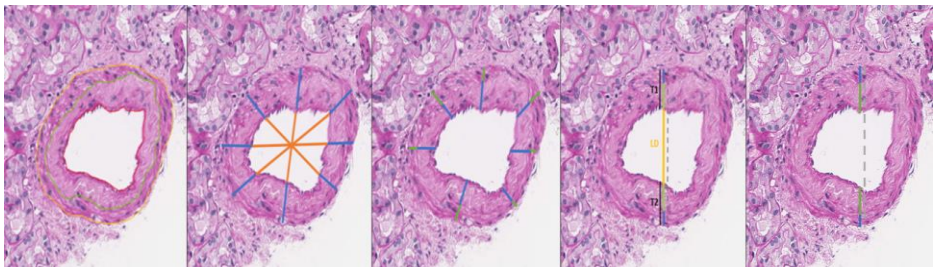

Finding the right fit:

A comparison of arteriosclerosis scoring systems in non-neoplastic kidney biopsies

Lars-Eirik Høgsaas, Class of 18A

Main supervisor: Hrafn Weishaupt, Ph.D.

Secondary supervisor: Sabine Leh, M.D., Ph.D.

Bergen, December 18th, 2023

Table of contents

1. Introduction 6
2 Materials and Methods 8
2.1 Literature search 8
2.2 Scoring systems 11
2.2.1 Banff 11
2.2.2 Remuzzi 12
2.2.3 Sethi 13
2.2.4 Maryland aggregate pathology index (MAPI) 13
2.2.5 Arteriosclerosis Index (ASI) 14
2.3 Acquisition and selection of raw images 15
2.4 Scoring procedure 15
2.5 Evaluation of scoring results 17
3 Results 19
3.1 Intra- and inter-rater variability (IARV and IERV) 19
3.2 Correlations across different scoring systems 21
3.2.1 Arteriosclerosis Index (ASI) compared to other systems 22
3.2.2 MAPI and the three semiquantitative scoring systems compared to each other 25
3.3 Missing values 28
4 Discussion 30
5 Conclusion 38
6 Appendix 41
7 References 43

Abbreviations

SP = Scoring person

ASI = Arteriosclerosis Index

MAPI = The Maryland Aggregate Pathology Index

VCI = Vascular Chronic Index

NEPTUNE = Nephrotic Syndrome Study Network

KA = Krippendorff's alpha

IERV = Inter-rater variability

IARV = Intra-rater variability

SCC = Spearman's Correlation coefficient

ICC = Intra Class Correlation coefficient

Abstract

Introduction: Arteriosclerosis is a pathological condition in the arteries which ultimately leads to a narrowing of the arterial lumen and a decreased and disrupted blood flow. Several different scoring systems are used by pathologists to score arteriosclerosis, but little is known about how they compare. This study aims to improve on this knowledge through the systematic identification of and comparison between the most relevant methods.

Materials and Methods: The first part of the study consisted of a literature search to identify scoring systems. These systems were then used by four scoring persons of different professional backgrounds to score a set of 60 renal arteries. The results were then compared using a number of parameters such as inter- and intra-rater variability. The fourth scoring person used the quantitative scoring systems. Results from scoring were then compared relative to scoring persons, scoring systems, time, and other variables.

Results: We found 11 different scoring systems, and several indications of differences between systems, especially in terms of their ability to score, and the associations between scores of different scoring systems.

Conclusion: Our results indicate that some scoring systems are favorable to others, but given the limitations of the study, most notably a lack of time and resources such as the number of scoring persons as well as a lack of a ground truth or other parameters to compare the results to, there is uncertainty around these findings. Future research is needed to further increase our understanding of the differences between these scoring systems.

Sammendrag

Innledning: Arteriosklerose er en patologisk tilstand i arteriene som til slutt fører til en innsnevring av arteriell lumen og nedsatt og forstyrret blodstrøm. Flere forskjellige skåringssystemer brukes av patologer for å skåre arteriosklerose, men lite er kjent om hvor mye de ligner hverandre. Denne studien har som mål å forbedre denne kunnskapen gjennom systematisk identifisering og sammenligning av de mest relevante skåringsmetodene.

Materialer og metoder: Første del av studien besto av et litteratursøk for å identifisere poengsystemer. Disse systemene ble deretter brukt av fire skåringspersoner med ulik faglig bakgrunn for å score et sett med 60 nyrearterier. Resultatene ble deretter sammenlignet ved hjelp av en rekke parametere som inter- og intra-rater-variabilitet. Den fjerde skåringspersonen brukte de kvantitative skåringssystemene. Resultatene fra skåringen ble deretter sammenlignet i forhold til skåringspersoner, skåringssystemer, tid og andre variabler.

Resultater: Vi fant 11 forskjellige skåringssystemer, og flere indikasjoner på forskjeller mellom systemene, blant annet fant vi variasjon i skåringssystemenes terskel for å vurdere en arterie som sklerotisk, og at noen systemer ser ut til å kreve mindre trening enn andre.

Konklusjon: Resultatene våre indikerer at noen skåringssystemer er overlegne andre, men gitt studiens begrensninger, særlig mangel på tid og ressurser slik som antall skåringspersoner, samt mangel på en «ground truth» eller andre parametere å sammenligne resultatene med, er det usikkerhet rundt disse funnene. Flere studier, fortrinnsvis med en «ground truth», er nødvendige for å ytterligere forbedre vår forståelse av forskjellene mellom disse poengsystemene.

1. Introduction

Arteriosclerosis is a pathological condition of the arterial wall. The name is derived from the Greek words “arteria” (meaning artery), “sclerosis” (meaning hardening), and “osis” (meaning diseased condition). It is characterized by abnormal arterial wall thickening, hardening, and loss of elasticity (2). Everyone develops some degree of arteriosclerosis over the span of a lifetime. So-called “fatty streaks” – plaques attaching to the inside of arterial walls – start appearing even in healthy individuals in childhood (3). There are however a number of factors that accelerate this process, including hypertension, hyperlipidemia, diabetes, and smoking. The pathogenesis of arteriosclerosis involves a complex interplay of endothelial dysfunction, lipid accumulation, inflammatory processes, and vascular smooth muscle cell proliferation (2, 3).

The artery wall consists of three main layers starting from the inside: Intima, media, and adventitia. Each layer has its own specialized functions. For instance, most of the muscle cells that are essential for controlling blood flow by contraction and relaxation are found in the medial layer (4). Arteriosclerosis is the thickening and hardening of one or more of these layers, and there are three main types, which are: Atherosclerosis, Mönckeberg medial calcific sclerosis, and arteriolosclerosis (2).

The most common and best-known type, atherosclerosis, involves the formation of plaques consisting of lipids, cholesterol, calcium and blood cells within the intimal layer of arteries. Over time, these atherosclerotic plaques can gradually accumulate and narrow the arterial lumen, reducing the space through which blood can flow. Plaque rupture can form a thrombosis which then gets carried away with the blood stream, potentially causing an acute vascular event like ischemic stroke, myocardial infarction, or even sudden cardiac death (2, 3, 5).

Another type is known as Mönckeberg medial calcific sclerosis. It is characterized by the calcification of the muscular layers in small and medium-sized arteries, and especially those in the medial layer, typically in extremities, without significant luminal narrowing or plaque formation. It is typically seen in patients with chronic kidney disease and diabetes, and often gives no symptoms (2, 5, 6).

A third type is known as arteriolosclerosis: This form affects smaller arteries and arterioles, and there are two subtypes: hyaline and hyperplastic arteriolosclerosis. Hyaline arteriolosclerosis is associated with mild to moderate hypertension and diabetes, where proteins like albumin leak into the vessel wall, causing thickening. Hyperplastic arteriolosclerosis, often seen in malignant hypertension, is characterized by concentric thickening of arterioles due to smooth muscle cell proliferation and collagen deposition (2, 5).

The arteries in the kidney, however, show a different type of arteriosclerosis affecting the arcuata and interlobular arteries. This manifestation differs from the common occurrence of atherosclerosis or calcifications in the media. Instead, the typical finding is widening and fibrosis of the intima, usually associated with narrowing of the lumen. The lamina elastica might become multilayered. Additionally, the media is either hypertrophic or withering with increased amount of fibrosis and a reduction of smooth muscle cells (7, 8).

The kidneys are well supplied with arteries, and arteriosclerosis here has been associated with tubular atrophy and glomerulosclerosis (9). Therefore, determining degrees of arteriosclerosis has both prognostic and predictive significance. This makes it important for nephropathologists to accurately assess degrees of arteriosclerosis.

Arteriosclerosis is usually graded semiquantitatively in non-neoplastic kidney biopsies, and different scoring systems are currently in use for such grading. However, research is lacking regarding the strengths and weaknesses of each system.

This combined literature- and experimental study aims to improve on this lack of knowledge by 1) describing scoring systems for arteriosclerosis in non-neoplastic kidney diseases, 2) illustrating their performance on randomly sampled images of arteries with various grades of sclerotic changes, 3) listing and elaborating on the advantages and disadvantages of each system, and finally, 4) making a recommendation of the most appropriate scoring system(s) for the assessment of arteriosclerosis.

2 Materials and Methods

2.1 Literature search

The first part of the study consisted of a literature search in order to find all the currently available scoring systems for assessment of arteriosclerosis in non-neoplastic cortical kidney biopsies. Specifically, the project started with pre-existing knowledge about four such scoring systems (Banff, Sethi, Remuzzi, ASI), and the literature search aimed to complete the list by discovering additional scoring systems.

Search algorithm in Pubmed on March 22nd, 2023: (scoring system[Title/Abstract]) AND (kidney biopsy[Title/Abstract] OR renal biopsy[Title/Abstract]): 58 results in total describing nine scoring systems, three of which were excluded because they referenced older scoring systems, five were excluded because they used the exact same scoring method as included systems, and one was excluded because it lacked specific instructions for scoring. The five remaining systems were divided based on whether or not they relied on semiquantitative assessment (“eyeballing”) or quantitative assessment with precise measurement tools. Three of them relied on the former, and two relied on the latter.

Inclusion criteria: Must have a well described method of scoring arteriosclerosis. Exclusion criteria: Method identical to or relying on an identical concept as another already included system. During the search, a number of scoring systems for renal quality did not involve the assessment of arteriosclerosis. These were not considered. Figure 1 shows the literature search process, and table 1 describes the various scoring systems found, whether they were included, and why.

