GeroScience (2019) 41:455–465
https://doi.org/10.1007/s11357-019-00088-w

ORIGINAL ARTICLE

Validation of a geropathology grading system for aging mouse studies

Jessica M. Snyder · Timothy A. Snider · Marcia A. Ciol · John E. Wilkinson · Denise M. Imai · Kerriann M. Casey · Jose G. Vilches-Moure · Christina Pettan-Brewer · Smitha P. S. Pillai · Sebastian E. Carrasco · Shabnam Salimi · Warren Ladiges

Received: 24 April 2019 / Accepted: 31 July 2019 / Published online: 29 August 2019

Abstract An understanding of early-onset mechanisms underlying age-related changes can be obtained by evaluating changes that precede frailty and end of life using histological characterization of age-related lesions. Histopathology-based information as a component of aging studies in mice can complement and add context to molecular, cellular, and physiologic data, but there is a lack of information regarding scoring criteria and lesion grading guidelines. This report describes the validation of a grading system, designated as the geropathology grading platform (GGP), which generated a composite lesion score (CLS) for comparison of histological lesion scores in tissues from aging mice. To assess reproducibility of the scoring system, multiple veterinary pathologists independently scored the same slides from the heart, lung, liver, and kidney from two different strains (C57BL/6 and CB6F1) of male mice at 8, 16, 24, and 32 months of age. There was moderate to high agreement between pathologists, particularly when agreement within a 1-point range was considered. CLS for all organs was significantly higher in older versus younger mice, suggesting that the GGP was reliable for detecting age-related pathology in mice. The overall results suggest that the GGP guidelines reliably distinguish between younger and older mice and may therefore be accurate in distinguishing between experimental groups of mice with more, or less, age-related pathology.

Jessica M. Snyder and Timothy A. Snider contributed equally to this work.

J. M. Snyder · C. Pettan-Brewer · W. Ladiges (✉)
Department of Comparative Medicine, University of Washington, Seattle, WA, USA
e-mail: wladiges@u.washington.edu

T. A. Snider
Department of Veterinary Pathobiology, Oklahoma State University, Stillwater, OK, USA

M. A. Ciol
Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA

J. E. Wilkinson
Department of Pathology, University of Michigan, Ann Arbor, MI, USA

D. M. Imai
Department of Veterinary Pathology, UC Davis, Davis, CA, USA

K. M. Casey · J. G. Vilches-Moure
Department of Comparative Medicine, Stanford University School of Medicine, Stanford, CA, USA

S. P. S. Pillai
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

S. E. Carrasco
Division of Comparative Medicine, Massachusetts Institute of Technology, Cambridge, MA, USA

S. Salimi
School of Medicine, University of Maryland, College Park, MD, USA

Keywords Geropathology · Aging · Mouse · Histopathology · Lesion grades · Geroscience

Introduction

Aging is the process of growing old. It is characterized by a progressive decline of various physiological functions in tissues and organs and is often associated with numerous neoplastic and chronic degenerative, non-neoplastic disorders. Why these conditions develop, and how they progress with increasing severity with increasing age, is not well understood. While clinical studies are highly informative relative to human health, they are time intensive and generally limited in depth and scope for understanding underlying basic mechanistic causes and associations. Extensive preclinical studies using animal models are being conducted that provide meaningful insight into molecular and cellular pathways mediating pathophysiology of aging and the development of parameters that can distinguish between healthy aging and unhealthy aging. The laboratory mouse is extensively used as a preclinical animal model for aging research (Nadon 2007; Brayton et al. 2012). Accumulated evidence from aging studies has solidly established it as a robust model of human aging (Sundberg et al. 2011; Yuan et al. 2011). Mouse models have been instrumental in uncovering basic mechanisms involved in aging processes as well as in extending understanding of disease-associated phenotypes in the human population (Vanhooren and Libert 2013; Koks et al. 2016). Mice have short lifespans and are economical to maintain for longitudinal studies and share many physiological and genetic attributes with humans (Sundberg et al. 2011; Ray et al. 2010; Yuan et al. 2011). In addition, large numbers of strain-specific mice can easily be generated for adequate cohort numbers to achieve statistical power in genetic- or pharmacologic-based aging intervention studies.

The traditional assessment of age-related mouse studies has generally consisted of molecular, physiological, and clinical phenotypes such as gene expression, lifespan, and frailty (Burch et al. 2014; Treuting et al. 2016; Kane et al. 2015; Ladiges et al. 2009). These endpoints have been very helpful in identifying gene-driven pathways and potential pharmacologic targets in late life conditions. However, to understand early-onset mechanisms underlying age-related changes, it is necessary to evaluate changes that precede frailty and end of life, such as response to physiological stress or histological characterization of age-related lesions. There is evidence to suggest that histological lesions may be detected before clinical phenotypes are apparent (Adissu et al. 2014). Histopathology-based information can therefore be extremely valuable. Pathology, as a component of aging studies in mice, complements and adds context to other molecular, cellular, and physiologic data (Ikeno et al. 2003; Treuting et al. 2016; Wilkinson et al. 2012). Unfortunately, even in studies with histopathological examination and the reporting of age-related lesion scores, a lack of information regarding which parameters are scored and detailed descriptions of the scoring criteria may impair critical evaluation of the pathologic results. This, coupled with differences in tissue and lesion evaluation, may also complicate comparison of results from study to study (Neff et al. 2013; Wilkinson et al. 2012) and deter more extensive pathology-based investigations.

