EDITORIAL COMMENTARY

Target hemoglobin trials in chronic kidney disease: design and interpretation issues

Robert N. Foley

Abstract Optimal management of anemia in patients with chronic kidney disease remains a divisive issue within the nephrology community. Because the evidence provided by successive randomized controlled trials has often proven to be incongruent, it is natural to consider whether methodological issues may be responsible. Using four large trials [US Normal Hematocrit, Canadian European Normalization of Hemoglobin, Cardiovascular Risk Reduction by Early Anemia Treatment with Epoetin Beta (CREATE) and Correction of Hemoglobin and Outcomes in Renal Insufficiency (CHOIR)], this review article highlights several methodological issues that may be important when trial evidence is translated into clinical practice. Issues discussed include heterogeneity of enrollment criteria, failure to conceal treatment allocation, generalizability of study interventions, systematic use of imbalanced co-interventions [especially dose of erythropoietin stimulating agent (ESA), confusion regarding stopping rules and interim analyses and failure to account for imbalances in important patient characteristics generated at randomization.

Keywords Hemoglobin · Target · Trial · Methodology

R. N. Foley (✉)
Chronic Disease Research Group,
914 South 8th Street, Suite D-253,
Minneapolis, MN 55404, USA
e-mail: RFoley@cdrg.org

R. N. Foley
University of Minnesota,
Minneapolis, MN, USA

Introduction

Optimal management of anemia in patients with chronic kidney disease remains a controversial and polarizing issue. This article focuses on design and methodological issues and is not meant to be a systematic review or meta-analysis of all available trials, one of which was published recently. Four large clinical trials are critiqued [US Normal Hematocrit [1] (USNH), Canadian European Normalization of Hemoglobin (CENH) [2], Cardiovascular Risk Reduction by Early Anemia Treatment with Epoetin Beta [3] (CREATE) and Correction of Hemoglobin and Outcomes in Renal Insufficiency [4] (CHOIR)], each performed in a relatively specific group of anemic patients with chronic kidney disease (CKD) and all including sample sizes exceeding 500. The reader is referred to the primary articles for detailed descriptions of each study. While anemia management clearly differs in pediatric and adult CKD patients [5, 6], large-scale trials have yet to be undertaken in pediatric CKD. This being said, trials from adult populations may help to inform current practice and the design of future trials in pediatric CKD.

Idiosyncrasies of target trials

Treatment target trials differ profoundly from double-blind, placebo-controlled trials, and these differences often impede trial interpretation. For example, target allocations are not usually concealed, a strategy that may help investigators to avail themselves fully of the therapeutic armamentarium to achieve treatment targets. Even trials that try to conceal target allocations tend towards being unmasked over time, as, all things being equal, patients requiring more intensive

efforts for an unknown target to be reached are more likely to have been assigned to the intensive treatment arm.

Non-randomly assigned treatments (co-interventions), that may themselves influence the primary outcome, are often used in the time interval between randomization and primary outcome assessment. Highly imbalanced co-interventions can cloud the mechanisms underlying beneficial or harmful effects in primary treatment comparisons. The extent of co-intervention in hemoglobin target trials is extreme, as one sets out with the intention of using different amounts of erythropoietin stimulating agents (ESAs), intravenous injections of iron, anti-hypertensive agents and transfusions.

Another type of trial which systematically introduces non-random elements involves immediate or delayed intervention. In the delayed intervention arm, treatment is predicated on a non-random element (time) that is not controlled by the study design. If one truly wishes to know the effects of hemoglobin level and duration of anemia, a factorial design would be required, such as one where patients are randomly assigned to one of four groups: maintain current hemoglobin level throughout; maintain current hemoglobin level, switch to higher hemoglobin level after a pre-specified time; immediate increase to higher hemoglobin level, maintain this level throughout; immediate increase to higher hemoglobin level, switch to lower hemoglobin level after the pre-specified time. In practice, such a design would be extremely difficult to implement.

