SUPPLEMENTARY DATA

Plasma Mucin-1 (CA15-3) Levels in Autosomal Dominant Tubulointerstitial Kidney Disease due to *MUC1* mutations

Authors

Petr Vylet'al, PhD, Kendrah Kidd, MS, Hannah C. Ainsworth, PhD, Drahomíra Springer, PhD, Alena Vrbacká, PhD, Anna Přistoupilová, PhD, Rebecca P. Hughey, PhD, Seth L. Alper, MD, PhD, Niall Lennon, PhD, Steven Harrison, PhD, Maegan Harden, PhD, Victoria Robins, RN, BSN, Abbigail Taylor, BS, Lauren Martin, MSW, Katrice Howard, Ibrahim Bitar, PhD, Carl D. Langefeld, PhD, Veronika Barešová PhD, Hana Hartmannová, PhD, Kateřina Hodaňová, PhD, Tomáš Zima, PhD, Martina Živná, PhD, Stanislav Kmoch, PhD, Anthony J. Bleyer, MD, MS

Corresponding Author

Anthony J. Bleyer, MD, MS

Wake Forest School of Medicine

Section on Nephrology

Winston-Salem, NC, USA 27157

Fax: 336-716-4318

Phone: 336-716-4650

e-mail: ableyer@wakehealth.edu

Supplementary Data:

Supplementary Methods:

Statistical Analysis: Equivalence testing of CA15-3 levels between the reference and study populations were evaluated using Kolmogorov-Smirnov's D_{max} statistic. To account for family structure, we completed 5000 random samplings with replacement based on family status for each pair of populations (reference and control, reference and ATDKD-*MUC1*). Families were randomly assigned to datasets, with sample numbers (within each dataset) corresponding to the approximate ratio of the original data (e.g. 85/6850 for ATDKD-*MUC1*/reference and 249/6850 for control/reference population). Comparing the reference and control populations showed comparable CA15-3 distributions with an average D_{max} of 0.05±0.02 and minimum and maximum values of 0.02 and 0.14, respectively. A greater deviation (as expected) was observed between the reference and ATDKD-*MUC1* populations with an average D_{max} of 0.09±0.03 and larger minimum and maximum values of 0.03 and 0.26, respectively (**shown in Supp. Fig 7**).

We computed a logistic regression model in the training dataset to provide a parametric estimate of the probability of ADTKD-*MUC1*. We then applied this model to the validation dataset. These predictions were calculated under two models: 1) using only CA15-3 as the predictor and 2) with both CA15-3 and rs4072037 genotype (coded under a dominant model for the T-allele). ADTKD-*MUC1* classifications in the validation dataset were assigned based on a probability > 0.5. To summarize, we computed the overall error rate and the area under the receiver operator characteristic (ROC) curve for each model.

ADTKD-*MUC1* prediction using CA15-3 and rs4072037: Given the significant variation of CA15-3 levels by rs4072037 genotype (**shown in Supp. Fig. 8, Supp. Fig. 9**), we compared predictions under two logistic regression models: Model 1 (CA15-3 only) and Model 2 (CA15-3 with rs4072037 genotype). Logistic regression analysis of the training data by Model 2 (CA15-3 with rs4072037 genotype) yielded a receiver-operator curve (ROC) with an area under the curve (AUC) statistically

greater (0.8972 than that for Model 1 (0.797) (**shown in Supp. Fig. 10**). We used these models to calculate parametric estimations of the probability for ADTKD-*MUC1* (**shown in Supp. Fig. 11**). ADTKD-*MUC1* classifications were assigned based on a probability > 0.5. Respective overall error rates for Models 1 and 2 were 0.217 and 0.116. Validation dataset (n=67) error rates were 0.134 (Model 1, CA15-3) and 0.164 (Model 2, CA15-3 and rs4072037; **shown in Supp. Fig. 12**). Although the predicted probability curves for ADTKD-*MUC1* diagnosis based on logistic regression Model 2 clearly separate the two rs4072037 genotypes, neither logistic regression model provided classification rules of the high certainty required for utility in clinical diagnosis. Specifically, the maximum probability for ADTKD-*MUC1* in Model 1 reached 0.62 for a serum CA15-3 plasma level of 5 U/mL.

Supp. Table 1. Characteristics of the study population training and validation sets

	Training Set (n=267)			Validation Set (n=67)		
	ADTKD-*MUC1*	Control		ADTKD-*MUC1*	Control	
		ADTKD-*UMOD*	Genetically unaffected		ADTKD-*UMOD*	Genetically unaffected
Individuals, n	70	107	90	15	28	24
Male, n (%)	34 (49)	43 (40)	42 (47)	6 (40)	7 (25)	7 (30)
Age (y)	41.5±13.3	40.5±13.5	41.6±16.6	51.1±15.0	42.±4	43.5±14.4
CA 15-3 U/mL	8.6±4.3	14.0±5.4	15.5±6.3	8.8±4.2	14.9±5.1	13.2±3.8
eGFR (min/mL/1.73^2)	51.0±28.4	46.1±27.8	95.7±22.2	48.04±28.8	55.2±32.2	97.2±18.2
rs4072037						
CC, n (%)	32 (46)	29 (27)	17 (18)	4 (27)	6 (21)	3 (13)
CT, n (%)	33 (47)	54 (51)	50 (56)	9 (60)	15 (54)	15 (65)
TT, n (%)	5 (7)	24 (22)	23 (26)	2 (13)	7 (25)	6 (26)

Supp. Table 2. rs4072037 Genotypes and Corresponding CA15-3 Levels in the Study Population. CA15-3 levels are given as mean ± sd. Control population combines ADTKD-*UMOD* and genetically unaffected individuals.

	ADTKD-*MUC1* (n=85)		**Control Population (n=249)**	
rs4072037 genotype[a]	**n (%)**	**CA15-3 U/mL**	**n (%)**	**CA15-3 U/mL**
CC[b]	36 (42)	11.5±3.9	55 (22)	18.9±5.0
CT[c]	42 (50)	6.7±3.3	134 (54)	14.9±4.9
TT[c]	7 (8)	5.8±2.2	60 (24)	9.9±4.1

[a]Genotype differences between ADTKD-*MUC1* and control population (p=0.04), coded as a dominant model for the T allele (P=0.05)

[b] For individuals with the CC genotype, CA15-3 levels significantly differed between ADTKD-*MUC1* and control populations (P<0.001).