In 2015, the National Institute on Aging funded the Geropathology Initiative (R24 AG047115, PI Ladiges) designed to enhance the integration of pathology into preclinical aging studies by providing an environment to promote learning and exchange of scientific information and ideas for the aging research community with an interest in pathological analysis through a series of symposia and network conferencing formats. The term "geropathology" was used to designate the study of aging and age-related lesions and diseases in the form of whole necropsies/autopsies, surgical biopsies, histology, and molecular biomarkers encompassing multiple subspecialties including geriatrics, anatomic pathology, molecular pathology, clinical pathology, and gerontology. An Anatomic Working Group was established to develop uniform ways of integrating pathology into mouse lifespan and healthy aging studies, for example, by providing consensus recommendations for standardizing the histological grading of lesions and performing statistical analyses designed to integrate pathology data with longitudinal and cross-sectional lifespan data and physiological function data for more relevant translation to human studies. The working group quickly transitioned into an active Geropathology Grading Committee (GGC) with the objective of developing pathology endpoints that could provide reliable and responsive readouts for aging processes and interventions targeting biology of aging, i.e., a pathology-based surrogate of aging, spanning young to old, using the mouse as a prototype animal model of aging.

The GGC considered that the ideal pathology assessment plan for aging mouse studies would provide the ability to comprehensively and efficiently detect and grade standard lesions in organs in an age-dependent manner, and generate a numerical index that would capture clinical as well as subclinical alterations at the tissue level. This index could then be used as a tool to compare animals in the same cohort and across various cohorts by tabulating composite lesion scores (CLS). Coupled with physiological and clinical pathology data, the pathology-based index could be a robust means of evaluating aging and aging intervention cohorts. This report describes the validation of guidelines for a scoring system to evaluate a series of organs in aging mice with the goal of assigning a numerical score representing the degree of organ age-related pathology, using two mouse strains (C57BL/6 and CB6F1) each at four different ages and evaluated by multiple pathologists.

Materials and methods

Histological grading system

A histological grading system, designated as the Geropathology Grading Platform (GGP) and developed by the Geropathology Grading Committee (GGC), was used to evaluate and score target organs from the two different mouse strains and four different age groups (described below). Organ-specific lesions selected for inclusion within the grading platform were based on the combined experience of pathologists within the grading committee as well as documented lesions that had been reported to naturally develop in mice as a function of age (Berridge et al. 2016; Frazier et al. 2012; Thoolen et al. 2010; Renne et al. 2009). Then, the GGC developed guidelines based on the intent to (1) detect the histological presence or absence of uncommon but potentially severe lesions and (2) determine the level of severity of common age-related lesions. Specific lesions were graded with a numerical score, with 0–1 representing presence or absence of a lesion, and 0–4 representing the increasing severity of a lesion (0 = none, 1 = minimal, 2 = mild, 3 = moderate, 4 = severe). It was then possible to add the individual lesion scores for that organ from each mouse to generate a composite lesion score (CLS). Neoplasms were graded separately. The proportion of mice with benign and malignant neoplasms and the proportion with specific neoplasms were calculated for each group. The presence of a benign neoplasm in all organs received a score of 1, and presence of extensive multifocal benign or malignant neoplasm(s) received a score of 2. In a few mice, not all anatomic structures (e.g., heart valves) or slides were available to score and thus, not all possible specific lesions could be assessed. Thus, to calculate a score for every mouse, we took the average score of the observed lesions for each mouse and called it standardized CLS. This approach assumed that the missing specific scores would have the same value as the average of the observed scores and avoids that a mouse missing a specific slide or anatomic structure would have lower scores than a mouse with all slides available, simply because of the missing data. In addition, for each mouse, we calculated the mean of the two or three standardized CLS (given by different pathologists), denominated averaged standardized CLS score.

This approach allowed multiple pathologists to read the same slides using a standardized grading system. In order to validate the GGP, two to three pathologists reviewed the same slide set for at least one of the four organs: the liver; heart; lungs; and kidney. In a blinded fashion, organ-specific lesions were either graded as present/absent for a score of 0 or 1 or assigned a severity score from 0 to 4. Neoplasms were scored from 0 to 2. Lesions for the heart that were scored as present (1) or absent (0) included atrial thrombosis, which represented a rare pathologic finding. Other lesions were scored by severity from 0 to 4, and included arteriosclerosis, cardiomyopathy/myocardial fibrosis, myocardial inflammation, myxomatous change of the valve(s), and lymphoid aggregates. Lesions for the lung scored as present (1) or absent (0) included airway metaplasia or hyperplasia, vascular hypertrophy, atelectasis, and pulmonary fibrosis. Lung lesions scored by severity from 0 to 4 included eosinophilic crystalline (acidophilic alveolar macrophage) pneumonia, alveolar histiocytosis, alveolar foam cells, heart failure cells (chronic passive congestion), interstitial pneumonia/pneumonitis, perivascular inflammation, bronchial/bronchiolar inflammation, and lymphoid aggregates (peribronchiolar, perivascular, and/or pleural/subpleural). Lesions for the liver scored as present (1) or absent (0) included central venous congestion (chronic passive congestion), Ito cell hyperplasia/lipidosis, and telangiectasia/angiectasis. Liver lesions scored by severity from 0 to 4 included hepatocellular degeneration/necrosis, hepatic lipidosis, periportal inflammation, bile duct hyperplasia/cysts, lymphoid aggregates, and microgranulomas. Lesions for the kidney scored as present (1) or absent (0) included infarction, mineralization, and amyloidosis. Lesions scored by severity from 0 to 4 included nephropathy, pyelonephritis/nephritis, and lymphoid aggregates.

