## Item S1. Supplemental Methods

This methods supplement provides conceptual overviews and analytic details concerning the three analytical approaches (individual level, Freedman-Prentice and trial level analyses) provided in this report. The general methodological perspectives taken by the authors for validation of surrogate endpoints are described in Joffe and Greene[1]. Stevens et al[2] overview the use of the methods of this report for the specific problem of validating proteinuria as a surrogate endpoint. A detailed presentation of statistical methods for validation of surrogate endpoints is provided in the book by Burzykowski et al[3]. Further background for trial-level approaches can be found in Daniels et al[4], Korn et al[5], and Gail et al[6], and additional background for the Prentice-Freedman criteria can be found in Prentice[7], Freedman et al[8], Frangakis et al[9] and Taylor at al[10]. Regulatory issues related to the validation of surrogate endpoints are described in Fleming et al and in Katz et al.[11-13]

Please note that in order to provide a more detailed presentation of our methods, we have adopted in this supplement a slight modification of the terminology of the primary manuscript, and below we use the term "study" to refer to one of the 23 distinct randomized trials, with the 3 classes of immunosuppressive studies grouped together as described in the primary manuscript, and use the expressions "treatment comparisons within each study" or "treatment comparison by study combinations" to refer to the specific pairwise comparisons between treatment groups within the studies. Thus, in this supplement, we refer to 23 studies, but 29 study by treatment-comparison combinations.

## A. Primary Analyses

Our primary analyses include three approaches: individual level, Prentice -Freedman and trial level analyses. For all three approaches, we first performed analyses at the study level and then performed joint analyses to summarize results across studies. The analyses of individual level association and the Prentice-Freedman criteria are limited by the potential for confounding and therefore throughout the manuscript we describe them first before we proceed to the trial level analysis. In this methods supplement, where our focus is on the statistical techniques, we will describe first our analyses of individual level and trial-level association, as the joint analyses for these approaches each employ Bayesian mixed effects models, followed by our analyses of the Prentice-Freedman criterion, where we employ a fixed effects model with stratification of the baseline hazard by study.

We first describe our rationale for the use of the Bayesian mixed effects models, (Section A.1), followed next by descriptions of our analyses of individual and trial-level association, (A.2, A.3, A.4), and then by our multi-level analyses relating the individual and trial-level parameters to baseline proteinuria (A.5) and finally the details of our analyses of the Prentice Freedman criteria (A.6).

## A.1. Rationale for Bayesian Analyses

Our primary reason for applying a Bayesian framework to implement joint mixed effects analyses is that a lower than expected variation in treatment effects on the clinical outcomes precluded adequate assessments of the uncertainty of our estimates under a conventional frequentist framework.[12] Specifically, the estimated between-study standard deviation of the log hazard ratios for the clinical outcome was identically 0 under frequentist mixed models. The estimate of a 0 standard deviation resulted because the between-study variation in the hazard ratios was smaller than expected by chance due to random sampling error, based on the sizes of the respective studies. As a result, we were not able to use the frequentist approach to determine confidence intervals for the pooled mean or the between-study standard deviation of the log hazard ratios for the clinical outcome, or for the analyses relating the treatment effects on the clinical outcome to treatment effects on change in log urine protein. The Bayesian approach allowed determination of posterior credible intervals to quantity our level of uncertainty in these quantities.[14] Although the interpretation of hypothesis testing differs between the Bayesian and the Frequentist frameworks, we can formally assess the compatibility of the data with a particular "null hypothesis" under the Bayesian framework by noting whether the Bayesian credible intervals include the value corresponding to the null hypothesis.

We also used the Bayesian framework for joint analyses of individual level association and of treatment effects on change in proteinuria to provide a consistent approach for all of our mixed effects analyses.

## A.2. Analyses of individual level association.

*Overview*. The purpose of these analyses was to characterize the epidemiologic association between the clinical outcome and early change in urine protein (UP) for individual patients, after controlling for the effect of treatment on the respective endpoints. Our analysis of individual level association across the studies consisted of two steps. In the first step, we obtained separate estimates of the individual level association within each study, represented by the log transformed hazard ratio (HR) relating the clinical outcome to the early change in log UP, controlling for baseline log UP and randomized treatment group. In the second step, we performed a meta-analysis to characterize the pooled overall geometric mean and the variation in the individual level association HR across all the 23 studies. A key feature of the second step is that it incorporates the precision of the estimate for each study, so that smaller studies had less influence than larger studies. The reported variability in the individual level association across studies reflects only the variation in the "true" individual level association parameters, and not the additional random sampling variation resulting from the limited sample sizes of the respective studies. We have provided single overall individual level association estimates for each study which incorporate all the treatment groups in the respective studies.

