IIB OR NOT IIB: ENDOMETRIAL BIOPSY EVALUATION IN HORSES USING THE KENNEY-DOIG SCALE

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
In Partial Fulfillment of the Requirements
For the Degree of Master of Science
In the Veterinary Pathology Department
Western College of Veterinary Medicine
University of Saskatchewan
Saskatoon

By

AUGUSTA JANE WESTENDORF

PERMISSION TO USE

In presenting this thesis/dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

DISCLAIMER

The Survey Monkey and OlyVia pages in this thesis were exclusively created to meet the thesis and/or exhibition requirements for the degree of Master of Science at the University of Saskatchewan. Reference in this thesis/dissertation to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis/dissertation in whole or part should be addressed to:

Head of the Department of Veterinary Pathology
1622 Western College of Veterinary Medicine, 52 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5B4 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks to my supervisor, Dr. Bruce Wobeser, who made this project possible. His continued support during both my Doctor of Veterinary Medicine and Master of Science degrees, as well as through multiple challenges encountered during the global pandemic, has been and will always be deeply appreciated.

To all of the participating pathologists in this study, thank you for all the time and effort you put into grading these slides; I could not have completed this project without you. Thank you to my advisory committee members Drs. Tasha Epp, Claire Card, Andy Allen, and Elemir Simko for your input and invaluable advice. To Dr. Epp, thank you for all of the research and mentoring you put into the statistical analysis of this project. I deeply appreciate every hour you volunteered to help troubleshoot and assist me with my methods and R code.

To my cousin, Jen Lindquist, thank you for both the emotional and statistical support while juggling two young toddlers. Thank you to all of my parents for sitting through hours of online presentation practices, sending care packages, and letting me come home not once, but twice, during the pandemic. And to Elliot, thank you for always making sure I had a coffee on my desk and food on my plate.

Finally, thank you to both the Townsend Equine Health Research Fund and the Interprovincial Graduate Fellowship Fund for their financial support and encouragement regarding this research.

DEDICATION

This thesis is dedicated to my two younger sisters, Zoe and Josephine, who have always supported my goals and ambitions no matter what. They have each grown into amazing young women who continue to inspire me every day. I hope that some of my work serves to inspire them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ICC..... Intraclass correlation coefficient

LH..... Luteinizing hormone

PAS..... Periodic acid-schiff

PDS..... Prairie Diagnostic Services

PMN..... Polymorphonuclear neutrophils

PMU..... Pregnant mare urine

WCVM..... Western College of Veterinary Medicine

WCVM/PDS..... Western College of Veterinary Medicine and Prairie Diagnostic Services

CHAPTER 1. Introduction and Literature Review

1.1 The Consequences of Mare Infertility

Within the equine world, planned breeding is a necessary and widespread practice. Across Canada, mares are bred either following a successful competitive career, the end of recreational use, forced retirement due to injury, for the purpose of propagating a breed, for commercial purposes, or just for the simple pleasure of having a foal born on farm. Yet throughout history, increasing pressure for improvements in esthetics and sport performance have outweighed the selection for reproductive efficiency in horses, contributing to the ever-present issue of infertility and the 'problem mare' (Waelchli, 1990; Woodward et al., 2012).

The reproductive process from conception to parturition is a long time frame in the mare with the average gestation lasting anywhere from 330 to 342 days (Bukowski & Aiello, n.d.). During this period, there are several events that can result in reproductive failure: early embryonic death, fetal resorption, placentitis, mare reproductive loss syndrome, abortion due to infectious agents, twin pregnancies, and premature delivery complications like placental separation (Bosh et al., 2009; Shideler et al., 1982). It is widely accepted that horses in general have one of the lowest reproductive efficiencies of the large domestic species (Mahon & Cunningham, 1982). In unmanaged herds, such as the feral Sable Island herd in Canada, foaling rates for horses aged three years and older were found to be 62%, while feral herds in the United States had pregnancy rates between 57% to 81.4% (Lucas et al., 1991; Wolfe et al., 1989). Domestic Thoroughbred breeding farms in Kentucky have reported live foaling rates of 55.2% to 82.9%, while Thoroughbred farms in the United Kingdom have reported similar foaling rates between 62.8% to 81.8%, with younger mares having higher reproductive success in both groups (W. R. Allen et al., 2007; Bosh et al., 2009). Similar seasonal foaling rates of approximately 80% have also been reported in the New Zealand Thoroughbred industry, while Australian Thoroughbred and Standardbred industries have reported 68.8% and 68.9% early pregnancy rates per cycle, respectively (Hanlon et al., 2012; Nath et al., 2010). Regardless of management, equine

reproductive rates fall well below the average for beef cattle operations within the United States where 91.5% of all bred cows go on to carry a calf to term (Animal and Plant Health Inspection Service, 2009). There are many factors that influence this difference in reproductive efficiency between species. As previously mentioned, equine breeding has primarily been driven by aesthetic and athletic value, unlike the cattle industry where selection has heavily favoured the ability to produce a calf. The seasonal nature of their oestrus cycles and longer gestational length compared to other species also reduces their reproductive efficiency. Additionally, the overall age of mares is much higher than in other production animal industries where older breeding females are often culled from the herd, and older mares are known to suffer high rates of embryonic loss and lower overall foaling rates (Bosh et al., 2009; Scoggin, 2015).

There are high financial costs associated with an eleven-month gestation comprising daily fees for mare management and husbandry, seasonal breeding costs, and routine farrier and veterinary care. This amounts to an expensive investment with a very real risk of no viable return. Considerable pressure is placed on clinicians to provide their clients with a relatively accurate estimate of a mare's ability to produce a live and healthy foal, enabling them to make educated and economical decisions. A thorough reproductive exam has long been employed to evaluate mares during pre-purchase exams, breeding soundness exams, following dystocia, before reproductive treatment, and in any situation where idiopathic infertility is suspected (R. Kenney & Doig, 1986; Love, 2011). The success and prognostic reliability of these examinations directly relate to the clinician's ability to integrate their findings and interpret the results of various diagnostic testing used.

1.2 Diagnosis of Infertility and the Breeding Soundness Examination

A robust reproductive exam is comprised of evaluation of a thorough breeding and management history, full physical examination, and a handful of specific diagnostic procedures including uterine visualization, palpation, cytology, bacterial and/or fungal culture and biopsy (Assad & Pandey, 2015; Riddle et al., 2007; Sertich, n.d.). The exam is designed to identify factors that may negatively impact a mare's fertility, including anatomical or conformational

abnormalities, endometrial diseases, hormonal imbalances, ovarian dysfunction, and other reasons that are outside the scope of this literature review.

Following a thorough investigation of a mare's reproductive history and overall physical exam, including special attention to the reproductive system, three specific diagnostic tests are routinely performed to complete the reproductive evaluation: uterine cytology, culture, and endometrial biopsy. While a combined approach involving uterine cytology and bacterial and/or fungal culture can help rule out certain conditions like acute endometritis and infectious agents, neither test reflects the true condition of the endometrium, an essential aspect when investigating issues concerning fertility. Endometrial biopsy, commonly one of the last diagnostic steps in the reproductive exam, is considered the gold standard for evaluating the structure and overall health of the endometrium (R. Kenney & Doig, 1986; Love, 2011; Overbeck et al., 2011).

The endometrial biopsy procedure involves aseptically passing a biopsy instrument through the cervix and into the uterus where a hand trans-rectally guides the basket of the instrument to the location within the uterus where a biopsy is desired (Love, 2011). While it is currently accepted that a single biopsy from the base of either uterine horn is generally representative of the whole uterus, conflicting studies exist, and it is recommended that any focal abnormalities found on physical exam should be biopsied individually (R. Kenney & Doig, 1986; Love, 2011; Overbeck et al., 2011; Sikora et al., 2017). Once the biopsy sample is obtained, it is placed in fixative and submitted for histopathologic analysis.

1.3 Basic Histologic Structure of the 'Normal' Endometrial Biopsy

When analyzing an endometrial biopsy, several different histopathologic features are of particular interest (Figure 1.1). First, the normal structure of the endometrium is evaluated. Three functional divisions can be discerned on endometrial biopsy.

The luminal epithelium of the uterus is partially ciliated and can range from low cuboidal to tall columnar cells. This layer provides the surface where maternal and fetal tissues will interact during pregnancy, providing a platform for the placental development.

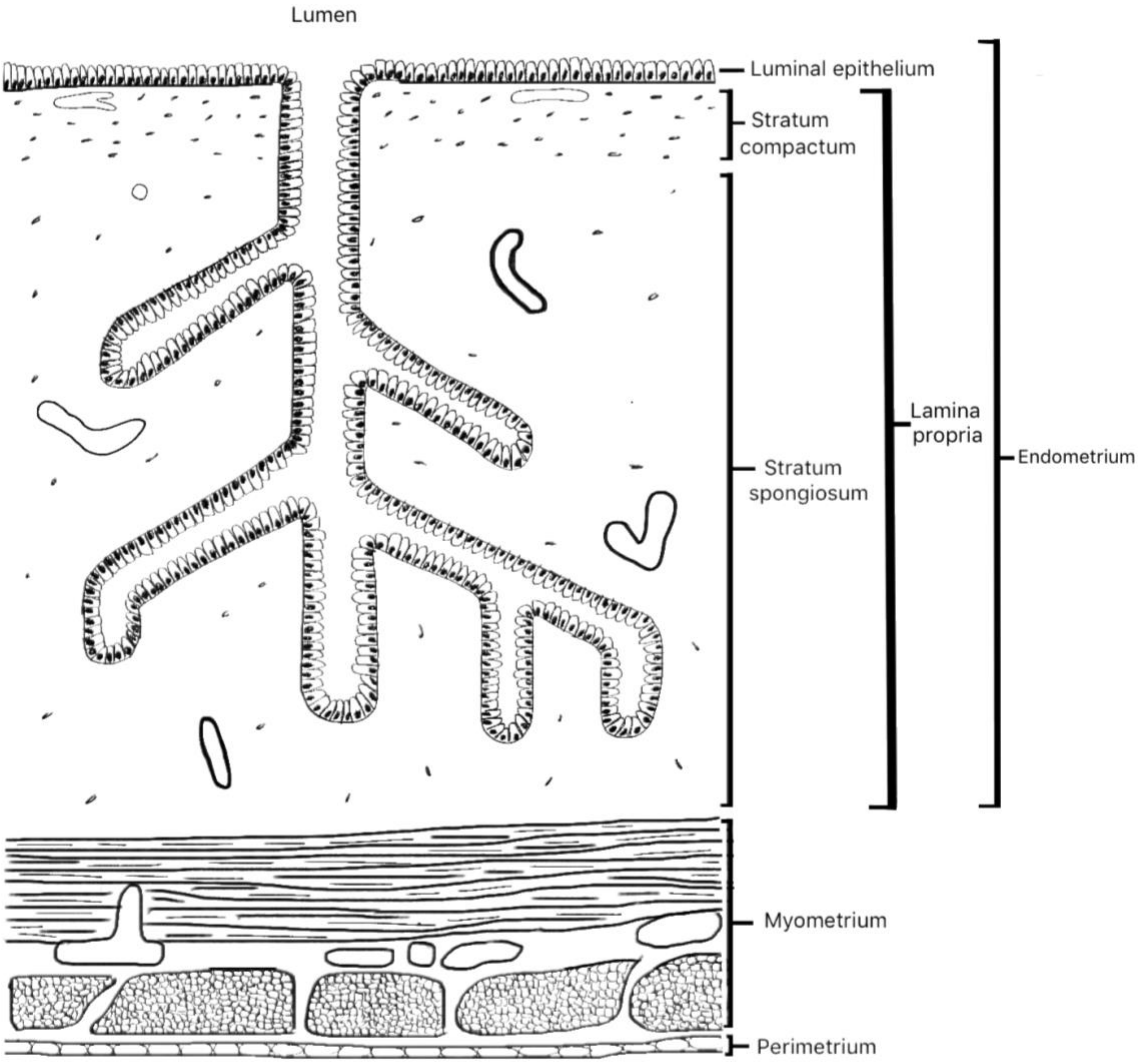

Figure 1.1 Schematic of a cross section of a normal equine endometrial biopsy adapted from Kenney, 1978b. (R. Kenney, 1978).

The lamina propria lies beneath the luminal epithelium and it can be further subdivided into the last two functional divisions of the endometrium. The stratum compactum is directly under the luminal epithelium and is separated from it by the epithelial basement membrane. Densely packed stromal cells comprise most of this layer, supporting the lumina of the endometrial glands that arise in the stratum spongiosum below and become continuous with the luminal epithelium above to empty into the uterine lumen. The stratum spongiosum is a gradual continuation of the compactum, but still visibly discernible; and is composed of loose connective tissue and sparse stromal cells surround tubular branched endometrial glands that grow and regress according to the mare's estrus cycle and seasonal effects. Glands appear mostly uniform in size and shape, with cuboidal to cylindrical epithelium containing basally located nuclei that may vary from round to ovoid in shape depending on the stage of the mare's cycle (Schöniger & Schoon, 2020). These glands are responsible for producing and secreting uterine fluid, also called uterine milk or histotroph. These secretions are rich in proteins and other nutrients that are essential to the early survival of the embryo during the pre-implantation period and before the placenta has developed. Blood and lymphatic vessels reside throughout the entire lamina propria with the larger vessels within the stratum spongiosum and smaller capillaries and venules located within the stratum compactum.

In an ideal endometrial biopsy, only the luminal epithelium and lamina propria of the endometrium should be present. A portion of myometrium may also be visible on occasion. The myometrium is composed of smooth muscle cells, nerves, and blood vessels and functions to aid in uterine clearance, blood supply, and drainage of lymph and blood from the uterus.

Adequate tissue must be present to be able to fully evaluate the luminal epithelium, stratum compactum, and stratum spongiosum. Observers may also be required to evaluate the histologic changes present and differentiate if the changes seen may be due to artifact from the biopsy procedure. Common artifactual change includes hyperemia, hemorrhage or edema within the lamina propria, loss of luminal epithelium from the shearing effect of the biopsy instrument's jaws, and intussusception or crush artifact of the endometrial glands (R. Kenney, 1978). Changes outside those induced by artifact must then be evaluated and determined if such variations are physiologic or pathologic and what their relative significance is for the mare's fertility.

1.4 Reproductive Cyclicity and Physiologic Variance in the 'Normal' Endometrial Biopsy

Mares are polyestrous long day breeders, meaning that they undergo repetitive estrous cycles during seasons with prolonged photoperiods. In the Northern Hemisphere, this physiologic breeding season corresponds to the late spring, summer, and early autumn, usually spanning from April to October in Western Canada. Outside of the breeding season, when mares are cycling in and out of estrus, three other recognized states occur during the year. The spring and autumn transition periods are characterized by changes where mares are either coming into or out of the breeding season, respectively. Between the autumn and spring transitions is a period of winter anestrus where the majority of mares are not actively cycling. The majority of biopsies are taken during the breeding season when issues concerning fertility are recognized, therefore a discussion involving the cyclic changes that occur during the estrous cycle where the mare cycles between diestrus, the luteal phase, and estrus, the follicular phase, is included in this review.

After the first ovulation occurs in the spring, the physiologic breeding season begins, characterized by repeating estrous cycles lasting usually between 18 to 25 days, averaging between 21 to 22 days per cycle. Each cycle is split into a five to seven-day estrus phase followed by 14 to 16 days of diestrus, with metestrus describing the transition between late estrus and early diestrus, and proestrus describing the transition from late diestrus into early estrus. The first day of the cycle is defined by ovulation (day 0), where a dominant follicle from the previous cycle ovulates in response to rising estradiol levels and a subsequent surge of luteinizing hormone (LH). The area vacated by the follicle within the ovary first fills with blood and fibrin strands to form a corpus hemorrhagicum before filling with luteal tissue by day five of the cycle. This tissue is termed the corpus luteum and produces progesterone. In this way, the mare transitions from estrus to diestrus. During diestrus, progesterone levels remain high through about day 16. Mares may experience one to two follicular waves that will grow and regress during this phase, unable to develop a dominant follicle under the influence of high progesterone. Near the end of diestrus, in the absence of an embryo and without maternal recognition of pregnancy, the endometrium produces prostaglandins such as prostaglandin F_2 alpha that is released into the systemic circulation. Prostaglandin F_2 alpha then reaches the ovary where it binds to the luteal cells resulting in regression of the corpus luteum This results in a rapid decline of progesterone, which

allows the maturation of a dominant follicle from the existing follicular wave. The dominant follicle grows and produces estradiol causing the behavioural signs of estrus, and inhibin that causes the regression of smaller follicles. Rising estrogen triggers an LH surge and ovulation. In this way, ovulation occurs and the cycle begins again.

The varying interactions of hormones during the estrous cycle also influence the appearance of the endometrium. During estrus, the predominant feature is a proliferative pattern characterized by actively dividing stromal and epithelial cells with noticeable mitotic figures (Aupperle et al., 2000; Gerstenberg et al., 1999; Schöniger & Schoon, 2020). The luminal epithelium grows to its peak height and is columnar in shape with elongated and oval shaped basally located nuclei, though these changes may be hard to asses based on how the biopsy is oriented for trimming and slide processing (R. Kenney & Doig, 1986; Schöniger & Schoon, 2020). Two particularly unique features occur during estrus that must be differentiated from either artifactual or pathologic change. First, a significant amount of stromal edema appears during early to mid-estrus, causing the endometrial glands to straighten and a relative decrease in overall gland density (R. Kenney, 1978; R. Kenney & Doig, 1986; Schöniger & Schoon, 2020). This must be differentiated from artifactually produced edema during the time of biopsy. Second, the appearance of inflammatory cells, specifically a low number of polymorphonuclear neutrophils (PMN), may be considered physiologic during estrus (R. Kenney, 1978; R. Kenney & Doig, 1986). The distribution of these PMNs must be examined carefully to determine if the presence of inflammation is physiologic or pathologic. During estrus, PMNs marginate within endometrial blood vessels, congregating and associating with the vascular endothelium. While margination is considered physiologic during estrus, migration of PMNs outside of blood vessels and through the lamina propria of the endometrium is considered pathologic and indicative of endometritis (R. Kenney, 1978; R. Kenney & Doig, 1986). Careful interpretation of PMN distribution is required to differentiate between what is considered normal change during estrus and what may be due to local disease.

While metestrus describes the period between late estrus and diestrus, it is not associated with any discernibly unique histologic changes and, therefore, this phase is not well described in the literature. Instead, histologic changes are described in the context of estrus, diestrus and proestrus where key characteristics can be seen.

In contrast to estrus, diestrus is characterized by a secretory phase where the majority of histologic changes are seen within the endometrial glands. The lack of stromal edema during diestrus causes a relative increase in gland density (R. Kenney & Doig, 1986; Killisch et al., 2017). Glands also appear more tortuous and have a wider lumen diameter compared to the straighter and narrower glands seen in estrus (R. Kenney, 1978; Schöniger & Schoon, 2020). Glandular epithelium becomes more columnar in shape, the nuclei have less dense chromatin, and the apical portion of glandular epithelial cells appear less distinct due to vacuolated cytoplasm (Schöniger & Schoon, 2020). While the basal-most portions of glandular epithelium still have evidence of active proliferation and cell division, the luminal epithelium does not and, instead, proceeds to shrink in height until approximately mid-diestrus when it begins to heighten again (R. Kenney, 1978; R. Kenney & Doig, 1986). As late diestrus approaches, the luminal epithelium reaches near peak height and the endometrial glands begin to involute, shrinking in size and becoming less tortuous, shifting the overall secretory profile back towards the proliferative changes seen in estrus (Schöniger & Schoon, 2020).