Trial participants

Generalizability of trials is largely a matter of judgment, and numerical methods are not available to guide this process. It is very useful to know the number of potentially eligible subjects, the number approached and the number recruited. Unfortunately, very few trials report these items. Table 1 details the enrollment criteria used in the four studies under consideration. It is evident that no two trials

Table 1 Enrollment criteria (*GFR* glomerular filtration rate, *E* exclusion criterion, *BP* blood pressure, *RRT* renal replacement therapy, *I* inclusion criterion, *TSAT* transferrin saturation)

Parameter	CHOIR [4]	CREATE [3]	USNH [1]	CENH [2]
Hemoglobin levels (g/dl)	<11.0	11.0 to 12.5	9.0 to 11.0	8.0 to 12.0
Previous epoetin	E	E	–	–
Estimated GFR (ml/min per 1.73 m^2 body surface area)	15.0 to 50.0	15.0 to 35.0	On hemodialysis	On hemodialysis for 3–18 months
Uncontrolled hypertension	E	BP≤170/95	Diastolic BP< 100 mmHg	Diastolic BP < 100 mm Hg
Active gastrointestinal bleeding	E	–	–	–
Iron overload	E	–	–	–
Frequent transfusions	E	–	–	E
Iron deficiency anemia	No 'refractory' iron deficiency	Ferritin ≥ 50 ng/ml	TSAT ≥20%	No 'uncorrected' iron deficiency
Cancer	E	–	–	E
Unstable angina	E	–	–	–
RRT expected ≤ 6 months	–	E	–	–
Valvular disease	–	E	E	–
Congestive heart failure	–	E	I	E
Myocardial infarction ≥ 3 months	–	E	I	E
Angiographic coronary artery disease	–	–	I	E
Major coronary event < 3 months	–	–	E	E
Stroke	–	I	–	–
Left ventricular volume index < 90 ml/m^2	–	–	――	I
C-reactive protein	–	≤15 mg/l	–	–
Pericardial disease	–	–	I	I
Cardiac amyloidosis	–	–	I	–
Androgen therapy	–	–	I	–
Seizure within previous year	–	–	–	I
Pregnant/lactating	–	–	–	I
Therapy with cytotoxic agent	–	–	–	I
Therapy with ≥ 10 mg/day prednisone	–	–	–	I

have identical enrollment criteria, and efforts to combine studies based on the premise of homogeneity of enrolled subjects may be suspect. Some notable differences between studies include: overt vs incipient anemia; overt vs absent cardiac disease; end-stage renal disease vs non-end-stage renal disease.

Interventions

Table 2 illustrates hemoglobin targets in the four trials under consideration and the therapeutic strategies used to meet these targets. Only CENH incorporated concealment of treatment targets. While concealing treatment allocation from patients is intuitively important for quality of life outcomes, concealment from patients, healthcare professionals and outcome assessors may be important for several other reasons. Investigator blinding may be important for rating 'hard' clinical events, even when these are subsequently adjudicated by blinded clinical event committees. For example, basic physiological constructs suggest that profound anemia and polycythemia predispose to congestive heart failure and vascular thrombosis, respectively. Knowing the assigned target hemoglobin level may influence site investigators when confronted with classically difficult diagnostic challenges, such as differentiating extracellular fluid volume expansion from congestive heart failure and non-specific chest pain from angina pectoris. Unfortunately, blinded event committees have no control over what is written in case records at the site level by unblinded investigators and cannot correct site-level biases.

Erythropoiesis stimulating agents (ESAs) are known to predispose to hypertension [7], and active surveillance of hemoglobin levels, ESA dose and blood pressure levels is desirable in clinical trials of different hemoglobin targets. Of the trials examined in this article, only CENH incorporated real-time, centrally-controlled, monitoring procedures for hemoglobin, iron, and blood pressure levels, with weekly or bi-weekly treatment recommendations transmitted to the study sites within days of measurement of these parameters.

Table 2 shows the different strategies used to achieve target hemoglobin levels. Once again, it is clear that the degree of heterogeneity between studies is large. Although CREATE and CHOIR both examined patients with CKD that did not require dialysis, the former used a strategy of delayed intervention in one treatment arm, while the latter used one of immediate intervention; in addition, the treatment strategy in the latter trial is impossible to emulate, as the epoetin dosing strategy was only described for the first 3 weeks of the study. In this regard, it is debatable whether many clinicians would begin treating epoetin-naïve