*Details*. The details of the two steps are as follows. In the first step, for each study *i*, we obtained an estimate of the log HR, denoted $\hat{\tau}_i$, and its model-based standard error $\sigma(\hat{\tau}_i)$, based on a Cox regression relating the clinical outcome to the early change in log UP after controlling for treatment assignment, baseline log UP, and the linear interaction of baseline log UP with follow-up time. For the three groups of studies within the immunosuppressive therapy class of

interventions, the baseline hazard functions of the Cox regressions were stratified for the individual trials within each of the three groups. In the second step we performed a meta-analysis of the individual level association parameter estimates from the first step using the Bayesian mixed effects model:

$$\tau_i \sim N(\tau, \sigma^2(\tau)),\ i = 1, 2, \ldots, 23, \qquad \text{(I1)}$$

$$\hat{\tau}_i \sim N(\tau_i, \sigma^2(\hat{\tau}_i)),\ i = 1,2, \ldots, 23. \qquad \text{(I2)}$$

The expression (I1) indicates that the true individual association log HRs, denoted by $\tau_i$ for $i$ = 1, 2, …, 23, are assumed to be normally distributed with overall mean $\tau$ and standard deviation $\sigma(\tau)$. The expression (I2) indicates for each study $i$, the estimated individual level association log HR $\hat{\tau}_i$ from the first step is normally distributed with mean $\tau_i$ and with the standard error $\sigma(\hat{\tau}_i)$ obtained in Step 1. We assumed that the true $\tau_i$ and estimated $\hat{\tau}_i$ are statistically independent between studies. Our Bayesian model stipulated a diffuse prior distribution for $\tau$ so that the Bayesian estimate of the overall individual level association would be determined primarily by the data, with little influence of the assumed prior distribution. The prior distribution for the variance $\sigma^2(\tau)$ was taken to be an inverse gamma distribution with shape and scale parameters of 0·261 and 0·00408, respectively, defined to assign a probability of 1/3 to a small level of heterogeneity, represented by a standard deviation in the log HRs smaller than 0·05, a probability of 1/3 to a large amount of heterogeneity, represented by a standard deviation in log HRs greater than 0·20, and a probability of 1/3 to an intermediate amount of heterogeneity, represented by a standard deviation between 0·05 and 0·20. The interpretation of the assumed prior distribution is based on the approximately equality between standard deviation of the log HRs and the coefficient of variation in the untransformed HRs.

Posterior distributions of the parameters for the Bayesian models for individual level association and for the trial level analyses described below were obtained by Markov Chain Monte Carlo (MCMC) simulation using the PROC MCMC of SAS Version 9.3.[15] Convergence was checked for each parameter by a) visual examination of MCMC chains, b) the Geweke diagnostic test of the equality of the mean level of the early and later part of the MCMC chains, and c) evaluation of autocorrelations and effective sample sizes for the MCMC chains.

### A.3. Trial-Level Analyses Relating Treatment Effects on the Clinical outcome and on Initial Change in UP

*Overview*. The purpose of these analyses was to address the question of whether treatment effects in RCTs of change in log UP can be used to predict treatment effects on the clinical outcome. The analyses of individual level association described above are subject to confounding if there are uncontrolled baseline or follow-up factors that influence both change in log UP and the clinical outcome. Thus, the question of whether treatment effects on change in log UP can be used to predict treatment effects on the clinical outcome is more directly addressed by trial level analyses, which can be viewed as meta-regressions in which treatment effects on the clinical outcome define the dependent variable, and treatment effects on early

change in proteinuria define the independent variable. In this section, we first describe preliminary meta-analyses which summarized the distributions of the treatment effects on change in log UP and on the clinical outcome, then describe the meta-regression which related the treatment effects on the change in log UP and on the clinical outcome to each other.