The period between late diestrus and early estrus is termed proestrus and it is mainly recognized by the increasing luminal epithelial height, the straightening of endometrial glands, and the reappearance of stromal edema (R. Kenney, 1978). During this early stage of stromal edema, glands become less dense and individual gland branches may appear to bunch or "nest", similar to the pathologic periglandular nesting seen with endometrosis (R. Kenney, 1978). Careful evaluation for surrounding layers of spindle-cell shaped fibroblasts is required to differentiate pathologic fibrotic nesting from the physiologic nesting induced by stromal edema.

Within the physiologic breeding season, mares continually cycle every 18 to 25 days. This cyclicity is reflected in the histologic structure of the endometrium. Therefore, biopsies submitted during this time should include the stage of the cycle the mare in question is in to help the pathologist accurately interpret the changes present. While many changes within the endometrium can be attributed to the time of year and respective stage of sexual cycle the mare is experiencing, there are a variety of pathologic changes that observers must be aware of in order to rule out certain endometrial diseases and causes of reduced fertility.

1.5 Endometrial Diseases and the 'Abnormal' Biopsy

When examining an endometrial biopsy there are several diseases that are included on the list of histopathologic differential diagnoses that could be contributing to reduced fertility in the mare. Knowledge of certain changes and their impact is essential for overall interpretation of the mare's uterine health. A brief summary of such diseases, consequences on uterine health, proposed etiologies, diagnosis and characteristic changes found on histopathologic evaluation of endometrial biopsies is discussed below.

1.5.1 Endometritis

Endometritis is an inflammatory condition of the uterus and can be classified into many different subtypes depending on the etiology of the condition or the histologic features. A host of factors can predispose a mare to developing endometritis including anything that compromises the natural uterine defense mechanisms. Of particular importance is a mare's perineal conformation. Sinking and sloping of the perineum may result in loss of a proper vulvar seal and increase the risk of aspiration of air, fecal material, or both into the reproductive tract. Urine pooling within the vaginal canal can also increase the risk of ascending bacterial or fungal contamination via the cervix into the uterus, thereby causing endometritis. Factors that reduce or prevent the uterus from properly expelling luminal contents also increase the risk of ascending infection, such as a pendulous uterus, scarring or adhesions that affect normal cervical patency, decreased myometrial contractions, retained fetal membranes and reduced lymphatic drainage (McCue, 2019; C. Scott, 2020). At a cellular level, reduced ability of neutrophils to phagocytize bacteria and weaker antibody-mediated immunity within the uterus have also been shown to increase the risk of endometritis (McCue, 2019).

Endometritis, regardless of cause, can have devastating consequences on uterine health; the inflammatory response produces damaging cytokines, causes scar tissue formation, alters secretions, impairs uterine clearance, and can cause premature luteolysis (Causey, 2006; T. Evans et al., 1998; Keller et al., 2006; McCue, 2019). Together, these factors alter the microenvironment within the uterus and pose a very real threat to a developing pregnancy.

Diagnosing endometritis can be difficult. Both cytology and culture, either alone or together, are commonly used when looking for evidence of endometritis as they offer more immediate results and are often more practical in the field setting during the breeding season. However, both tests are less sensitive than histology and only offer samples from the endometrial surface and uterine lumen, potentially missing inflammatory cell populations within the endometrium itself, infectious agents deep within endometrial glands, or suspended in thick resistant biofilm adhered to the endometrial surface. For these reasons, endometrial biopsy is the best route for a definitive diagnosis (LeBlanc et al., 2007; Love, 2011; Nielsen et al., 2012; Overbeck et al., 2011; Riddle et al., 2007). Histology also offers the added advantage where special histochemical staining techniques can be used to more easily identify bacteria, yeasts, and or fungal hyphae such as Gomon ‘s methamine silver stain, periodic acid-schiff (PAS) stain, and gram stain (C. Scott, 2020).

Generally, endometritis is defined as the presence of inflammatory cells within the endometrium (R. Kenney, 1978; R. Kenney & Doig, 1986; McCue, 2019; Schöniger & Schoon, 2020). Histologically this inflammation may be further characterized by the number and type of the inflammatory cells present as well as the distribution of those cells. Different authors have suggested varying cut offs for what method of quantification most accurately detects true endometritis, as the presence of some inflammatory cells may be physiologic due to estrus or as a transient post-breeding response. For example, Kenney suggested that the presence of any neutrophils migrating through the lamina propria indicate endometritis, while other authors specify that the presence of neutrophils during proestrus or estrus may just be physiologic, but if neutrophils are present during diestrus it is indicative of endometritis (R. Kenney, 1978; H. Schoon et al., 1992). Still other authors have set the threshold for endometritis at either one or more, or three or more PMNs averaged over five high powered fields at a magnification of 400x (Buczkowska, Kozdrowski, Nowak, Raś, Staroniewicz, et al., 2014; LeBlanc et al., 2007). While the exact number of inflammatory cells denoting endometritis may be subjective, the inflammatory cell type further describes the type of endometritis present and may help determine the etiology of the condition.

As previously mentioned, the presence of neutrophils marginating within vessels has been described as a normal variant during estrus (R. Kenney, 1978). However, neutrophils migrating

within the stratum compactum or spongiosum is considered pathologic and may be due to persistent breeding-induced endometritis, bacterial endometritis, or subclinical endometritis (R. Kenney, 1978; McCue, 2019; Schöniger & Schoon, 2020). The presence of neutrophils usually denote acute infection and may be present with or without an inciting infectious agent. The presence of sparse and relatively few lymphocytes can be considered normal in the endometrium (Snider et al., 2011). Increased numbers of lymphocytes, the presence of plasma cells or macrophages, are all indicative of a more chronic inflammatory process compared to neutrophilic inflammation (Snider et al., 2011).

Eosinophilic endometritis, while less common than neutrophilic or lymphoplasmacytic endometritis, is characterized by the presence of eosinophils within the endometrium. Recently, work has been done to propose thresholds for the diagnosis of eosinophilic endometritis, as a few eosinophils within the endometrium are considered physiologic during the normal estrous cycle (Grimm et al., 2017). Grimm et al. proposed that anything greater than eleven eosinophils per high powered field in the stratum compactum, or five eosinophils per high powered field in the stratum spongiosum, when averaging counts over ten high powered fields, is indicative of eosinophilic endometritis (Grimm et al., 2017). The exact pathogenesis for this particular form of endometritis is still unclear. While eosinophilic endometritis is traditionally thought to be most commonly caused by fungal agents such as *Candida albicans* or *Aspergillus fumigatus*, other conditions have also been associated with eosinophilic inflammation in the endometrium (McCue, 2019; Snider et al., 2011). Conditions such as allergic reactions, pneumovagina, urine pooling and administration of exogenous progestins have all been found in cases of eosinophilic endometritis (Grimm et al., 2017; McCue, 2019; Schöniger & Schoon, 2020; Snider et al., 2011). Eosinophilic endometritis has also been seen in cases where concurrent bacterial endometritis, endometrosis, or endometrial maldifferentiation are present (Grimm et al., 2017). Similar to other endometrial diseases of the mare, more work is needed to investigate the exact mechanisms behind eosinophilic endometritis and its effect on fertility.

1.5.2 Endometrosis

Endometrosis, also known as chronic degenerative endometritis, is a condition where inappropriate deposition of collagen and fibrosis within the endometrium affects the tissue's proper function. This process is considered the most common histologic culprit of sub-fertility in the mare (Flores et al., 1995; Hanada et al., 2014; Kilgenstein et al., 2015; Ricketts & Alonso, 1991b; Ricketts & Barrelet, 1997).

While the exact pathogenesis of endometrial fibrosis is still unclear, there are a multitude of possible etiologic factors that have been proposed to initiate and/or perpetuate the fibrotic process. These include both inflammatory and non-inflammatory conditions such as endometritis, vasculitis, hypoxia, dysfunctional wound healing, release of growth factors and cytokines, increased production of reactive oxidative species, and mechanical stress (Hoffmann, Bazer, et al., 2009; R. Kenney, 1978; R. Kenney & Doig, 1986; Lee & Nelson, 2012; Schmitt-Gräff et al., 1994; Snider et al., 2011; Walter et al., 2003). Regardless of etiology, diagnosis of this disease can only be made via endometrial biopsy and subsequent histologic examination.

During the early stages of endometrosis, stromal cells within the lamina propria become more organized and collagen is deposited throughout the stratum spongiosum and within the basement membrane of the luminal epithelium (R. Kenney & Doig, 1986). Initially, a change in the size and patterning of stromal cells is seen. These cells become enlarged and ovoid-shaped and start to lose their random distribution, arranging in layers near glandular epithelia (Hanada et al., 2014; Hoffmann, Bazer, et al., 2009; R. Kenney, 1978; R. Kenney & Doig, 1986). These stromal cells then begin to synthesize increased amounts of collagen and extracellular matrix (ECM), usually associated with either the basal lamina of the luminal epithelium or more frequently, the glandular epithelia deeper in the stratum spongiosum (Buczkowska, Kozdrowski, Nowak, Raś, & Mrowiec, 2014; Flores et al., 1995; Hanada et al., 2014; R. Kenney, 1978; R. Kenney & Doig, 1986). As fibrosis progresses and becomes more severe, collagen synthesis begins to slow, eventually ceasing as stromal cells shrink, becoming elongated spindle-shaped quiescent cells that can arrange in concentric layers around affected glands to form 'nests' of coiled glands (T. Evans et al., 1998; Hanada et al., 2014; Hoffmann, Bazer, et al., 2009; R. Kenney & Doig, 1986; Walter et al., 2001). The term periglandular fibrosis is used to describe this general patterning of fibrosis around endometrial glands, while the term 'fibrotic nests' is

used to label the torturous and often dilated glands surrounded by layers of fibrosis (R. Kenney, 1978). Eventually, endometrial glands can become effectively obstructed by periglandular fibrosis, resulting in constriction of the lumen, accumulation of glandular secretion, and overall cystic dilation of the gland (R. Kenney & Doig, 1986). The basal lamina of these glands slowly deteriorates and there is a decreased response of glandular epithelia to the hormonal cycle of the mare, resulting in an increased prevalence of glandular epithelial maldifferentiation, atrophy, and even necrosis (Hoffmann, Ellenberger, et al., 2009; Lehmann et al., 2011; McCue, 2019; Schöniger & Schoon, 2020). The disruption of endometrial gland secretion has a direct effect on fertility as it alters normal histotroph production and the nutrition of the conceptus during early gestation (W. Allen, 1992; Lehmann et al., 2011).

Other possible sequalae of endometrial fibrosis involve direct damage to the endometrial surface resulting in reduced embryonic mobility and implantation, and delayed placental development (T. Evans et al., 1998; Ferreira-Dias & King, 1994; Lehmann et al., 2011; Mambelli et al., 2014). Overall, fibrosis alters the environment within the uterus by impeding the mechanics of uterine clearance, immunity and embryonic movement and implantation, as well as the nutritional supplementation of the embryo and developing fetus. While these effects on their own are particularly significant for fertility, the progressive and irreversible nature of endometrosis gives mares with severe endometrial fibrosis a particularly poor prognosis regardless of treatment.

1.5.3 Endometrial Cysts

There are two different types of cysts that are known to occur within the endometrium and can be seen with histology. Glandular cysts are dilated endometrial glands contained within the lamina propria of the endometrium and are microscopic in nature, being only millimeters to a centimeter in diameter. Lymphatic cysts, on the other hand, result from enlarged lymphatic vessels within the endometrium or myometrium and can range from microscopic lymphatic lacunae to large singular to multilobular structures over ten centimeters in diameter that can protrude into the uterine lumen itself. Both types of cysts have been associated with decreased fertility in the mare, though most authors argue that the presence of the either type of cyst itself is

only secondary to underlying mechanisms that are the primary problem in regards to sub-fertility (W. Allen, 1992; R. Kenney, 1978; Love, 2011; McCue, 2019; Stanton et al., 2004).

Glandular cysts can be a normal finding during pregnancy, however, their presence can also be pathologic in the non-pregnant uterus (Stanton et al., 2004). Endometrial fibrosis can form around endometrial glands and block the flow of glandular secretions, resulting in endometrial cysts that may be filled with retained glandular material. This disruption of glandular function including reduced histotroph production and impaired ability of the glandular epithelium to respond to hormonal cycling of the mare have been suggested to directly contribute to early embryonic loss and overall sub-fertility (W. Allen, 1992; Lehmann et al., 2011; Stanton et al., 2004).

Endometrial lymphatic cysts begin as enlarged lymphatic vessels within either the endometrium or myometrium and are termed lymphatic lacunae. These lymphatic lacunae result from some form of reduction in lymphatic drainage including physical obstruction of a lymphatic vessel, decreased myometrial contractions, or gravitational pooling of lymph fluid due to an enlarged or pendulous uterus (R. Kenney, 1978; McCue, 2019; Stanton et al., 2004). While some researchers have found evidence that lymphatic cysts do not significantly affect fertility, others have suggested contradictory theories (W. Allen, 1992; Eilts et al., 1995; Love, 2011; McCue, 2019; Stanton et al., 2004). As they grow into cyst-like structures, lymphatic cysts can protrude into the uterine lumen and may negatively affect both the conceptus and placental development resulting in early embryonic loss. Larger cysts may obstruct the migration of an embryo through the uterus, thereby blocking the maternal recognition process necessary to prevent luteolysis and establish pregnancy (W. Allen, 1992; Love, 2011; McCue, 2019; Stanton et al., 2004). These cysts are most commonly found at the base of a uterine horn, the site of embryo fixation and development. This presents multiple potential problems for the embryo as implantation beside a cyst can reduce early blood flow and nutrition as the cyst reduces functional endometrial space adjacent to the embryo. Cysts also reduce the surface area available for placental development, reducing overall placental exchange between the maternal and fetal circulation (McCue, 2019; Stanton et al., 2004). Finally, the presence of cysts within the endometrium can complicate the diagnosis of pregnancy as cysts are often singular, round and hypoechoic on ultrasonography, much like an early conceptus. This may result in a false pregnancy diagnosis, or in an assumed

diagnosis of twins where a clinician may inadvertently manually reduce a single conceptus while assuming that the cyst is a second viable pregnancy (McCue, 2019; Stanton et al., 2004).

While ultrasonography and hysteroscopy are often the best diagnostic methods for larger lymphatic cysts, endometrial biopsy and histology are needed to diagnose glandular cysts, smaller lymphatic cysts or lymphatic lacunae. As previously mentioned, glandular cysts will appear as dilated endometrial glands often with surrounding periglandular fibrosis and/or inspissated eosinophilic material representing retained glandular secretions. Lymphatic lacunae and lymphatic cysts, on the other hand, are distinct histologic findings.

Lymphatic lacunae describe enlarged lymphatic vessels and must be differentiated from physiologic edema during estrus or from artifact created by forceps during the biopsy procedure. To help differentiate between physiologic edema and lymphatic lacunae, knowledge of the mare's stage of cycle at time of biopsy as well as evaluating the distribution of the edema is helpful, as physiologic edema is uniform and widespread throughout the endometrium, reducing the overall glandular density, while lymphatic lacunae are distributed less uniformly (R. Kenney, 1978). Biopsy artifact creates edema by directly damaging lymphatic vessels, therefore, searching for evidence of endothelial damage as well as the presence of hemorrhage within the lumen of lymphatic vessels will help discern between artifact and true lymphatic lacunae (R. Kenney, 1978). As lymphatic lacunae enlarge into cysts, their histologic appearance changes from a simple enlarged vessel to a large dilated structure where the outer lining of the cyst is consistent with the luminal columnar epithelium of the endometrium separated from an inner lining of squamous to cuboidal cells by a band of normal lamina propria tissue containing endometrial glands and abnormal hyaline (Brook & Frankel, 1987; Stanton et al., 2004).

In sum, both types of cysts have been found in healthy mares as well as sub-fertile problem mares with the relative number, size and location contributing to the significance of these cysts on overall fertility. Endometrial biopsy not only allows for definitive diagnosis of glandular cysts and smaller lymphatic lacunae, but also allows for evaluation of underlying endometrial pathology such as endometrosis that may be the inciting cause for such cystic development and carry a more serious prognosis regarding mare fertility.

1.5.4 Non-seasonal Endometrial Atrophy

Endometrial atrophy is a normal physiologic finding during the winter anestrus period (Doig et al., 1981; R. Kenney, 1978). A degree of endometrial atrophy is also considered physiologic during the spring and fall transition periods, as the hypothalamic-pituitary-gonadal axis of the mare begins to change and the endometrium subsequently responds (Doig et al., 1981; R. Kenney, 1978). However, once the mare is in the midst of the breeding season and is cycling regularly between true estrus and diestrus, any endometrial atrophy is considered pathologic as it reflects a reduced ability of the endometrium to respond to hormonal stimulus and maintain a pregnancy (Doig et al., 1981; R. Kenney, 1978; R. Kenney & Doig, 1986).

Diagnosis of endometrial atrophy requires procurement of an endometrial biopsy and histologic evaluation, though manual palpation revealing a relatively thin uterine wall with reduced presence of endometrial folds is another indicator (R. Kenney, 1978). Histologic findings are characterized by a relative thinning of the endometrial layers with glandular changes including quiescent epithelium with no evidence of active secretory or proliferative differentiation, inspissated secretions within the glandular lumina, and an overall reduction in gland number (R. Kenney, 1978). Etiologies of endometrial atrophy during the breeding season include ovarian dysfunction, severe endometrosis, senility, and unknown causes (R. Kenney, 1978; Ricketts & Barrelet, 1997). This atrophy can be widespread throughout the uterus or can be focally distributed, sometimes affecting a single uterine horn (R. Kenney, 1978). Atrophic endometrium fails to support pregnancy in a variety of ways including an overall decrease in glandular function and reduced surface area for placental attachment (Ferreira-Dias & King, 1994; R. Kenney, 1978).

Not to be confused with endometrial atrophy is the condition of endometrial hypoplasia, where the endometrium has failed to develop fully and lacks the ability to respond to normal hormonal influence. This is a congenital condition that becomes apparent in mares whose ovaries have failed to become active following sexual maturation (R. Kenney, 1978). Since the endometrial histologic features of this disease are very similar to that of endometrial atrophy, diagnosis of endometrial hypoplasia requires further investigation beyond an endometrial biopsy including a thorough physical examination for other evidence of congenital abnormalities and genetic testing (R. Kenney, 1978).

1.5.5 Endometrial Maldifferentiation

Endometrial maldifferentation refers to a disorder describing several abnormal differentiation patterns of endometrial glands in relation to the stage of a mare's reproductive cycle. Broadly, an affected endometrium can be categorized as having either unequal or irregular maldifferentiation.

Unequal maldifferentiated endometrium are those that display two distinct patterns of glandular differentiation within the biopsy. A predominant pattern is evident that reflects the normal sexual cycle (e.g., proliferative in estrus and secretory in diestrus) while groups of glands can be found in the biopsy with differing epithelial differentiation (H. Schoon et al., 2000). These glands have been found to express varying levels of steroid receptors and certain epithelial proteins compared to normally differentiated glandular epithelia (Häfner et al., 2001; H. Schoon et al., 2000). While the etiology of this glandular epithelial disorder is unknown, it has been suggested that any changes to affected glands resulting in altered steroid receptor expression may contribute to inappropriate differentiation (H. Schoon et al., 2000).