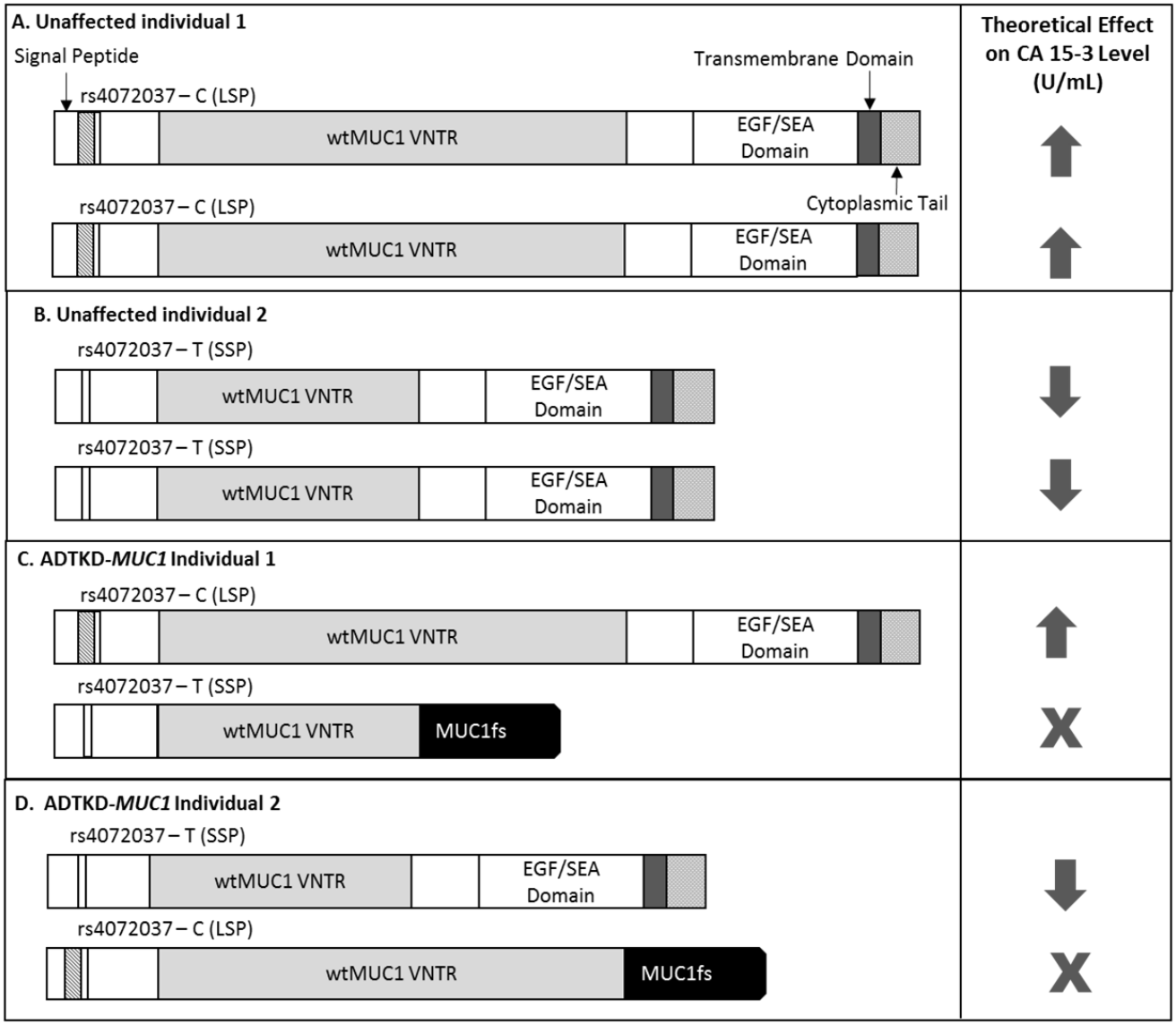
[C] For individuals with the T allele (CT or TT), CA15-3 levels significantly differed between ADTKD-*MUC1* and control populations (P<0.001). Note: there were too few individuals with TT genotype to stratify model by TT genotype.

Supp. Table 3. ADTKD-*MUC1* individuals with CA15-3 levels less than 5 U/mL. Low plasma MUC1 levels were identified in individuals at varying levels of eGFR.

	Age (years)	eGFR (ml/min/1.73^2)	CA15-3 (U/mL)
Training			
1	36.2	88.2	1.8
2	37.9	98.6	2.8
3	52.0	66.0	2.8
4	43.3	55.9	3.3
5	44.2	64.0	3.6
6	53.9	34.6	3.8
7	35.0	21.5	3.8
8	33.0	40.0	3.8
9	40.3	34.4	3.9
10	31.0	89.1	4.1
11	23.4	84.7	4.1
12	47.2	80.1	4.2
13	40.9	36.6	4.2
14	34.5	77.3	4.3
15	65.4	22.6	4.3
16	54.8	Peritoneal dialysis began at age 45	4.7
17	45.3	65.6	4.9
Validation			
1	26.9	49.2	3.3
2	62.5	82.5	4.6
Mean	42.5±11.4	60.6±24.6	3.8±0.8