Source of mouse tissues

Paraffin-embedded blocks and hematoxylin and eosin (HE)–stained glass slides of tissues from C57BL/6 to CB6F1 (BALB/cBy × C57BL/6) male mice were obtained from the Geropathology Rodent Tissue Bank at the University of Washington. The mice were wild type, from the National Institute on Aging (NIA) contract facility (Charles River) in age groups of 4, 12, 20, and 28 months, and originally used in a 4-month physiological assessment study. During this period, mice were housed at the University of Washington under a 12:12 h light:dark cycle in individually ventilated cages (Allentown, Allentown, NJ) containing corncob bedding (Andersons, Maumee, OH) and Nestlets nesting material, and fed irradiated rodent chow (Rodent Diet, Lab Diet, St. Louis, MO) with autoclaved, acidified (pH 2.4–2.8) water. Physiologic assessments in this original study were performed in succession over the 4-month period and included 3-day wheel running, echocardiography, rotarod, open field activity, cognitive radial water tread maze, grip strength, indirect calorimetry, corneal opacity, and a 2-week tumor response procedure following subcutaneous injection of B16F0 melanoma cells (ATCC) 2 weeks prior to euthanasia (Pettan-Brewer et al. 2012). After 4 months and at the time of euthanasia by CO_2, the cohort ages were 8, 16, 24, and 32 months. Tissues were collected, weighed, and placed in 10% neutral buffered formalin. Samples were routinely processed, paraffin embedded, sectioned at 4–5 μm thickness, stained with HE, and deposited in the Geropathology Rodent Tissue Bank. Blocks and HE-stained slides were randomly selected from the tissue bank for this validation study such that each strain and age cohort represented 12 mice.

Statistical analysis

The aim of this study is to show that the GGP provides CLS, standardized CLS, and averaged standardized CLS that are valid (measure age-related lesions as desired) and reliable. To accomplish that, the proportion of agreement between two or three pathologists and the proportions of higher or lower scores by the type of lesion within an organ were calculated to assess agreement or reliability in scoring between independent pathology readers (Fayers and Machin 2007). The results for two or three pathologists were compared using the proportion of agreement on the exact value of CLS and on the CLS within one point of difference. Pathologic lesions occur on a continuum, so differences in scores are inevitable with borderline lesions. Medians of CLS between two readers were compared using a Wilcoxon signed rank test for paired data. Finding statistically significant median differences would be evidence that agreement between the raters has not been yet achieved.

Another form of instrument validation is the concept of known-groups (Fayers and Machin 2007). For this validation, we compared groups of mice that might, in principle and from what is known in the literature, produce different lesion score values according to their age. The expectation was that scores from the grading system would be larger for older mice, as a function of developing more lesions with increasing age. For this validation, we used the averaged standardized CLS scores (mean of the two or three pathologist scores) as the final score for a mouse and compared the scores by age and strain through visual display and a two-way analysis of variance (ANOVA) including an interaction of age by strain, with post hoc comparisons using Scheffe's method. When the interaction was not statistically significant, it was dropped of the final model. Here, finding statistical significance for age or strain would be evidence that the averaged standardized score is detecting differences in strain or age groups, which is an instrument characteristic that is desirable.

Significance level was kept to 0.05 for all statistical tests, since this is an exploratory validation study. Analyses were performed in SPSS version 25 for Mac and figures were produced using RStudio version 1.1.383.

Results

Agreement of lesion scores varied by organ

Composite lesion score (CLS) agreement between two or three pathologists varied from organ to organ. CLS was standardized by dividing it by the number of scored lesions to account for missing scores for certain lesions and to allow for comparison of scores between the organs. When three instead of two pathologists were assessing the same slide, comparisons were made by two pathologists at a time. The agreement did vary not only by organ using standardized CLS but also when individual lesion scores were compared among pathology readers, agreement varied by lesion.

The CLS for the heart scored by two independent pathologists showed exact agreement 54% of the time (Table 1A). However, there was some variation for the scoring of individual lesions. Using arteriosclerosis as a lesion example, there were 96 slides with a possible range of scores from 0 to 4, but only scores of 0 to 2 were observed in the sample of slides. Both pathologists agreed 66% of the time, while reader 1 scored a higher value than reader 2 in 33% of the slides and reader 2 scored higher in 1% of the slides. Even though the pathologists agreed only 54% of the time on the exact total CLS, the proportion increased to 91% when considering agreement within 1 point of difference. When comparing the CLS for the two pathologists, the test was statistically significant ($p < 0.001$), signaling that the median difference between the two readers was not zero.

Table 1 The range of observed composite lesion scores (CLS) and proportion of agreement for two independent pathology readers are shown for the (A) heart and (B) lungs. The total number of slides read for each lesion was 97 except for valvular myxomatous change because, based on section-to-section variability, valves were not present in all sections. The reader agreement columns provide insight into the reader-dependent aspects of the data and how each reader is scoring a specific lesion compared with the second reader starting with agreement followed by how often reader 1 scored higher or lower than reader 2. The greater the difference, the more likely there is a need to further adjust the lesion guidelines

A. Heart lesions	Potential range	Observed range	Proportion of		
			Agreement	Reader 1 > Reader 2	Reader 1 < Reader 2
Arteriosclerosis	0–4	0–2	0.66	0.33	0.01
Cardiomyopathy/myocardial fibrosis	0–4	0–2	0.91	0.06	0.03
Myocardial inflammation	0–4	0–1	0.96	0.03	0.01
Valvular myxomatosis[†]	0–4	0–2	0.89	0.05	0.05
Lymphoid aggregates	0–4	0–1	0.92	0.03	0
Atrial thrombosis	0–1	0	1	0	0
Tumor	0–2	0–2	0.99	0	0.01
CLS	0–23	0–7	0.54	0.42	0.04
CLS, 1 point*			0.91	0.07	0.02
B. Lung lesions					
Alveolar acidophilic macrophage pneumonia	0–4	0–3	0.88	0.01	0.11
Alveolar histiocytosis	0–4	0–2	0.59	0.20	0.21
Foam cells	0–4	0–1	0.97	0.01	0.02
Heart failure cells (chronic passive congestion)	0–4	0–3	0.99	0	0.01
Interstitial pneumonia	0–4	0–2	0.66	0.14	0.21
Perivascular inflammation	0–4	0–4	0.46	0.04	0.50
Bronchial/bronchiolar inflammation	0–4	0–2	0.64	0.01	0.35
Airway metaplasia or hyperplasia	0–1	0–1	0.88	0.01	0.10
Vascular hypertrophy	0–1	0–1	0.99	0	0.01
Pulmonary fibrosis	0–1	0–1	0.96	0	0.04
Atelectasis	0–1	0	1	0	0
Lymphoid aggregates	0–4	0–4	0.45	0.49	0.06
Tumor	0–2	0–2	0.90	0.05	0.05
CLS	0–39	0–12	0.26	0.43	0.31
CLS, 1 point*			0.65	0.15	0.21