Table 2 Interventions

| Parameter	CHOIR [4]	CREATE [3]	USNH [1]	CENH [2]
Blinding	No	No	No	Patients
Real-time central monitoring	No	No	No	Yes
Hemoglobin targets (g/dl)				
Low	10.5 to 11.0	No epoetin until < 10.5	10.0 ± 1	9.5 to 11.5
High	13.0 to 13.5	Immediate epoetin 13.0 to 15.0	14.0 ± 1	13.5 to 14.5
Treatment strategies	High and low target groups received 10,000 units epoetin alfa once weekly for 3 weeks. Thereafter, a 'prespecified' dosing algorithm was used, not reported in primary publication	Dose reviewed every 4 weeks. If hemoglobin increment < 0.5 g/dl, ↑ by 25–50%; if hemoglobin increment > 0.5 g/dl, ↓ by 25–50%	With high target, dose initially ↑ 1.5-fold on study entry; thereafter if 2-week Δ hematocrit < 2%– ↑ by 25%; if >2% ↓ 25 U/kg. With low target, adjust by 10–25 U/kg	Epoetin-naïve patients in the higher target level initially given 150 IU/kg per week. For hemoglobin deviating from target, Δ 25% or 25 IU/kg
Primary outcome	Time to the first occurrence of death, myocardial infarction, hospitalization for congestive heart failure (excluding renal replacement therapy), or stroke	Time to first occurrence of sudden death, myocardial infarction, acute heart failure, stroke, transient ischemic attack, angina pectoris resulting in hospitalization for 24 h or more or prolongation of hospitalization, complication of peripheral vascular disease, or cardiac arrhythmia resulting in hospitalization for 24 h or more	Time to death or a first non-fatal myocardial infarction	Echocardiographic left ventricular volume index

patients with chronic kidney disease with 10,000 units of epoetin per week, especially when the intent was to maintain current hemoglobin levels.

Analytical plans

With regard to primary outcomes, no two studies were alike (Table 2). Some of the implications of composite endpoints (used in three of these studies) may be worth exploring. In the CHOIR study, for example, the primary end point was the time to first occurrence of death, myocardial infarction, hospitalization for congestive heart (that did not include renal replacement therapy) and stroke; renal replacement therapy was a censoring event. With this design, subjects with the most severe episodes of heart failure (those requiring dialysis) and subjects receiving renal replacement therapy are censored. From an analytical perspective, they are treated identically to subjects exiting the study without any clinical event. With composite outcomes, only the first event is counted when a patient experiences multiple components of the composite outcome in rapid sequence. Thus, patients in whom myocardial infarction is quickly followed by congestive heart failure are treated identically to those with myocardial infarction alone.

Sample size considerations are important in clinical trials. It is critical to note the planned minimum detectable between-group difference in primary outcome rates, as well as that observed in practice. Plans for interim analyses and stopping rules deserve close scrutiny. Stopping rules are frequently employed in trials, with pre-defined mathematical rules used to detect unexpectedly early but clear differences in primary outcome rates. Figure 1 shows a typical example, using the O'Brien–Fleming boundary method [8] for a trial planning a total of four treatment comparisons; alpha, the chance of a false positive result, is set at the conventional level of 0.05 in this two-sided comparison. It is important to note that, because of multiple comparisons, each separate analysis consumes a portion of the available alpha (hence the term 'alpha spending function'). Finally, it should also be noted that two-sided boundary P values for rejecting the null hypothesis are considerably less than 0.05 at all analyses except the last.

Among the four trials under review, USNH reported comparative findings that accounted for multiple comparisons; this issue is moot for CENH and CREATE, given that the null hypothesis was not rejected. For CHOIR, four interim analyses were planned; the "data and safety monitoring board recommended that the study be terminated in May 2005 at the time of the second interim analysis, even though neither the efficacy nor the futility boundaries had been crossed, because the conditional power for demonstrating a benefit for the high-hemoglobin group by
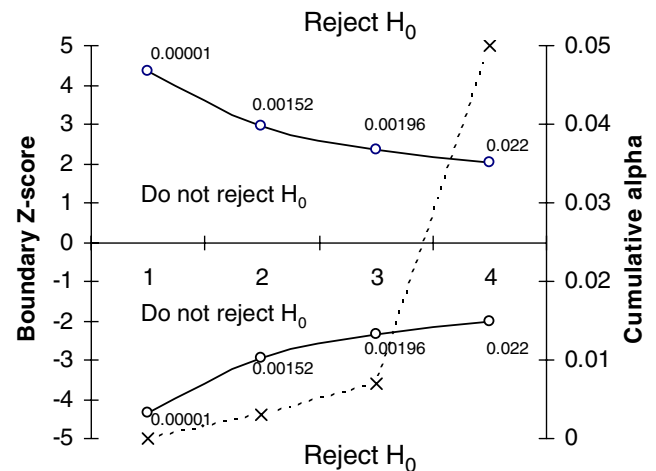
Fig. 1 O'Brien–Fleming boundary method for a clinical trial with power 0.80 and conventional alpha 0.05 in which four primary outcome comparisons (three interim and one final) were planned to occur at equal intervals. *Open circles* represent Z-score values to reject the null hypothesis at each analysis, and the corresponding figures are the one-sided nominal P values. x represents the cumulative alpha spent at each interim analysis. For two-tailed boundaries, multiply by a factor of 2

