Plasma glycosaminoglycans as diagnostic and prognostic biomarkers in surgically treated renal cell carcinoma

SUPPLEMENTAL INFORMATION

SUPPLEMENTAL METHODS

Study design. This study is reported in compliance with the STARD and REMARK guidelines.

The sample size for this study was powered using Bayesian estimation to achieve 99.5% probability that the primary endpoint was met, in that the 95% highest density interval (HDI) of GAG scores in RCC should lie outside the region of practical equivalence (ROPE) with the HDI of GAG scores in healthy controls. To this end, the scores in RCC were assumed to follow a HDI previously reported in [19] for metastatic ccRCC. Likewise, the scores in healthy controls were assumed to follow a HDI previously reported in [19] for healthy subjects. The ROPE was defined as the interval centered on 0 +/- 10% of the mean plasma score in metastatic ccRCC in [19], which was equal to 1.09 (i.e., if the mean difference between scores in RCC vs. healthy was 0 +- 0.1 then the distributions were defined as practically equivalent). The power analysis showed that a 99.5% probability was achieved for samples sizes exceeding $N > 40$ samples for RCC and $N > 10$ for healthy controls. To compare GAG scores in other types of renal masses as controls, a 15% rate of other renal masses was assumed in a consecutive series of surgically resected renal masses, of which 50% were assumed to be oncocytomas. Given that the minimum sample size for controls was estimated in $N = 10$ samples, in order to achieve this size for oncocytoma the sample size for RCC must exceed $N > 133$ samples according to the above assumptions. Therefore, a consecutive series of surgically treated renal masses of at least 156 patients was deemed powered to meet the primary endpoint using either healthy subjects or other renal masses as controls.

The study was double-blinded in that laboratory measurements of the GAG profile were performed at the University of Modena and Reggio Emilia, Modena, Italy independently and masked from clinical data relative to the samples collected at the Memorial Sloan Kettering Cancer Center, New York City, United States and vice versa. Laboratory and clinical data were uploaded by the data analysts to an electronic database designed for this study so that the respective contributing centers were unable to browse, edit, or delete information provided by the other center. Database administrator rights were granted exclusively within Chalmers University of Technology, Gothenburg, Sweden.

The study was carried out in accordance with the recommendations in the guidelines of the Institutional Review Board at the Memorial Sloan Kettering Cancer Center, New York City, United States and that participants provided written informed consent in accordance with the Declaration of Helsinki. The present study obtained ethical permission from the Institutional Review Board at the Memorial Sloan Kettering Cancer Center, New York City, United States on November 5th, 2012 (# 12-237).

Study participants. Data collection was retrospective. Inclusion criteria were: patients with radiographic finding of renal mass; healthy volunteers without any history of malignancy. Exclusion criteria were: no records on date of surgery; a pre-operative sample was obtained 50 days or earlier with respect to the date of surgery; absence of pre-operative samples following filtering out outliers or laboratory assay failures. Participants were enrolled at the Memorial Sloan Kettering Cancer Center, New York City, United States between May 25th 2011 and 18th February 2014. Eligible participants were identified based on radiographic findings and formed a consecutive series. For subjects with a renal mass, plasma samples were collected up to 50

days before primary surgery. A convenience sub-cohort of these subjects was followed longitudinally and samples were collected during a follow-up visit between 1 and 30 months after first surgery. For healthy volunteers, plasma samples were collected among relative of cancer patients. Sample draws were not performed at pre-defined times during the day.

Glycosaminoglycan measurements. Whole blood samples were collected in EDTA-coated tubes. The tubes were centrifuged (1100 g for 10 minutes) and the plasma was extracted and collected in a separate tube. All samples were stored at -80 °C. Shipment was performed in dry ice.

Laboratory measurements of the GAG profile quantified 19 properties: the total concentration of chondroitin sulfate (CS) in μg/mL, total concentration of heparan sulfate (HS) in μg/mL, total concentration of hyaluronic acid (HA) in μg/mL, the mass fraction of 8 sulfation patterns of CS (0s CS, 2s CS, 6s CS, 4s CS, 2s6s CS, 2s4s CS, 4s6s CS, Tris CS) and of 8 sulfation patterns of HS (0s HS, 2s HS, 6s HS, Ns HS, Ns6s HS, Ns2s HS, 2s6s HS, Tris HS). In addition, the charge of CS and HS was calculated as the weighted sum of all mass fractions, where the weight equals the number of charges present in a given disaccharide, and three additional CS ratios were calculated, 6s/0s CS, 4s/0s CS and 4s/6s CS. The overall GAG profile therefore consisted of 24 properties. These measurements were performed using capillary electrophoresis with laser induced fluorescence, as previously described [21-23] and as adopted in our previous study [19]. Samples with incomplete GAG profile measurements and sample with ambiguous labels were discarded. We refer to these events as laboratory assay failure. Technical replicates were performed for 80 samples as described in the section "Reproducibility analysis".