In comparison, irregular maldifferentiation describes endometrium where there are no identifiable glandular epithelial differentiation patterns. Instead, most of the glands display variable characteristics of both proliferative and secretory differentiation, making it impossible to assign any specific pattern to the biopsy as a whole (H. Schoon et al., 2000). Possible causes for this pattern of maldifferentiation include ovarian disorders such as corpus luteal cysts, granulosa-thecal cell tumors or teratomas, ovarian inactivity secondary to intense physical performance, and exogenous hormonal therapy (Ellenberger et al., 2002; Kilgenstein et al., 2015; Klug et al., 1997; H. Schoon et al., 2000).

Regardless of the type of endometrial maldifferentiation, definitive diagnosis can only be made by endometrial biopsy. Histologic diagnosis involves evaluating glandular epithelial cells for indications of proliferative or secretory differentiation based on overall gland diameter, luminal size, epithelial height, and various characteristics of epithelial cell nuclei and cytoplasm, and interpreting which patterns appear to be the most consistent throughout the biopsy (H. Schoon et al., 2000). While evaluation on routine hematoxylin and eosin stained biopsies is possible, especially when following schematic diagrams given by Schoon et al.,

immunohistochemistry to assess steroid receptor expression variability and certain protein markers is highly useful (Kilgenstein et al., 2015; H. Schoon et al., 2000; H. Schoon & Schoon, 2003). Both types of endometrial maldifferentiation are thought to negatively affect fertility as affected glands may not secrete adequate glandular proteins resulting in changes within the uterine microenvironment that can result in early embryonic loss (H. Schoon & Schoon, 2003). Additionally, biopsies with diagnosed endometrial maldifferentiation are more likely to have other endometrial comorbidities such as endometritis, endometrosis, or angiopathies, all conditions that can further negatively impact fertility (Ellenberger et al., 2002; Häfner et al., 2001; H. Schoon et al., 2000).

1.5.6 Angiopathies

Another group of endometrial diseases involves changes found in the uterine blood vessels. Angiopathies refer to both inflammatory and non-inflammatory histopathologic changes within and around the wall of uterine arteries and/or veins. Of particular importance to mare fertility is angiosis or angiosclerosis, the non-inflammatory condition characterized by increased deposition of elastin or collagen within or around blood vessel walls.

While the exact pathogenesis of angiosis is unknown, the prevalence of the condition has been shown to be higher in older, multiparous mares with other concomitant endometrial pathology (Grüninger et al., 1998; Schöniger & Schoon, 2020). Increasing severity of angiosis has also been associated with reduced uterine blood flow and suggested to directly affect fertility by affecting the maternal recognition of pregnancy, placental formation and development, as well as affecting endometrial gland function (Esteller-Vico et al., 2015; Grüninger et al., 1998). Severe cases of angiosis can even weaken uterine arterial walls to the point that they may rupture, risking abortion or even fatal hemorrhage (Grüninger et al., 1998; Schöniger & Schoon, 2020).

Diagnosis of angiosis can only be made via endometrial biopsy and histology. On routine hematoxylin and eosin stained sections, the characteristic histologic feature involves thickening of the vessel wall expanded with eosinophilic extracellular matrix and a loss of discreet cellular detail (Schöniger & Schoon, 2020). The exact type of angiosis can then be further characterized based on what types of blood vessels are affected, what histopathologic changes are present (e.g.,

collagen or elastin deposition), and which specific layer of the blood vessel is affected (Schöniger & Schoon, 2020). Histochemical stains can help to both qualitatively and quantitatively evaluate the degenerative changes present in a biopsy sample. For example, picrosirius red stain will differentially stain collagen and elastin fibers within and around vessel walls (Schöniger & Schoon, 2020). Similar to glandular and lymphatic cysts, the presence of angiosis usually occurs in concert with other degenerative endometrial conditions such as endometrosis, and together these changes carry a poor prognosis for fertility.

1.5.7 Endometrial Neoplasia

Neoplasia involving the endometrium of the mare appears to be exceedingly rare. Since the 1970s only a handful of case reports have been published and include endometrial adenocarcinoma, fibrosarcoma, and metastatic lymphoma (Canisso et al., 2013; Claes et al., 2015; Freeman et al., 1997; Govaere et al., 2011; Gunson et al., 1980; Lopez et al., 2018; Thompson et al., 2014).

Of the three uterine adenocarcinoma case reports found in the literature, definitive diagnosis was obtained by histopathologic findings from post mortem histology. Common findings included abnormal neoplastic glandular epithelium that frequently invaded the myometrium and lymphatic vessels, often resulting in metastasis to local lymph nodes and other organs (Gunson et al., 1980; Lopez et al., 2018; Thompson et al., 2014). All mares presented with abnormal transrectal palpation and ultrasonographic findings that included identifiable masses and retention of free fluid within the body and horns of the uterus. Additionally, all three mares had progressive clinical signs that warranted euthanasia based on the presumed antemortem diagnosis, evidence of metastasis and associated poor prognosis (Gunson et al., 1980; Lopez et al., 2018; Thompson et al., 2014). Therefore, uterine adenocarcinoma, though extremely rare in horses, is not only a reason for infertility, but also carries consequences for overall systemic health of the mare.

Aside from relatively rare endometrial neoplasms, leiomyomas of the underlying myometrium appear to be the most common uterine neoplastic disorder, though they are not considered a likely cause of sub-fertility in the mare due to their infrequent occurrence and the

relatively common prevalence of other previously discussed endometrial disease (R. M. Kenney, 1978; McCue, 2019).

1.6 Hallmarks of Endometrial Pathology as Originally Described by Kenney:

Endometrial histopathology has a long history of development and, over the past 40 years, suggestions by various authors involving which histologic lesions are significant to fertility and how to quantitatively measure these lesions have been made to help histopathologists more accurately analyze biopsies (Snider et al., 2011). In 1978, Kenney proposed a unique categorical system for the uniform assessment of equine endometrial biopsies that would build the foundation for endometrial evaluation used today (R. Kenney, 1978).

This original system identified key histologic lesions associated with several, but not all, of the previously mentioned endometrial diseases, as some were yet to be described in the literature. In particular, Kenney focused on histologic changes associated with endometritis, endometrosis, endometrial cysts, and endometrial atrophy. Kenney suggested that the degree of these histologic lesions reduced fertility, and therefore, used them as guidelines for classifying the overall severity of pathology in a given endometrial sample (R. Kenney, 1978; R. Kenney & Doig, 1986; Love, 2011). These lesions included the degree of inflammation, fibrosis, glandular distension, and lymphatic lacunae present in the sample (Figure 1.2). Kenney used the severity and distribution of inflammation, fibrosis, glandular distension, and lymphatic lacunae in a given biopsy to classify it within a three-category scoring system that correlated biopsy grade with a foaling rate prognosis.

Category I endometrium was described as appearing normal or with mild, scattered pathologic change (R. Kenney, 1978). Category II was a broader category, encompassing more histologic changes such as mild to moderate inflammation, fibrosis, and scattered lymphatic lacunae and setting multiple different qualifying thresholds for these changes. Specifically, inflammatory cell infiltration was broken down into either appearing as moderate and diffusely spread throughout the endometrium or limited to scattered groups or foci throughout the endometrium (R. Kenney, 1978). Category II fibrosis was further defined as either involving

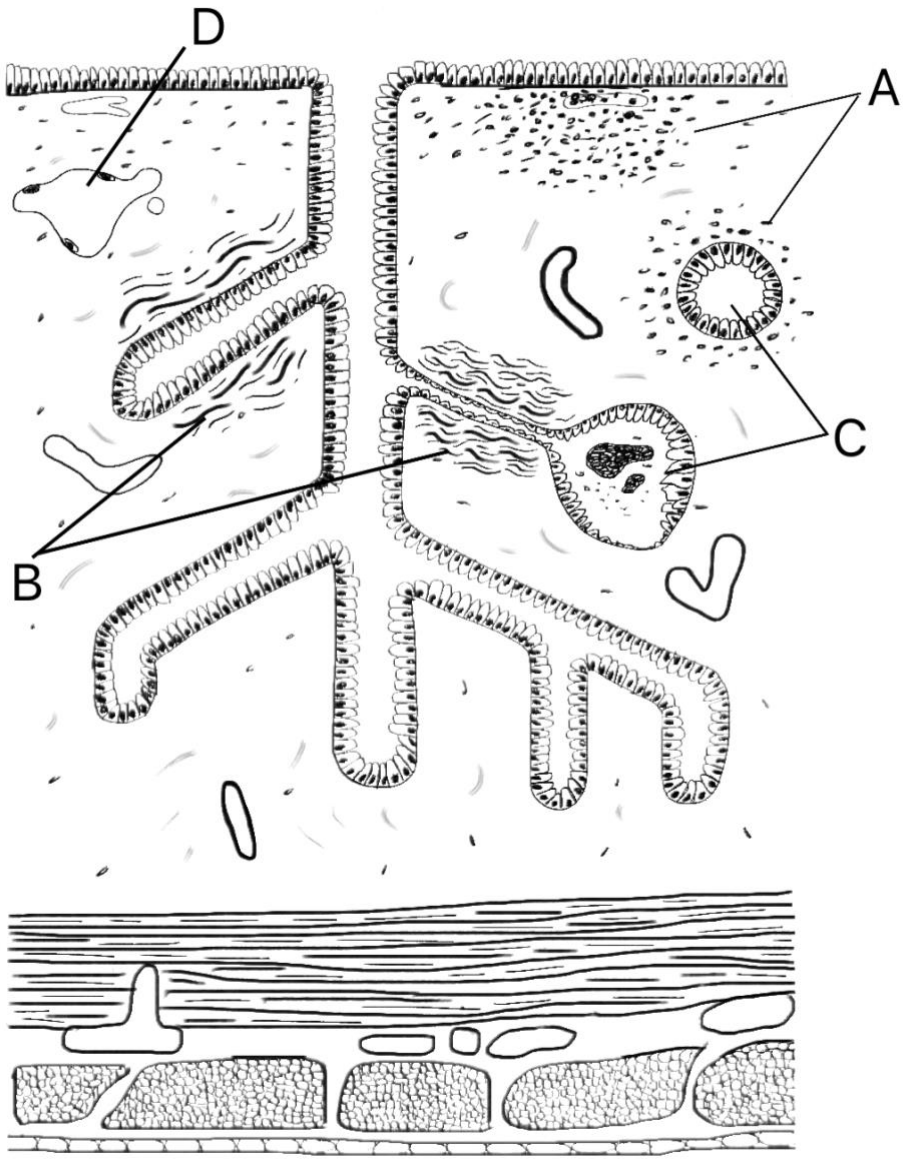

Figure 1.2. Schematic showing a cross section of an abnormal equine endometrial biopsy adapted from Kenney, 1978b. Features include specific hallmarks of pathology described by Kenney. A: Inflammation: characterized by increased populations of polymorphonuclear cells, lymphocytes, plasma cells, macrophages, siderophages, eosinophils and mast cells. B: Fibrosis: characterized by organized layers of stromal cells and increased collagen deposition. C: Cystic dilation: constriction of gland lumen resulting in enlargement of the terminal end and accumulation of secretions. D: Enlarged lymphatic lacunae.

frequent individual gland branches or forming a maximum of three fibrotic nests counted per five mm linear field. Both inflammation and lymphatic lacunae were considered reversible and Kenney noted that should either of these characteristics decrease following treatment, a subsequent endometrial biopsy from the same mare may be re-classified as a category I. Endometrium that appeared either hypoplastic or atrophied during the breeding season was also included in this category and, again, could re-classified to a less severe category if it returned to a normal glandular cycle. Category III encompassed the most severe pathologic specimens. It described samples with widespread irreversible change including diffuse periglandular fibrosis with five or more fibrotic nests per linear field. Diffuse and heavy inflammatory cell infiltration and large lymphatic lacunae that disrupted uterine wall consistency were also included in this category along with mares with gonadal hypoplasia and ensuing hormonal disruption of endometrial development (R. Kenney, 1978).

After defining each grade, Kenney then categorized 244 samples from mares according to his proposed system and collected foaling data from the subsequent foaling season in an attempt to correlate the categories with fertility. In one group of mares, Kenney found that category I, II, and III mares foaled at rates of 68%, 51%, and 11%, respectively (R. Kenney, 1978). When controlling for other factors like differences in management and veterinary care, he found an even wider disparity between foaling rate prognoses with 92%, 67% and 4.3%. Kenney concluded that together with clinical findings from both an in-depth history and a reproductive exam, histologic grades could be used to develop an 'epicrisis' or overall prognosis for the mare's potential breeding career (R. Kenney, 1978). This initial system showed a significant relationship between foaling rates and biopsy grade and the Kenney scale became widely used throughout the United States over the following five years (Doig et al., 1981).

1.7 The Evolution of the Kenney-Doig Scale:

As the use of endometrial biopsies gained popularity among equine practitioners, Doig, McKnight, and Miller conducted a study to test the association between endometrial biopsy grade and foaling rates (Doig et al., 1981). They began by using the newly proposed Kenney scale, but soon found that fibrosis was better described in four categories as opposed to three: absent, mild, moderate, or severe as opposed to Kenney's original absent to mild, mild to moderate, or

moderate to severe. As well as modifying the categorization, Doig et al. also found a significant correlation between the number of years a mare had been barren and her prospective foaling rate. Mares with discernible fibrosis and only one year of barren history were found to have a foaling percentage of 82%, 62%, and 0% when classified into Kenney categories I, II, and III, respectively, while mares with a greater than two year barren history had 53%, 28%, and 0% (Doig et al., 1981). This led the authors to suggest two important modifications to the existing Kenney scale: creating an additional category to accommodate more distinct criteria regarding fibrosis, and including the years barren as an important modifier when assigning a final biopsy grade (Doig et al., 1981).

Between 1981 and 1986, two groups of researchers worked to further solidify the relationship between increasing Kenney grade and decreasing fertility. Shideler et al found similar foaling rates as Kenney's original study, citing 61%, 48.3%, and 35.1% for Kenney grade I, II and III ranked endometrial biopsies (Shideler et al., 1982). In the same year, de la Concha-Bermejillo and Kennedy released a retrospective study correlating Kenney rank and fertility (de la Concha-Bermejillo et al., 1982). They found foaling rates of 78%, 55%, and 35% for category I, II, and III classified mares that added to the growing support behind Kenney's system (de la Concha-Bermejillo et al., 1982).

While these studies corroborated the general trend Kenney had originally demonstrated, Kenney aimed to further refine his system. Kenney and Doig published a joint paper outlining a new version of the Kenney scale including the modifications Doig et al. had suggested in 1981 (Table 1.1) (R. Kenney & Doig, 1986). The most profound change arose from splitting the middle category II into two separate categories, creating four categories from the previous three. Categories I and III remained similar to the original categories described by Kenney. Category II experienced the most change as it was split into Category IIA and IIB.

Category IIA was described as having one of four basic hallmarks. The first included mild to moderate inflammation that could either be spread diffusely throughout the stratum compactum or in scattered foci throughout the lamina propria (R. Kenney & Doig, 1986). Fibrosis that frequently surrounded individual gland branches or formed fibrotic nests that averaged less than two nests per 5.5 mm linear field in a minimum of four fields was another independent qualifier for category IIA. The third characteristic included lymphatic lacunae that

Table 1.1 Description of the histologic grading system proposed by Kenney and Doig including the modification to the original Kenney categories (R. Kenney & Doig, 1986)

Category	Pathologic Changes				Special Considerations
	Glandular Atrophy	Inflammation	Fibrosis	Lymphatic Lacunae	
I	None	Slight and sparsely scattered	Slight and sparsely scattered around individual branches Slight and sparsely scattered nests	Slight and sparsely scattered	If more than 1 pathologic change is present, then the biopsy is moved to category IIA
IIA	Partial atrophy present late in the physiologic breeding season	Slight to moderate diffuse infiltrations of the stratum compactum or Scattered but frequent within the stratum compactum or spongiosum	Frequent and scattered distribution Individual gland branches of any number of layers, but usually only 1 to 3 layers An average of only 2 or less nests in 4 or more 5.5 mm linear fields	Lacunae that are large enough to produce a palpable change within endometrial folds or the uterine wall	If more than 1 pathologic change is present, then the biopsy is moved to category IIB History of barrenness > 2 years then the biopsy is moved to category IIB If inflammation, lymphatic change, or glandular atrophy is corrected via treatment, the biopsy is able to move to a category I
IIB	Not described	Widespread, diffuse, moderately severe foci	Widespread with uniform distribution Individual branches with 4 or more layers An average of 2 to 4 nests in 4 or more 5.5 mm linear fields	Not described	If more than 1 pathologic change is present, then the biopsy is moved to category III If the inflammation or lymphatic change is corrected via treatment, the biopsy is able to move to a category IIA
III	Deep atrophy present within the physiologic breeding season	Widespread, diffuse and severe distribution	Uniformly widespread distribution An average of 5 or more fibrotic nests in 4 or more 5.5 mm linear fields	Severe enough to produce a "jelly-like" feel to the endometrial folds or uterine walls	If more than 1 pathologic change is present, the prognosis is more severe If the inflammation or lymphatic change is corrected via treatment, the biopsy may move to a category to IIB unless fibrosis is also present. Fibrosis at this stage is considered severe enough that the biopsy is **unable** to move categories and will remain a category III

were large enough to produce palpable change within the endometrial wall. Finally, any endometria that appeared to be atrophic despite being within the breeding season were considered the fourth factor that could place a biopsy within the IIA group. Any one of these four signs of mild inflammation, fibrosis, lymphatic lacunae, and non-physiologic atrophy qualified the biopsy as a category IIA. Importantly, they were considered additive and the presence of two or more of these signs within the same biopsy required the grade to be re-classified as a category IIB.

While the presence of two or more of any of the previously mentioned mild changes classified a biopsy as category IIB, other singular but more severe histologic characteristics could assign a biopsy into the same group. Moderate periglandular fibrosis that comprised four or more layers around glandular branches or an average of two to four fibrotic nests per 5.5 mm linear field in a minimum of four fields qualified an endometrial sample as a category IIB (R. Kenney & Doig, 1986). Inflammation alone could qualify a sample as IIB if the changes deemed moderate instead of mild. The additive criteria also applied to category IIB; more than one moderate pathologic change would re-classify the biopsy as a category III.

Category III was also described with four characteristics in mind. Severe inflammation with a diffuse distribution was the first qualifier. Widespread fibrosis was a second, either uniformly spread throughout individual gland branches or producing an average of five or more fibrotic nests per 5.5 linear field in a minimum of four fields (R. Kenney & Doig, 1986). Included submission history involving lymphatic lacunae large enough to lend a gelatinous-like texture to the endometrial wall on transrectal palpation was considered another qualifying sign for category III. Finally, if the mare in question had diffuse and severe endometrial glandular atrophy despite being within the physiologic breeding season, the biopsy was considered a category III. While the presence of only one of these severe pathologic features was considered sufficient to categorize the biopsy as a category III, the presence of two or more of these changes together made the prognosis far worse than if only one change was present (R. Kenney & Doig, 1986). It was also made clear that inflammation, lymphatic lacunae, and endometrial atrophy are all potentially reversible changes that could benefit from treatment and allow a mare be re-classified into one of the less severe categories on subsequent biopsies. Fibrosis, on the other hand, is irreversible and, therefore, prognoses involving these mares should be guarded as treatment is often not successful (R. Kenney & Doig, 1986).