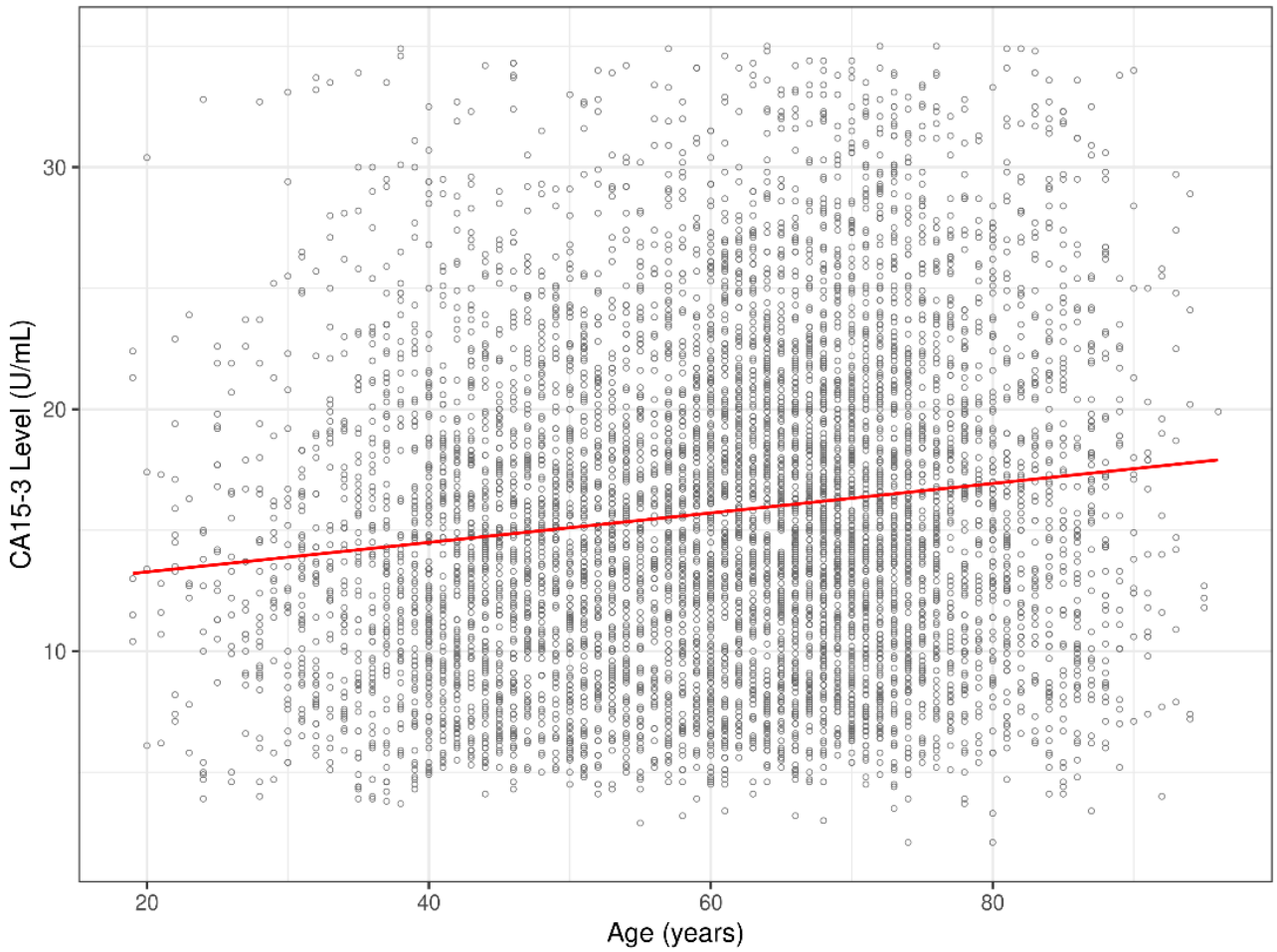
Supp. Fig. 1. Theoretical effect of rs4072037 phase on plasma CA15-3 levels.

MUC1 genotypes of an unaffected individual homozygous for the LSP rs4072037 SNP variant and of two individuals with *ADTKD-MUC1* heterozygous for the rs4072037 SNP variant. Plasma MUC1 levels (at right in panels) are affected only by the *wtMUC1* allele, as the mutated *MUC1fs* allele produces a protein retained intracellularly. Plasma MUC1 levels in the unaffected individuals homozygous for rs4072037 (**Supp. Fig. 1. a**) are determined by both C alleles or (**Supp. Fig. 1. b**) both T alleles. For ADTKD-*MUC1* individual 1 (**Supp. Fig. 1. c**), the C allele is in phase with wtMUC1, which is detected by the CA 15-3 assay. The VNTR is longer and the CA 15-3 level is increased. For ADTKD-*MUC1* Individual 2 (**Supp. Fig. 1. d**), the T allele is in phase with the wtMUC1, which is detected by the CA 15-3 assay. The VNTR associated with the T variant is in general shorter and will contribute to a lower CA 15-3 level. For simplicity, only one VNTR length is shown for each rs4072037 allele C and T.



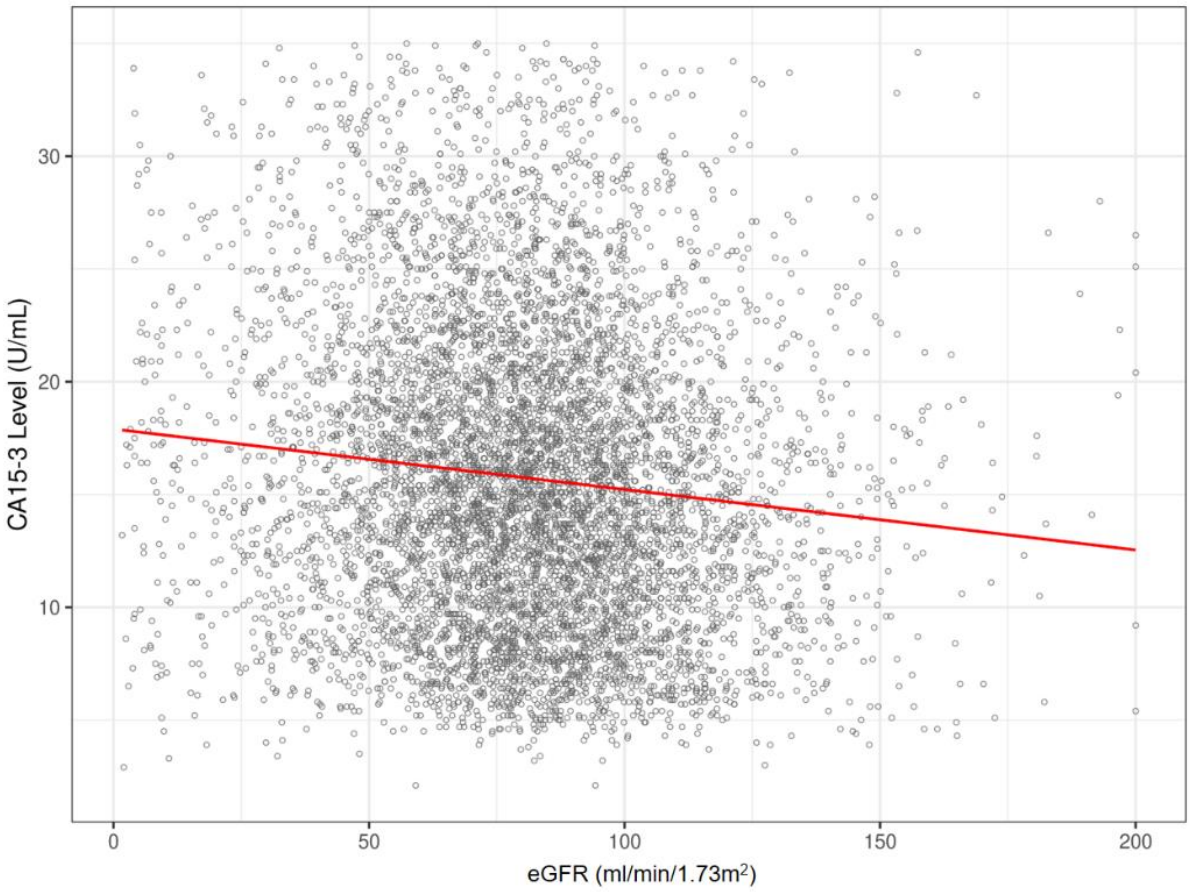
Supp. Fig. 2. Serum CA15-3 levels by age in the reference population.