An outlier was defined *post-hoc* by first calculating the Mahalanobis distance of the coordinates of each sample along the top two principal components in a principal component analysis of all available GAG profiles; and next by labeling a sample as outlier if its Mahalanobis distance was greater or equal to 9.210, which is 99%-quantile of the Chi-Square distribution with 2 degrees of freedom (e.g. equal to the number of principal components used to calculate the Mahalanobis distance). Only pre-operative and healthy samples were included in the outlier analysis. Subjects with outlier samples were excluded from the study.

Exploratory data analysis. Principal component analysis (PCA) was performed across all pre-operative RCC and healthy control samples by using either CS or HS only properties. GAG measurements were centered by the mean and scaled by the standard deviation prior (PCA). PCA was performed using the native R function prcomp. Unsupervised hierarchical clustering was performed on the same datasets using Pearson correlation to compute clustering distance and the Ward's minimum variance as agglomeration method in order to find compact, spherical clusters [24]. This was implemented using the R function pheatmap from the homonymous library. The enrichment of selected histopathologic features in the clusters emerging from hierarchical clustering was tested using the proportional equality tested, implemented using the native R function prop.test.

Glycosaminoglycans scores derivation and accuracy evaluation. Two different scoring systems using selected properties in the plasma GAG profile were used: the previously developed GAG score and the new GAG score derived in this study. For diagnostic use, the index test was defined as the test that classifies a subject as having RCC if the formula to

calculate the previous or new GAG score in a sample from that subject returns a score greater than a pre-specified cut-off for plasma GAG scores. In the case of the previous GAG score, the following formula was applied to calculate the score [19]:

$$\text{Plasma score} = \frac{[6s\ CS] + CS_{tot}}{\dfrac{3}{10}\dfrac{[4s\ CS]}{[6s\ CS]} + [Ns\ HS]}$$

where $[6s\ CS]$ is the mass fraction of 6-sulfated chondroitin sulfate, $[4s\ CS]$ is the mass fraction of 4-sulfated chondroitin sulfate, $[Ns\ HS]$ is the mass fraction of N-sulfated heparan sulfate, and CS_{tot} is the total concentration of CS in µg/mL. The pre-specified cut-off score above which the sample was deemed to have RCC was 0.234 [19]. A pre-specified variation of this score omitting the $[Ns\ HS]$ term was also tested although no cut-off was pre-specified. The index test was conducted on all 175 pre-operative RCC samples vs. 19 healthy controls from this cohort or vs. 25 healthy controls from historical cohorts.

In the case of the new GAG score, the formula was derived using penalized regression using Lasso[25] to determine which plasma GAG properties were robustly associated with RCC vs. healthy controls, as previously described[19]. To this end, 38% of the total pre-operative RCC samples from this study's cohort were randomly selected to form a discovery set. This subset was obtained from the interim analysis of this study. Feature selection was performed by reducing the total set of 19 GAG properties to those properties either associated to RCC in the current cohort during exploratory data analysis or in the previous study: the 6s CS fraction, the 6s/4s CS ratio (here normalized by total 4s and 6s CS), the total CS concentration, the CS charge and the 0s/Ns HS ratio (in $\log_2$ scale to increase robustness). Feature selection yielded superior goodness-of-fit than penalized regression using all available GAG properties as independent variables, likely due to internal correlations in the GAG profiles. Using these 5 properties as independent variables, a penalized regression was applied separately to the following cohorts: the discovery set plus all healthy controls from this study's cohorts; and historical cohorts with metastatic ccRCC and healthy samples obtained in either Sweden or Italy from our previous study[19]. Each regression was controlled for over-fitting by performing 20-fold cross-validation. The median coefficient for each independent variable in the three regressions, as determined by the penalty value λ at which that error is within 1 standard error of the minimum, was rounded to build the formula for the new GAG score. An optimal cut-off for the new GAG score was then computed so to maximize classification accuracy in the discovery set vs. healthy controls from this study's cohort. For validation purposes, the index test was defined as above and the tested population was the remaining 62% of the total pre-operative RCC samples. Note that only the index test sensitivity could be validated in this population.