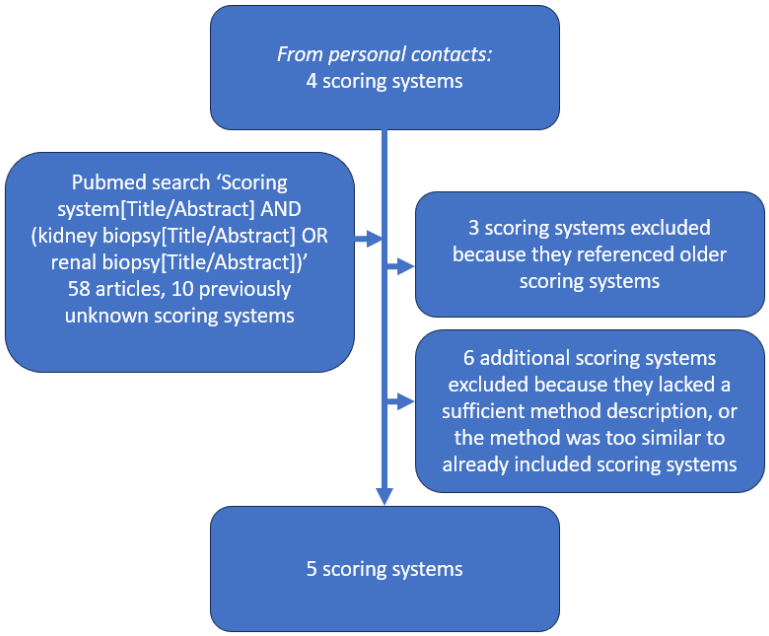

Figure 1. Decision tree showing the literature search process. A total of five different scoring systems were included in the study.

Table 1. An overview of the scoring systems that were identified and considered for the study. There were five included systems, and six excluded systems.

Scoring system	Type	Reference	Notes
Included			
Banff	Semiquantitative	(1, 10-13)	Leuven and Amyloid Score reference Banff with respect to arteriosclerosis grading, see references listed next to these systems farther down in this table.
Remuzzi et al (Pirani)	Semiquantitative	(14-18)	ISGFN scoring system references Remuzzi (Pirani) with regard to arteriosclerosis grading, see references
Sethi et al	Semiquantitative	(8)	
MAPI	Quantitative	(19)	
ASI	Quantitative	(20)	
Excluded			
NEPTUNE	Semiquantitative	(21)	Same principle as for Sethi et al, just 3 instead of 2 grades.
Tervaert et al	Semiquantitative	(22, 23)	Identical method to Sethi et al and Joh et al, only slightly different scoring thresholds
Joh et al	Semiquantitative	(24)	Identical method to Sethi et al and Tervaert et al, only slightly different scoring thresholds
VCI	Semiquantitative	(25)	Lacking sufficient method details for implementation
Oxford Cl.	Semiquantitative	(26, 27)	Identical to Sethi
Zhang et al	Semiquantitative	(28)	Identical to Banff

2.2 Scoring systems

The following is a description of the five scoring systems included in this study. In the first three described, a general principle of examining the whole artery and making a judgement based on the overall impression applies. The scoring persons (SP) must base their score only on what they can see without the use of any measurement tools. In the latter two scoring systems, precise measurements are made as further described under the sections of each respective scoring system.

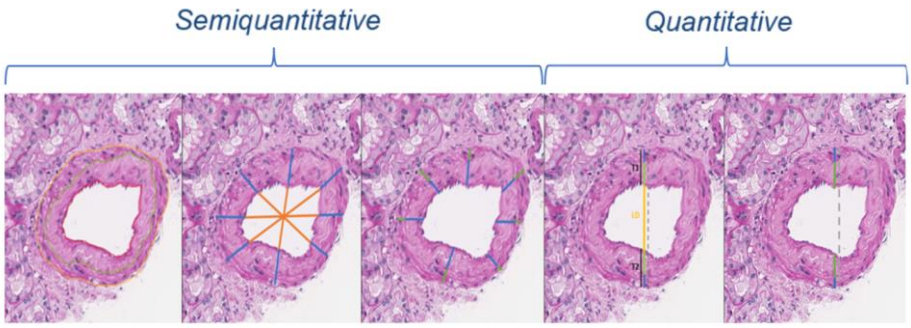

Figure 2. Illustrations indicating how the same image of a cross-sectionally cut renal artery is assessed differently using the various scoring systems. From the left to right: Banff; Remuzzi; Sethi; MAPI; ASI.

2.2.1 Banff

The Banff scoring system is widely used to record the status of kidney transplants as well as non-neoplastic kidney biopsies and is by many considered the gold standard (1, 10). Figure 4 details how the Banff system works with respect to arteriosclerosis grading.

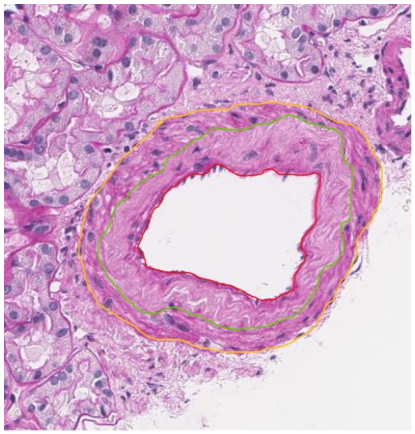

Figure 3. Artery annotated based on the Banff method. The green line marks the border between intima and media, the yellow line marks the border between media and adventitia.

Specifically, the figure depicts a renal artery with arteriosclerosis, as is clearly visible by the thickening of the intimal layer. In the Banff system, the degree of luminal narrowing is considered, and based on this, four different grades are specified: 0, 1, 2, and 3. The grades are defined as follows (1, 10): No arterial narrowing: 0. Mild narrowing up to 25 percent: 1. Moderate narrowing up to 50 percent: 2. Severe narrowing above 50 percent: 3.

2.2.2 Remuzzi

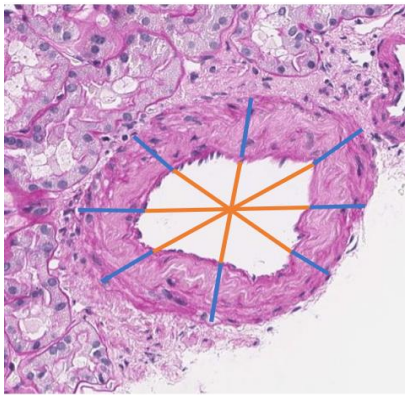

Figure 4 Artery annotated based on the Remuzzi method. Blue lines illustrate the wall thickness, orange lines the lumen diameter.

According to Remuzzi, it is the relationship between the thickness of the arterial wall, including both media and intima, and the diameter of the lumen that determines the score. In figure 5, approximations of the luminal diameter are marked in orange, while the blue lines indicate wall thickness. As previously described, the scoring persons do not select any specific areas of each artery image for comparison (i.e. select just one blue line and one orange line and compare those), but rather consider a general impression of the relationship between luminal diameter and arterial wall thickness based on the overall appearance of the vessel (i.e. compare “all”

possible blue and orange lines; not just the ones annotated in the illustration). Remuzzi et al. define 3 different grades (14, 15): No increased wall thickness: 0. Increased wall thickness but to a degree that is less than the diameter of the lumen: 1. Wall thickness that is equal to or slightly greater than the diameter of the lumen: 2. Wall thickness that far exceeds the diameter of the lumen with extreme luminal narrowing or occlusion: 3.

2.2.3 Sethi

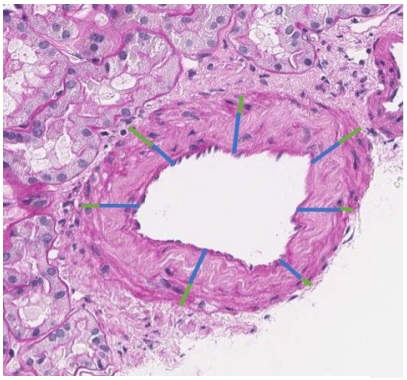

Figure 5 artery annotated based on the Sethi method. Blue lines indicate intima thickness, green lines media thickness.

In the scoring system proposed by Sethi et al, the intimal layer thickness is compared to the medial layer thickness. The lumen is not relevant. In figure 6, the blue and green lines indicate intima thickness and media thickness, respectively. It is clear that both media and intima thickness vary greatly within the same image, so the scoring persons follow the same principle here as with the Remuzzi system, forming a general impression of the media-intima relationship (i.e. comparing "all" possible blue lines with "all" possible green lines). Sethi et al defines 2 scores (8): A score of 0 if the intima thickening is less than the thickness of the media, and a score of 1 if it is equal to it or greater.

2.2.4 Maryland aggregate pathology index (MAPI)

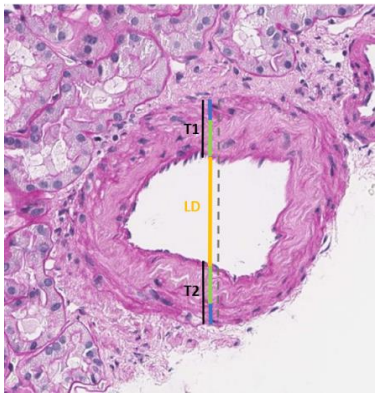

Figure 6. Artery annotated based on the MAPI method. The black lines mark the entire wall thickness, the orange line the lumen diameter.

MAPI measures the entire thickness of the arterial wall in two places, adds them up and then divides them by the diameter of the lumen. There are only two possible scores: 0 or 2.

The threshold is set to a ratio of 0.5 between the thickness of the two arterial walls and the diameter of the lumen; anything below yields a score of 0, and anything equal to or above will yield 2 points. The entire wall is measured, and not just the intimal layer (19).

2.2.5 Arteriosclerosis Index (ASI)

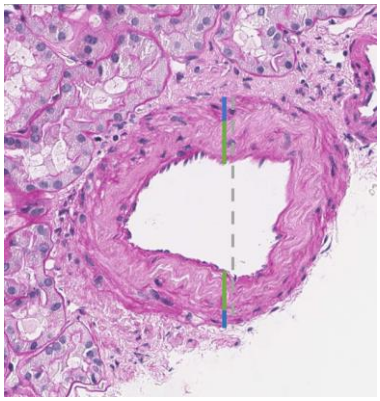

Figure 7 Artery annotated based on the method of Arteriosclerosis Index.

