U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Sciences
Office of Biostatistics

SECONDARY STATISTICAL REVIEW MEMORANDUM

Biomarker Name	Total Kidney Volume (TKV)
Context of Use	Baseline TKV can be applied as a prognostic biomarker for clinical trial enrichment in Autosomal Dominant Polycystic Kidney Disease (ADPKD)
Submitter	Polycystic Kidney Disease Outcomes Consortium
Briefing Documents Date(s)	March 20, 2014
Medical Division	Division of Cardio-Renal Drug Products
From:	Sue-Jane Wang, Ph.D.
Date:	May 14, 2015

Subject: Qualification of Total Kidney Volume (TKV) as a prognostic biomarker for the following Proposed Context of Use

- **General Area:** Clinical trial enrichment in Autosomal Dominant Polycystic Kidney Disease (ADPKD)
- **Target Population for Use:** Patients with ADPKD
- **Stage of drug development for use:** All clinical stages of ADPKD drug development, including proof of concept, dose-ranging, and confirmatory clinical trials
- **Intended application:** Baseline TKV can be applied as a prognostic biomarker that, in combination with patient age and baseline estimated Glomerular Filtration Rate (eGFR), can be used to help identify those ADPKD patients who are at the greatest risk for a substantial decline in renal function defined as (1) 30% worsening of eGFR, (2) 57% worsening of eGFR (equivalent to doubling of serum creatinine), or (3) End-Stage Renal Disease (ESRD, defined as dialysis or transplant). This biomarker will be used as an inclusion criterion in clinical trials to identify patients likely to show a clinically relevant decline in kidney function during the duration of the trial. Data are provided showing the calculated risk of each of these outcomes of declining renal function depending on age, total kidney volume, and baseline eGFR. Tables will be used by clinical trial researchers to determine the inclusion criteria to help select patients who are likely to reach the clinical endpoint of interest within a timeframe practical for the trial. These criteria include the optimum age, TKV, and eGFR for selecting subjects to be enrolled in the clinical trial.

I. Background

Following the completion of the statistical review (dated February 18, 2015) written primarily by Dr. John Lawrence, the clinical team met with the statistical team on March 13, 2015 to further discuss the question of what the added value for clinical utility can be obtained by baseline TKV as a prognostic biomarker in terms of clinical trial enrichment in patients with Autosomal Dominant Polycystic Kidney Disease (ADPKD), that is, whether baseline TKV can be used as a prognostic biomarker together with patient age and baseline estimated Glomerular Filtration Rate (eGFR) to help identify the ADPKD patients who are at the greater risk of having a substantial decline in renal function (a confirmed 30% decline in eGFR). This secondary review is written to stipulate more for this question.

II. Improved predictive performance from baseline TKV

In the original statistical review dated February 18, 2015, Table 9 indicates that there is a statistically significant improvement in the predictive performance quantified by Uno's C-statistics (2011) from Model-2 ("C no TKV") to Model-3 ("C with log(TKV)"). This improvement is consistently shown across the follow-up time in years. The comparison between the base model (Model-2) and the expanded model (Model-3) to assess improvement on predictive performance is important. This is because prior to a clinical trial, a subject's event status during the course of the trial is unknown.

C-statistic for time to event outcome measures the probability of concordance between patient's risk score (computed based on baseline covariates, in our case, it is a weighted combination of effects from baseline covariates, e.g., age, eGFR, tkv, through Cox regression) and the time to event. Thus, the model that has a higher probability of concordance will have a larger value of C-statistic. We employ this C-statistic to quantify an added value, if any, of a baseline covariate or marker for predicting an event risk on top of existing covariates.

A sensitive tool for selecting patients for enrollment into the trial is one that yields a higher predicted probability of event for subjects who are at a higher risk of event and a lower predicted probability of event for subjects who are at a lower risk of event; equivalently, a lower predicted probability of no-event for subjects who are at a higher risk of event and a higher predicted probability of no-event for subjects who are at a lower risk of event.

- Predicted probability of having a confirmed 30% decline in eGFR

If baseline TKV has no added value, both Model-2 and Model-3 are expected to yield very similar predicted probability of having a confirmed 30% decline in eGFR and very similar predicted probability of not having a confirmed 30% decline in eGFR. A crude comparison of such predicted probability can be made graphically by plotting the predicted probabilities from the two models in a scatter plot with a 45 degree line on which the points are the predicted probabilities from the two models that are equal.