Statistical association analysis. The difference in each of the GAG properties included during feature selection between RCC and healthy groups, adjusted by cohort, was assessed using Bayesian estimation on a bivariate linear model [26,27]. The highest density interval (HDI) for the mean difference in a given GAG property between RCC versus healthy was calculated under the following assumptions: scores are sampled from a t-distribution of unknown and to be estimated normality (i.e., degrees of freedom); high uncertainty on the prior distributions; the marginal distribution is well approximated by a Markov chain Monte Carlo (MCMC) sampling

with no thinning and chain length equal to 100,000. MCMC sampling was considered satisfactory if at least 10,000 MCMC samples were effectively obtained. The estimation was performed using JAGS [26,27].

The association between the new GAG score or any of its 5 constituent GAG properties and selected histopathologic variables was assessed with the Mann-Whitney test for categorical variables (tumor grade, stage, and histological subtype) and a t-test on the linear regression coefficient for continuous variables (tumor size). In the former case, the reference category was grade 2 or low for tumor grade, stage I or II for tumor stage, and clear-cell histology for tumor histological subtype, and the shift to such reference is the reported statistics. In the latter case, the reported statistics is not the regression coefficient, but the Pearson correlation coefficient for ease of interpretation. p-values < 0.05 were considered significant.

Survival analysis. Survival was calculated as the time between the date of first surgery and the time of event. The time of event is defined as right-censoring (date of last follow-up without the event) or as date of death in case of overall survival (OS) and date of recurrence in case of recurrence-free survival (RFS). Recurrence was defined as radiological evidence of one or more metastatic lesions. The population was defined as all surgically treated RCC cases in the case of overall survival and as the subset of this population with no evidence of metastasis before surgery in the case of recurrence-free survival. The last date of follow-up without the event was assessed between June and November 2017. Univariate and multivariate survival analyses were performed by fitting a Cox proportional hazard model to estimate the hazard ratio for the variables of interest and the 95% confidence interval. The log-rank statistical test was utilized to determine the significance of the regression. Initial candidate variables considered for regression of survival using a univariate Cox model as above: age (continuous, in years), tumor grade (categorical, Low or II vs. High or III or IV, 35 missing data), TNM stage (categorical, I or II vs. III or IV), tumor size (continuous, in cm), SSIGN score (integer, 57 missing data, OS only), Leibovich score (integer, 54 missing data, RFS only), the new GAG score and its 5 constituent GAG properties (continuous, arbitrary unit). Missing data were omitted. A multivariate Cox model was pre-specified by first using variables reaching statistical significance in the univariate analysis and then by subjecting this model to penalized regression using the Lasso with 50-fold cross-validation. No other pre-specified multivariate Cox model with validated prognostic factors were considered given that the univariate analysis on two established nomograms (SSIGN and Leibovich score for OS and RFS, respectively) can be used as reference. The validity of the proportional hazard assumption was checked using a two-sided t-test between transformed survival time and the scaled Schoenfeld residuals. The sample size was not powered specifically for this study, because no prior knowledge on the prognostic value of the plasma GAG score was available for surgically treated RCC at the time of design of this study. We checked for severe overfitting by performing internal validation of the Lasso-penalized multivariate models for OS and RFS using a bootstrapping algorithm (1,000 bootstraps) and observing the change in Somers' D rank correlation (D_{xy}) statistics in the original datasets as opposed to the test set. The so-corrected D_{xy} is transformed into the concordance index ($C_{index} = \frac{D_{xy}}{2} + 0.5$) and reported together with the original C-index to both convey a metric for the predictive discrimination of each model and its optimism (i.e. if the

corrected C-index is substantially lower than the original C-index, then the model is fitting survival with excessive optimism and vice versa).

One of the constituent GAG properties of the new GAG score, namely CS total, was also used to dichotomize patients into two groups, "Low" versus "High" score, where an optimal value for CS total was used as cut-off. The optimal value was searched using maximally selected rank statistics. Kaplan-Meier survival curves were fitted for the two groups, and the statistical significance for survival difference was evaluated using the log-rank test. Two- or three-year survival rates were calculated as the survival probability at the start of the time interval that includes the Kaplan-Meier fit for 24 or 36 months for RFS or OS respectively. In addition, we repeated this analysis for the following cases: first, we summed the tumor size normalized to 5 cm to the CS total value and searched an optimal cut-off for this metric; second, by assigning patients into "Low" vs. "High" groups based on whether the tumor size was below or above 5 cm. A multivariate Cox model using CS total and tumor size for both OS and RFS was finally built to assess concordance when using variables only known pre-operatively.

Survival analyses were performed using the packages *survival, glmnet, maxstat* and *rms. p*-values < 0.05 were considered significant.