This is the only system using a continuous score with no upper limit. It measures the relationship between the intima and media in the same fashion as Sethi, the former layer being divided by the latter. To make the measurement more precise, the comparison is based on two separate measurements along the wall added together. This allows the SP to select two representative areas of the vessel. This score is not used to grade arteriosclerosis independently but has been used to compare relative degrees of it (20).

2.3 Acquisition and selection of raw images

Digital slides from 60 consecutive non-neoplastic kidney biopsies were retrieved and anonymized. These 60 biopsies consisted of 15 biopsies of each grade of the AS score (0 – 3). The grade of AS was defined based on the Banff scoring system in the diagnostic setting. Only sections stained with periodic-acid Schiff were used. The artery with the most severe degree of sclerosis was annotated in QuPath (29). These arteries were used for evaluation in the current study. See figure 7 (p. 16) for an illustration of the procedure for selecting the artery to be used for grading.

2.4 Scoring procedure

Each of the five scoring systems were used to assess the 60 selected arteries. The three semiquantitative scoring systems (Banff, Remuzzi, Sethi) were used by three scoring persons (SPs). Each SP used each scoring system twice with a washout period of two weeks, as further described below. The two quantitative scoring systems (ASI, MAPI) were used by a fourth SP to grade the 60 images once each.

A number of concerns were accounted for during the scoring process: In order to minimize bias, an independent scorer used the quantitative scoring systems. The three SPs using the semiquantitative systems came from different professional backgrounds: A non-clinician (denoted SP1), a general pathologist (denoted SP2), and a nephropathologist (denoted SP3). This selection of SPs was made in order to test the reliability of each system in relation to the professional background of each scorer. For the two quantitative systems, a single measurement conducted by one person was considered sufficient, since the scoring was quantitative and based on the use of precise measurement tools.

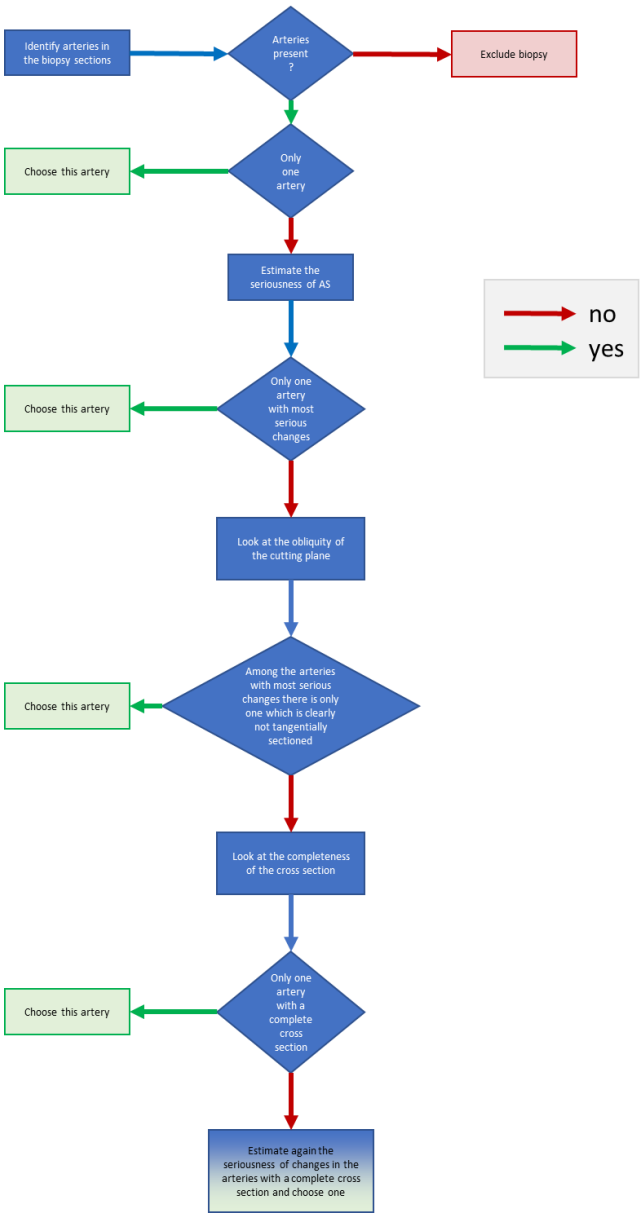

Figure 8. Procedure for selecting the artery to be used for grading.

Since the first three scoring systems mentioned do not rely on precise measurements, we assumed that there would be higher inter- and intra-rater variability. To investigate this variability, we set up the study as following: 1) We used three different raters, each with different academic expertise. 2) In order to control for intra-rater bias, each round of image scoring had a washout period, each round was conducted no less than two weeks apart to prevent memorization of prior results. 3) Also, to control for intra-rater variability, each of the three scoring systems was used twice by each scoring person to control for any possible random variance in the same test person. In the case of the other two scoring systems, we operated with the assumption that they would be free of bias as they rely on objective measurements. The scoring persons were allowed to conduct each round of image scoring within the span of one week, and each round of scoring was estimated to take between one and three hours. To make the scoring process easier to manage for each SP, they were given one week to complete each scoring system. Table 1 shows the plan for image scoring that the three SPs of the semiquantitative systems followed.

Table 2. Each test person had one week to finish each round of image scoring, and a two-week wash-out period prior to the subsequent round. SP1-3 = Scoring persons 1-3; scoring person 1 is a non-clinician, scoring person 2 is a general pathologist, and scoring person 3 is a nephropathologist.

	Banff Round 1	**Remuzzi Round 1**	**Sethi Round 1**	**Banff Round 2**	**Remuzzi Round 2**	**Sethi Round 2**
SP1	Week 1	Week 4	Week 7	Week 12	Week 15	Week 18
SP2	Week 1	Week 4	Week 7	Week 10	Week 13	Week 16
SP3	Week 1	Week 4	Week 7	Week 10	Week 13	Week 16

2.5 Evaluation of scoring results

The various systems were evaluated through a number of comparisons. Specifically, the results from each individual system were compared relative to different SPs (inter-observer variability) and relative to different scoring rounds (intra-observer variability). The different scoring systems were also compared to each to each other visually using scatter plots, box-and-whisker plots, and

contingency tables. Where relevant, Spearman's correlation coefficient (SCC) and p-values were also calculated.

We investigated inter-rater variability (IERV) for each of the three semiquantitative scoring systems (Banff, Remuzzi, and Sethi). We did this by using Krippendorff's alpha (hereby denoted KA) (30) to measure the rate of agreement between three sets of 60 scores – one set for each SP. Next, we investigated intra-rater variability (IARV) by having the SPs repeat the scoring process for all three scoring systems, so that we could compare each SP's two sets of scores for the same system.

Further, we wanted to investigate how well the five scoring systems agreed with each other, by directly comparing each system with the others via box-and-whisker plots and scatter plots. This was done to give a visual presentation of the relationship between each system's scores. For the scatter plots, a regression line was added, with the formula for these lines added above the scatter plots. Each comparison has the Spearman's correlation coefficient calculated.

For both IERV and IARV, our statistical measurement instrument of choice was Krippendorff's alpha (KA). This was not only because KA is considered a reliable measure of both IERV and IARV for a variety of different types of data sets, but also because it was especially well suited for our study for the following reasons: Our data set is ordinal, since each value is a number corresponding to the degree of arteriosclerosis in an artery. These numbers are ordered, that means a grade 1 arteriosclerosis is less severe than a grade 2 arteriosclerosis. The scoring systems provide specific instructions for assessing a score, but are also inevitably to some degree based on each SP's subjective judgement. KA, as opposed to other measurements such as Intra Class Correlation (ICC), is well suited for quantifying inter-rater variability for such data sets. Additionally, KA is ideal for handling incomplete data sets, which is beneficial in this study since each SP had the choice of abstaining from scoring arteries that were too difficult or impossible to score (30). Our data set therefore contains a number of arteries where no score was given. For all calculations using KA, we set the seed to '2023' and bootstraps to 10 000.

3 Results

3.1 Intra- and inter-rater variability (IARV and IERV)

Each SP scored the data set of 60 arteries twice with each semiquantitative system, and with a washout period of no less than 2 weeks. We wanted to see how consistent the scoring systems were between the first and second time. As further explained in chapter 1.5, we used KA to analyze this. According to KA, the higher the result of the calculation – which goes from 0 to 1 –, the stronger the agreement between the two compared sets of data. A result of 1 means a perfect agreement, while a result of 0 indicates no agreement. A result between 0.66-0.81 is considered "tentatively acceptable agreement", while anything above 0.81 is considered "acceptable agreement") (30).

Table 3. Inter- and intra-rater variability calculated using Krippendorff's alpha coefficient, CI 95%, bootstraps 10k, seed '2023'.

	Sethi	**Remuzzi**	**Banff**
IARV			
SP1	0.821	0.918	0.643
SP2	0.767	0.870	0.816
SP3	0.771	0.897	0.896
IERV			
SP1+2	0.685	0.578	0.562
SP1+3	0.752	0.639	0.682
SP2+3	0.827	0.548	0.692
SP1+2+3	0.746	0.588	0.631

Without knowing which scoring method is superior, it was logical to assume that Sethi would have the strongest intra- and inter-rater correlation given its binary score model. Table 3 shows that it did have the strongest inter-rater correlation, but had a relatively weak intra-rater correlation, below Remuzzi and similar with Banff, which is arguably a more difficult system to use. This

indicates that Sethi is a somewhat less reliable method relative to the two other systems, especially since the SPs had twice as many scores to choose from using the two other systems.

The results further show that Remuzzi had the lowest intra-rater variability (meaning the strongest agreement from the first to second round) for all three SPs. With the maximum average KA value achievable being 1.0, Remuzzi had the highest value with an average score of 0.895, with Sethi and Banff being essentially tied at 0.786 and 0.785, respectively. For the nephropathologist (SP3), Remuzzi and Banff were the systems with the strongest agreement.

Interestingly, SP1 – the non-clinician – had the most consistent scoring of the three SPs in Sethi and Remuzzi, while SP3 – the nephropathologist – was most consistent with the Banff system, followed by SP2 – the general pathologist – and then SP1. Possible reasons for this pattern are mentioned in the discussion. Sethi had the strongest inter-rater correlation for all comparisons between SPs. It is however as previously mentioned, a binary score system, and these results must be interpreted with this in mind.