Another important change within the Kenney-Doig system that varied from the original scale proposed by Kenney involved the inclusion of the number of years barren as an important modifier when determining an endometrial biopsy grade (R. Kenney & Doig, 1986). For example, mares that had been graded a category IIA but had also experienced greater than two years of barrenness may re-classified to a category IIB. Finally, Kenney and Doig still emphasized the importance of the evaluating clinician to develop an overall 'epicrisis', the final summation of each case that interpreted the mare's history, physical exam findings, microbiological results, and biopsy grade to provide a more accurate fertility prognoses and direct possibly therapy (R. Kenney & Doig, 1986).

Ultimately, Kenney and Doig published a simplified chart to correspond each biopsy categorization with expected foaling rates (Table 1.2) (R. Kenney & Doig, 1986). These guidelines are currently in place throughout North America and continue to be used by histopathologists on a daily basis, acting as the standard for analyzing equine endometrial biopsies. Armed with this information, clinicians can use the biopsy results to help make a number of decisions including possible treatment modalities, ideal breeding technique (e.g., natural versus artificial insemination), and ultimately the economic viability of a sub-fertile horse based on her projected foaling success.

1.8 Criticisms of the Kenney-Doig Scale

While the Kenney-Doig scale has remained the predominant system for classifying endometrial biopsies in North America, the scale has received criticism throughout its use involving unvalidated prognostic relevance, subjective category guidelines, and inadequate inclusion of other pertinent endometrial pathologies.

1.8.1 Endometrial Pathology Not Included in the Kenney-Doig Scale

Over the past two decades, a large group of researchers from Germany, known as the Schoon group, have been searching for different histopathologic characteristics they feel are missing from the Kenney-Doig scale. The Schoon group has described and explored recently

Table 1.2. Expected foaling rates of mares according to categorization of endometrium (R. Kenney & Doig, 1986). Pathologic change is additive; more than one mild pathologic feature would qualify an endometrial biopsy as a category IIB, while more than one moderate pathologic feature would qualify as category III.

Category	Degree of Endometrial Change	Expected Foaling Rate (Percent)
I	Absent	80-90
IIA	Mild	50-80
IIB	Moderate	10-50
III	Severe	10

recognized endometrial pathologic changes such as endometrial maldifferentiation and angiopathies. They have made a compelling case to include such histopathologic changes in the categorization of biopsies when evaluating mare fertility (H. Schoon & Schoon, 2003). Multiple other studies from this group attempted to identify endometrial pathologic markers that influence fertility and that could be used to better evaluate a mare's reproductive potential (Aupperle et al., 2004; Grüninger et al., 1998; Hoffmann, Bazer, et al., 2009; Hoffmann, Ellenberger, et al., 2009; Kilgenstein et al., 2015; Lehmann et al., 2011; H. Schoon et al., 2000; H. Schoon & Schoon, 2003). However, more work is needed to investigate the relationship between these new endometrial pathologic changes and their potential effects on fertility, as well as their relationship with other endometrial diseases. Depending on this future work, modifications to the Kenney-Doig scale may be warranted. As of today, these histologic features and the suggested modifications to the Kenney-Doig scale appear to be used in Germany, however, they have yet to be incorporated into routine endometrial biopsy evaluation in North America.

1.8.2 Validity of the Kenney-Doig Scale and Prognostic Value

In regards to histopathologic grading systems, validity refers to how well the measurement reflects the truth, or how well the measure of interest correlates with an accurate prognosis (Silcocks, 1983). Since the modification of Kenney's scale into the guidelines proposed by Kenney and Doig, there are only a short list of studies designed to test the prognostic relationship between the four proposed Kenney-Doig categories and the associated estimated foaling rates. Of note is the retrospective study that Snider et al described as the only current evaluation of the Kenney-Doig scale (Snider et al., 2011). In 1990, Waelchli evaluated 192 biopsies from mares that had presented to the University of Zurich Veterinary College Gynecological Clinic. Using the Kenney-Doig scale, he categorized each biopsy and compared the grades to foaling rates obtained in the following breeding season. He found foaling rates of 70, 42, 18 and 0% for categories I, IIA, IIB and III respectively, each within the ranges proposed by Kenney and Doig (Waelchli, 1990). While this may be the only study that evaluates biopsy grade and expected foaling rates using the exact Kenney-Doig guidelines, other studies have since been done with slight modifications to study design and biopsy categorization.

A study by Held and Rorhbach demonstrated that as severity of biopsy grade increases, the chance of pregnancy decreases (1991). More recently, Nielsen et al found a similar negative correlation between biopsy grade and pregnancy status as of gestation day 70 (Nielsen et al., 2012). While both studies adhered closely to the Kenney-Doig classification system, both reported fertility as the ability for a mare to either become pregnant or maintain pregnancy to a specified point in gestation. Kenney originally defined fertility as the ability to conceive, maintain and deliver a live foal, therefore he measured seasonal foaling rates against biopsy grades for a more accurate prognosis concerning reproductive potential (R. Kenney, 1978). Studies using pregnancy as their primary measure may overestimate the fertility of each category of mare as fetal loss later in gestation is not accounted for.

Another important study is that of Ricketts and Alonso who evaluated 530 sub-fertile mares via endometrial biopsy both before and after uterine treatment and correlated them with subsequent foaling rates (1991a). While the researchers used both the same classification standards and fertility measure as Kenney and Doig, they used a paired biopsy design as opposed to a single biopsy. Ultimately, Ricketts and Alonso found that their two most severely graded pre-treatment groups had significantly higher foaling rates than expected (1991a). This led the authors to conclude that paired biopsy, both before and after uterine treatment, is a more accurate diagnostic tool than a single pre-breeding biopsy (Ricketts & Alonso, 1991a).

With a different colleague, Ricketts carried out a retrospective review where over 4000 biopsies from Thoroughbred mares were classified into 15 groups based on the presence of different combinations of six histopathologic descriptors (Ricketts & Barrelet, 1997). Again, though this study also confirmed that increasing severity of biopsy grade correlates with decreasing fertility, these authors did not use the exact classification system as Kenney and Doig and therefore extrapolation between the two is difficult.

There does exist a thesis project from this author's own academic institution that evaluated the use of the Kenney-Doig scale prospectively in a herd of fertile mares (Manning, 2002). Endometrial biopsies were taken from a herd of post-partum mares from a Pregnant Mare Urine (PMU) ranch in the spring of 1996 and then compared to foaling rates from the following season. Surprisingly, the researchers found no association between biopsy grade and foaling outcomes. When they compared the amount of inflammation and fibrosis in each biopsy to

foaling outcome, they found that foaling outcome was only significantly correlated with fibrosis (Manning, 2002). They suggested that post-partum mares have a reversible physiologic inflammation present in the uterus that does not have the same detrimental effect on reproductive efficiency as endometrial fibrosis. Given this, they suggested that the Kenney-Doig scale itself should be modified to put more emphasis on irreversible endometrial fibrosis, and less on inflammation that is potentially physiologic and reversible, when grading endometrial biopsies (Manning, 2002). While this study provided prospective work on the Kenney-Doig system, it evaluated biopsies collected during the post-partum period while most submitted biopsies are submitted pre-breeding or after issues of sub-fertility. It also focused on a relatively young mare population that had been specifically bred for reproductive efficiency, unlike the vast majority of older, sub-fertile reproductive cases that present for work-up in clinical practice. Therefore, drawing comparisons among this population to mixed mare populations submitted to diagnostic laboratories is difficult.

Of this list of studies, variables concerning the method of measurement such as using gestation day as a benchmark for fertility versus foaling rate, the exact method of biopsy classification, the mare population, the number of biopsies procured, and the effect of uterine therapy have all differed from the original studies by Kenney and Doig used to stratify the different biopsy categories with a percent foaling rate prognosis. As such, the Kenney-Doig system's validity remains arguably under tested (Snider et al., 2011).

1.8.3 Repeatability and Subjectivity of the Kenney-Doig Scale

Aside from potential issues with validity, there have been studies in the literature that have voiced concerns regarding the vague category definitions and the subjective evaluation of histologic features that may affect how different observers categorize similar endometrial biopsies using the Kenney-Doig scale.

A study by Evans et al. argued that the guidelines for evaluating fibrosis in particular during Kenney-Doig categorization of endometrial biopsies is semi-quantitative and subjective (T. Evans et al., 1998). They advocated for more objective and reproducible measures, exploring

the potential of using collagen-specific stains and image analysis to better measure the amount of fibrosis in a given biopsy (T. Evans et al., 1998).

In two separate studies by Ricketts and Alonso, mares were found to be older than expected or to have higher foaling rates than expected given their Kenney-Doig classification (Ricketts & Alonso, 1991a, 1991b). While the authors mentioned that these discrepancies could be due to the mare population examined, they could also be due to observer differences in interpretation and categorization of the studied biopsies.

An in-depth 2011 review describes the history, creation, and application of the Kenney-Doig scale (Snider et al., 2011). The authors specifically outlined a number of limitations including the subjective criteria and lack of consistency in grading, the controversy over the representativeness of a single biopsy for the entire endometrium, and the inadequate number of retrospective studies that have since evaluated the system (Snider et al., 2011). They suggested that new techniques such as immunohistochemistry, in situ hybridization, and polymerase chain reaction should be investigated as possible avenues to lend both objective data to the categories and identify new histologic markers that may correlate with reduced fertility (Snider et al., 2011).

1.8.4 Evaluating and Improving the Kenney-Doig Scale

Throughout the literature, three common themes emerged. First, since the publication of the modified Kenney-Doig scale, no retrospective evaluation of the prognostic ability of the system has been done in Canada, or even North America. Waelchli undertook a retrospective analysis on 192 biopsies from the University of Zurich Veterinary College, but his study appears to stand alone (Waelchli, 1990). As Snider et al. described, "just one peer-reviewed evaluation of the scheme seems inadequate" (Snider et al., 2011). Second, given new insights into the pathogenesis of endometrial fibrosis and the development of novel diagnostic techniques such as immunohistochemistry, other biomarkers exist that may allow for more accurate quantification of fibrosis and give better prognostic information regarding mare fertility (Schöniger & Schoon, 2020; H. Schoon & Schoon, 2003). And finally, concerns have been raised regarding the use of the Kenney-Doig scale and the categorization of the biopsies themselves. Multiple authors have offered 'differences in interpretation' or 'subjective guidelines' as an explanation for

disagreement between the system and their results, or voiced concerns over the broad use of the Kenney-Doig scale and issues with observer variability (T. Evans et al., 1998; Ricketts & Alonso, 1991a, 1991b; Schlafer, 2007; Snider et al., 2011). This leads to the subject of inter and intra-rater agreement and whether the Kenney-Doig scale is reproducible between observers or within observers. To date, no studies have been undertaken to assess the inter or intra-agreement using the Kenney-Doig scale.

1.9 The Concept of Inter-rater and Intra-rater Agreement and Histopathology:

Histopathology is a unique field within both human and veterinary medicine. While qualitative descriptions of various lesions are routine for complete necropsy reports in ascertaining a cause of death, certain areas involving surgical biopsy histopathology use structured scoring systems similar to the Kenney-Doig scale to communicate prognostic information to submitting clinicians. Examples of such histopathologic scoring systems include those used to evaluate chronic hepatitis, endometrial carcinomas, cervical dysplasias, mammary tumours, and mast cell tumours (Bergeron et al., 1999; de Vet et al., 1990; Fadare et al., 2013; Ishak et al., 1995; Kiupel et al., 2011; Malpica et al., 2005; Matos et al., 2012; Munkedal et al., 2016; Northrup, Howerth, et al., 2005; Robbins et al., 1995; Scholten et al., 2004; Westin et al., 1999).

For a scoring system to be considered accurate, it must be both valid and repeatable (Cross, 1998; Silcocks, 1983). While proving validity of a scoring system involves measuring categorized biopsies against a predictor outcome, for example tumor grades with patient survival times, measuring repeatability involves examining the amount of observer variability produced by the system. The subject of observer variability, including inter and intra-rater agreement, is particularly important in a field like histopathology where the majority of measurements and evaluations are based on qualitative to semi-quantitative guidelines and assessments made on a subjective basis. Other diagnostic fields such as the clinical pathology areas of hematology and biochemistry, use more quantitative parameters with numerical values. The histopathologist, however, is often left relying on less discreet guidelines. Additionally, there is commonly no gold standard or reference range available for a histopathologist to compare their own interpretation

against, as their diagnosis is often regarded as the gold standard itself (Brothwell et al., 2003; Langley, 1978). Therefore, there is substantial room for the possibility of observer variability to be present in histopathology, and as such inter- and intra-rater agreement studies have been employed as a mechanism to act as both quality control and justification for the continued use or proposed alteration of certain histopathologic scoring systems in both human and veterinary medicine (Bergeron et al., 1999; Brothwell et al., 2003; de Vet et al., 1990, 1995; Fadare et al., 2013; Geboes, 2000; Ishak et al., 1995; Kiupel et al., 2011; Langley, 1978; Malpica et al., 2005; Northrup, Howerth, et al., 2005; Robbins et al., 1995; Scholten et al., 2004; Thomas et al., 1983).

To measure inter-rater agreement, a common study design involves a set group of observers tasked with evaluating the same set of clinical biopsies and a separate third-party coder comparing their diagnoses of each slide. Intra-rater agreement is measured in a test-retest design where individual observers evaluate the same clinical biopsies at two separate time points with sufficient time between evaluations to control for any slide recognition. These types of studies may or may not involve blinding where the same observer is unaware that subsequent biopsies are repeat observations. Similar to how histopathologic systems have evolved over the years, the statistical methods used to communicate the magnitude of both inter- and intra-rater agreement observed in these studies have also changed.

1.9.1 Methods of Measuring Agreement: The Evolution of Kappa Statistics

The subject of inter-rater and intra-rater agreement first became popular in respect to testing the repeatability of psychiatric diagnoses and public survey interview coding (Light, 1971). Initially, statistics involving testing association such as Pearson's Chi-square test was used to measure the agreement between observers. However multiple statisticians recognized that a strong association between observer answers does not necessarily mean strong agreement. For example, if two observers had perfectly opposing diagnoses for a given set of biopsies, there would be a strong negative association between their grading distributions however there would be little to no agreement. Therefore, agreement was recognized as a 'special case of association' and statistics such as Pearson's Chi-square were deemed inappropriate for its measurement (Cohen, 1960; Light, 1971).

Another early popular method for measuring agreement was percent agreement. In the simplest method, where only two raters are involved and there are only two categories or decisions to be made, the observations can be coded as 0's and 1's with a column added to calculate the difference between the two raters' observations, leaving either a difference of 1 for opposing observations or a 0 value where the two raters agree. The number of instances a 0 occurs in this category is then counted and divided by the total number of observations and multiplied by 100 to produce a percent agreement that reflects the proportion of agreement between the two raters. The percent of observations where the raters disagree can then be calculated by subtracting the percent agreement from 1. While this method provides a relatively easy way of measuring agreement, especially for numerical data, there is a fundamental flaw in this method of analysis when examining nominal data. For any given observation, there is a chance that either observer could simply guess the same diagnosis or categorization as the other observer. Without accounting for the amount of agreement that could be due to chance, percent agreement could significantly overestimate the true agreement occurring between two observers (Cohen, 1960; Light, 1971; McHugh, 2012).

Based on this knowledge, several different statistical approaches were developed to try and account for the proportion of agreement that could be expected between observers due to chance alone (Cartwright, 1956; W. Scott, 1955). However the most popular and widely used statistical method for measuring agreement was developed in 1960, Cohen's kappa coefficient (Cohen, 1960).

Cohen's kappa coefficient offered a way to measure the observed proportion of agreement against the expected proportion of agreement based on the nominal data while accounting for any random agreement due to chance alone, resulting in a coefficient that ranged between -1.0 and +1.0 where -1.0 indicated perfect disagreement, 0 indicated random agreement, and +1.0 indicated perfect agreement. As with any statistical test, Cohen's kappa involved meeting certain assumptions. First, the units or observations made were independent of each other. Second, the categories of the nominal scale being tested were independent and mutually exclusive. Third, the observers were assumed to independent of one another (Cohen, 1960). Additionally, Cohen specified a list of factors to consider when deciding to use the kappa coefficient and in its interpretation.

When calculating agreement, Cohen pointed out that the magnitude of agreement does not equate to the 'correctness' of certain observers (Cohen, 1960). In other words, just because two observers agree more than another observer pair does not mean that one pair is more accurate than the other. He also described that comparisons between observers should only be made when they are considered 'equally competent' to use the nominal scale examined (Cohen, 1960). There also must not be any constraints on any observer that limits the distribution of observations they can assign. Finally, and most importantly for the consideration of histopathologic scoring systems, Cohen's kappa coefficient does not account for any ordered difference in categories of the scale examined. Therefore, Cohen's kappa is accurate when measuring the agreement concerning nominal scales, but when the scale is ordinal, such as a histopathologic scoring system like the Kenney-Doig system where a higher category is associated with a more severe disease state, Cohen's kappa coefficient is unable to account for the magnitude of disagreement between differentially weighted categories (Cohen, 1960). While it is possible to still use Cohen's kappa on ordinal scales, the exact agreement may be over or under-estimated depending on how many categories are involved in the scale examined and by how many categories each observation differs by between observers.

To account for such ordinal scales, Cohen augmented the formula for the kappa coefficient to produce a weighted version of kappa (Cohen, 1968). This weighted kappa allows researchers to award partial agreement to observations that may only differ by one or two categories, while assigning little to no agreement for observations that differ by a higher number of categories. For example, if two observers were to assign the same biopsy a category I and II respectively, this would produce higher levels of partial agreement than if they assigned the biopsy a category I and III. This is accomplished by weighting how far the observations deviate off perfect agreement by a specified amount and incorporating those deviations into the kappa formula. In this way, agreement measured by weighted kappa is considered more accurate than the original Cohen's kappa coefficient when evaluating the observer agreement for an ordinal scale.

A particularly important limitation of both the unweighted and weighted kappa coefficients is that they can only be used to measure agreement between a pair of observers. As inter-rater agreement studies became more popular, there was a clear need to be able to measure

agreement across more than two observers. Several derivations of kappa have been used to accomplish this. Possibly the simplest method is that of Light, where pairwise kappa values are computed for each observer pair in the study and then the arithmetic mean of these kappa values is calculated to provide an 'average level of agreement' for the entire group (Light, 1971). Light's kappa offers a single value to describe the level of agreement produced across multiple raters and can be used for both the unweighted and weighted versions of kappa, providing a useful way to compare agreement across multiple studies without having to examine each pairwise kappa value. Together, Cohen's unweighted and weighted pairwise kappa statistics in addition to Light's average kappa provide a comprehensive method to measure and compare levels of agreement produced by various nominal and ordinal scales.

