \). If the within-patient variability is a small fraction of the total variability, then the weighted average is nearly the patient’s observed BP. On the other hand, if the within-patient variability is a large fraction of the total variability, then the weighted average is nearly the average BP across people rather than the individual’s observed BP.

Regression calibration was applied with proportional hazards regression in two ways. The first way applied the Cox model to events occurring after 6 months of follow-up, whereas the second way included all events and used SBP as a time-varying covariate. With either method, we used regression-calibrated...
SBP and treatment arm as covariates and regarded statistically significant treatment arm hazard ratios as evidence that the effect of treatment was not mediated entirely through SBP. The advantage of using month-6 SBP is that 6 months is a sufficiently short time that it is almost like a baseline covariate. An advantage of the time-varying covariate analysis is that it uses the most proximal and arguably the most relevant BP to determine current risk. A disadvantage of time-varying covariates is that the variable used to determine current risk is measured in people still at risk (i.e., being event free thus far), and the at-risk groups may become less comparable across arms over time.

3. Results

3.1. Comparability of participants with a given follow-up SBP

Tables I and II show results of regression and logistic regression analyses comparing baseline characteristics of amlodipine and chlorthalidone participants (Table I) and lisinopril and chlorthalidone participants (Table II) with the same year-1 SBP. Tables III and IV show similar information for the lowest and highest quintiles of year-1 SBP. For participants achieving the same year-1 SBP on amlodipine and chlorthalidone, amlodipine participants tend to have higher baseline creatinine but lower baseline SBP, diastolic blood pressure, glomerular filtration rate (GFR), and prevalence of CHD. Whether amlodipine or lisinopril participants with the same year-1 SBP had higher prevalence of high density lipoprotein (HDL) < 35 depended on the year-1 SBP. This is borne out by Tables III and IV, showing that for participants in the lowest quintile of year-1 SBP, a higher percentage of chlorthalidone than amlodipine participants had HDL < 35 (13.7% versus 12.2%), but for participants in the highest quintile of year-1 SBP, a lower percentage of chlorthalidone than amlodipine participants had HDL < 35 (8.3% versus 10.7%).

<table>
<thead>
<tr>
<th>Table I. Results of regression and logistic regression analyses comparing baseline characteristics of amlodipine and chlorthalidone participants after adjusting for year-1 SBP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline characteristic</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>SBP</td>
</tr>
<tr>
<td>DBP</td>
</tr>
<tr>
<td>Creatinine</td>
</tr>
<tr>
<td>GFR</td>
</tr>
<tr>
<td>Age 65 years or older</td>
</tr>
<tr>
<td>HDL cholesterol</td>
</tr>
<tr>
<td>Prior history of CHD</td>
</tr>
</tbody>
</table>

CHD, coronary heart disease; DBP, diastolic blood pressure, GFR, glomerular filtration rate; HDL, high density lipoprotein; SBP, systolic blood pressure.

<table>
<thead>
<tr>
<th>Table II. Results of regression and logistic regression analyses comparing baseline characteristics of lisinopril and chlorthalidone participants after adjusting for year-1 systolic blood pressure.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline characteristic</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>SBP</td>
</tr>
<tr>
<td>Height</td>
</tr>
<tr>
<td>Potassium</td>
</tr>
<tr>
<td>GFR</td>
</tr>
<tr>
<td>Triglycerides</td>
</tr>
<tr>
<td>Age 65 years or older</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Left ventricular hypertrophy on ECG</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Diabetes</td>
</tr>
<tr>
<td>Aspirin</td>
</tr>
</tbody>
</table>

ECG, electrocardiogram; GFR, glomerular filtration rate; SBP, systolic blood pressure.
### Table III. Baseline characteristics for participants in the lowest quintile of SBP at year-1.