As shown in Figure 1 and Figure 2, the predicted probabilities of having a confirmed 30% decline in eGFR regardless of whether subjects have an event or not do not evenly spread around the 45 degree line. It appears that approximately 42% ADPKD patients have a larger predicted probability at year 3 (Figure 1) and 44% at year 5 (Figure 2) from Model-3 than from Model-2; both are significantly different from 50% with a nominal p-value < 0.0001. This indicates the predictive performance differs between Model-2 and Model-3. It appears that statistical uncertainty on the improved predictive performance increases from year 3 to year 5.

Figure 1. Predicted probabilities of having a confirmed 30% decline in eGFR at year-3 between Model-2 and Model-3

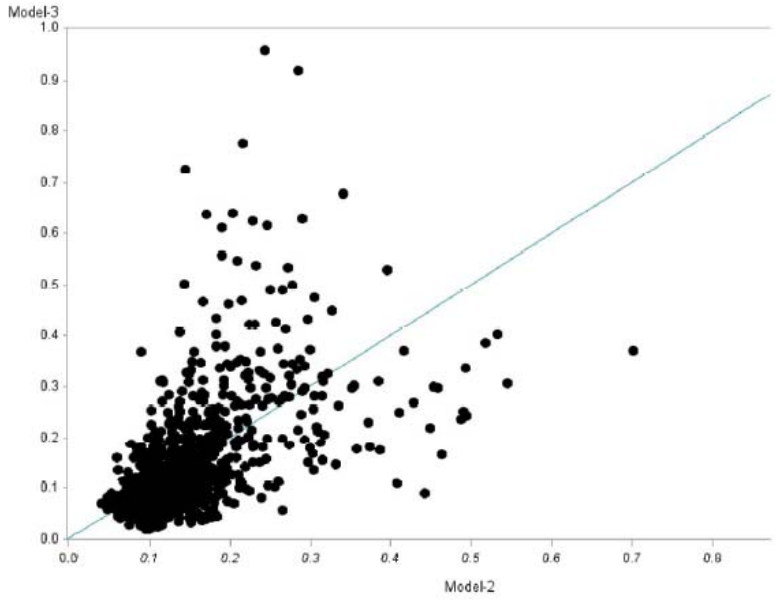
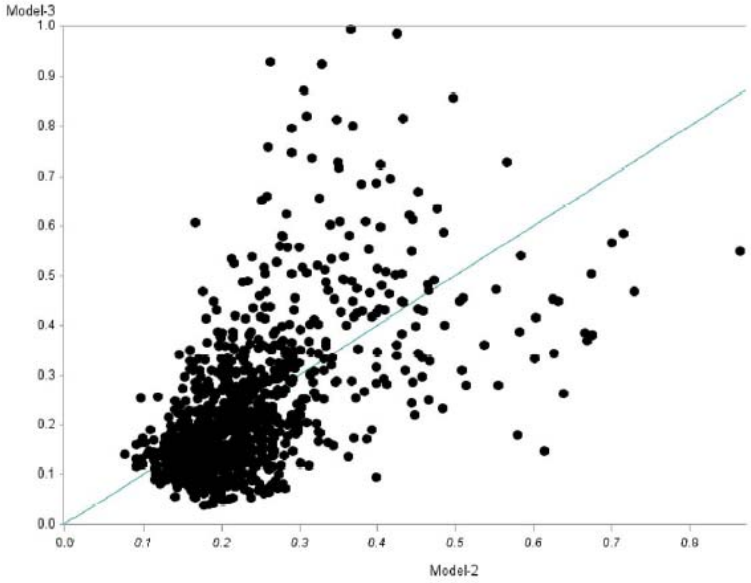
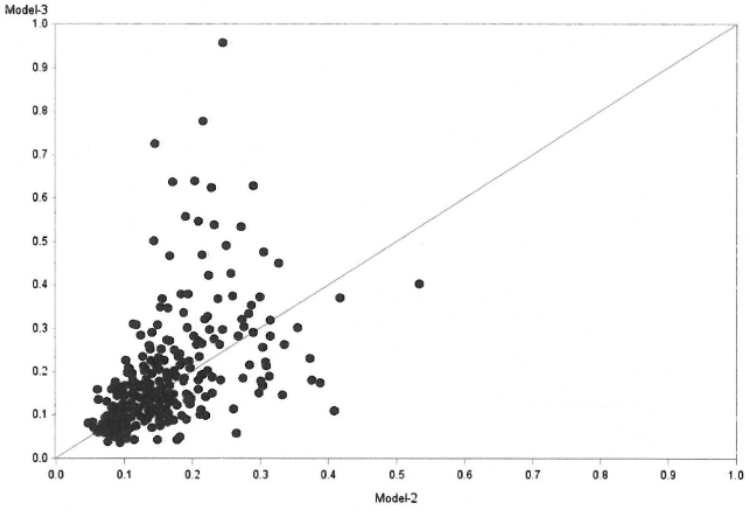


