CASE 9

Stopping the Randomized Aldactone Evaluation Study Early for Efficacy

Janet Wittes
Jean-Pierre Boissel
Curt D. Furberg
Desmond Julian
Henri Kulbertus
Stuart Pocock

ABSTRACT

The Randomized Aldactone Evaluation Study (RALES) was a randomized double-blind placebo-controlled trial designed to test the hypothesis that addition of daily spironolactone to standard therapy would reduce the risk of all-cause mortality in patients with severe heart failure as a result of systolic left ventricular dysfunction. The Data Safety Monitoring Board (DSMB) for RALES reviewed data on safety and efficacy throughout the trial using pre-specified statistical stopping boundaries for efficacy. To ensure that the data were complete, the DSMB requested successive "mortality sweeps." At the time of these sweeps, all RALES investigators determined the vital status of participants at their clinics. Therefore, the data that the DSMB saw included a much higher percentage of the deaths than would have been observed without these sweeps. At the DSMB's fifth meeting, the data showed 351 deaths in the placebo group and 269 in the spironolactone group for an estimated hazard ratio of 0.78 ($p = 0.00018$). The board recommended early termination of the trial because the observed Z-value of 3.75 exceeded the pre-specified critical value of 2.79 and the data on mortality showed consistency among subgroups and across time. The sweeps had identified 31 deaths that likely would not have been reported by the time of the meeting. Subsequent data collection identified an additional 46 deaths that had occurred by the time the study ended. Even when the endpoint of a randomized clinical trial is mortality, routine methods of data collection and reporting are unlikely to identify all events in a timely manner. The experience from RALES provides an example of the importance of active follow-

up of patients to ensure that a DSMB is observing a high proportion of the events that have actually occurred.

INTRODUCTION AND BACKGROUND

In "heart failure," the heart is incapable of maintaining cardiac output adequate to accommodate metabolic requirements and venous return. The heart fails either because it is subjected to an overwhelming pressure or volume overload, because myocardial contractility is depressed (e.g., in myocardiopathy or intoxication), or because a significant loss of contractile tissue has occurred (e.g, after a myocardial infarction).[1] The condition can lead to a rise of pressure in the return veins, both on the systemic and the pulmonary sides. The resulting engorgement of pulmonary veins and capillaries can cause dyspnea, a difficulty with breathing, which is the most common symptom of heart failure. Heart failure also involves a fall of cardiac output, which can cause fatigue and activate the sympathetic nervous system with, consequently, an increase in heart rate and vasoconstriction of arteries and veins.

The fall in cardiac output and the increase in sympathetic drive lead to reduced effective renal blood flow. Through the renin–angiotensin system, this reduced flow induces a rise in the levels of angiotensin II, a vasoconstrictor, which stimulates aldosterone secretion by the cortex of the adrenal gland. Aldosterone is a hormone that, by its action on the distal renal tubule, promotes retention of sodium and accompanying water, while increasing potassium excretion. Consequently, blood volume increases, leading to the potential development of peripheral edema and pulmonary congestion. In addition to its renal action, aldosterone exerts a large number of potentially deleterious effects on the cardiovascular system. The New York Heart Association categorizes patients with heart failure into four classes depending on the severity of their symptoms, principally, dyspnea:[2]

Class I	patients withstand normal physical activity without symptoms;
Class II	patients develop symptoms on moderate or severe exertion only;
Class III	symptoms are present even on mild exertion;
Class IV	symptoms are present at rest.

A relationship between functional capacity and survival in heart failure is well established. In the early 1990s, studies showed the annual mortality rate of Class IV patients to be above 50% while the annual mortality rate in Class III patients varied between 10% and 45%.[3]

Until the mid-1980s, treatment was not evidence-based. Because fluid retention is the hallmark of heart failure, diuretics were the principal agents

used for its treatment. Digitalis, a positive inotrope that boosts cardiac contraction, was also commonly prescribed. Vasodilators, in particular, nitrates, prazosin, and ACE inhibitors, were recently introduced with a view to unload the heart, thereby improving cardiac function. In 1986 and 1987, the first trials to demonstrate a benefit of vasodilator therapy on mortality were published.[4,5]