As with individual level association, each of these analyses contained two steps. In this case, the first step provided separate estimates of the treatment effects on the clinical outcome and on change in log UP for each independent treatment comparison in each study. The second step consisted either of a Bayesian meta-analysis, when evaluating the distribution of the treatment effects on each endpoint individually, or a Bayesian multi-level analysis, when relating the treatment effects on the two endpoints to each other. Importantly, in each case the Bayesian analyses of the second steps estimated the variation of the *true treatment effects* on the respective endpoints after discounting the additional variation due to random sampling error that resulted from the limited sample sizes of the studies. The Bayesian meta-regression relating the treatment effects to each other also took into account the correlation between the deviations of the estimated vs. the true treatment effects for the two endpoints. Such a correlation in sampling error is a concern, as individual level associations can lead to nonzero correlation between the estimated treatment effects on the two endpoints even in the absence of any true treatment effects on either endpoint in any study.[10] The impact of this spurious correlation can be particularly severe for studies with the small sample sizes that were typical of many of the randomized trials included in this report.

*Details for Meta-Analysis of Treatment Effects on Change in log UP*. In the first step, separate linear regression analyses were performed in each study to relate the early change in log UP to the randomized treatment assignment, without covariate adjustment. For the three groups of studies of immunosuppressive therapy, separate intercepts were fit for each of the contributing trials. For the IDNT and AASK studies, which both included three drug treatment groups, the linear regressions were performed with separate indicator variables to produce separate treatment effect estimates for the ACE/ARB and CCB treatments vs. control, and the output included an estimate of the correlation in the sampling error between the two treatment effect estimates. In all, 5 of the 23 studies had more than one independent treatment comparison, yielding a total of 29 study by treatment comparison combinations. These included 3 comparisons for the AASK Study (2 drug intervention comparisons and 1 blood pressure control comparison), and 2 comparisons each for MDRD Studies A and B (1 diet and 1 blood pressure comparison for each study), the ABCD (1 drug comparison and 1 blood pressure comparison), and the IDNT (2 drug intervention comparisons).

For each treatment comparison within each study, the linear regression produced an estimate of the log of the geometric mean ratios (GMR) of the follow-up vs. baseline geometric mean UP levels between the treatment and control groups, denoted $\hat{\gamma}_i$, along with a model based estimate of its standard error $\sigma^2(\hat{\gamma}_i)$.

The Bayesian model for the second step is expressed:

$$\gamma_i \sim N(\gamma, \sigma^2(\gamma)),\ i = 1, 2, \ldots, 29, \qquad \text{(U1)}$$

$$\hat{\gamma}_i \sim N(\gamma_i, \sigma^2(\hat{\gamma}_i)),\ i = 1,2, \ldots, 29. \quad \text{(U2)}$$

Analogous to the model for individual level association, we assumed a diffuse prior for $\gamma$ and an inverse gamma prior for $\sigma^2(\gamma)$ which assigned probabilities of 1/3 to small, intermediate, and large values of $\sigma(\gamma)$ defined by $\sigma(\gamma) < 0{\cdot}05$, $0{\cdot}05 \leq \sigma(\gamma) < 0{\cdot}20$, and $\sigma(\gamma) \geq 0{\cdot}20$. We assumed that all the true treatment effects $\gamma_i$ were mutually independent. We also assumed that the deviations of the estimated from the true treatment effects $\hat{\gamma}_i - \gamma_i$ were mutually independent (including those resulting from different factors from the same study in factorial designs) , with the exception of the aforementioned AASK and IDNT drug group comparisons, where deviations between the observed and true treatment effects were assumed to be correlated as estimated from the linear regressions in Step 1.

In a separate Bayesian mixed effects model, we expanded the common mean treatment effect $\delta$ in expression U1 to allow different values $\gamma_k$, $k = 1, 2, .., 5$, for the five intervention types, where each $\gamma_k$ was assumed to have an independent diffuse prior distribution. The results of each these analyses were exponentiated to characterize the distribution of the HRs for the treatment vs. control groups across the study by treatment comparison combinations.