Serum CA15-3 levels for 6,850 individuals are shown. While there was an association between serum CA15-3 vales and age ($p<0.001$), this accounted for only 1.2% of the overall variation, as can be seen by the wide variation in serum CA15-3 levels at each age.



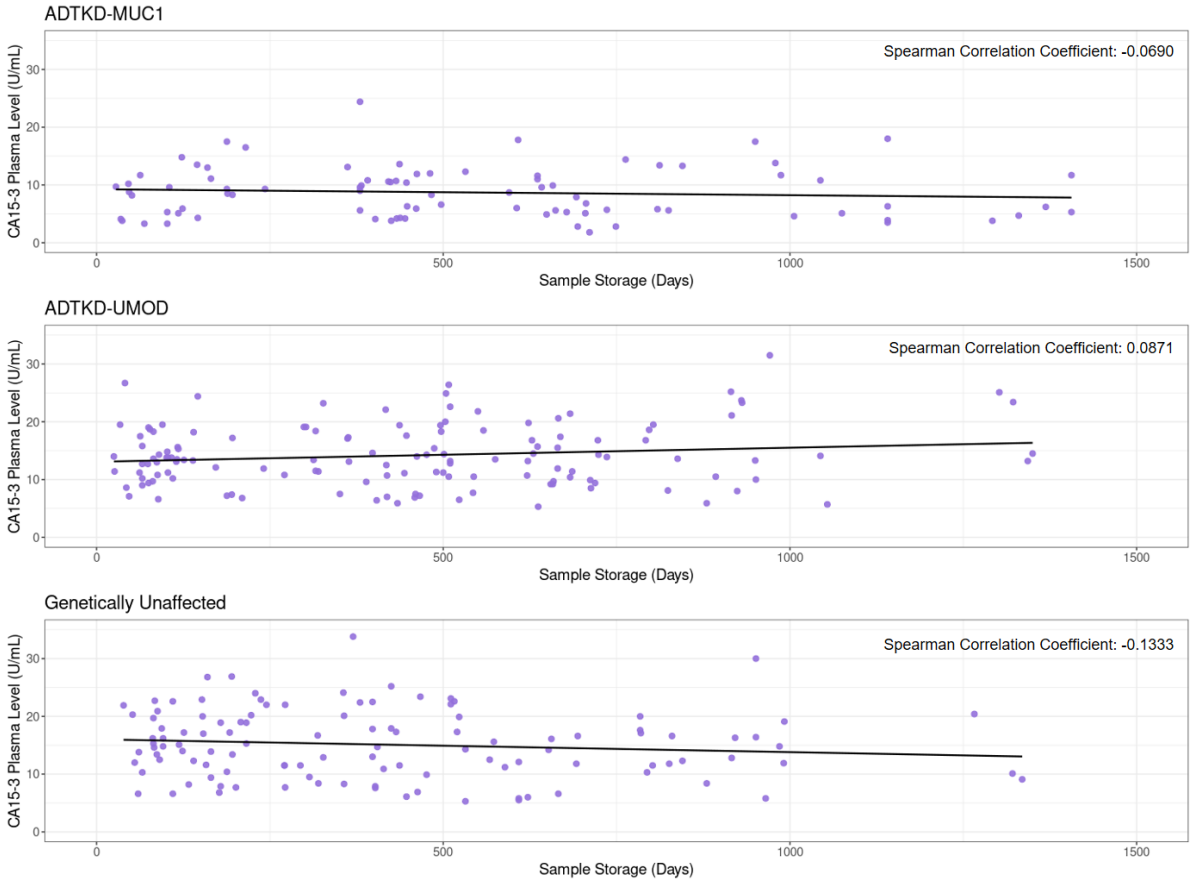
Supp. Fig. 3. eGFR by CA15-3 levels in the reference population.

Serum CA15-3 levels for 6,850 individuals are shown. While there was an association between CA15-3 and eGFR ($p<0.001$), this accounted for only 1.1% of the overall variation, as can be seen by the wide variation in serum CA15-3 levels at each level of eGFR.