<table>
<thead>
<tr>
<th></th>
<th>Chlorthalidone</th>
<th>Amlodipine</th>
<th>p-value*</th>
<th>Lisinopril</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 2819</td>
<td>N = 1297</td>
<td></td>
<td>N = 1482</td>
<td></td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>66.2 ± 7.5</td>
<td>66.7 ± 7.9</td>
<td>0.048</td>
<td>66.2 ± 7.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Black (%)</td>
<td>33.9</td>
<td>32.3</td>
<td></td>
<td>27.4</td>
<td></td>
</tr>
<tr>
<td>Women (%)</td>
<td>43.6</td>
<td>41.2</td>
<td></td>
<td>40.6</td>
<td></td>
</tr>
<tr>
<td>Years of education (mean ± SD)</td>
<td>11.1 ± 4.1 (n = 2660)</td>
<td>11.1 ± 4.1 (n = 1216)</td>
<td></td>
<td>11.2 ± 4.0 (n = 1409)</td>
<td></td>
</tr>
<tr>
<td>Eligibility risk factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cigarette smoker (%)</td>
<td>22.9</td>
<td>21.7</td>
<td></td>
<td>22.2</td>
<td></td>
</tr>
<tr>
<td>History of CHD (%)</td>
<td>28.1 (788/2800)</td>
<td>27.7 (357/1289)</td>
<td></td>
<td>29.7 (439/1476)</td>
<td></td>
</tr>
<tr>
<td>History of MI or stroke (%)</td>
<td>25.0</td>
<td>25.4</td>
<td></td>
<td>25.3</td>
<td></td>
</tr>
<tr>
<td>History of revascularization (%)</td>
<td>14.9</td>
<td>15.0</td>
<td></td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td>Type 2 diabetes (%)</td>
<td>32.4</td>
<td>32.5</td>
<td></td>
<td>30.4</td>
<td></td>
</tr>
<tr>
<td>HDL cholesterol &lt; 35 (%)</td>
<td>13.7</td>
<td>12.2</td>
<td></td>
<td>14.2</td>
<td></td>
</tr>
<tr>
<td>Left ventricular hypertrophy by ECG (%)</td>
<td>14.8</td>
<td>17.2</td>
<td>0.048</td>
<td>12.6</td>
<td>0.051</td>
</tr>
<tr>
<td>MI by ECG (%)</td>
<td>5.2 (128/2449)</td>
<td>7.0 (77/1094)</td>
<td>0.033</td>
<td>5.3 (68/1278)</td>
<td></td>
</tr>
<tr>
<td>Atrial fibrillation (%)</td>
<td>0.4 (9/2449)</td>
<td>0.9 (10/1094)</td>
<td>0.040</td>
<td>0.7 (9/1278)</td>
<td></td>
</tr>
</tbody>
</table>

CHD, coronary heart disease; ECG, electrocardiogram; HDL, high density lipoprotein; MI, myocardial infarction; SBP, systolic blood pressure; SD, standard deviation.

*The denominators for values in the table appear at the top of each column with exceptions noted. p-values are for comparisons with the chlorthalidone arm. If the p-value is absent, then its value was greater than 0.05.
### Table IV. Baseline characteristics of participants in the highest quintile of systolic blood pressure at year-1.

<table>
<thead>
<tr>
<th></th>
<th>Chlorthalidone</th>
<th>Amlodipine</th>
<th>$p$-value*</th>
<th>Lisinopril</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N = 2241$</td>
<td>$N = 1462$</td>
<td></td>
<td>$N = 1896$</td>
<td></td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>67.9 ± 7.9</td>
<td>67.5 ± 7.7</td>
<td></td>
<td>67.6 ± 7.8</td>
<td></td>
</tr>
<tr>
<td>Black (%)</td>
<td>39.8</td>
<td>42.4</td>
<td></td>
<td>44.4</td>
<td>0.003</td>
</tr>
<tr>
<td>Women (%)</td>
<td>51.4</td>
<td>49.4</td>
<td></td>
<td>49.9</td>
<td></td>
</tr>
<tr>
<td>Years of education (mean ± SD)</td>
<td>10.8 ± 4.0 ($n = 2080$)</td>
<td>10.9 ± 3.8 ($n = 1360$)</td>
<td></td>
<td>10.7 ± 3.9 ($n = 1752$)</td>
<td></td>
</tr>
<tr>
<td>Eligibility risk factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cigarette smoker (%)</td>
<td>20.2</td>
<td>20.2</td>
<td></td>
<td>21.7</td>
<td></td>
</tr>
<tr>
<td>History of CHD (%)</td>
<td>25.2 (561/2225)</td>
<td>22.1 (321/1454)</td>
<td></td>
<td>21.7 (407/1875)</td>
<td>0.008</td>
</tr>
<tr>
<td>History of MI or stroke (%)</td>
<td>23.7</td>
<td>23.3</td>
<td></td>
<td>21.6</td>
<td></td>
</tr>
<tr>
<td>History of revascularization (%)</td>
<td>13.6</td>
<td>12.0</td>
<td></td>
<td>11.9</td>
<td></td>
</tr>
<tr>
<td>Type 2 diabetes (%)</td>
<td>40.7</td>
<td>41.4</td>
<td></td>
<td>38.1</td>
<td></td>
</tr>
<tr>
<td>HDL cholesterol &lt; 35 (%)</td>
<td>8.3</td>
<td>10.7</td>
<td>0.017</td>
<td>9.9</td>
<td></td>
</tr>
<tr>
<td>Left ventricular hypertrophy by ECG (%)</td>
<td>18.8</td>
<td>19.2</td>
<td></td>
<td>20.7</td>
<td></td>
</tr>
<tr>
<td>MI by ECG (%)</td>
<td>4.6 (88/1926)</td>
<td>5.1 (63/1236)</td>
<td></td>
<td>5.5 (91/1646)</td>
<td></td>
</tr>
<tr>
<td>Atrial fibrillation (%)</td>
<td>0.4 (8/1926)</td>
<td>0.6 (7/1236)</td>
<td></td>
<td>0.6 (10/1646)</td>
<td></td>
</tr>
</tbody>
</table>