*Details for Multi-Level Analyses Relating Treatment Effects on the Clinical Outcome*. The analyses of the treatment effects on the clinical outcome had the same structure as the analyses of the treatment effect on the early change in log UP, with Cox regressions of the clinical outcome replacing the linear regressions of change in log UP in the first step of the 2-step analysis. For the 3 groups of immunosuppressive trials, the baseline hazard function was stratified by the specific contributing trials. For each treatment comparison within each study, the Cox regression produced an estimate of the log HR for the treatment effect, denoted $\hat{\delta}_i$, along with a model based estimate of its standard error $\sigma(\hat{\delta}_i)$.

The second step was based on the Bayes model:

$$\delta_i \sim N(\delta, \sigma^2(\delta)),\ i = 1, 2, \ldots, 29, \quad \text{(C1)}$$

$$\hat{\delta}_i \sim N(\delta_i, \sigma^2(\hat{\delta}_i)),\ i = 1,2, \ldots, 29. \quad \text{(C2)}$$

For each of the 29 study by treatment-comparison combinations $i$, $\delta_i$ denotes the true treatment effect on the clinical outcome expressed as the log HR for the treatment vs. control groups, $\delta$ denotes the mean treatment effect across the 29 study by treatment-comparison combinations, $\sigma(\delta)$ represents the standard deviation of the true treatment effects about their overall mean, after discounting random sampling error associated with the limited sample sizes of the respective studies. We applied the same independence assumptions and the same prior distributions to the treatment effects on the clinical outcome as we did for the treatment effects on the early change in log UP. We also considered an analogous expanded model allowing separate mean treatment effects $\delta_k$ for the five intervention types.

*Details of Multi-Level Analyses Relating Treatment Effects on the Clinical Outcome to Treatment Effects on Change in UP*. In the first step, we performed separate Cox regressions of the clinical outcome and linear regressions of early change in log UP to obtain treatment effect estimates $\hat{\delta}_i$ and $\hat{\gamma}_i$ for each study by treatment comparison combination as described above. However, rather than using model-based estimates of the standard errors in these estimates, we applied bootstrap resampling for each study by treatment comparison to estimate the full covariance matrix for the deviations of each $\hat{\delta}_i$ and $\hat{\gamma}_i$ from the true treatment effects $\delta_i$ and $\gamma_i$. This covariance matrix defined the standard errors of $\hat{\delta}_i$ and $\hat{\gamma}_i$ as well as the correlation between these estimates. A total of 800 bootstrap samples were generated for each trial, with estimates $\hat{\delta}_{ib}$ and $\hat{\gamma}_{ib}$ obtained from each sample *b*, *b* = 1, 2, …, 800, and the covariance matrix was computed empirically based on the joint distribution of the $\hat{\delta}_{ib}$ and $\hat{\gamma}_{ib}$. By using the bootstrap approach we were able to avoid the need to validate complex joint models for the early change in log UP and time to the clinical outcome for each study. However, in order to assure convergence of the Cox models for each bootstrap sample, it was necessary to exclude three studies with fewer than 15 clinical events (Van Essen, A3; HKVIN, A9; and Praga, A11).

The Bayesian model in the second step retained expressions C1 and U1 (with *i* ranging from 1 to 26 instead of 29 to account for the three omitted studies), but replaced C2 and U2 by the assumption that the pair ($\hat{\gamma}_i$, $\hat{\delta}_i$) is bivariate normal with mean ($\gamma_i$, $\delta_i$) and covariance matrix given by the bootstrap procedure. The model also includes the meta-regression:

$$\delta_i = \beta \times \gamma_i + \mu_\delta + \varepsilon_{\delta i}, \qquad i = 1, 2, \ldots, 26 \qquad \text{(M1)}$$

where β represents the meta-regression coefficient relating the treatment effects $\delta_i$ on the clinical outcome to the treatment effects $\gamma_i$ on early change in log UP, $\mu_\delta$ represents the intercept of this regression and the $\varepsilon_{\delta i}$ represent residuals from the regression model. The parameters $\beta$ and $\mu_\delta$ were assumed to have diffuse prior distributions, and the $\varepsilon_{\delta i}$ were assumed to be independent of each other and normally distributed with mean 0 and standard deviation $\sigma_\varepsilon$, where the prior distribution of $\sigma^2_\varepsilon$ was taken to be same inverse gamma distribution which assigned probabilities of 1/3 each to small, intermediate, and large values of $\sigma^2_\varepsilon$ as described above for σ($\gamma$) and σ($\delta$). The key parameter from the meta-regression is the slope coefficient $\beta$, which characterizes the proportional change in the HR defining the treatment effect on the clinical outcome associated with a given proportional change in the magnitude on the treatment effect on the geometric mean change in UP.