Reproducibility analysis. Technical replicates were performed on 80 samples from a quasi-random subset of 40 patients so to encompass a pair of pre-op and post-op samples for each patient. The quasi-randomization aimed at balancing RCC histologies and presence of metastases at surgery in the subset. The coefficient of variability and least-square linear regression was computed for the new GAG score in the original batch (1st batch) vs. the replicated batch (2nd batch) for all 80 samples. The change of new GAG score post-op vs. pre-op was computed as the $\log_2$ ratio. The direction of change was dichotomized into "increase" vs. "decrease" depending if the $\log_2$ ratio post-op vs. pre-op was positive or negative. The concordance in direction of change for each patient according to 1st vs. 2nd batch was tested with the Fisher's Exact Test in which an odds-ratio greater than 1 is suggestive of concordance. *p*-values < 0.05 were considered significant
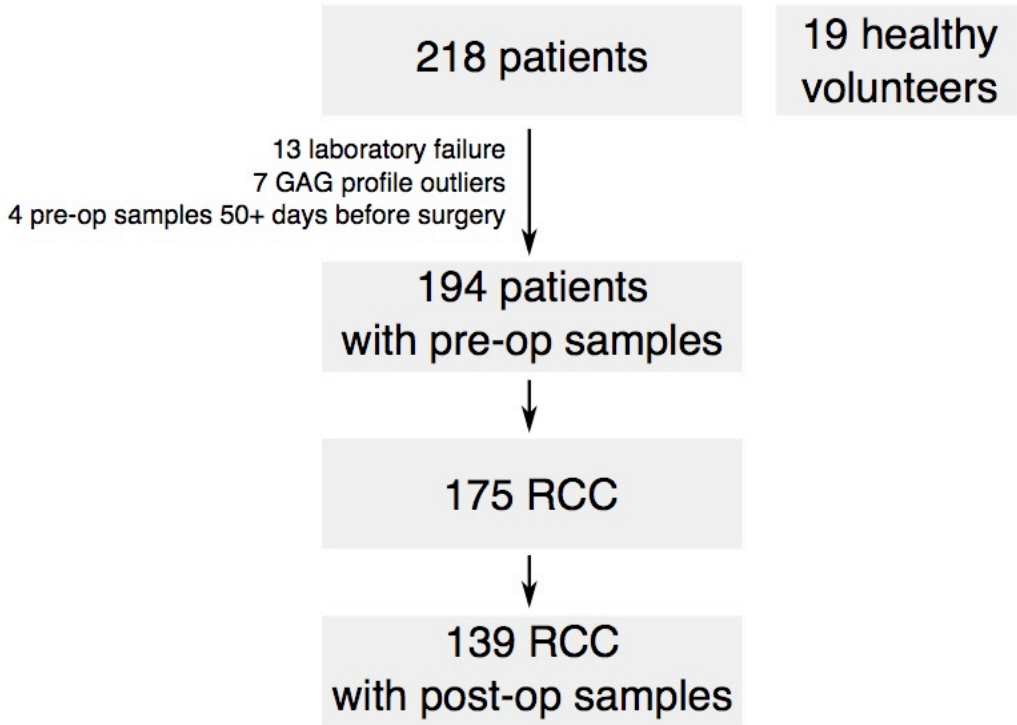
SUPPLEMENTAL FIGURES