CHD, coronary heart disease; ECG, electrocardiogram; HDL, high density lipoprotein; MI, myocardial infarction; SBP, systolic blood pressure; SD, standard deviation.

*The denominators for values in the table appear at the top of each column with exceptions noted. $p$-values are for comparisons with the chlorthalidone arm. If the $p$-value is absent, then its value was greater than 0.05.
Similarly, Table II shows important differences between lisinopril and chlorthalidone participants after adjustment for year-1 SBP; lisinopril participants are more likely to be male, whereas chlorthalidone participants tend to have higher baseline SBP, height, potassium, GFR, triglycerides, and prevalence of diabetes.

These results show that there are important differences across arms in measured covariates among participants achieving the same follow-up BPs. It is likely that there are differences among unmeasured covariates as well, highlighting the difficulty of comparing outcomes among participants with similar follow-up BPs.

3.2. Meta-regressions

The top panel of Figure 3 shows a plot of the amlodipine to chlorthalidone hazard ratio (vertical axis, using a ‘log’ or ‘doubling’ scale) for the primary outcome (fatal CHD or nonfatal MI) for each cluster as a function of the cluster’s difference in mean follow-up BP between amlodipine and chlorthalidone (horizontal axis). Also shown is the weighted regression line relating the SBP difference to the log hazard ratio. The bottom panel of Figure 3 is a similar plot for the comparison of lisinopril to chlorthalidone. In both plots, the points cannot be distinguished from random scatter about the horizontal line corresponding to a hazard ratio of 1. In other words, there is no visible relationship between BP differences of different clusters and hazard ratios. This visual observation is confirmed by the meta-regression results for the primary outcome, which show a non-significant relationship between SBP difference and log hazard ratio for fatal CHD or nonfatal MI for the comparison of both amlodipine to chlorthalidone ($p = 0.61$) and lisinopril to chlorthalidone ($p = 0.95$).

The non-significant effect of SBP differences between arms on outcomes was seen for every outcome, highlighting the major problem with the meta-regression approach, namely a serious loss of power from using cluster rather than individual participant as the unit of analysis. Even within an arm there was very little relationship between achieved BP and Kaplan–Meier year-6 event rates.

There was one intriguing result for HF. Figure 4 reveals that although there is a non-significant relationship between SBP difference and the hazard ratio for HF, the intercept is statistically significant for the comparison of amlodipine to chlorthalidone ($p < 0.0001$; Figure 4(a)). This indicates that the amlodipine to chlorthalidone hazard ratio is greater than 1 even for the same degree of BP lowering; that is, the improved hazard ratio for chlorthalidone compared with amlodipine would have been seen

---

**Figure 3.** Relationship between year-1 systolic blood pressure (SBP) difference and hazard ratio for fatal coronary heart disease (CHD) or nonfatal myocardial infarction (MI) for 77 clusters. Each circle represents a cluster. Each cluster was formed by aggregating one or more ALLHAT clinics to ensure at least 10 events. The treatment group abbreviations are A = amlodipine, C = chlorthalidone, and L = lisinopril. The intercept and slope, with standard errors in parentheses, are A/C 0.063 (0.066) and −0.014 (0.026) and L/C 0.006 (0.074) and 0.001 (0.020).
even if the two agents reduced BP by the same amount. A similar phenomenon occurred for the comparison of lisinopril to chlorthalidone, but it did not reach statistical significance ($p = 0.09$).