We also fit the Bayesian model with the following modified version of M1 to relate the treatment effect on the clinical outcome to the approximate treatment effect on the change in UP in absolute units of grams/day:

$$\delta_i = \beta \times [\mathrm{BUP_i} \times \gamma_i] + \mu_\delta + \varepsilon_{\delta i}, \qquad i = 1, 2, \ldots, 26, \qquad \text{(M1)}$$

where $BUP_i$ represents the median baseline UP for the $i^{th}$ study by treatment comparison combination.

### A.4. Treatment effect Ratios:

*Overview.* Analyses of ratios of the treatment effects on proteinuria to treatment effects on clinical outcomes can be viewed as a variant of the Trial Level analysis for the setting in which there is limited variation in the treatment effects on the surrogate endpoint. The treatment effect ratios have also been referred to as relative effects. The meta-regression coefficient $\beta$ of the trial level analysis described above characterizes how *variation* in the treatment effect on the clinical outcome relates to *variation* in the treatment effect on the early change in log UP. If, as was the case for the RCTs considered in this report, the variation in the treatment effects on the early change in UP is relatively small, or is not clearly discernible due to small sample sizes in the individual trials, then the meta-regression coefficient will be imprecisely estimated, and this aspect of the trial level analysis will be uninformative. However, if relatively consistent treatment effects in the same direction are observed both for change in log UP *and* the clinical outcome across a wide range of treatments, this consistency may be viewed as providing evidence that similar reductions of UP by different methods were associated with similar effects of the treatments on the clinical outcome. The degree of this consistency can be assessed by considering the variation between intervention types in the ratio of treatment effects on the clinical and surrogate endpoints. Even if the variation in treatment effects on change in UP is relatively small, some evidence supporting validity of change in UP as a surrogate may be provided if the ratios of treatment effects on the clinical outcome and on change in UP exhibit a low variation between different types of treatment. Because the treatment effect ratios are generally highly imprecise for individual trials, we have limited their presentation in the primary manuscript to the ratios of pooled treatment effects for each of the 5 intervention types,

*Details for Analysis of Treatment Effect Ratios.* For each of the 26 study by treatment comparison combinations *i* with at least 15 events, we defined the treatment effect ratio as $\exp(\rho_i) = \exp(\delta_i)/\exp(\gamma_i)$, where $\delta_i$ denotes the log HR for the clinical outcome and $\gamma_i$ denotes the difference between the treatment and control groups in the mean changes in log UP from baseline to early follow-up. We estimated $\exp(\rho_i)$ as $\exp(\hat{\rho}_i) = \exp(\hat{\delta}_i)/\exp(\hat{\gamma}_i)$ where $\hat{\delta}_i$ and $\hat{\gamma}_i$ are as defined above, and we estimated the standard error in $\hat{\rho}_i$ by bootstrap resampling. We then applied the Bayesian mixed effects model to pool results across the study by treatment combinations belong to each of the 5 intervention types:

$$\rho_i \sim N(\rho_k, \sigma^2(\rho)),\ i = 1, 2, \ldots, 26, \qquad \text{(R1)}$$

$$\hat{\rho}_i \sim N(\rho_i, \sigma^2(\hat{\rho}_i)),\ i = 1,2, \ldots, 26. \qquad \text{(R2)}$$

where $\rho_k$ denotes the mean of the log transformed treatment effect ratios across the study by treatment comparison combinations belonging to the *k*th intervention type, for $k = 1, 2, 3, 4, 5$, $\sigma(\rho)$ represents the standard deviation of the log of the true treatment effect ratios about the means for each treatment type. The interpretation of $\sigma(\rho)$, which is the key parameter characterizing the amount of variation in the treatment effect ratios, is facilitated by noting that

it is the approximate coefficient of variation of the treatment effect ratios across studies. We assumed a diffuse prior for $\rho$ and the same inverse gamma prior for $\sigma^2(\rho)$ as for $\sigma^2(\delta)$ and $\sigma^2(\gamma)$. We assumed that the true log treatment effect ratios $\rho_i$ are mutually independent conditional on the treatment type, and that the $\hat{\rho}_i$ are also mutually independent (including $\hat{\rho}_i$ corresponding to different factors from the same studies in factorial designs) with the exception of the 3-arm drug group comparisons from the AASK and IDNT trials, in which case the correlation in the estimated treatment effect ratios were estimated from the bootstrap.