*Starting with agreement followed by how often reader 1 scored higher or lower than reader 2

† stand for "greater than" and "less than" depending on the point of direction

For the lungs, lesions scored by two independent pathologists showed exact agreement 26% of the time (Table 1B). This agreement increased to 65% within one point of difference. When comparing the CLS for the two pathologists, the test was statistically significant ($p < 0.001$), signaling that the median difference between the two readers was not zero. Tumor score agreement between the two pathologists was high at 90%.

When three pathologists independently scored liver lesions, there was exact agreement only 5% of the time, but this increased to 13, 19, and 36% when comparing scores by any two pathologists (Table 2A). Agreement between any two pathologists varied from 42, 49, and 71% when considering scores within 1 point. All pairwise comparisons of CLS resulted in statistically significant difference between two pathologists ($p < 0.001$ for all). The three pathologists had a good overall agreement in tumor score (92%).

Results for lesion assessment of kidney tissue scored by three independent pathologists showed that agreement was mostly moderate to high for specific lesions, but exact agreement by all three on the CLS was only 29% among the three readers. Exact agreement between any two pairs of pathologists varied from 43, 44, and

Table 2 The range of observed CLS and proportion of agreement for three independent pathology readers are shown for the (A) liver and (B) kidney. The total number of slides read for each lesion was 93–96. For agreement, and two-by-two comparisons, the order of presentation is Readers 1 and 2, Readers 1 and 3, and Readers 2 and 3. The reader agreement columns provide insight into the reader-dependent aspects of the data and how each reader is scoring a specific lesion compared to the second reader or third reader

A. Liver lesions	Potential range	Observed range	Proportion of		
			Agreement*	Reader 1 > 2 Reader 1 > 3 Reader 2 > 3	Reader 1 < 2 Reader 1 < 3 Reader 2 < 3
Hepatic degeneration/necrosis	0–4	0–4	0.78/0.75/0.89	0.20/0.20/0.04	0.02/0.05/0.05
Hepatic lipidosis	0–4	0–4	0.47/0.57/0.59	0.48/0.36/0.09	0.04/0.04/0.32
Central venous congestion	0–1	0	1 for all	0 for all	0 for all
Periportal inflammation	0–4	0–4	0.48/0.54/0.87	0.52/0.42/0	0./0.04/0.13
Bile duct hyperplasia/cysts	0–4	0–4	0.32/0.33/0.81	0.65/0.62/0.13	0.03/0.05/0.07
Lymphoid aggregates	0–4	0–4	0.62/0.64/0.75	0.13/0.15/0.16	0.25/0.21/0.19
Microgranuloma	0–4	0–4	0.56/0.54/0.68	0.20/0.0.03/0.00.03/0.03/0.03	0.24/0.33/0.25
Ito cell hyperplasia	0–1	0–1	0.89/0.78/0.82	0/0/0.01	0.07/0.19/0.15
Telangiectasia	0–1	0–1	0.97/0.97/0.98	0.02/0.04/0.04	0.03/0.03/0.01
Tumor	0–2	0–2	0.96/0.95/0.95		0.02/0.01/0.01
CLS	0–30	0–9/0–21	0.13/0.19/0.36	0.76/0.64/0.18	0.12/0.17/0.46
CLS, 1 point**			0.42/0.49/0.71	0.53/0.40/0.06	0.05/0.11/0.23
B. Kidney lesions					
Nephropathy	0–4	0–4	0.52/0.44/0.56	0.14/0.35/0.40	0.35/0.21/0.04
Pyelonephritis	0–4	0–3	0.98/0.97/0.99	0.02/0.02/0	0/0.01/0.01
Infarct	0–1	0–1	0.95/0.98/0.95	0.04/0.01/0.01	0.01/0.01/0.04
Lymphoid aggregates	0–4	0–4	0.69/0.62/0.57	0.09/0.02/0.09	0.22/0.35/0.33
Mineralization	0–1	0–1	0.81/0.76/0.76	0.08/0.22/0.23	0.10/0.02/0.01
Amyloid	0–1	0–1	0.98/0.98/1	0.02/0.02/0	0/0/0
Tumor	0–2	0–2	0.98/0.84/0.84	0/0/0.01	0.02/0.16/0.15
CLS	0–17	0–10/0–7	0.43/0.44/0.47	0.16/0.22/0.31	0.25/0.24/0.22
CLS, 1 point**			0.77/0.85/0.88	0.06/0.04/0.08	0.17/0.10/0.04

*Proportions comparing two readers are presented in the following order: Readers 1 and 2, Readers 1 and 3, and Readers 2 and 3

**Proportion of agreement within one point of difference. Proportion of agreement within one point was not calculated for all three readers at once

47%, while agreement within 1 point varied from 77, 85, and 88% (Table 2B). Pairwise comparisons of the CLS yielded p values of 0.003, 0.05, and 0.16, signaling that some pathologists agreed among themselves but not with the third pathologist.