Beyond Sethi, the second strongest agreement was from comparing the IERV Banff scores of SP2 and SP3 at 0.692, while Remuzzi had the lowest IERV scores (highest result was 0.639 between SP1 and SP3; see table 3). For Remuzzi, this indicates a relatively poor agreement, while Banff once again correlated relatively high apart from between SP1 and SP2. The most accurate scoring systems based on IARV and IERV KA measurements in combination, factoring in that one system has a binary score while the two others have a four-point score, the result from this analysis indicates that Remuzzi and Banff are similarly accurate with regard to IARV and IERV, while Sethi underperforms with regard to IARV, which indicates that this method is less reliable.
Banff appears to be more reliable in terms of IERV among the pathologists (SP2 and SP3) than with comparisons involving the non-clinician (SP1). This is likely because it is a more technically challenging system, which in turn becomes more reliable with practice. More on this in the discussion (chapter 4).

3.2 Correlations across different scoring systems

Arteriosclerosis Index (ASI) and the Maryland Aggregate of Pathology Index (MAPI) are interesting systems to compare to the other systems because they are based on the use of measurement tools, see further explanations in section 1.2. We used scatter plots to visualize the comparisons between the two quantitative scoring systems and the comparisons of the three semiquantitative scoring systems to each of the two quantitative systems.

Additionally, Spearman's correlation coefficient (SCC) was calculated for each comparison in order to quantify the correlations. The SCC measures the strength and direction of two different variables, and, unlike other correlation coefficients such as Pearson's, it is nonparametric, meaning it does not rely on any assumption of a distribution of normality. Different sources operate with different threshold values for what is considered a strong correlation, but in general, scores higher than 0.4 are widely considered to correlate (31, 32). Figure 10-11 illustrate all the ASI comparisons with other scoring systems.

3.2.1 Arteriosclerosis Index (ASI) compared to other systems

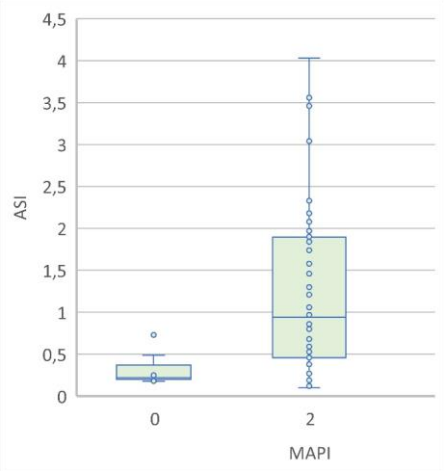

Figure 9. Box-and-whisker plot comparing MAPI to ASI. Horizontal line = median. Box = mid two quartiles. Top and bottom whiskers, respectively, are the minimum and maximum values excluding outliers. Data points outside of whiskers are outliers. P-value <0.001 (based on independent samples Welch T-test).

Assuming that ASI and MAPI measure similar properties and are both quantitative and thus likely to be more reliable, an initial hypothesis was that both systems should produce highly comparable scores. The box-and-whisker plot in figure 10 compares the two systems. It shows approximately a one-quarter overlap between the two samples (those scored as 0 and those scored as 2 in MAPI), meaning that one-quarter of arteries scored as 2 in MAPI have ASI scores that overlap with the ASI scores of the arteries scored as 0 in MAPI. Starting from an ASI score of approximately 0.48 and upwards, no arteries were scored as 2 in MAPI with the exception of a single outlier. About three-quarters of the arteries were in this group.

So, figure 9 shows that when the ASI score is higher than 0.48, the MAPI score is almost always 2, and when the ASI score is lower than 0.48, the MAPI score is 0 in about 75 percent of cases.

The question then becomes: Are MAPIs scores of 2 in the below 0.48 ASI range mostly false positives or not?

Based on the results from comparing MAPI to the other systems (see chapter 3.2.2), there are reasons to believe that this system is highly sensitive, and therefore gives almost no false negatives, but potentially some false positives. At the same time, the scatter plots in figure 10 comparing ASI to the other systems strongly indicate that ASI would be a highly unspecific system, since the spread of ASI scores is wide for almost all the scores in the other systems, and widen more for each point increase in the other systems. Even when other systems score a 3 out of 3, ASI measures anything from 4.02 and downward close to 0. Using the other systems as a reference for reliability, this makes it highly likely that ASI, which lacks a defined cut-off for defining an artery as arteriosclerotic, would give many false positives regardless of which cut-off was defined for this system. Therefore, there are some indications to claim that MAPI might often be more reliable than ASI, and that MAPI might be both a highly sensitive and specific system.

While what ASI measures is associated with what MAPI measures, the two systems can disagree to an unreasonable extent. This is clear from the fact that many of the MAPI scores of 2 are given very low ASI scores. Utilizing MAPI as a reference for reliability, if a cut-off was to be made for ASI, it should not be lower than 0.48 since this is the point at which the highly sensitive system MAPI begins to give scores of 0. Thus, if ASI gives a score below 0.48, the score is not reliable, and the artery must be scored using another system.

Figure 10 contains scatter plots comparing each of the three semiquantitative scoring systems to Arteriosclerosis Index (ASI). These were interesting comparisons to make because ASI is a quantitative scoring system, and the only system that uses a continuous scale as opposed to categories. In the case of Sethi, which only has two possible scores, the p-value is directly calculated using the independent samples Welch T-test. For Remuzzi and Banff, which have four

possible scores, the Spearman's correlation coefficient was calculated, and a p-value was derived from it.

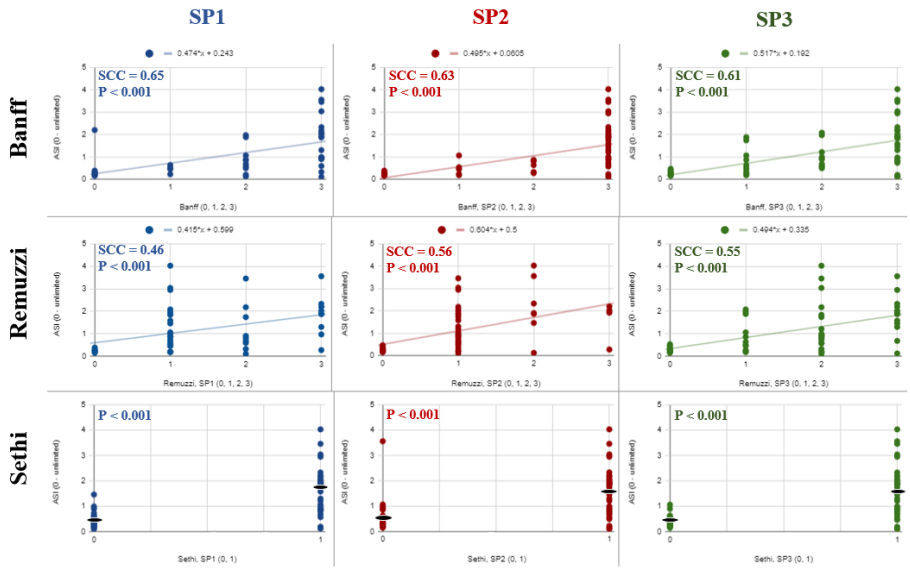

Figure 10. Scatter plots comparing the three semiquantitative scoring systems to the ASI system. Each column shows the scores of the same SP; each row shows the scores of the same scoring system. The Spearman's correlation coefficient (SCC) is calculated and included in the top left corner of the plots with Remuzzi and Banff. The p-value is added in the top left corner of each plot for all systems. Where the SCC is calculated, the p-value is derived from it. For Sethi, which has a binary score, the p-value is directly calculated using the independent samples Welch T-test. Each scatter plot for Banff and Remuzzi include a regression line, with the equation for the regression line directly above each plot. Mean values for the Sethi plots are visualized by black-and-white oval symbols.

The scatter plots in figure 10 indicate that ASI gives many false negatives relative to the scores of the three other systems, but not necessarily many false positives. There appears to be considerable agreement between all three systems with ASI when the ASI scores are relatively high, but not when the scores are below a certain threshold. Given the fact that this is also the trend when comparing ASI to MAPI (see figure 9), it might be considered more probable that ASI tends to give false negatives. Banff might be considered the system that most strongly agrees with ASI, since higher ASI scores appear to better correlate with higher Banff scores, but the regression lines

and SCC values are comparable with those of Remuzzi, so no definitive conclusions can be drawn. The p-values, all of which are less than 0.001, indicate that the differences between the scores of ASI and the other systems is not due to randomness.

3.2.2 MAPI and the three semiquantitative scoring systems compared to each other

The contingency tables in figure 12 display the relationships between MAPI and the three semiquantitative scoring systems, Sethi, Remuzzi, and Banff. Heat maps are added to better visualize the distributions of scoring combinations.

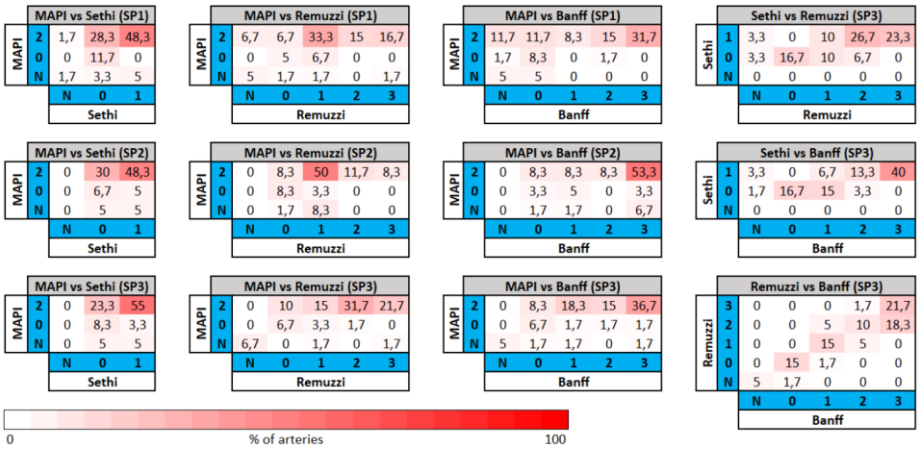

Figure 11. Contingency tables comparing the scores of four scoring systems (MAPI, Sethi, Remuzzi, and Banff). Each value is the percentage of the total number of combinations. Arteries that were not scored are also included (denoted as N). The first 3x3 tables from the left compare the MAPI scores (y-axis) to the scores of Sethi, Remuzzi, and Banff, respectively (x-axis). The rightmost column consists of tables comparing the scores of the three semiquantitative systems (Sethi, Remuzzi, and Banff) to each other. Each row of tables contains the scores of the same SP, except for the tables in the rightmost column, which are only based on SP3 scores (nephropathologist). Each table has a heat map to visualize the distributions of scoring combinations. All scores are based on the first round of scoring.