1.9.2 Methods of Measuring Agreement: Interpretation of Kappa Statistics

With the growing popularity of kappa statistics, it became clear that some sort of standard was needed to enable researchers to compare their kappa values and interpret these values in the context of relative strength of agreement as they ranged between 0 and +1.0. Landis and Koch suggested a breakdown of the possible kappa values with associated descriptors measuring the strength of agreement: a kappa value of <0.00 is poor agreement, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement (Landis & Koch, 1977). While the authors admitted these benchmarks may be arbitrary, their suggestions have become the common standard for interpreting both the unweighted and weighted kappa coefficients. However, despite the interpretation the cut-offs offer, there is still ambiguity in the literature regarding what kappa threshold should be regarded as an acceptable standard of agreement for the relative industry. Some researchers argue that an inter-rater agreement of <0.60 is unacceptable given the amount of disagreement assumed with kappa coefficients below that value (McHugh, 2012).

In a study using an analysis of histopathologic changes, there does not appear to be a consensus on which kappa values or interpretations of agreement are deemed 'too low' for any given scoring system, making said system unacceptable for routine diagnostic use. Therefore, to accurately interpret the reproducibility of a given histopathologic scoring system, multiple studies

are required to enable the comparison of kappa values produced by different observer groups and comparison of these kappa values to those found in other accepted histopathologic scoring systems is required. Ultimately, the observed level of agreement produced by a scoring system must be put into context of the accepted level of agreement found in other systems that are commonly used and accepted by histopathologic laboratories.

1.9.3 Methods of Measuring Agreement: The Evolution of Intraclass Correlation Coefficients

Another common method of measuring agreement is the intraclass correlation coefficient (ICC) first introduced by Fisher as a modification of the Pearson Chi-Square test (Koo & Li, 2016). Since then, the method behind ICCs have changed with the calculation now done by mean squares obtained through analysis of variance (Koo & Li, 2016). Similar to the weighted kappa statistic, ICCs can be used for measuring agreement using ordinal scales by incorporating the magnitude of agreement when categories are differentially weighted. Where weighted kappa and ICCs differ, however, are the many variants of ICC available depending on a variety of factors.

Through the work of Shrout, Fleiss, McGraw, and Wong, there are 10 different ICC forms that can be used to measure different aspects of inter- and intra-rater agreement based on the study design and research question (McGraw & Wong, 1996; Shrout & Fleiss, 1979). First, the researcher must specify the ICC as either a one-way or two-way effects model, where a one-way is appropriate for different subsets of raters and subjects and a two-way is used in a fully crossed design where the same raters evaluate the same subjects. Second, the researcher must decide if they are interested in measuring absolute agreement where the observers' answers must be identical, or if they are interested in the consistency or pattern of agreement observed. Third, whether the ratings were provided by several coders or a single coder must be specified. Finally, the researcher must decide between a random or mixed effects model. Random effects models are meant for randomly sampled coders whose observations can be extrapolated to a general population whereas a mixed effects model is for those studies where observers were not randomly selected or whose observations cannot be generalized to the population. Based on these specifications, ICCs offer a variety of different ways to evaluate observer agreement.

Of particular note, ICCs allow for specifying whether there is a single coder or a group of coders, which is important when considering the subject of intra-rater agreement. A downfall of both versions of Cohen's kappa when used for measuring intra-rater agreement is its inherent assumption that the raters are independent of each other. When measuring agreement in a test-retest fashion on the same observer, there is no independence. Hence, the percent of agreement that could occur due to chance is lower when considering an individual making repeat observations compared to two separate individuals making single observations to compare to one another. ICCs do not make this assumption of rater independence and offer alternate forms to specifically account for a single observer.

1.9.4 Methods of Measuring Agreement: Interpretation of Intraclass Correlation Coefficients

Since there are various forms of ICCs available for use, it is important for researchers to specify which particular version of ICC was used in assessing agreement to ensure that the proper choice was made based on the researchers' study design (Hallgren, 2012; Koo & Li, 2016). Additionally, there are various cut offs used for interpreting the strength of agreement based on the ICC value, including using similar benchmarks as for kappa statistics. While a two-way mixed, single-measures, consistency ICC has been shown to be equivalent to a weighted kappa statistics calculated using quadratic weights, other forms of ICCs are not and may not be comparable to the standard interpretations of agreement often used with kappa values (Hallgren, 2012). Instead, Koo and Li suggest interpreting ICC values using the following cut-offs: ICC value <0.50 is poor agreement, 0.50 – 0.75 is moderate, 0.75 – 0.90 is good, and >0.90 is considered excellent agreement (Koo & Li, 2016).

Similar to kappa statistics, there is no recognized gold standard for which ICC values or levels of agreement are deemed acceptable for quality control in diagnostic histopathology. Again, researchers are left with exploring the literature for other histopathologic inter- and intra-rater agreement studies to look for trends and interpret their own findings in the context of industry standards in order to draw conclusions as to whether the histopathologic scoring system in question is adequately repeatable.

1.9.5 Observations and Potential Reasons for Disagreement in Histopathology

When there is evidence of low or suboptimal agreement in a histopathologic system, there are a variety of reasons that could be contributing to this disagreement. First, one must consider the actual study design and whether there was an appropriate sample size and selection of statistical analysis appropriate for the type of observer agreement examined. Barring any inconsistencies in actual study design, there are three main factors that could contribute to troubles with inter-rater agreement: the scoring system used, the tissue samples examined, and the observers themselves.

Histopathologic scoring systems are built by categorizing biopsies into grades or categories based on guidelines that describe aspects of histologic features present in a given sample. One problem that has been identified in the literature is that these categories may be creating discreet categories for histologic phenomena that follow a more continuous spectrum, therefore the lines of division outlined by the scoring system may be interpreted differently based on the progressive nature of certain histopathology (Cross, 1998; de Vet et al., 1990; Thomas et al., 1983). Additionally, the criteria used to distinguish between these categories can be vague and based on subjective evaluation rather than more objective and quantitative measures (Mosli et al., 2015; Northrup, Howerth, et al., 2005; Thomas et al., 1983). For example, what one observer may deem as mild inflammation might be considered moderate inflammation to another observer. It has also been suggested that scoring systems involving the summation of multiple different types of histopathology into a single category may also increase observer variation as the integration of multiple factors increases the chance of error, a phenomena observed in the grading of chronic hepatitis using the Histological Activity Index scoring system (Desmet et al., 1994; Ishak et al., 1995; McHugh, 2012; Scholten et al., 2004).

Aside from the guidelines for categorization provided by histopathologic scoring systems, the biopsies available for grading can also influence observer variation. Even if the same biopsy is viewed by multiple raters, there can be enough heterogeneity within the tissue that one observer may not agree with another's overarching diagnosis (Geboes, 2000; Northrup, Howerth, et al., 2005). In other words, if an overall mildly affected biopsy has a relatively small area with severe changes, some observers may choose to still categorize the biopsy as a lesser, milder grade while other observers may feel that the biopsy belongs in a more severe grade. The overall

quality of the biopsy can also affect observer variation as artifacts such as shearing, sloughed epithelium, damaged endothelium and hemorrhage may be confused by some observers as actual pathologic change (Geboes, 2000; Mosli et al., 2015). Ultimately, characteristics involved with the actual biology of the tissue, the artifact produced by obtaining the biopsy, the quality of the slide during routine preparation and staining, or even the exact tissue slice used for a glass slide from a given tissue sample block could impact observer variation. Specific to the Kenney-Doig system, missing clinical submission history can also affect the grade assigned to an endometrial biopsy. If the mare's history of barrenness is unknown, or omitted during submission, that can affect how the biopsy is categorized. If the mare's stage of sexual cycle at the time of the biopsy procedure is not known or omitted, this can also lead to differences in interpretation of the significance of certain pathologic change such as mild inflammation.

Finally, there are inherent observer traits that can also contribute to differences in agreement when using histopathologic scoring systems. Differing levels of expertise among observers could affect inter-rater agreement, especially between observers of lower levels of experience and those that may be highly specialized in certain areas of histopathology (Fadare et al., 2013; Malpica et al., 2005). This point is reinforced by Cohen's reminder during the use of kappa statistics that observers should be 'equally competent' in rating subjects to produce the most accurate measurement of inter-rater reliability (Cohen, 1960). Aside from differences in expertise and caseloads, there may be psychological biases affecting inter-rater agreement of certain histopathologic scales. A number of studies have found that some observers gravitate towards certain ends of scoring systems, or alternatively, tend to lump their diagnoses in the more ambiguous middle categories and avoid extreme ranges (Cross, 1998; Kiupel et al., 2011; Thomas et al., 1983). Aside from personal bias, there may also be professional differences in opinion when using systems that involve the integration of several distinct lesions into a single diagnostic grade (Brothwell et al., 2003; de Vet et al., 1990; Scholten et al., 2004). For example, some observers may argue that mild amounts of fibrosis are more significant than moderate to severe amounts of inflammation, resulting in different interpretations of scoring system guidelines and ultimately increasing observer variation in its use.

While intra-rater agreement is obviously not affected by differences between observers, issues involving the scoring system such as subjective guidelines with little references to discreet

and quantifiable pathologic change can also contribute to observer variability within the same individual at different time points. Inconsistent gradings by the same observer can also be affected by biopsies that appear 'on the fence' between two categories, or by lower quality slides with high amounts of artifact. Many of the same reasons for low inter-rater agreement may also affect the intra-rater agreement of any histopathologic scoring system.

1.9.6 Inter-rater and Intra-rater Agreement and the Kenney-Doig Scale

Based on the previously mentioned factors that have contributed to high observer variability in other histopathologic scoring systems, and the concerns raised in the literature regarding the subjectivity of the Kenney-Doig scale guidelines, there is a possibility that certain categories within the Kenney-Doig scale may be acting as a 'catch all' and as such, a potential source of inter-rater and intra-rater disagreement. Previous studies have reported anywhere from 43.8% to 93.2% of endometrial biopsies examined as within either of the two middle-ranked grades of sub-category IIA or IIB (Kabisch et al., 2019; Kilgenstein et al., 2015; Nambo et al., 2014; Ricketts & Alonso, 1991a; Schilling, 2017; Waelchli, 1990). Ultimately, there is a chance that different pathologists may not agree on what constitutes a category IIA endometrium versus a category IIB, or that the same pathologist may not agree with their own diagnosis of a given biopsy at two different time points. If this is the case, how reliable can the routine histopathology grading of equine endometrial biopsies be if pathologists cannot agree? Given the prognostic significance associated with each discreet category, a difference of even one Kenney-Doig grade could considerably alter client or referring clinician decisions regarding the mare's future. An investigation is warranted as to whether pathologists across North America are consistently ranking a given biopsy into the same Kenney-Doig category.

While the majority of recent research has been directed towards finding new histologic characteristics that influence mare fertility and exploring the potential pathogenesis of these processes, the foundational framework of the Kenney-Doig scale remains untested. The system must be assessed concerning repeatability, measuring both inter- and intra-observer agreement using acceptable statistical analysis such as kappa and ICC statistics to allow for proper comparison of observer agreement found in other histopathologic scoring systems. In this way,

an evidence-informed decision concerning whether or not the Kenney-Doig scale produces acceptable levels of observer variability may be made.

1.10 Objectives: Assessing the Use and Repeatability of the Kenney-Doig Scale

Given the histopathologic experience of the pathologists here at the Western College of Veterinary Medicine Veterinary Pathology department, we set out to tackle the issue of intra- and inter-pathologist agreement concerning the Kenney-Doig scale. The study was broken down into two main goals.

1. To describe the past trend in endometrial biopsy grading by conducting a retrospective analysis of the distribution of grades given by pathologists over the past twenty years from the Western College of Veterinary Medicine and Prairie Diagnostic Services.
2. To evaluate the inter- and intra-rater agreement of the Kenney-Doig scale by comparing the grading of the same set of equine endometrial biopsies between pathologists and within pathologists at two separate time points using kappa and ICC statistics. Additionally, to investigate the inter- and intra-agreement of certain histologic features used to define Kenney-Doig categories and determine if some of these histologic features are more useful at predicting the final Kenney-Doig grade assigned.

CHAPTER 2. Retrospective Review: Grading Tendencies of Pathologists at the Western College of Veterinary Medicine and Prairie Diagnostic Services When Using the Kenney-Doig Scale

2.1 Abstract

The Kenney-Doig scale is a histopathologic grading system used as the international standard for assessing endometrial pathology and communicating prognostic fertility information for equine breeding prospects. The descriptive modifiers used for the scale are potentially subjective and as such they may not produce repeatable results between different observers and may contribute to certain grades of the scale acting as a 'catch all'. To investigate the frequencies of Kenney-Doig grades assigned at the Western College of Veterinary Medicine and Prairie Diagnostic Services (hereafter referred to as WCVM/PDS), a retrospective analysis of all equine endometrial submissions was completed from records between 1998 and 2018. Of 726 biopsies, the following grading distribution was found: 46/726 (6.3%) as category I, 307/726 (42.3%) as category IIA, 326/726 (44.9%) as category IIB, and 47/726 (6.5%) as category III. For comparison purposes, a retrospective review of the literature was completed and six different studies reporting Kenney-Doig grading distributions were included. Chi-square analysis showed significant differences between the grading distribution found at WCVM/PDS and each grading distribution reported in the six studies. To account for differences in mare populations, individual grading distributions were generated for five pathologists at the WCVM/PDS. Fisher's exact test between these five Kenney-Doig grading distributions revealed significant differences in grading tendencies, suggesting the presence of observer variation. This study suggests the need for prospective inter and intra-observer agreement studies investigating the repeatability of the Kenney-Doig scale.

2.2 Introduction

The Kenney-Doig scale for evaluating equine endometrial biopsies is built upon categorizing different histologic markers such as inflammation and fibrosis as either absent, mild, moderate, or severe. The severity of each histologic marker and the combination of these markers, in conjunction with the mare's age and reproductive history, determine which grade an endometrial biopsy will ultimately receive (R. Kenney & Doig, 1986). The grade and type of histologic changes present in the biopsy help guide therapy and economic decisions when it comes to breeding the mare in question.

Though the Kenney-Doig scale is still considered the industry standard for grading endometrial biopsies, criticisms involving the subjectivity of the category guidelines have surfaced throughout the literature (T. Evans et al., 1998; Ricketts & Alonso, 1991a, 1991b; Snider et al., 2011). The guidelines set by Kenney and Doig to determine whether the inflammation or fibrosis present in a biopsy qualifies as absent, mild, moderate, or severe are potentially vague and not specifically defined. In particular, the two middle categories of IIA and IIB involve a wide range of qualifying lesions with potential overlap between the two grades. Given the overlap between the IIA and IIB categories, and the wide prognostic value associated with each grade, these two middle ranks may act as a "catch all" for borderline biopsies. In other histopathologic grading systems, observers have been noted to predominantly grade within the middle categories of scales and avoid the extremes, possibly due to reluctance in upgrading a borderline biopsy to the next severe category (Cross, 1998; Kiupel et al., 2011; Northrup, Howerth, et al., 2005; Thomas et al., 1983).

To investigate a possible tendency of histopathologists to assign the middle grades of the Kenney-Doig system, a retrospective evaluation of the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS) equine endometrial biopsy submissions was performed to generate an institution-wide grading distribution curve. Individual grading frequencies were also generated from the top five contributing pathologists to investigate for significant differences in individual observer grading tendencies. A retrospective evaluation of the literature was also performed to attempt to shed light on what other institutions experience when it comes to Kenney-Doig grading distributions.

2.3 Materials and Methods

2.3.1 Retrospective Analysis of Endometrial Biopsies at the WCVM/PDS

Endometrial biopsy records were collected from the Veterinary Diagnostic Services software (1998-2014) and the Prairie Diagnostic Services Casebook 2 (2014 – 2018), both institute-wide computerized databases used for pathology submission record keeping at the WCVM/PDS, two affiliated diagnostic laboratories operating under the same department at the University of Saskatchewan, Canada. A free text search engine was used for the record history, final diagnosis, and comments for terms including "endometrial biopsy", "uterine biopsy", "endometrium", or "Kenney" and performed on all equine surgical biopsy submissions from 1998 to 2018. The signalment, submission history, final diagnosis including Kenney-Doig grade, histopathologic comments and the reporting pathologist were recorded. In rare cases, biopsies submitted in-hospital at the WCVM had incomplete histology reports and final biopsy grades were cross-referenced from the VetNet record keeping service at the WCVM Veterinary Teaching Hospital. The final biopsy grades were then plotted onto a bar graph to show the institution-wide distribution of grading of endometrial biopsies. Separate distribution curves were then generated for five individual pathologists who had contributed a minimum of 50 biopsy submissions over the allotted time period. A single coder was responsible for transcribing all data and submission history from the database software therefore the identities of the pathologists were not blinded from the coder. Letter identifiers were assigned to each pathologist after their total Kenney-Doig grades were recorded to maintain confidentiality (A, B, C, D, and E).

2.3.2 Retrospective Review of Kenney-Doig Grades Reported in the Literature

To collect a database of similar published material the following inclusion criteria were set: the study must have graded endometrial biopsies according to the scale proposed by Kenney and Doig, they must have reported the grading distribution in the publication (either in total counts or in frequency proportions), and the mare sample size must have been greater or equal to 150 in number (R. Kenney & Doig, 1986). If studies reported multiple grading distributions, such as comparing distributions based on the use of barren history or before and after uterine therapy,

the distribution that best represented the situation observed at the WCVM/PDS was used for more accurate comparison. If studies reported between grade diagnoses, for example a category I to IIA, these were excluded and only discreet Kenney-Doig categories were used. If the grading distributions were reported as raw count data, these were recorded, and if reported only as a percentage, the raw count data was calculated from the overall sample size and each category percentage. If the study did not report percentage frequencies for their Kenney-Doig distribution data, these were calculated from the raw counts and overall sample size, and used for descriptive comparison.

2.3.3 Statistical Analysis

Raw counts for the 5 individual WCVM pathologist grading distributions were uploaded into the statistical software R (R Core Team, http://www.R-project.org). An initial Fisher's exact test was performed on all 5 grading distributions, followed by separate Fisher's exact tests to compare each pathologist's grading distribution in a pairwise fashion. A Bonferroni adjustment was made to calculate a new p-value threshold for significance to adjust for the increased possibility of observing a significant difference due to chance with multiple comparisons (0.05/6 = acceptable p-value threshold < 0.008).

Similarly, raw counts for the grading distribution of each of the 6 studies identified in the literature, as well as the counts for the grading distribution found at the WCVM/PDS, were inputted into R. Chi-square analysis was done on all 7 grading distributions, followed by separate pairwise chi-square tests between each study's distribution and that found at the WCVM/PDS. Again, a Bonferroni adjusted p-value was calculated for an adjusted significance threshold to adjust for multiple comparisons and the possibility of increased error (0.05/10 = acceptable p-value threshold < 0.005) .

2.4 Results

A total of 755 biopsy submissions were recorded; however, 726 were included in the distribution curve analysis since 29 submissions had been assigned "between" Kenney-Doig

grades, for example, between a grade I and a grade IIA. A total of 726 equine endometrial biopsy submissions were included in the final distribution analysis of the WCVM/PDS database.

The 726 biopsies were obtained from mares ranging in age of 3 to 30 years of age with 85/726 submissions having an "unknown age" (Appendix B: Figure B.1). Biopsies were submitted for reasons including routine pre-breeding soundness exams, as part of pre-purchase examinations, in response to a documented abortion or early embryonic death, and to investigate various clinical signs or histories of sub-fertility. The majority of samples were obtained from light horse breeds, particularly Quarter Horses, Thoroughbreds, and varying Warmblood breeds. Other contributing breeds included various draft breeds mainly comprised of Clydesdales and Percherons as well as smaller breeds including ponies and Miniature Horses (Appendix B: Figure B.2).