The scoring system distinguished age differences in multiple organs but strain differences only in the kidney

When standardized CLS was averaged between pathologists, mostly significant increases in scores were seen with increasing age in all the four organs (Fig. 1) showing the expected pattern for known-groups validation analysis. For this analysis, the averaged standardized CLS from either two or three pathologists was the response variable in the ANOVA. For all models, there was no statistically significant interaction between age and strain. Specifically, for the heart (Fig. 1A), post hoc multiple comparisons showed differences between ages 8 and 24 and 32 months, between 16 and 24 and 32 months, and between 24 and 32 months ($p \leq 0.04$ for all). For the lungs (Fig. 1B), post hoc multiple comparisons showed differences between ages 8 and 16, between 24 and 32 months (all $p \leq 0.05$), between 16 and 32 months ($p < 0.001$), and 24 and 32 months ($p = 0.008$). For the liver (Fig. 1C), post hoc multiple comparisons showed differences between ages 8 and 24 and 32 months ($p < 0.001$ for both), between 16 and 24 and 32 months ($p < 0.001$ for both). For the kidney (Fig. 1D), post hoc multiple comparisons showed differences between all age groups ($p < 0.03$ for all), except between ages 8 and

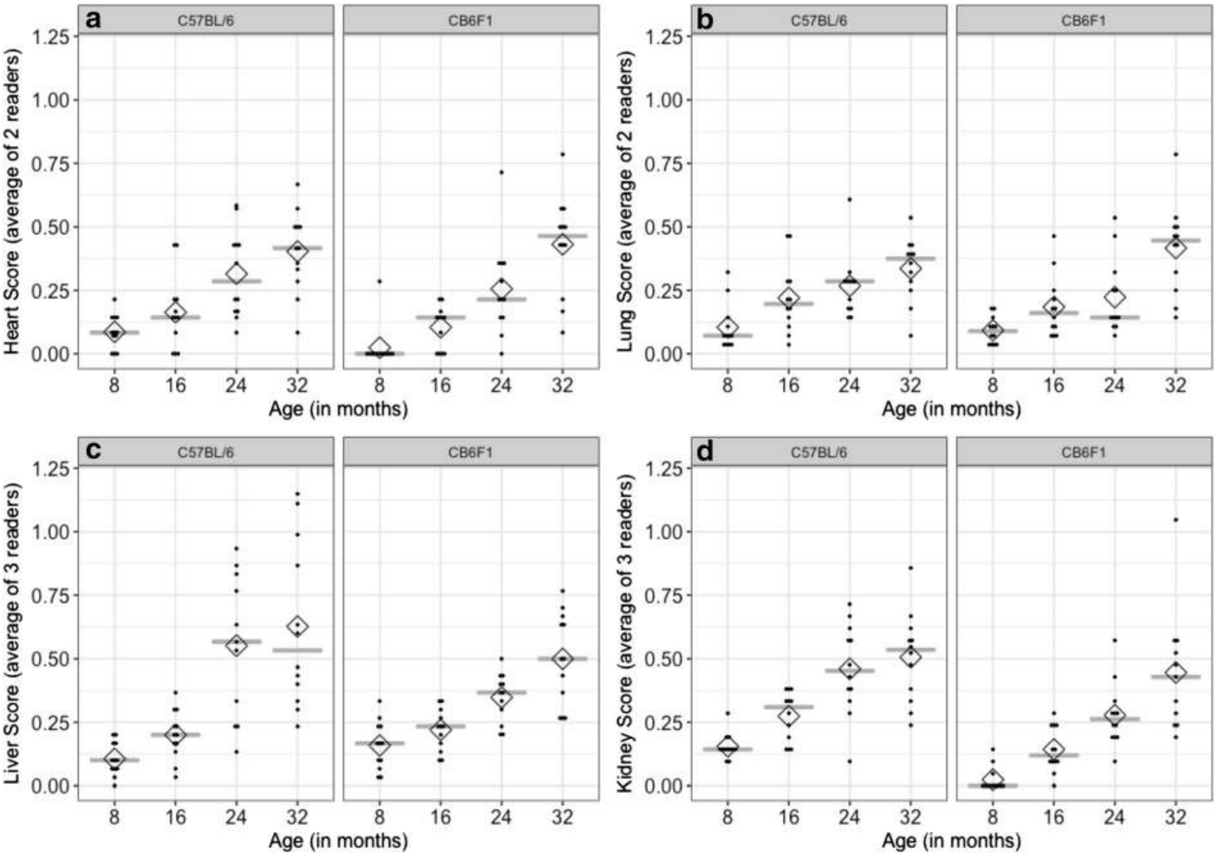

Fig. 1 Distribution of averaged standardized CLS in C57BL/6 and CB6F1 mice stratified by age group shows that lesion grading can distinguish between older and younger age groups in the **a** Heart; **b** Lungs; **c** Liver; and **d** Kidney. The "score" is the averaged standardized CLS from each mouse. Each mouse's score is depicted by a dot. The gray line depicts the median score and the diamond depicts the mean score in each age group and mouse strain. Increasing values in median and mean by age group can be observed, while the two strain groups (side-by-side plots) show little difference in medians, except for Kidney scores. Lesion gradings were different between the two strains in the kidney **(d),** but not the other three organs

16 months ($p = 0.09$). Clearly, there was an increase in the median and mean scores with increasing age for all the four organs from both mouse strains. These results show that the averaged standardized CLS can find differences between mean scores according to age groups.

Significant differences in lesion scores between tissues from the two strains were seen in the kidney as expected, but not in any of the other three organs, when standardized CLS was averaged between pathologists (Fig. 1). Kidneys from the C57BL/6 strain had averages and medians significantly higher than kidneys from the CB6F1 strain ($p = 0.001$). These results show that the averaged standardized CLS can find differences in the studied strains for kidney lesions.