On the basis of the then understood physiopathology of heart failure, a logical approach to treatment would have been to add a drug that blocks aldosterone receptors. At that time, however, physicians were reluctant to prescribe aldactone, an aldosterone inhibitor, to patients with heart failure because of the potential for serious elevations in potassium levels (hyperkalemia) among those receiving an ACE inhibitor, a class of agents that had quickly become one of the mainstays of treatment. Addressing this potential problem, a study published in 1996 showed that treatment with a low dose of spironolactone, an aldosterone-receptor blocker, in conjunction with standard dose of an ACE inhibitor, a loop diuretic, and digoxin was well tolerated and did not lead to serious hyperkalemia.[6]

PROTOCOL DESIGN

The establishment of the safety of low-dose spironolactone in patients with heart failure led to the design of the double-blind Randomized Aldactone Evaluation Study (RALES), a trial that aimed "to test the hypothesis that daily treatment with 25 mg of spironolactone would significantly reduce the risk of death from all causes among patients who had severe heart failure as a result of systolic left ventricular dysfunction and who were receiving [the then] standard therapy, including an ACE inhibitor, if tolerated."[7]

RALES took place in 195 centers from 15 countries. Sponsored by Searle, the manufacturer of spironolactone, the study had an academic executive committee chaired by Bertram Pitt, M.D., and an independent Data Safety Monitoring Board (DSMB) chaired by Desmond Julian, M.D. Collectively, the DSMB had expertise in cardiology, epidemiology, biostatistics, and clinical trials. Spironolactone had been in use since 1960, so its adverse event profile was well known. The most common adverse events are gynecomastia and other feminizing symptoms in males. As described above, the most serious expected adverse event associated with spironolactone is hyperkalemia. The role of the DSMB was to monitor safety, especially with respect to the potential for hyperkalemia, and to assess whether to recommend stopping the study early for efficacy.

The DSMB was originally blind to treatment code. Several of the members argued for unblinding the groups immediately, but given that the opinion was not unanimous the reports to the DSMB were designed with the treat-

ment groups for most variables labeled as A and B. Because increased rates of gynecomastia and hyperkalemia would unmask the A and B assignments, these two adverse events were labeled X and Y. The DSMB reserved the right to unblind itself should it feel the need.

DATA MONITORING EXPERIENCE

During the trial, Searle provided data to Statistics Collaborative, which prepared reports to the DSMB. The board had a predefined statistical guideline for stopping for efficacy. The guideline specified that early in the trial, stopping for efficacy would require very strong evidence favoring spironolactone. As the trial progressed, the standard for efficacy would become less stringent. Overall, the probability of declaring benefit if spironolactone and placebo had identical effects on mortality was 0.025. Technically, the guidelines were based on an O'Brien–Fleming boundary[8] for efficacy at a two-sided α-level of 0.05. Because the standard O'Brien–Fleming boundary requires looking at the data at equal increments of numbers of deaths and there was no practical way to schedule the meetings to ensure equal numbers of deaths at each meeting, the Lan–DeMets use function[9] was employed. This function allows flexibility in planning meetings without sacrifice of the stringency of the type I error rate. Figure 1 shows the boundaries used.

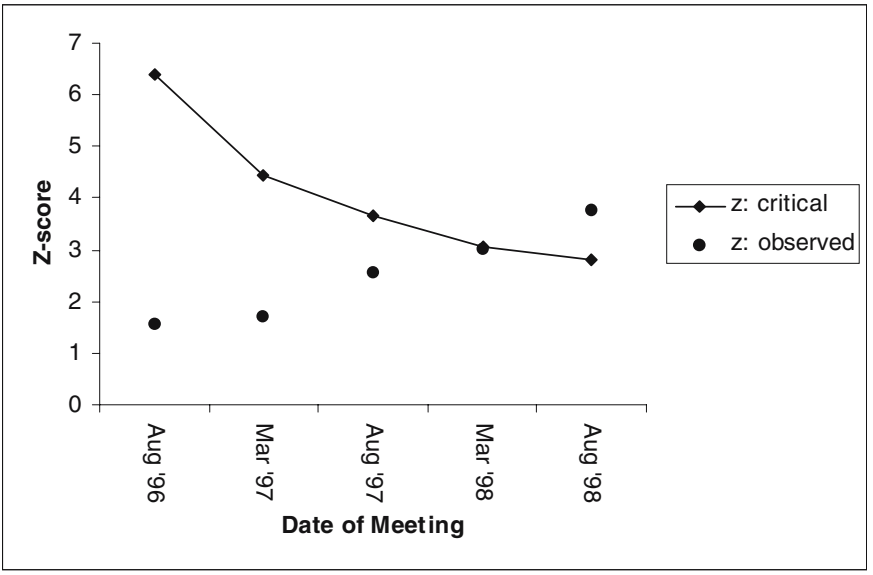

Figure 1 Monitoring boundaries and observed Z-values at the five interim analyses.