Figure 2. Predicted probabilities of having a confirmed 30% decline in eGFR at year-5 between Model-2 and Model-3



When only those subjects with a confirmed 30% decline in eGFR from Figure 1 is considered, the percent of subjects whose probabilities of having the event (a confirmed 30% decline in eGFR) predicted by Model-3 is larger than that predicted by Model-2, see Figure 3. This percentage is 57% at year 3 and is 56% at year 5.

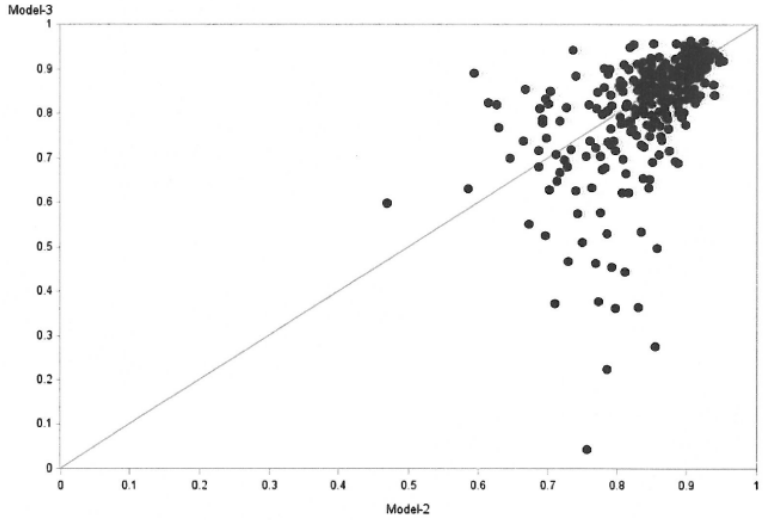
Figure 3. Predicted probabilities of having a confirmed 30% decline in eGFR in those subjects with the event at year-3 between Model-2 and Model-3



- Predicted probability of not having a confirmed 30% decline in eGFR

Equivalently, a comparison between the prediction models can be made on the predicted probability of not having a confirmed 30% decline in eGFR in the various subgroups defined by age, baseline eGFR and baseline TKV between Model-2 (without baseline TKV) and Model-3 (with baseline TKV).

Figure 4. Predicted probabilities of not having a confirmed 30% decline in eGFR in those subjects without the event at year-3 between Model-2 and Model-3



When only those subjects without a confirmed 30% decline in eGFR from Figure 1 is considered, it becomes clear that a larger percent of subjects has the probabilities of not having the event (a confirmed 30% decline in eGFR) predicted by Model-3 than Model-2, see Figure 4. This percentage is 65% at year 3 and is 63% at year 5.

The predicted probabilities of not having a confirmed 30% decline in eGFR from Model-2 and from Model-3 produced by Dr. John Lawrence are summarized in Table 1. Note that the results in "Based on Model-3" of Table 1 are reproduced from Table 13 of the original statistical review. In Table 1, for subjects with baseline TKV < 1L, the probability of not having a confirmed 30% decline in eGFR predicted from Model-3 is slightly higher to higher than that from Model-2 over the follow-up time. In contrast, for subjects with baseline TKV ≥ 1L, their average predicted probabilities of not having a confirmed 30% decline in eGFR from Model-3 is generally lower than that from Model-2. If the baseline TKV at 1L cutoff value can be considered for classifying ADPKD patients at a relatively lower risk (< 1L) versus a relatively higher risk (≥ 1L), then, these observations suggests an improvement from Model-2 to Model-3.