Discussion

This report describes a grading system designated as the geropathology grading platform (GGP) and shows that it can be useful to distinguish age-related differences regarding the absence or presence and severity of specific histological lesions. The GGP consists of guidelines for a scoring system to evaluate organs from aging mice with the goal of assigning a numerical score representing the presence and degree of organ age-related pathology. The data provide validation of the GGP using the heart, lung, liver, and kidney from two different strains (C57BL/6 and CB6F1) of male mice at 8, 16, 24, and 32 months of age. The scoring system includes not only lesions that are suspected to have a negative effect on systemic health and homeostasis (clinically significant lesions) but also lesions that have no known negative effect but occur or increase with age (incidental age-related lesions).

Scoring systems for histological lesions need to be reliable and reproducible in order to be valid (Ward and Thoolen 2011). In our study, averaged standardized composite lesion scores (CLS) for all organs were significantly higher in older versus younger mice, suggesting that the GGP was reliable for detecting age-related pathology in mice. There was more variation in the scores of older-aged mice most likely as a natural consequence of age, but variability in slide reading may have also been a contributing factor. Nevertheless, different pathologists were independently able to assess reproducibility by scoring the same histology tissues with an overall favorable agreement between scores, particularly, when a 1-point range was considered.

Unexpectedly, pathologist agreement was better for some organs (heart, kidney) than others (lung, liver). This result may have been influenced by the number of parameters scored for each organ (Gibson-Corley et al. 2013). If too few parameters are scored, or if a smaller range of severity scores is used to distinguish between lesions of different severities, then, the scoring system may not efficiently distinguish between groups. If too many parameters are incorporated into the scoring system, exact agreement among scorers may be more difficult to achieve as differences between scorers for each parameter may be compounded by the number of parameters. Studies examining agreement among pathologists in scoring single variables show a range from 47 to 94% (Eaton et al. 2007; Koelink et al. 2018; Liang et al. 2014). In the GGP scoring system, which aims to score a variety of age-related lesions at the whole organ level, it is difficult to limit the number of parameters scored, and variation in the individual parameters within each organ is compounded when multiple parameters (up to 12 for lung) are included in the final composite lesion score. This may have contributed to the lower inter-reader agreement in exact score in some cases. Indeed, the organs with the least agreement (liver and lung) were those with the most parameters scored. Lesions were scored on a scale of 0–4 rather than 0–3 in order to differentiate between minimal and mild lesions. There may have been different interpretations among pathologists as to the difference between normal and minimal or between minimal and mild lesions. Further refinement of the descriptive terminology in the GGP guidelines will help to improve agreements in this area.

Assessment of the heart resulted in the highest degree of agreement between pathologists. Importantly, the presence or absence of severe and/or unusual lesions, such as atrial thrombosis and neoplasms, had near 100% agreement between pathologists. Arteriosclerosis was the parameter with the lowest agreement. However, scores for this parameter were generally low, and most of the pathologists' scores were within one point. The observed range for heart CLS (0–7) was far below the possible range (0–53) and therefore did not allow for validation of high values within the GGP. This also indicated that although there was a significant difference in CLS between young mice and old mice, there was little evidence of severe cardiac pathology as detected on HE slides at any age in either strain in this study.

For the lung, agreement between pathologists was high for some lesions, but not for others. The lowest agreements were for perivascular inflammation and lymphoid aggregates, which had less than 50% agreement, followed closely by perivascular inflammation. As lymphoid aggregates often occupied a perivascular space, some degree of descriptive overlap is present between lymphoid aggregates and perivascular inflammation, and this likely contributed to lower agreement in lesion scores between pathologists. As with the heart, the observed range (0–13) of the exact CLS for the lung was far below the possible range (0–45) and did not allow for validation of high values within the GGP. Alveolar acidophilic macrophage pneumonia was one of the parameters with more severe lesions observed in this population of older age mice, with scores of 0–3 reported on a scale of 0–4. There was 88% agreement between pathologists scoring this lung lesion.

For the liver, the three pathologists had greater than 90% agreement for tumors. Exact agreement between all three pathologists was less than 50% for hepatic lipidosis, periportal inflammation, and bile duct hyperplasia, while agreement between two of the readers was greater than 80% for several parameters. One pathologist was consistently in disagreement with the other two, who were closer to each other in scoring. As a result, the final CLS for the liver had a lower exact agreement between all three pathologists, although agreement improved when considering scores within 1 point of each other at 42 to 71%. The discrepancy in scoring may mean that either the definitions of the lesions were not clear or that the reading was too subjective to allow for better agreement. For example, the description of "hepatic lipidosis" could be further clarified as "microvesicular" or "macrovesicular", given the difference in underlying pathogenesis for these two lesions. The range of scores was higher for the liver compared with the heart and lung (scores of 0–37 possible; 0–21 observed).

For the kidney, agreements among three pathologists were generally high. However, there was less than 50% exact agreement for nephropathy and lymphoid aggregates. Here, agreement between two pathologists was not always the same across the specific lesions. Reader 1 sometimes agreed more with reader 2 and other times with reader 3, and sometimes readers 2 and 3 agreed more with each other than with reader 1. Consequently, there was low agreement on the exact CLS, but a higher agreement for CLS within 1-point difference of agreement. As with the liver, the range of observed scores was also higher for the kidney (0–29 possible; 0–14 observed). The difference in renal lesion scores between C57BL/6 and CB6F1 mice was an interesting observation; although, this study was not designed to investigate differences between the strains.

Certain lesions had high variability. Lymphoid aggregates, which were evaluated for all organs, are a good example. For the lung, liver, and kidney, agreement among all three pathologists in lymphoid aggregate scores ranged from 45 to 55%. In these three organs, lymphoid aggregates were commonly seen, with scores ranging from 0 to 4. The only organ for which there was good agreement in lymphoid aggregates was the heart, in which these were infrequently seen. This suggests that the definition for lymphoid aggregates should be refined.