Endometrial biopsy grades were distributed as follows: 46/726 (6.3%) as category I, 307/726 (42.3%) as category IIA, 326/726 (44.9%) as category IIB, and 47/726 (6.5%) as category III (Figure 2.1). The majority (87.2%) of the biopsies were graded as either of the two middle ranked categories IIA and IIB. In total, 48 different pathologists and 2 board-certified theriogenologists contributed to the grading of these submissions. While the WCVM/PDS is a teaching institute and junior pathology residents may have written initial histology reports for a proportion of submissions, all histology reports and assigned Kenney-Doig grades are reviewed and finalized by a practicing senior pathologist.

Five pathologists were identified that had graded at least 50 endometrial biopsies and individual distribution curves were generated for each of them (Figure 2.2). All five pathologists ranked the majority of their grades in one of the middle-ranked categories IIA and IIB. Notably, pathologist B and D displayed almost exact opposite grading distributions, with pathologist B assigning category IIA most frequently compared to pathologist D assigning category IIB as the predominant middle rank. The initial Fisher's exact test performed on all five grading distributions of Pathologists A through E revealed significant differences within the group (p-value <0.001). Additional Fisher's exact tests comparing the individual grading distributions of Pathologists A through E in a pairwise fashion found that all pathologists grading distributions were significantly different from each other (p-value < 0.001, with Bonferroni adjusted threshold at 0.005) except for Pathologist A and C (p-value = 0.006), and Pathologist A and E (p-value =

0.065) (Appendix B: Table B.1). Grading distributions were transformed into percentage frequencies for better visual comparison (Figure 2.3).

Six studies were identified from the literature that met the inclusion criteria (Kabisch et al., 2019; Kilgenstein et al., 2015; Nambo et al., 2014; Ricketts & Alonso, 1991a; Schilling, 2017; Waelchli, 1990). Grading distributions were transformed into percentage frequencies for better visual comparison (Figure 2.4). Four of the six studies had the majority of biopsies (>50%) graded within the two middle ranked categories, but with variable splits between whether IIA or IIB as the most commonly assigned grade. For example, Ricketts and Alonso assigned 87.2% of biopsies studied as a category IIB in contrast to Nambo et al. who assigned 70.8% of their biopsies to category IIA (Nambo et al., 2014; Ricketts & Alonso, 1991a). An initial chi-square analysis on all six Kenney-Doig grading distributions from the literature and that of WCVM/PDS found significant differences between grading trends with $X^2 = 1628.2$ (18) (p-value <0.001). Individual chi-square tests comparing each distribution from the literature against that at WCVM/PDS found that every identified Kenney-Doig grading distribution reported was significantly different from the distribution at WCVM/PDS (p-values <0.001, with Bonferroni adjusted threshold at 0.008) (Appendix B: Table B.2).

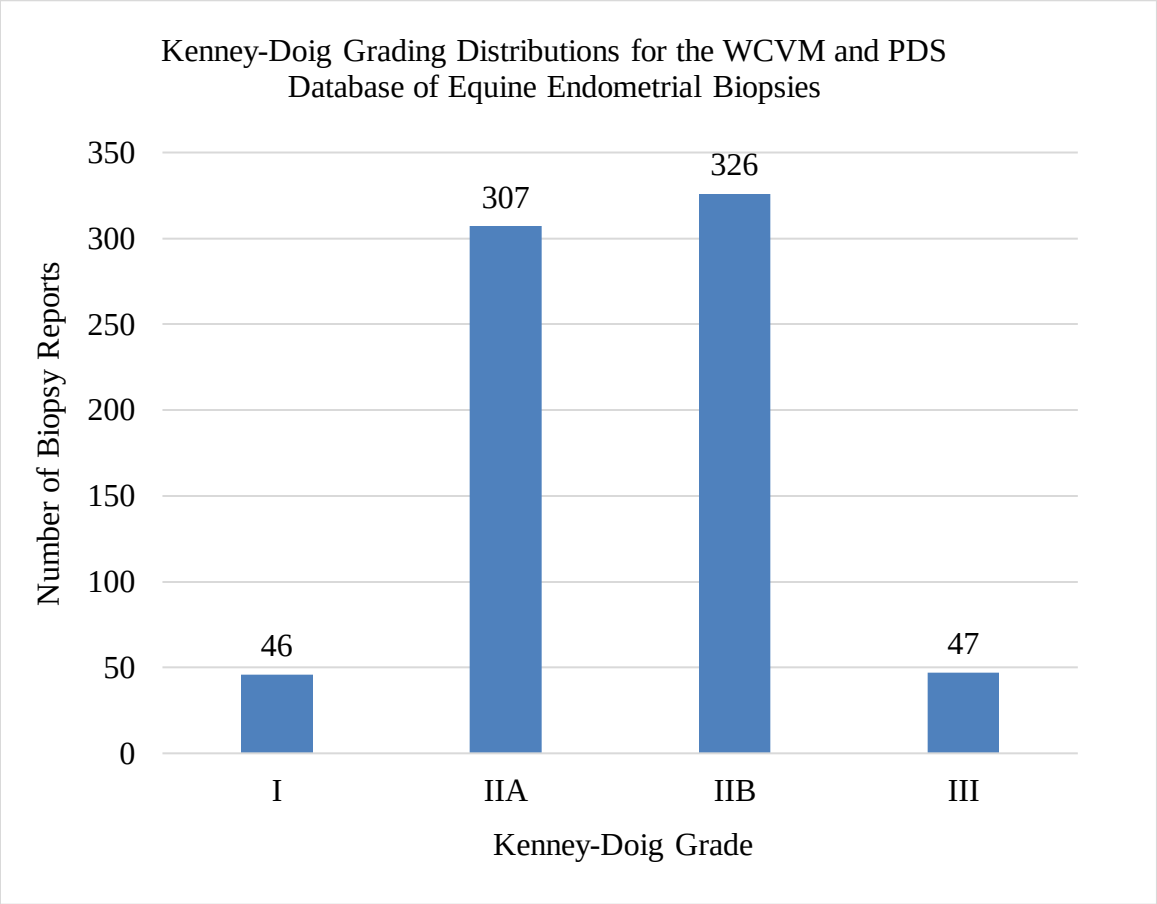

Figure 2.1. Graphical representation of the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS) institution-wide Kenney-Doig grading distribution of 726 equine endometrial biopsies submitted between 1998 and 2018.

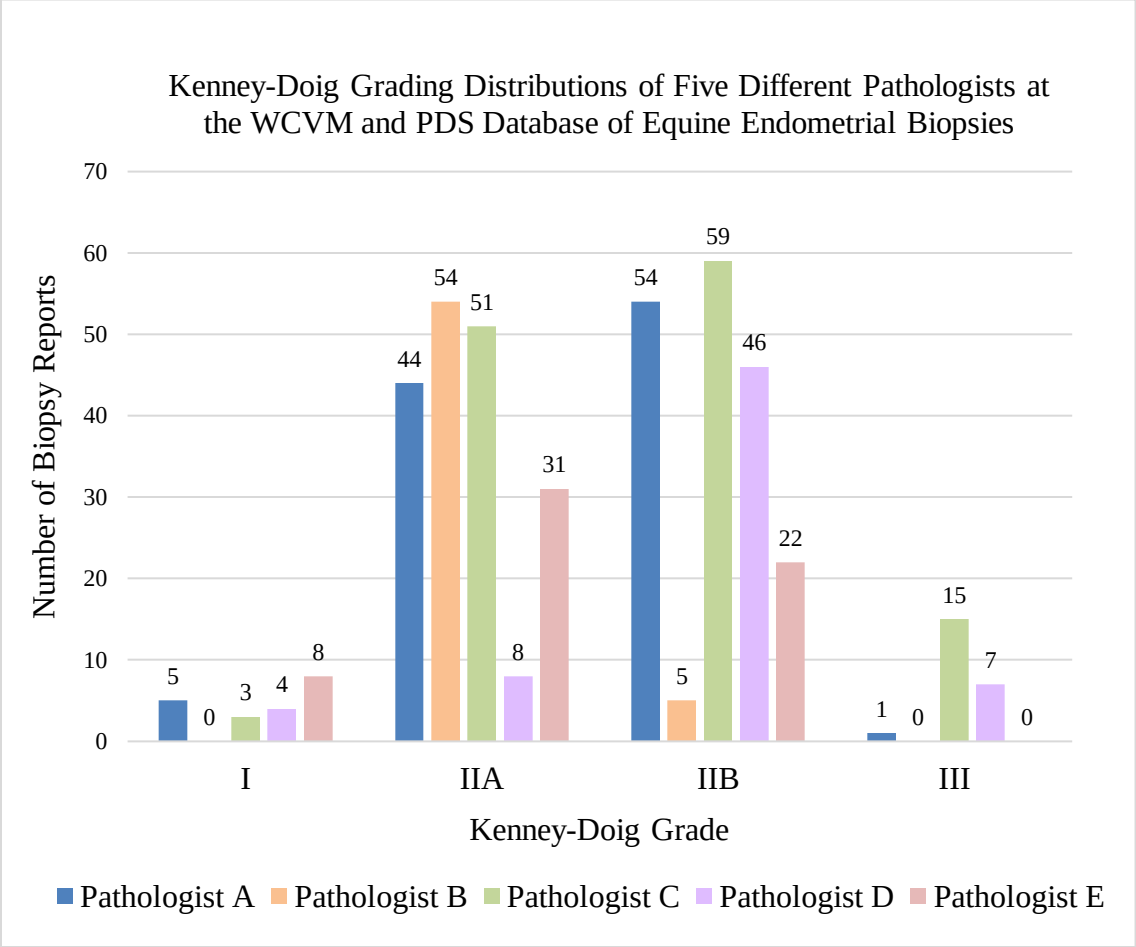

Figure 2.2. Graphical representations of Kenney-Doig grading distributions for five different pathologists at the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS). Criteria for inclusion of pathologists for analysis included grading a minimum of 50 equine endometrial biopsies within the period of study. Pathologist A: n = 104, Pathologist B: n = 59, Pathologist C: n = 128, Pathologist D: n = 65, Pathologist E: n = 61 where n refers to the total number of endometrial biopsy grades assigned.

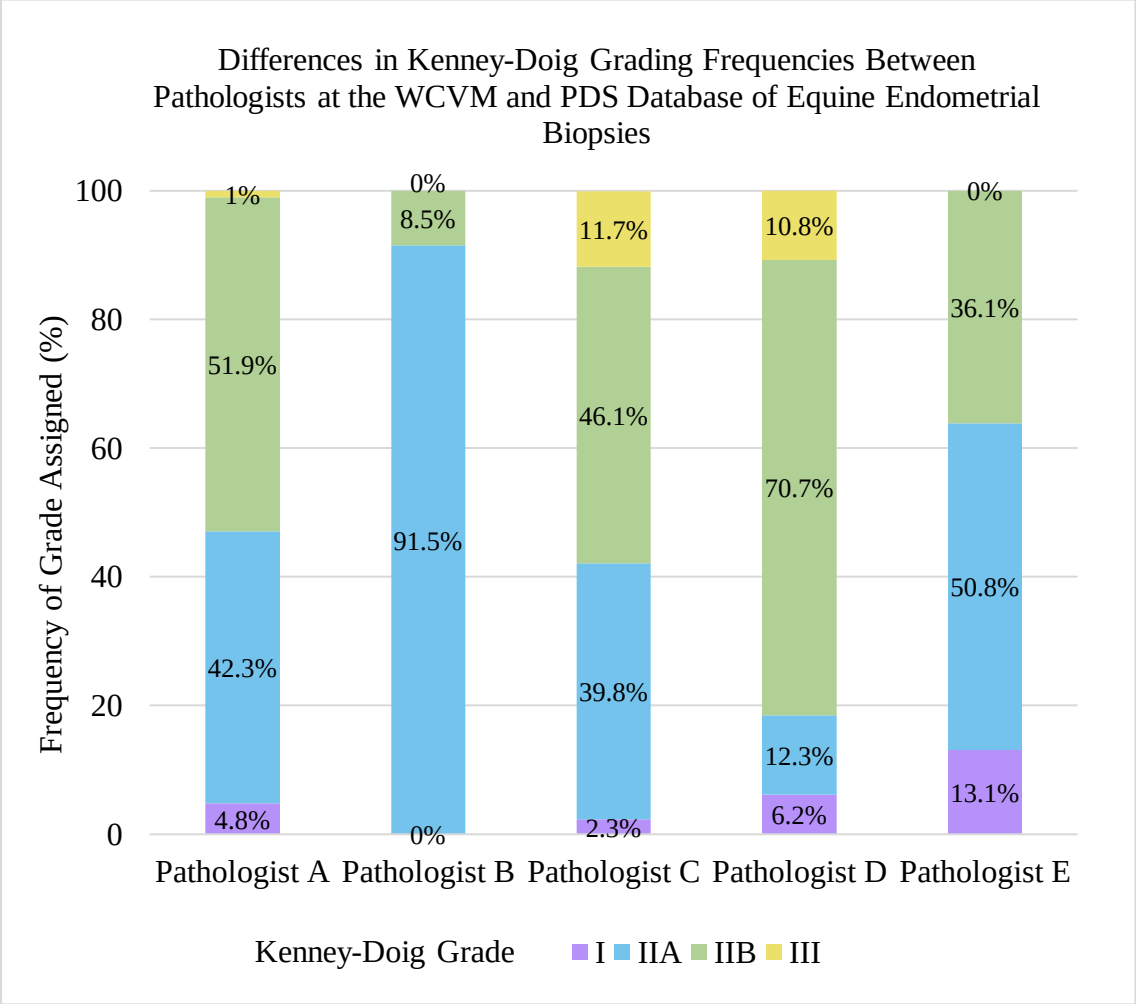

Figure 2.3. Differences in frequency distributions in Kenney-Doig grades assigned by five different pathologists to equine endometrial biopsies submitted to the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS) between 1998 and 2018. Criteria for inclusion of pathologists for analysis included grading a minimum of 50 equine endometrial biopsies within the period of study. Fisher's exact test revealed significant differences (p-value <0.001) between the grading distributions of Pathologist A vs B, A vs D, B vs C, B vs D, B vs E, C vs D, C vs E, and D vs E.

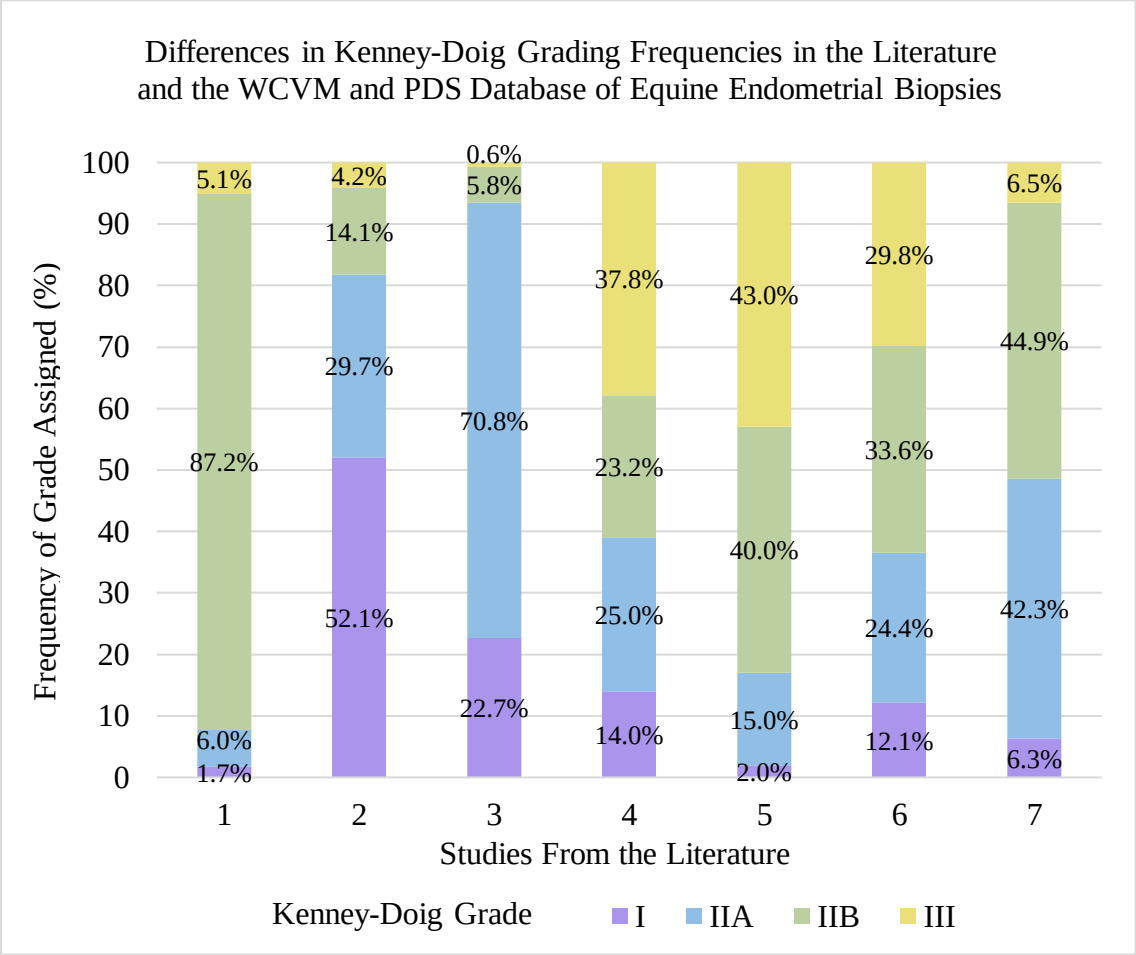

Figure 2.4. Graphical representation of frequency distributions describing Kenney-Doig grading in six studies found in the literature and that found at the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS). 1) Ricketts and Alonso (n = 530)(Ricketts & Alonso, 1991a), 2) Waelchli (n = 192)(Waelchli, 1990), 3) Nambo et al. (n = 154)(Nambo et al., 2014), 4) Kilgenstein et al. (n = 164)(Kilgenstein et al., 2015), 5) Kabisch et al. (n = 819)(Kabisch et al., 2019), 6) Schilling (n = 8795)(Schilling, 2017), **7)** WCVM/PDS (n = 726). Chi-square analysis revealed that all six studies had significantly different (p-value <0.001) Kenney-Doig distributions when compared to the WCVM/PDS individually. Criteria for inclusion of literature studies included use of the Kenney-Doig scale, documentation of grade distribution, and a mare population of n ≥ 150.

2.5 Discussion

The Kenney-Doig grading distribution generated for the WCVM/PDS equine endometrial biopsy database suggests that the categories IIA and IIB are predominantly assigned compared to the two extreme categories of I and III. For comparison purposes, six other studies were included from the literature that had graded at least 150 biopsies and used the Kenney-Doig scale for categorization and all were found to be significantly different from the grading distribution found at the WCVM/PDS. To be able to critically evaluate why these significant differences between grading distributions in the literature and the distribution at the WCVM/PDS exist, multiple variables that may affect the Kenney-Doig grade assigned to a given biopsy must be taken into account. Two overarching factors determine the Kenney-Doig grade: the presence and nature of endometrial pathology in the biopsy presented for grading evaluation and the accurate quantification of said pathology. To put it simply, the mare and the observer determine the Kenney-Doig grade, however a variety of confounding variables may influence the mare or the observer and ultimately the assigned grade.

