

UNIVERSITY OF CALGARY

Clinical activity score as an endpoint in randomized clinical trials of immunosuppressive
medications in anti-neutrophil cytoplasm antibody associated vasculitis: a validation
study

by

Michael Walsh

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

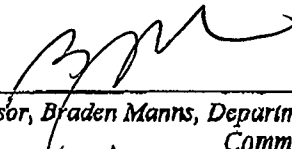
CALGARY, ALBERTA

JUNE 2009

© Michael Walsh 2009

UNIVERSITY OF CALGARY
FACULTY OF GRADUATE STUDIES

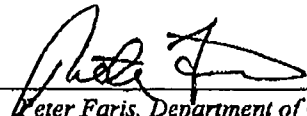
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled " Clinical activity score as an endpoint in randomized clinical trials of immunosuppressive medications in anti-neutrophil cytoplasm antibody associated vasculitis: a validation study " submitted by Michael Walsh in partial fulfilment of the requirements of the degree of Masters of Science.



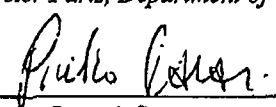
Supervisor, Braden Manns, Departments of Medicine and
Community Health Sciences



David Payne, School of Clinical Medicine, University of
Cambridge



Peter Faris, Department of Community Health Sciences

→ 

Pietro Ravani, Departments of Medicine and Community
Health Sciences

MAY 20/2009.

Date

ABSTRACT

Background: This study investigated the validity of a relapse of anti-neutrophil cytoplasm (ANCA) associated systemic vasculitis (AASV) as a surrogate endpoint for the clinical endpoint of end-stage renal disease (ESRD) or death.

Materials and Methods: Data from three randomized controlled trials of low toxicity immunosuppression compared to standard oral cyclophosphamide immunosuppression were used. Relapses were primarily defined using criteria from the Birmingham Vasculitis Activity Score (BVAS) according to the original study protocols. Alternative definitions were developed to optimize the use of BVAS information. Validity was assessed using the operational criteria developed by Ross Prentice that rely on significance tests of four associations between the treatment (low toxicity vs. standard oral cyclophosphamide immunosuppression), the surrogate endpoint (relapse of disease) and the clinical endpoint (ESRD or death). Generalized linear and latent multi-level models were utilized as a meta-analytical tool to assess Prentice's criteria.

Results: No definition of relapse could be considered valid. Failure to be accepted as valid by Prentice's criteria was potentially due to inadequate statistical power, true equivalence in terms of efficacy of the treatments, inherently flawed criteria, or truly invalid surrogate endpoints.

Conclusions: The effects of treatments on relapses can not be presumed to accurately reflect the effects of the treatment on ESRD or death. Further studies are required to identify a valid surrogate endpoint for trials of immunosuppressive treatments of AASV.

ACKNOWLEDGEMENTS

Neither this thesis nor this degree would ever have been completed without the support of some truly amazing people. Braden Manns was far more than a supervisor; he rekindled my interest in asking, “why?” and taught me the importance of addressing whether we are making patients live better and/or live longer every time we act as a physician or conduct clinical research. If I have any success in research, much of the credit will go to Braden’s endless efforts on my behalf.

David Jayne introduced me not only to the world of vasculitis and vasculitis research but to an amazing worldwide family of vasculitis researchers. Without his support, none of this research would have been possible but more importantly, by taking a chance and introducing my wife to this network of vasculitis researchers, he has undoubtedly affected the career course of our entire family for the better. Peter Faris introduced me to joys of multi-level models and renewed my confidence to continue after each crisis of faith in statistics.

I would also like to thank Brenda Hemmelgarn and Bruce Culleton who ensured I had superlative research mentors throughout my training and gave me far more opportunities than I deserved.

My wife, Dorothy is the real reason this thesis could be completed. She provided enough love, support, encouragement, and nutrition to sustain me through ten graduate degrees. And finally, my darling baby boy, Eoin, despite his complete absence of vocabulary, motivates me to try harder and be better every day. Amazing.

DEDICATION

To my parents who instilled the belief that knowledge is more important than “things” and to my wife that supports that belief even on the days she wishes she didn’t.

TABLE OF CONTENTS

APPROVAL PAGE.....	II
ABSTRACT.....	III
ACKNOWLEDGEMENTS.....	IV
DEDICATION.....	V
TABLE OF CONTENTS.....	VI
OVERVIEW	1
OBJECTIVES	2
BACKGROUND	3
Assessing the efficacy of treatments	3
Endpoints.....	4
<i>Clinical Endpoints</i>	4
<i>Non-clinical Endpoints</i>	4
<i>Why Use a Surrogate Endpoint?</i>	5
<i>Examples of Surrogate Endpoints Used Successfully</i>	6
Validating Putative Surrogate Endpoints	8
<i>Statistical Considerations</i>	8
Why Putative Surrogates Fail Validation	15
Trial Considerations	19
ANTI-NEUTROPHIL CYTOPLASM ANTIBODY ASSOCIATED SYSTEMIC VASCULITIS (AASV).....	21
AASV: History and Clinical Features	21
Potential Endpoints for Trials in AASV.....	23
Contemporary Clinical Trials in AASV	24
Disease Activity.....	26
<i>The Birmingham Vasculitis Activity Score</i>	26
<i>Other Disease Scoring Utilities</i>	27
Relapse as a Surrogate Endpoint in AASV Deserves Further Study	28
METHODS	28
Hypothesis	28
Data Sources	29
<i>CYCAZAREM</i>	29
<i>NORAM</i>	30
<i>CYCLOPS</i>	31

Defining the Clinical Endpoint.....	31
Defining the Surrogate Endpoint.....	32
<i>Primary Surrogate Endpoint: Protocol Relapses</i>	32
<i>Secondary Surrogate Endpoints</i>	33
Analyses	36
<i>Descriptive Statistics of Source Data</i>	36
<i>Analyses of the Primary Hypothesis</i>	37
<i>Prentice's Criteria: operational definitions</i>	38
RESULTS	43
Patients	43
Clinical Endpoints	44
The Surrogate Endpoint.....	46
<i>Protocol Defined Relapses</i>	46
<i>Alternatively Defined Relapses</i>	47
<i>Summary of Putative Surrogate Endpoints</i>	51
Evaluating Prentice's First Criterion – The Clinical Endpoint is Associated with the Treatment.....	51
Prentice's Second Criterion – The surrogate endpoint is associated with the treatment	52
<i>Protocol Defined Relapses</i>	52
<i>Protocol Defined Major Relapses</i>	53
<i>Peak BVAS Defined Relapses</i>	54
<i>Weighted Score Defined Relapses</i>	54
<i>Renal Relapses</i>	55
<i>Summary of Prentice's Second Criterion</i>	55
Prentice's Third Criterion – The surrogate is associated with the clinical endpoint.....	56
<i>Protocol Defined Relapse</i>	56
<i>Protocol Defined Major Relapse</i>	57
<i>Peak BVAS Defined Relapse</i>	58
<i>Weighted Peak BVAS</i>	58
<i>Renal Relapse</i>	60
<i>Summary of Prentice's Third Criterion</i>	60
Prentice's Fourth Criterion – The effect of the treatment on the clinical endpoint is accounted for by the effect on the surrogate	60
<i>Protocol Defined Major Relapse</i>	61
<i>Peak BVAS Defined Relapse</i>	62
<i>Weighted BVAS Defined Relapse</i>	62
<i>Summary of Prentice's Fourth Criterion</i>	63

Stability of Models	63
DISCUSSION	63
Appropriateness of Pooling Data.....	64
Predicting the Clinical Endpoint	65
The Association Between the Treatment and the Clinical Endpoint.....	66
The Association Between the Surrogate Endpoint and the Clinical Endpoint.....	67
The Association Between the Treatment and the Surrogate Endpoint.....	68
Treatment Effects Captured by Surrogate Endpoints	68
Interpretation	69
LIMITATIONS.....	70
SIGNIFICANCE AND FUTURE RESEARCH.....	72
TABLES	75
FIGURES.....	90
APPENDICES	119
Appendix 1. Birmingham Vasculitis Activity Score	120

List of Tables

Table 1. Organ damage often associated with anti-neutrophil cytoplasm antibody associated systemic vasculitis by organ system.....	76
Table 2. Summary of major randomized control trials in AASV by study short title for consideration in the validation of relapse as a surrogate endpoint.....	77
Table 3. Definitions of surrogate endpoints used for validation.....	78
Table 4. Baseline characteristics by trial and for overall pooled cohort.....	79
Table 5. Organ system involvement at baseline in European Vasculitis Study Group clinical trials of ANCA associated vasculitis.....	80
Table 6. Frequency of the composite clinical endpoint of death or end-stage renal disease and each component of the composite by trial.....	81
Table 7. Characteristics of patients who reached or did not reach the clinical outcomes in trials of ANCA associated vasculitis.....	82
Table 8. The association of baseline characteristics with the composite clinical endpoint of ESRD or death using multi-level, multi-variable logistic regression.....	83
Table 9. Protocol defined relapses for trials of AAV.....	84
Table 10. Differences in the organ involvement according to clinical composite endpoint of ESRD or death.....	85
Table 11. Results of the logistic regression model using peak system activity as a predictor of death or ESRD used to develop weighted BVAS.....	86
Table 12. Summary of the association of each definition of a relapse identified as a putative surrogate endpoint for composite clinical endpoint of death or ESRD.....	87
Table 13. Summary of Prentice's second criterion.....	88
Table 14. Summary of Prentice's third criterion.....	89

List of Figures and Illustrations

Figure 1. Hierarchy of endpoints in randomized control trials.....	91
Figure 2. Prentice's criteria.....	92
Figure 3 The different ways in which a putative surrogate endpoint may be invalid.....	94
Figure 4 Representative pathology of the ANCA associated systemic vasculitides.	95
Figure 5 Summary of the proposed pathogenesis of vasculitis.	96
Figure 6 Plan for the assessment of defining relapses with the pooled data set.	97
Figure 7 Patient flow through validation study.....	98
Figure 8 Probability of death or ESRD predicted from a saturated logistic regression model using age and baseline eGFR as predictor variables.....	99
Figure 9 Probability of death or ESRD predicted from a saturated logistic regression model using age and baseline eGFR as predictor variables.....	100
Figure 10 Peak Birmingham Vasculitis Activity Scores (BVAS) after the successful induction of remission in patients who did not reach death or ESRD (No) and those that did reach death or ESRD (Yes) arranged by trial.....	101
Figure 11 The predicted probability of death or ESRD by peak BVAS.....	102
Figure 12 The predicted probability of death or ESRD by weighted BVAS.....	103
Figure 13 Predicted probability of death or ESRD by protocol defined relapse.	104
Figure 14 The predicted probability of death or ESRD over a range of baseline estimated glomerular filtration rate (eGFR) for patients with and without a relapse defined by a peak BVAS	105
Figure 15 Predicted probability of death or ESRD from logistic regression using a weighted BVAS definition for relapse and eGFR as predictor variables demonstrating effect modification of the effect of relapse by eGFR.....	106
Figure 16 Predicted probability of death or ESRD from logistic regression using protocol defined major relapse status and age as independent variables.....	107

List of Abbreviations

Abbreviation	Definition
AASV	Anti-neutrophil cytoplasm antibody associated systemic vasculitis
ANCA	Anti-neutrophil cytoplasm antibody
AZA	Azathioprine
BVAS	Birmingham Vasculitis Activity Score
CYC	Cyclophosphamide
CYCLOPS	Short name for a randomized control trial comparing daily oral cyclophosphamide to pulse intravenous cyclophosphamide for induction remission in vasculitis
CYCAZAREM	Short name for a randomized control trial comparing daily oral cyclophosphamide to azathioprine after induction remission in vasculitis
eGFR	Estimated glomerular filtration rate
ENT	Ears, nose, and throat
ESRD	End-stage renal disease
EUVAS	European Vasculitis Study Group
MPA	Microscopic polyangiitis
MTX	Methotrexate
NORAM	Short name for a randomized control trial comparing daily oral cyclophosphamide to methotrexate for induction remission in vasculitis
RCT	Randomized controlled trial
WG	Wegener's granulomatosis

OVERVIEW

Anti-neutrophil cytoplasm antibody (ANCA) associated systemic vasculitis (AASV) is an important cause of kidney failure and death. Before treatment with immunosuppressive medications and steroids were used, the expected 5 year mortality was 90% ^{1,2}. Although treatment is associated with improved survival, it is also associated with serious adverse events in up to half of treated patients ³. Randomized clinical trials (RCTs) testing the efficacy of new treatments that may reduce kidney failure and mortality with less toxicity are a priority in this field.

Randomized clinical trials are the gold standard by which the clinical efficacy of medications is judged ⁴. Classically, RCTs are designed to judge efficacy by comparing the impact of a new medication to placebo or active comparator on an endpoint that is of known importance to a patient's survival or function. These endpoints are known as clinical endpoints ⁵. Typical clinical endpoints include death, cardiac death, myocardial infarction, stroke and end-stage renal disease. Evaluating the impact of a medication on these endpoints may require a study that lasts many years and/or requires many patients due to the relatively infrequent occurrence of these events ⁶. Therefore, trials based on these endpoints are both time and resource intensive. Given these issues and the concern by some that this delays the availability of new and potentially lifesaving medications, the use of surrogate endpoints rather than clinical endpoints has become common.

A surrogate endpoint is a variable or parameter that accurately predicts the occurrence of a clinical endpoint in a test of an intervention ⁵. Surrogate endpoints typically occur more

quickly than clinical endpoints thus increasing the number of endpoints experienced at any point in time and improving statistical power and trial efficiency. A valid surrogate endpoint may therefore allow a RCT to recruit fewer patients and follow them for less time, decreasing the resources required to perform a RCT and reach a valid conclusion ⁷.

Given the high mortality and morbidity associated with AASV as well as increasing interest from pharmaceutical companies in marketing new medications for autoimmune disease, RCT design in AASV is becoming increasingly important ⁸. Currently there are no validated surrogate endpoints for use in RCTs in AASV ⁹. Despite this, disease activity scores, biochemical parameters and immunologic markers are often used as trial endpoints. The objective of this study is to use data from existing RCTs in AASV and their long-term follow-up to ascertain the validity of putative surrogate endpoints for the composite clinical endpoints of interest (end-stage renal disease [ESRD] or all-cause mortality).

OBJECTIVES

Primary: To examine the validity of using the occurrence of a relapse of AASV, as defined by the Birmingham Vasculitis Activity Score (BVAS), within 18 months of enrollment in a clinical trial as a surrogate endpoint for the composite clinical endpoint of death or ESRD at any time.

Secondary: To examine the validity of using the following parameters as surrogate endpoints in clinical trials of immunosuppressive medications in AASV:

- The occurrence of relapse of AASV, utilizing an optimized BVAS definition of relapse, within 18 months of enrollment in a clinical trial as a surrogate endpoint for the clinical endpoint of ESRD or death at any time.
- Renal relapse defined by at least two of the following: 1) increase in proteinuria defined by BVAS question R2 ($>1\text{g}/24\text{ hours}$), 2) increase in haematuria of at least 10 cells/high powered field and/or 3) increase in creatinine by 20% within 18 months of enrollment in a clinical trial as a surrogate endpoint for the clinical endpoint of time to ESRD or death at any time.

BACKGROUND

Assessing the efficacy of treatments

By the late 1980s, RCTs using clinical endpoints were well established as the gold standard approach for assessing the efficacy of medical treatments⁴. These trials are, however, costly and require long completion times owing to the relative infrequency of the endpoints^{6, 7, 10}. This presents significant challenges in generating results in a timely manner. These issues are a concern to industry, regulators, patients, and medical researchers alike. For example, today's trials for treatment of heart disease with the clinical composite primary endpoints of cardiovascular death, myocardial infarction, and stroke are often multi-national and can involve as many as 20 000 subjects recruited from 50 or more centers and followed for a median of 2–5 years⁶. These studies cost many millions of dollars. In the context of drug approval, such trials make it almost impossible for all but the largest sponsors to carry out the entire approval process¹¹.

Endpoints

Clinical Endpoints

Clinical endpoints are events that affect the way a patient functions, feels or survives⁵. In lay terms, they may be considered events of importance to a patient. Examples of such endpoints include: all cause mortality, cardiovascular mortality, stroke, myocardial infarction, ESRD and health related quality of life (HRQOL). The occurrence of such events may take years from the time a therapy is started. This slow accrual of endpoint events requires a large sample size and/or a long follow-up time to ensure adequate power are achieved to show a statistically significant difference. Advances in population health and the standard of care as well as earlier diagnoses may also be reducing the rate of occurrence of clinical endpoints in many diseases⁷. This further elevates the sample size and follow-up time requirements of clinical trials. Given that the completion of a RCT may be the rate limiting step between the discovery and implementation of a new therapy, and that the number and time to clinical endpoints is crucial, there has been much interest in non-clinical endpoints.

Non-clinical Endpoints

A non-clinical endpoint is any event or parameter that is not a clinical endpoint but is used to define the outcome of a RCT. Typically, non-clinical endpoints are biomarkers. Biomarkers are defined by the National Institutes of Health as characteristics that are objectively measured as an indicator of the normal biological process, pathogenic process or pharmacologic response to a therapeutic intervention⁵. The use of biomarkers as

endpoints in early phase trials of drug development is well established. Some biomarkers are used as a substitute for a clinical endpoint in RCTs evaluating the efficacy and role of a therapeutic. A biomarker that substitutes a clinical endpoint and predicts the clinical benefit of the therapy is termed a surrogate endpoint⁵. For a biomarker to be accepted as a surrogate endpoint, it must be validated. For the purpose of this discussion, those biomarkers that are used as a surrogate without prior validation will be termed putative surrogate endpoints. Finally, some biomarkers may blur the lines between being purely a biomarker and having clinical significance. Such biomarkers are often referred to as intermediate surrogate endpoints^{10, 12}. This array of endpoints forms a hierarchy for the information they provide on the efficacy of a therapy (Figure 1).

Why Use a Surrogate Endpoint?

Although clinical trials powered to show a difference or non-inferiority in a clinical endpoint are the gold standard, trials utilizing an appropriately validated surrogate endpoint may be considered equivalent in some cases. Surrogate endpoints are particularly important in cases where clinical endpoints occur infrequently. Surrogate endpoints generally occur earlier than clinical endpoints and have the added potential advantages of being easier to measure and typically occur on a continuous scale thus improving statistical power^{7, 13}. These properties reduce the time, number of participants necessary, the cost requirements and ultimately the difficulty of performing a RCT. These qualities are attractive to investigators given the limited funding available for RCTs and the pressure to publish research. Similarly, these qualities are attractive to the pharmaceutical industry as regulatory agencies may accept trials utilizing a surrogate

endpoint for the licensing of some medications. Wittes and co-authors provide several examples of these qualities for putative surrogate endpoints in cardiovascular disease⁶. In their calculations, a RCT of an antihypertensive powered to the clinical endpoint of the prevention of stroke would require 25,000 patients to be followed for 5 years. A similar magnitude of benefit could be demonstrated if the same medication were used in a RCT powered to the endpoint of diastolic blood pressure lowering only 200 patients followed for 2 years are required. A surrogate endpoint therefore reduces the time and money required to license a medication. However, before considering these practical qualities of utilizing a surrogate endpoint, it must first be proven the equivalent (or near equivalent) of the clinical endpoint it is intended to replace.

Examples of Surrogate Endpoints Used Successfully

Two examples of surrogate endpoints commonly used with success have been CD4 cell count and viral load quantification in human immunodeficiency virus (HIV)/acquired immune deficiency syndrome (AIDS) trials of anti-retroviral drugs and regression of tumor size in chemotherapeutic trials in oncology literature.

In early AIDS literature, the use of anti-retroviral medications in comparison were shown to improve survival and that elevations in patients CD4 cell counts or suppression of their HIV viral loads correlated well with both treatment allocation (active drug vs. placebo) and survival time^{14, 15}. This corresponded well to the known pathology of AIDS as the virus infected CD4 cells which then lead to the immunodeficiency and subsequent immunodeficiency related death. CD4 cell counts and viral loads were then accepted as

valid surrogate endpoints in trials of anti-retroviral medications in AIDS research and led to the fast-track use and marketing of several medications for the treatment of AIDS.

Similarly in oncology literature, regression of tumor size was noted to be closely associated with survival time in trials of chemotherapeutics ¹⁶. Again, this relationship was supported by the pathology of the disease where the harmful effects of some cancers are mediated by a combination of local mechanical effects, metastatic effects and the production of detrimental cytokines all of which are directly proportional to total tumor burden. Tumor regression and time-to-tumor growth were, therefore, accepted as valid surrogate endpoints in trials of cancers such as ovarian and colon cancer.

In AIDS and oncology literature, the use of surrogates has more recently been met with some controversy. In certain drug trials, the above surrogate endpoints may have only a weak relationship to the clinical endpoints they are thought to predict ¹⁷⁻¹⁹. Whether this is a function of the specificity of these surrogate endpoints to the medication or the clinical scenario in which they were originally validated or a change in the mechanisms by which these diseases cause death (e.g. medication toxicity may have become a major component of the pathway to death) is not yet understood.

Validating Putative Surrogate Endpoints

Statistical Considerations

Statistical models have been developed to test the validity of putative surrogates. Many models have been proposed but most have roots in the seminal work by Ross Prentice, who proposed operational criteria for the statistical validation of surrogate endpoints¹³. Briefly, Prentice's criteria outline that the putative surrogate must be "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint."¹³ Functionally, this requires that the distributions of both the surrogate endpoint response and the clinical endpoint response must be dependent on the treatment allocated, that the surrogate should capture the dependence of the clinical endpoint on the treatment allocation and that the surrogate must have some prognostic implication on the clinical endpoint.

Operationally, Prentice's criteria are defined in statistical terms. Perhaps the easiest way to understand the operational criteria is to transform them into simple regression equations with an interpretation of the salient features of each regression equation. Prentice's first description of his criteria employed time-to-event clinical endpoints with bivariate surrogate endpoints in single trials. Buyse et al illustrated these criteria using linear and logistic regression models which are somewhat more transparent and shall be used in this description²⁰. Because satisfying Prentice's criteria is dependent on

hypothesis testing, linear or logistic regression models can be used so long as the appropriate hypothesis test is used.

The operational criteria first state that the clinical endpoint (T) must be dependent on the treatment allocation (Z). In the following equation T is assumed to be a normally distributed, continuous ratio or interval parameter and Z is assumed to be a nominal/categorical variable:

$$T_i = \mu_T + \beta Z_i + \varepsilon_{Ti} \quad (\text{Equation 1})$$

β must not be zero in this model to show that T is dependent on Z. ε in all equations represents the error component of the linear predictor for the regression model.

The second operational criterion requires the surrogate endpoint (S) (another normally distributed, continuous ratio or interval parameter) to also be dependent on treatment allocation (Z):

$$S_i = \mu_S + \alpha Z_i + \varepsilon_{Si} \quad (\text{Equation 2})$$

Again, α must not equal zero in this model to show the dependence of S on Z.

The third operational criteria requires the clinical endpoint (T) to be dependent on the surrogate endpoint (S):

$$T_i = \mu' + \gamma_Z S_i + \varepsilon'_i \quad (\text{Equation 3})$$

Here, γ_Z cannot equal zero to show the dependence of T on S.

Lastly, the entire effect of treatment allocation (Z) on the clinical endpoint (T) must be captured by the surrogate endpoint (S):

$$T_i = \mu' + \beta_S Z_i + \gamma_Z S_i + \varepsilon'_i \quad (\text{Equation 4})$$

In this model, γ_Z cannot be equal to zero and β_S must not be proven to be anything other than zero. The hypothesis testing of these parameters is typically carried out by performing a t-test of the coefficient with the null hypothesis being that the coefficient is zero. Thus, rejecting the null hypothesis by virtue of a highly significant Wald test is generally required except in the case of β_S where the failure to reject the null hypothesis is required. Importantly, Equations 1, 2, 3 and 4 demonstrate the nomenclature specific to generalized linear models which are appropriate for normally distributed continuous endpoints but are also applied to bivariate endpoints ²⁰.

These criteria have been the source of much debate amongst statisticians and clinical trialists. Chief amongst these have been the arguments that the operational criteria may only be valid for binary endpoints ²¹, that the criteria may be unrealistically stringent ²² and that examples of perfect surrogates (i.e. those that perfectly predict clinical endpoints) could be found that actually failed Prentice's criteria ⁷. Regardless of the debate, they have spawned many techniques for the validation of surrogate endpoints. While Prentice's criteria are generally accepted as valid, the major concern is that their rigidly stringent nature may reject useful putative surrogate markers.

Freedman et al advocated one commonly used method for exploring the validity of a putative surrogate endpoint, known as the proportion of treatment effect explained (PTE) ²³. Rather than hypothesis testing as used by Prentice, which requires the acceptance of a null hypothesis on the basis of failing to reject the null hypothesis, PTE quantifies the proportion of the effect of the treatment on the clinical endpoint that is explainable by the

proportion of the effect of the treatment on the surrogate endpoint (and corresponding confidence intervals). This approach has raised concerns that the number of patients needed to have reasonable confidence in the surrogate endpoint may be as high as the clinical endpoint ²⁴. Concerns have also been raised that the PTE may be based on models which do not fit the data ²⁵.

Buyse et al proposed an adaptation of the PTE which overcomes some of the limitations of PTE, specifically difficulties with differences in unit intervals between surrogate and clinical endpoints ²⁰. This new measure of surrogacy required joint linear models to estimate the regression coefficients for the effect of the treatment on the clinical endpoint and the effect of the treatment on the surrogate endpoint. These joint models are exactly those specified in Equations 1, 2 and 4 above in Prentice's operational criteria but the treatment of the regression coefficients no longer relies on hypothesis testing. The beta coefficient in Equation 1 estimates the effect of the treatment on the surrogate endpoint while the alpha coefficient in Equation 2 estimates the effect of the treatment on the surrogate endpoint. If the units of the endpoints are the same (i.e. either continuous parameters in the same units or both dichotomous parameters), the ratio of these two coefficients thus estimates the effect of the treatment on the clinical endpoint relative to the effect on the surrogate endpoint. This ratio was termed the relative effect (RE). The RE represents the validity of the surrogate endpoint at the level of the trial with a strong surrogate endpoint producing a RE close to 1. Intuitively, the RE is the slope of a regression line between the $\ln(OR_{ZT})$ and the $\ln(OR_{ZS})$. Narrow confidence intervals around the RE imply it will accurately predict the magnitude of effect that a treatment

should have on a clinical endpoint. Trial-level surrogacy is thought to represent the quality of the surrogate endpoint to predict the trial level results of the treatment effect on the clinical endpoint (i.e. treatment Z_1 is more effective than treatment Z_2 at preventing clinical endpoint T).