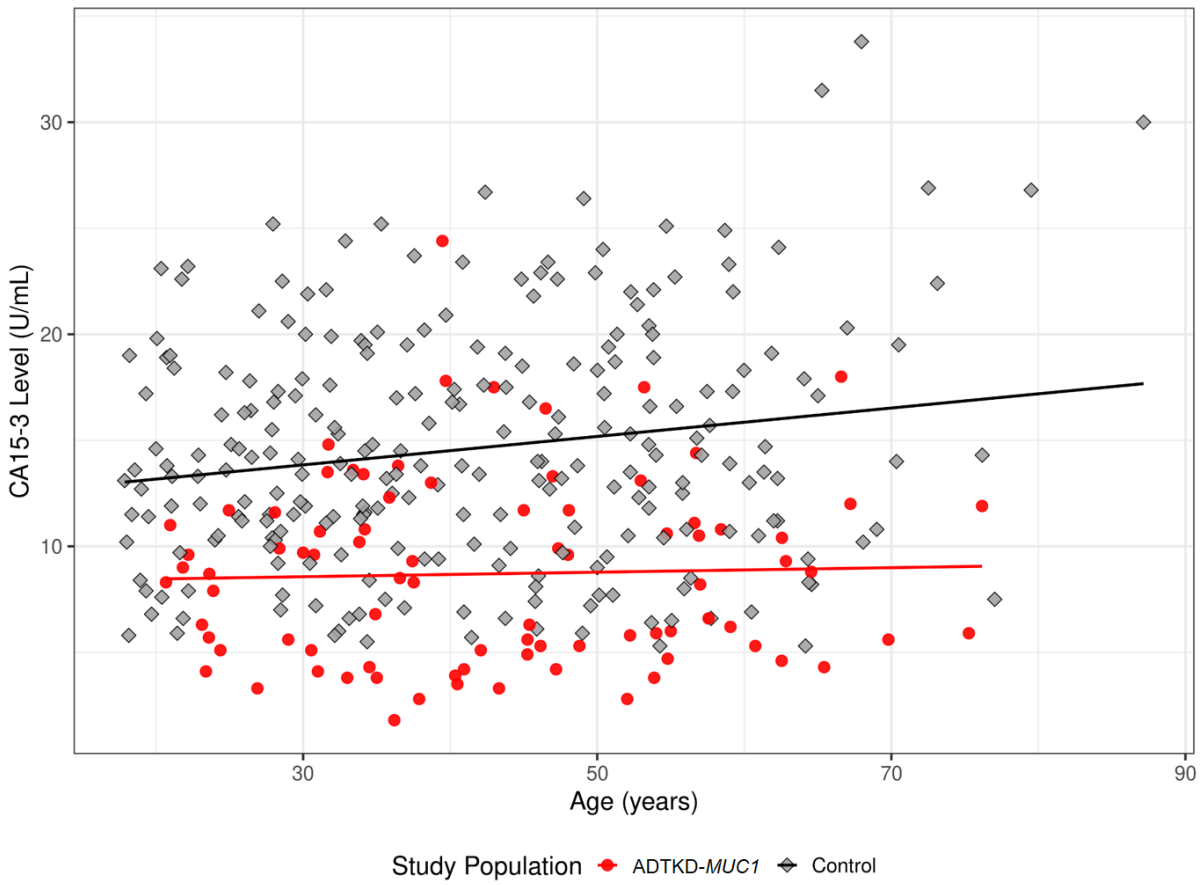
Supp. Fig. 4. CA15-3 levels according to length of storage time (days).

To test the effect of storage on CA15-3 levels, an analysis was performed to check for any correlation between plasma CA15-3 levels and the time interval from sample collection to analysis. The correlation for ADTKD-*MUC1*, ADTKD-*UMOD* and genetically unaffected study populations are shown below. The calculated Spearman correlation coefficients are very small, indicating no effect of storage on plasma CA15-3 levels.



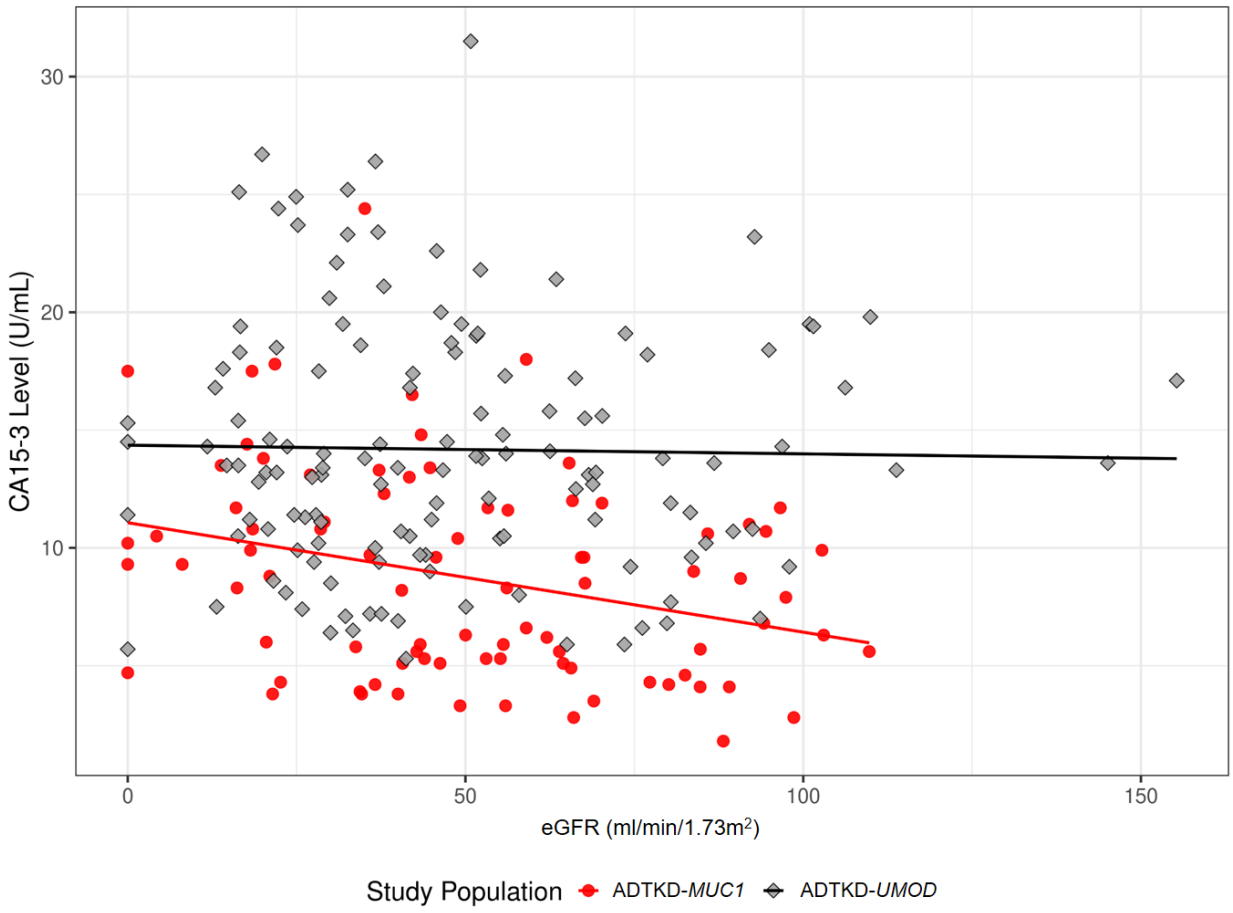
Supp. Fig. 5. CA15-3 levels according to age in patients with ADTKD-*MUC1* and controls (ADTKD-*UMOD* and genetically unaffected individuals).