When examining variables that may affect a particular mare and therefore the quality of her endometrium, certain factors must be discussed such as the age of the mare, reproductive history, breed, and performance use. Differences in these factors may affect the prevalence and severity of endometrial pathology present based on known associations between advancing age, increasing number of years barren, breed differences and high level athletic performance with higher prevalence of endometrosis and susceptibility to endometritis, both of which contribute to a more severe Kenney-Doig grade (T. Evans et al., 1998; Flores et al., 1995; Hanada et al., 2014; Hurtgen, 2006; Kabisch et al., 2019; R. Kenney & Doig, 1986; Kilgenstein et al., 2015; Ricketts & Alonso, 1991b; Schilling, 2017; Waelchli, 1990). These population demographics are therefore expected to significantly influence the Kenney-Doig grading distribution seen for any given sample of mares. A discussion regarding the differing demographics between the mare populations of each study found in this retrospective review must be had to extrapolate comparisons between the reported Kenney-Doig grading distributions in each study and that found at the WCVM/PDS.

It is well established that the prevalence of endometrosis increases as mares age (T. Evans et al., 1998; Flores et al., 1995; Hanada et al., 2014; Held & Rohrbach, 1991; Kabisch et al.,

2019; Kilgenstein et al., 2015; Ricketts & Alonso, 1991b; Waelchli, 1990). Mare populations that consist of predominantly older animals may therefore be expected to have a higher prevalence of endometrosis and more severe Kenney-Doig grades. The studies by Kabisch et al. and Schilling both reported Kenney-Doig distributions supporting this theory.

Kabisch et al. conducted a retrospective analysis on 816 biopsies submitted from exclusively mares that were 20 years of age or older (Kabisch et al., 2019). Without taking into account the history of barrenness, the majority of biopsies (83%) of mares 20 to 24 years of age were classified as either category IIB or III, with 43% of those placed in category III (Kabisch et al., 2019). In the second older age group of mares between 25 and 32 years of age, a similar distribution was found where 41% and 52% of biopsies were classified as IIB and III respectively, with the remaining 8% split evenly between the categories I and IIA (Kabisch et al., 2019). Schilling found a similar trend after analyzing 9120 Kenney-Doig grades assigned to endometrial biopsies of mares aged 1 to over 20 years of age, with 8795 of these biopsies falling within a single Kenney-Doig category (Schilling, 2017). After breaking these biopsies up into five different age groups, Schilling reported opposite Kenney-Doig grading distributions for the youngest and oldest groups of mares. The youngest group, 1 to 5 years of age, were mostly assigned in category I or between I and IIA, with very few assigned to category III. This is in stark contrast to the over 20 year old age group where the majority were assigned to category IIB or III and very few in category I or IIA (Schilling, 2017). Together, both studies provide strong evidence that the age range of a mare population must be taken into account when comparing two different Kenney-Doig distributions and may explain why both distributions had such a higher prevalence of category III mares compared to the WCVM/PDS.

In addition to accounting for differences in ages, the reproductive histories of the population must also be considered. If the majority of biopsies submitted to a particular facility are from clinically healthy mares for routine pre-breeding examinations, one may expect a higher prevalence of categories I and IIA to be assigned. In contrast, if the majority of samples are from mares with documented abortion, dystocia, or sub-fertility issues, a higher prevalence of the more severe categories IIB and III may be expected. In particular, the duration of barrenness might also influence the Kenney-Doig grade assigned.

There is ample evidence in the literature that the longer a mare remains barren, the higher the likelihood of endometrosis and issues with sub-fertility (Doig et al., 1981; Kabisch et al., 2019; R. Kenney & Doig, 1986; Lehmann et al., 2011; H. Schoon et al., 2000; Waelchli, 1990). Due to the significance of this finding, Kenney and Doig included barrenness as an important modifier when grading biopsies; a mare with a duration of barrenness greater than two years automatically moved the mare into the next higher category of endometrial pathology (R. Kenney & Doig, 1986). This modification however is dependent on the submitting clinician providing an accurate reproductive history to accompany the endometrial biopsy. Examination of submission histories for biopsies at the WCVM/PDS show a wide variation of reproductive histories provided, from no history given to detailed histories including the sexual cycle the mare was in at the time of sampling. This not only affects what grade the pathologist will assign, but ultimately the overall grading distribution of the database. The study by Kabisch et al. is an important example of how knowledge of a mare's barrenness influences the Kenney-Doig grade assigned.

While the retrospective analysis by Kabisch et al. focused on mares older than 20 years of age, it also reported two different Kenney-Doig distributions for those biopsies where the history of barrenness was known (Kabisch et al., 2019). This data reinforced two important concepts. First that only 76% of the submission histories included a duration of barrenness, suggesting that pathologists may commonly find themselves grading biopsies within the Kenney-Doig system without the proper modification of barrenness (Kabisch et al., 2019). Secondly, that including the duration of barrenness visibly affects the grading distribution. While the grading distribution without barrenness accounted for showed 15%, 40% and 43% in categories IIA, IIB and III respectively, after taking into account this vital reproductive history the distribution of categories IIA, IIB and III changed to 6%, 24% and 68% of submissions (Kabisch et al., 2019). In the field, however, clinicians are often faced with the difficult task of evaluating a mare with no known breeding history, and consequently pathologists will grade such a biopsy without being able to take barrenness into account. For the purposes of our study, the grading distribution that did not account for barrenness was used from Kabisch et al.'s study to compare against that at the WCVM/PDS, as complete breeding histories were not always available for biopsy submissions.

Without a complete reproductive history, other information such as whether or not the mare has been previously treated for endometrial conditions such as endometritis are not always available to the evaluating pathologist.

In the study by Ricketts and Alonso, 530 mares aged between 3 and 23 years old were used to assess any changes in Kenney-Doig grades in endometrial biopsies procured before and after targeted uterine therapy (Ricketts & Alonso, 1991a). All mares were biopsied due to a history of sub-fertility or genital abnormality. Of the Kenney-Doig grades assigned to the pre-treatment group, 87% fell within category IIB which the authors attributed to the expected endometrial pathology of barren mares at the end of an unsuccessful breeding season (Ricketts & Alonso, 1991a). A unique aspect of this study was the repeat assessment of biopsies after targeted uterine therapy in those mares where treatment was indicated. Kenney-Doig grades assigned to this post-treatment group showed a more even distribution between the two middle categories, though the number of biopsies assigned to one of the two extreme ends of the scale remained small (Ricketts & Alonso, 1991a). These distributions support the idea that certain endometrial pathologies are treatable however it is difficult to extrapolate which treatment group more accurately reflects what should be expected for the submission population seen at the WCVM/PDS. The majority of submissions seen at the WCVM/PDS were not repeat biopsies before and after treatment, or this information was not indicated on the submission history. Therefore, we chose to compare the pre-treatment group from Ricketts and Alonso's study with the Kenney-Doig grading distribution at the WCVM/PDS, as the majority of mares submitted did not indicate previous targeted uterine therapy. The mares submitted to the WCVM/PDS were instead from varied reproductive backgrounds including routine pre-breeding biopsies that one may expect to have a less severe Kenney-Doig grade than submissions based on a history of barrenness or sub-fertility like those used by Ricketts and Alonso. While the age range of the mares used in this study is similar to those submitted to the WCVM/PDS, other differences such as the exclusive use of sub-fertile mares and targeted uterine therapy may explain the differences observed between the two grading distributions. It is, however, difficult to confirm if the mares in Ricketts and Alonso's study population were exclusively sub-fertile, as differences in breeding techniques and stallion semen quality can also be responsible for reduced conception rates and this was not addressed in the study (van Buiten et al., 2003). Therefore, the study populations between Ricketts and Alonso and that at WCVM/PDS may contain a more similar mix of mares

regarding fertility. Without controlling for stallion and breeding technique factors, comparing the relative fertility of either mare population is difficult.

Aside from age and reproductive history, a mare's breed may also influence the likelihood of endometrial disease. Throughout the years various studies have assessed the effects of inbreeding on reproductive performance. Whether a rare breed such as the small population of Black Forest Draught Horses in Germany, or an international breed with a larger gene pool such as Thoroughbreds, many studies agree that inbreeding has a negligible effect on foaling rates and is probably negated by artificial selection imposed by breeders (Cothran et al., 1984; Mahon & Cunningham, 1982; Müller-Unterberg et al., 2017). Additionally, other researchers that have conversely reported decreased conception rates due to inbreeding, have suggested that the main culprit lies in reduced stallion fertility, and not in the mare population (Dini et al., 2020). Regardless of the conflicting evidence regarding inbreeding, differences in breed may not be expected to significantly influence Kenney-Doig grading distributions in specific mare populations based on simple physiologic differences between different gene pools. There is, however, a unique variable seen between equine breeds that does not involve fundamental genetic differences in fertility; the exclusive career uses in the equine industry for specific breeds and the associated management. The most common example of this concept is the Thoroughbred racing industry.

While the Thoroughbred breed may not have any genetic advantages or disadvantages when it comes to reproductive fertility, many Thoroughbreds are subjected to high levels of athletic performance early in life. Rigorous training schedules result in lean body conditions, depleting fat reserves throughout the body that can result in sinking and sloping of the vulva and poor perineal conformation (Hurtgen, 2006). This compromises the vulvar seal and predisposes these mares to vaginal and uterine contamination leading to conditions such as pneumovagina and endometritis. Additionally, Thoroughbreds may be selected for a flatter pelvic conformation to optimize stride length and speed, which also contributes to poor perineal conformation (Back & Clayton, 2013). Ultimately, racing Thoroughbreds are commonly affected by poor perineal conformation and mares often undergo vulvoplasties in attempts to decrease air and fecal contamination of the reproductive tract. Intense exercise and resulting exhaustion can also cause the vulvar seal to fail, causing even mares with good perineal conformation to have

pneumovagina. It may be argued that any breed with a commonly associated high-level sport, such as Thoroughbreds and flat track racing, Standardbreds and pacing, or Warmbloods and eventing, may have specific conformational changes associated with their athletic career that could influence the prevalence and severity of endometrial pathology present. With this in mind, comparisons between the Kenney-Doig grading distribution of a study such as that done by Nambo et al. to the WCVM/PDS must be done carefully.

Nambo et al. used exclusively Thoroughbred mares when categorizing 154 endometrial biopsies according to Kenney and Doig (Nambo et al., 2014). While it was not specified whether these mares were currently racing or had previously retired off the track, one might assume that at least a portion of the population sampled had some form of previous racing career. This is in contrast to the breed demographics seen at the WCVM/PDS where the most common breed is the Quarterhorse followed by a mixture of lighter horse breeds such as Thoroughbreds and Arabians and various heavier draft breeds. When comparing the Kenney-Doig grading distribution found by Nambo et al. where the majority of samples were category IIA (70.8%) to that of the WCVM/PDS where category IIA only comprised 42.3%, this difference in breed representation must be considered as it may affect the prevalence and quality of endometrial disease seen (Nambo et al., 2014).

Regardless of breed, the level of physical exercise and type of training regime a mare is engaged in may also affect the prevalence of endometrial disease. Any mare, regardless of breed, may experience loss of perineal fat reserves in response to higher levels of training resulting in poor vulvar conformation and increased risk of endometritis (Hurtgen, 2006). Exogenous progesterone supplementation, often used on performance mares to suppress estrus behaviour during the show season, has also been associated with higher susceptibility to endometritis and possible hormonal dysfunction affecting the endometrium (Burger et al., 2008; M. Evans et al., 1986; Kilgenstein et al., 2015). Performance mares also may contend with higher levels of both psychological and physical stress, higher body temperatures for prolonged periods of time during exercise, and higher prevalence of corticosteroid use for joint maintenance than the average pleasure horse, all of which may contribute to altered hormonal status and changes to the endometrial glandular function (Burger et al., 2008; Kilgenstein et al., 2015). In the study by Kilgenstein et al., endometrial biopsies were examined from a group of exclusively retired

performance horses (Kilgenstein et al., 2015). While the prevalence of endometritis and endometrosis appeared similar to other studies cited using pleasure mares, Kilgenstein et al. found a much higher prevalence of endometrial maldifferentiation in these sport mares. (Kilgenstein et al., 2015; H. Schoon et al., 1997, 1999). Endometrial maldifferentiation is a phenomenon characterized by an irregular pattern of differentiation in endometrial glandular epithelium postulated to affect glandular secretions and overall fertility (H. Schoon et al., 1997, 1999). While the Kenney-Doig scale does not include endometrial maldifferentiation, it does take into account inflammatory lesions such as those caused by endometritis and generalized non-physiologic glandular atrophy, both changes that performance mares may be more prone to developing (Burger et al., 2008; M. Evans et al., 1986; Hurtgen, 2006). Studies involving exclusive use of performance mares are lacking in the literature, making it unclear whether high level athleticism may exhibit significant effects on Kenney-Doig grades as opposed to populations of pleasure or mares. In sum, drawing parallels between the mare population of Kilgenstein et al.'s study and that of the WCVM/PDS may be confounded by the effects of strenuous exercise and competition.

After accounting for all the different variables within mare populations, one study in the literature most closely resembles the mare population found at WCVM/PDS. The study by Waelchli included 192 mares of ages between 4 and 24 years of age, breeds including European Warmbloods, Thoroughbreds, light draft types, Standardbreds and Arabians, and had mixed reproductive histories including barren, maiden, postabortion, lactating, previous foaling history but left open for the previous season, and unknown (Waelchli, 1990). Mares were from a mixture of farms with different breeding management conditions. While these demographics arguably best reflect the mare population seen at the WCVM/PDS, the Kenney-Doig grading distributions reported for both populations were still significantly different. Waelchli reported the majority (52.1%) of biopsies as classified as category I, almost ten times the 6.3% of category I biopsies classified at the WCVM/PDS (Waelchli, 1990). Though Waelchli had a similar prevalence of category IIA and III biopsies, the study only reported 14.1% as category IIB, much less than the 44.9% of category IIB grades assigned at the WCVM/PDS (Waelchli, 1990). While the previously mentioned studies had obvious differences in mare demographics that may explain the differences shown in Kenney-Doig grading distributions, the mare population in Waelchli's study contained a similar heterogenous mix regarding age, breed, use, and reproductive history

compared to the WCVM/PDS population and still had a significantly different Kenney-Doig grade distribution. An important factor remains that has yet to be discussed; the observer or pathologist responsible for assigning the Kenney-Doig grades.

As previously mentioned, the Kenney-Doig grading distribution found at the WCVM/PDS has a higher proportion of the two middle categories of IIA and IIB. Aside from the recently discussed demographic variation within the mare populations themselves, other factors that may influence the Kenney-Doig grading distribution include observer bias and variation. When using histopathologic grading scales, it has been shown that observers often gravitate towards assigning middle ranks (Cross, 1998; Kiupel et al., 2011; Northrup, Howerth, et al., 2005; Thomas et al., 1983). This may occur due to differing views among observers when it comes to the determining the threshold between histopathologic grades, hesitation to increase the grade severity when only a small portion of the biopsy is affected, or subjectivity involving the characteristics of the grading system on the whole (Brothwell et al., 2003; Kiupel et al., 2011; Northrup, Howerth, et al., 2005). How pathologists grade a biopsy that displays heterogenous changes that are sparse and inconsistent may also affect how they grade, or even whether they examine the entirety of every tissue sample present on the slide, or multiple sections from the same biopsy sample. Other sources of observer variation may stem from differences in training and mentorship between pathologists, resulting in different interpretations or quantifications of certain endometrial pathologic features. Given the vague and subjective guidelines of the Kenney-Doig system, particularly of the middle two categories, observer bias and variation may contribute to the higher proportion of categories IIA and IIB seen at the WCVM/PDS. To investigate this, individual grading distributions from five contributing pathologists at the WCVM/PDS were generated and compared by Fisher's exact test, allowing for analysis of different grading tendencies within the same local mare population.

The individual grading distributions generated from the top five contributing pathologists in this retrospective review illustrate obvious differences in grading tendencies between most of the pathologists when compared in a pairwise fashion. All five pathologists assigned more of either one or both of the middle grades over the extreme ends of the scale with category I or III. This could be explained as a reflection of the population demographics each pathologist happened to evaluate, as previously discussed, or due to avoidance tendencies when it comes to utilizing either end of the scale, or due to blurred lines between what should differentiate a

category I or III from the middle IIA or IIB. When looking at the shapes of the distribution curves between pathologists, there are obvious differences in whether a category IIA or IIB are assigned. This is well illustrated particularly between Pathologist B and Pathologist D where the grading tendencies between category IIA and IIB appear to be in exact opposition to each other. This suggests some fundamental differences in regard to using the Kenney-Doig scale between observers.

The Kenney-Doig scale predominantly relies on subjective descriptors such as absent, mild, moderate or severe when quantifying endometrial pathology (R. Kenney & Doig, 1986). While more objective measurements exist to describe varying severities of periglandular fibrosis, the majority of histopathologic features depend on the observer's concept of what determines absent versus mild, or mild versus moderate and so on. Accurate use of the Kenney-Doig scale depends on individuals assigning the same descriptive modifier to given histologic changes and agreeing on the same final diagnostic category designation. The differences seen in this review between the five pathologists' Kenney-Doig grading tendencies begs the question, "Do they agree?"

Inter-rater agreement, whether separate individuals agree on a diagnosis, and intra-rater agreement, whether the same individual agrees on the diagnosis of a given biopsy at different time points, are fundamental to the reliability of a grading system. Without reproducibility between separate observers, clinicians cannot accurately interpret Kenney-Doig grades assigned by different diagnostic laboratories, or different pathologists within each laboratory. While inter-rater and intra-rater agreement studies have become popular in testing the reproducibility of histopathologic grading of certain types of cancer or degenerative conditions in both human and veterinary medicine, no studies exist on the observer agreement concerning the Kenney-Doig scale (Bergeron et al., 1999; Cross, 1998; de Vet et al., 1990, 1995; Fadare et al., 2013; Ishak et al., 1995; Kiupel et al., 2011; Matos et al., 2012; Morris, 1994; Munkedal et al., 2016; Northrup, Howerth, et al., 2005; Robbins et al., 1995; Scheuer, 1997; Scholten et al., 2004; Silcocks, 1983). While differences in mare populations make it difficult to draw conclusions between the differences in Kenney-Doig grading distributions of the six studies identified in the literature and that of WCVM/PDS, examining the overall grading tendencies at the WCVM/PDS and those of the top 5 contributing pathologists suggest observer variation may also play a factor and overall

subjectivity of the Kenney-Doig scale in general. The disproportionate assignment of category IIA and IIB endometrial biopsies at the WCVM/PDS, and the significant differences seen between individual pathologist grading tendencies, suggest that a study assessing the inter-rater and intra-rater agreement concerning the use of the Kenney-Doig scale is warranted.

2.6 Transition Statement

Inter- and intra-observer agreement studies have been used extensively to evaluate many histopathologic grading schemes in both human and veterinary medicine, however no studies exist testing the observer agreement of the Kenney-Doig scale (Bergeron et al., 1999; Cross, 1998; de Vet et al., 1990, 1995; Fadare et al., 2013; Ishak et al., 1995; Kiupel et al., 2011; Matos et al., 2012; Morris, 1994; Munkedal et al., 2016; Northrup, Howerth, et al., 2005; Robbins et al., 1995; Scheuer, 1997; Scholten et al., 2004; Silcocks, 1983). Given the distribution of category IIA and IIB endometrial biopsies at the WCVM/PDS, and the significant differences seen between individual pathologist grading tendencies, a prospective study assessing the inter-rater and intra-rater agreement concerning the use of the Kenney-Doig scale was completed.