the scheduled end of the study was less than 5% for all plausible values of the true effect for the remaining data" [5]. Formal reporting of actual test statistics at each interim analysis, perhaps plotted against the planned boundary conditions, would considerably enhance the critical appraisal process for CHOIR. Another approach that might enhance interpretation would be to report event numbers within treatment groups, risk ratios and P values at each of the interim analyses, as well as the final analysis. Needless to say, the decision to terminate a trial also includes patient-safety variables that are not a part of the primary study outcome. While both CREATE and USNH were terminated early without formally reaching a primary outcome stopping boundary, it is evident that there were safety issues in both trials, with imbalanced rates of renal replacement therapy in the former and imbalanced rates of vascular access thrombosis in the latter. At the time that CHOIR was terminated, it was difficult to see any rationale for trial termination, as no differences in primary outcome rates were apparent and non-primary major adverse event rates were similar in both arms.

With well-conducted trials, primary outcome events should be collated in as timely a manner as possible; in particular, when decisions about primary outcome-based stopping rules are about to be applied at a planned interim analysis, it is reasonable to expect that a high proportion of the primary events will be available to data, safety and monitoring boards. For example, it would be useful to know how many events accrued between the last interim

analysis and study termination. This information is not available for USNH and CREATE. While not available in the original study publication, this important information is contained in a clinical study report for CHOIR, available at the US National Institutes of Health website: http://www.clinicaltrials.gov [9, 10]. In that report, it is apparent that a total of 37 primary outcome events were counted at the first interim analysis, 145 at the last interim analysis and 222 shortly thereafter at the termination of the study. It is notable that 35% extra primary events were added in the short interval between the last interim analysis and study completion.

The over-arching rationale of randomization is to generate two groups which are identical, for all measured and all unmeasured patient characteristics. It is worth considering, in advance, whether plans will be in place to deal with the likely occurrence of imperfect randomization. With perfect randomization, the true effect of an intervention should be insensitive to adjustment for baseline characteristics. If apparent treatment effects disappear when adjustment is made for baseline characteristics, the only safe course of action is to conclude that the null hypothesis cannot be rejected. Of the four studies, USNH and CENH formally reported plans for baseline covariate adjustment for primary outcome comparisons of the study interventions [2, 3].

Trial results

Even with well-designed trials, clear and comprehensive reporting can greatly help interpretation. The ability to publish supplementary information on journal websites means that a trade-off between readability and comprehensiveness should no longer be necessary. It is important that all patients entering a trial be accounted for and attributed at the conclusion of the trial. Similarly, when sequential measures are being compared, it is important to report the numbers being assessed, as well as the timing of these assessments.

In Table 3, baseline characteristics in the four major studies are compared by treatment arm. Statistically significant differences in baseline characteristics were present in each of the four studies, as shown in Table 3. It seems natural to question whether imbalances in hypertension, history of coronary artery bypass surgery (CHOIR [5]), body mass index, beta blocker use (CREATE [4]), angina pectoris (USNH [2]) and age (CENH [3]) could have biased treatment comparisons; by extension, having observed these imbalances, it seems natural to expect analyses of the primary outcomes that control for these imbalances.

Table 3 also shows a comparison of hemoglobin levels achieved in the four studies. Inspection of the temporal

evolution of hemoglobin levels showed that initial rates of rise of hemoglobin in the first month were greater in CHOIR [5] than in USNH [2], CENH [3] and CREATE [4]. During the maintenance phase, none of the studies consistently achieved planned hemoglobin levels in the high target groups; statistical separation, however, was clear. Table 3 also compares epoetin doses used to achieve and maintain hemoglobin levels. Epoetin doses were remarkably higher in the two studies from the United States of America [2, 5]. In the non-dialysis studies, these differences may have arisen, in part, by design, as all patients in CHOIR were treated with 10,000 U/week of epoetin alfa for 3 weeks [5], five-times the recommended starting dose in the early intervention arm of CREATE [4]. Regarding blood pressure levels, a between-groups difference in blood pressure levels was seen in the CHOIR study [4]. The other three studies showed equivalent blood pressure levels, albeit with the requirement for more antihypertensive agents in the high target arm in CENH [1–3].