The board did not specify a boundary for safety; instead it relied on its collective judgment to recommend early termination if spironolactone showed a net adverse effect.

The original protocol specified an event-driven trial. Investigators would randomize patients 1 : 1 to spironolactone or placebo and stop recruitment at a pre-specified number of events. A trial with this design is called an "information time" trial because the design specifies the number of deaths, defined as the "total information." At each look, the DSMB would calculate the "information time" as the fraction of deaths that had occurred thus far relative to the total planned deaths.

The first patient was randomized on March 24, 1995. A protocol amendment, approved in early 1996, changed the planned end of the trial to December 31, 1999. Thus, the trial was now based on calendar time instead of total events. Consequently, the calculations for the interim analysis had to be based on an unknown total number of deaths.

Each DSMB meeting began with an open session for the investigators and the sponsor to report about the status of the trial. At the closed session, attended only by the DSMB and the statisticians reporting to it, the DSMB reviewed the data.

The Emerging Data

On August 24, 1996, at the DSMB's first meeting with an interim analysis, a difference in mortality between the two groups emerged, with 70 deaths in one group and 52 in the other (see Table 1). The Z-value was far from statistically significant ($z = 1.58$; critical $z = 6.38$, nominal p-value = 0.11 Note: the "critical Z" is the predetermined boundary that must be exceeded for the drug to be deemed effective.); nonetheless, the board expressed the view that such a large difference in the direction of increased mortality in the spironolactone group would lead to concern about safety. Consequently, the DSMB unblinded itself and learned that the lower event rate was occurring in the treated group. The board recommended continuing the trial with no change in protocol.

Recruitment ended as planned on December 31, 1996. At that time, a total of 1,663 patients had been randomized, 841 to receive placebo and 822 to receive spironolactone.

At the second meeting, which took place on March 17, 1997, the reported deaths were now 136 and 109 in the placebo and spironolactone groups, respectively, for a hazard ratio of 0.83 ($Z = 1.69$; critical $Z = 4.43$; nominal p-value = 0.092). Again, the board noted the reduction in mortality; however, in light of the non-statistically significant finding, it again recommended continuing the trial with no change in protocol.

Table 1 Observed and Projected Number of Deaths and Summary Statistics at Each Interim Analysis

Look number	Meeting date	Observed Deaths		Hazard ratio	Estimated information time	Z-value		Observed two sided p-value
		Placebo	Spironolactone			Critical	Observed	
Interim analyses with the sweeps as they occurred								
1	24-Aug-96	70	52	0.76	0.12	6.38	1.58	0.11
2	17-Mar-97	136	109	0.83	0.24	4.43	1.69	0.092
3	25-Aug-97	224	175	0.80	0.34	3.67	2.55	0.011
4	30-Mar-98	304	241	0.81	0.48	3.04	3.02	0.0026
5	24-Aug-98	351	269	0.78	0.57	2.79	3.75	0.00018
Estimated interim analyses that would have occurred without the sweeps								
4a	30-Mar-98	281	222	0.81	0.45	3.16	2.93	0.0034
5a	24-Aug-98	333	256	0.79	0.55	2.81	3.59	0.00034
Interim analysis cutoffs that would have occurred had the true numbers and times of deaths been known								
1b	24-Aug-96	81	59	0.75	0.14	5.88	1.80	0.072
2b	17-Mar-97	189	140	0.76	0.26	4.24	2.75	0.0060
3b	25-Aug-97	257	201	0.80	0.38	3.46	2.82	0.0048
4b	30-Mar-98	330	254	0.79	0.51	2.95	3.56	0.00038
5b*	24-Aug-98	383	283	0.76	0.60	2.72	4.46	0.000008

* These are the data in the paper describing the final results.[8]