There was a question of how the estimates from Model-2 and/or Model-3 fair against Kaplan-Meier (K-M) estimates. Note that in Table 14 of the original statistical review, the K-M estimates for each subgroup (defined by age, baseline eGFR and baseline TKV) are generated within that subgroup only; therefore, these estimates use the subgrouping information. To address this question further, Table 2 presents the estimated predicted probabilities of not having a confirmed 30% decline in eGFR calculated at year-3 from K-M estimates that ignore the baseline covariates completely, Model-2 and Model-3, respectively. Certainly, these K-M estimates derived not incorporating any baseline covariates do not differ among subgroups.

For subjects with TKV < 1L, Table 2 shows that the predicted probabilities of not having a confirmed 30% decline in eGFR are lowest with K-M estimates, followed by Model-2, then followed by Model-3 in the sub-category defined by eGFR ≥ 50 mL/min/1.73m^2 (relatively lower risk) in both age < 40 years and age ≥ 40 years. For subjects with TKV ≥ 1L, a reversed trend is observed in the sub-category defined by eGFR < 50 mL/min/1.73m^2 (relatively higher risk) in both age < 40 years and age ≥ 40 years. These predicted probabilities suggest that both baseline eGFR and baseline TKV may have better predictive performance than age.

[The remaining space of this page is intended to be blank]

Table 1. Predicted probability of not having a confirmed 30% decline in eGFR in subgroups defined by age, baseline eGFR and baseline TKV by follow-up time in years using Model-2 (without baseline TKV) and Model-3 (with baseline TKV)

Predicted probability	TKV < 1L	TKV < 1L	TKV < 1L	TKV < 1L	TKV ≥ 1L	TKV ≥ 1L	TKV ≥ 1L	TKV ≥ 1L
	Age < 40yrs	Age < 40yrs	Age ≥ 40yrs	Age ≥ 40yrs	Age < 40yrs	Age < 40yrs	Age ≥ 40yrs	Age ≥ 40yrs
Followup time (years)	eGFR ≥ 50	eGFR < 50	eGFR ≥ 50	eGFR < 50	eGFR ≥ 50	eGFR < 50	eGFR ≥ 50	eGFR < 50
Based on Model-2								
1	0.964	0.935	0.969	0.901	0.969	0.952	0.969	0.916
2	0.931	0.879	0.945	0.831	0.941	0.911	0.943	0.854
3	0.870	0.798	0.893	0.719	0.882	0.832	0.843	0.741
4	0.825	0.749	0.850	0.644	0.837	0.773	0.811	0.659
5	0.784	0.692	0.821	0.588	0.800	0.725	0.821	0.603
Based on Model-3								
1	0.971	0.982	0.977	0.954	0.960	0.916	0.961	0.914
2	0.944	0.965	0.958	0.919	0.922	0.851	0.928	0.851
3	0.895	0.938	0.916	0.853	0.846	0.735	0.858	0.730
4	0.860	0.920	0.879	0.803	0.787	0.654	0.800	0.640
5	0.828	0.897	0.854	0.765	0.738	0.594	0.758	0.581

Table 2. Predicted probability of not having a confirmed 30% decline in eGFR in subgroups defined by baseline covariates (age, eGFR and TKV) at year-3 using Kaplan-Meier estimates ignoring covariates, Model-2 (without baseline TKV), Model-3 (with baseline TKV)

Predicted probability	TKV < 1L	TKV < 1L	TKV < 1L	TKV < 1L	TKV ≥ 1L	TKV ≥ 1L	TKV ≥ 1L	TKV ≥ 1L
	Age < 40yrs	Age < 40yrs	Age ≥ 40yrs	Age ≥ 40yrs	Age < 40yrs	Age < 40yrs	Age ≥ 40yrs	Age ≥ 40yrs
Estimation method	eGFR≥50	eGFR<50	eGFR≥50	eGFR<50	eGFR≥50	eGFR<50	eGFR≥50	eGFR<50
From K-M estimates	0.855	0.855	0.855	0.855	0.855	0.855	0.855	0.855
From Model-2	0.870	0.798	0.893	0.719	0.882	0.832	0.843	0.741
From Model-3	0.895	0.938	0.916	0.853	0.846	0.735	0.858	0.730