The RE was complimented by a proposed measure of individual patient level surrogacy termed the adjusted association (AA). The AA is the relationship between the clinical and surrogate endpoints after adjustment for the treatment effect. The AA is thought to represent the ability of the surrogate endpoint to predict the occurrence of the clinical endpoint in a given patient (i.e. how likely patient j who experiences surrogate endpoint S is to experience clinical endpoint Z). The AA can be calculated in at least two ways. The first is via the regression equation that models the effect of S on T with adjustment for Z in Equation 4 above. In this equation the γ_Z term represents the AA with a large value representing a strong association and a perfect individual level surrogate should have a γ_Z of ∞ . Alternatively, the AA can be represented by the coefficient of determination of Equation 3 ($R^2_{\text{individual}}$). In this context, an AA approaching one demonstrates a strong association. The RE therefore connects the treatment effects at the population-averaged level while the AA (γ_Z) connects them at the individual specific level (random-effects model terms)²⁰.

The paradigm of using a trial level and individual level measure of surrogacy was further extended by Buyse et al to a meta-analytical approach²⁶. This approach allows the validation of a surrogate endpoint when individual trials did not show significant

differences on a clinical endpoint (considered a requisite by some) and allows validation in the absence of Prentice's first criteria^{25,27}. This approach allows the quantification of parameters similar to the RE and the AA estimated for a single trial. In the meta-analytic approach, rather than the ratio of the coefficients for the treatment effect in the joint linear models, the correlation between the variances of the correlations is analogous to the RE and is a measure of trial level surrogacy. Rather than the simple ratio of the β coefficients of joint linear regression model, the covariance of the β coefficients of joint multilevel linear models is used. This correlation produces the coefficient of determination or R^2_{trial} . Here, trial level surrogacy describe how consistent trial results may be using the surrogate or clinical endpoint between several trials with the overall importance of each trial in determining the strength of this association determined by the number of patients and number of events within that trial. Because trial events are necessarily influenced by treatment effects, the impact of treatment allocation is a key component of determining trial level surrogacy.

At the individual patient level, the correlation between the random error term variances is used to calculate the parameter analogous to the AA. This coefficient of determination is called the $R^2_{\text{individual}}$. This measure of individual patient level surrogacy describes how likely the occurrence of the surrogate in a patient will be followed by the clinical endpoint. This is not necessarily the same as trial level surrogacy as a surrogate can perform well overall in a trial while having little predictive accuracy for any given patient due to the averaging effect over numerous trials (i.e. a surrogate that is only weakly predictive in individual patients but is consistent among all trials may be a good surrogate

at the trial level). Furthermore, while the strength of the trial level surrogacy is dependent on the treatment effect, the individual level is only a measure of the association between the surrogate and the clinical endpoint irrespective of the treatment itself. Ideally, both parameters should be close to 1 to demonstrate the validity of the surrogate although exact cut-off “levels” of the coefficient of determination to determine validity have not been specified. This approach has been adapted for situations where the clinical endpoint is continuous but the surrogate is binary or vice versa and for time-to-event analyses²⁸⁻³⁰.

Although the meta-analytical method has its roots in Prentice's criteria, Buyse and colleagues argue their operational criteria are simpler²¹. Only two relationships must be shown: 1) that the surrogate endpoint is strongly associated with the clinical endpoint (a function of R^2_{trial}) and 2) that the treatment effect on the surrogate endpoint is highly correlated to the treatment effect on the clinical endpoint (a function of the $R^2_{\text{individual}}$). Buyse et al argue that a strong association between the surrogate and clinical endpoint here would suggest a strong biologic link between the surrogate and clinical endpoint²¹. It should be noted that these statistical associations cannot prove causation, i.e. just because both a surrogate and clinical endpoint are highly associated with a treatment and with each other does not prove that the occurrence of the surrogate endpoint is the mechanism by which death occurs. While associations at both the individual and trial level strengthen the argument for causation they cannot disprove a non-causative correlation, they only make it less likely by using more stringent statistical criteria than a

simple association. Essentially, causality cannot be tested but may be less likely in the absence of a strong statistical association.

Why Putative Surrogates Fail Validation

The validity of a surrogate is thought to rely on its close relationship to the causal mechanism of the disease which results in the clinical endpoint ³¹. Additionally, the effect of the therapy in question must be fully captured by the surrogate endpoint as laid out in Prentice's criteria (Figure 2). Putative surrogates may therefore fail to be validated by a number of mechanisms (Figure 3).

If the surrogate endpoint does not lie on the causal pathway between the disease etiology and the final endpoint, it is unlikely to be valid (Figure 3b) ³¹. This would be the expected scenario in cases where the pathobiology is poorly understood and a biomarker associated with the disease has little or no actual bearing on the clinical endpoint. Similarly, surrogates that lie on only one of several causal pathways to the clinical endpoint with the majority of the intervention effect on a separate causal pathway are unlikely to be valid (Figure 3c). One might expect this in a scenario of using blood pressure as a surrogate for a trial examining the effect of lipid lowering agents on cardiovascular mortality.

Although both hyperlipidemia and blood pressure lie on the causal pathway between atherosclerotic disease and cardiac death, the surrogate in this scenario is unlikely to be affected given the mechanism of action of the intervention.

Surrogates that lie on the causal pathway affected by the intervention but with the majority of the disease effect occurring by a separate causal pathway are also unlikely to be valid (Figure 3d) ³¹. This was seen in the use of serum lipid levels as a surrogate for cardiovascular outcomes in patients treated with hormone replacement therapy. HRT was associated with improvements in patient's lipid profile with a reduction of low-density lipoprotein cholesterol and an increase in high-density lipoprotein cholesterol ³², ³³. In studies of lipid lowering medications, these measures had been proven to be valid and reliable surrogates for cardiac morbidity ³⁴⁻³⁶. However, in two large randomized control trials, the Heart and Estrogen/Progestin Replacement Study (HERS) and the Women's Health Initiative Study (WHI), HRT did not reduce the incidence of cardiovascular events despite a significant improvement in patient's lipid profile ³⁷⁻³⁹. Similar results were seen in a trial evaluating sodium fluoride for fracture prevention. Sodium fluoride had been noted to improve bone mineral density ⁴⁰, a surrogate endpoint often used in trials of bisphosphonate medications ⁴¹⁻⁴³. Combined with promising epidemiologic data on fracture rates in patients taking sodium fluoride, it was assumed that bone mineral density would be a surrogate endpoint for the clinical endpoint of fracture rates ^{40, 44}. However, a randomized trial was done comparing the effect of sodium fluoride to placebo on fracture rates and despite an improvement in bone mineral density, there was no difference in fractures ⁴⁵. This may be attributed to bone mineral density being only one of many potentially important factors, for example bone architecture and fall risk, in the prevention of fractures ⁴⁵.

Lastly, surrogates that lie on one causal pathway but with an intervention affecting not only that pathway but several others, including directly influencing the clinical endpoint, are unlikely to be valid ³¹. This may occur when the causal mechanism is complex and the intervention has pleotropic effects that the surrogate does not encompass or when the intervention has an associated toxicity with a powerful effect on the clinical endpoint. The best known example of the latter is the Cardiac Arrhythmia Suppression Trial (CAST) study ⁴⁶. By the 1980's post-myocardial infarction arrhythmias were noted to be a major source of mortality ⁴⁷⁻⁵¹. The use of class Ic anti-arrhythmic drugs such as encainide and flecainide were noted to abolish these arrhythmias. Thus, the suppression of post-infarction arrhythmias became an accepted surrogate for death in evaluating anti-arrhythmic drugs. The CAST study evaluated the effect of encainide and flecainide on post-infarction mortality and found their use to be associated with an excess mortality compared to placebo. The ability of these medications to not only suppress premature ventricular beats but to also be pro-arrhythmic was one of the factors attributed to the excess mortality ⁵². Because these medications had already been in widespread use prior to the CAST study, the excess mortality caused by their use in the United States is estimated by some to exceed the number of lives lost by the United States in the Vietnam War ⁵³.

Several other examples of failed surrogate endpoints exist in the nephrology literature. Prominent among these are the use of coronary calcification in the assessment of the efficacy of non-calcium phosphate binding medications, the use of hemoglobin levels in the assessment of different doses of erythropoietic medications and the use of vascular

access monitoring for early endovascular procedures for dialysis accesses. In terms of non-calcium phosphate binding medications, coronary calcification was shown to be a predictor of cardiac mortality in non-ESRD patients ^{54, 55}. Non-calcium phosphate binding medications were subsequently shown to be associated with the regression of coronary calcification ⁵⁶. A large RCT using the clinical endpoint of all-cause mortality did not, however, demonstrate any survival advantage for the use of non-calcium phosphate binding medications ⁵⁷.

With respect to the role of hemoglobin levels in the prescription of erythropoietic agents, observational studies of ESRD patients demonstrated those with higher hemoglobin values survive longer than those with lower hemoglobin values ^{58, 59}. Erythropoietic agents that increase hemoglobin were therefore assumed to improve mortality because hemoglobin level was used as a putative surrogate. Again, in a large RCT, the use of erythropoietic agents to normalize hemoglobin not only showed no benefit in survival but actually appeared to cause harm ⁶⁰. It should be noted that the example of erythropoietic agents can be variably interpreted with respect to surrogacy as the hemoglobin level is most often the assigned treatment rather than the outcome of the trial. It is, however, commonly referred to as a surrogate endpoint in both the general medical and nephrology literature.

Another example in the nephrology literature can be seen with the use of hemodynamic measures of vascular accesses in hemodialysis patients. Access blood flow rates are known to fall as stenoses develop in the distal vessels of hemodialysis patients with

vascular grafts for dialysis access and are an independent predictor of graft failure⁶¹. The correction of graft stenoses improves blood flow rates and presumably reduces the likelihood of graft survival, an important outcome for patients receiving hemodialysis. A randomized trial examining the impact of ultrasound monitoring with angioplasty of stenoses in patients with low access blood flow, however, failed to demonstrate any improvement in access survival despite improvements in access flow (the surrogate endpoint) after the intervention⁶²⁻⁶⁴.

These examples not only demonstrate the necessity to validate a surrogate endpoint but, in several instances, also that a surrogate endpoint is only validated for therapies with the same or very similar mechanisms of action and adverse effects as that tested in the validation study.

Trial Considerations

Validation of a putative surrogate endpoint, irrespective of the statistical methods used, requires a RCT. Traditionally, this RCT of a new therapy is powered to detect a difference in a clinical endpoint. Concurrently, the putative surrogate endpoint is measured. At the end of the trial, if the effects of the intervention on the surrogate endpoint are similar in magnitude and direction to the effect of the intervention on the clinical endpoint, the surrogate would be considered valid. In subsequent trials of the same or similar medications for the same disease, the surrogate could be used to improve the efficiency of the trial.

This method was used in the validation of hypertension as a surrogate for cardiovascular death in clinical trials of anti-hypertensives and low density lipoprotein cholesterol level as a surrogate for cardiovascular death in trials of cholesterol lowering agents⁶⁵⁻⁷⁰. Use of these surrogates can reduce clinical trials to a duration of less than one year involving fewer than 1000 subjects compared to the clinical endpoint trials which require 5 times as many patients and 5 times the duration of follow-up⁶. Thus, valid surrogate endpoints may decrease the cost and time required of a clinical trial. Some surrogate endpoints have also been advocated as decreasing patient discomfort when invasive tests are required for clinical endpoint assessment^{13,71} and to reduce the ethical dilemma of requiring significant numbers of patients to progress to the endpoint of death⁵. This final point is contentious as it is often used in arguments to justify the provision of new medications despite the lack of evidence of their efficacy and it is rarely acknowledged that a newer medication may have significant toxicity that actually harms patients.

In the 1990's, on the basis of these arguments, regulatory agencies for new medication came under tremendous pressure to "fast-track" the approval of medications for certain cancers and Acquired Immunodeficiency Syndrome (AIDS)⁷². The premise was that these diseases were incurable and universally fatal. The Food and Drug Authority (FDA) allowed the licensing of medications for these diseases after proof of efficacy on the basis of two separate trials which used the same putative surrogate⁷². If both trials demonstrated the efficacy of the therapy on the surrogate endpoint then it would be provided conditional approval. Once conditional approval was granted, a third trial utilizing a clinical endpoint was mandated to validate the surrogate endpoint. This policy

has more recently come under scrutiny due to ongoing licensing of some medications with negative clinical endpoint results despite previous positive surrogate endpoint results⁵⁷. This is relevant in the current era of drug development in the context of chronic, rare diseases such as AASV. Rare diseases pose challenges in performing RCTs due to the limited patient population to study and due to the low event rates for clinical endpoints despite large burdens of morbidity and poor quality of life. These challenges are magnified by the proliferation of new treatments that have a strong biologic rationale for their efficacy and pressure from patients with these diseases and their care providers to access these new therapies. Valid surrogate endpoints to efficiently evaluate new therapies are therefore of great interest.

ANTI-NEUTROPHIL CYTOPLASM ANTIBODY ASSOCIATED SYSTEMIC VASCULITIS (AASV)

AASV: History and Clinical Features

AASV is comprised of two clinical entities, Wegener's granulomatosis and microscopic polyangiitis. These two clinical entities have been united under the term AASV due to the relatively recent discovery of their shared diagnostic biomarker, the ANCA immunofluorescence and enzyme linked immunosorbent assay tests⁷³. For the purpose of clinical studies, WG and MPA are often grouped together as AASV due to commonalities in their presentation, treatment and likely pathogenesis.

AASV, as a collective, is a multi-system disease with a highly variable presentation (Table 1)⁷⁴. The classically described manifestations of AASV can be broadly

categorized as those affecting vital organs and those affecting non-vital organs. Classic vital organ manifestations include: rapidly progressive renal failure characterized by the nephritic syndrome (hematuria, proteinuria and hypertension), mononeuritis multiplex, and lung hemorrhage. Classic non-vital organ manifestations included: fever, malaise, weight loss, arthritis, vasculitic rashes and sinusitis with erosions.

Morphologically, the symptoms of AASV are explained by the presence of acute necrotizing granulomas of the upper and lower respiratory tract, necrotizing or granulomatous vasculitis affecting small to medium-sized vessels (e.g., capillaries, venules, arterioles, and arteries), most prominent in the lungs and upper airways and focal necrotizing, often crescentic, glomerulitis (Figure 4) ^{75,76}. While incompletely understood, the role of ANCA is central to the most accepted models of the pathogenesis of AASV (Figure 5) ⁷⁷⁻⁸³. In this model, ANCA binds to and activates primed neutrophils in the presence of tumor necrosis factor α resulting in a respiratory burst. This causes the release of chemokines, cytokines and reactive oxygen species which damage nearby endothelium and recruits more neutrophils, perpetuating the cycle of neutrophil activation and endothelial damage.

Early observations of AASV were largely limited to case series due to its relative rarity and difficulties identifying patients and classifying the disease ⁸⁴. These factors also resulted in a delay in diagnosis, and subsequent treatment, until signs and symptoms of AASV were more flagrant ⁸⁵. This likely contributed to the extremely high mortality rates in early case series of this disease as it was uniformly fatal within five years of diagnosis.

By 1967, corticosteroids emerged as the standard of care due to several observations that mortality was significantly reduced ^{2, 86, 87}. However, sustained remission was uncommon with corticosteroids alone and, following a trend seen in many severe autoimmune diseases, toxic immunosuppressive medications such as nitrogen mustard, cyclophosphamide, methotrexate and azathioprine were co-administered ⁸⁸⁻⁹². This combination treatment appeared to significantly improve survival and the mechanism by which these medications appeared to improve survival was by controlling disease activity ^{2, 92, 93}. As patients began surviving longer the cumulative toxicities of the medications became apparent. Death due to severe infections, myelosuppression and malignancy as well as treatment associated infertility became recognized problems ^{94, 95}. Much of the focus of research in AASV then changed from maximally suppressing the patient's immune system to balancing adequate immunosuppression with reduced treatment related toxicity.

Potential Endpoints for Trials in AASV

Commonly considered measurements for clinical trials in AASV include death, ESRD, doubling of creatinine or a reduction of glomerular filtration rate (GFR), the successful induction of clinical remission, relapse of clinical disease or changes in proteinuria, hematuria, ANCA titre or markers of inflammation such as the erythrocyte sedimentation rate or C-reactive protein level. Some of these are difficult to classify as either clinical or non-clinical in nature (Figure 1).

In the case of “relapse of clinical disease”, in favor of considering relapse a clinical endpoint is that it may be associated with discomfort, disfigurement or damage to vital organs ^{96,97}. It could therefore be considered important to patient function and quality of life on these grounds. Additionally, patients assessed as having relapsing disease are often treated with aggressive and toxic medications which may impact on patient’s quality of life and are associated with an increased risk of serious infection ⁹⁸. However, the assessment of relapse typically involves assessing disease activity based on clinical judgment which is inherently subjective. Relapsing disease also occurs on a tremendous clinical spectrum from almost imperceptible changes in symptoms or laboratory measures to dramatic and life threatening presentations. These factors make relapse of disease difficult to categorize as a clinical versus a non-clinical endpoint.

Contemporary Clinical Trials in AASV

RCTs were a rarity in AASV until the 1990’s when small trials were reported evaluating the efficacy of low toxicity alternatives to daily oral cyclophosphamide and alternative medicines to maintain remission ⁹⁹⁻¹⁰⁴. In the early 1990’s the European Vasculitis Study Group (EUVAS) identified areas of controversy in the treatment of AASV and subsequently designed and implemented a series of RCTs comparing differing regimens/routes and immunosuppressive medications ^{73, 105}. These trials were largely powered as non-inferiority trials for less toxic medications using clinical remission and relapse of disease as target endpoints. A single EUVAS trial evaluated the use of a therapy, plasma exchange, on the clinical endpoints of death or dialysis dependency ¹⁰⁶. At a similar time in the United States, the use of adjuvant biological therapy for vasculitis

was evaluated in a blinded, placebo controlled trial ¹⁰⁷. This trial utilized a sustained clinical remission and relapse rates as primary endpoints.

The classification of remission in studies of AASV was frequently based solely on the clinical investigators judgment. To make assessments more objective and allow the gradation of severity of disease activity, clinical scores, such as the Birmingham Vasculitis Activity Score (BVAS) are employed in contemporary trials ¹⁰⁸. These scores attempt to use standardized definitions to describe clinical features of the disease and determine the overall activity. Cut-off points based on the activity score are used to classify patients as having active or inactive disease. The determination of these cut-off points is largely arbitrary, however. Additionally, generating the scores is marked by inter-observer variability ^{96, 109}. Remission defined by any clinical score has not been subjected to surrogate endpoint validation.

The concept of disease activity as a surrogate is, however, quite appealing. The immune system is based on systems of redundant mechanisms. Effective therapies commonly impact several mechanisms. A comprehensive clinical activity index may represent the final common pathway for the impact of a therapy on the disease ¹¹⁰. It also seems plausible that additional disease activity will result in additional and accrued target organ damage which likely contributes to all-cause mortality. Also, the treatment of clinically apparent activity is generally associated with increased exposure to toxic medication which may in themselves predispose patients to serious infections and extra risk of malignancy. Thus, therapies which induce a clinical remission as judged by an activity

score, may capture the majority of factors contributing to the effect of a therapy on AASV and mortality.

Disease Activity

The Birmingham Vasculitis Activity Score

Scoring utilities also present many challenges. As mentioned above, their accuracy and reliability rely on the judgment of clinicians generating the score. This may result in varying sensitivity and inter-observer reliability of a given scoring system. Although the scores should represent only the clinical manifestations directly attributable to active AASV, the signs and symptoms of infections and scars resulting from previously active AASV are often difficult to distinguish from active AASV ¹⁰⁹. The optimal scoring system and score cut-points to classify remission are debated.

The Birmingham Vasculitis Score (BVAS) was first published in 1994 and is considered the standard for disease scoring in AASV (Appendix 1) ¹¹¹. The BVAS was first validated in specialty clinics and consists of 66 clinical items grouped by 9 organ systems. Both individual items and each organ system have been weighted according to a perceived clinical value ¹⁰⁸. These weights were not derived nor have they been validated against long-term prospective data but in short term prospective data and retrospectively derived BVASs are associated with all cause mortality ^{108, 112}. Importantly, the BVAS requires that all points generated are due to active AASV. Given the overlap between active AASV, damage due to previous AASV and infection, it takes considerable clinical judgment and experience to appropriately use the BVAS. Due to these difficulties, it has

been widely recognized that BVASs have substantial inter-rater variability and patients with chronic damage from previously active AASV tend to be over-scored ¹¹³.

Other Disease Scoring Utilities

Other disease scores have been used in vasculitis. The Disease Extent Index (DEI), developed in Germany, gives one point for each organ involved with no preferential weighting ¹¹⁴. The organ systems scored in the DEI include all organs represented by the systems in the BVAS so that the DEI can be calculated from the BVAS component scores (any system that does not score zero would receive one or two points on the DEI depending on the system). The Five Factor Score (5F Score), which also gives one point for each vital organ involved, was developed in classical and microscopic polyarteritis nodosa and Churg-Strauss disease in France ¹¹⁵. The 5F score was shown to have prognostic value but has not been used extensively in trials outside of France and remission and relapse can only be very crudely scored. Lastly, organ scarring producing permanent dysfunction, known as damage, has also been given a scoring utility called the Vasculitis Damage Index (VDI) ¹¹⁶. Sixty-four symptoms or signs of organ dysfunction permanently affected (i.e. at least 3 months) by vasculitis are scored with this utility. Although the VDI suffers from the same difficulties with reliability and attributability as the BVAS and is undergoing revisions, it was used extensively in AASV trials.

Relapse as a Surrogate Endpoint in AASV Deserves Further Study

The development of medications for use in autoimmune disease is a growing field. Given the rapid proliferation of new medications proposed for the treatment of AASV, there are likely to be several evaluative trials for these medications in the near future. The validation of a surrogate endpoint for such trials would greatly enhance the appropriate evaluation and licensing of new medications. Conversely, the rejection of commonly used endpoints due to a lack of surrogate validity would greatly inform clinical trialists in this area for future trial design. I propose to examine the validity of relapses within 18 months of treatment as defined by the BVAS as a surrogate endpoint for the composite clinical endpoint of ESRD or death in randomized trials of immunosuppressive therapy in AASV using the long-term outcome data available for the EUVAS clinical trials.

METHODS

Hypothesis

A relapse of AASV, as defined the BVAS, within 18 months of enrollment in a clinical trial is not a valid surrogate endpoint for the composite clinical endpoint of ESRD or all-cause mortality.

The overall work plan is summarized in Figure 6. This study uses Prentice's criteria as a framework to test the validity of relapses within the first 18 months of study as a surrogate endpoint for the composite clinical endpoint of ESRD or death at any time during study follow-up.

Data Sources

Data from clinical trials completed by the EUVAS were utilized for this study. This data included both published and unpublished randomized control trials of immunosuppressive medications (see Table 2). Trials under the short titles of: CYCAZAREM, NORAM, and CYCLOPS were included^{73, 117-119}. These studies broadly tested the efficacy of a traditional regimen of oral cyclophosphamide on the endpoints of the induction of remission and the prevention of relapse compared to less toxic alternatives in incident patients with AASV. All three trials used the BVAS to grade disease activity. All three studies were of 18 months duration with a long-term observational follow-up period (mean ~ 5 years). I used relapse data up to 18 months. Very few clinical endpoints were encountered during the 18 months of study while up to 20% of participants developed a clinical endpoint during the extended observational period. I therefore used both the original 18 month trial period and long-term observational period for data on the clinical endpoint. Other data elements included in all databases include participants' date of birth, sex, and longitudinal data on the presence of ANCA, abnormalities in urinalysis, renal function, or proteinuria and subtype and treatment type, date and dose.

CYCAZAREM

CYCAZAREM tested the hypothesis that incident patients treated with long-term cyclophosphamide (CYC) had superior disease control to those treated with a 'weaker'

immunosuppressive, in this case azathioprine (AZA) in patients with moderate AASV¹²⁰. This trial was powered to demonstrate equivalence. Patients in this study were enrolled and treated with a uniform induction regimen of oral CYC and oral prednisolone. Patient who achieved remission were then randomized to either continue receiving oral CYC or to AZA in addition to prednisolone. These maintenance of remission medications were continued until the conclusion of the study. 155 patients were enrolled and 143 reached remission within 6 months and were subsequently randomized. All patients were followed for 18 months after the time of enrollment. There was no significant difference in the time to relapse or proportion of relapses at the end of the study period (10 in the CYC group vs 11 in the AZA group). Only one death occurred after randomization.

NORAM

NORAM tested the hypothesis that induction treatment with oral CYC and prednisolone is more effective at inducing and maintaining remission than methotrexate (MTX) and prednisolone in patients with mild AASV (defined as no vital organ involvement)¹²¹. Both groups had all AASV related medications (MTX, CYC and prednisolone) discontinued at 12 months. 100 incident patients were randomized in this study and followed for 18 months. Remission rates did not differ on the basis of treatment assignment (89.5% of those treated with MTX and 93.5% of those treated with CYC). Patients treated with MTX had a shorter time to relapse with 69.5% relapsing by 18 months compared to 46.5% in the CYC group. Two patients in each group died.

CYCLOPS

CYCLOPS tested the hypothesis that intravenous pulses of CYC were as effective at inducing remission and maintaining remission as oral CYC in patients with moderate AASV ¹¹⁷. Both groups were given 3 months of intensive CYC therapy followed by 3 more months of less intensive CYC therapy (a reduction in the frequency of pulses for the intravenous group and a reduction in the dose for the oral group) before being converted to an indefinite period of AZA and prednisolone. 136 of a planned 150 patients were recruited and followed for 18 months. The results of the study have not been published but the preliminary analyses have been completed. Preliminary results suggest there is no difference between the intravenous and oral groups with respect to the induction of remission (96% versus 93% respectively). Additionally, although the study is likely to be published as demonstrating non-inferiority of the intravenous regimen at maintaining remission, approximately double the relapse events were noted in the intravenous group (13 versus 6 in the oral group). The total number of relapses was well below expectations and thus the statistical tests were much lower power than anticipated. Of interest, although there were more relapses in the intravenous group, only 5 patients died compared to 8 patients in the oral group during the initial study period.