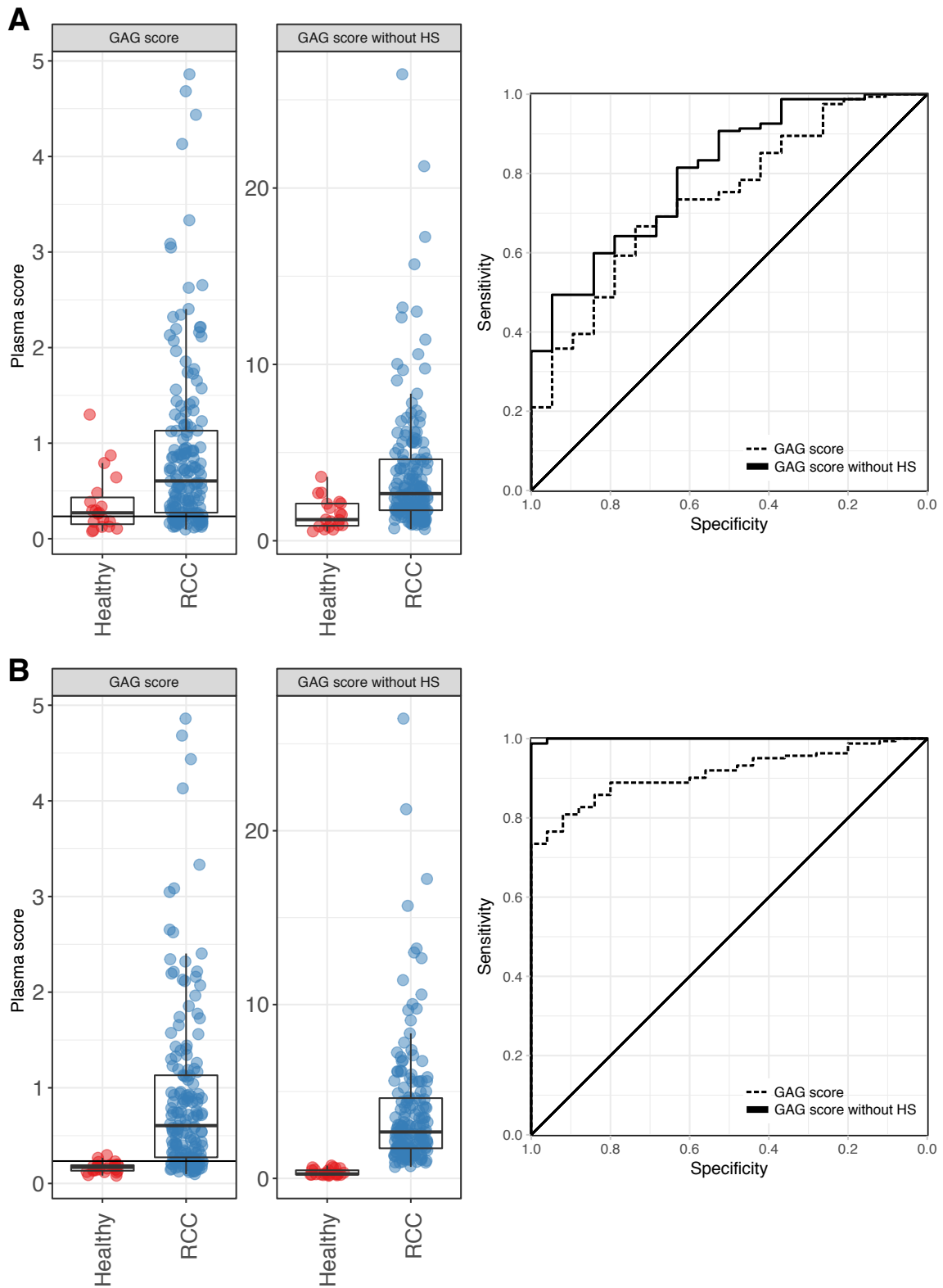**Figure S1.** Flow diagram of participants in the study.