At the time of the third meeting on August 25, 1997, data were still strongly favoring spironolactone, with 224 deaths in the placebo group and 175 in the spironolactone group for a hazard ratio of 0.80. Although the p-value was now nominally statistically significant ($p = 0.011$), the observed Z-statistic of 2.55 was quite far from the critical value of 3.67 defined by the O'Brien–Fleming boundary. At that meeting the board prepared itself for a crossing of the boundary. Given the strong trends observed thus far and the consistent patterns emerging over subgroups of interest, the board predicted that the data would cross the pre-specified stopping boundaries before the planned end of the study. It was, however, somewhat uncertain about the reliability of the data. According to the protocol, investigators were to report deaths within 24 hours of occurrence; because the interval between protocol-specified visits was every three months through one year of study follow-up and every six months thereafter, the DSMB suspected that information about deaths might be delayed. The board believed it highly likely that the number of deaths was being undercounted. If the probability of late reporting of deaths were equal in the placebo and spironolactone groups, this delay would reduce the power of the statistical tests at the interim analyses. More seriously, if deaths in the placebo group were reported with more, or less, alacrity than deaths in the spironolactone groups, the apparent effect size might be either over- or underestimated. While the double-blind nature of the study should afford considerable protection against differential reporting; nonetheless, if adverse events or better functioning were leading one group to have more frequent contact with the study staff than the other group, a bias in the reporting of events, including deaths, could occur.

The board was concerned lest it make a decision at one of the next meetings to recommend stopping the trial only to learn later, when all the deaths were reported, that the observed effect size was incorrect. To prevent crossing the statistical boundary with uncertainty remaining about the number of unreported deaths, the board requested that each investigator provide a census, or a "sweep," of vital status as of December 31, 1997. To avoid alerting the sponsor and the investigators of the reason for its request, the board worded its request in terms of the need for a "standard two-year" accounting of data. Anticipating a crossing of the boundary for efficacy, it also requested that at each meeting of the DSMB, the sponsor and the Principal Investigator routinely remain available for another open session at the end of the closed session.

The request for a sweep required considerable effort on the part of the sponsor and the investigators. Each investigator had to contact every participant, a task that was somewhat daunting, partly because it was unexpected.

After the March 1998 meeting, where the boundary was almost crossed (observed $Z = 3.02$; critical $Z = 3.04$), the board requested another "sweep"

just prior to its subsequent meeting. It also discussed the data it wanted to see at the next meeting with the view toward assuring that, should it recommend stopping, it would have considered all reasonable likely criticisms of an early stop. It discussed writing a press release, methods of informing the investigators of the early stop, and approaches to early publication of the results.

Finally, at the fifth meeting in August 1998, the observed Z-value was 3.75 while the critical Z-value required to cross the boundary was 2.79. The board's planning at its previous meeting allowed it to proceed deliberately at this last meeting. Although the data had crossed the boundary, the DSMB carefully considered the totality of the evidence available to it in deciding whether to recommend stopping the trial for efficacy. In particular, it reviewed effects in subgroups of interest; it considered the strength and internal consistency of the secondary endpoints; and it assessed the likelihood that the data would be reversed when the complete information became available. Given the consistency of the results and the strong effect on mortality, the board recommended early termination. Because it had requested that the sponsor and the Principal Investigator be present after the closed session, the board was able to report the data to them immediately. The board, the sponsor, and the Principal Investigator drafted a letter to the Steering Committee and a press release describing the data.

Ending the Study

The study ended smoothly because, having anticipated that the study would stop early, the DSMB set in motion actions to facilitate the process. The sweeps had identified a sizable increase in the number of deaths reported at the fourth and fifth interim analysis. While no can one know how many deaths would have been reported had the sweeps not occurred, the statistical group performed computer simulations to assess the likely effect of the sweeps. The simulations showed about an 8% increase in the number of reported deaths at each of the fourth and fifth meetings.[10] When several months later all the data were complete, another 46 deaths were identified. These deaths strengthened the inference so that the Z-statistic changed from 3.75 (for a p-value of 0.00018) at the DSMB meeting to 4.46 ($p = 0.000008$) when all the data had been collected. The estimated hazard ratio was 0.78 when the DSMB recommended stopping the study; the final estimate was 0.76.

LESSONS LEARNED

Several lessons emerged from the RALES trial. First, blinding in a study of this type is difficult. Even if one believes that a DSMB should be blind to

treatment (which most of the authors of this chapter do not), the actual process of blinding is cumbersome. The nature of the adverse events are often so clear that blinding requires complicated efforts on the part of the statistical center. Moreover, this process clouds the ability of the DSMB to weigh the risks and benefits of therapy.