Defining the Clinical Endpoint

The clinical endpoint of interest for this study will be the development of the composite endpoint of ESRD or all-cause mortality over the entire patient follow-up, including the extended observational period beyond the originally reported RCTs. ESRD was defined

as the permanent need for renal replacement therapy such as hemodialysis, peritoneal dialysis or renal transplantation. This endpoint is recognized as having a significant impact on patient function and survival and thus is a valid clinical endpoint ¹¹.

Additionally, while the impact of AASV on the renal system was a common cause of death in the past, renal replacement therapies for patients with ESRD have delayed death due to renal dysfunction. The effects of AASV on other organs such as the lungs and heart do not have readily available organ replacement therapies to delay death. For these reasons, ESRD and death are appropriate to consider as a composite endpoint.

Defining the Surrogate Endpoint

Primary Surrogate Endpoint: Protocol Relapses

No well defined putative surrogate has emerged in the literature for trials in AASV ¹²².

Events/biomarkers of potential importance include the successful induction of remission (at any time), the time to induction of remission, the occurrence and timing of relapses of disease after initially inducing remission and the loss of ANCA positivity. For this study, the occurrence of a relapse within 18 months of onset of study constituted the putative surrogate endpoint for all analyses. The time period of 18 months is of considerable interest as the included studies all used 18 months of study as the primary analysis time-point and have thus proven it is feasible to perform 18 month studies. The occurrence of a relapse of disease is likely of biologic significance as it is associated with increased disease activity and therefore the potential accrual of organ scarring and exposure to toxic therapies to control disease. The definition of a relapse of disease is frequently based on the BVAS. The score which defines relapse has, however, varied between studies and has largely been determined by expert opinion. I considered the protocol definition of a

relapse as the primary surrogate endpoint (protocol relapse). This definition was harmonized for EUVAS trials and endpoints were adjudicated for each trial. A relapse was defined as the recurrence or first appearance of at least three BVAS items or the recurrence or first appearance of at least one of the 24 BVAS items that indicate threatened function of a vital organ (i.e. renal involvement, lung hemorrhage, cardiac involvement, major gastrointestinal involvement, or neurological involvement) attributable to active vasculitis.

Secondary Surrogate Endpoints

Four alternative definitions of a relapse using BVAS criteria were also evaluated (Table 3). The first was to investigate protocol defined major relapses (i.e. major protocol relapse). This definition was identical to protocol relapses except that only those relapses that affected vital organ function were considered a major protocol relapse (i.e. those that affected three BVAS items but did not include a vital organ were excluded). A second alternative definition was based on a threshold value of the overall BVAS (i.e. any score greater than the threshold defined a relapse). The threshold value was to be chosen on the basis of a natural inflection point seen in spline analysis but as no such threshold existed, the 75th percentile of scores was chosen as an *a priori* alternative. These relapses are referred to as “peak BVAS relapses”. The third was to redefine a BVAS based on the association of BVAS items with the clinical endpoint. This reweighted score was then investigated for a threshold value that would define a relapse. These relapses are referred to as “weighted BVAS relapses”. This was meant to optimize the BVAS information as a surrogate by finding a BVAS definition of a relapse that was strongly associated with the

clinical endpoint, a procedure closely related to the first operational criteria Buyse and colleagues proposed in their meta-analytical framework for validating surrogate endpoints (i.e. ensure there is an association between the surrogate and the clinical endpoint)²⁶. Finally, relapse will be defined by the occurrence of renal manifestations of AASV within 18 months of study. Specifically, a renal relapse is defined by at least two of the following: 1) increase in proteinuria defined by BVAS question R2 ($>1\text{ g}/24\text{ hours}$), 2) increase in haematuria of at least 10 cells/high powered field and/or 3) increase in creatinine by 20%.

The associations between surrogate and clinical endpoints were assessed first by categorizing both the surrogate endpoint (presence of relapse within 18 months of study) and the composite endpoint at any time as binary variables in the combined data set of all trials. No adjustment was made in these analyses for the non-independence of participants in the same/different trials (i.e. trials were not considered to act as a higher level of organization of data as in subsequent analyses). The relative odds of developing the clinical endpoint for those who developed the surrogate endpoint compared to those who did not along with the corresponding confidence intervals were used to quantify the association and to test the significance of the association. P values less than or equal to 0.05 were considered significant. Data were also stratified by trial to ensure homogeneity of the point estimates across trials and this absence of heterogeneity was characterized with the Mantel-Hanzel test.

Determining Thresholds for Peak BVAS Relapses

The association between peak BVAS and the composite clinical endpoint were initially tested using GLLMs with the highest overall BVAS at any study visit for each patient as a predictor of death or ESRD. The distribution of the peak scores for patients who reached a clinical endpoint were compared to those who did not using box plots and measures of central tendency (mean and median). Gross differences in the distributions of these scores were assessed to aid in the identification of any naturally occurring cut points in the BVAS that identify a relationship between the scores and the clinical endpoint. Utilizing median spline analyses, the peak BVASs were separated into 6 segments using 5 knots. Each segment was compared to the sum of the previous segments with respect to the association between that segment's BVAS and the composite clinical endpoint in logistic regression. These spline specific odds ratios were plotted against their corresponding BVAS value to assess a change in the odds ratio. The Wald statistic was used to compare the odds ratio of one segment to the preceding segments. The first segment that is statistically different from the preceding ($p \leq 0.05$) will be considered the threshold value for a relapse based on the total BVAS. *A priori*, I decided that in the absence of a naturally occurring threshold in BVAS by spline analysis, the score that defines the 75th percentile BVAS would be used as the threshold to define a relapse.

Determining Thresholds for Weighted BVAS Relapses

Further exploration was done using a weighted BVAS. A weighted score was generated for all assessments. The assessment with the highest score was used for each patient and in the event that two assessments had the same score, the first assessment was used. The

weighted score was generated by using the organ involvement at each time point as independent variables in a forced entry logistic regression model where the composite clinical endpoint is the dependent variable. Forced entry was used to limit the estimate inflation noted with backwards elimination models^{123, 124}. The beta coefficient of the final models was used to generate the weighted score by adding together each coefficient if the associated organ is affected at a particular assessment. The weighted BVAS were used in spline analysis identical to the method described for peak BVAS relapses above. Relapses were defined by the spline analysis as above the inflection point. As with peak BVAS relapses, in the absence of a naturally occurring inflection, the weighted BVAS corresponding to the 75th percentile was chosen as the threshold to define a relapse. The newly defined relapses based on the weighted BVAS were assessed for their association with the clinical outcome in a logistic regression model using the composite clinical endpoint as the dependent variable and relapse as the independent variable.

Analyses

Descriptive Statistics of Source Data

Patients from all three trials were described to ensure that they were sufficiently similar to justify pooling them. The distributions of data within each trial data set and the aggregate data set of all trials were initially explored. Typical demographic factors, factors of known prognostic importance in AASV and other factors thought to reflect the disease process were included. Continuous variables such as age, eGFR, C-reactive protein titre, and baseline BVAS were described with box plots, histograms and measures

of central tendency suitable to the distribution of the data. Categorical variables such as gender, type of ANCA and organ involvement were assessed as fractions. It was assumed that significant clinical heterogeneity between trials would exist with respect to renal function as it was used as a stratifying variable during trial design (i.e. patients with poor renal function were not eligible for the NORAM study). Factors significantly different between studies were used as covariables in the final validation models.

The number of surrogate and clinical endpoints were tabulated for each trial and each treatment arm and for the aggregate of all trials. Again, heterogeneity between trials was expected given differences in their target populations.

Analyses of the Primary Hypothesis

No single method used to assess the validity of a surrogate endpoint dominates this field of biostatistics. Given the data came from three independent but related trials, a meta-analytic approach seemed natural. The literature surrounding the meta-analytic approach is dominated by theories of Buyse and colleagues and includes approaches to continuous linear data, time-to-event data and count data surrogate endpoints with time-to-event clinical endpoints^{26, 29, 30}. Conspicuously, approaches to dichotomous endpoints are not well described in the meta-analytic framework by Buyse and colleagues, possibly due to the paucity of accessible meta-analytic tools available for this type of data and the difficulty in interpreting covariance structures for these data. I therefore relied on hypothesis testing as originally suggested by Prentice's criteria but adapted the hypothesis tests to a meta-analytic framework using generalized linear modeling.

In considering a meta-analytical approach, the appropriateness of pooling the data from both a logical/clinical viewpoint and a statistical viewpoint must be considered.

Clinically, the three primary studies are differentiated by the exclusion of patients with overt renal disease in the NORAM trial as compared to CYCLOPS and CYCAZAREM which do include patients with renal disease. This may have changed the overall phenotype of patients seen between these trials with respect to the type and meaning of relapses. NORAM was noted to have a much higher relapse rate than either CYCAZAREM or CYCLOPS using the protocol defined relapse definitions. This may be a sign of significant between study variability (heterogeneity) of a magnitude greater than can be explained by chance alone. In this case, it could be considered statistically inappropriate to pool these studies. However, if these differences were explained by known variables that differed between studies (e.g. renal function) the use of these variables as independent variables in the regression models would be appropriate. Additionally, the use of random effects models will make allowances for unknown or latent variables that explain differences in between trial relapse rates thus allowing for direct comparisons of the treatment effects across trials if necessary.

Prentice's Criteria: operational definitions

The Use of Multi-level Models

Prentice's operational criteria for a single trial are demonstrated in Equations 1 through 4 above (pages 9 and 10). In the case of a meta-analytic framework the analysis of this

association must consider the potential non-independence of the patients in the trial from one another. In particular, patients in this study come from three different trials and also come from different countries, either of which may create a dependence between observations. Thus, the simple approach in Equation 1 must be adapted to a multi-level approach that allows the model to consider this potential non-independence. For my data, which has a dichotomous clinical endpoint and a dichotomous surrogate endpoint, generalized linear multi-level modeling (GLMM) is probably the most appropriate and flexible. GLMM makes use of a linear predictor equation that is canonically linked to the dichotomous endpoint by a binomial distribution function ¹²⁵. Other linking functions can be used for non-dichotomous outcome data or non-canonical links can be considered for dichotomous endpoints. The dependence of observations within a hierarchical unit is represented by a common intercept for each hierarchical unit that is allowed to differ from other units of the same level (i.e. a random intercept model) ^{125, 126}. It is also possible to allow the slope of lines for each hierarchical unit to vary (i.e. a random intercept model). Each random component to the model has an associated error term and the error of the random components is assumed to follow a standard normal distribution and have a zero mean. The general form of the GLMM is:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \beta_{1j} + \beta_2 x_{ij} + \mu_{1j} \\ \mu_{1j} &\sim N(0, \Omega_\mu) \end{aligned} \quad \text{Equation 5}$$

Where π , a binomial variable, is the dependent variable of patient i in j^{th} trial and μ_{1j} is a randomly occurring, normally distributed error component of the intercept with a mean value of 0 and a variance of Ω_μ . B_{1j} is thus the mean intercept across trials. If the variance of the intercept is high relative to its standard error then it is likely that the grouping of hierarchical units provided a more accurate model than using simple logistic regression.

A Wald test can be applied to this variance parameter assuming it follows a normal distribution.

Prentice's First Criterion

Prentice's first criterion stipulates there must be a significant relationship between the clinical endpoint and the treatment. In this study, an association between the low toxicity treatment and death or dialysis is required. The model specifying this relationship is as follows:

$$\begin{aligned} \text{logit}(T_{ij}) &= \beta_{1j} + \beta_2 Z_{ij} + \mu_{1j} \\ \mu_{1j} &\sim N(0, \Omega_\mu) \end{aligned} \quad \text{Equation 6}$$

Where S is the occurrence of the surrogate endpoint within the first 18 months of study and Z is the treatment allocation to either the low-toxicity medication or standard medication. A relationship will be shown so long as the β_2 coefficient can be shown to be significantly different from zero. Additional independent variable that are of prognostic significance for the clinical endpoint will be included as additional terms.

Prentice's Second Criterion

Prentice's second criterion stipulates there must be a significant relationship between the surrogate endpoint and the treatment. In this study an association between the low toxicity treatment and a relapse by the BVAS is required. The model specifying this relationship is as follows:

$$\begin{aligned} \text{logit}(S_{ij}) &= \beta_{1j} + \beta_2 Z_{ij} + \mu_{1j} \\ \mu_{1j} &\sim N(0, \Omega_\mu) \end{aligned} \quad \text{Equation 7}$$

Where S is the occurrence of the surrogate endpoint within the first 18 months of study and Z is the treatment allocation to either the low-toxicity medication or standard medication. A relationship will be shown so long as the β_2 coefficient can be shown to be significantly different from zero. Additional independent variable that are of prognostic significance for the surrogate endpoint will be included as additional terms.

Prentice's Third and Fourth Criterion

Prentice's third criterion stipulates there must be a significant relationship between the clinical endpoint and the surrogate endpoint. In this study an association between the composite of ESRD or death and a BVAS defined relapse is required. The fourth criterion stipulates that this relationship accounts for all of the effect of the treatment on the clinical endpoint. In the model, this is accounted for by simply adding in a treatment term to the model showing the relationship of the surrogate endpoint to the clinical endpoint. The model specifying this relationship is as follows:

$$\begin{aligned} \text{logit}(T_{ij}) &= \beta_{1j} + \beta_2 Z_{ij} + \beta_3 S_{ij} + \mu_{1j} & \text{Equation 8} \\ \mu_{1j} &\sim N(0, \Omega_\mu) \end{aligned}$$

Where T is the occurrence of the composite clinical endpoint at any time, S is the occurrence of the surrogate endpoint within the first 18 months of study and Z is the treatment allocation to either the low-toxicity medication or standard medication. A relationship will be shown so long as the β_3 coefficient can be shown to be significantly different from zero while the β_2 coefficient cannot be shown to be different from zero. Additional independent variable that are of prognostic significance for the clinical endpoint will be included as additional terms.

Fitting GLMM Models

GLMM models will be fit in Stata version 10 using the GLLAMM command (generalized linear and latent multilevel models) developed by Rabe-Hesketh and Skrondal ¹²⁷. This model fitting procedure allows the specification of dichotomous endpoints with random slope and random coefficient components in the linear predictor as well as specification of the linking function and an arbitrary number of levels. The GLLAMM command uses the method of maximum likelihood to estimate the coefficients of the specified model. Fitting multi-level models with maximum likelihood is often problematic due to the presence of random factors in the likelihood function. GLLAMM integrates out these random variables using adaptive Gaussian quadrature initially before the maximum likelihood. The program allows the specification of the number of integration points for the adaptive quadrature with a default of 8 points in adaptive quadrature. This program also allows the specification of several possible links between the linear predictor and the dependent variable. For the GLMMs, I will use the canonical link function for binomial dependent variables, the logit function.

GLLAMM and multi-level models in general, can be difficult to fit. In the event of a non-converging model, several strategies may be used to enhance the ability of the algorithm to find a fit for the model. Refitting the model using non-adaptive quadrature with an increased number of integration points may resolve issues of non-convergence during the initial phase of the model fitting procedure. Refitting the model with random coefficient variables as fixed coefficient variables may simplify the model fit as may reducing the

number of model parameters overall. Model fitting can also be more difficult in situations where there are relatively few higher order hierarchical units. This is a concern in the data as there are only three study level units. Independent predictor variable for each outcome will be added to the assessment of each of Prentice's criteria. *A priori* these will include age and renal function (expressed as estimated glomerular filtration rate [eGFR]) at baseline, sex and diagnostic subtype (i.e. Wegener's granulomatosis or microscopic polyangiitis). All independent variables will be retained in all models (i.e. forced entry) to avoid estimate inflation associated with the selection of variables based on univariate significance tests. Once the appropriate model has been fit, the results of a given coefficient will be considered significant if the corresponding p-value on a Wald test is less than or equal to 0.05. No model diagnostics will be performed as the GLLAMM model has few assumptions and there are no well accepted model diagnostic tests.

RESULTS

Patients

Of the 391 patients enrolled in the three EUVAS trials, 360 reached a remission: 87/100 in NORAM, 143/155 in CYCAZAREM, and 130/136 CYCLOPS, and 294 (82%) had both long term follow-up data and detailed BVAS data available for the 18 months of primary follow-up (Figure 7). Notably, patients with missing detailed BVAS data still had the presence/absence of protocol defined relapses noted. Patients lost to follow-up were more likely to have MPA, and had a lower eGFR than those included reflecting higher loss to follow-up rates in the CYCAZAREM and CYCLOPS trials (Table 4). Between the three trials, patients in the NORAM trial were younger, more likely to have

WG, and had a higher eGFR than those in CYCAZAREM or CYCLOPS. The median overall follow-up time for clinical outcomes was 6 years ranging between 4 and 9 years with the longest follow-up times in CYCAZAREM and the shortest follow-up times in CYCLOPS.

Organ system involvement at study entry indicated by baseline BVASs is summarized in Table 5. Patients in NORAM almost universally had ENT involvement, a function of the high proportion of WG patients, compared to less than half of patients in CYCAZAREM and CYCLOPS. Conversely, almost all patients in CYCAZAREM and CYCLOPS had renal involvement while only one-third of those in NORAM did, again, a function of the entry criteria for each trial. Involvement of the remaining organ systems was similar between trials.

Clinical Endpoints

53 of 294 evaluable patients developed the composite of ESRD or death with a median time of 2.7 years (IQR 1.1 – 4.6 years). 23 patients developed ESRD with a median time to ESRD of 1.7 years (IQR 0.5 – 4.0). 37 patients died with a median time to death of 3.3 years (IQR 1.7 – 5.1 years). Individual and composite clinical endpoints were more common in patients from CYCAZAREM and CYCLOPS compared to NORAM (Table 6). The composite endpoint was more common in patients with MPA (24%) compared to WG (15%) although this was likely due to significantly lower baseline eGFR in the MPA group (median 25 vs 65 ml/min/1.73 m²; $p<0.001$) leading to a higher incidence of ESRD in patients with MPA (15%) compared to WG (6%; $p=0.01$) as compared to death

(15% MPA vs 11% WG; $p=0.17$). Patients who reached the composite had a median age at study entry of 66 years (IQR 54 – 72) compared to 54 years (45 – 64) for those who did not ($p<0.001$). The median baseline eGFR was 26 ml/min/1.73m² (IQR 15 - 49) year in patients who reached the composite compared to 55 (IQR 27 - 79) years in those who did not ($p<0.001$). These differences were consistent between trials although there were too few events in the NORAM trial to make a meaningful comparison to CYCAZAREM and CYCLOPS. Differences in baseline variables were also consistent for both components of the composite clinical endpoint (Table 7).

Baseline characteristics associated with an increased risk of death or ESRD included age ($p=0.05$), sex ($p=0.04$), and eGFR ($p=0.001$). There was insufficient evidence that diagnostic subtype was associated with an increased risk of death or ESRD ($p=0.07$) (Table 8). A significant interaction between age and baseline eGFR was found ($p=0.002$). Lower levels of baseline eGFR were associated with a blunting of the increasing risk due to increasing age (i.e. patients with very poor renal function had a high risk of death irrespective of age) (Figure 8). Similarly, the risk conferred by increasing age was blunted by very low baseline eGFR to the point that very old patients with very low eGFR were at lower risk of the composite endpoint than very old patients with normal eGFR (Figure 9). These risk estimates, however, were based on very few data points at the extreme of age and eGFR and model fit was poor. Point estimates for the effect of baseline characteristics were largely unchanged when comparing the above model based only on the 294 patients included in the surrogate endpoint analyses to the same model using all patients with long-term data but missing detailed BVAS data (Table 8).

Multi-level models did not suggest there was significant heterogeneity between trials. There was minimal variance in the estimates of the trial level random effects after adjustment for baseline eGFR and age. The p value for the likelihood ratio test comparing the multi-level model to simple logistic regression models was 0.28 in the unadjusted model and 1.0 in the adjusted model and the log likelihood went from -138.5 to -119.2. This combination of changes suggests that the fixed effect components of baseline eGFR and age largely accounted for between trial differences in the composite endpoint.

The Surrogate Endpoint

Protocol Defined Relapses

A total of 85 protocol defined relapses occurred in 81 patients between the three trials. Sixty-four percent (52/81) of patients with a relapse were from NORAM compared to only 16% (13/81) from CYCAZAREM and 20% (16/81) from CYCLOPS patients. Protocol defined major relapses occurred in 47 (16%) patients, of which 27 were in NORAM, 11 were in CYCAZAREM and 9 were in CYCLOPS. Minor relapses occurred in 38 patients, of which 28 were in NORAM, 3 were in CYCAZAREM and 7 were in CYCLOPS. Multiple relapses occurred in 3 patients in NORAM, 1 patient in CYCAZAREM, and 0 patients in CYCLOPS. Protocol defined relapses were more common in WG (70/193 patients) compared to MPA (11/101 patients). This difference may have been driven predominantly by the NORAM trial which accounted for the most relapses and was almost exclusively WG patients. Stratifying by trial, protocol defined relapses continued to be slightly more common in WG patients compared to MPA

patients (Table 9). Patients who had any protocol defined relapse were younger (median 54 years vs 58 years; $p=0.08$), more frequently had involvement of the mucous membranes (e.g. mouth ulcers) at baseline ($p=0.03$), and ENT system ($p<0.001$), and had a higher baseline eGFR (median 73 vs 35 ml/min/1.73 m²; $p<0.001$) than patients who did not have a protocol defined relapse. Baseline BVASs were similar between the two groups (18 vs 16; $p=0.64$). These differences were broadly similar between trials.

Alternatively Defined Relapses

Peak Scores

Fifty-four percent of all patients had no recorded disease activity after the induction of remission (peak BVAS of zero) and 57% of all patients never had a peak BVAS greater than 1. Both CYCAZAREM and CYCLOPS patients were more likely to maintain a peak BVAS ≤ 1 after the induction of remission compared to NORAM patients ($p=0.007$), a finding consistent with the higher incidence of protocol define relapses in the NORAM study. Peak BVAS after the successful induction of remission were similar between trials although scores in CYCAZAREM, the only trial to use oral cyclophosphamide induction in both arms, were slightly lower than NORAM and CYCLOPS ($p=0.02$ for difference between trials by ANOVA). There was no gross difference in peak BVAS between those that reached the composite endpoint and those that did not on either visual inspection (Figure 10) or comparing statistically ($p=0.37$) preventing the assignment of a relapse definition based on these simple analyses and reducing the likelihood of a strong association between peak BVAS and the clinical endpoint (i.e. Prentice's third criterion).

Spline analysis of peak BVASs after remission did not reveal a significant change in the odds ratio of reaching the composite clinical endpoint in crude or adjusted logistic regression (Figure 11). The 75th percentile peak BVAS (peak BVAS>6) was chosen as the definition of a relapse as decided *a priori*. Using this definition, 20% of patients had a relapse, 27% in NORAM, 12% in CYCAZAREM and 21% in CYCLOPS. There was no significant difference in the proportion of relapses experienced by patients with WG compared to MPA (21 vs 17%; $p=0.37$), median GFR for those with and without a relapse (60 vs 43 ml/min/1.73 m²; $p=0.24$) or median age (57 vs 57 years; $p=0.88$).

Weighted Score

Some activity items may be associated with the clinical endpoint while others may not or be associated to a lesser degree. These potential differences in association between components of the BVAS and the clinical endpoint make weighting components of the score important. This was demonstrated in differences in the type of disease activity present after the induction of remission in patients that reached the clinical endpoint compared to those that did not. In patients that reached the clinical endpoint, compared to those who did not, activity in the abdomen and neurologic system was more common ($p=0.003$ and 0.06 respectively) (Table 10). This could imply that relapses in some organ systems are associated with the composite clinical endpoint while high activity scores from others are not associated with the clinical endpoint.

Each scale of the BVAS was determined to be active or inactive at the time of peak BVAS after remission. Activity for each organ system was used as a dichotomous predictor variable for the clinical endpoint as well as baseline eGFR and age in logistic regression (Table 11). The regression coefficients from this logistic regression were then used as a weight and each patient had a weighted BVAS calculated by summing the weights of each organ system that had activity. The weighted BVAS ranged from -1.21 to 3.84 with a median score of 0 (IQR 0 – 0.07) in all patients. Patients who reached the composite endpoint had a median score of 0 (IQR 0 to 0.35) compared those who did not reach the composite (median score 0; IQR 0 to 0.0). The scores between those who did and did not reach the composite endpoint were significantly different ($p < 0.001$).

Median spline analysis of the weighted score as a predictor of death or ESRD demonstrated a constantly increasing predicted probability of the endpoint as score increased. Risk increased continuously with the weighted BVAS (Figure 12). A score > 0.81 define the 75 percentile of scores in patients without the clinical endpoint and was defined as a relapse by this weighted score and designated a weighted BVAS relapse. Twenty-three patients had a weighted BVAS relapse, 7% in NORAM, 8% in CYCAZAREM and 8% in CYCLOPS. There was no difference in weighted BVAS relapses between MPA and WG patients (8% each) nor were there differences in median age (52 vs 58 years; $p = 0.81$) or median baseline eGFR (46 vs 50 ml/min/1.73 m²) in those with and without a weighted BVAS relapse. In logistic regression adjusted for baseline eGFR and age, a weighted BVAS relapse resulted in a 4.5 fold risk of death or ESRD (OR 4.52, 95% CI 1.63 to 12.5; $p = 0.004$). There was no evidence of an interaction

between the weighted BVAS relapse and age ($p=0.24$). The area under the curve of the receiver operating curve for this model was 0.76. The Hosmer-Lemeshow test for goodness-of-fit resulted in a p value of 0.84 demonstrating predicted number of deaths was not significantly different from the observed number.

The Mantel-Haenszel test of the crude association of the weighted BVAS relapse with the composite endpoint did not demonstrate any heterogeneity between trials ($p=0.89$).