**A.5. Multi-Level Analyses Relating Surrogacy Parameters to Baseline UP.**

*Overview***.** The purpose of these analyses was to explore the dependence of the results of the individual and trial-level analyses above on baseline UP. A key feature of the preceding analyses is that the change in UP from baseline to early follow-up is expressed on the logarithmic scale. This logarithmic transformation improves the statistical properties of the analyses by reducing positive skewness in UP. The use of the log transformation also presumes that a given percent change in UP has the same implications for the clinical outcome irrespective of the level of baseline UP. However, because the same percent change in UP translates to a larger absolute change in UP (expressed in grams/day) at higher levels of baseline UP, it is also plausible that a given change in log UP would have a larger effect at higher levels of baseline UP, which would represent an interaction between baseline UP and the change in log UP. In addition, several reports from individual RCTs and from meta-analyses of RCTs have noted the presence of interactions between the treatment and baseline UP for the clinical outcome.[16-18] Accordingly, we extended the above individual-level and trial-level analyses to multi-level analyses which explore the dependence of the results on baseline UP. In particular, we investigated models which related baseline UP to a) the Cox regression coefficient relating the clinical outcome to change in log UP (individual association), b) the treatment effects on change in log UP, and c) the treatment effects on the clinical outcome. For these analyses, we categorized baseline UP as either < 1 g/day, 1-3 g/day, or > 3 g/day.

We also considered an alternate formulation of the trial level analysis in which related treatment effects on the clinical outcome to approximate treatment effects on the absolute change in UP.

*Details*. We first describe the multi-level analyses characterizing the dependence of treatment effects on the clinical outcome on the baseline proteinuria category, and then describe modifications of this model for assessing the dependence of treatment effects on change in proteinuria and of individual level association on the baseline UP category. For the clinical outcome, we first estimated log HRs (and associated standard errors) relating the clinical outcome to treatment assignment separately for each of the three baseline UP subgroups within each of the 29 study by treatment-comparison combinations as described in Section A2. Because the analyses were stratified both by the baseline UP subgroup, we were able to assume that the sampling errors of the log HRs were statistically independent between the three UP subgroups for each study. In cases where a given study had too few subjects to provide an estimate of the log HR in one or more of the UP subgroups, we assigned a log HR of 0 and a very large SE of 10, which assured that that the studies in question would not have a detectable influence the final result for that UP subgroup. We then fit the Bayes model:

$$\delta_{ki} = \mu_k + \beta_i + A_k \times \xi_i, k = 1, 2, 3, \ i = 1, 2, \ldots, 29, \quad \text{(J1)}$$

$$\beta_i \sim N(0, \sigma^2(\mu)), \quad i = 1, 2, \ldots, 29, \quad \text{(J2)}$$

$$\xi_i \sim N(0, \sigma^2(\xi)), \quad i = 1, 2, \ldots, 29, \quad \text{(J3)}$$

$$\hat{\delta}_{ki} \sim N(\delta_{ki}, \sigma^2(\hat{\delta}_{ki})), 1, 2, 3, i = 1,2, \ldots, 29, \quad \text{(J4)}$$

where $A_1 = -1$, $A_2 = 0$, and $A_3 = 1$ represent coefficients of the $\xi_i$ in for the baseline UP < 1 g/day and baseline UP > 3 g/day subgroups, $\delta_{ki}$ denotes the true treatment effect on the clinical outcome (expressed as the log HR) for the $k^{th}$ baseline UP subgroup in the $i^{th}$ study, the $\mu_k$ represent the mean log HRs across the 29 study by treatment- comparison combinations for the three baseline UP groups, the $\beta_i$ account for variation in the overall log HRs between studies, and the $\xi_i$ account for variation in the baseline UP group by treatment interactions across studies. The objective of the analysis was to determine the posterior distributions of the $\mu_k$ and of the difference between the $\mu_k$ in order to characterize the dependence of the treatment effect HRs on the baseline UP categories.