The overall results of this study suggest that the GGP guidelines reliably distinguish between younger and older mice and may therefore be accurate in distinguishing between experimental groups of mice with more, or less, age-related pathology. This exciting but preliminary observation needs to be further validated by additional studies. For example, the available set of slides did not cover all possible values of the scoring system. The heart and lung had lower total scores, even in the oldest mice, and some lesions were not represented in any mice. The consequence is that when the specific lesion scores are summed up, they did not cover all the possibilities of the final CLS to test reliability in the presence of larger lesion burdens. Because of the small range of scores for the heart and lung, values for agreement within a 1-point range should be interpreted with caution. Additional studies are needed to address these issues.

To be useful in preclinical studies, scoring platforms must be sensitive enough to discern that changes in lesion scores in a treatment group are not due to reader limitations in the grading system but from the treatment itself. As shown in this study, numerous lesion scores were reader-dependent, particularly for organs such as the liver, so there is a need to fine-tune the grading guidelines to reduce that dependence. Definitions for lesion scores such as lymphoid aggregates, perivascular and periportal inflammation, and microgranulomas need to be modified to induce more uniformity among readers, and more training must be done. Geropathology workshops are conducted by the Geropathology Research Network to help serve this purpose (Ladiges et al. 2016). An example of future workshop theme topics includes "minimum clinical significant change", defined as the minimal change that is considered a real change instead of simply a random variation in the score.

There are several issues related to interpreting and implementing results from this study, for example, the lack of gender comparison. Lesion scores were assigned in organs from only male mice, because tissues from female mice of the same strain and age groups were not readily available. Therefore, studies are needed to score lesions in organs from aging female mice. The same platform will be used in order to validate the system for female mice, but the scores could likely vary from males. A second issue is that the mice from which the organs were collected were involved in an unrelated physiological assessment study over a 4-month time period. Whether any of these individually, or in combination with others including an invasive tumor inoculation of the last two weeks of life, had any effects on lesion scores is not known, but all mice in all cohorts had the same procedures performed so lesion scores should represent any effects across all ages and both strains. Even though the mice were not naïve, they most likely would be representative of certain types of cross-sectional drug studies, where mice are evaluated with physiological assessments at the end of the study before tissues are collected for histopathology. In this regard, extensive physiological data are available to correlate with lesion scores in the various cohorts (Ge et al. 2017). Finally, the histology platform reported in this study includes only four organs. Lesion scores from additional organs, such as the skeletal muscle, pancreas, the head and brain, and reproductive organs, could have an impact on increasing the robustness of the GGP. Work is ongoing to incorporate these into the GGP.

It must be emphasized that the GGP is a scoring system currently designed to assess the presence and severity of age-related lesions, especially in the context of aging intervention studies. Grading guidelines for longitudinal lifespan studies are more complex because mice die at different times making it challenging to tabulate any semblance of a composite lesion score. For cross-sectional studies, the GGP provides a reasonably quick and comprehensive screening approach which can be used as an endpoint for pharmacological response and also to highlight the need for more in-depth histologic evaluations. A full understanding of the pathogenesis of aging- and/or toxicology-mediated lesions would require more extensive pathologic investigation. There also may be situations where there is a need to customize the GGP to help address specific research objectives related to certain organs or lesions not adequately addressed in the GGP. It is also possible to customize the GGP for longitudinal lifespan studies, but extensive effort will be needed for a workable system.

Acknowledgements We thank John Morton, Kateryna Tonyuk, and Jo Jiang for technical assistance.

Funding information This work was financially supported by NIA grants AG047115 and AG013280.

References

Adissu HA, Estabel J, Sunter D, Tuck E, Hooks Y, Carragher DM, Clarke K, Karp NA, Project SMG, Newbigging S, Jones N, Morikawa L, White JK, McKerlie C (2014) Histopathology reveals correlative and unique phenotypes in a high-throughput mouse phenotyping screen. Dis Model Mech 7: 515–524. https://doi.org/10.1242/dmm.015263

Berridge BR, Mowat V, Nagai H, Nyska A, Okazaki Y, Clements PJ, Rinke M, Snyder PW, Boyle MC, Wells MY (2016) Non-proliferative and proliferative lesions of the cardiovascular system of the rat and mouse. J Toxicol Pathol 29:1S–47S. https://doi.org/10.1293/tox.29.3S-1

Brayton CF, Treuting PM, Ward JM (2012) Pathobiology of aging mice and GEM: background strains and experimental design. Vet Pathol 49:85–105. https://doi.org/10.1177/0300985811430696

Burch JB, Augustine AD, Frieden LA, Hadley E, Howcroft TK, Johnson R, Khalsa PS, Kohanski RA, Li XL, Macchiarini F, Niederehe G, Oh YS, Pawlyk AC, Rodriguez H, Rowland JH, Shen GL, Sierra F, Wise BC (2014) Advances in geroscience: impact on healthspan and chronic disease. J Gerontol A Biol Sci Med Sci 69(Suppl 1):S1–S3. https://doi.org/10.1093/gerona/glu041

Eaton KA, Danon SJ, Krakowka S, Weisbrode SE (2007) A reproducible scoring system for quantification of histologic lesions of inflammatory disease in mouse gastric epithelium. Comp Med 57:57–65

Fayers FM, Machin D (2007) Scores and measurements: validity, reliability, sensitivity. In: Fayers FM, Machin D (eds) Quality of life: the assessment, analysis and interpretation of patient-reported outcomes, 2nd edn. Wiley, Oxford, pp 77–108

Frazier KS, Seely JC, Hard GC, Betton G, Burnett R, Nakatsuji S, Nishikawa A, Durchfeld-Meyer B, Bube A (2012) Proliferative and nonproliferative lesions of the rat and mouse urinary system. Toxicol Pathol 40:14S–86S. https://doi.org/10.1177/0192623312438736