Similarly the multi-level model with a random-effects trial level intercept demonstrated a standard deviation of intercepts that approached zero and had a large standard error and the LR test compared to a fixed effects model logistic regression model without a trial variable was not significant ($p=1.0$). When the eGFR and age adjusted model was fit to each trial separately, the point estimates were broadly similar between trials but lack of statistical power in separate trials created extremely broad confidence intervals. The weighted BVAS relapse had the weakest point estimates for association with the clinical endpoint in the NORAM trial and the strongest in the CYCLOPS trial.

Renal Relapses

Twenty-seven (9%) patients had a renal relapse. There was no significant association between this definition of relapse and the composite endpoint in crude analyses ($p=0.27$). Multivariable logistic regression adjusted for age and baseline eGFR did not significantly alter the results of the crude analysis (OR 1.47, 95% CI 0.56 – 3.87; $p=0.43$). It appears unlikely that any candidate definition for a renal relapse will meet the requirements for a surrogate endpoint.

Summary of Putative Surrogate Endpoints

The association of the each putative definitions of a relapse with the composite clinical endpoint of death or dialysis is summarized in Table 12. It is notable that no definition of relapse occurred in the majority of patients who reached the composite endpoint.

Evaluating Prentice's First Criterion – The Clinical Endpoint is Associated with the Treatment

Prentice's first criterion is assessed by examining the relationship between the treatment assignment and the composite clinical endpoint. The treatments were grouped as standard therapy vs low toxicity therapy (3 limbs each). There was no difference in the proportion of patients that reached the composite endpoint when comparing standard to low toxicity therapy (16% vs 20%; $p=0.45$). These results were consistent between trials and between components of the composite endpoints. In multivariable logistic regression adjusted for baseline eGFR, age, sex and diagnostic subtype, low-toxicity treatment was associated with an increased risk of death or ESRD (OR 1.75, 95% confidence interval 0.9-3.4; $p=0.08$). These results were not robust and confidence intervals calculated from bootstrapped models reduced significance levels to $p=0.1$. Additionally, there was evidence of significant confounding by all covariates and the removal of any covariate resulted in a significant reduction in the point estimate for the treatment classification.

There was no evidence of trial level heterogeneity. Multi-level models using trials as a random-effects parameters did not demonstrate significant variance in intercepts and did

not differ from simple logistic regression models without trial level effects (LR test $p=1.0$).

Given the lack of a robust association between treatment and the clinical endpoint, Prentice's first criterion is not fulfilled. Failing to fulfill the first criterion does not make it possible to validate the putative surrogate endpoints. However, this may be a function solely of inadequate power in which case the addition of future studies to the meta-analytic framework may render exploring Prentice's other criteria a useful exercise. Additionally, if the associations seen between the treatment groups and the clinical endpoint are strongly affected by the addition of putative surrogate endpoint that is strongly associated with the clinical endpoint, there would be a suggestion of meeting Prentice's criteria. Furthermore, if all patients (or almost all patients) who experience a surrogate endpoint also reach the clinical endpoint, and no patients (or almost none) who do not reach the surrogate endpoint do not reach the clinical endpoint, the surrogate endpoint will at least be strongly predictive of the clinical endpoint and it is more likely that a treatment that alters the surrogate endpoint would also alter the clinical endpoint.

Prentice's Second Criterion – The surrogate endpoint is associated with the treatment

Protocol Defined Relapses

Low toxicity therapy was associated with an increased proportion of protocol defined relapses in crude analysis (OR 1.93, 95% CI 1.14 – 3.25; $p=0.01$). The crude relationship was consistent between trials with a Mantel-Haenszel test statistic ($p=0.23$). However, a

mixed effects models predicting protocol defined relapse using low toxicity treatment as a predictor suggested significant heterogeneity when trials were included as a random-effects parameter. The standard deviations of intercepts at the trial level were 1.01 with a standard error of 0.43 for low toxicity treatment. The LR tests for models with random effects compared to a simple logistic regression were highly significant ($p < 0.001$). Adding fixed effects for baseline eGFR, age, sex and diagnosis as parameters that were important and different between the trials significantly reduced the between trial variability (standard deviation 0.74 with a standard error of 0.35). In the final model, adjusted for diagnostic subtype and eGFR at baseline, low toxicity therapy was associated with an increased risk of protocol defined relapses (OR 2.27, 95% CI 1.22 – 4.21; $p = 0.009$). These analyses would indicate that treatment is associated with the putative surrogate endpoint of protocol defined relapses but that the strength of this relationship may vary between trials based on unknown (latent) factors.

Protocol Defined Major Relapses

Low toxicity therapy was associated with more protocol defined major relapses than standard therapy in crude analysis (OR 1.92, 95% CI 1.0 – 3.8; $p = 0.04$). There was no significant heterogeneity in the crude association between trials on the basis of the Mantel-Haenszel test statistic ($p = 0.65$) but there was significant trial level variability in the associated random-effects model (LR test compared to simple logistic regression $p = 0.001$). Trial level variability was largely accounted for by differences in baseline eGFR, age and diagnosis as evidenced by a reduction in the standard deviation of the random intercept from 0.61 (standard error 0.30) to 0.38 (standard error 0.29). The LR

test of the random effects model compared to the simple logistic regression model was $p=0.13$. The adjusted OR for low toxicity compared to standard therapy was 2.13 (95% CI 1.08 – 4.22; $p=0.03$).

There is an association between treatment and the putative surrogate endpoint of protocol defined major relapse. The odds of a major relapse were also affected by diagnostic subtype and by unknown trial or patient level differences.

Peak BVAS Defined Relapses

There was no association between low toxicity therapy and peak BVAS defined relapse (OR 0.92, 95% CI 0.52 – 1.63; $p=0.77$). There was marginal evidence of heterogeneity between trial strata in crude analysis with a Mantel-Haenszel test value of $p=0.08$. This was due to a reduced risk of relapse from low toxicity treatment in the CYCAZAREM trial compared to increased risk of relapse in the other two trials. This was also reflected in the random effects model where the standard deviation of the intercept associated with the trials was 0.32 with a standard error of 0.21 and a non-significant LR test ($p=0.09$) for the comparison of the random-effects model to simple logistic regression model without trial level effects. There remained no association between low toxicity treatment and this definition of relapse, however and there was no significant confounding by baseline eGFR, age, or diagnosis.

Weighted Score Defined Relapses

Low toxicity compared to standard toxicity therapy was not associated with a significantly increased risk of relapses defined by the weighted BVAS (crude OR 0.91, 95% CI 0.39 – 2.13; $p=0.83$). There was no evidence of trial level heterogeneity in crude analysis (Mantel-Haenszel test $p=0.97$). Adjusted estimates were not significantly different from unadjusted estimates (OR 0.91, 95% CI 0.38 – 2.17; $p=0.84$). There was no evidence of heterogeneity between trials using a random effects model (between trial intercept standard deviation approached 0 with standard error 0.23) and no difference between multi-level and simple logistic regression models (LR test $p=1.0$). There was no evidence of confounding by baseline eGFR, age or diagnosis.

These results suggest there is no association between treatment and a relapse defined by the weighted BVAS.

Renal Relapses

There was no association between low toxicity and standard therapy and renal relapse in crude analysis (OR 1.23, 95% CI 0.59 – 2.54; $p=0.58$). There was no evidence of trial level heterogeneity in crude analysis (Mantel-Haenszel test $p=0.72$) or from the trial level random-effects intercept in the multi-level mode. There was no evidence of confounding or effect modification by baseline eGFR, age or diagnosis. These results suggest there is no association between treatment type and the putative surrogate of renal relapse.

Summary of Prentice's Second Criterion

Low-toxicity treatment was associated with the putative surrogate endpoint of protocol defined relapses and protocol defined major relapses (Table 13). There was no association between treatment and peak BVAS relapses, weighted BVAS relapses or renal relapses. The lack of congruence between the association between treatment and protocol/major protocol relapses and treatment and peak BVAS/weighted BVAS/ renal relapses may be due to differences in statistical power (there were fewer weighted BVAS and renal relapses than other definitions of relapse), only a very weak association between the treatment and alternate relapse definitions (i.e. treatment had little effect on the potentially more serious relapses associated with peak BVAS or weighted BVAS definitions). Additionally, BVAS scores themselves were not adjudicated as the presence/absence of protocol defined flares were so there may be more noise in the detailed BVAS derived relapse definitions compared to protocol defined flares. Finally, these RCTs were not blinded and the subjective nature in the assessment and ascertainment of relapses may have been biased in protocol defined relapses but such a bias may have been decreased in the more objective detailed BVAS derived definitions.

--

Prentice's Third Criterion – The surrogate is associated with the clinical endpoint

Protocol Defined Relapse

There was no crude association between protocol defined relapse and the composite clinical endpoint (OR 0.83, 95% CI 0.31 – 1.13; $p=0.11$). This association was, however, subject to both confounding by age and baseline eGFR. The crude point estimate changed from 0.83 to 1.63 (95% CI 0.73-3.6; $p=0.23$) after adjustment. There was no significant interaction between age and relapse ($p=0.16$) or between eGFR and relapse ($p=0.58$).

Estimates of the association were similar between the NORAM and CYCAZAREM trials (OR 1.24 and 1.20 respectively) but the estimate for CYCLOPS was much higher (8.65) although the confidence intervals included the point estimates for the other two trials (lower limit 1.20). These broad confidence intervals result from small numbers of events and increase the fragility of the estimates while reducing utility of statistical tests for heterogeneity. Having said that, there was no suggestion of heterogeneity by the inclusion of a random effects intercept for the trials (LR test for difference from model without random intercepts $p=1.0$).

Protocol defined relapses do not appear to be significantly associated with death or ESRD and thus are unlikely to be an adequate surrogate endpoint.

Protocol Defined Major Relapse

There was no crude association between protocol defined major relapse and the composite clinical endpoint (OR 1.49, 95% CI 0.70 – 3.16; $p=0.30$). This association was, however, subject to both confounding and effect modification by baseline factors. The point estimate of the association became 2.45 (95% CI 1.05 – 5.76; $p=0.04$) after adjustment for baseline eGFR and age. There was no interaction between age and major relapse ($p=0.29$) or eGFR and major relapse ($p=0.74$). As with all relapses, the point estimates for the association between major relapse and death or ESRD were consistent between NORAM and CYCAZAREM (2.26 and 1.70 respectively) but much higher and with a very broad confidence interval in CYCLOPS (22.1). Again, however, there was

no suggestion of heterogeneity based on the results of including a random effects intercept for trials (LR test for difference from model without random intercepts $p=1.0$).

The results suggest protocol defined relapses may fulfill Prentice's third criterion and require further consideration for Prentice's fourth criterion.

Peak BVAS Defined Relapse

There was no crude association between peak BVAS relapses and the composite clinical endpoint (OR 1.61, 95% CI 0.80 – 3.21; $p=0.18$). There was no evidence of confounding or effect modification by age ($p=0.40$). There was borderline evidence of effect modification by baseline eGFR ($p=0.07$) (Figure 14). Patients with peak BVAS relapses had little effect on death or ESRD in patients with preserved renal function. In a model adjusted for baseline eGFR and age, a peak BVAS was not significantly associated with the clinical endpoint in patients with a baseline eGFR > 61 ml/min/1.73 m².

Relapse defined by a peak BVAS may be associated with death or ESRD in patients with a low baseline eGFR. This requires further exploration in Prentice's fourth criterion.

Weighted Peak BVAS

The weighted BVAS defined relapse had a strong association with the clinical endpoint in crude analysis (OR 3.32, 95% CI 1.35 – 8.13; $p=0.009$). There was no evidence of effect modification by age ($p=0.93$). There was evidence of possible effect modification by baseline eGFR by significance testing ($p=0.09$). This resulted in an increased risk of

death or ESRD for patients with a baseline eGFR <71 ml/min/1.83 m² that experienced a weighted BVAS relapse but no increased risk in those with a baseline eGFR >71 ml/min/1.83 m² (Figure 15). This interaction is consistent with interaction observed in peak BVAS defined relapses where relapses in patients with low baseline eGFR were associated with death and ESRD but not in patients with higher baseline eGFR levels. Given the majority of NORAM patients had higher eGFR at baseline than patients from CYCAZAREM and CYCLOPS, and patients in NORAM were less likely to die or develop ESRD, these patients may be driving the interaction. Because the interaction is questionable on the basis of hypothesis testing, it is reasonable to report an adjusted OR in the absence of the interaction. The adjusted OR for weighted BVAS relapse is 3.8 (95% CI 1.4 – 10.2; $p=0.007$).

Although the interaction between eGFR and relapse was not seen in regression analyses stratified by trial, the point estimates between trials were broadly similar. This suggests the individual trial data may be underpowered to detect such an interaction. There was no suggestion of between trial heterogeneity based on the results of including a random effects intercept for trials (LR test for difference from model without random intercepts $p=1.0$).

Therefore, there is evidence that a relapse definition based on a weighted BVAS is associated with the clinical endpoint although this relationship is complex due to a potential interaction between relapses and eGFR.

Renal Relapse

Renal relapse was not associated with the clinical endpoint in crude analysis (OR 1.68, 95% CI 0.67 – 4.21; $p=0.27$). This association was not significantly confounded by nor was there effect modification by baseline eGFR or age. There was significant heterogeneity between trials in unadjusted analyses (Mantel-Haenszel test $p=0.02$) but none suggested by the random effects parameters at the trial level in adjusted multi-level analyses. There was no evidence that renal relapses were associated with the clinical endpoint.

Summary of Prentice's Third Criterion

Candidate surrogates that appear to have an association with the clinical endpoint include protocol defined major relapses, relapses defined by a peak BVAS and relapses defined by a weighted peak BVAS (Table 14). The relationship between the peak BVAS and weighted BVAS relapse definition was modified by baseline eGFR.

Prentice's Fourth Criterion – The effect of the treatment on the clinical endpoint is accounted for by the effect on the surrogate

This set of analyses was restricted to the three definitions of relapse that had an association with the clinical endpoint (protocol defined major relapses, peak BVAS and weighted BVAS). The baseline association between the low toxicity therapy and the composite clinical outcome was 1.70 (95% CI 0.90 – 3.21; $p=0.10$). As this is not statistically significant at the predefined level of 0.05 the following analyses can only be regarded as exploratory.

Protocol Defined Major Relapse

The addition of the surrogate endpoint to the model using low toxicity treatment as a predictor of the composite clinical endpoint resulted in a change of the treatment point estimate from 1.70 to 1.64 with a change in the p value from 0.10 to 0.14. The surrogate endpoint point estimate changed from 2.45 (p=0.04) to 2.31 (p=0.06). These represent very mild changes overall.

There is no evidence that protocol defined major relapses are a valid surrogate endpoint for the composite clinical endpoint of death or ESRD.

Peak BVAS Defined Relapse

The addition of the surrogate endpoint to the model using low toxicity treatment as a predictor of the composite clinical endpoint resulted in a change of the treatment point estimate from 1.70 to 1.75 with a change in the p value from 0.10 to 0.09. The surrogate endpoint point estimate changed from 6.45 (p=0.01) to 6.64 (p=0.01) (both calculations assuming a baseline eGFR of 0, the point estimates for the interaction were unchanged by the inclusion of the low toxicity treatment term). These represent essentially no change overall.

There is no evidence that peak BVAS >6 defined relapses are a valid surrogate endpoint for the composite clinical endpoint of death or ESRD.

Weighted BVAS Defined Relapse

The addition of the surrogate endpoint to the model using low toxicity treatment as a predictor of the composite clinical endpoint resulted in a change of the treatment point estimate from 1.70 to 1.88 with a change in the p value from 0.10 to 0.06. The surrogate endpoint point estimate changed from 31 (p=0.007) to 39 (p=0.005) (both calculations assuming a baseline eGFR of 0, the point estimates for the interaction were unchanged by the inclusion of the low toxicity treatment term). These represent essentially no change overall.

There is no evidence that a relapse defined by a weighted peak BVAS is a valid surrogate endpoint for the composite clinical endpoint of death or ESRD.

Summary of Prentice's Fourth Criterion

The inclusion of any putative surrogate endpoint did not change the association between the therapy and the clinical endpoint. This suggests that the putative surrogate endpoints do not capture the causal pathway between either low toxicity or non-oral cyclophosphamide therapy and the increased risk of death or ESRD.

Stability of Models

The modeling process was limited by the small number of outcomes for each model. This can result in fragile results (i.e. small changes in the data by the addition or loss of a few subjects can dramatically alter the effect estimates). This concept was illustrated by bootstrap fitting of confidence intervals. Bootstrap models lead to large fluctuations in confidence intervals and underscored the caution required to interpret any statistically significant results.

DISCUSSION

This study found no evidence that relapse of AASV disease activity, measured using the BVAS, is a valid surrogate endpoint for the composite clinical endpoint of death or end-stage renal disease. This result was consistent across all definitions of relapse despite the use of optimized definitions that were highly associated with the clinical endpoint. Furthermore, the primary definition of relapse, which was used in RCTs, failed essentially all of Prentice's criteria.

Given the small number of clinical endpoint events and the lack of an established association between the treatment and the clinical endpoint, these results may be due to inadequate statistical power rather than a true lack of surrogacy. Irrespective of this limitation, the finding that protocol defined relapses, a commonly accepted RCT endpoint, was not associated with the clinical endpoint should provoke caution in the interpretation of RCTs that use protocol defined relapses as an endpoint. Additionally, this study found a potentially important interaction between severe relapses (defined either by a very high BVAS or a weighted BVAS) and baseline eGFR and age (Figures 14, 15, and 16). This interaction suggests that patients that are more fragile (i.e. older or have poor renal function at trial entry) a clinical endpoint if they have a severe relapse. Conversely, relapses in patients that are less fragile (i.e. younger and/or better renal function at baseline) are less likely to suffer a clinical endpoint if they have a relapse. This could be because patients who have relapses have fewer severe manifestations of disease or because they tolerate the treatment for vasculitis better than fragile patients so have better disease control after a relapse without excess adverse effects of therapy. These findings have not previously been described.

Appropriateness of Pooling Data

The use of pooled data is largely dependent on ensuring that all sources of data are consistent in key areas. In this study key areas include the disease and the treatment. Although the eligible diagnoses were the same for each trial, there was a preponderance of Wegener's granulomatosis in the NORAM trial and MPA in the CYCLOPS trial. Despite these differences in diagnostic subtype, the outcomes in the trials appeared quite

homogenous after controlling for baseline renal function and age on the basis of very little between trial variability in the random effects model. This may be criticized when considering the small number of trials available for analysis given that the trial is the unit of analysis for inter-trial variability. The tests of between trial variability therefore have very low power and would require one of the three trials to be an extreme outlier to detect variability. Similarly, it is readily apparent that the treatment regimens in each trial are not identical even though the principles behind them (ways of limiting cyclophosphamide exposure) are consistent. Small but true differences in the effect sizes between treatment regimes, particularly when working with small data sets, may add significant variability to the pooled effect estimate. This appears unlikely to be the case given the relative consistency of the results of this study within each trial.

Predicting the Clinical Endpoint

Several other studies have described important prognostic factors for the development of death or ESRD in patients with AAV. These prognostic factors have included age, presenting renal function, sex, involvement of the lungs, diagnostic subtype, ANCA pattern, and treatment with only glucocorticoids^{128, 129}. Disease relapses have not been well characterized as a risk factor for death or ESRD.

This study demonstrated that the clinical composite endpoint is associated with several baseline factors and confirmed the associations of female sex, and diagnostic subtype with death and ESRD. I also confirmed that age and renal function at the time of diagnosis are associated with death and ESRD but are complicated by an interaction.

Although this interaction was highly statistically significant, its practical implications are difficult to discern. The interaction suggests that patients with low eGFR are at high risk of the clinical endpoint irrespective of age while those with preserved renal function have a risk that increases with age. This may suggest that poor renal function is the major driving factor in determining the risk of death and ESRD in patients with AASV.

Alternatively, if the majority of clinical endpoints were ESRD then it would be unsurprising that poor renal function was the major determinant of the clinical endpoint. However, death was more common than ESRD as a clinical endpoint and patients that developed ESRD frequently went on to die within the study period.

The Association Between the Treatment and the Clinical Endpoint

Prior to this study, the three included studies were largely interpreted as demonstrating non-inferiority between low-toxicity regimens and standard immunosuppressive regimens; NORAM on the basis of induction remission, CYCAZAREM on the basis of relapses, and CYCLOPS on the basis of induction remission. This study demonstrates that in a model fully adjusted for case mix, low-toxicity treatments may be associated with an increase in death or ESRD. This is particularly important since the number of clinical events in each of these studies separately was inadequate to draw conclusions about the effects of the treatment on clinical endpoints. However, the results must be interpreted with extreme caution in light of the limitations of the study. There was significant missing BVAS data which limited the number of patients for the primary analysis, an important consideration since those with missing data were likely to have more severe disease and therefore were more likely to die. Also, this study only considers

patients who entered a remission. The importance of excluding patients who did not reach a remission and/or did not have complete BVAS data is demonstrated by using patient's baseline and follow-up data irrespective of adequate BVAS data or remission status and refitting the candidate model. In this scenario, treatment type is not associated with composite clinical endpoint (OR 1.37, 95% CI 0.78 – 2.39; $p=0.27$).

The Association Between the Surrogate Endpoint and the Clinical Endpoint

Even after adjustment for the differences in the case mixes enrolled in each trial, the protocol defined relapses were not associated with the clinical endpoint. This is extremely important given these relapses are often regarded as the primary endpoint for studies of AAV. The lack of an association implies that surrogacy is extremely unlikely in the case of protocol defined relapse. Definitions of relapse restricted to more severe relapses, as evidenced either by the designation “major”, or by a high total BVAS, or by the weighted BVAS are associated with the clinical endpoint. One or more of these endpoints may therefore be valid surrogates but require testing in the context of a treatment that reduces death and ESRD in AAV. Although the BVAS has previously undergone forms of validation, it has not been validated as a predictor of death or ESRD and the prognostic value of longitudinal measures has not been performed. This study used limited longitudinal data (i.e. at least one value after a value associated with remission induction). Additionally, the use of spline models to determine a threshold at which the risk of death or ESRD increases to dichotomize BVASs adds legitimacy to the definition of a relapse as a prognostic marker of death or ESRD. Unfortunately, the spline

models did not show unequivocal inflection points in the risk of death or ESRD limiting the utility of the threshold values.

The Association Between the Treatment and the Surrogate Endpoint

The weak association between the treatment and protocol defined relapses was not seen in the individual trials. This novel finding may be due to the meta-analytic framework of the study (i.e. increased statistical power). This association was not maintained, however, when relapse was given a more restrictive definition such as with peak BVAS, weighted BVAS or renal relapse definitions used in this study. Conversely, peak BVAS and weighted BVAS definitions were more strongly associated with the clinical endpoint than protocol defined relapses. Low toxicity may increase the risk of minor relapses which have very little prognostic importance (compared to the relapses defined in a more restrictive sense). Failing Prentice's criteria then would be an expected consequence. Similarly, if the low-toxicity treatment does not substantially increase the risk of relapses that are prognostically important, I would also expect the surrogate to fail Prentice's criteria as they are based on demonstrating superiority.

Treatment Effects Captured by Surrogate Endpoints

The fully adjusted model including both surrogate endpoints and the treatment terms demonstrated these predictors were independently associated with the clinical endpoint rather than the expected reduction in significance of the treatment, when the surrogate endpoint was included. This indicates the two variables likely act on separate causal

pathways towards the composite clinical endpoint. This concept is reinforced by the finding that treatment appears to exert its effect on the ESRD portion of the composite endpoint while relapse appears to exert its effect on the death portion of the composite endpoint.

Interpretation

Protocol defined relapses failed Prentice's criteria. This may be due to the lack of statistical power to demonstrate low-toxicity altered the risk of the clinical endpoint. Low toxicity treatment was expected to have either no effect on ESRD and death or to improve ESRD and death. It is therefore not surprising that no association was found and this may in fact be due to a lack of difference in risk between the treatment groups. In such a scenario, no statistical tests have been developed to validate a surrogate and given the amount of precision that would be required to validate "equivalent" therapies, such a validation exercise seems unlikely as it would require enormous statistical power.

Low toxicity therapy did demonstrate an association with protocol defined relapses. Given the relatively small number of events and concerns over imprecision in the measurement of relapse, this finding could be due to chance. This possibility is reinforced by the fragility of the results as demonstrated by the wide confidence intervals and results with bootstrapped models. The surrogate also failed Prentice's third criteria as it was not associated with the clinical endpoint. This may be due to a very unrestricted definition of relapse which would create noise in the association between the surrogate and the clinical endpoint (i.e. clinically unimportant events may have been recorded as relapses).

Alternatively, the treatment for relapses may be extremely efficacious. One might argue that relapses are important because they lead to death or ESRD in untreated patients but this does not improve their validity as a surrogate endpoint. Because treatment of relapses is the standard of care, the failure of treated relapses to predict ESRD or death would still make it inadequate as a surrogate endpoint since neither the relapse nor the treatment of the relapses were strongly associated with the clinical endpoint.

LIMITATIONS

The primary limitation of this study was a general lack of statistical power despite using the largest database of clinical trial data in AAV in the world. This lack of power was driven by the inability to obtain adequate data from additional French and American studies, the missing data in the available trials, and the relative lack of endpoint events in the data. This is not surprising when one considers the included trials were not designed to test a superiority hypothesis between interventions (although one might assume that given enough patients and time a lower toxicity regimen that is truly non-inferior with respect to efficacy would result in superior clinical outcomes) so the treatments were expected to have very small differences in event rates and therefore be associated with very small effect sizes. Unfortunately, no methods of validating surrogate endpoints have been developed to deal with non-inferiority studies. Given non-inferiority studies rely on reasonably restricted confidence intervals and surrogate endpoint validation methods often suffer from broad confidence intervals when estimating the relationship between the surrogate and the clinical endpoint, it seems unlikely that such methods will be feasible.

Other limitations to consider are that the treatments explored are different in all studies as are the patient populations at least between NORAM and other studies as mentioned above in Section 5.4.2. These factors introduce clinical heterogeneity that may make the pooling of these studies inappropriate. Although the traditional therapy arm remains fairly standardized, the alternative therapy arms are quite different. The argument to combining these trials is that all the alternative arms use immunosuppressive medications that are reportedly less toxic than the traditional arm. The assessment of variance at the study level will help determine to what extent these differences may have affected the results of the analyses.