We used essentially the same formulation to characterize the dependence on baseline UP of the treatment effects on the early change in log UP and of individual level associatoin. We also evaluated the relationship of the baseline UP category with change in UP on its raw scale in grams/day by approximating the mean treatment effect on change in UP in g/day within the $k^{th}$ baseline UP group as $GM(BUP_k) \times \exp(\mu_k)$, where in this case $\mu_k$ represents the log of the geometric mean ratio comparing the early changes in log(UP) between the treatment and control groups, and $GM(BUP_k)$ represents the geometric mean baseline UP for the $k^{th}$ baseline UP subgroup across the 29 study by treatment-comparison combinations. Finally, we performed a separate set of analyses corresponding to those described above in which baseline UP was treated as a continuous variable (based on log transformed baseline UP) rather than grouping it into three categories. These analyses were used to provide separate estimates of the relationship of baseline UP with individual level association and with treatment effects on early change in UP and the clinical outcome which corresponded to within study and between study interactions with baseline UP. The within study analyses can be viewed as controlling for study, and consequently also for the treatment type, as each study by treatment combination included only a single treatment type. The results of these analyses are provided in Supplemental Table 6.

### A.6. Prentice-Freedman Approach.

*Overview*. As in the analyses described above, our analyses of the proportion of the treatment effect explained (PTE) under the Prentice-Freedman approach can also be viewed as comprising two steps: First, separate analyses estimated the PTEs for each treatment comparison within each study, followed by pooled analyses that provided PTEs for each of the five intervention types. However we used stratified Cox regression analyses rather than random effects Bayesian analyses to provide pooled estimates of the PTEs across studies. This is because estimation of the

PTE requires fitting two separate Cox regressions with different predictor variables for each study – first without and then with adjustment for the early change in log UP. Because it is unlikely that both of these Cox models can hold simultaneously,[19] a Bayesian analysis under a single unified model is problematic. By contrast, it is possible to estimate Cox regression coefficients under alternative models, both for individual studies and for pooled analyses incorporating multiple studies, and then perform statistical inferences for the PTEs using robust empirical estimates of standard errors which are independent of the assumed models. Even though, strictly speaking, it is unlikely that the assumptions of both Cox regression models with and without adjustment for change in UP are simultaneously satisfied, the Cox regression coefficients are well-defined under both Cox regressions, and valid statistical inferences for the computed PTEs can be obtained.[19]

*Details*. In the first step of the analyses, we fit separate Cox regressions for each independent treatment comparison within each study, first adjusting only for baseline log UP, and subsequently adjusting for the early change in log UP in addition to baseline log UP. The PTE was calculated within each study as 1 minus the ratio of the log transformed Cox regression coefficients for the treatment with and without adjusting for early change in proteinuria. We used the method of Lin, Fleming and DeGruttola[19] to obtain a robust sandwich-type empirical estimate of the covariance matrix of the estimated Cox regression coefficients for the treatment, with and with adjustment for change in log UP, and then applied the delta method to estimate the standard error and obtain 95% confidence limits for the PTE. Because the interpretation of the PTE requires the presence of a treatment effect with adjustment for the surrogate, we reported the PTEs only for those studies in which the p-value for the analysis not controlling for change in UP was smaller than 0·10. We subsequently obtained pooled PTEs and associated 95% confidence intervals for each of the five intervention types by repeating the above procedure for joint analyses of the studies in each of the intervention types, where the baseline hazards of the Cox regressions were stratified by study. Because each study had no more than one treatment comparison within a given treatment type, we were able to apply the simplifying assumption of independent observations for each patient in the pooled analyses of each treatment type. Finally, noting that the PTE is subject to bias from confounding from factors that jointly influence both the change in log UP and the clinical outcome[1,9,10], we repeated the above analyses adjusting for five baseline covariates (age, sex, serum creatinine and mean arterial pressure) in addition to baseline log UP.