Figure S2. Boxplots of plasma GAG scores as calculated according to the published formula in healthy groups vs. 175 RCC pre-operative samples - with or without the inclusion of Ns HS in the formula - and their corresponding receiving operating characteristic (ROC) curve. A) Healthy samples (*N*=19) from this cohort. B) Healthy samples (*N*=25) from the historical cohorts in [18].
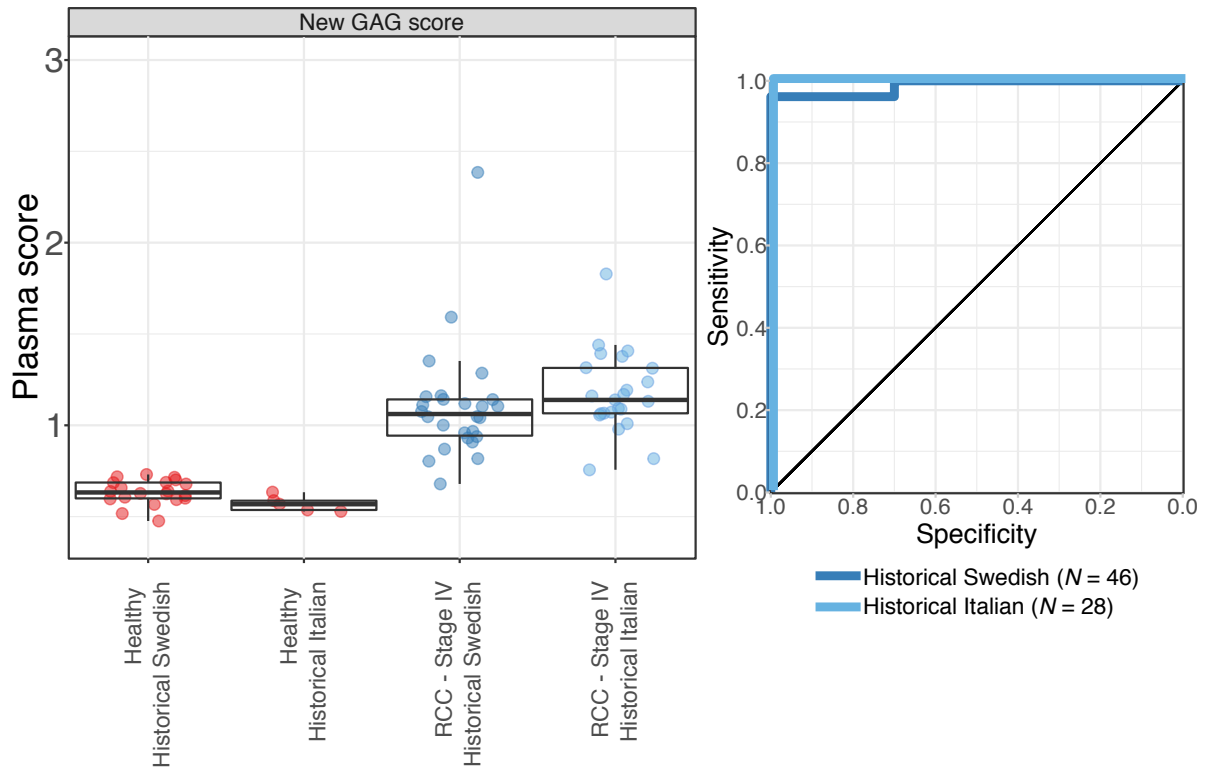
Figure S3 – Boxplot for the new plasma score in 26 Stage IV ccRCC versus 20 healthy (historical Swedish cohort) and 23 Stage IV ccRCC versus 5 healthy (historical Italian cohort), previously published in [18]. The corresponding ROC curve of the classification in each cohort is shown of the right.
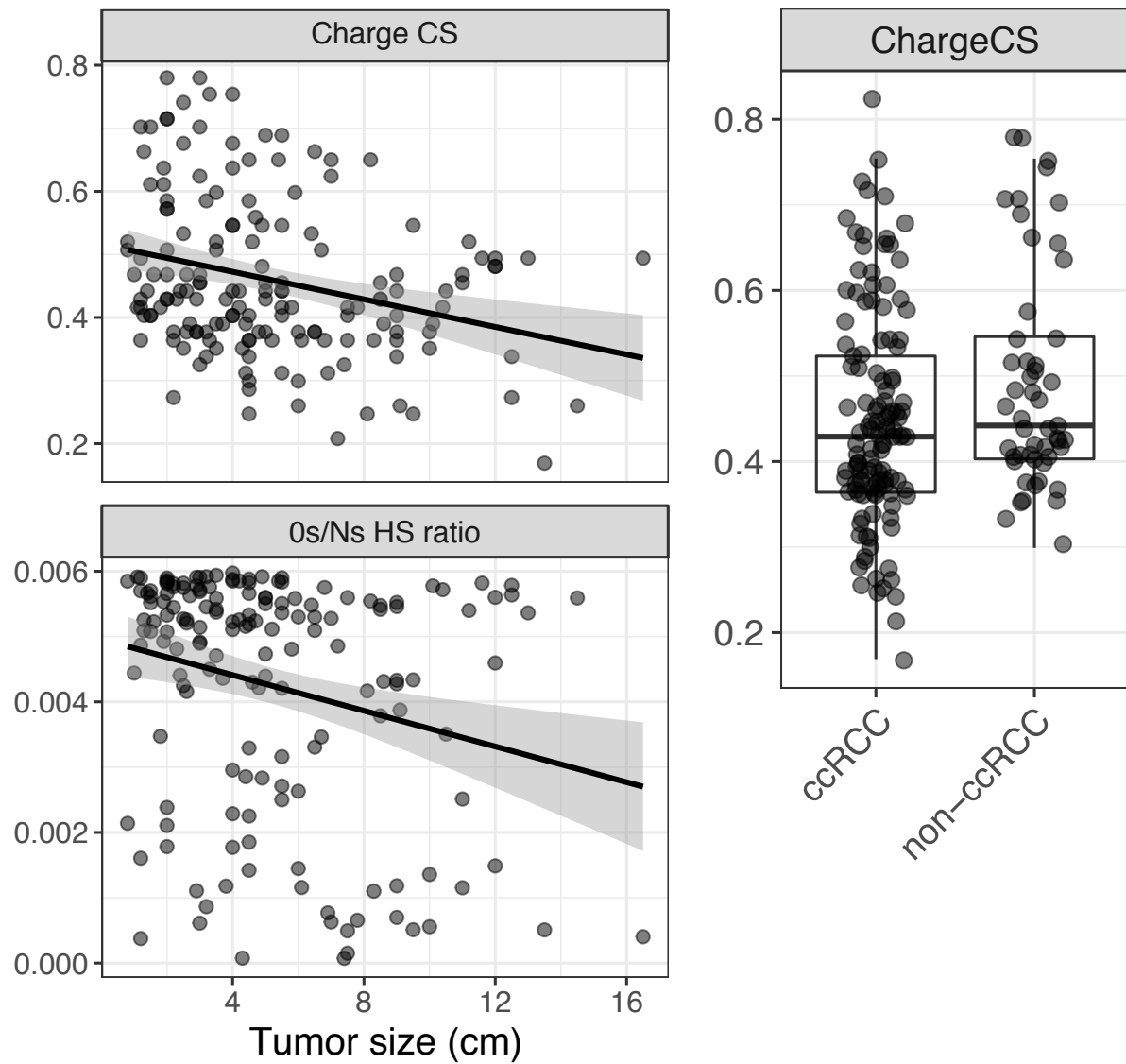
Figure S4 – Significant correlations between GAG properties in the new GAG score and clinicopathologic features of pre-operative RCC samples (*N*=175). On the left, the correlation between tumor size (in cm) and the CS charge or the 0s/Ns HS ratio. The least-square regression line is shown with the 95% confidence interval (grey shade). On the right, the Charge CS grouped in boxplots by ccRCC vs. nccRCC histology.