Although this study used three of the largest RCTs in AASV and used advanced statistical techniques, only approximately 300 patients were analyzed. This is a small study compared to cardiovascular and oncology studies of surrogate endpoint validation and the precision of the estimates reflect this fact.

The difference between BVASs and true disease activity must also be considered when interpreting the eventual results of this study. The intention of the BVAS determined relapses is to detect a significant loss of disease control. The score is an attempt to quantify the magnitude of disease activity at a point in time. The relationship between true disease activity and the BVAS is determined, and limited by, our understanding of the disease and the recording clinician's abilities and acumen. The accuracy of the score may therefore be limited and there may be significant measurement error. The results of this study are necessarily limited to the use of a BVAS derived relapse definition and not

to relapses of disease in general. It should be noted, however, that the BVAS is the current standard for measuring disease activity in clinical trials and this study will reflect the actual practice of clinical trials in AASV.

Lastly, no universally accepted technique for validating a surrogate endpoint or universally accepted definition of validity for a surrogate endpoint exists. The techniques for this study rely heavily on adaptations of the work by Prentice and draw on concepts from later work by Buyse et al. In essence, however, this analysis assesses only Prentice's criteria and is thus guilty of the shortcomings of Prentice's criteria discussed above. A parsimonious finding using at least one approach other than Prentice's criteria would strengthen any findings from this analysis. Future work could confirm this approach.

SIGNIFICANCE AND FUTURE RESEARCH

Future trials of immunosuppression cannot rely on relapse of any definition as the sole indicator of treatment efficacy at the present time. Although modified versions of BVAS defined relapses show promise at predicting clinical endpoints they require rigorous testing in large trials to validate their prognostic significance. In the absence of any known valid surrogate endpoints, future trials in AASV should be powered for a clinical endpoint such as ESRD or death. Within the context of such a trial the evaluation of other putative surrogates is possible.

The design of RCTs for therapies in AASV is still in their infancy. The trials used in this validation study have been groundbreaking and the backbone to the current standard of

care for patients with AASV. This study is unable to offer unequivocal support for the original findings of the included studies based on the disconnect between the protocol defined relapses and the clinical endpoint. Ongoing clinical trials utilizing relapse as a primary endpoint must be interpreted with caution given the uncertainty surrounding the clinical significance of relapses of any definition. Due to the potential interactions noted in this study, the interpretation of such results will be even more difficult as relapses may have very different meanings in different populations. This study should impact the planning of future trials in AASV. Any future trials should incorporate disease activity measurements with the purpose of comparing these measurements to clinical endpoints. With enough clinical endpoint data it may be possible to validate relapses using the models outlined in this study. If relapses can be shown to be a valid surrogate of death or dialysis, future trials may benefit from a reduced number of patients and follow-up time. For example, if a trial were planned to examine the impact of a new therapy on death or dialysis in high risk patients (such as CYCLOPS patients), given 25% are expected to reach the clinical endpoint within 5 years and a powerful therapy would generally be considered to have a 30% relative risk reduction, 2102 patients (1051 per group) may be required. Given protocol defined relapses occur at a rate of 10% per year and assuming the same relative risk reduction, in the same amount of time only 650 patients would be required to arrive at the same answer. This would have major implications on the cost and feasibility of trials in AASV. This may, however, be an unrealistic situation given protocol relapses were not associated with the clinical endpoint. More strenuous definitions of relapse, such as weighted BVAS relapses, occurred in only 8% of this study population in 18 months, a rate of approximately 5% per year and nearly identical to the

death rate for this population. In this scenario, there would be no advantage to using the surrogate endpoint and a significant risk of incorrect conclusions due to inaccuracies in the relationship between the treatment, the surrogate endpoint and the clinical endpoint.

Lastly, this study represents the first individual patient data meta-analysis in the realm of vasculitis. As recently as 10 years ago, RCTs in AASV were thought to not be feasible by many investigators. The emergence of RCTs has elevated the level of evidence for the treatment of AASV. Progressing to individual patient data meta-analyses, a process that requires the cooperation of a network of investigators, represents another step forward in elevating the evidence behind the treatment of AASV and an increase in the cooperation of investigators with an interest in this rare disease – a trend I hope to see continue.

Furthermore, this represents one of the largest collated experiences with long-term follow-up in AASV and will be a valuable addition to the general AASV literature. This also has value as a role model to investigators in other rare diseases demonstrating that RCTs and clinical endpoints can and should be investigated despite the relative paucity of patients available for study.

TABLES

Table 1. Organ involvement often associated with anti-neutrophil cytoplasm antibody associated systemic vasculitis by organ system.

Organ System	Manifestations
General	Fever, weight loss, malaise
Ophthalmologic / Otolaryngologic	Uveitis, retro-orbital granulomata, sensorineural deafness, hydotympanum
Upper Respiratory Tract	Sinusitis, rhinitis, nasal crusting, septal erosions/perforations, tracheal/subglottic stenosis
Lower Respiratory Tract	Bronchial stenosis, cavitating lung lesions, diffuse alveolar hemorrhage, lung masses, retrosternal masses
Nervous System	Mononeuritis multiplex, mononeuropathy, stroke like syndromes
Cardiovascular	Coronary aneurysms, myocardial infarction, myocarditis, pericarditis
Gastrointestinal	Nausea, intestinal ulceration and bleeding
Musculoskeletal	Arthralgia, arthritis, rash (typically leukocytoclastic)
Renal	Rapidly progressive renal failure, nephritic syndrome, abnormal urine sediment

Table 2. Summary of major randomized control trials in AASV by study short title for consideration in the validation of relapse as a surrogate endpoint. With the exception of WGET, all studies were performed by the European Vasculitis Study Group.

Study Name	Patients	Interventions	Endpoint	Endpoint Measure	Duration (months)	Relapses Arm 1/ Arm2	Deaths Arm 1/ Arm 2	Result
CYCAZAREM (2003)	155	Oral CYC vs. AZA after CYC induction	Relapse	BVAS	18	11/10	8	Equivalent at preventing relapse
NORAM (2005)	100	Oral CYC vs. oral MTX	Remission; Relapse	BVAS	18	20/32	2/2	Equivalent at inducing remission; CYC more effective at preventing relapse
CYCLOPS* (2008)	133	Oral CYC vs IV CYC	Remission; Relapse	BVAS	18	7/13	8/5	Equivalent at inducing remission;

CYC = cyclophosphamide, AZA = azathioprine, MTX = methotrexate, BVAS = Birmingham Vasculitis Activity Score. * unpublished data

Table 3. Definitions of surrogate endpoints used for validation.

Surrogate Endpoint	Analysis	Definition
Protocol Relapse	Primary	The recurrence or first appearance of at least three non vital organ threatening BVAS items or the recurrence or first appearance of at least 1 of the 24 BVAS items that are indicative of threatened function of a vital organ attributable to active vasculitis within 18 months of treatment.
Major Protocol Relapse	Secondary	The recurrence or first appearance of at least 1 of the 24 BVAS items that are indicative of threatened function of a vital organ attributable to active vasculitis within 18 months of treatment.
Peak BVAS Relapse	Secondary	The occurrence of an overall BVAS in excess of threshold value within 18 months of treatment. The threshold value is determined by the assessment of association of the overall BVAS with the clinical endpoint.
Rewighted BVAS Relapse	Secondary	The occurrence of a reweighted BVAS in excess of a threshold value within 18 months of treatment. The reweighted BVAS is calculated using weights determined by the association of each organ system to the clinical endpoint. The threshold value is determined by the assessment of association of the reweighted BVAS with the clinical endpoint.
Renal Relapse	Secondary	The occurrence of at least two of the three following BVAS renal items within 18 months of treatment: 1) increase in proteinuria (>1g/24 hours), 2) increase in haematuria of at least 10 cells/high powered field and/or 3) increase in creatinine by 20%.

Table 4. Baseline characteristics by trial and for overall pooled cohort.

	NORAM (n=91)	CYCAZAREM (n=107)	CYCLOPS (n=96)	Total (n=294)	Excluded (n=76)
Median Age (IQR)	53 (40 – 61)	58 (49 – 67)	61 (49 – 69)	59 (47 – 67)	60 (50 – 67)
Sex (% Female)	53	56	41	56	52
% WG/% MPA	94/6	63/37	38/62	61/39	42/58
Median eGFR ml/min/1.73m ² (IQR)	80 (64 – 95)	32 (19 – 63)	32 (17 – 51)	43 (22 – 74)	34 (18 – 65)
Median BVAS (IQR)	13 (9 – 20)	16 (7 – 24)	15 (4 – 22)	15 (9 – 21)	19 (12 – 24)
Median Systems Involved (IQR)	4 (3 – 4)	4 (3 – 5)	3 (2 – 4)	4 (2 – 4)	3 (2 – 4)
Median Follow-up Time (years)	6 (5 – 7)	8.5 (8 – 9)	5 (4 – 6)	6 (4.5 – 8.5)	4.5 (3 – 5.5)

Table 5. Organ system involvement at baseline in European Vasculitis Study Group
clinical trials of ANCA associated vasculitis.

System	NORAM	CYCAZAREM	CYCLOPS	All Trials
Systemic	95	94	90	94
Cutaneous	21	30	23	25
Mucous	42	35	24	34
Membranes				
ENT	85	52	38	59
Chest	55	44	43	47
Cardiac	0	5	7	4
Abdominal	3	2	6	4
Renal	34	97	100	78
Neurologic	21	25	17	21

Table 6. Frequency of the composite clinical endpoint of death or end-stage renal disease (ESRD) over extended trial follow-up and each component of the composite by trial.

Endpoint	NORAM	CYCAZAREM	CYCLOPS	All Trials
Composite	11%	20%	23%	18%
ESRD	1%	10%	15%	9%
Death	11%	14%	13%	13%

Table 7. Characteristics of patients who reached or did not reach the clinical outcomes in trials of ANCA associated vasculitis. Values represent the median (interquartile range) or percent of patients with the characteristi

Trial		Composite		Death		ESRD	
		Yes	No	Yes	No	Yes	No
Age (years)	NORAM	63 (59 – 68)	53 (39 – 60)	63 (59 – 68)	53 (39 – 60)	59 (NA)	53 (40 – 61)
	CYCAZAREM	66 (50 – 72)	56 (48 – 66)	69 (64 – 72)	55 (47 – 64)	45 (36 – 67)	59 (49 – 68)
	CYCLOPS	67 (53 – 72)	60 (48 – 67)	68 (58 – 73)	61 (46 – 68)	54 (38 – 69)	61 (51 – 69)
Female (%)	NORAM	64%	52%	64%	52%	0%	54%
	CYCAZAREM	40%	61%	36%	60%	40%	55%
	CYCLOPS	36%	42%	33%	39%	33%	40%
WG (%)	NORAM	91%	94%	91%	94%	100%	96%
	CYCAZAREM	53%	64%	73%	64%	50%	69%
	CYCLOPS	31%	41%	17%	42%	33%	43%
MPA (%)	NORAM	9%	6%	9%	6%	0%	4%
	CYCAZAREM	47%	34%	27%	36%	50%	31%
	CYCLOPS	69%	59%	83%	58%	67%	57%
eGFR (ml/min/ 1.73m ²)	NORAM	73 (60 – 91)	82 (68 – 95)	73 (60 – 91)	82 (68 – 95)	64 (NA)	82 (68 – 97)
	CYCAZAREM	24 (13 – 33)	36 (21 – 70)	29 (24 – 49)	35 (19 – 64)	19 (13 – 23)	33 (20 – 61)
	CYCLOPS	18 (14 – 29)	36 (21 – 61)	21 (15 – 38)	31 (17 – 47)	15 (11 – 24)	33 (20 – 44)

Table 8. The association of baseline characteristics with the composite clinical endpoint of ESRD or death using multi-level, multi-variable logistic regression. Point estimates are given as odds ratios (95% CI). Odds ratios for age, GFR and BVAS are for each 1 year, ml/min/1.73 m² and point increment respectively. Point estimates for age and GFR must consider the interaction term in models for the composite endpoint. Composite = patients included in primary analyses. Composite (all patients) = patients with long-term data but may be missing detailed BVAS data required for validation studies.

	Endpoint			
	Composite	Composite (all pts)	Death	ESRD
Age	0.95 (0.91 – 1.00)	0.98 (0.95 – 1.02)	1.07 (1.01 – 1.13)	0.95 (0.92 – 0.98)
Female	0.49 (0.25 – 0.96)	0.57 (0.33 – 0.99)	0.52 (0.24 – 1.14)	0.30 (0.10 – 0.89)
MPA	2.0 (0.94 – 4.3)	1.29 (0.70 – 2.3)	1.95 (0.79 – 4.9)	2.07 (0.68 – 6.3)
GFR	0.87 (0.8 – 0.95)	0.92 (0.87 – 0.98)	1.01 (0.99 – 1.02)	0.91 (0.87 – 0.96)
Age*GFR	1.002 (1.00 – 1.003)	1.00 (1.00 – 1.06)	NA	NA
BVAS	1.06 (1.01 – 1.1)	1.02 (0.99 – 1.06)	1.06 (1.02 – 1.11)	1.06 (0.99 – 1.14)

Table 9. Protocol defined relapses for trials of AAV by diagnostic category and trial.

Trial	WG	MPA	Total
NORAM	51/86 (59%)	1/5 (20%)	52/91 (57%)
CYCAZAREM	9/70 (13%)	4/37 (11%)	13/107 (12%)
CYCLOPS	10/37 (27%)	6/59 (10%)	16/96 (17%)
Total	70/193 (36%)	11/101 (11%)	81/294 (28%)

Table 10. Differences in the organ involvement in patients with any disease activity according to the BVAS after the induction of remission according to whether they reached the clinical composite endpoint of ESRD or death. P values computed by chi-squared test.

System Involved	Composite Endpoint Reached		P value
	No (%)	Yes (%)	
Systemic	25	28	0.61
Cutaneous	4	6	0.52
Mucous Membranes	8	8	0.86
ENT	22	13	0.17
Chest	13	17	0.43
Cardiac	1	0	0.51
Abdominal	0.4	6	0.003
Renal	18	25	0.29
Neurologic	6	13	0.06

Note: this table is interpreted as 13% of patients that did not reach the clinical endpoint had at least one item of chest involvement after the induction of remission compared to 17% of patients that did reach the composite clinical endpoint.

Table 11. Results of the logistic regression model using peak system activity as a predictor of death or ESRD used to develop weighted BVAS.

	Univariable	P value	Full Model	P value
	Coefficient		Coefficient	
Systemic	0.43	0.23	0.27	0.59
Cutaneous	0.43	0.55	-0.50	0.58
Mucous	-0.12	0.84	-0.50	0.47
Membranes				
ENT	-0.19	0.68	-0.65	0.28
Chest	0.58	0.19	0.55	0.36
Cardiac	Dropped	NA	Dropped	NA
Abdominal	2.87	0.02	2.64	0.05
Renal	0.29	0.45	0.21	0.65
Neurologic	1.09	0.04	1.14	0.05

Table 12. Summary of the association of each definition of a relapse identified as a putative surrogate endpoint for composite clinical endpoint of death or ESRD. The association of each definition of relapse is presented as the OR for death or ESRD adjusted for baseline eGFR and age

Definition	# with relapse	% Patients with Relapse that Reached ESRD/Death	% Patients without Relapse that Reached ESRD/Death	% Patients with ESRD/Death that had a Relapse	% Patients without ESRD/Death that had a Relapse
Protocol Defined	81	16	18	24	28
Major Protocol Defined	47	23	17	21	15
Peak BVAS	58	24	17	26	18
Weighted BVAS	23	39	16	17	6
Renal	27	27	17	13	8

Table 13. Summary of Prentice's second criterion, the association between treatment and the surrogate endpoint for varying definitions of relapse in ANCA associated systemic vasculitis.

Relapse Definition	Association with Treatment (Adjust Odds Ratio)	P value	Conclusion
Protocol	2.3 (1.2 – 4.2)	0.009	Fulfills
Major Protocol	2.1 (1.1 – 4.2)	0.03	Fulfills
Peak BVAS	0.9 (0.5 – 1.6)	0.77	Fails
Weighted BVAS	0.9 (0.4 – 2.2)	0.84	Fails
Renal	1.2 (0.6 – 2.5)	0.58	Fails

Table 14. Summary of Prentice's third criterion, the association between the surrogate endpoint and the clinical endpoint for varying definitions of relapse in ANCA associated systemic vasculitis.

Relapse Definition	Association with Clinical Endpoint (Adjust Odds Ratio)	P value	Conclusion
Protocol	1.63 (0.7 – 3.6)	0.23	Fails
Major Protocol	2.5 (1.1 – 5.8)	0.04	Fulfills
Peak BVAS	1.9 (0.9 – 4.0)	0.09	Possibly Fulfills*
Weighted BVAS	3.8 (1.4 – 10.2)	0.007	Fulfills*
Renal	1.3 (0.7 – 2.1)	0.42	Fails

* Estimates based on adjustment without consideration of potential interactions

FIGURES

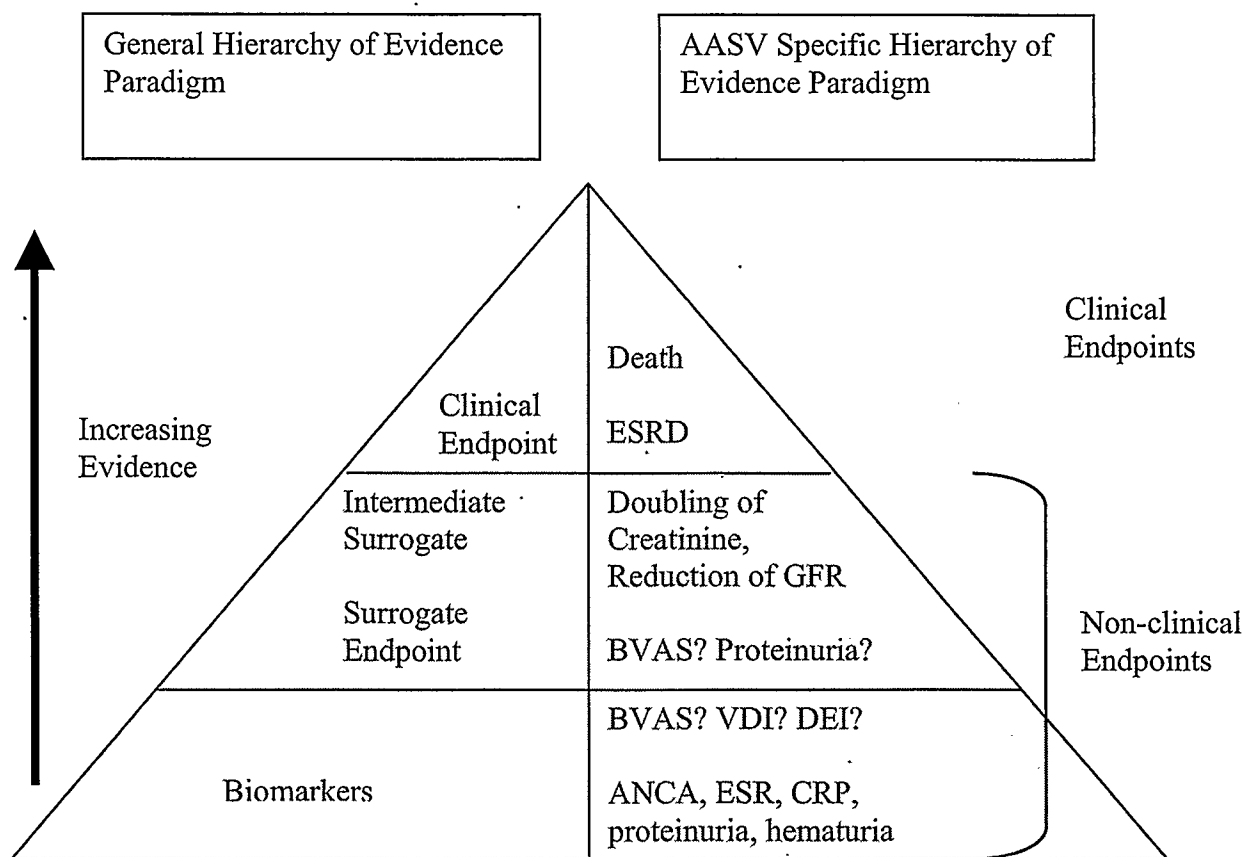


Figure 1. Hierarchy of endpoints in randomized control trials and trials in nephrology ESRD = end-stage renal disease; GFR = glomerular filtration rate; ANCA = anti-neutrophil cytoplasm antibody; ESR = erythrocyte sedimentation rate; CRP = C-reactive protein.

First Criterion ($T_i = \mu_T + \beta Z_i + \varepsilon_{Ti}$):

Treatment is associated with the clinical endpoint

$$Z \rightarrow T$$

Second Criterion ($S_i = \mu_S + \alpha Z_i + \varepsilon_{Si}$):

Treatment is associated with the surrogate endpoint

$$Z \rightarrow S$$

Third Criterion ($T_i = \mu' + \gamma_Z S_i + \varepsilon'_i$): Surrogate endpoint is associated with the clinical endpoint

$$S \rightarrow T$$

Fourth Criterion ($T_i = \mu' + \beta_S Z_i + \gamma_Z S_i + \varepsilon'_i$): the effect of the treatment on the clinical endpoint is mediated by the surrogate endpoint

$$Z \rightarrow S \rightarrow T$$

Figure 2. Prentice's criteria conceptually and operationalized (in brackets). Z = treatment, T = clinical endpoint, S = surrogate endpoint.

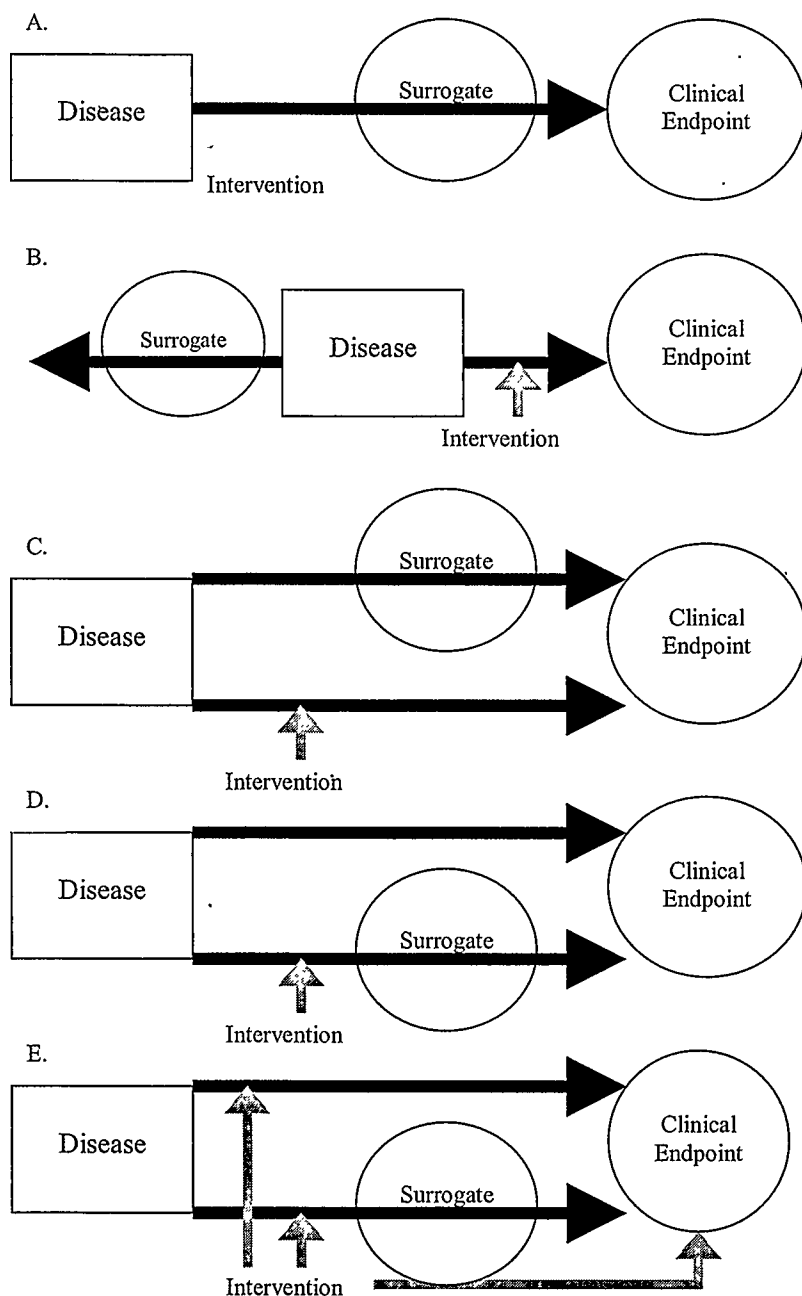


Figure 3 The different ways in which a putative surrogate endpoint may be invalid. Black arrows represent a causal mechanism between disease and an endpoint. Red arrows represent the effects of an intervention or therapy. A) valid surrogate; B) surrogate does not lie on the causal pathway to the clinical endpoint as with biomarkers that are clinical correlates alone; C) the surrogate lies on one of many causal pathways to the clinical endpoint but not the one that the intervention impacts; D) the surrogate lies on the causal pathway to the clinical endpoint affected by the intervention but other pathways are as or more important in determining the clinical endpoint; E) the intervention has many effects on the clinical endpoint, only some of which are captured by the surrogate.