III. Observed added clinical utility of baseline TKV as a prognostic biomarker

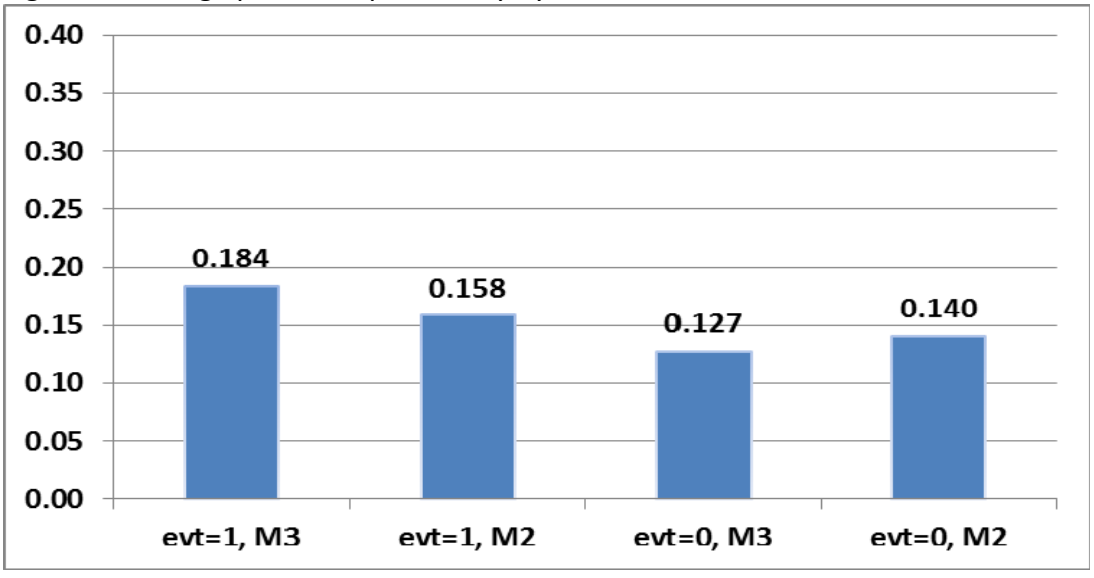
We select year-3 as the time point to elucidate potentially additional clinical utility when the baseline TKV is incorporated besides age and baseline eGFR for clinical trial enrichment consideration. To assess if there is an added clinical utility of baseline TKV as a prognostic biomarker, we will focus on the comparison between Model-2 (not including baseline TKV) and Model-3 (including baseline TKV). Below we quantify the differences between Model-2 and Model-3 via model discrimination and potential reclassification.

- Model Discrimination

Differences between Model-2 and Model-3 can be quantified by looking into numerical changes in predicted probability in a few different ways. We first compare the average predicted probability in subjects with and in subjects without a confirmed 30% decline in eGFR.

Figure 5 depicts the predicted probabilities for subjects experiencing a confirmed 30% decline in eGFR (evt=1) and for subjects without experiencing the event (evt=0) at year-3 with Model-2 (M2) and Model-3 (M3). By including baseline TKV in the model, there is a 2.6% absolute increase in the average predicted probability for the subjects experiencing a confirmed 30% decline in eGFR and a 1.3% absolute decrease in the average predicted probability for the subjects without experiencing a confirmed 30% decline in eGFR. This results in an absolute change of 3.9% (= 2.6% + 1.3%). Interpretation of this absolute change depends on the background event rate observed in the data. A relative measure, (0.184 – 0.127) / (0.158 – 0.14) – 1 = 2.2, suggests that the difference in predicted probabilities between events and non-events is about 2.2-fold more in Model-3 compared to Model-2. The concept introduced by Pencina et al. (2008) is borrowed. Instead of stating that such absolute change or relative change indicates an improvement from Model-2 to Model-3, we only state that there is model discrimination.

Figure 5. Average predicted probability by event status for Model-2 and Model-3 at year-3



- Potential Reclassification

As an alternative, we can look into the potential upward shifting and downward shifting movements of the predicted probabilities if subjects are to be reclassified using some cutoff values. Two cutoffs are chosen to illustrate the potential reclassification from Model-2 to Model-3: a 3-risk-category cutoff of 10% and 15%, and another 3-risk-category cutoff of 10% and 20%. For example, using a 3-risk-category cutoff of 10% and 15%, there are three risk categories: (i) predicted probability (pp) < 0.10, (ii) 0.10 ≤ pp < 0.15, and (iii) pp ≥ 0.15. The "shift-upward" indicates that subjects classified into (i) with Model-2, for example, are reclassified into (ii) or (iii) with Model-3. Conversely, "shift-downward" indicates that subjects who are classified as (ii), for example, with Model-2 are reclassified into (i) with Model-3.