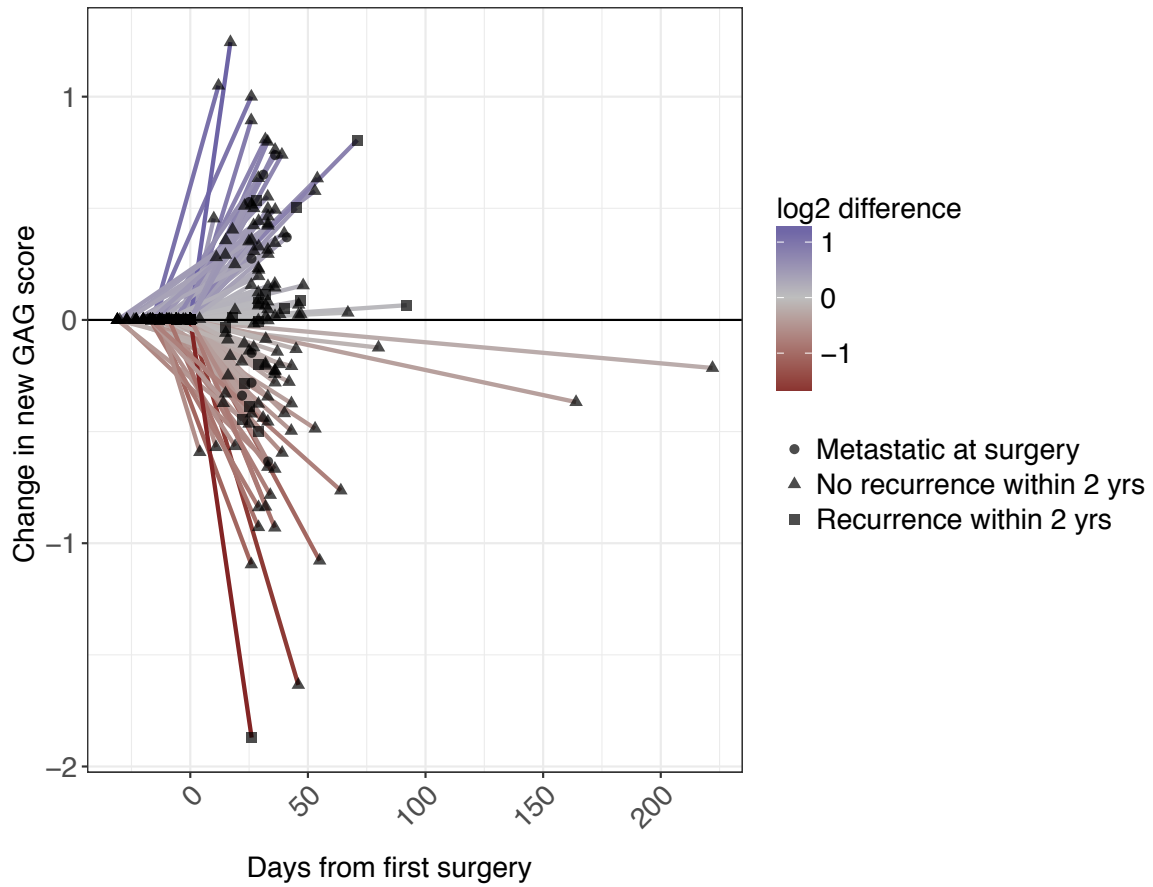
Figure S5 – The change in new GAG score after surgery was calculated as the log$_2$ difference in the new GAG score between the first available sample post-operatively and the most recent available sample pre-operatively for 139 RCC patients. Each patient is represented by a line, whose colored is scaled to the log$_2$ difference, and the shape of the extremes of each line represents the outcome of the corresponding patient within 2 years from first surgery.
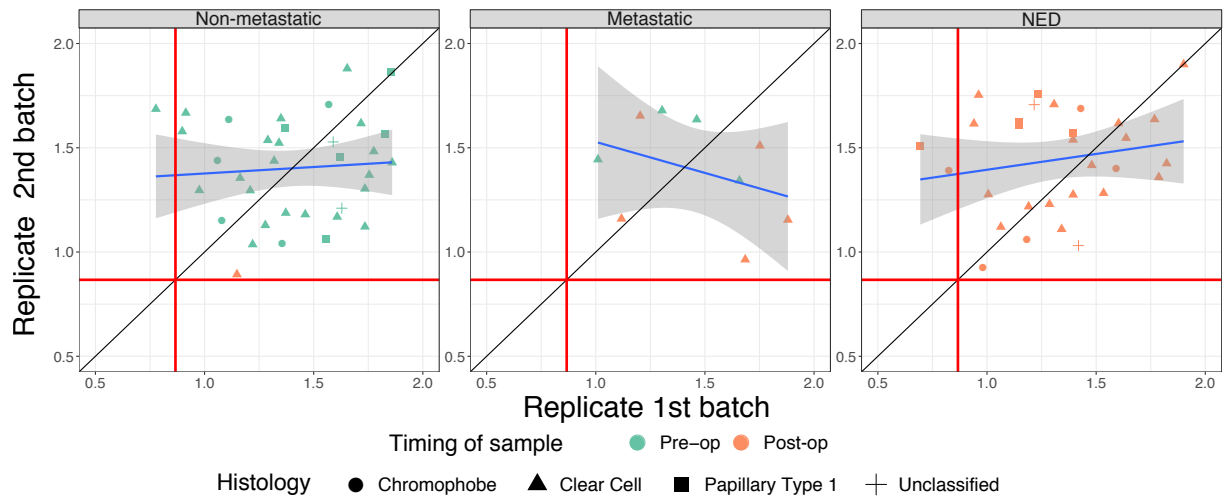
Figure S6 – Estimation of the new GAG score in 80 samples assessed in duplicates and their correlation in the 1st vs. 2nd batch. Samples are grouped by stage. The black diagonal line represents the line of identity, while the blue line represents the least-squares linear regression of the new GAG score in the 1st vs. 2nd batch. The red lines define the threshold below which a sample is classified as "healthy". Each sample is represented by a point whose color indicates the timing of sampling (pre-operative vs. post-operative) and shape indicates the RCC histological subtype.