Since the systems included in this study measure arteriosclerosis using different methods and have different numbers of possible scores (0-1 in Sethi; 0-2 in MAPI; 0-indefinite in ASI; 0-3 in

Remuzzi and Banff), it is not a straightforward process to compare the results in figure 12. For example, there is considerable uncertainty about what a score of 1 in Sethi constitutes in the other two systems. Generally, though, one can assume (still, with some uncertainty) that a one-point-increase in the score of one system should roughly correspond to the same in another system if the numbers of possible scores are the same in both systems. This is only the case for MAPI and Sethi (two possible scores) and Banff and Remuzzi (four possible scores). If the numbers are not the same, this assumption cannot be made. The safest and most useful assumption we can make is that regardless of system, a score of 0 should roughly mean the same in as in any other system (i.e no clinically relevant arteriosclerosis) and a score above 0 should roughly mean the same (clinically relevant arteriosclerosis).

To make the analysis more readable, the denotation (y = score 1 & x = score 2) is used when discussing the combinations of different scoring systems. Also, all results are in percentages if nothing else is specified.

The tables comparing MAPI to the other systems reveal a number of patterns. Although there are some differences between the SPs, the general patterns found are largely the same. MAPI is clearly the most sensitive of all systems, as evidenced by the fact that the top row of each table (corresponding to a score of 2 in MAPI) is associated with relatively high percentages of scores of 0 in the other systems, compared with the relatively few percentages of a score of 2 in the other systems found in the middle row of each table (corresponding to MAPI scores of 0). In other words, when there is disagreement between MAPI and another system, and where one of the disagreeing systems gives a score of 0, MAPI is in the majority of cases not the system giving a 0, so it is likely the more sensitive system.

When comparing the disagreement with MAPI between the three semiquantitative systems, the most useful comparisons are those involving at least one score of 0, based on the previously described assumptions that a score of 0 should roughly mean the same in every system (no clinically relevant arteriosclerosis), and a score above 0 should roughly mean the same in every system (clinically relevant arteriosclerosis). Based on this, we can summarize the percentage of

disagreement and compare them between the three systems. An overview of the disagreements between systems is given in table 5.

Table 4. The sum of all disagreements between MAPI and the three semiquantitative systems (in percent), not including mismatch due to combinations with "no score". Combinations (y = 0 & x = 0) and (y = 2 & x >0) are considered as agreement, others are included as disagreement. In parentheses: All disagreement including mismatch due to combinations with "no score".

	Sethi	**Remuzzi**	**Banff**
MAPI disagreement, SP1	28.3 (38.3)	13.4 (25.2)	13.4 (31.8)
MAPI disagreement, SP2	35 (45)	11.6 (21.6)	16.6 (26.7)
MAPI disagreement, SP3	26.6 (36.6)	15 (18.4)	13.4 (18.5)

The same trends are evident for all three SPs, with Sethi having the highest disagreement, while Remuzzi and Banff have more similar disagreement. When combinations including "no score" are included, all SPs rate Sethi has having the most disagreement, followed by Banff, and then Remuzzi.

Remuzzi is special in this context because it uses the same guiding principle for scoring as MAPI does, the difference being that the latter uses precise measurements and a mathematical formula to arrive at a score, while the former is based solely on "eye balling". It is therefore no surprise that Remuzzi has the overall lowest rates of disagreement with MAPI. With the same logic, it is also no surprise that Sethi, which uses a very different method, has the highest rates of disagreement. The fact that Remuzzi and Banff agree so well when directly compared, indicates that the methods more closely correlate with each other than they do with Sethi, which is supported by their similar rates of disagreement with MAPI.

The fact that the most associated scores between MAPI and Remuzzi is 2x1 for SP1 and SP2, and 2x2 for SP3 (as opposed to being 2x3) might simply come from the fact that Remuzzi measures arteriosclerosis by using a different method than the two other systems. Another explanation is that Remuzzi operates with different cut-offs for the different degrees of arteriosclerosis. This is supported by the results of directly comparing Banff with Remuzzi, where the highest score of 3

only occurs in 23.4 percent of Remuzzi scores, but 40 percent in Banff. The opposite trend is the case for the score of 2, where the percentages are 33.3 vs 16.7. For scores 1 and 0, the percentages are identical or almost identical (1: 20 vs 21.7; 0: both have 16.7), and they have the same number of unscored arteries (both have 5). In other words, Remuzzi and Banff seem to be strongly associated in their ability to differentiate an arteriosclerotic artery from a healthy one (going from a score of 0 to 1), but they differ in assessing degrees of arteriosclerosis: An artery scored as a 3 in Banff will only be a 3 in Remuzzi about half the time and a 2 in the other half, but an artery scored as a 3 in Remuzzi will almost always be scored as a 3 in Banff.

Sethi differs from the two other systems in its ability to differentiate arteries with arteriosclerosis from healthy arteries (going from score 0 to 1). When Remuzzi and Banff give a score of 0, Sethi always gives a 0. Opposite, when Sethi gives a score of 0, Banff and Remuzzi only give a 0 about half the time, and a 1 or a 2 in the other half. In other words, it takes more for Sethi to define an artery as arteriosclerotic. Whether these arteries should be considered arteriosclerotic or not, is not entirely clear. The fact that Sethi is most associated with the ASI system, which in turn is a likely unspecific system, implies that also Sethi might also be more unspecific.

Reviewing differences between scoring persons, we see the same general trends between all three SPs except for in the MAPI vs Remuzzi comparison. SP1 and SP2 used Remuzzi differently than SP3, scoring between 1/3 and ½ of arteries as a 1, as opposed to SP3, who only scored 1/8 of arteries as a 1, and ½ of arteries as a 2 or 3.

3.3 Missing values

Another measure of scoring system quality is comparing the number of arteries where an SP felt unable to apply the scoring system, resulting in a missing score. The heat map in figure 13 offers

a visualization of cases where arteries could not be scored by the respective scoring system. The table displays the complete dataset and both the first and second scoring from all scoring persons.

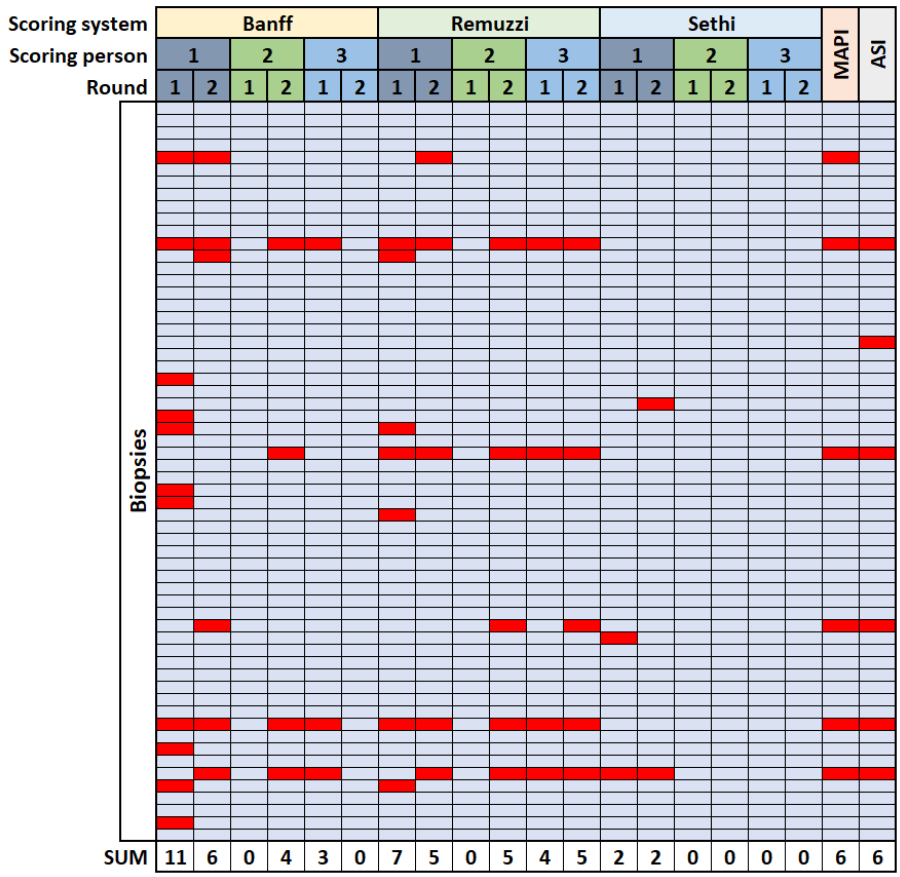

Figure 12. Heat map visualizing the missing values in the dataset. Each red rectangle within a column indicates a missing value. The bottom row shows the total number of missing values per column.

The heat map reveals a clear difference between SP1 and the two other SPs, as SP1 abstained from scoring more arteries than SP2 and SP3 in every system. This is true for both Banff, Remuzzi, and Sethi. Across all three systems, SP1 abstained from scoring 11 percent of arteries on average, as

opposed to 0 and 4 percent for SP2 and SP3, respectively. A likely reason for this is SP1s lack of experience using these scoring systems, making it difficult to score the most challenging arteries.

Reasons for not scoring certain images include the following: Uncertainty about borders; only partial artery; difficulty determining the angle at which the artery was cut; uneven wall- or wall layer thickness. For example, there could be great variation in the intimal thickness along a stretch of artery wall. ASI and Sethi have an advantage in that they do not rely on the lumen for scoring. This enables them to also score partial arteries. Sethi was the clear winner as far as the ability to score arteries is concerned, as the pathologists were able to score all 60 arteries with this system, and the non-clinician 58 of them. With 54 scored arteries, the ASI system was more comparable to the other systems in its number of successful scores. Since ASI relies on specific measurements of the intimal and medial layers, it can be more challenging to know where to measure as some arteries appear with significant variation in wall layer thickness. Identifying the basal membrane was also challenging in some arteries, which made it difficult to know where the transition between intima and media was.