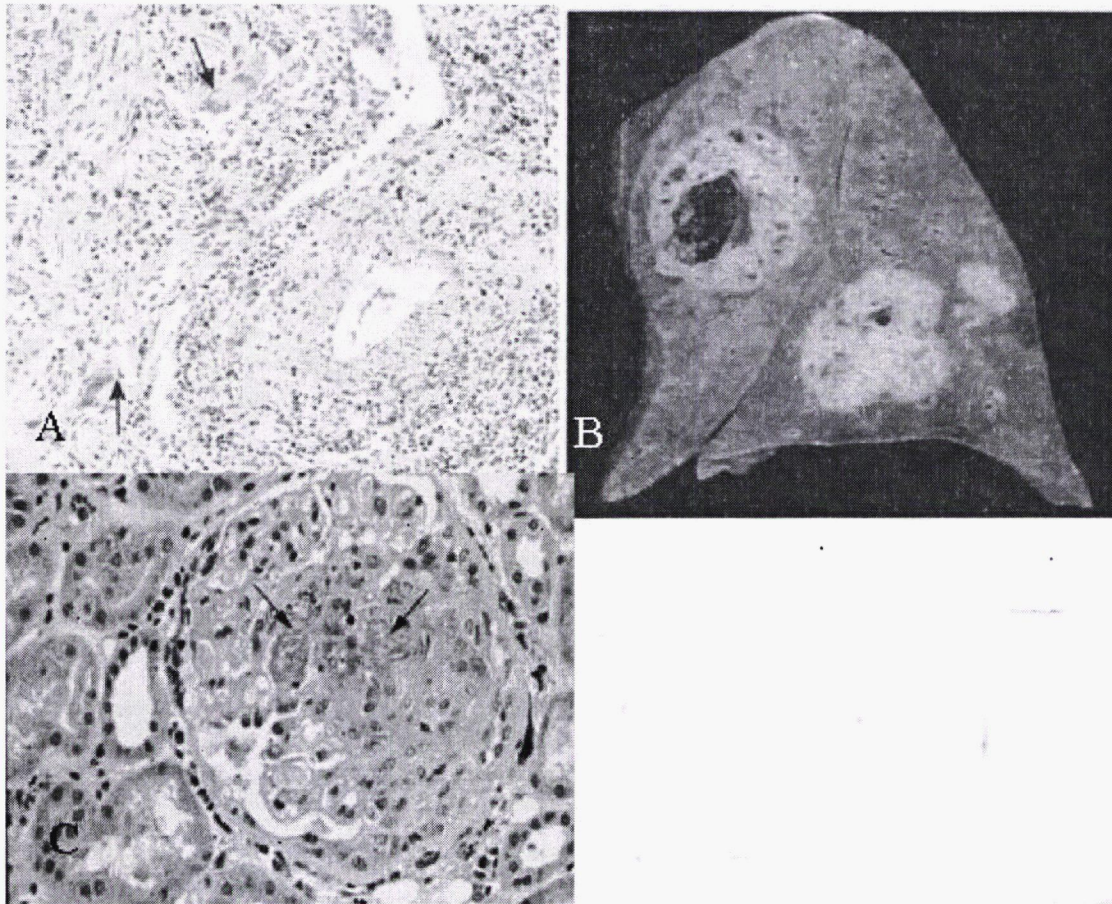


Figure 4 Representative pathology of the ANCA associated systemic vasculitides. A) Histologic sample demonstrating inflammation (vasculitis) of a small artery along with adjacent granulomatous inflammation, in which epithelioid cells and giant cells (*arrows*) are seen. B) Gross photo from the lung of a patient with fatal Wegener granulomatosis, demonstrating large nodular lesions. C) Light micrograph of renal histology demonstrating a fresh segmental necrotizing lesions with bright red fibrin deposition (*arrows*). Adapted from Robbins and Cotran's Pathologic Basis of Disease 7th Ed (A & B) and UpToDate ver 15.2 (C).

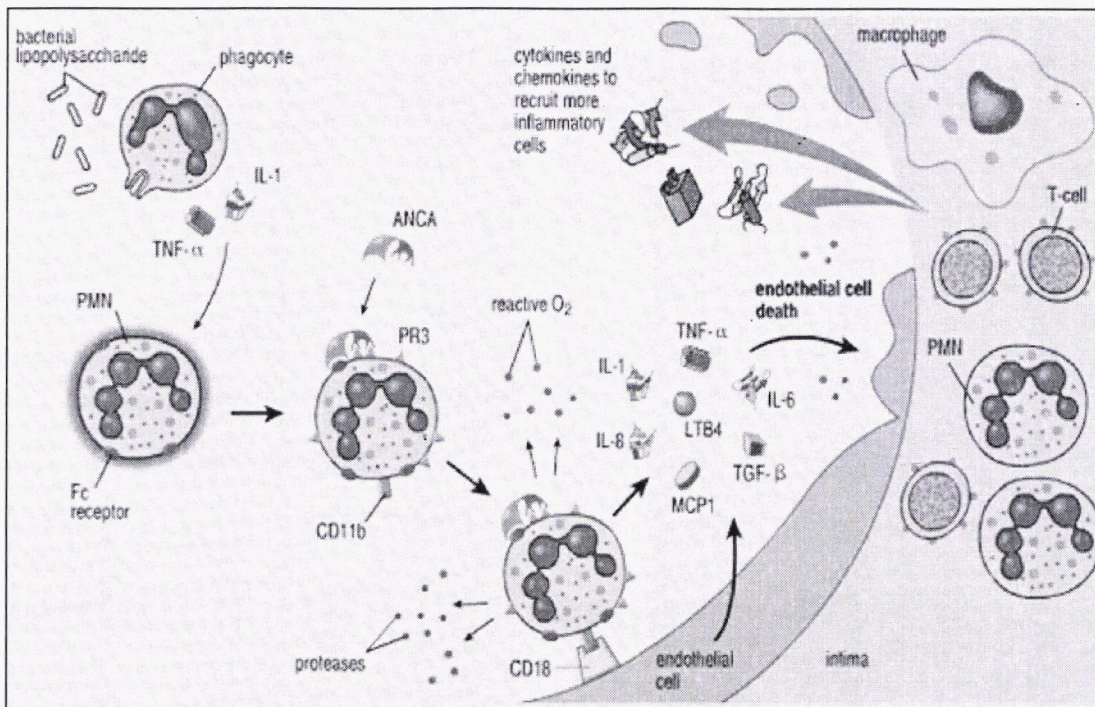


Figure 5 Summary of the proposed pathogenesis of vasculitis. In this model, neutrophils are primed by an antigenic exposure (e.g. bacterial lipopolysaccharides) which causes the expression of cell surface proteinase 3 (PR3). ANCA then bind cell surface PR3 causing an upregulation of pro-inflammatory cytokines (IL-1, IL-8, TNF- α , etc...) and degranulation of damaging proteases and reactive oxygen species. Concurrently, these circulating neutrophils migrate to endothelial walls via CD18-CD11b interactions. Migration to the endothelium is followed by endothelial damage and further upregulation of pro-inflammatory cytokines with subsequent migration of more neutrophils and lymphocytes. (Adapted from *Immune Mechanisms in Rheumatology*, 2001).

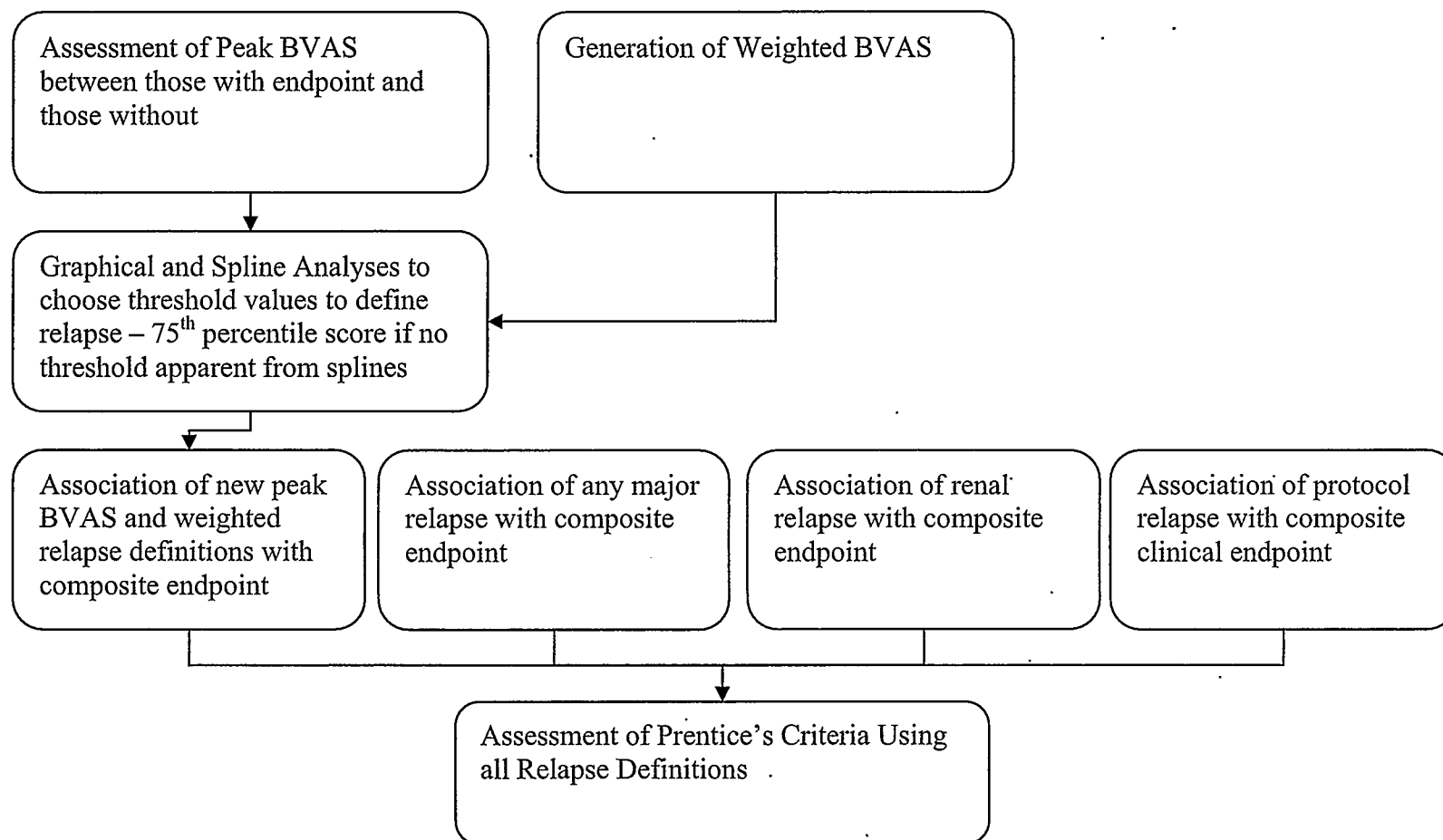


Figure 6 Plan for the assessment of defining relapses with the pooled data set.

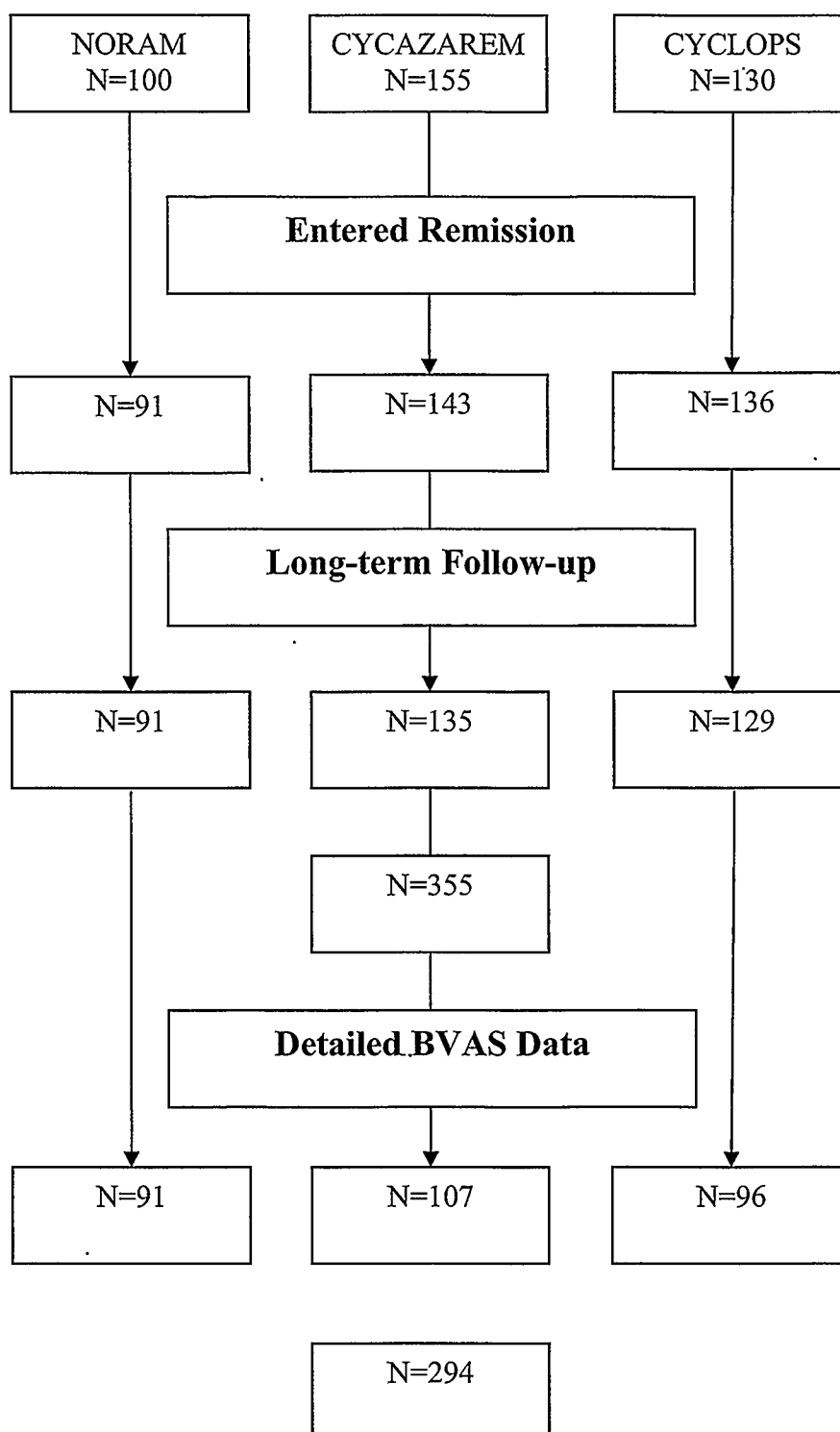


Figure 7 Patient flow through validation study.

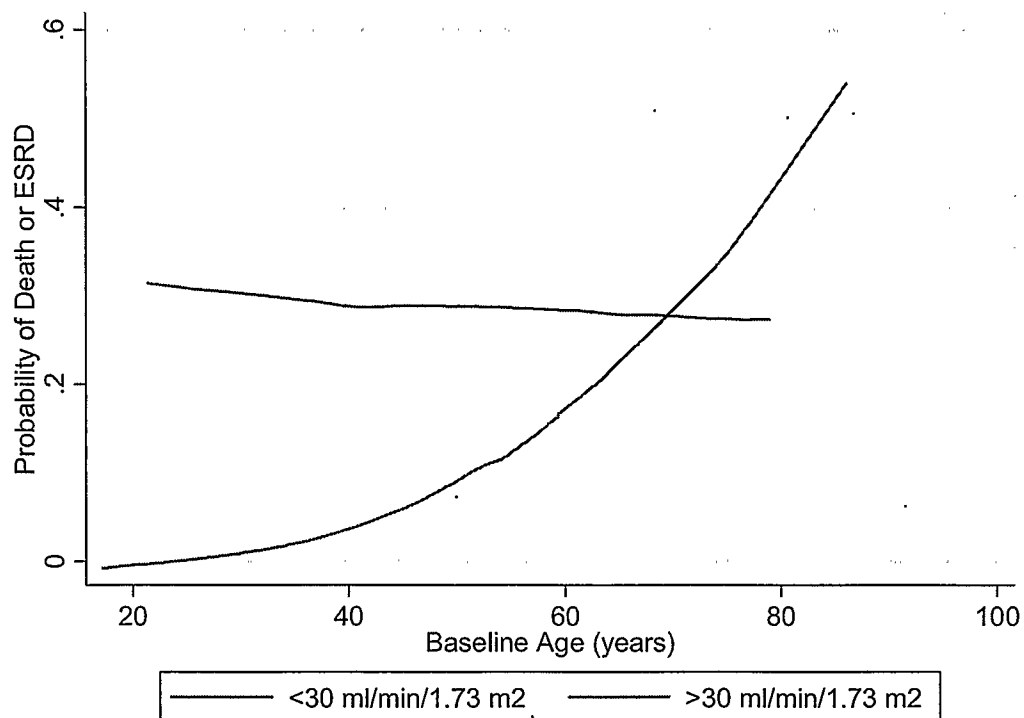


Figure 8 Probability of death or ESRD predicted from a saturated logistic regression model using age and baseline eGFR as predictor variables. A significant interaction between age and eGFR ($p=0.002$) is demonstrated. In patients with preserved eGFR at baseline, age is a large determinant of risk but in patients with a low baseline eGFR, age has little impact on risk.

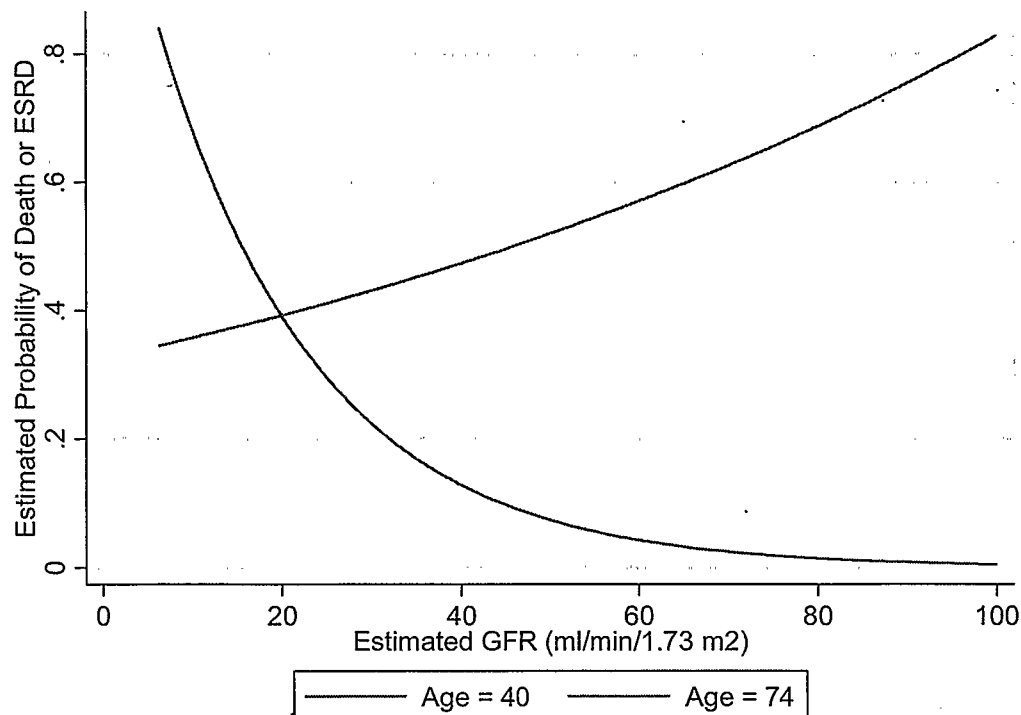


Figure 9 Probability of death or ESRD predicted from a saturated logistic regression model using age and baseline eGFR as predictor variables. A significant interaction between age and eGFR ($p=0.002$) is demonstrated. In younger patients, lower eGFR confers increased risk but in older patients, a low baseline eGFR confers less risk. Few data points exist for patients at the extremes of age and low eGFR, therefore some model instability may exist in probability estimates below 20 ml/min/1.73 m².

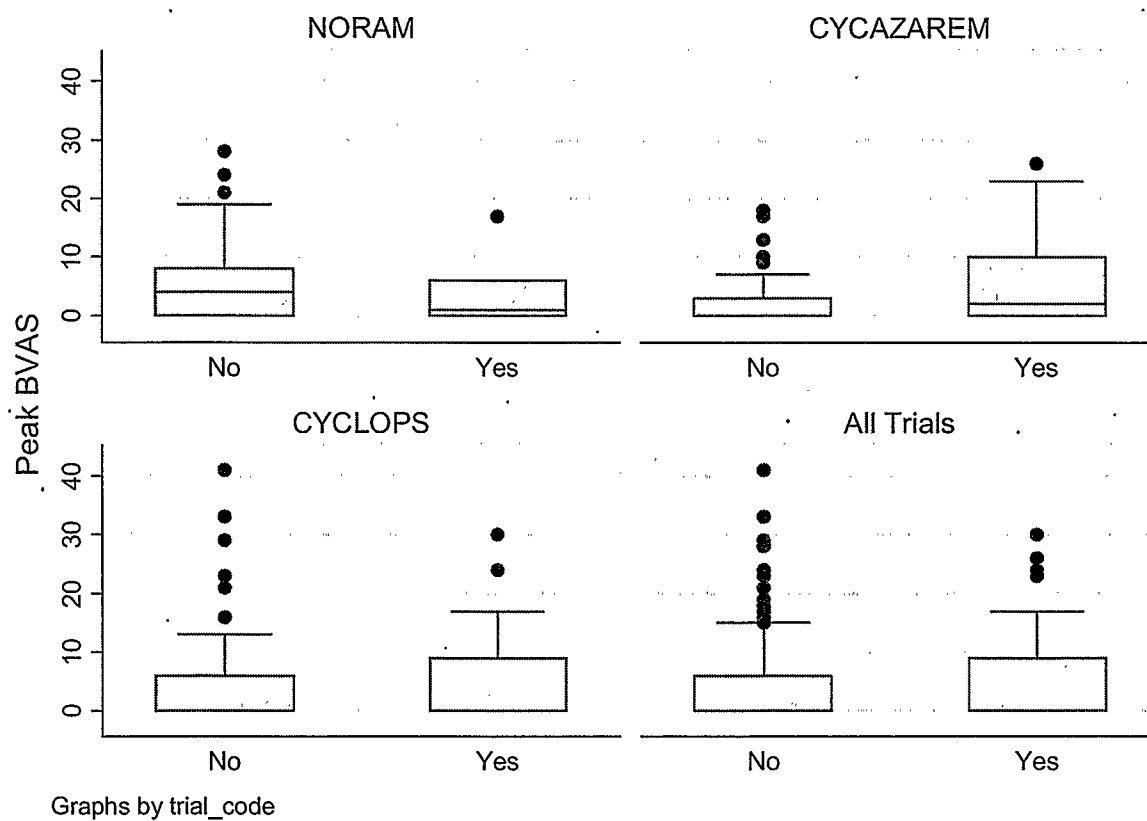


Figure 10 Peak Birmingham Vasculitis Activity Scores (BVAS) after the successful induction of remission in patients who did not reach death or ESRD (No) and those that did reach death or ESRD (Yes) arranged by trial.

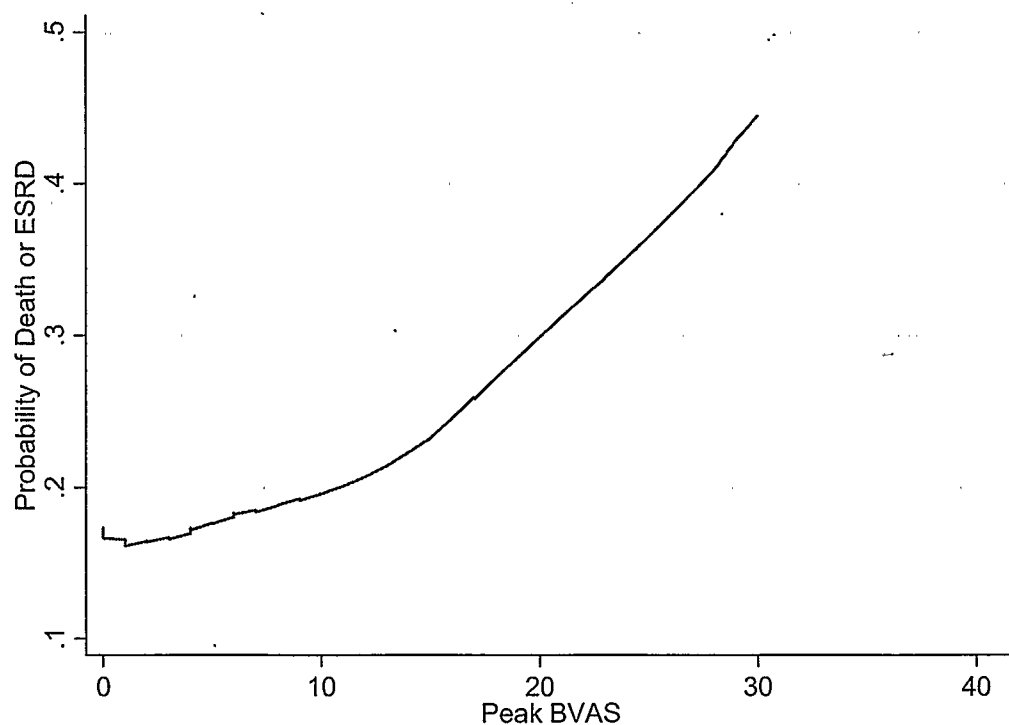


Figure 11 The predicted probability of death or ESRD by peak BVAS after the induction of remission according to median spline analysis by logistic regression adjusted for baseline eGFR and age. Risk appears to increase continuously with a rising score.