To see if our results were sensitive to the number of clusters formed, we re-ran analyses after combining clusters to form 30 larger clusters. Results were similar in that the only significant results were for HF. Again the amlodipine to chlorthalidone hazard ratio for HF was greater than 1 even for no difference in BP lowering between the arms. For the comparison of lisinopril with chlorthalidone with respect to HF, the slope relating the hazard ratio to BP differences reached statistical significance ($p = 0.036$).

### 3.3. Regression calibration

Table V shows the results of the proportional hazards model with and without including uncalibrated or calibrated month-6 SBP.

For lisinopril versus chlorthalidone, inclusion of the month-6 SBP dampened treatment effects for most outcomes. For example, the 1.15 L/C hazard ratio for stroke without including month-6 SBP became 1.05 and 1.01 when month-6 SBP was included in an uncalibrated or calibrated way, respectively. Similarly, the $p$-value of 0.021 without including month-6 SBP became 0.47 and 0.86 when uncalibrated and calibrated month-6 SBPs were included, respectively, suggesting that chlorthalidone’s stroke advantage over lisinopril is entirely consistent with what would be expected from its SBP lowering advantage. The same was true of combined CVD, HF, and hospitalized/fatal HF. The treatment hazard ratios diminished, and none was significant when month-6 SBP was included. Interestingly, the opposite occurred for the primary outcome, CHD. The 0.99 L/C hazard ratio was nowhere near being statistically significant until the month-6 SBP was included, after which the L/C hazard ratios became 0.92 and 0.91 for uncalibrated and calibrated SBPs, respectively. This result was marginally statistically significant for uncalibrated SBP, suggesting that lisinopril may have a BP-independent benefit on CHD that offsets its inferior BP lowering.

For amlodipine versus chlorthalidone, HF and hospitalized/fatal HF were the only outcomes for which there was a statistically significant difference. Amlodipine’s disadvantage remained highly statistically significant whether or not adjustment was made for either calibrated or uncalibrated month-6 SBP, although the estimated hazard ratio was attenuated. This suggests that BP differences do not completely explain chlorthalidone’s advantage over amlodipine with respect to HF or hospitalized/fatal HF.

For lisinopril versus chlorthalidone, time-varying covariate analysis showed a similar attenuation pattern to that shown in Table V for stroke but no such consistent pattern for other outcomes (combined CVD and HF; Table VI). Table VI was consistent with Table V in showing an advantage of
**Table V. Results of proportional hazards regression applied to events after 6 months.**

<table>
<thead>
<tr>
<th>Month-6 SBP</th>
<th>Lisinopril versus chlorthalidone</th>
<th>Amlodipine versus chlorthalidone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
<td>95% CI</td>
</tr>
<tr>
<td>CHD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>0.99</td>
<td>(0.91, 1.08)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>0.92</td>
<td>(0.83, 1.01)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>0.91</td>
<td>(0.82, 1.00)</td>
</tr>
<tr>
<td>Stroke</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.15</td>
<td>(1.02, 1.30)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.05</td>
<td>(0.92, 1.20)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.01</td>
<td>(0.88, 1.16)</td>
</tr>
<tr>
<td>Combined CVD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.10</td>
<td>(1.05, 1.16)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.04</td>
<td>(0.98, 1.10)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.02</td>
<td>(0.97, 1.09)</td>
</tr>
<tr>
<td>HF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.20</td>
<td>(1.09, 1.34)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.08</td>
<td>(0.96, 1.21)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.05</td>
<td>(0.93, 1.18)</td>
</tr>
<tr>
<td>Hospitalized/fatal HF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.11</td>
<td>(0.99, 1.24)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.00</td>
<td>(0.88, 1.14)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>0.98</td>
<td>(0.86, 1.11)</td>
</tr>
</tbody>
</table>

The month-6 systolic blood pressure, either uncalibrated or calibrated, was used as a covariate (except in the rows labeled ‘not used’). CHD, coronary heart disease; CI, confidence interval; CVD, cardiovascular disease; HF, heart failure; HR, hazard ratio; SBP, systolic blood pressure.