Banff had a very high number of missing values, but only for the non-clinician (SP1) – 11 percent vs. zero for SP2 and three percent for SP3. The two other systems, Remuzzi and MAPI, had similar rates of scored arteries, which makes sense since these rely on the same method of scoring. Remuzzi had the second highest rated arteries after Sethi, indicating that it is also a system with a high ability to score arteries.

4 Discussion

This study intended to improve on the lack of knowledge of the different scoring systems used by pathologists to score arteriosclerosis in non-neoplastic kidney biopsies. It sought to do so through a series of four main steps starting with a literature search to identify scoring systems, selecting and testing a suitable number of the described systems against a fixed set of randomly sampled renal arteries, and, finally, extracting as much information as possible about the systems from the analysis of the results of these tests.

Of the five systems included, all three semiquantitative systems (Sethi, Remuzzi, Banff) were used by each of three scoring persons of different medical backgrounds, a non-clinician (SP1), a general pathologist (SP2), and a nephropathologist (SP3). In addition, a fourth scoring person used the two quantitative systems (ASI, MAPI). This section will go through each of the four main aims and elaborate on why we believe we did or did not meet each aim. Also, important limitations and prospectives are discussed.

The literature search yielded an additional ten scoring systems beyond the four we already had knowledge of. These were categorized and described in table 1. We believe that table 1 contains a useful overview of the different scoring systems currently in use by pathologists to score degrees of arteriosclerosis, and thus that we reached this aim. We could have included more scoring systems from the literature search into the experimental phase, but we decided against this primarily for the following reasons: 1) Lack of resources: We did not have enough time and manpower for a larger study. 2) The five included scoring systems covered all the different methods of scoring that we discovered, so including additional scoring systems whose only difference was the scoring range and scoring cut-offs would be of limited return value for our resources.

The second aim, which was to execute the experimental phase, was also completed successfully. This was the most time-consuming phase, as there were four scoring persons involved, three of which had to score three systems – and do so twice – a total of six sessions of scoring with a washout period of no less than two weeks between each.

The decision to include the two quantitative scoring systems was made based on the assumption that such systems are more accurate because they allow the use of precise measurement tools, they could therefore be used as a more (read: not entirely) objective standard, in the absence of a known "ground truth" – a key limitation of our study. The scoring person using ASI and MAPI was, however, not a trained pathologist. There is a question of possible implications of this, but we believe it would not affect the scores substantially because the scoring systems in question are based on precise measurement tools and a minimum of subjective interpretation, especially in the

case of ASI, which uses a scaled point system. In addition, the annotations on which the measurements were based, were also validated by an external nephropathologist.

The three other scoring persons were intentionally selected based on their professional background in order to investigate the effect of prior training on the use of the scoring systems. We were able to draw several conclusions about the utility of the scoring systems based on this, such as Banff likely being a more reliable method for those with experience using the system, but a less reliable system for those lacking prerequisite expertise.

Other potential limitations include factors such as the state of mind of the scoring person on the particular day of scoring, the different scoring persons might have spent different amounts of time scoring each artery, and other factors like mood and concentration can vary from day to day, and indeed throughout the day. Such variables were not controlled for, and may or may not have had a significant impact on each scoring person's ability to score appropriately.

Our third aim was to elaborate on the strengths and weaknesses of each system. We are confident that we also achieved this aim, although little can be claimed with full confidence due to factors such as limited resources and the fundamental differences between the scoring systems which make comparing them challenging, and the lack of a ground truth, which made evaluations of absolute performance impossible. For Sethi and Remuzzi, it was the non-clinician (SP1) that scored the most consistently of all. Conversely, the nephropathologist had the most consistent scoring using Banff, while the non-clinician performed the least consistent here.

This pattern could be explained by the fact that Sethi and Remuzzi arguably involve easier measurements than Banff: In the former two systems, scoring is based on comparing distances (thickness of walls, wall layers, and/or luminal diameter), whereas Banff considers reduction of luminal area, which is a less straight-forward process. Banff is, as previously mentioned, the most widely used classification system worldwide (1), and was also used for the scoring of the arteries in the diagnostic setting, therefore SP3 had more prior experience using this system than any other system.

As shown in figure 13 it could be argued that Banff uses a method that requires more training in order to master. The luminal radius as an indication for luminal area can be deceiving because a relatively modest reduction in luminal radius has a relatively large impact on the area. For example, a luminal radius reduction of 29 percent may not seem substantial, but it actually corresponds to a 50 percent reduction in luminal area, which yields the highest possible score in the Banff system. The non-clinician performed the poorest of the scoring persons using Banff, and among the scoring systems used by the non-clinician, Banff performed the poorest. The opposite was the case for the general pathologist and the nephropathologist. This is in line with the idea that Banff is the most difficult system to use.

Figure 13. The Banff scoring method is based on determining the degree of luminal narrowing. The illustration shows how a relatively low reduction in luminal radius has a relatively large reduction in luminal area The figure was reproduced with permission under Creative Commons license 4.0 (CCBY) from Roufosse 2018 (1).

If an SP has a low intra-rater variability in a particular scoring system, that does not necessarily mean that they score accurately with said system. It only means that they agree with themselves, so even if an SP misunderstands the instructions and makes mistakes, but then repeats the same

mistakes in the second round of scoring, they might still get a low intra-rater variability. That is not to exclude the possibility that SP1 may have in fact scored more appropriately than the others. Also, more experienced SPs might have had more difficulty ignoring other types of information that they observe in the artery biopsies (e.g. changes at the cellular level that should not be considered when scoring). SP1, who had no prior knowledge of scoring arteries, might have had fewer "distractions" from the more or less straight-forward scoring instructions of each system. It also makes sense, then, that when the instructions become less intuitive (e.g. Banff), a less experienced scorer (SP1) might start to have difficulties following the instructions, while the exact opposite trend was evident for SP2 and SP3, who performed best using Banff. A scoring system that is easy to understand and use even for a non-clinician has a great advantage over other systems since it can be used not only by experienced pathologists. This was the case for Sethi.

Our fourth aim was to make a recommendation of the most appropriate scoring systems included in the study. The following are the most prominent arguments we have for making such a recommendation.

Since all the systems involved are used in clinical practice today, our expectation was that all of them would prove reasonably accurate and reliable, but that some systems might be somewhat more accurate and reliable than others given that the scoring systems measure different things and use different ways of scoring them. It would not be reasonable to assume that all these systems would happen to be exactly as accurate and reliable when they actually measure different parts of the artery, have different ways of measuring, different score ranges, and different cut-offs. We also expected the quantitative methods to show a stronger correlation with each other than the semiquantitative ones, since they involve using precise measurements.

Our first expectation proved correct: There are many differences between the scoring systems. Our second expectation proved incorrect: MAPI and ASI did not match as well as expected if they truly measured the same thing. Of the two, MAPI is the more favorable of the two since all the results in combination strongly indicate that it might be a highly sensitive (giving few false negatives) and possibly also highly specific (giving few false positives) system. ASI, on the other hand, appears less reliable as it gives very low scores for arteries that all the other four systems score as

moderately high or even maximally high. ASI appears to be comparably sensitive to MAPI, but far less specific, as evidenced by the box-and-whisker plot comparing the two systems, combined with the scatter plots comparing ASI to the semiquantitative systems. In all ASI comparisons, ASI appears to underestimate the degrees of arteriosclerosis in many arteries relative to the other systems. Without further evaluating the performance of this system against a ground truth, we would therefore not recommend its use over that of the other systems.

MAPI, on the other hand, appeared reliable in every way we were able to measure it except for the number of missing values, which was moderately high. The main question regarding this system is how specific it is. We found no evidence to suggest that MAPI is unspecific. About one-quarter of MAPI scores disagree with ASI. When they disagree, it is almost always the case that MAPI has a "positive" score, and ASI a "negative" score (meaning a score in the same range as it gives for a MAPI score of 0). In theory, a system that appears to be highly sensitive could potentially be moderately or even minimally sensitive but appear highly sensitive because it includes many "false positives" (low specificity) that deceive us in the analysis. If we examine the method of MAPI, however, an issue with too low specificity is not likely. The wall of the artery is measured in two places, the two widths added and then divided by the diameter of the lumen. In other words, MAPI follows the same scoring principal as Remuzzi, but uses exact measurements and a formula to arrive at a score. Since MAPIs cut-off for scoring a 0 is lower than Remuzzi, it is likely a more sensitive system, and definitely not less sensitive. MAPI has the fewest numbers of arteries scored as 0 of all the systems. MAPIs sensitivity is therefore likely close to optimal (100 percent).

The scatter plots comparing the three semiquantitative systems to ASI indicate that many of the ASI scores are false negatives, so the semiquantitative systems indirectly favor MAPIs specificity over ASI. In order to further investigate specificity, we would need other parameters to compare our results to, such as clinical data from the patients from whom the biopsies were taken, or a ground truth that all the systems could be compared against.

Pathologists have limited time to review arteries, and using a system with measurement tools is time-consuming. Quantitative scoring systems such as MAPI are therefore generally less practical.

The great advantage of the semiquantitative scoring systems is that they do not rely on such measurements.

Of the three systems in this category, Sethi has the advantage of being able to score the most arteries, even partial arteries without an intact luminal space. Additionally, it has the lowest observed IERV value. However, since it is a binary system with limited room for variation, direct comparison with other systems might potentially be biased. It had a lower IARV than the two other systems, which indicates that the system is less concise than the other two. It has the same scoring method as ASI (comparing intima to media), so it was no surprise that it appeared to match the best with this system. Of all semiquantitative systems, Sethi is the one with the most scores of 0. Given its association with ASI, which our results indicate might be less specific than the other systems, a relevant question is whether Sethi is also less specific. MAPI, which is the only other binary scoring system, disagrees more with Sethi than the other two semiquantitative systems when compared using the premises outlined in chapter 3.2.2. This could mean that Sethi fails to identify some arteries with low – yet clinically relevant – degrees of arteriosclerosis, but it could also mean that it more appropriately identifies low degrees of arteriosclerosis as clinically irrelevant.