**B. Sensitivity Analyses.**

The purpose of these analyses was to address potential violations of parametric assumptions required for using the hazard ratio as the metric for characterizing treatment effects and individual level association, including:

a) Similar individual level association in the treatment and control groups within each study
b) Proportional hazards for treatment comparisons and for the early change in log UP in the Cox regression models

c) An approximately linear association between the log HR and early change in log transformed UP

The sensitivity analyses included:

i) Fitting extended Cox regression models which incorporated allowed potentially different hazard ratios for the treatment effects and for change in log UP during early and later follow-up times
ii) Fitting extended Cox regression models with linear spline terms to allow for nonlinear effects of early change in log UP
iii) Fitting extended Cox regression models for individual level association which stratified the baseline hazard function by treatment group rather than fitting a coefficient for treatment
iv) Fitting extended Cox regression models evaluating individual level association separately in each treatment group

We also compared the model-based estimated standard errors from the Cox regression analyses to bootstrap estimates of the standard errors, which should be robust to violations of proportional hazards and other modeling assumptions. In general, while some deviations from modeling assumptions did occur in some of the studies, none of these deviations was substantial enough to effect the interpretation of the overall results.

Finally, we have focused in this report on the complete clinical composite of doubling of serum creatinine, ESRD, or death. We included death in the primary composite outcome because of its clinical relevance and due to the risk of bias from informative censoring of deaths stemming from the strong association between renal disease progression and cardiovascular mortality risk, and the expectation that change in proteinuria and several of the evaluated interventions may influence both cardiovascular and renal outcomes. We conducted a parallel set of analyses using renal composite of doubling of serum creatinine or ESRD as the outcome while censoring death. Results with the renal composite closely paralleled those for the full composite including death.

**Item S1 References**

**1.** Joffe MM, Greene T. Related causal frameworks for surrogate outcomes. *Biometrics.* 2009;65(2):530-538.
**2.** Stevens LA, Greene T, Levey AS. Surrogate end points for clinical trials of kidney disease progression. *Clin J Am Soc Nephrol.* 2006;1(4):874-884.
**3.** Burzykowski T, Molenberghs G, Buyse M, eds. *The Evaluation of Surrogate Endpoints*. New York: Springer; 2005.
**4.** Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in medicine.* 1997;16(17):1965-1982.
**5.** Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. *Statistics in medicine.* 2005;24(2):163-182.
**6.** Gail M, Pfeiffer R, van Houwelingen H, Carroll R. On Meta-Analytic Assessment of Surrogate Outcomes. *Biostatistics.* 2000;1:231-246.

**7.** Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine.* 1989;8(4):431-440.
**8.** Freedman L, Graubard B, Schatzkin L. Statistical Validation of Intermediate Endpoints for Chronic Diseases. *Statistics in medicine.* 1992;11:167-178.
**9.** Frangakis C, Rubin D. Principal stratification and causal inference. *Biometrics.* 2002;58:21-29.
**10.** Taylor JM, Wang Y, Thiebaut R. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics.* 2005;61(4):1102-1111.
**11.** Fleming TR. Surrogate endpoints and FDA's accelerated approval process. *Health Aff (Millwood).* 2005;24(1):67-78.
**12.** Katz R. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics.* 2004;1(2):189-195.
**13.** Thompson A. Proteinuria as a surrogate end point--more data are needed. *Nat Rev Nephrol.* 2012;8(5):306-309.
**14.** Steel PDG, Kammeyer-Mueller J. Bayesian Variance Estimation for Meta-Analysis: Quantifying Our Uncertainty. *Organizational Research Methods.* 2007.
**15.** *SAS/STAT® 9.3 User's Guide* [computer program]. Cary, NC: SAS Institute Inc; 2011.
**16.** Klahr S, Levey AS, Beck GJ, et al. The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. Modification of Diet in Renal Disease Study Group. *N Engl J Med.* 1994;330(13):877-884.
**17.** Wright JT, Jr., Bakris G, Greene T, et al. Effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the AASK trial. *JAMA.* 2002;288(19):2421-2431.
**18.** Jafar TH, Stark PC, Schmid CH, et al. Progression of chronic kidney disease: the role of blood pressure control, proteinuria, and angiotensin-converting enzyme inhibition: a patient-level meta-analysis. *Ann Intern Med.* 2003;139(4):244-252.
**19.** Lin D, Fleming T, De Gruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine.* 1997;16:1515-1527.