**Table VI. Results of proportional hazards regression with systolic blood pressure as a time-varying covariate (except in the rows labeled ‘not used’).**

<table>
<thead>
<tr>
<th>SBP</th>
<th>Lisinopril versus chlorthalidone</th>
<th>Amlodipine versus chlorthalidone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
<td>95% CI</td>
</tr>
<tr>
<td>CHD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.00</td>
<td>(0.92, 1.10)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.01</td>
<td>(0.92, 1.11)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.01</td>
<td>(0.92, 1.11)</td>
</tr>
<tr>
<td>Stroke</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.17</td>
<td>(1.03, 1.32)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.07</td>
<td>(0.94, 1.22)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.07</td>
<td>(0.94, 1.22)</td>
</tr>
<tr>
<td>Combined CVD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.13</td>
<td>(1.07, 1.19)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.10</td>
<td>(1.04, 1.17)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.10</td>
<td>(1.04, 1.17)</td>
</tr>
<tr>
<td>HF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.23</td>
<td>(1.11, 1.37)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.26</td>
<td>(1.12, 1.41)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.25</td>
<td>(1.12, 1.40)</td>
</tr>
<tr>
<td>Hospitalized/fatal HF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not used</td>
<td>1.13</td>
<td>(1.00, 1.27)</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>1.16</td>
<td>(1.02, 1.32)</td>
</tr>
<tr>
<td>Calibrated</td>
<td>1.16</td>
<td>(1.02, 1.31)</td>
</tr>
</tbody>
</table>

CHD, coronary heart disease; CI, confidence interval; CVD, cardiovascular disease; HF, heart failure; HR, hazard ratio; SBP, systolic blood pressure.
chlorthalidone over lisinopril with respect to combined CVD and HF that remained even after adjustment for BP differences. Also consistent with Table V was that adjustment for SBP differences did not eliminate chlorthalidone’s advantage over amlodipine with respect to HF. Thus, the main difference in findings between Tables V and VI was that chlorthalidone’s overall advantage over lisinopril did not diminish when follow-up SBP was included in Table VI.

Interestingly, the effect of regression calibration itself was minimal in both the month-6 SBP and time-varying covariate analyses. In most cases, there was little difference whether SBP was calibrated or uncalibrated because the great majority of variability in BP was estimated to be from true person-to-person differences rather than measurement error.

4. Discussion

Determining the extent to which between-arm differences in event rates are explained by BP differences is a difficult one. One problem is that if one drug lowers BP more than another, then participants achieving a given follow-up BP on one drug may differ in important ways from those achieving the same BP on another drug. We observed such differences in the ALLHAT. For example, participants achieving a given year-1 SBP on chlorthalidone had higher baseline SBP than those achieving the same year-1 SBP on amlodipine or lisinopril. Thus, at the very least, we must adjust for baseline differences in any analyses comparing participants achieving a given SBP in different arms. Of course one can only adjust for measured baseline covariates.

One common approach in clinical trials is to include follow-up BP as a time-varying covariate in a Cox model that includes other baseline covariates. The problem with this approach is that measurement error and transient BP changes can make it appear that between-arm outcome differences are not entirely explained by BP differences when, in fact, they are (Figure 2).

Investigators of the Valsartan Antihypertensive Long-term Use Evaluation (VALUE) trial [11] hypothesized that, for the same BP control, the angiotensin-receptor blocker valsartan would reduce cardiac events more than amlodipine. BP in the amlodipine arm was 2.1 mmHg lower than in the valsartan arm after 6 months. Faced with the same issue of how to adjust endpoint results for differential achieved BP, VALUE investigators adopted a serial median matching approach [12]. The valsartan participant with month-6 SBP closest to the median of all 7543 valsartan participants was matched with an amlodipine participant with similar SBP (within 2 mmHg), age, sex, and prior history of CHD, stroke, and diabetes. From the remaining 7542 valsartan participants, the one closest to the median of the 7542 SBPs was selected and again matched to an amlodipine participant as described earlier. This process continued until no more matching amlodipine participants could be found. The resulting 5006 matching pairs were then analyzed. Although the method has the advantage of making fewer assumptions than using a model with follow-up BP as a time-varying covariate, we feel it is unattractive for several reasons, primary among them being that it discarded about one-third of all VALUE participants, and in a nonrandom and possibly biased way. Because amlodipine lowered BP more than valsartan, it is likely that no amlodipine participants would match the very highest BPs in the valsartan arm. Similarly the very lowest amlodipine BPs would probably be too low to match any valsartan participants. Furthermore, comparability of valsartan and amlodipine participants in the pairs is assured only on the selected covariates.