Ge X, Ciol MA, Pettan-Brewer C, Goh J, Rabinovitch P, Ladiges W (2017) Self-motivated and stress-response performance assays in mice are age-dependent. Exp Gerontol 91:1–4. https://doi.org/10.1016/j.exger.2017.02.001

Gibson-Corley KN, Olivier AK, Meyerholz DK (2013) Principles for valid histopathologic scoring in research. Vet Pathol 50: 1007–1015. https://doi.org/10.1177/0300985813485099

Ikeno Y, Bronson RT, Hubbard GB, Lee S, Bartke A (2003) Delayed occurrence of fatal neoplastic diseases in Ames dwarf mice: correlation to extended longevity. J Gerontol A Biol Sci Med Sci 58:291–296

Kane AE, Hilmer SN, Boyer D, Gavin K, Nines D, Howlett SE, de Cabo R, Mitchell SJ (2015) Impact of longevity interventions on a validated mouse clinical frailty index. J Gerontol A Biol Sci Med Sci 71:333–339. https://doi.org/10.1093/gerona/glu315

Koelink PJ, Wildenberg ME, Stitt LW, Feagan BG, Koldijk M, van 't Wout AB, Atreya R, Vieth M, Brandse JF, Duijst S, te Velde AA, D'Haens GRAM, Levesque BG, van den Brink GR (2018) Development of reliable, valid and responsive scoring systems for endoscopy and histology in animal models for inflammatory bowel disease. J Crohns Colitis 12:794–803. https://doi.org/10.1093/ecco-jcc/jjy035

Koks S, Dogan S, Tuna BG, Gonzalez-Navarro H, Potter P, Vandenbroucke RE (2016) Mouse models of ageing and their relevance to disease. Mech Ageing Dev 160:41–53. https://doi.org/10.1016/j.mad.2016.10.001

Ladiges W, Ikeno Y, Niedernhofer L, McIndoe RA, Ciol MA, Ritchey J, Liggitt D (2016) The geropathology research network: an interdisciplinary approach for integrating pathology into research on aging. J Gerontol A Biol Sci Med Sci 71:431–434. https://doi.org/10.1093/gerona/glv079

Ladiges W, Van Remmen H, Strong R, Ikeno Y, Treuting P, Rabinovitch P, Richardson A (2009) Lifespan extension in genetically modified mice. Aging Cell 8:346–352. https://doi.org/10.1111/j.1474-9726.2009.00491.x

Liang W, Menke AL, Driessen A, Koek GH, Lindeman JH, Stoop R, Havekes LM, Kleemann R, van den Hoek AM (2014) Establishment of a general NAFLD scoring system for rodent models and comparison to human liver pathology. PLoS One 9:e115922. https://doi.org/10.1371/journal.pone.0115922

Nadon NL (2007) Animal models in gerontology research. Int Rev Neurobiol 81:15–27. https://doi.org/10.1016/S0074-7742(06)81002-0

Neff F, Flores-Dominguez D, Ryan DP, Horsch M, Schroder S et al (2013) Rapamycin extends murine lifespan but has limited effects on aging. J Clin Invest 123:3272–3291. https://doi.org/10.1172/JCI67674

Pettan-Brewer C, Morton J, Coil R, Hopkins H, Fatemie S, Ladiges W (2012) B16 melanoma tumor growth is delayed in mice in an age-dependent manner. Pathobiol Aging Age Relat Dis 2. https://doi.org/10.3402/pba.v2i0.19182

Ray MA, Johnston NA, Verhulst S, Trammell RA, Toth LA (2010) Identification of markers for imminent death in mice used in longevity and aging research. J Am Assoc Lab Anim Sci 49: 282–288

Renne R, Brix A, Harkema J, Herbert R, Kittel B, Lewis D, March T, Nagano K, Pino M, Rittinghausen S, Rosenbruch M, Tellier P, Wohrmann T (2009) Proliferative and nonproliferative lesions of the rat and mouse respiratory tract. Toxicol Pathol 37:5S–73S. https://doi.org/10.1177/0192623309353423

Sundberg JP, Berndt A, Sundberg BA, Silva KA, Kennedy V, Bronson R, Yuan R, Paigen B, Harrison D, N. Schofield P (2011) The mouse as a model for understanding chronic diseases of aging: the histopathologic basis of aging in inbred mice. Pathobiol Aging Age Relat Dis 1. https://doi.org/10.3402/pba.v1i0.7179

Thoolen B, Maronpot RR, Harada T, Nyska A, Rousseaux C, Nolte T, Malarkey DE, Kaufmann W, Küttler K, Deschl U, Nakae D, Gregson R, Vinlove MP, Brix AE, Singh B, Belpoggi F, Ward JM (2010) Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. Toxicol Pathol 38:5S–81S. https://doi.org/10.1177/0192623310386499

Treuting PM, Snyder JM, Ikeno Y, Schofield PN, Ward JM, Sundberg JP (2016) The vital role of pathology in improving reproducibility and translational relevance of aging studies in rodents. Vet Pathol 53:244–249. https://doi.org/10.1177/0300985815620629

Vanhooren V, Libert C (2013) The mouse as a model organism in aging research: usefulness, pitfalls and possibilities. Ageing Res Rev 12:8–21. https://doi.org/10.1016/j.arr.2012.03.010

Ward JM, Thoolen B (2011) Grading of lesions. Toxicol Pathol 39: 745–746. https://doi.org/10.1177/0192623311408622

Wilkinson JE, Burmeister L, Brooks SV, Chan CC, Friedline S, Harrison DE, Hejtmancik JF, Nadon N, Strong R, Wood LK, Woodward MA, Miller RA (2012) Rapamycin slows aging in mice. Aging Cell 11:675–682. https://doi.org/10.1111/j.1474-9726.2012.00832.x

Yuan R, Peters LL, Paigen B (2011) Mice as a mammalian model for research on the genetics of aging. ILAR J 52:4–15