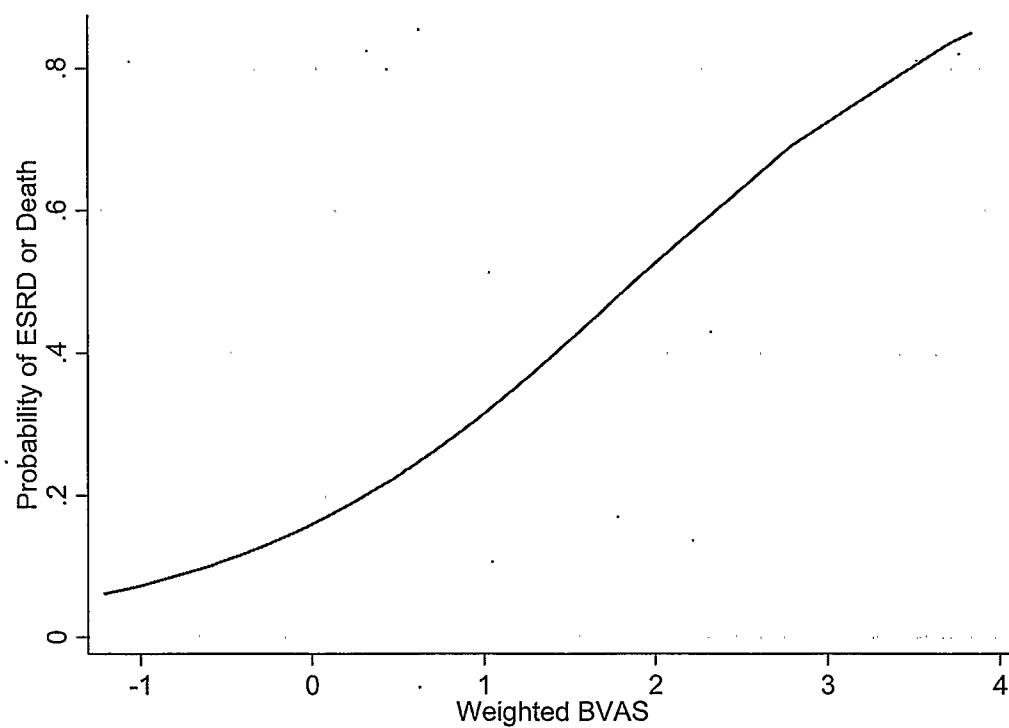


Figure 12 The predicted probability of death or ESRD by weighted BVAS after the induction of remission according to median spline analysis logistic regression. Risk appears to increase continuously with a rising score.

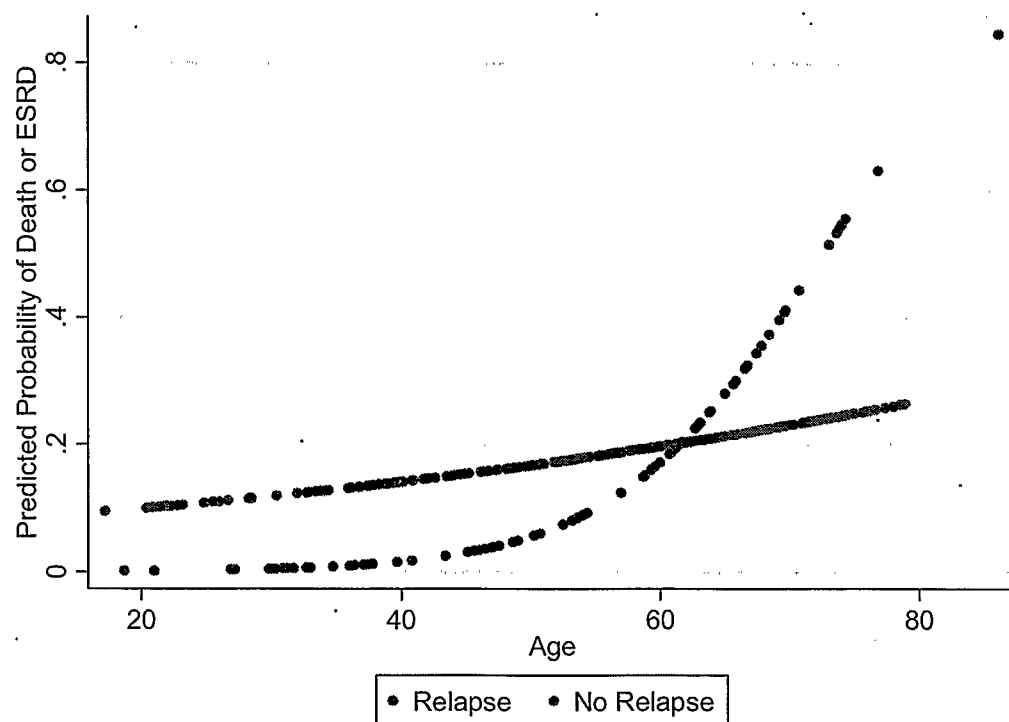


Figure 13 Predicted probability of death or ESRD from logistic regression using protocol defined relapse status and age as independent variables. The sharp increase in the probability of death or ESRD in elderly patients with a relapse may suggest an element of fragility in these patients (i.e. inability to tolerate recurrent disease or its treatment).

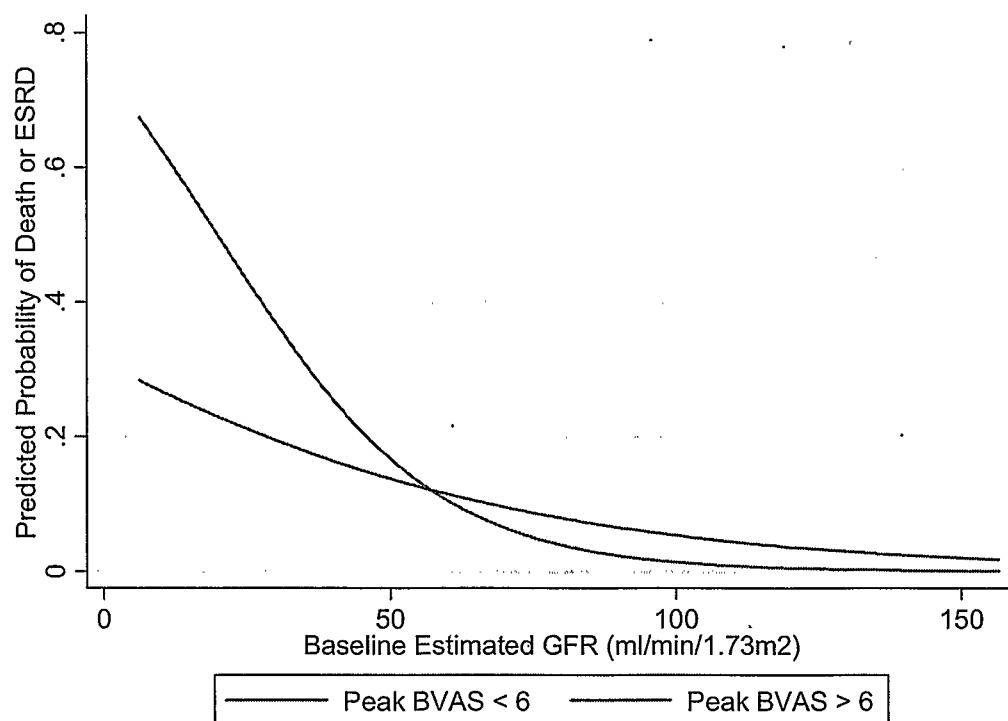


Figure 14 The predicted probability of death or ESRD over a range of baseline estimated glomerular filtration rate (eGFR) for patients with and without a relapse defined by a peak BVAS of greater than 6 in trials of ANCA associated vasculitis. The sharp increase in the probability of death or ESRD in patients with poor renal function with a relapse may suggest an element of fragility in these patients (i.e. inability to tolerate recurrent disease or its treatment).

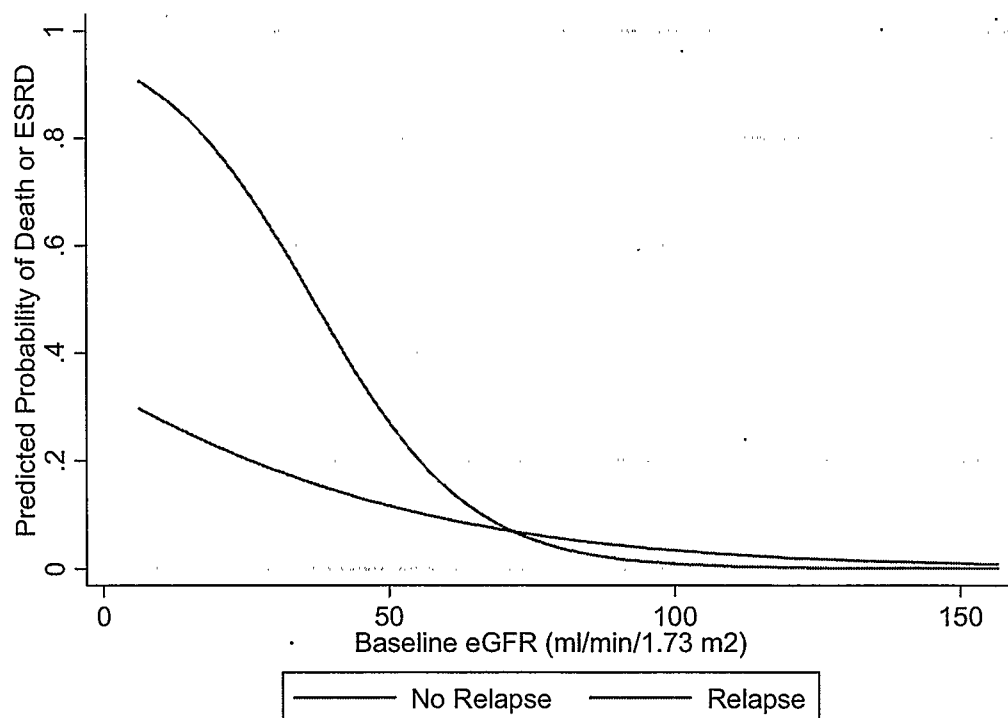


Figure 15 Predicted probability of death or ESRD from logistic regression using a weighted BVAS definition for relapse and eGFR as predictor variables demonstrating effect modification of the effect of relapse by eGFR. The sharp increase in the probability of death or ESRD in patients with poor renal function with a relapse may suggest an element of fragility in these patients (i.e. inability to tolerate recurrent disease or its treatment).

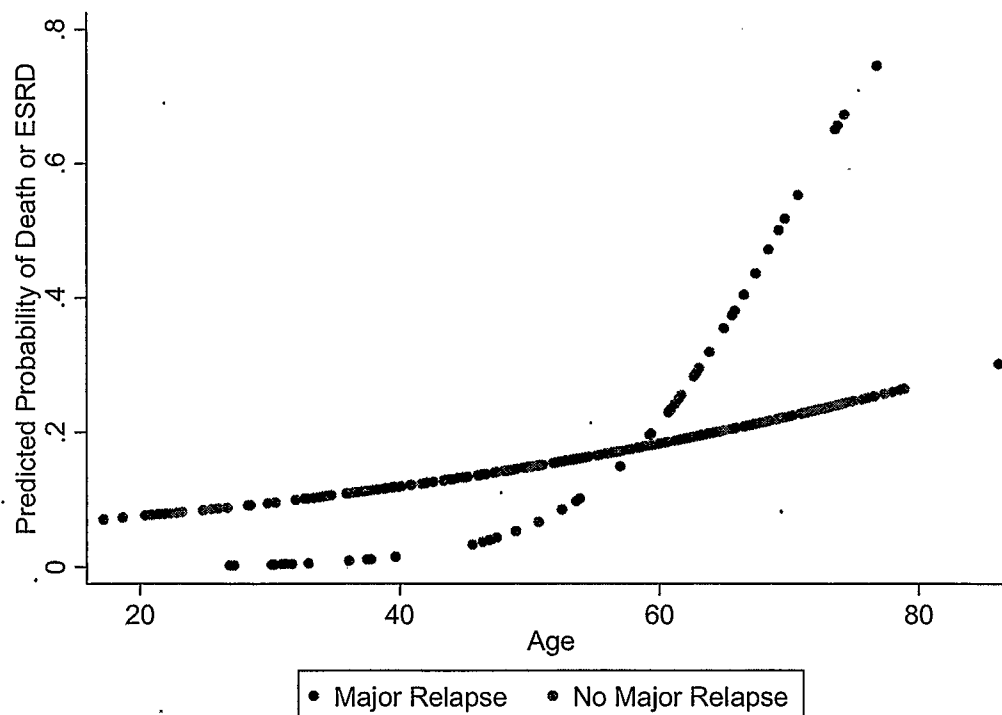


Figure 16 Predicted probability of death or ESRD from logistic regression using protocol defined major relapse status and age as independent variables. The sharp increase in the probability of death or ESRD in elderly patients with a relapse may suggest an element of fragility in these patients (i.e. inability to tolerate recurrent disease or its treatment).

Reference List

1. Walton EW. Giant cell granuloma of the respiratory tract (Wegener's granulomatosis). *Br Med J* 1958; 2:265-270.
2. Frohnert PF, Sheps SG. Long-term follow-up study of periarteritis nodosa. *Am J Med* 1967; 43:8-11.
3. Westman KW, Bygren PG, Ericsson UB, Hoier-Madsen M, Wieslander J, Erfurth EM. Persistent high prevalence of thyroid antibodies after immunosuppressive therapy in subjects with glomerulonephritis. A prospective three-year follow-up study. *Am J Nephrol* 1998; 18(4):274-279.
4. Friedman L, Furberg C, DeMets DL. Introduction to Clinical Trials. *Fundamentals of Clinical Trials*. 3 ed. New York: Springer-Verlag; 1998 p. 1-15.
5. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001; 69(3):89-95.
6. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med* 1989; 8(4):415-425.
7. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat Med* 2006; 25(2):183-203.
8. Walsh M, Tonelli M, Jayne D, Manns B. Surrogate end points in clinical trials: the case of anti-neutrophil cytoplasm antibody-associated vasculitis. *J Nephrol* 2007; 20(2):119-129.
9. Walsh M, Tonelli M, Jayne D, Manns B. Surrogate end points in clinical trials: the case of anti-neutrophil cytoplasm antibody-associated vasculitis. *J Nephrol* 2007; 20(2):119-129.
10. Temple R. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, editors. *Clinical Measurement in Drug Evaluation*. New York: Wiley; 1995 p. 1-22.
11. Manns B, Owen WF, Jr., Winkelmayr WC, Devereaux PJ, Tonelli M. Surrogate markers in clinical studies: problems solved or created? *Am J Kidney Dis* 2006; 48(1):159-166.
12. Blue JW, Colburn WA. Efficacy measures: surrogates or clinical outcomes? *J Clin Pharmacol* 1996; 36(9):767-770.
13. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989; 8(4):431-440.

14. Fischl MA, Richman DD, Grieco MH et al. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial. *N Engl J Med* 1987; 317(4):185-191.
15. Fischl MA, Richman DD, Hansen N et al. The safety and efficacy of zidovudine (AZT) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type 1 (HIV) infection. A double-blind, placebo-controlled trial. The AIDS Clinical Trials Group. *Ann Intern Med* 1990; 112(10):727-737.
16. Ellenberg S, Hamilton JM. Surrogate endpoints in clinical trials: cancer. *Stat Med* 1989; 8(4):405-413.
17. Hughes MD, Daniels MJ, Fischl MA et al. CD4 cell count as a surrogate endpoint in HIV clinical trials: a meta-analysis of studies of the AIDS Clinical Trials Group. *AIDS* 1998; 12(14):1823-1832.
18. Collette L, Burzykowski T, Schroder FH. Prostate-specific antigen (PSA) alone is not an appropriate surrogate marker of long-term therapeutic benefit in prostate cancer trials. *Eur J Cancer* 2006; 42(10):1344-1350.
19. Johnson KR, Ringland C, Stokes BJ et al. Response rate or time to progression as predictors of survival in trials of metastatic colorectal cancer or non-small-cell lung cancer: a meta-analysis. *Lancet Oncol* 2006; 7(9):741-746.
20. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; 54(3):1014-1029.
21. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. Statistical validation of surrogate endpoints: problems and proposals. *Drug Information Journal* 2000; 34:447-454.
22. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* 1994; 13(9):955-968.
23. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992; 11(2):167-178.
24. Lin DY, Fleming TR, De G, V. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat Med* 1997; 16(13):1515-1527.
25. Bycott PW, Taylor JM. An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. *Control Clin Trials* 1998; 19(6):555-568.

26. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; 1(1):49-67.
27. Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol* 2003; 3:16.
28. Molenberghs G, Geys H, Buyse M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Stat Med* 2001; 20(20):3023-3038.
29. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Applied Statistics* 2001; 50:405-422.
30. Renard D, Geys H, Molenberghs G et al. Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics* 2003; 30:235-247.
31. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996; 125(7):605-613.
32. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med* 1991; 20(1):47-63.
33. Stampfer MJ, Colditz GA, Willett WC et al. Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the nurses' health study. *N Engl J Med* 1991; 325(11):756-762.
34. Andrews TC, Raby K, Barry J et al. Effect of cholesterol reduction on myocardial ischemia in patients with coronary disease. *Circulation* 1997; 95(2):324-328.
35. Kjekshus J, Pedersen TR. Reducing the risk of coronary events: evidence from the Scandinavian Simvastatin Survival Study (4S). *Am J Cardiol* 1995; 76(9):64C-68C.
36. Shepherd J, Cobbe SM, Ford I et al. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group. *N Engl J Med* 1995; 333(20):1301-1307.
37. Hulley S, Grady D, Bush T et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA* 1998; 280(7):605-613.

38. Grady D, Herrington D, Bittner V et al. Cardiovascular disease outcomes during 6.8 years of hormone therapy: Heart and Estrogen/progestin Replacement Study follow-up (HERS II). *JAMA* 2002; 288(1):49-57.
39. Rossouw JE, Anderson GL, Prentice RL et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA* 2002; 288(3):321-333.
40. Riggs BL, O'Fallon WM, Lane A et al. Clinical trial of fluoride therapy in postmenopausal osteoporotic women: extended observations and additional analysis. *J Bone Miner Res* 1994; 9(2):265-275.
41. Eastell R, Barton I, Hannon RA, Chines A, Garnero P, Delmas PD. Relationship of early changes in bone resorption to the reduction in fracture risk with risedronate. *J Bone Miner Res* 2003; 18(6):1051-1056.
42. Harris ST, Watts NB, Jackson RD et al. Four-year study of intermittent cyclic etidronate treatment of postmenopausal osteoporosis: three years of blinded therapy followed by one year of open therapy. *Am J Med* 1993; 95(6):557-567.
43. Liberman UA, Weiss SR, Broll J et al. Effect of oral alendronate on bone mineral density and the incidence of fractures in postmenopausal osteoporosis. The Alendronate Phase III Osteoporosis Treatment Study Group. *N Engl J Med* 1995; 333(22):1437-1443.
44. Riggs BL, Seeman E, Hodgson SF, Taves DR, O'Fallon WM. Effect of the fluoride/calcium regimen on vertebral fracture occurrence in postmenopausal osteoporosis. Comparison with conventional therapy. *N Engl J Med* 1982; 306(8):446-450.
45. Riggs BL, Hodgson SF, O'Fallon WM et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med* 1990; 322(12):802-809.
46. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med* 1989; 321(6):406-412.
47. Kostis JB, Byington R, Friedman LM, Goldstein S, Furberg C. Prognostic significance of ventricular ectopic activity in survivors of acute myocardial infarction. *J Am Coll Cardiol* 1987; 10(2):231-242.
48. Moss AJ, DeCamilla JJ, Davis HP, Bayer L. Clinical significance of ventricular ectopic beats in the early posthospital phase of myocardial infarction. *Am J Cardiol* 1977; 39(5):635-640.

49. Moss AJ, Davis HT, DeCamilla J, Bayer LW. Ventricular ectopic beats and their relation to sudden and nonsudden cardiac death after myocardial infarction. *Circulation* 1979; 60(5):998-1003.
50. Ruberman W, Weinblatt E, Goldberg JD, Frank CW, Shapiro S. Ventricular premature beats and mortality after myocardial infarction. *N Engl J Med* 1977; 297(14):750-757.
51. Ruberman W, Weinblatt E, Goldberg JD, Frank CW, Chaudhary BS, Shapiro S. Ventricular premature complexes and sudden death after myocardial infarction. *Circulation* 1981; 64(2):297-305.
52. Anderson JL, Platia EV, Hallstrom A et al. Interaction of baseline characteristics with the hazard of encainide, flecainide, and moricizine therapy in patients with myocardial infarction. A possible explanation for increased mortality in the Cardiac Arrhythmia Suppression Trial (CAST). *Circulation* 1994; 90(6):2843-2852.
53. Moore T. *Deadly Medicine*. Simon and Schuster; 1995.
54. Detrano R, Hsiai T, Wang S et al. Prognostic value of coronary calcification and angiographic stenoses in patients undergoing coronary angiography. *J Am Coll Cardiol* 1996; 27(2):285-290.
55. Keelan PC, Bielak LF, Ashai K et al. Long-term prognostic value of coronary calcification detected by electron-beam computed tomography in patients undergoing coronary angiography. *Circulation* 2001; 104(4):412-417.
56. Chertow GM, Raggi P, McCarthy JT et al. The effects of sevelamer and calcium acetate on proxies of atherosclerotic and arteriosclerotic vascular disease in hemodialysis patients. *Am J Nephrol* 2003; 23(5):307-314.
57. Suki W, Zabaneh R, Cangiano J, Reed J, Fischer D, Swan S. The DCOR trial: a prospective, randomized trial assessing the impact on outcomes of sevelamer in dialysis patients. *American Society of Nephrology Renal Week* . 11-13-2005.
Ref Type: Abstract
58. Foley RN, Parfrey PS, Harnnett JD, Kent GM, Murray DC, Barre PE. The impact of anemia on cardiomyopathy, morbidity, and mortality in end stage renal disease. *Am J Kidney Dis* 1998; 31:53-61.
59. Locatelli F, Pisoni R, Combe C et al. Anaemia in haemodialysis patients of five European countries: association with morbidity and mortality in the Dialysis Outcomes and Practice Patterns Study (DOPPS). *Nephrology Dialysis and Transplantation* 2004; 19:121-132.

60. Besarab A, Bolton WK, Browne JK et al. 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* 1998; 339(9):584-590.
61. Singh N, Ahmad S, Wienckowski JR, Murray BM. Comparison of access blood flow and venous pressure measurements as predictors of arteriovenous graft thrombosis. *J Vasc Access* 2006; 7(2):66-73.
62. Moist LM, Churchill DN, House AA et al. Regular monitoring of access flow compared with monitoring of venous pressure fails to improve graft survival. *J Am Soc Nephrol* 2003; 14(10):2645-2653.
63. Dember LM, Holmberg EF, Kaufman JS. Randomized controlled trial of prophylactic repair of hemodialysis arteriovenous graft stenosis. *Kidney Int* 2004; 66(1):390-398.
64. Robbin ML, Oser RF, Lee JY, Heudebert GR, Mennemeyer ST, Allon M. Randomized comparison of ultrasound surveillance and clinical monitoring on arteriovenous graft outcomes. *Kidney Int* 2006; 69:730-735.
65. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994; 344(8934):1383-1389.
66. Abernethy J, Borhani NO, Hawkins CM et al. Systolic blood pressure as an independent predictor of mortality in the Hypertension Detection and Follow-up Program. *Am J Prev Med* 1986; 2(3):123-132.
67. Amery A, Birkenhager W, Brixko P et al. Mortality and morbidity results from the European Working Party on High Blood Pressure in the Elderly trial. *Lancet* 1985; 1(8442):1349-1354.
68. Amery A, Birkenhager W, Brixko P et al. Influence of antihypertensive drug treatment on morbidity and mortality in patients over the age of 60 years. European Working Party on High blood pressure in the Elderly (EWPHE) results: sub-group analysis on entry stratification. *J Hypertens Suppl* 1986; 4(6):S642-S647.
69. Dahlof B, Lindholm LH, Hansson L, Schersten B, Ekbom T, Wester PO. Morbidity and mortality in the Swedish Trial in Old Patients with Hypertension (STOP-Hypertension). *Lancet* 1991; 338(8778):1281-1285.
70. Staessen J, Fagard R, Amery A. Isolated systolic hypertension in the elderly: implications of Systolic Hypertension in the Elderly Program (SHEP) for clinical practice and for the ongoing trials. *J Hum Hypertens* 1991; 5(6):469-474.

71. Rolan P. The contribution of clinical pharmacology surrogates and models to drug development—a critical appraisal. *British Journal of Clinical Pharmacology* 1997; 44:219-225.
72. Food and Drug Administration. Guidance for Industry Fast Track Drug Development Programs -Designation, Development, and Application Review. 2004.
Ref Type: Report
73. Rasmussen N, Jayne DRW, Abramowicz D, Andrassy K, Bacon PA, Cohen. European therapeutic trials in ANCA-associated systemic vasculitis. *Clinical & Experimental Immunology, Supplement* 1995; . 101(1).
74. King T, Stone J. Clinical manifestations and diagnosis of Wegener's granulomatosis and microscopic polyangiitis. In: Rose B, editor. *UpToDate*. 15.2 ed. Waltham, MA: 2007.
75. Jennette J, Olson J, Schwartz M, Silva F. *Heptinstall's Pathology of the Kidney*. 6 ed. Lippincott Williams & Wilkins; 2006.
76. Schoen F. Blood Vessels. In: Kumar V, Abbas A, Fausto N, editors. *Robbins and Cotran Pathologic Basis of Disease*. 7 ed. Philadelphia: Elsevier Saunders; 2005.
77. Little MA, Smyth CL, Yadav R et al. Antineutrophil cytoplasm antibodies directed against myeloperoxidase augment leukocyte-microvascular interactions in vivo. *Blood* 2005; 106(6):2050-2058.
78. Little MA, Bhangal G, Smyth CL et al. Therapeutic effect of anti-TNF-alpha antibodies in an experimental model of anti-neutrophil cytoplasm antibody-associated systemic vasculitis. *J Am Soc Nephrol* 2006; 17(1):160-169.
79. Savage CO, Pottinger BE, Gaskin G, Pusey CD, Pearson JD. Autoantibodies developing to myeloperoxidase and proteinase 3 in systemic vasculitis stimulate neutrophil cytotoxicity toward cultured endothelial cells. *Am J Pathol* 1992; 141(2):335-342.
80. Savage CO, Gaskin G, Pusey CD, Pearson JD. Myeloperoxidase binds to vascular endothelial cells, is recognized by ANCA and can enhance complement dependent cytotoxicity. *Adv Exp Med Biol* 1993; 336:121-123.
81. Xiao H, Heeringa P, Hu P et al. Antineutrophil cytoplasmic autoantibodies specific for myeloperoxidase cause glomerulonephritis and vasculitis in mice. *J Clin Invest* 2002; 110(7):955-963.
82. Xiao H, Heeringa P, Liu Z et al. The role of neutrophils in the induction of glomerulonephritis by anti-myeloperoxidase antibodies. *Am J Pathol* 2005; 167(1):39-45.