We also look into how the predicted probability shifts considering all possible cutoff values data provides, i.e., "Infinite cutoffs". The advantage of using "Infinite cutoffs" is that the total reclassification is not cutoff-dependent. However, the shortcoming is that it may lose its practical interpretation in terms of a biomarker's context of use for its clinical utility.

For the subjects with a confirmed 30% decline in eGFR at year-3, Model-3 yields a larger percentage for shift-upward than for shift-downward as compared with Model-2. Conversely, for the subjects without a confirmed 30% eGFR decline at year-3, Model-3 yields a larger percentage for shift-downward than for shift-upward. These numerical trends are observed in Table 3. Thus, the total reclassification as a result of adding baseline TKV as a predictor in the model (i.e., from Model-2 to Model-3) can then be calculated by summing the difference (% shift-upward - % shift-downward) for the subjects with a decline and the difference (% shift-downward - % shift upward) for the subjects without a decline.

Table 3. Examples of potential reclassification from Model-2 to Model-3 at year-3

Cutoff used	Shift-upward	Shift-downward
Among those who have a confirmed 30% decline in eGFR		
10% and 15% 3-risk-category	22%	19%
10% and 20% 3-risk-category	23%	16%
Infinite cutoffs*	Difference (% shift-upward – % shift-downward) = 12%	
Among those who do not have a confirmed 30% decline in eGFR		
10% and 15% 3-risk-category	12%	32%
10% and 20% 3-risk-category	11%	28%
Infinite cutoffs*	Difference (% shift-downward – % shift-upward) = 31%	

For instance, using the cutoffs of 10% and 20% shown in Table 3, the total reclassification in Table 4 is obtained from Table 3, (23% - 16%) + (28% -11%) = 23%. The advantage of using the cutoffs is to put the clinical utility of baseline TKV as a prognostic biomarker in context, such as, identify an ADPKD patient as potentially being at low risk if the predicted probability is less than 10%, at high risk if the predicted probability is at least 20% and higher, or at moderate risk otherwise. However, this requires that the cutoff value applied is clinically meaningful.

Table 4. Examples of total reclassification from Model-2 to Model-3 at year-3

Cutoff used	Total reclassification	95% confidence interval
10% and 15% 3-risk-category	22%	(13%, 31%)
10% and 20% 3-risk-category	23%	(15%, 32%)
Infinite cutoffs*	43%	(29%, 56%)

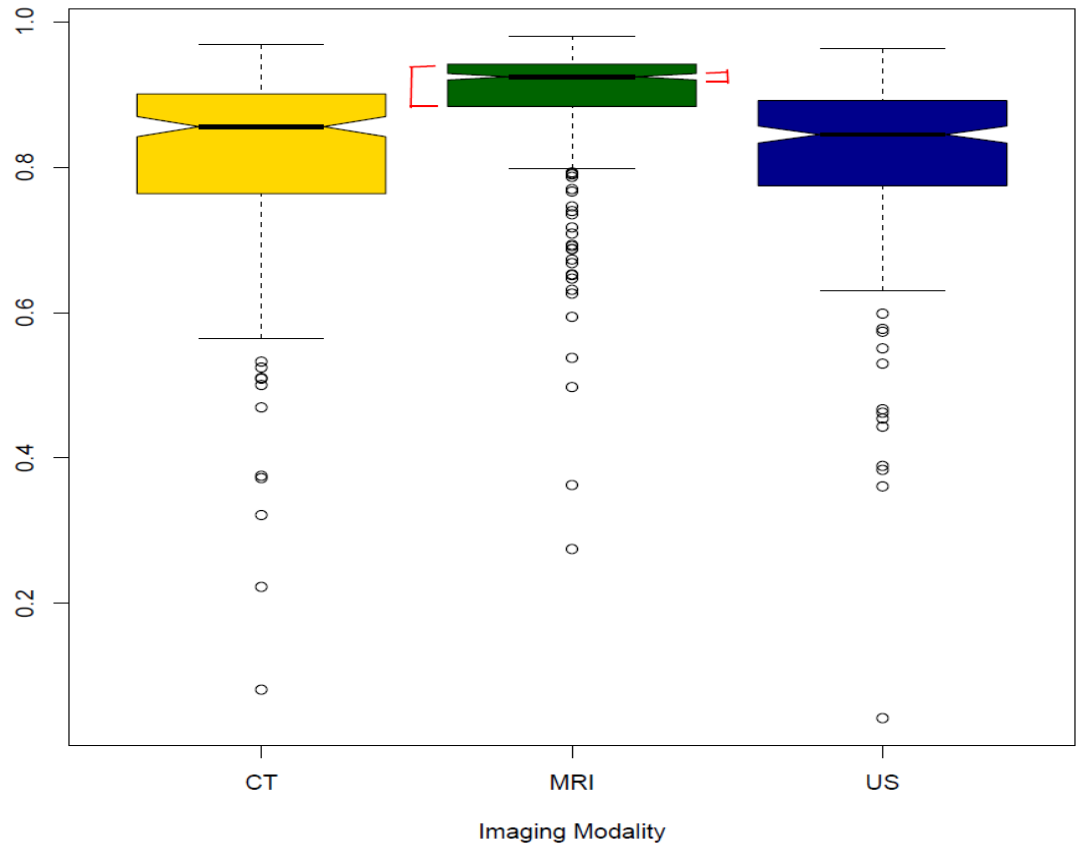
*can be viewed as "as many risk categories as there are estimated predicted probabilities"