Control individuals (n=245) are denoted by gray diamonds with a black trend line and patients with ADTKD-*MUC1* are denoted by red circles with a solid red trend line (n=85).



Supp. Fig. 6. CA15-3 levels according to estimated glomerular filtration rate (eGFR) in the ADTKD-*MUC1* and ADTKD-*UMOD* populations.

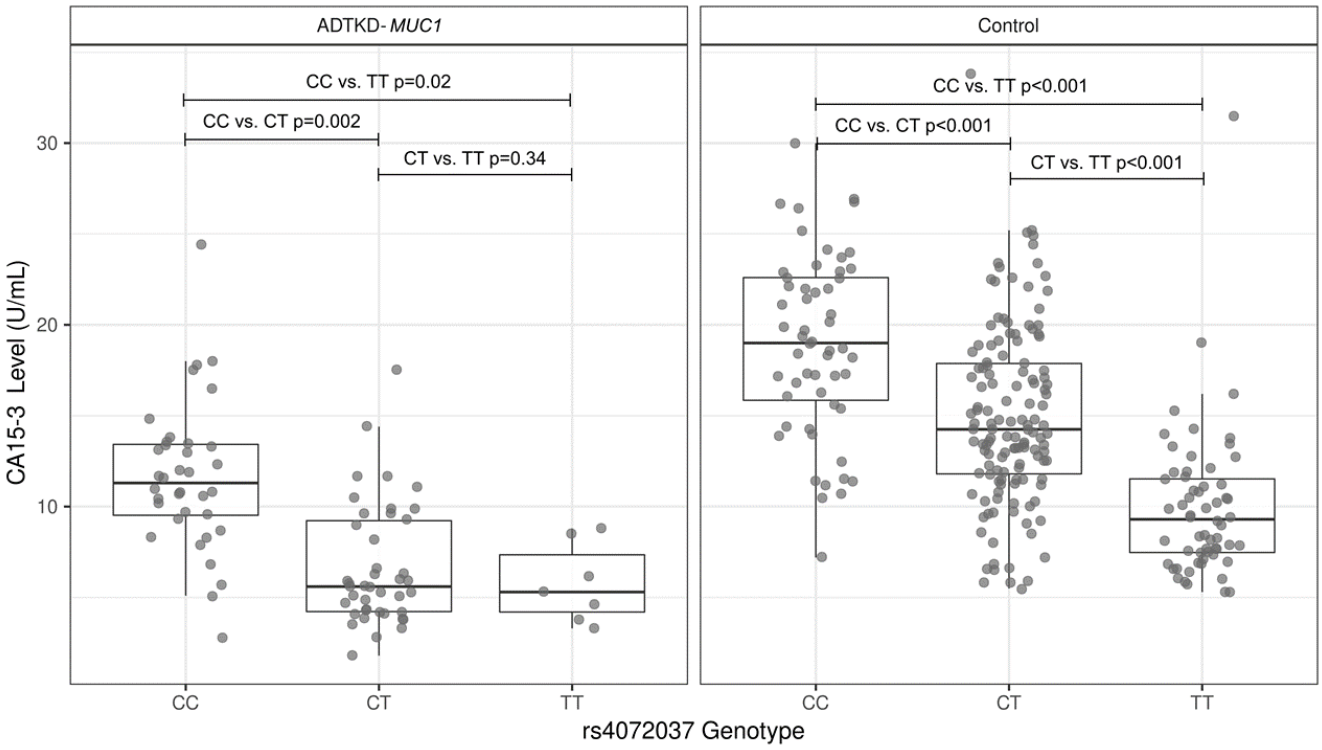
ADTKD-*UMOD* individuals (n=135) are denoted by gray diamonds with a black trend line and patients with ADTKD-*MUC1* are denoted by red circles with a solid red trend line (n=85).



Supp. Fig. 7. Histograms of Kolmogorov-Smirnov D_{max} .

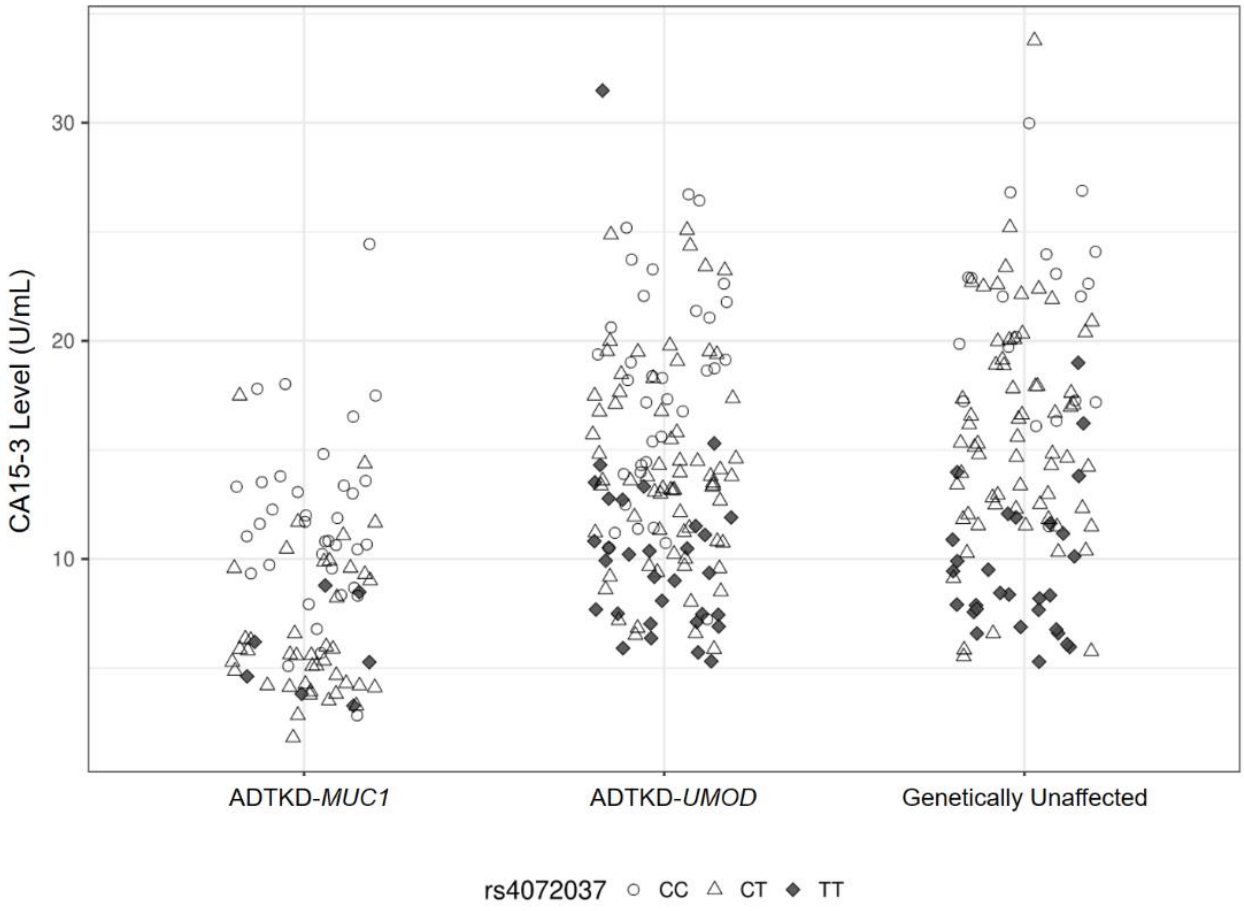
Equivalence of CA15-3 distributions between the reference population and the study populations (controls and ATDKD-*MUC1*) were compared using the Kolmogorov-Smirnov's D_{max}. To account for family structure in the study data, a resampling approach was applied. In each comparison (reference vs controls and reference vs ATDKD-*MUC1*), we completed 5000 random samplings with replacement on family status. The distribution of D_{max} for each set of comparisons is shown. The reference and control populations (gray histogram) showed comparable distributions (with small D_{max}) compared to reference and ATDKD-*MUC1* populations (blue histogram).