83. Xiao H, Schreiber A, Heeringa P, Falk RJ, Jennette JC. Alternative complement pathway in the pathogenesis of disease mediated by anti-neutrophil cytoplasmic autoantibodies. *Am J Pathol* 2007; 170(1):52-64.
84. Knight A, Ekbom A, Brandt L, Askling J. Increasing incidence of Wegener's granulomatosis in Sweden, 1975-2001. *J Rheumatol* 2006; 33(10):2060-2063.
85. Abdou NI, Kullman GJ, Hoffman GS et al. Wegener's granulomatosis: survey of 701 patients in North America. Changes in outcome in the 1990s. *J Rheumatol* 2002; 29(2):309-316.
86. Mundy WL, Walker WGJ, Bickerman HA, Beck GJ. Periarthritis nodosa: report of a case treated with ACTH and cortisone. *Am J Med* 1951; 11:630-638.
87. Collagen Diseases and Hypersensitivity Panel (MRC). Treatment of polyarteritis nodosa with cortisone: results after three years. *Br Med J* 1960; 1399:1400.
88. Hollander D, Manning RT. The use of alkylating agents in the treatment of Wegener's granulomatosis. *Ann Intern Med* 1967; 67(2):393-398.
89. Fauci AS, Wolff SM, Johnson JS. Effect of cyclophosphamide upon the immune response in Wegener's granulomatosis. *N Engl J Med* 1971; 285(27):1493-1496.
90. Novack SN, Pearson CM. Cyclophosphamide therapy in Wegener's granulomatosis. *N Engl J Med* 1971; 284(17):938-942.
91. Fauci AS, Wolff SM. Wegener's granulomatosis: studies in eighteen patients and a review of the literature. *Medicine (Baltimore)* 1973; 52(6):535-561.
92. Leib ES, Restivo C, Paulus HE. Immunosuppressive and corticosteroid therapy of polyarteritis nodosa. *Am J Med* 1979; 67:941-945.
93. Fauci AS, Haynes BF, Katz P, Wolff SM. Wegener's granulomatosis: prospective clinical and therapeutic experience with 85 patients for 21 years. *Ann Intern Med* 1983; 98(1):76-85.
94. Fauci AS, Haynes BF, Katz P, Wolff SM. Wegener's granulomatosis: prospective clinical and therapeutic experience with 85 patients for 21 years. *Ann Intern Med* 1983; 98(1):76-85.
95. Hoffman GS, Leavitt RY, Fleisher TA, Minor JR, Fauci AS. Treatment of Wegener's granulomatosis with intermittent high-dose intravenous cyclophosphamide. *The American Journal of Medicine* 1990; 89(4):403-410.
96. Luqmani R. Measuring disease activity and outcomes in clinical studies. *Cleveland Clinic Journal of Medicine* 2002; . 69(SUPPL. 2).

97. Mukhtyar C, Hellmich B, Jayne D, Flossmann O, Luqmani R. Remission in antineutrophil cytoplasmic antibody-associated systemic vasculitis. *Clinical & Experimental Rheumatology* 2006; . 24(6 SUPPL. 43).
98. Jayne D. Current attitudes to the therapy of vasculitis. *Kidney and Blood Pressure Research* 2003; 26(4):231-239.
99. Pusey CD, Rees AJ, Evans DJ, Peters DK, Lockwood CM. Plasma exchange in focal necrotizing glomerulonephritis without anti-GBM antibodies. *Kidney international* 1991; . 40(4).
100. Pall A. Controlled trial of pulse cyclophosphamide (pcy) and prednisolone (pp) versus continuous cyclophosphamide (ccy) and prednisolone (cp) in the treatment of systemic vasculitis [abstract]. *J Am Soc Nephrol* 1992; 3(3):317.
101. Stegeman CA, Tervaert JWC, de Jong PE, Kallenberg CGM. Trimethoprim-sulfamethoxazole (Co-trimoxazole) for the prevention of relapses of Wegener's granulomatosis. *New England Journal of Medicine* 1996; . 335(1).
102. Adu D, Pall A, Luqmani RA et al. Controlled trial of pulse versus continuous prednisolone and cyclophosphamide in the treatment of systemic vasculitis. *Qjm: Monthly Journal of the Association of Physicians* 1997; . 90(6).
103. Guillevin L, Cordier J-F, Lhote F, Cohen P, Jarrousse B, Royer I. A prospective, multicenter, randomized trial comparing steroids and pulse cyclophosphamide versus steroids and oral cyclophosphamide in the treatment of generalized Wegener's granulomatosis. *Arthritis & Rheumatism* 1997; . 40(12).
104. Haubitz M, Schellong S, Gobel U, Schurek HJ, Schaumann D, Koch KM. Intravenous pulse administration of cyclophosphamide versus daily oral treatment in patients with antineutrophil cytoplasmic antibody-associated vasculitis and renal involvement: A prospective, randomized study. *Arthritis & Rheumatism* 1998; . 41(10).
105. Jayne DR, Rasmussen N, Jayne DR, Rasmussen N. Treatment of antineutrophil cytoplasm autoantibody-associated systemic vasculitis: initiatives of the European Community Systemic Vasculitis. *Mayo Clinic Proceedings* 1997; 72(8):737-747.
106. Jayne D, Gaskin G, Rasmussen N et al. Randomized trial of plasma exchange or high-dosage methylprednisolone as adjunctive therapy for severe renal vasculitis. *J Am Soc Nephrol* 2007; 18:2180-2188.
107. Etanercept plus standard therapy for Wegener's granulomatosis. *N Engl J Med* 2005; 352(4):351-361.
108. Luqmani RA, Bacon PA, Moots RJ et al. Birmingham Vasculitis Activity Score (BVAS) in systemic necrotizing vasculitis. *QJM* 1994; 87(11):671-678.

109. Flossmann O, Bacon PA, de GK et al. Development of comprehensive disease assessment in systemic vasculitis. *Ann Rheum Dis* 2007; 66:283-292.
110. Walsh M, Tonelli M, Jayne D, Manns B. Surrogate endpoints in clinical trials: the case of anti-neutrophil cytoplasm antibody associated systemic vasculitis. *J Neph* 2007; 2.
111. Hellmich B, Flossmann O, Gross WL et al. EULAR recommendations for conducting clinical studies and/or clinical trials in systemic vasculitis: Focus on ANCA-associated vasculitis. *Ann Rheum Dis* 2006.
112. Gayraud M, Guillevin L, le TP et al. Long-term followup of polyarteritis nodosa, microscopic polyangiitis, and Churg-Strauss syndrome: analysis of four prospective trials including 278 patients. *Arthritis Rheum* 2001; 44(3):666-675.
113. Luqmani R. Evaluation of vasculitis disease activity in Europe. *Eur J Intern Med* 2001; 12(5):401-402.
114. De Groot K, Gross W, Herlyn K, Reinhold-Keller E. Development and validation of a disease extent index for Wegener's granulomatosis. *Clin Nephrol* 2001;31-38.
115. Guillevin L, Lhote F, Gayraud M et al. Prognostic factors in polyarteritis nodosa and Churg-Strauss syndrome. A prospective study in 342 patients. *Medicine* 1996;17-28.
116. Exley AR, Bacon P, Luqmani R, Kitas GD, Savage C, Adu D. Development and initial validation of the Vasculitis Damage Index for the standardized clinical assessment of damage in the systemic vasculitides. *Arthritis Rheum* 1997;371-380.
117. De Groot K, Jayne D, Tesar V, Savage C. Randomised controlled trial of daily oral versus pulse cyclophosphamide for induction of remission in ANCA-associated systemic vasculitis. *Kidney & Blood Pressure Research* 2005; 28:195.
118. DeGroot K, Rasmussen N, Bacon PA et al. Randomized trial of cyclophosphamide versus methotrexate for induction of remission in early systemic antineutrophil cytoplasmic antibody-associated vasculitis. *Arthritis & Rheumatism* 2005; 52(8):2461-2469.
119. Jayne D, Rasmussen N, Andrassy K et al. A randomized trial of maintenance therapy for vasculitis associated with antineutrophil cytoplasmic autoantibodies. *N Engl J Med* 2003; 349(1):36-44.
120. Jayne D, Rasmussen N, Andrassy K et al. A randomized trial of maintenance therapy for vasculitis associated with antineutrophil cytoplasmic autoantibodies. *N Engl J Med* 2003; 349(1):36-44.

121. DeGroot K, Rasmussen N, Bacon PA et al. Randomized trial of cyclophosphamide versus methotrexate for induction of remission in early systemic antineutrophil cytoplasmic antibody-associated vasculitis. *Arthritis & Rheumatism* 2005; 52(8):2461-2469.
122. Walsh M, Tonelli M, Jayne D, Manns B. Surrogate end points in clinical trials: the case of anti-neutrophil cytoplasm antibody-associated vasculitis. *J Nephrol* 2007; 20(2):119-129.
123. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004; 66(3):411-421.
124. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49(12):1373-1379.
125. Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using Stata. College Station: StataCorp; 2005.
126. Twisk J. Applied multilevel analysis. Cambridge: Cambridge University Press; 2006.
127. Rabe-Hesketh S, Skrondal A, Pickles A. GLLAMM manual. Berkley: Berkley Electronic Press; 2004.
128. Hogan SL, Nachman PH, Wilkman AS, Jennette JC, Falk RJ. Prognostic markers in patients with antineutrophil cytoplasmic autoantibody-associated microscopic polyangiitis and glomerulonephritis. *J Am Soc Nephrol* 1996; 7(1):23-32.
129. Westman KW, Bygren PG, Olsson H, Ranstam J, Wieslander J. Relapse rate, renal survival, and cancer morbidity in patients with Wegener's granulomatosis or microscopic polyangiitis with renal involvement. *J Am Soc Nephrol* 1998; 9(5):842-852.

APPENDICES

Appendix 1. Birmingham Vasculitis Activity Score

VASCULITIS ACTIVITY SCORE

o Tick box **only** if abnormality is **newly present** since last assessment or **worse** in the last **few weeks** (use the Vasculitis Damage Index, VDI to score items of damage)
 □ Tick box **only** if abnormality is due to **active** (but not new or worse) vasculitis
 ◇ Tick box if more information (specialist opinion/tests) is requested
 @ oral/axillary temperatures; rectal temperatures are 0.5°C higher

DEMOGRAPHY

Trial Number

Visit Date / /

Investigator

PERSISTENT

NEW/WORSE

PERSISTENT

NEW/WORSE

1. GENERAL

□ (none)

malaise ☐ ☐
 myalgia ☐ ☐
 arthralgia/arthritis ☐ ☐
 headache ☐ ☐
 fever (< 38.5°C) @ ☐ ☐
 fever (≥ 38.5°C) @ ☐ ☐
 wt loss (≥ 2kg) ☐ ☐

2. CUTANEOUS

□ (none)

Infarct ☐ ☐
 purpura ☐ ☐
 other skin vasculitis ☐ ☐
 ulcer ☐ ☐
 gangrene ☐ ☐
 multiple digit gangrene ☐ ☐

3. MUCOUS MEMBRANES/EYES

□ (none)

mouth ulcers ☐ ☐
 genital ulcers ☐ ☐
 significant proptosis ☐ ☐
 red eye- conjunctivitis ☐ ☐
 red eye- epi/scleritis ☐ ☐
 blurred vision ☐ ☐
 sudden visual loss ☐ ☐
 ophthalmic opinion ☐ ◇
 no active vasculitis ☐ ☐
 uveitis ☐ ☐
 retinal exudates ☐ ☐
 retinal haemorrhage ☐ ☐

4. ENT

□ (none)

Nasal obstruction ☐ ☐
 Bloody nasal discharge ☐ ☐
 Nasal crusting ☐ ☐
 Sinus involvement ☐ ☐
 Hearing loss ☐ ☐
 Hoarseness/stridor ☐ ☐
 ENT opinion ☐ ◇
 no active vasculitis ☐ ☐
 Granulomatous sinusitis ☐ ☐
 Conductive hearing loss ☐ ☐
 Sensorineural hearing loss ☐ ☐
 Significant Subglottic inflammation ☐ ☐

5. CHEST

□ (none)

persistent cough ☐ ☐
 dyspnoea or wheeze ☐ ☐
 Haemoptysis/haemorrhage ☐ ☐
 chest radiology performed ☐ ◇
 no active vasculitis ☐ ☐
 nodules or cavities ☐ ☐
 pleural effusion/pleurisy ☐ ☐
 Infiltrate ☐ ☐
 massive haemoptysis or alveolar haemorrhage ☐ ☐
 respiratory failure ☐ ☐

6. CARDIOVASCULAR

□ (none)

aortic incompetence ☐ ☐
 pericardial pain/rub ☐ ☐
 ischaemic cardiac pain ☐ ☐
 congestive cardiac failure ☐ ☐
 cardiology opinion/tests ☐ ◇
 no active vasculitis ☐ ☐
 pericarditis ☐ ☐
 myocardial infarct/angina ☐ ☐
 cardiomyopathy ☐ ☐

7. ABDOMINAL

□ (none)

severe abdominal pain ☐ ☐
 bloody diarrhoea ☐ ☐
 surgical opinion/tests ☐ ◇
 no active vasculitis ☐ ☐
 gut perforation/infarct ☐ ☐
 acute pancreatitis ☐ ☐

8. RENAL

□ (none)

hypertension (diastol>95) ☐ ☐
 proteinuria >1+>0.2g/24h ☐ ☐
 haematuria>1+>10rbc/ml ☐ ☐
 creatinine 125-249 umol/l ☐ ☐
 creatinine 250-499 umol/l ☐ ☐
 creatinine >500 umol/l ☐ ☐
 rise in creatinine >30% or fall in creatinine clearance>25% ☐ ☐

9. NERVOUS SYSTEM

□ (none)

organic confusion/dementia ☐ ☐
 seizures(not hypertensive) ☐ ☐
 stroke ☐ ☐
 cord lesion ☐ ☐
 sensory peripheral neuropathy ☐ ☐
 cranial nerve palsy ☐ ☐
 motor mononeuritis multiplex ☐ ☐

10. OTHER

☐ ☐

GLOSSARY for BVAS

GENERAL RULE: disease features are scored only when they are due to active vasculitis, after exclusion of other obvious causes (e.g. infection, hypertension, etc.). If the feature has occurred afresh or represents a recent deterioration of status since last visit, it is scored in the NEW/WORSE boxes. It is essential to apply these principles to each item below. Scores have been weighted according to the severity which each symptom or sign is thought to represent. Tick box (Persistent) if the abnormality indicates the presence of active (but not new or worse) vasculitis. For some features, further information (from specialist opinion or further tests) is required if abnormality is newly present or worse. Remember that in most instances, you will be able to complete the whole record when you see the patient. However, on occasions, you may require further information before entering some items. We would suggest that you leave these items blank, and once the information is available, please remember to take the time to fill in the information. For example, if the patient has new onset of stridor, you would usually ask an ENT colleague to investigate this further to determine whether or not it is due to active Wegener's granulomatosis.

DERIVATION of BVAS.1 (new/worse) BVAS.2 (persistent) scores. The data from the score sheet will be used to derive indices of disease activity as follows:

BVAS.1 - This represents a score of new/worse disease activity attributable to vasculitis

BVAS.2 - This represents a score of disease activity due to persisting or grumbling disease, which is neither new nor worse, compared to the previous assessment..

Scores are calculated using the values given to each item as shown; each section has a maximum score, corresponding to the total value for BVAS (new/worse) and BVAS (persistent).

TERM	DEFINITION	BVAS persistent	BVAS new/worse
1. General			
Maximum scores		2	3
Malaise	A general feeling of tiredness, illness & discomfort.	1	1
Myalgia	Pain in the muscles	1	1
Arthralgia or arthritis	Pain in the joints or joint inflammation;	1	1
Headache	New, unaccompanied & persistent	1	1
Fever <38.5	Documented oral/axillary temperature elevation. Rectal temperatures are 0.5 C higher	1	1
Fever >=38.5	Documented oral/axillary temperature elevation. Rectal temperatures are 0.5 C higher	2	2
Weight Loss	At least 2kg loss of body weight (not fluid) having occurred since last assessment or in the 4 weeks not as a consequence of dieting	2	2

2. Cutaneous			
Maximum scores		3	6
Infarct	Area of tissue necrosis or splinter haemorrhages	1	2
Purpura	Petechiae (small red spots), palpable purpura, or ecchymoses (large plaques) in skin or oozing (in the absence of trauma) in the mucous membranes.	1	2
Other skin vasculitis	e.g., livedo reticularis, nodules etc.	2	2
Ulcer	Open sore in a skin surface.	1	4
Gangrene	Extensive tissue necrosis (e.g. digit)	1	6
Multiple digit gangrene	Extensive tissue necrosis occurring in more than one digit or limb.	2	6

3. Mucous membranes/eyes			
Maximum score		3	6
Mouth ulcers	Ulcers localised in the mouth. Exclude other causes, such as drugs, Crohn's disease, pemphigus etc.	1	1
Genital ulcers	Ulcers localised in the genitalia or perineum.	1	1
Significant proptosis	Protrusion of the eyeball due to significant amounts of inflammatory in the orbit. This may be associated with diplopia due to infiltration of extra-ocular muscles.	2	4
Red eye conjunctivitis	Inflammation of the conjunctivae (exclude infectious causes); (specialist opinion not usually required)	1	1
Red eye (Epi)scleritis	Inflammation of the sclerae (specialist opinion not usually required)	1	2
Blurred vision	Significant impairment of vision.	2	3
Sudden visual loss	Sudden loss of vision requiring ophthalmological assessment.	-	6
Ophthalmic opinion	To diagnose & score retinal exudates, haemorrhages, uveitis & reason for sudden visual loss. This data must be entered on score sheets subsequently.	-	-
Uveitis*	Inflammation of the uvea (iris, ciliary body, choroid) confirmed by ophthalmologist.	-	6
Retinal exudates*	Any area of soft retinal exudates (exclude hard exudates) seen on ophthalmoscopic examination.	-	6
Retinal haemorrhages*	Any area of retinal haemorrhage seen on ophthalmoscopic examination.	-	6

4. ENT			
Maximum scores		3	6
Nasal obstruction	A history of nasal blockage	1	2
Bloody nasal discharge	Blood stained secretions from the nose, irrespective of severity, or frequency & severity of previously occurring bleeding since last visit.	2	4
Nasal crusting	Discharge of large serous or serosanguinous crusts from either nostril.	2	4
Sinus involvement	Tenderness or pain over paranasal sinuses or X-ray evidence of sinusitis. If nasal bridge collapse is observed, this may be recorded separately (in 10. Other)	1	2
Hearing loss	Significant new hearing loss requiring specialist opinion.	-	3
Hoarseness/stridor	Increasing hoarseness & inspiratory stridor.	2	5
ENT opinion	To ascribe otitis media, deafness, or diagnose subglottic involvement due to vasculitis. This data can be entered on score sheets subsequently.	-	-
Granulomatous sinusitis*	Characteristic appearance on nasal examination	-	4
Conductive hearing loss*	Any hearing loss due to middle ear involvement preferably confirmed by audiometry.	-	3
Sensorineural hearing loss*	Deafness attributable to auditory nerve or cochlear damage.	-	6
Significant subglottic inflammation*	Inspiratory stridor with significant narrowing of subglottic space confirmed by further examination (usually by an ENT specialist) or by radiological assessment	-	6

5. Chest			
Maximum scores		3	6
Persistent cough	Cough for more than 2 weeks (other causes for the cough having been excluded e.g. infection)	1	2
Dyspnoea or wheeze	Shortness of breath or audible wheeze on exercise, by history &/or clinical examination.	1	2
Haemoptysis/haemorrhage	Production of blood stained sputum. Other causes (e.g. infection, cancer) should be excluded.	1	3
Chest radiology performed	A chest radiograph should be performed if there are significant signs or symptoms to suggest chest disease or in the presence of a generalised flare - to determine the following three:	-	-
Nodules or cavities*	New lesions, detected by CXR.	-	3
Pleural effusion/pleurisy*	Pleural pain &/or friction rub on clinical assessment or new onset of radiologically confirmed pleural effusion. Other causes (e.g. infection, cancer) should be excluded.	-	4
Infiltrate	By CXR, CT scan.	-	4
Massive haemoptysis/Alveolar haemorrhage	Major pulmonary bleeding, with shifting pulmonary infiltrates & usually associated with signs of shock; other causes of bleeding should be excluded.	-	6
Respiratory failure	Dyspnoea which is sufficiently severe as to require artificial ventilation; arterial blood gases should be performed to confirm the presence of hypoxaemia & or hypercapnia.	3	6

6. Cardiovascular			
Maximum scores		3	6
Aortic incompetence	Significant aortic valve regurgitation, detected clinically or echocardiographically.	2	4
Pericardial pain/rub	Pericardial pain &/or friction rub on clinical assessment.	2	3
Ischaemic cardiac pain	Typical clinical history of cardiac pain. Consider the possibility of more common causes (e.g. atherosclerosis)	2	4
Congestive cardiac failure	By history or clinical examination	2	4
Cardiology opinion or tests	Specialist opinion/tests are usually required to determine the following features	-	-
Pericarditis*	Pericardial pain &/or friction rub on clinical assessment.	-	4
Myocardial infarction/angina*	Typical history of cardiac pain.	-	6
Cardiomyopathy*	Heart failure by history or clinical examination	-	6

7. Abdominal			
Maximum scores		4	9
Severe abdominal pain	Of recent onset & attributed to vasculitis.	2	3
Bloody diarrhoea	Of recent onset, not due to known inflammatory bowel disease, etc.	2	3
Surgical opinion/ tests	Specialist opinion/tests required to determine the cause of abdominal pain or diarrhoea if they are of recent onset or worse since last visit.	-	-
Gut perforation/infarction*	Typical pain & peritonism includes gall bladder or appendix. Confirmed by X-ray or at surgery.	-	9
Acute pancreatitis*	Typical history & clinical examination findings of acute abdominal pain & tenderness with guarding. Confirmed by elevated serum amylase & a surgical opinion	-	9

8. Renal			
Maximum scores		6	12
Hypertension	Diastolic BP>95, accelerated or not, with or without retinal changes.	1	4
Proteinuria	>1+ on urinalysis; >0.2g/24 hours. Infection should be excluded.	2	4
Haematuria	>1+ on urinalysis; >10 rbc/ml, or red cell casts seen on urine microscopy. Infection should be excluded.	3	6
Creatinine 125-249	Serum creatinine values 125-249 umol/l at first assessment only.	2	4
Creatinine 250-499	Serum creatinine values 250-499 umol/l at first assessment only.	3	6
Creatinine >=500	Serum creatinine values 500 umol/l or greater at first assessment only.	4	8
Rise in creatinine > 30% or creatinine clearance fall > 25%	Significant deterioration in renal function attributable to active vasculitis.	-	6

9. Nervous system			
Maximum scores		6	9
Organic confusion/Dementia	Impaired orientation, memory or other intellectual function in the absence of metabolic, psychiatric, pharmacological or toxic causes.	1	3
Seizures (not hypertensive)	Paroxysmal electrical discharges in the brain & producing characteristic physical changes including tonic & clonic movements & certain behavioural changes.	3	9
Stroke	Cerebrovascular accident resulting in focal neurological signs such as paresis, weakness, etc. A stroke due to other causes (eg atherosclerosis) should be considered & appropriate neurological advice is recommended	3	9
Cord lesion	Transverse myelitis with lower extremity weakness or sensory loss (usually with a detectable sensory level) with loss of sphincter control (rectal & urinary bladder).	3	9
Sensory Peripheral neuropathy	Sensory neuropathy resulting in glove &/or stoking distribution of sensory loss. Other causes should be excluded (e.g. idiopathic, metabolic, vitamin deficiencies, infectious, toxic, hereditary).	3	6
Cranial nerve palsy	Isolated acute cranial nerve palsy, excluding sensorineural hearing loss, or optic nerve lesion secondary to retro-orbital mass.	3	6
Motor mononeuritis multiplex	Simultaneous neuritis of many peripheral nerves, only scored if motor involvement. Other causes should be excluded (diabetes, sarcoidosis, carcinoma, amyloidosis).	3	9

10. Other			
Significant features attributable to active vasculitis not listed above.			
Total maximum score		33	63



2008-05-29

Dr. Braden Manns
Department of Medicine
Foothills Medical Centre
Calgary
Alberta

OFFICE OF MEDICAL BIOETHICS
Room 93, Heritage Medical Research Bldg
3330 Hospital Drive NW
Calgary, AB, Canada T2N 4N1
Telephone: (403) 220-7990
Fax: (403) 283-8524
Email: omb@ucalgary.ca

Dear Dr. Manns:

RE: Clinical Activity Score as an Endpoint in Randomized Clinical Trials of Immunosuppressive Medications in Anti-Neutrophil Cytoplasm Antibody Associated Vasculitis: A Validation Study

Ethics ID: E-21536

Student: Dr. M. Walsh

The above-noted proposal including the Protocol has been submitted for Board review and found to be ethically acceptable.

Please note that this approval is subject to the following conditions:

- (1) consent for access to personal identified health information in chart review is not required on grounds considered under
- (2) a copy of the informed consent form must have been given to each research subject, if required for this study;
- (3) a Progress Report must be submitted by **May 29, 2009**, containing the following information:
 - i) the number of subjects recruited;
 - ii) a description of any protocol modification;
 - iii) any unusual and/or severe complications, adverse events or unanticipated problems involving risks to subjects or others, withdrawal of subjects from the research, or complaints about the research;
 - iv) a summary of any recent literature, finding, or other relevant information, especially information about risks associated with the research;
 - v) a copy of the current informed consent form;
 - vi) the expected date of termination of this project.
- 4) a Final Report must be submitted at the termination of the project.

Please note that you have been named as the principal collaborator on this study because students are not permitted to serve as principal investigators. Please accept the Board's best wishes for success in your research.

Yours sincerely,


Glenys Godlovitch, BA(Hons), LLB, PhD
Chair, Conjoint Health Research Ethics Board

GG/emcg

c.c. Ms. Gladys Glowacki (Health Records) Ms. Donna McDonald (RTA)
Services Dr. M. Walsh (Student)
Office of Information & Privacy Commissioner

Dr.J.Conly (information)

Research