Regardless of using specific cutoffs or not, Table 4 strongly suggests that there is a tangible total reclassification from use of Model-2 to use of Model-3 regarding the potential added value of baseline TKV, given age and baseline eGFR, in predicting the likelihood of having a confirmed 30% decline in 3 years for ADPKD patients.

IV. Precision of TKV measures by imaging modalities reflected on predicted probability

From the four data sources (Mayo, Emory, CRISP, Colorado), about 50% of the subjects had TKV measurements taken by MRI, 25% by CT and 25% by US, see Table 5 and the text on page 10 of the original statistical review of February 18, 2015). The predicted probability of not having a confirmed 30% decline in eGFR by imaging modality with Model-3 is depicted in Figure 6 shown with three notched box-plots.

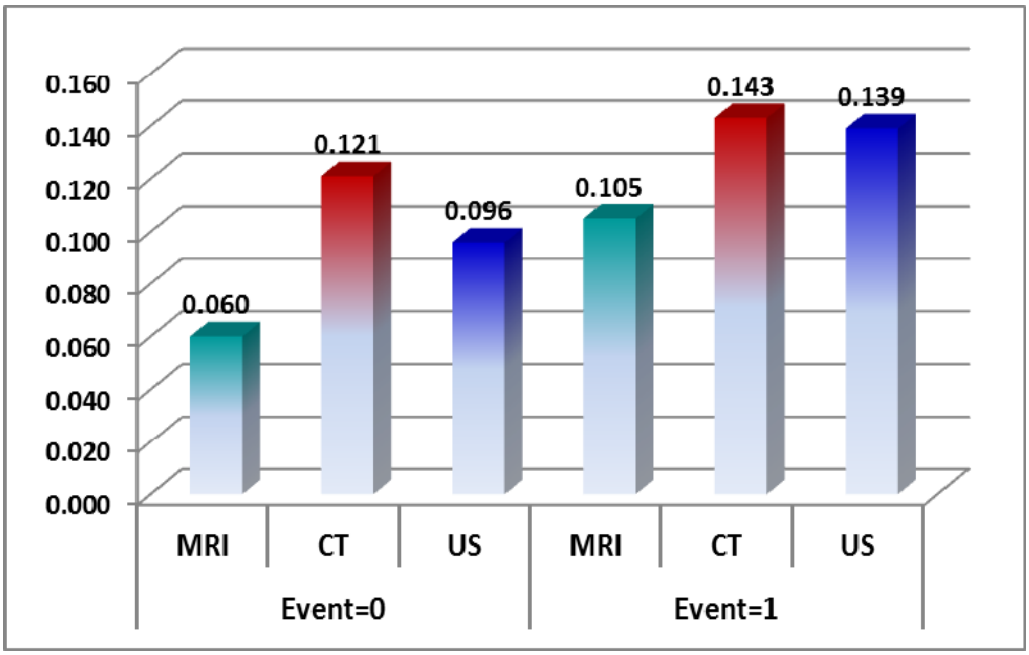
Figure 6. Predicted probability of not having a confirmed 30% eGFR decline (Model-3 at year-3)



From Figure 6, if two boxes' notches do not overlap, it suggests that there is strong evidence that the medians differ between the two distributions (Chambers et al., 1983). The interval estimates of the predicted probability for the MRI stratum based either on the notch height (the small distance shown in red line range to the right of MRI notched boxplot) or the interquartile range (the distance shown in red line range to the left of MRI notched boxplot) are much more precise than those by CT or US. This precision pattern is similarly observed with Model-2 (results not shown).