Supp. Fig. 8. Serum CA15-3 levels by rs4072037 and ADTK-MUC1 status. Distribution of plasma MUC1 levels according to rs4072037 genotype. Plasma levels are generally lower in ADTKD-*MUC1* and lower in individuals with the TT genotype. The distribution of plasma MUC1 levels for the C/T genotype in the controls is between that of the CC and TT genotypes. For ADTKD-*MUC1*, the distribution of the C/T genotype clusters near the T/T genotype, likely reflecting a higher frequency in individuals of C/T genotype of C alleles in phase with the mutant *MUC1*. P-values account for the family structure of the data; the illustrated means do not.



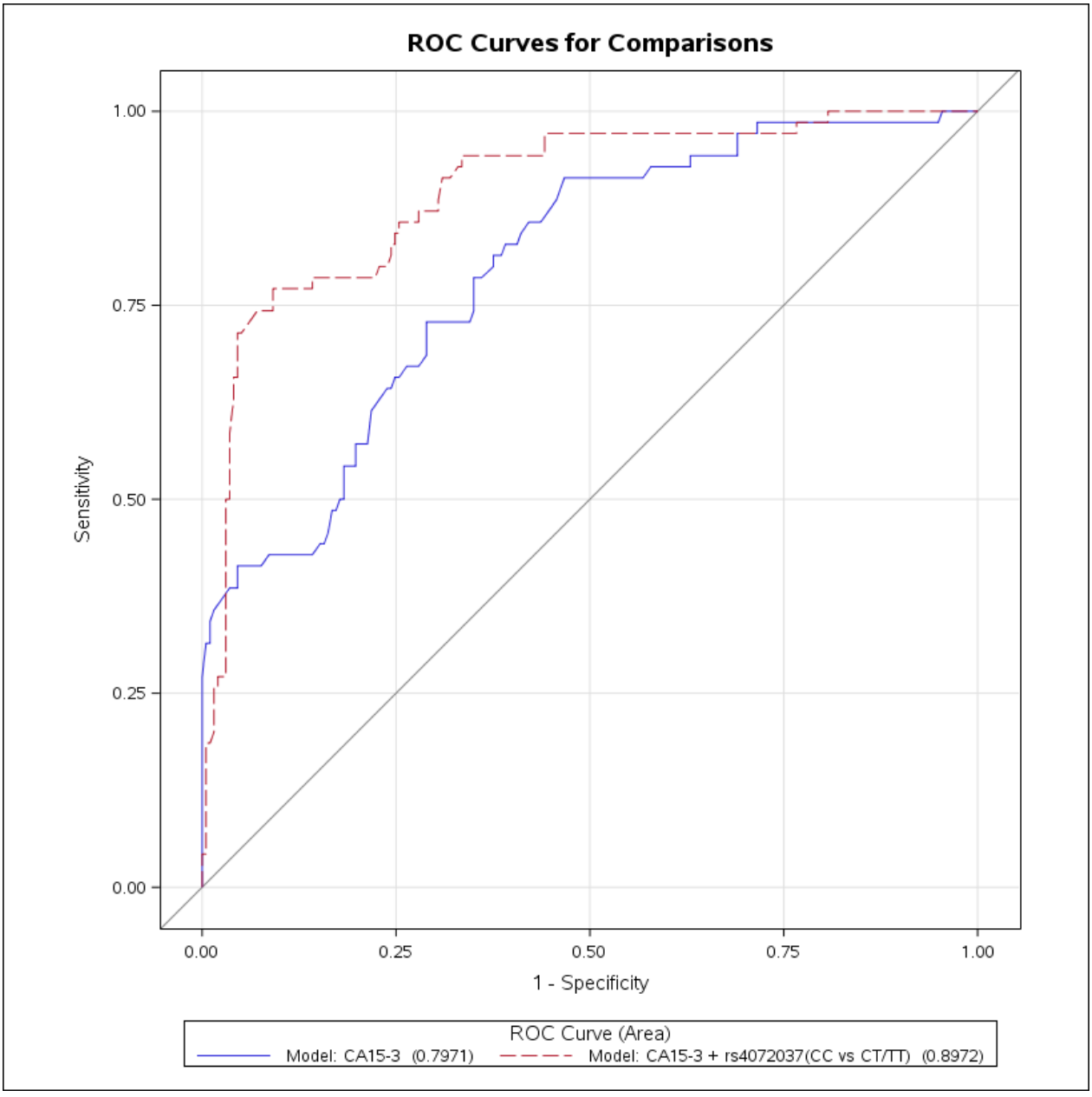
Supp. Fig. 9. Plasma CA15-3 levels by diagnosis and rs4072037 genotype.