An alternative approach for a trial like the ALLHAT with hundreds of clinics is the meta-regression approach. It tends to achieve balance across measured and unmeasured covariates because it compares all randomized participants. It also ameliorates the measurement variability problem because the SBP difference is the average across all participants in the cluster, and averaging diminishes measurement error. The down side is that treating clusters as the unit of analysis reduces power. Our strategy was to examine the slope and intercept of the regression line relating the log hazard ratio for different clinical events to BP differences at the clinics. The intercept determines whether the effect on the clinical event is explained entirely by BP differences. We found only one outcome for which the effect appeared not to be entirely explained by BP, and that outcome was HF for the comparison of amlodipine with chlorthalidone. The estimated amlodipine to chlorthalidone hazard ratio for the same degree of BP lowering was greater than 1, indicating a lower risk of HF for chlorthalidone even if the BP reduction had been the same.

Regression calibration analysis is arguably the best approach because it maintains the individual participant as the unit of analysis and thereby preserves power. This analysis confirmed the apparent BP-independent benefit of chlorthalidone compared with amlodipine with respect to HF, both overall and for hospitalized/fatal cases. For other outcomes, there was no compelling evidence for differential non-BP effects of different drugs on outcome.
The problems we outlined here have been considered in the context of verifying other surrogate outcomes such as ejection fraction or cholesterol for heart disease [13] or CD4 count and CD4 percent in HIV [14, 15]. Prentice [16] proposed some of the earliest criteria for a valid surrogate outcome, namely that a surrogate should be predictive of outcome and that it should capture the entire effect of a treatment. This latter condition specifies that the conditional distribution of time to a ‘true’ clinical outcome like death, given treatment and the surrogate, should depend only on the value of the surrogate. Under these and one additional technical condition, a test of treatment effect on the surrogate is a valid test of treatment effect on the true clinical outcome. Freedman et al. [17] suggested an early measure of the proportion of treatment effect explained (PTE) by a surrogate, and Lin et al. [15] extended their work to survival settings. The PTE approach has received some criticism [18], and we elected to use a somewhat less formal method of describing the change in the treatment effect regression coefficient when BP was included in the model.

The ALLHAT was not powered to detect whether differences in cardiovascular endpoints are entirely explained by BP differences, especially considering that the BP differences of 0.8–2 mmHg would be expected to account for only a 3%–8% difference in stroke [7, 19] and a 2%–5% difference in HF [20, 21]. Clearly, hypothesis testing cannot tell the entire story. That is why we also described the size of change in the regression coefficient when the regression-calibrated surrogate was included versus excluded from the model, although we recognize that the estimated variability of this change is large.

We hope that we have conveyed the difficulty in trying to assess whether the effect of treatments on clinical events is explained entirely by BP differences. It is important to take measurement error and transient fluctuations into account and to analyze the data in different ways to check for consistency. The only BP-independent effects that were consistently observed using different analyses were those of amlodipine versus chlorthalidone with respect to HF.

5. Funding/support

The ALLHAT study was supported by contract NO1-HC-35130 with the National Heart, Lung, and Blood Institute (NHLBI). The ALLHAT investigators received contributions of study medications supplied by Pfizer (amlodipine besylate and doxazosin mesylate), AstraZeneca (atenolol and lisinopril), and Bristol-Myers Squibb (pravastatin), and financial support provided by Pfizer Inc.

6. Financial disclosures

Michael H. Alderman has received research support from Sankyo. Samuel S. Blumenthal has held a financial interest in Abbott Laboratories, Bristol-Myers Squibb, Merck, Pfizer, Roche, and Wyeth and has received honoraria from Boehringer Ingelheim. William C. Cushman has consulted for Calpis, Daichi Sankyo, Gilead Colorado, Novartis, Noven, Pharmacepuica, Sanofi Aventis, Sciele, Takeda, Theravance; has received honoraria from Bristol-Myer Squibb, Forest Pharmaceuticals, King; and has received research support from GlaxoSmithKline, King, Novartis, and Merck. Barry R. Davis has consulted for Amgen, Forest Pharmaceuticals, and Takeda. Richard A. Preston has consulted for Gilead Colorado and has received research support from Abbott Laboratories, Bayer Corporation, Forest Laboratories, Merck, Novartis, Schering Plough, Shionogi, and Takeda. Richard S. Castaldo, Jeffrey A. Cutler, Charles E. Ford, Stanley S. Franklin, Curt D. Furberg, David Gordon, James F. Graumlich, Valory Pavlik, and Michael Proschan have no financial interests to report.

References