Another lesson related to ascertainment of the endpoint during a study. In trials of mortality, one might assume that because ascertainment of the primary endpoint–death–is simple, the accruing data should be complete. RALES provides an example where this assumption does not hold. Ideally, studies should devise methods to ensure a very short delay between the occurrence of an event and its reliable documentation in the dataset. One such method is performing periodic sweeps assessing the primary endpoint for each person. Such a process, while cumbersome, can be essential to decision-making. RALES showed some evidence of differential reporting in the two groups. In the placebo group, 32 of the total of 383 deaths, or 8.4%, were reported after the last sweep; the comparable numbers for the spironolactone group were 14 of 283, or 4.9% ($p = 0.09$). Differential reporting is likely greater in unblinded studies.

The choice of how to monitor the study–by information time or calendar time–may seem statistically arcane, but in RALES we had to confront the choice explicitly because the study changed from one based on information time (800 deaths) to one based on calendar time. Even though the study was based on calendar time, we chose to monitor it on the basis of information time because in a long-term follow-up study, monitoring by information time is more statistically efficient. We, of course, did not know the number of deaths that would have occurred if the study were to continue until its planned end. Therefore, at each meeting of the DSMB, the statistical group calculated the expected total number of deaths projected from the observed survival patterns thus far. To ensure that the decision to stop early was insensitive to the estimated total, the statisticians provided a range of information fractions consistent with the data thus far and reported the boundaries for this range. Had the DSMB used calendar time instead, the boundary would have been crossed at the meeting of March 1998 (data not shown).

Finally RALES confirmed the importance of careful planning and of frequent communication with the study sponsor and the Principal Investigator. The DSMB's foresight enhanced its ability to recommend stopping the trial early and to make clear conclusions. Data from trials rarely leap over the monitoring boundaries; instead, a DSMB usually has highly suggestive evidence several meetings before the boundary is crossed. Positioning itself to make an orderly decision helps the credibility of a study. Furthermore, the availability of the sponsor and the Principal Investigator at the DSMB's meetings helped foster mutual understanding of the roles of everyone involved.

ACKNOWLEDGMENTS

The authors of this chapter consist of the members of the DSMB for RALES and the statistician (J.W.) who presented the data to the DSMB. Dr. Bertram Pitt was the Principal Investigator; Dr. Alfonso Perez and Dr. Barbara Roniker were the clinical monitors at Searle.

REFERENCES

1. Julian DG, Cowan JC, McLenachan JM. 1998. *Cardiology,* 7th Edition. WB Saunders, London.
2. Criteria Committee of the New York Heart Association: Diseases of the Heart and Blood Vessels (Nomenclature and Criteria for Diagnosis). 1964. Little, Brown, Boston.
3. Gradman AH, Deedwania PC. 1994. Predictors of mortality in patients with heart failure. *Cardiol Clin* 12:25–35.
4. Cohn JN, Archibald DG, Ziesche S, Franciosa JA, Harston WE, Tristani FE, Dunkman WB, Jacobs W, Francis GS, Flohr KH, et al. 1986. Effect of vasodilator therapy on mortality in chronic congestive heart failure. Results of a Veterans Administration Cooperative Study. *N Engl J Med.* 314:1547–1552.
5. The CONSENSUS Trial Study Group. 1987. Effects of enalapril on mortality in severe congestive heart failure. Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS). *N Engl J Med.* 316:1429–1435.
6. The RALES Investigators. 1996. Effectiveness of spironolactone added to an angiotensin-converting enzyme inhibitor and a loop diuretic for severe chronic congestive heart failure (the Randomized Aldactone Evaluation Study [RALES]). *Am J Cardiol* 78:902–907.
7. Pitt B, Zannad F, Remme WJ, Cody R, Castaigne A, Perez A, Palensky J, Wittes J for The Randomized Aldactone Evaluation Study Investigators. 1999. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *N Engl J Med.* 341:709–717.
8. O'Brien P, Fleming T. 1979. A multiple testing procedure for clinical trials. *Biometrics* 35:549–556.
9. Lan K, DeMets D. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 14:1927–1931.
10. Wittes J, Palensky J, Asner D, Julian D, Boissel J, Furberg C, Kulbertus H, Pocock S, Roniker B. 2001. Experience collecting interim data on mortality: an example from the RALES study. *Curr Control Trials Cardiovasc Med* 2:59–62.