With unadjusted analysis, CREATE and USNH showed no differences in the primary outcome of cardiovascular events [2, 4]; in contrast, there were more cardiovascular events with the high hemoglobin target in CHOIR [4]. Of the three trials examining cardiovascular outcomes, only USNH reported in the primary study publication a comparison of the primary outcome with adjustment for baseline variables, findings almost identical to those without covariate adjustment [2]. With CHOIR, the clinical study report alluded to above did report a comparison of primary outcomes in which adjustment had been made for several baseline covariates, such as age, gender, race, renal history, National Health and Nutrition Examination Survey (NHANES) scores for clinical variables, diabetes, glomerular filtration rate, albumin, reticulocyte count, and iron status. In that analysis, no efficacy differences were present, as the adjusted P value was 0.111, as opposed to 0.03 in the unadjusted analysis [9, 10]. This finding is noteworthy. With truly random assignment of patient-related characteristics that effect the primary outcome, adjustment for these characteristics should have no effect when randomly assigned treatments are compared. In consequence, it is impossible to refute the hypothesis that imbalanced assignment of non-treatment factors led to the observed disparity in event rates seen on unadjusted analysis. Once again, one is struck by the heterogeneity of the trial results, at least with unadjusted analytical strategies.

Quality of life was a secondary outcome in each of these trials. As such, these trials cannot be viewed as definitive for quality of life, even though the sample size in each of these studies was large. In addition, none of these trials concealed treatment allocation completely from patients, site investigators and outcome assessors. Three of these

Table 3 Study results (*CABG* coronary artery bypass graft, *CV* cardiovascular, *NHANES* National Health and Nutrition Examination Survey, *GFR* glomerular filtration rate, *LASA* linear analog self assessment, *KDQ* kidney disease questionnaire, *SF-36* short-form 36-item health survey, *LV* left ventricular, *LVH* left ventricular hypertrophy, *FACIT* functional assessment of chronic illness therapy, *KDQOL* kidney disease quality of life)

Parameter	CHOIR [4]	CREATE [3]	USNH [1]	CENH [2]
Success of randomization	↑ Hypertension and CABG in high target	↑ Body mass index and use of beta blockers in immediate treatment	↑ Angina pectoris in normal hematocrit	↑ Age in higher target
Hemoglobin levels (g/dl)				
Increment during month 1	>1 in both arms	≈0.3 in high target	≈0.3 in high target	≈0.3 in high target
During maintenance phase	High, ≈ 13.0 Low, ≈ 11.0	High, ≈ 13.3 Low, ≈ 11.4	High, ≈13.3 Low, ≈10.1	High, ≈ 13.0 Low, 10.8
Epoetin doses	High, 11,000 IU/week Low, 5,000 IU/week	High, 5000 IU/week Low, 2,000 IU/week	High, 450 IU/kg/week Low, 150 IU/kg/week	High, 150 IU/kg/week Low, 75 IU//kg/week
Blood pressure	High, ↑ by 0.2 mmHg Low, ↓ by 0.7 mmHg	No difference	No differences	No differences, but ↑ anti-hypertensive agents with high target
Primary outcome, unadjusted	↑CV events in high target. Hazards ratio ($P=0.03$)	No difference in CV events	No difference in CV events	Unadjusted analysis was not reported
Primary outcome, adjusted for baseline characteristics	Not reported in primary publication. However, clinical study report, available 12 September 2006, reported findings from adjusted analysis. With adjustment for demographic variables, such as gender, race, and age, and baseline clinical variables, such as renal history, NHANES scores, diabetes mellitus status, GFR, albumin, reticulocyte count, and iron status, hazards ratio were no longer statistically significant ($P=0.111$)	Analysis not reported	No difference in CV events	No difference in LV volume index when adjustment made for age, epoetin use at randomization, and baseline LVH
Quality of life	Patients not blinded. Similar evolution of LASA, KDQ, and SF-36 scores in both treatment arms, except for the score for the emotional role subscale of the SF-36, which was significantly higher in the low-hemoglobin group. Timing of quality of life assessments not reported	Patients not blinded. Quality of life by SF-36 better with high hemoglobin target at year 1 (general health, mental health, physical function, physical role, social function, and vitality) and at year 2 (general health, vitality)	Quality of life was assessed but not formally reported in primary publication	Patients were blinded. SF-36 vitality scores improved more in the higher than in the lower target group. No significant between-group differences were seen in the time course of FACIT fatigue scores and the KDQOL quality of social interaction scores
Major non-primary adverse events	No differences	↑ Renal replacement therapy with higher target	↑ Vascular access loss with higher target	↑ Stroke with higher target