We also compare the standard deviation of the predicted probability among the three imaging modalities for Model-3, see Figure 7. The standard deviation is the smallest for MRI modality and the largest for CT modality. The standard deviation for US modality is closer to that for CT modality. The standard deviation pattern of the predicted probability is similarly observed between those having the event of a confirmed 30% decline in eGFR and those not. Such pattern is also observed with Model-2 (results not shown).

Figure 7. Precision by imaging modality expressed as standard deviation for Model-3 at year-3



V. Summary and Recommendation

We have shown a statistically significant improvement with use of baseline TKV based on C-statistic in the primary statistical review performed by Dr. John Lawrence. In this secondary review, we further compare the predictive performances between Model-2 and Model-3 by year-3 and quantify the added clinical utility of baseline TKV as a prognostic biomarker through model discrimination and potential reclassification by year-3.

Based on these further findings, we show that compared to Model-2 at year-3, Model-3 tends to yield a higher predicted probability of not having a confirmed 30% decline in eGFR for ADPKD patients with TKV < 1L and yield a lower predicted probability of not having a confirmed 30% decline in eGFR for ADPKD patients with TKV ≥ 1L. This finding is further illustrated via model discrimination. In addition, the total potential reclassification is shown to be positive, i.e., there is a higher percentage of being reclassified to higher risk categories for those having a confirmed 30% decline in eGFR and a higher percentage of being reclassified to lower risk categories for those not having a confirmed 30% decline in eGFR.

In this secondary review, we have also highlighted the caveats on the precision of the TKV imaging modality data reflected on the predicted probabilities at year-3. Specifically, the interval estimates of the predicted probability with MRI are much more precise than those by CT or US. The standard deviation of the predicted probability is the smallest for MRI modality and the largest for CT modality. The standard deviation for US modality is closer to that for CT modality. These observed patterns are similar between Model-2 and Model-3.

We conclude by repeating the concerns raised in the original statistical review on the heterogeneity of ADPKD populations and data quality included in the four data sources (Mayo, Emory, CRISP, and Colorado).

a) Differences in the sampling mechanism. The population was not a random sample from the population of people with diagnoses of ADPKD. In some cases, family members were recruited based on genotyping although they did not necessarily have a diagnosis. In addition, the people recruited for a clinical trial are not a random sample from the population;

b) In a clinical trial, baseline eGFR will often be calculated as an average of two or more values taken within a short time (one or two weeks) before randomization or first dose of study drug. This is done in part because of a moderate level of within subject variability. The average is more reliable. In this dataset, only one baseline value was used and it could have been measured many months (up to a year) after the defined time 0 (the time of the baseline TKV measurement);

c) Because an event cannot occur unless a measurement is observed, the timing and frequency of serum creatinine measurements can affect the event rate. In a clinical trial, serum creatinine measurements are taken at defined time points such as every 3 months or every 4 months. Although missing data is inevitable, there is an expectation that most people will follow that schedule. In this dataset, they were taken haphazardly whenever chosen by the subject or physician. In some cases, measurements were taken every day for many days in a row; in other cases, there were gaps of a year or more between measurements;

d) In a clinical trial, the endpoint of confirmed 30% change from baseline would usually need to be confirmed by the very next subsequent measurement. If the qualifying and

confirmatory measurement were not consecutive, then, an event does not count. However, in this dataset, event could count even when the qualifying and confirmatory events are years apart with many non-confirmatory measurements in between;

e) In a clinical trial, an event of confirmed 30% worsening of eGFR would also include more severe endpoints including need for initiation of dialysis or transplant. In this dataset, subjects with these more severe events were censored at the time of the event.

Sue-Jane Wang, May 14, 2015

Sue-Jane Wang, Ph.D.
Associate Director, Pharmacogenomics and Adaptive Design
Office Liaison for the CDER Biomarker Qualification Program
Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research

cc:
Lisa LaVange, Ph.D.
Director
Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research

References:

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 2011;30(10):1105-1117.

Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predicted ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008;27(2):157-172.

Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making 2006;26(6):565-574.

Chambers JM, Cleveland WS, Kleiner B, Tukey P A. *Graphical Methods for Data Analysis.* 1983. Wadsworth International Group, Belmont, CA.