Remuzzi uses the same method of scoring as MAPI, and is closely associated with it. It has the highest IARV value despite having four different possible scores, but a low IERV value. When reviewing the differences in scores between the scoring persons, the low IERV is explained by the fact that SP3 scored more 2s and 3s while SP1 and SP2 scored more 1s, with a similar ratio of 0s to above 0s. SP3 has the most experience using Banff, a system which we have showed more frequently scores 3s than Remuzzi does, and it is possible that SP3 is impacted by this, especially because the first round of scoring was done using Banff. We attempted to minimize the bias-effect of this by having a washout period of no less than two weeks, but cannot rule it out. Ideally, we should have done the Banff system in the last round to further reduce this bias effect. If there was a bias effect, then the IERV should have been higher. Remuzzi had the second fewest missing values after Sethi, meaning it could score most or all of the arteries regardless of scoring person. It was relatively weakly associated with ASI.

Banff is a truly unique system since it has a different method of scoring than all four other systems, as discussed earlier in this chapter. It is likely a more challenging method, and therefore not an entirely unexpected finding that the non-clinician (SP1) had a very high number of missing values in this system. It relies on an intact lumen for scoring, which renders it useless for scoring the subset of arteries without intact lumen. The pathologists (SP2 and SP3) therefore also had some missing values. Its IARV was relatively low for the non-clinician, while high for the pathologists. The IERV value was generally higher than Remuzzi but lower than Sethi. Given the potential bias effect previously described (which could have also affected the Sethi values involving SP3, not just Remuzzi), there is some uncertainty around the IERV values. Banff had the strongest association with MAPI and Remuzzi, but differed from the latter in that it more often gave the maximum score of 3. This was the case regardless of scoring person.

Remuzzi and Banff appear to be evenly reliable scoring systems despite using completely different methods of scoring. There is a difference in how they score degrees of established arteriosclerosis (whether an artery should be 2 or 3), but they are very similar in determining whether an artery has any clinically relevant arteriosclerosis or not (0 or above 0). Banff appears to require more training than Remuzzi. In summary, both of these systems can be recommended for use in clinical practice, but Banff requires more training.

An important limitation of this study is that all the results are relative to each other because all comparisons were between the five scoring systems, with no ground truth or clinical data to compare the results to. Our results are therefore useful in order to say something about the relationships between the systems, such as the fact that MAPI appeared more sensitive than the semiquantitative systems, but we cannot say anything with certainty about whether it is overly sensitive (i.e. scoring normal arteries as sclerosed), accurately sensitive, or even too little sensitive (missing sclerosed arteries). Despite the lack of a ground truth, we might approximate a ground truth by comparing many scoring systems and identifying common patterns among them. The more scoring systems are compared, the stronger the approximation to a ground truth.

A limitation was the number of scoring persons. Due to lack of resources, we only had one scoring person per category, which prohibits us from drawing general conclusions about the differences in

Kommentert [WHNH1]: How to distinguish higher sensitive from lower specificty? I have to think about this myself a bit. See comment below

Kommentert [LH2R1]: What do you think of the edited text?

Kommentert [WHNH3R1]: I made one suggestion in the end of the sentence

Kommentert [WHNH4]: You could also discuss that a system can not be more sensitive than 100% (at which point it would detect all the sclerosed arteries), but if it still detects more than those it means it has decreases specificity. For instance, just as an example, Sethi, Banff, and MAPI might all have a sensitivity of could also have 100% sensitive (just as MAPI) but MAPI might have a lower specificity, meaning it scores more non-sclerosed arteries as sclerosed? Would that look different in the results?

scoring system performance between non-pathologists, general pathologists, and nephropathologists. A suggestion for a new study would be to include three scoring persons per group (three nephropathologists, three pathologists, and three non-clinicians), which would make it possible to compare group averages more representable of each group. Still, many of the patterns in our analysis are the same when comparing the different scoring persons, suggesting that a larger-scale study would come to many of the same conclusions as in this study.

Further studies are required in order to increase our understanding of how different arteriosclerosis scoring systems compared. Future research should seek to avoid the limitations affecting the current study, especially with regard to the lack of a ground truth.

5 Conclusion

In conclusion, this study has reached all four of its main aims. The first aim was to collect and describe the most commonly used scoring systems for grading arteriosclerosis in non-neoplastic kidney biopsies. We were able to find and describe 11 different systems. Our second aim was to use the scoring systems on a randomly sampled set of non-neoplastic kidney arteries. Of the five scoring systems included in this study, two were quantitative (ASI and MAPI) and three were semiquantitative (Sethi, Remuzzi, and Banff). Three scoring persons of different professional backgrounds used each scoring system twice to score the same set of 60 arteries. A fourth scoring person used the two quantitative systems to score the same arteries. This allowed us to compare the scoring systems relative to professional background of the scoring person, time, other semiquantitative scoring systems, and to quantitative scoring systems. We also examined missing values. Among the measures to minimize bias was a washout period of no less than two weeks between every round of scoring.

Aims three and four were to elaborate on strengths and weaknesses of the different scoring systems, and to make a recommendation of the most favorable scoring systems. Of the two quantitative systems, MAPI appeared as the more favorable based on the overall findings, especially because it is likely highly sensitive, and no evidence was found to suggest that it is unspecific. Since ASI was an outlier relative to the four other systems, usually by giving low scores

when the other systems gave higher scores, it is possibly a sensitive, yet highly unspecific system, and can therefore not be recommended unless it proves reliable when compared to a ground truth.

As pathologists have limited time to score arteries, semiquantitative scoring systems have a clear advantage. Of the three systems under this category, Banff and Remuzzi appeared equally useful and reliable. They have similar ability to score arteries, with approximately the same number of missing scores of the sample arteries. They appear similarly consistent in scoring across time (IARV) and between scoring persons (IERV), but Banff is less consistent when used by a non-clinician, suggesting that it requires more training. Remuzzi and Banff score arteries as arteriosclerotic or non-arteriosclerotic (scores 0 vs 1-3) similarly, but differ in their scoring of degrees of arteriosclerosis (scores 2 vs 3). The implications of their difference in arteriosclerosis degree scoring remains unknown. Based on our findings, we can generally recommend Remuzzi and Banff for scoring arteriosclerosis, but Banff appears to require more training. It is important to note that these recommendations might have been different if we had established a ground truth.

The main strength of the third semiquantitative scoring system, Sethi, was its ability to score arteries. It had only two missing scores for the non-clinician, and no missing scores for the pathologists. It had a relatively low IARV value, indicating that the system might be somewhat difficult to use. The high IERV value could potentially be biased because Sethi only has two possible scores, limiting the potential for variation. Otherwise, the IERV value indicates that it is in fact consistent between different scoring persons. It uses the same scoring method as ASI, and was somewhat associated with this system. Sethi and ASI gave relatively low scores more frequently than the other systems. Little can be concluded with about the specificity of these scoring systems, but there are indications that ASI might be a less specific system because it frequently gives relatively low scores when all other systems give moderate to high scores. Sethi is the system that most often scored arteries as non-sclerotic (score of 0), indicating that it considers mild arteriosclerosis as clinically irrelevant. If this premise is true, Sethi would be a more favorable system. There is considerable uncertainty surrounding these results because they are based on relative comparisons between scoring systems.

This study lacked a ground truth or other parameters to compare the data to, such as clinical information about the patients from whom the artery biopsies were taken. Further studies are required in order to investigate the relationships between these scoring systems, especially regarding specificity. It is unclear how a reliable ground truth can be established in this context. One possibility is to compare the system performances with relevant clinical data connected to the patients the artery biopsies were taken from. When comparing many scoring systems with each other, one might be able to approximate a ground truth by identifying common patterns and excluding outliers.

Future studies will benefit from including more scoring persons in each scoring group (non-clinician, general pathologist, and nephropathologist), finding more ways to compare the scoring systems, including other parameters of data to compare the scoring results to, such as clinical information connected to the patients the biopsies were taken from, and establishing a ground truth to measure the scoring systems up against.

6 Appendix

Table 5 Complete data set of Sethi, Remuzzi, and Banff.

Biopsy	SP1 B1	SP2 B1	SP3 B1	SP1 B2	SP2 B2	SP3 B2	SP1 R1	SP2 R1	SP3 R1	SP1 R2	SP2 R2	SP3 R2	SP1 S1	SP2 S1	SP3 S1	SP1 S2	SP2 S2	SP3 S2
1	3	3	2	3	3	2	2	1	2	2	1	2	0	0	0	0	0	0
2	1	2	1	2	2	1	1	1	1	1	1	1	0	0	0	0	0	0
4	3	3	3	3	3	3	2	1	2	1	1	2	0	1	1	1	1	1
6	3	3	3	3	3	3	2	1	2	2	1	2	1	1	1	1	1	1
7		3	3		3	3	1	1	3		2	3	1	1	1	1	1	1
11	0	3	0	1	2	0	0	0	0	0	0	0	0	1	0	0	0	0
17	2	3	2	3	3	2	2	1	3	2	1	3	1	1	1	1	1	1
18	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	1
19	2	1	1	2	1	1	1	1	1	1	1	1	1	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	3	3	3	3	3	3	3	1	2	3	1	2	1	0	1	1	1	1
22		3				3		1					1	1	1	1	1	1
23	0	2	1		1	1		1	2	2	1	2	0	0	0	0	0	1
24	2	2	3	2	2	3	1	1	2	1	1	3	1	0	1	1	0	1
25	3	3	3	3	3	3	3	2	3	3	3	3	1	1	1	1	1	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	3	3	3	3	3	3	1	2	2	1	1	2	1	1	1	1	1	1
29	1	1	1	2	2	2	1	1	1	1	1	1	0	0	1	0	0	1
30	3	3	3	3	3	3	3	2	3	3	2	3	1	1	1	1	1	1
31	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	1
32	3	3	3	3	3	3	1	1	2	1	1	2	1	1	1	1	1	1
37	3	3	2	2	3	2	1	1	2	1	1	2	0	1	1	1	1	1
40		3	3	2	3	3	1	1	2	1	1	1	1	1	1	1	1	1
43	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0
45	3	3	2	3	3	1	2	1	2	2	1	2	0	0	0		0	1
46		3	3	3	2	3	1	2	3	1	2	3	0	1	1	1	0	1
50		3	2	2	3	2		1	2	1	1	2	1	1	1	1	1	1
51	1	1	1	1	1	1	0	1	1	1	1	2	0	0	0	0	0	0
55	0	0	0	0		0		0					0	0	0	0	0	0
56	3	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61		3	3	2	2	3	1	1	3	1	1	3	1	1	1	1	1	1
64		3	1	3	3	3	2	1	2	2	2	3	1	1	1	1	0	1
67	2	3	3	2	3	3		2	3	1	2	3	1	1	1	1	1	1
68	2	3	1	2	2	3	1	1	1	1	1	1	1	1	1	1	1	1
72	3	3	3	3	3	3	3	1	3	3	3	2	1	1	1	1	1	1
73	0	2	1	3	3	3	3	3	1	3	3	3	0	0	0	0	0	0
75	3	3	3	3	3	3	3	2	3	3	3	3	1	1	1	1	1	1
76	3	3	2	2	2	2	1	1	1	1	1	2	1	1	1	1	1	1
83	1	1	1	2	2	1	1	1	0	1	1	0	0	0	0	0	0	0
84	2	0	1	1	1	1	1	1	2	1	1	2	0	0	0	0	0	1
85	0	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	1
89	0	1	1		1	1	0	1	1	0			0	0	0	0	0	1
91	3	3	3	3	2	3	3	2	3	3	2	3		0	1	1	0	1
92	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
94	2	3	2	2	2	2	1	1	2	1	1	2	1	1	1	1	1	1
96	2	1	2	1	1	2	1	1	1	1	1	1	0	1	1	0	0	1
97	3	3	2	3	2	1	1	1	2	2	1	2	1	1	1	1	0	1
99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
106		3				2		1					1	1	1	1	1	1
107	0	3	3	3	3	3	1	1	2	2	1	2	1	1	1	1	1	1
110		2	3	2	2	3	2	1	2	1	1	2	1	1	1	1	0	1
111	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	1
114	0	3				2	3	1						0	0		0	0
117		3	3	3	3	3		3	3	3	3	3	1	1	1	1	1	1
120	2	3	2	2	3	2	1	1	1	1	1	1	1	1	1	1	1	1
121	3	3	3	3	3	3	2	1	2	2	1	2	1	1	1	1	1	1
122		3	3	3	3	3	1	1	2	1	1	2	0	1	1	1	1	1
123	2	3	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1