trials formally reported quality of life comparisons, albeit with incompletely overlapping arrays of instruments [2–4]. If one accepts that potential quality-of-life effects are mediated through changes in hemoglobin levels, it is important to know the timing of these assessments, as one would not expect to see quality of life differences before hemoglobin levels have separated. Two of the trials, including the single trial with patient blinding, reported quality life benefits with higher hemoglobin targets [2, 3]. In contrast, the CHOIR study showed no clear benefit, and even a potential loss of quality life with the higher hemoglobin target [in the emotional role subscale of the short-form 36-item (SF-36) instrument]; these findings are difficult to interpret, as the effect of imbalanced baseline characteristics on quality of life comparisons is unknown, and the numbers of patients and timing of quality of life assessments were not reported [4].

Two of the studies had echocardiographic outcomes, and higher hemoglobin targets had no effect on ventricular size [2, 3]. With regard to higher target assignment, three of the trials showed more unanticipated, non-primary, serious adverse events. For example, the risk of replacement therapy was greater in CREATE [3], in spite of similar rates of change of glomerular filtration rate; vascular access loss was greater in USNH [1], and stroke was greater in CENH [2].

Conclusions

Even though a mere four trials were considered in this article, the differences between trials were remarkable at each level considered. Enrollment criteria, interventions studied, blinding, success of randomization and study findings showed important differences between studies. It is debatable, then, whether techniques such as meta-analysis should be employed in this universally heterogeneous environment. Similarly, it seems inadvisable to base one's therapeutic approach entirely on the findings of a single study. Presumably, depending on the specific question, studies with better methodology should carry more weight when study findings are applied to patients. It would also seem reasonable to add a note of caution about studies employing interventions few clinicians would consider in day-to-day practice. Unfortunately, despite many years of trials, the ideal approach to treating anemia in patients with chronic kidney disease remains to be decided and continues to generate controversy. Considerably more research is needed, and the quality of the evidence that accrues from this research will be entirely dependent on the quality of the methods used. In the published literature, methodological issues deserve careful attention before research findings are translated into clinical practice.

References

1. Besarab A, Bolton WK, Browne JK, Egrie JC, Nissenson AR, Okamoto DM, Schwab SJ, Goodkin DA (1998) The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. N Engl J Med 339:584–590
2. Parfrey PS, Foley RN, Wittreich BH, Sullivan DJ, Zagari MJ, Frei D (2005) Double-blind comparison of full and partial anemia correction in incident hemodialysis patients without symptomatic heart disease. J Am Soc Nephrol 16:2180–2189
3. Drüeke TB, Locatelli F, Clyne N, Eckardt KU, Macdougall IC, Tsakiris D, Burger HU, Scherhag A, CREATE Investigators (2006) Normalization of hemoglobin level in patients with chronic kidney disease and anemia. N Engl J Med 355:2071–2084
4. Singh AK, Szczech L, Tang KL, Barnhart H, Sapp S, Wolfson M, Reddan D, CHOIR Investigators (2006) Correction of anemia with epoetin alfa in chronic kidney disease. N Engl J Med 355:2085–2098
5. Bamgbola OF, Kaskel FJ, Coco M (2008) Analyses of age, gender and other risk factors of erythropoietin resistance in pediatric and adult dialysis cohorts. Pediatr Nephrol doi:10.1007/s00467-008-0954-3
6. Port RE, Mehls O (2008) Erythropoietin dosing in children with chronic kidney disease: based on body size or on hemoglobin deficit? Pediatr Nephrol doi:10.1007/s00467-008-0955-2
7. Vaziri ND (1999) Mechanism of erythropoietin-induced hypertension. Am J Kidney Dis 33:821–828
8. O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. Biometrics 35:549–556
9. US National Institutes of Health. http://www.clinicaltrials.gov
10. The Veritas Clinical Trial Register. http://download.veritasmedicine.com/PDF/CR004588_CSR.pdf