Within the study population, CA15-3 levels varied by ADTKD-*MUC1* status (ADTKD-*MUC1* vs ADTKD-*UMOD* + genetically unaffected) (p<0.001) but not between ADTKD-*UMOD* and genetically unaffected (p=0.39). The CC genotype of rs4072037 was significantly associated with higher CA15-3 levels (versus CT) within each group (ADTKD-*MUC1* p=0.002, ADTKD-*UMOD* p=0.002, genetically unaffected p<0.001).



Supp. Fig. 10. Receiver operating characteristic (ROC) curves for predictive models.

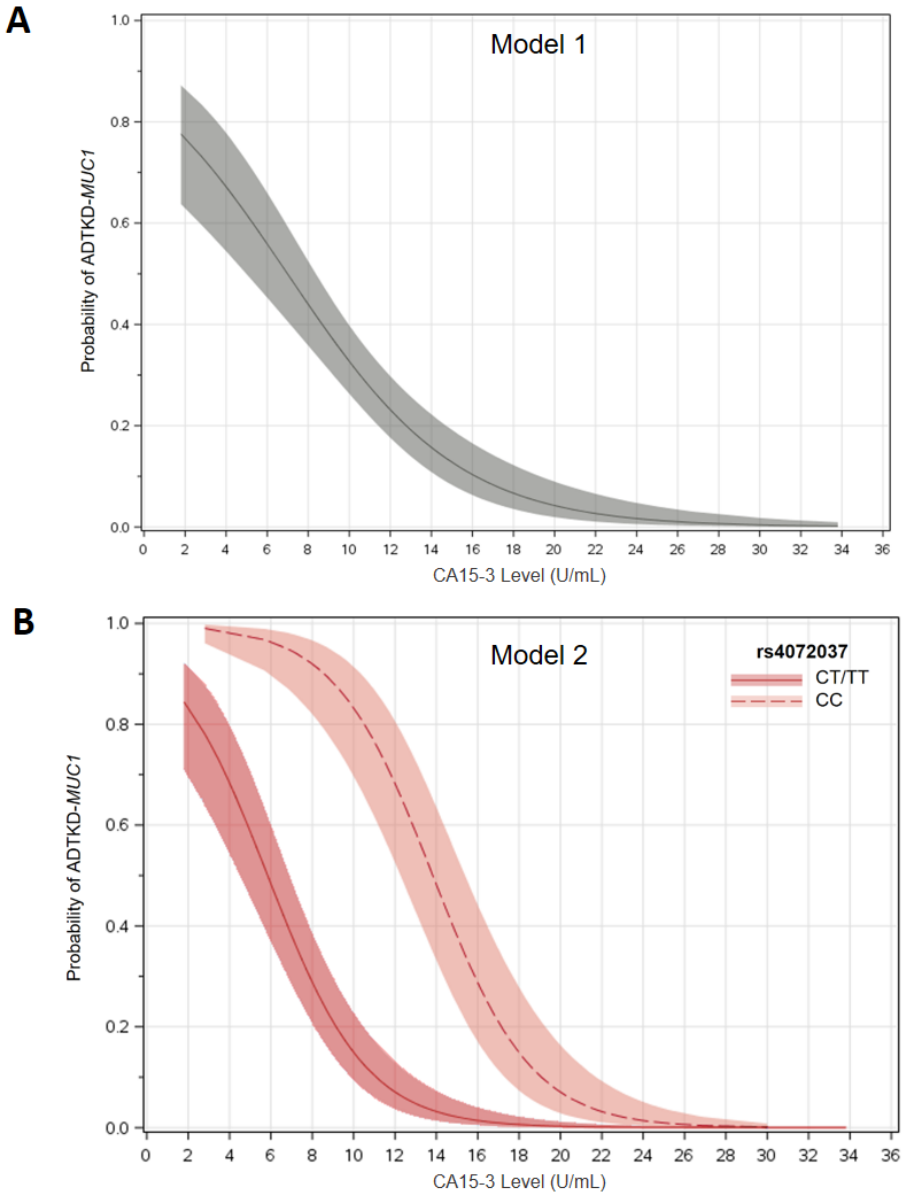
Two models for ADTKD-*MUC1* prediction were compared using the training dataset. Model 1 (shown as a red dotted line) used only CA15-3 serum levels as a predictor and Model 2 (shown as solid blue line) also included rs4072037 (dominant model for the T allele). Model 2 showed a greater area under the ROC Curve at 0.8972 compared to Model 1 (0.7971).



Supp. Fig. 11. Parametric estimation of the probability of ADTKD-*MUC1* genotype as a function of CA15-3 plasma level..

Predictions for ADTKD-*MUC1* were calculated under two models: 1) using only CA 15-3 as the predictor and 2) with both CA 15-3 and rs4072037 genotype (coded under a dominant model for the T-allele). **Supp. Fig. 11. a** shows the parametric estimation of probability for ADTKD-*MUC1* using Model 1 on the training dataset. **Supp. Fig. 11. b** shows the parametric estimation of probability for ADTKD-*MUC1* using Model 2, stratified by the rs4072037 genotype (CC and CT+TT), which shows clear separation. A probability of 0.80 corresponds to a CA15-3 level of 2.5 (U/mL) in CT/TT and 10.5 (U/mL) for the CC genotype.

Overall, we note that as a clinical diagnostic tool, the logistic regression models failed to provide classification rules with high certainty.



Supp. Fig. 12. Classifications of ADTKD-*MUC1* using a parametric estimation of probability in a validation set of 67 individuals.

Predictions of ADTKD-*MUC1* genotype were calculated under two models: 1) using only CA 15-3 as the predictor and 2) with both CA 15-3 and rs4072037 genotype (coded under a dominant model for the T-allele). For both models, ADTKD-*MUC1* status was assigned based on a probability greater than 0.5. **Supp. Fig. 12. a**. Classifications for Model 1 within the validation dataset. **Supp. Fig. 12. b**. Classifications for Model 2 within the validation dataset. Misclassifications are shown in red. While the overall error rates (proportion of misclassifications) were low for each model (0.134 in Model 1 and 0.164 in Model 2), we note that these classifications did not provide the same level of certainty (per classification –shown in **Supp. Fig. 11**) as did use of the CA15-3 thresholds of 5 and 20 (U/mL) (shown in **Fig. 2**).