Table 6 Complete data set of MAPI and ASI.

Biopsy	Arteriosclerosis Index score	MAPI score
1	0,59	2
2	0,63	2
4	0,1	2
6	3,46	2
7	2,95	
11	0,22	0
17	0,68	2
18	0,18	0
19	1,06	2
20	0,38	2
21	0,97	2
22		
23	0,32	2
24	0,86	2
25	2,33	2
26	0,28	2
28	4,03	2
29	0,46	2
30	1,87	2
31		2
32	3,04	2
37	0,99	2
40	1,84	2
43	0,46	2
45	0,9	2
46	1,46	2
50	1,21	2
51	0,22	2
55		
56	0,32	2
59	0,25	0
61	1,58	2
64	1,74	2
67	0,12	2
68	0,86	2
72	1,3	2
73	0,27	2
75	1,9	2
76	2,08	2
83	0,53	2
84	0,2	2
85	2,2	2
89		
91	3,56	2
92	0,22	0
94	0,57	2
96	0,49	0
97	0,91	2
99	0,15	2
103	0,18	0
106		
107	0,19	2
110	0,8	2
111	2,02	2
114		
117	1,92	2
120	1,97	2
121	2,18	2
122	0,73	0
123	1,88	2

7 References

1. Roufosse C, Simmonds N, Clahsen-van Groningen M, Haas M, Henriksen KJ, Horsfield C, et al. A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology. Transplantation. 2018;102(11):1795-814.
2. Kumar V, Abbas AK, Fausto N, Mitchell R. Robbins Basic Pathology. 8th ed. Kumar V, Abbas AK, Fausto N, Mitchell R, editors. Philadelphia, USA: Saunders, an imprint of Elsevier Inc.; 2007. 960 p.
3. McGill HC, Jr., McMahan CA, Herderick EE, Malcom GT, Tracy RE, Strong JP. Origin of atherosclerosis in childhood and adolescence. Am J Clin Nutr. 2000;72(5 Suppl):1307s-15s.
4. Zhao Y, Vanhoutte PM, Leung SW. Vascular nitric oxide: Beyond eNOS. J Pharmacol Sci. 2015;129(2):83-94.
5. Dos Santos VP, Pozzan G, Castelli V, Caffaro RA. Arteriosclerosis, atherosclerosis, arteriolosclerosis, and Monckeberg medial calcific sclerosis: what is the difference? J Vasc Bras. 2021;20:e20200211.
6. Micheletti RG, Fishbein GA, Currier JS, Fishbein MC. Mönckeberg sclerosis revisited: a clarification of the histologic definition of Mönckeberg sclerosis. Arch Pathol Lab Med. 2008;132(1):43-7.
7. Tracy RE, Berenson G, Wattigney W, Barrett TJ. The evolution of benign arterionephrosclerosis from age 6 to 70 years. Am J Pathol. 1990;136(2):429-39.
8. Sethi S, D'Agati VD, Nast CC, Fogo AB, De Vriese AS, Markowitz GS, et al. A proposal for standardized grading of chronic changes in native kidney biopsy specimens. Kidney Int. 2017;91(4):787-9.
9. Rule AD, Semret MH, Amer H, Cornell LD, Taler SJ, Lieske JC, et al. Association of kidney function and metabolic risk factors with density of glomeruli on renal biopsy samples from living donors. Mayo Clin Proc. 2011;86(4):282-90.
10. Racusen LC, Solez K, Colvin RB, Bonsib SM, Castro MC, Cavallo T, et al. The Banff 97 working classification of renal allograft pathology. Kidney Int. 1999;55(2):713-23.
11. Rubinstein S, Cornell RF, Du L, Concepcion B, Goodman S, Harrell S, et al. Novel pathologic scoring tools predict end-stage kidney disease in light chain (AL) amyloidosis. Amyloid. 2017;24(3):205-11.
12. Naesens M, Kuypers DR, De Vusser K, Vanrenterghem Y, Evenepoel P, Claes K, et al. Chronic histological damage in early indication biopsies is an independent risk factor for late renal allograft failure. Am J Transplant. 2013;13(1):86-99.
13. Navarro MD, López-Andréu M, Rodríguez-Benot A, Ortega-Salas R, Morales ML, López-Rubio F, et al. Significance of preimplantation analysis of kidney biopsies from expanded criteria donors in long-term outcome. Transplantation. 2011;91(4):432-9.
14. Remuzzi G, Grinyò J, Ruggenenti P, Beatini M, Cole EH, Milford EL, et al. Early experience with dual kidney transplantation in adults using expanded donor criteria. Double Kidney Transplant Group (DKG). J Am Soc Nephrol. 1999;10(12):2591-8.
15. Remuzzi G, Cravedi P, Perna A, Dimitrov BD, Turturro M, Locatelli G, et al. Long-term outcome of renal transplantation from older donors. N Engl J Med. 2006;354(4):343-52.
16. Fogo AB, Bostad L, Svarstad E, Cook WJ, Moll S, Barbey F, et al. Scoring system for renal pathology in Fabry disease: report of the International Study Group of Fabry Nephropathy (ISGFN). Nephrol Dial Transplant. 2010;25(7):2168-77.
17. L'Imperio V, Smith A, Pisani A, D'Armiento M, Scollo V, Casano S, et al. MALDI imaging in Fabry nephropathy: a multicenter study. J Nephrol. 2020;33(2):299-306.

18. Snoeijs MG, Boonstra LA, Buurman WA, Goldschmeding R, van Suylen RJ, van Heurn LW, et al. Histological assessment of pre-transplant kidney biopsies is reproducible and representative. Histopathology. 2010;56(2):198-202.
19. Munivenkatappa RB, Schweitzer EJ, Papadimitriou JC, Drachenberg CB, Thom KA, Perencevich EN, et al. The Maryland aggregate pathology index: a deceased donor kidney biopsy scoring system for predicting graft failure. Am J Transplant. 2008;8(11):2316-24.
20. Kawamoto M, Yamada SI, Gibo T, Kajihara R, Nagashio S, Tanaka H, et al. Relationship between dry mouth and hypertension. Clin Oral Investig. 2021;25(9):5217-25.
21. Zee J, Liu Q, Smith AR, Hodgin JB, Rosenberg A, Gillespie BW, et al. Kidney Biopsy Features Most Predictive of Clinical Outcomes in the Spectrum of Minimal Change Disease and Focal Segmental Glomerulosclerosis. J Am Soc Nephrol. 2022;33(7):1411-26.
22. Hoshino J, Mise K, Ueno T, Imafuku A, Kawada M, Sumida K, et al. A pathological scoring system to predict renal outcome in diabetic nephropathy. Am J Nephrol. 2015;41(4-5):337-44.
23. Tervaert TW, Mooyaart AL, Amann K, Cohen AH, Cook HT, Drachenberg CB, et al. Pathologic classification of diabetic nephropathy. J Am Soc Nephrol. 2010;21(4):556-63.
24. Joh K, Muso E, Shigematsu H, Nose M, Nagata M, Arimura Y, et al. Renal pathology of ANCA-related vasculitis: proposal for standardization of pathological diagnosis in Japan. Clin Exp Nephrol. 2008;12(4):277-91.
25. Jiang L, Liu G, Lv J, Huang C, Chen B, Wang S, et al. Concise semiquantitative histological scoring system for immunoglobulin A nephropathy. Nephrology (Carlton). 2009;14(6):597-605.
26. Roberts IS, Cook HT, Troyanov S, Alpers CE, Amore A, Barratt J, et al. The Oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. Kidney Int. 2009;76(5):546-56.
27. Zhang Y, Sun L, Zhou S, Xu Q, Xu Q, Liu D, et al. Intrarenal Arterial Lesions Are Associated with Higher Blood Pressure, Reduced Renal Function and Poorer Renal Outcomes in Patients with IgA Nephropathy. Kidney Blood Press Res. 2018;43(2):639-50.
28. Zhang Y, Yang C, Zhou X, Hu R, Quan S, Zhou Y, et al. Association between thrombotic microangiopathy and activated alternative complement pathway in malignant nephrosclerosis. Nephrol Dial Transplant. 2020.
29. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017;7(1):16878.
30. Krippendorff K. Estimating the Reliability, Systematic Error and Random Error of Interval Data. Educational and Psychological Measurement. 1970;30(1):61-70.
31. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69-71.
32. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018;18(3):91-3.