SUPPLEMENTAL TABLES

Table S1. Cross-tabulation of the index test with the reference standard for different definitions of the index test.

Previous GAG score (pre-specified cut-off)	RCC	Healthy
> 0.234	140	11
<= 0.234	35	8
New GAG score (Discovery set, exploratory cut-off)	**RCC**	**Healthy**
> 0.87	67	1
<= 0.87	0	18
New GAG score (Validation set, pre-specified cut-off)	**RCC**	**Healthy**
> 0.87	101	-
<= 0.87	7	-

Table S2. Clinicopathological features in the pre-operative RCC patients as split between the discovery and validation set. Distributions are summarized as median and interquartile ranges in brackets.

Factors	Discovery	Validation	
	N = 67	*N* = 108	
Age [years]	55 [48-63]	62 [55-68]	*p*<0.01
Gender			
Female	17	31	*p*=0.76
Male	50	77	
Ethnicity			
White American	59	100	*p*=0.46
African American	2	5	
Asian American	1	1	
Other/Not Available	5	1	
Histological subtype			
Clear cell	46	78	*p*=0.74
Non-clear cell	21	30	
Chromophobe	8	9	
Mucinous tubular and spindle cell	1	1	
Papillary Type I	8	11	
Papillary Type II	1	2	
Papillary (unspecified)	0	4	
Unclassified	3	3	
Tumor size [cm]	4.3 [2.7 – 6.7]	4.6 [2.9 – 7.3]	*p*= 0.65
AJCC stage			
Stage I	37	57	*p*=0.99
Stage II	3	3	
Stage III	20	32	
Stage IV	7	15	
Not Available	1	1	
Grade			
Fuhrman nuclear grade	39	65	
2	14	25	*p*=1
3	26	35	
4	5	17	
Other grading system	17	17	
High	3	4	
Low	6	5	
Not Available	13	22	