CHAPTER 3. Prospective Study: Measuring Observer Variation When Grading Equine Endometrial Biopsies with the Kenney-Doig Scale

3.1 Abstract

Histopathologic grading scales exist in both human and veterinary medicine as diagnostic tools to aid in the prognostic evaluation of certain diseases and pathology. To be considered accurate and useful, these grading systems must be both valid and reliable. Inter and intra-rater variability has been identified in multiple systems and negatively affects the reliability of these scales. The Kenney-Doig grading scale, used to associate equine endometrial pathology with prognostic estimation of a mare's reproductive potential, has not been evaluated for inter or intra-rater variability in the literature. To assess whether the Kenney-Doig system produces reliable and consistent results between and within observers, eight American College of Veterinary certified pathologists were recruited to blindly grade the same set of 63 digitized equine endometrial biopsy slides as well as blindly re-evaluating 21/63 of these slides at a later time point. Cohen's kappa values for pairwise comparison of final Kenney-Doig grades ranged from -0.052 to 0.458 (unweighted) and 0.082 to 0.638 (weighted), with an average Light's kappa of 0.187 (unweighted) and 0.359 (weighted) across all eight pathologists, 0.143 (unweighted) and 0.330 (weighted) for pathologists practicing at different institutions, and 0.216 (unweighted) and 0.464 (weighted) for pathologists at the same institution. Intra-class correlations measuring intra-rater agreement ranged from 0.116 to 0.774 with an average of 0.553 for all eight pathologists. This study shows only slight to moderate inter-rater agreement and poor to good intra-rater agreement is produced by the Kenney-Doig scale, suggesting that the system may be unreliable and subject to significant observer variability.

3.2 Introduction

Many histopathologic scales exist in both human and veterinary medicine to correlate histologic evidence of certain disease processes with a given prognosis and to aid in guiding treatment. Histopathologic scales are particularly common in oncology where surgical biopsies of tumors are graded and associated with risks of metastasis and other prognostic traits. To be able to trust the results of these scales, validity and reproducibility studies are done to both validate the prognoses associated with each grade and to investigate whether significant observer variation exists between different pathologists using the scales (Cross, 1998). Prospective inter- and intra-observer agreement studies have been used throughout the years to assess the reproducibility of these systems and identify the need for grading guideline modification (Bergeron et al., 1999; Fadare et al., 2013; Ishak et al., 1995; Kiupel et al., 2011; Munkedal et al., 2016; Northrup, Howerth, et al., 2005; Robbins et al., 1995; Scholten et al., 2004). While several studies have focused on proving the prognostic validity of the Kenney-Doig scale, no studies exist that investigate the reproducibility of this system.

As previously discussed, retrospective evaluation of the Kenney-Doig grading distribution of biopsies submitted to the WCVM/PDS revealed a non-uniform spread of grades with a heavy lumping of biopsies within the two middle IIA and IIB categories. This bell-shaped curve may be expected in a mixed mare population where biopsies are being submitted from both young and old mares, as well as healthy and sub-fertile mares. This may result in the majority of mares being diagnosed with mild to moderate endometrial pathology and relatively few in the extreme categories. While the effect of mare population dynamics on grading distribution cannot be excluded, observer variation may also be contributing to the significantly higher prevalence of middle-ranked grades, as has been seen in other histopathologic grading systems in veterinary medicine (Kiupel et al., 2011; Northrup, Howerth, et al., 2005). Examination of the individual grading tendencies between five of the top contributing pathologists at the WCVM/PDS revealed significant differences in grade proportions, suggesting that observer bias and possible disagreement regarding use of the scale may be occurring.

The goal of this study was to investigate the inter and intra-rater agreement between and within eight different pathologists, as well as between observers practicing at different institutions and within the same institution, when grading the same blinded set of endometrial

biopsies using the Kenney-Doig scale. Grading scores and histologic feature evaluations were recorded for these slides and inter-observer agreement was examined using Cohen's kappa statistics, a popular choice for observer variation studies where the observed agreement is compared to the expected agreement due to chance and a number between -1 and 1 is reported, with -1 being perfect disagreement, 0 being no agreement and 1 being perfect agreement (Cross, 1996; Landis & Koch, 1977; Malpica et al., 2005; Silcocks, 1983). Intra-observer agreement was evaluated using intraclass correlation coefficients, a statistic similar to Cohen's kappa where the resulting coefficient is reported between -1 and +1, and is interpreted similarly to Cohen's kappa (Koo & Li, 2016). Logistic regression modelling was also used to investigate whether certain histologic features may have higher predictive value for a given Kenney-Doig category based on the collected data. Once identified, these histologic features with high predictive value were investigated for consistent assignment to a given category based on the additive nature of the Kenney-Doig criteria.

3.3 Materials and Methods

3.3.1 Slide Selection and Digitization

Sample size was determined using recommendations from Bujang and Baharum for unweighted kappa statistical analysis (Bujang & Baharum, 2017). A 4x4 table was selected for the four possible Kenney-Doig categories available, the power set at 90% with an alpha level of 0.05. Estimates of kappa coefficients based on the null hypothesis of no agreement ($K_1 = 0.0$) and the alternate hypothesis of expected moderate agreement ($K_2 = 0.40$) were set and used to select sample size. This resulted in a suggested minimum sample size of n = 25, however due to the unequal frequency distributions observed in the retrospective study and the possibility of participating pathologists' to grade in similar distributions, this number was doubled according to suggestions from Bujang and Baharum (Bujang & Baharum, 2017). Weighted kappa values were not taken into account for sample size as weighted kappa statistics are more sensitive at measuring agreement than unweighted and therefore require smaller sample sizes in general. Based on the proposed blocking of the surveys described below and the projected timeline for the

project, an overall inter-rater agreement sample size of 63 was set, with an additional 21 slides of previously graded slides to be used for intra-agreement.

Endometrial biopsy submissions from the Prairie Diagnostic Services Casebook 2 (2014 – 2018) and the Veterinary Diagnostic Services (1998-2014) databases were selected via random number generator (Microsoft Excel) to select 16 random biopsies from each of the four Kenney-Doig categories as assigned by the original overseeing PDS or WCVM pathologist, for a total of 64 slides. The original hematoxylin and eosin stained glass slide for each biopsy was pulled and examined for preliminary inclusion criteria: adequate amount of physical tissue in the sample and appropriate amount of histologic tissue layers (sufficient intact epithelium, stratum spongiosum, stratum compactum, and glandular density). If the original glass slide could not be located, another slide was prepared from the paraffin block containing the original surgical biopsy by the PDS histology laboratory using routine slide preparation and hematoxylin and eosin staining technique. Finally, all slides were screened for adequate staining intensity and recut as previously described if they did not meet appropriate standards.

All 64 slides were then digitally scanned using a 20X objective on the Olympus VS120 Virtual Microscope by the WCVM Imaging Suite**.** The resulting scans were uploaded to a locally maintained database for access via Olympus OlyVia software (OLYMPUS, 2020) for viewing.

Sixty-three of the 64 digitized slides were randomly selected to be used in the experiment, leaving a single Kenney-Doig category III slide excluded from use to maintain an even distribution of slides between three different administered groups of slides. This resulted in 16 slides of categories I through IIB and 15 slides of category III within the slide set. These slides were then randomly partitioned into three groups of 21 slides. Each group of 21 slides was administered at separate time points to decrease the total number of slides graded per sitting and control for fatigue bias. A fourth group of slides was also created by randomly selecting within each Kenney-Doig category from the 63 previously selected slides. This fourth group was used for testing intra-agreement of pathologists at different time points and contained five slides for each category I, IIB, and III and six slides for category IIA.

3.3.2 Survey Design and Implementation

To collect Kenney-Doig grading data, surveys were designed for each of the four blocks of slides using Survey Monkey software (San Mateo, California, USA) (http://www.surveymonkey.com). Each survey page included the same questions per slide (Appendix A: Table A.1). Pathologists were asked to evaluate the amount of inflammation, fibrosis, lymphatic lacunae, and glandular atrophy as either absent, mild, moderate, or severe, and to assign a final Kenney-Doig grade. Optional comment boxes were included for each slide for additional description if deemed necessary. The first survey also included some basic questions regarding the pathologist's experience with the Kenney-Doig scale.

Slides were embedded into each survey page with a hyperlink to the corresponding viewing window in the online OlyVIA server (Appendix A: Figure A.1). Pathologists were asked to log into the OlyVIA server at the beginning of each session in a separate tab and then access the corresponding slide to each survey page with the embedded hyperlink. In this fashion, pathologists were able to view and grade the same slides via Survey Monkey and OlyVIA. Pathologists were blinded to the original Kenney-Doig grade assigned to each slide by the WCVM/PDS database. Slides were also randomly ordered so as not to reflect the different proportions of Kenney-Doig grades included in the slide set (ie. 16 slides of category I followed by 5 slides of category IIB). Instead, each survey section consisted of 21 randomly selected slides from the entire 63 slide set. Pathologists were also unaware of any 'repeat' slides and were not told whether some slides were repeated later on in surveys.

Surveys were administered via email invitation over the course of 17 months, with appropriate breaks in between to control for fatigue bias and to accommodate scheduling conflicts. The fourth survey which consisted of a mixture of previously graded slides was administered at least two weeks after the last completed survey to control for slide recognition, with some slides having been graded over 12 months prior.

3.3.3 Pathologist Recruitment and Participation

In total, eight American College of Veterinary Pathology certified pathologists were recruited for the experiment. Five pathologists were from different academic or private institutions, four of these within Canada and the final one from the United States. Three additional pathologists were recruited from the same institution as one of the original five. This resulted in three observer groups: eight pathologists total, five from separate institutions (inter-institution group), and four from the same institution (intra-institution group).

Each pathologist was provided with a copy of the original Kenney-Doig paper to peruse if desired and a brief introductory statement to the study (R. Kenney & Doig, 1986). Pathologists were instructed to complete the surveys to the best of their abilities, focusing on only the histologic pathology to determine the overall Kenney-Doig grade as reproductive history and/or age of the mare was not provided for many of the submissions, therefore no clinical history was given for any of the slides.

3.3.4 Statistical Analysis

Individual observer scores for histologic features and Kenney-Doig grades for each slide were tabulated using Microsoft Excel (Microsoft 365 Subscription, version 16.38 20061401). This data was uploaded into R (R Core Team, 2020), and the additional "irr" package installed to calculate a variety of statistics for levels on inter-rater agreement and intra-rater agreement. Descriptive statistics for frequency distributions of assigned histologic evaluations and final Kenney-Doig grades were tabulated. Inter-rater agreement was evaluated using two different versions of Cohen's kappa, unweighted and weighted. Unweighted kappa values measure the amount of agreement without accounting for the number of categories pathologists may differ by, while weighted kappa values were used to look more specifically at the magnitude of agreement by placing more emphasis on a disagreement of more than one Kenney-Doig category and awarding partial agreement when differing by only one category. For example, weighted kappa values will measure lower agreement between two pathologists that grade the same biopsy a category I and a category III than an unweighted kappa, as the unweighted statistic does not differentiate between a disagreement of one versus two or more categories. Therefore, both

unweighted and weighted kappa values were used, with the weighted kappa value using a quadratic formula based on the "squared" option for computing weighted kappa statistics in R.

Unweighted and weighted kappa values were computed between each pathologist observer pair and the arithmetic mean of these kappa values were used to show an average agreement coefficient as described by Light for all eight pathologists, the inter-institution group, and the intra-institution group (Light, 1971). Unweighted Light's kappa values were computed using the Light's kappa function in R, while weighted Light's kappa values were computed manually using Microsoft Excel. Kappa statistics were interpreted using the standards suggested by Landis and Koch (Table 3.1) (Landis & Koch, 1977).

Percent agreement was also calculated for all eight pathologists. The same kappa values and percent agreement analysis was also used to determine inter-rater agreement concerning the evaluation of histologic features including inflammation, fibrosis, glandular atrophy, and lymphatic lacunae.

Intraclass correlation coefficients (ICC) and 95% confidence intervals were used to measure intra-rater agreement between each pathologist at two separate time points assessing the same 21 biopsy slides. ICCs were chosen as the statistic best representing intra-rater agreement as the two measurements made were related instead of independent; they were made by the same individual at two different time points which arguably involves less percent agreement due to chance than two individuals making separate observations at the same time, as is assumed by the kappa statistic.

ICCs and their 95% confidence intervals were calculated using the same R software and "irr" package previously mentioned, with the "model" set to two-way, "type" set to absolute agreement, and "unit" set to single comparisons to compare each individual grading attempt versus averaging them for consistency evaluation. ICCs were interpreted using the standards suggested by Koo and Li (Table 3.2) (Koo & Li, 2016).

This analysis was repeated for the intra-rater agreement of the evaluation of the different histologic features including inflammation, fibrosis, glandular atrophy, and lymphatic lacunae. The arithmetic mean was calculated for all ICC values using Microsoft Excel, demonstrating the

average intra-rater agreement for all eight pathologists concerning Kenney-Doig grades and histologic features.

Logistic regression modelling was used to assess the level of association of each histologic feature with the four different Kenney-Doig categories. Models and calculated predictive probabilities associated with each model were done using R (R Core Team, 2020). Three models were built to accommodate comparison of different Kenney-Doig categories as a binary outcome; category IIA and above compared to below; Category IIB and above compared to below; and category III to below. Mean predicted probabilities from the three logistic models for each individual predictor variable were calculated and a single table of values for each outcome category were created by combining the relevant three logistic model values. Due to low inter and intra-rate agreement, adjustment for pathologist or slide to account for clustering were not considered in these models.

Histologic features identified as having high predictive value were examined for usage concerning the additive nature of the Kenney-Doig scale. Observations made by all eight pathologists on the original 63 slides were counted using Microsoft Excel to explore how different descriptive modifiers were being categorized into final Kenney-Doig grades.

Table 3.1. Guidelines used for interpretation of Cohen's kappa coefficient (Landis & Koch, 1977).

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Table 3.2. Guidelines used for interpretation of intraclass correlation coefficient (Koo & Li, 2016).

Intraclass Correlation Coefficient	Strength of Reliability
<0.50	Poor
0.50 – 0.75	Moderate
0.75 – 0.90	Good
>0.90	Excellent

3.4 Results

3.4.1 Inter-rater Agreement of Kenney-Doig Grades

Frequency distributions for Kenney-Doig grades assigned to the same set of 63 endometrial biopsies revealed a wide variability in grading tendencies (Figure 3.1 and Appendix B: Table B.3). None of the eight participating pathologists' grading distributions approached the frequency distribution of 25.4% category I, 25.4% category IIA, 25.4% category IIB, and 23.8% category III that was designed based off the original diagnoses of the selected slide set from the PDS and WCVM database. While five of the eight pathologists frequently utilized category III, category I remained the most under-utilized category and the majority of slides (50.79% to 88.89%) were assigned to either category IIA or IIB. Ultimately, only 1/63 of the biopsy slides was assigned the same Kenney-Doig grade from all eight pathologists while 6/63 biopsy slides were assigned all four possible Kenney-Doig grades.

Unweighted Cohen's kappa coefficients for pairwise comparison between all eight pathologists revealed a range of -0.052 to 0.458 with an average Light's kappa coefficient of 0.187 for the entire group, indicating poor to moderate agreement among different pathologists with an average of only slight agreement (Table 3.3 and Appendix B: Table B.4). Weighted kappa coefficients for pairwise comparison between all eight pathologists revealed higher levels of agreement, with a range of 0.082 to 0.638 indicating slight to substantial agreement and an average Light's weighted kappa coefficient of 0.359 indicating overall fair agreement among raters (Table 3.3 and Appendix B: Table B.4). Low levels of agreement across the group were reflected in percent agreement calculations (Appendix B: Table B.5).

The inter-institution group of pathologists had unweighted kappa values ranging from -0.052 to 0.458 with an average unweighted Light's kappa of 0.143 (Table 3.3 and Appendix B: Table B.4). In comparison, the intra-institution group showed unweighted kappa values ranging from 0.120 to 0.333 with an average unweighted Light's kappa of 0.216 (Table 3.3 and Appendix B: Table B.4). Weighted kappa values for the inter-institution group ranged from 0.082 to 0.638 with an average of 0.330 while the intra-institution group ranged from 0.376 to 0.542 with an average of 0.464 (Table 3.3 and Appendix B: Table B.4). Overall, the intra-institution group

showed on average higher levels of agreement than those at different institutions, with fair agreement seen in unweighted calculations and moderate agreement in weighted calculations.

3.4.2 Inter-rater Agreement of Histologic Features

Frequency distributions for the use of the descriptive modifiers absent, mild, moderate or severe for all four histologic features including inflammation, fibrosis, glandular atrophy, and lymphatic lacunae were tabulated (Appendix B: Tables B.6 to B.9). Unweighted and weighted pairwise kappa values were also calculated (Appendix B: Tables B.10 to B.13). The average agreement for both unweighted and weighted kappa values were calculated and summarized across all eight pathologists, the inter-institution group, and the intra-institution group (Table 3.4). Percent agreement for each pathologist concerning the different histologic features were also calculated (Appendix B: Tables B.14 to B.17).

Across all eight pathologists, evaluation of inflammation showed the highest average unweighted and weighted kappa values at 0.183 and 0.386 respectively, indicating slight to fair agreement. Evaluation of glandular atrophy revealed the lowest average unweighted kappa value of 0.112 indicating slight agreement. However, the lowest average weighted kappa value was calculated from the evaluation of lymphatic lacunae at 0.237 indicating fair agreement. Findings were similar in the inter-institution group where inflammation had the highest levels of agreement at 0.150 and 0.359 for average unweighted and weighted kappa values and lymphatic lacunae showing the lowest levels of agreement at 0.065 and 0.229 for average unweighted and weighted kappa values. Consistent with the other groups, inflammation produced the highest average kappa values indicating fair to moderate agreement among the intra-institution group (0.283 unweighted kappa and 0.482 weighted kappa), while glandular atrophy and lymphatic lacunae produced the lowest average unweighted and weighted kappa values respectively, resulting in only slight to fair agreement among pathologists.

3.4.3 Intra-rater Agreement of Kenney-Doig Grades

Intra-rater agreement concerning Kenney-Doig grades assigned to the same set of 21 slides at two different time points was measured by calculating ICC values for each pathologist with accompanying 95% confidence intervals (Appendix B: Table B.18). ICC values ranged from 0.116 to 0.774, with an average ICC of 0.553 for the entire group (Table 3.5). Interpretation based off the single ICC values indicate a range of poor to good agreement concerning intra-rater reliability among the studied slide set. Confidence intervals for ICC values showed larger variation, widening the interpretation from as little as poor agreement to as much as excellent intra-rater agreement within the group.

3.4.4 Intra-rater Agreement of Histologic Features

The intra-rater agreement for each of the individual pathologists concerning the evaluation of histologic features of inflammation, fibrosis, glandular atrophy, and lymphatic lacunae using descriptive modifiers of either absent, mild, moderate, or severe was analyzed using ICC statistics (Appendix B: Tables B.19 to B.22). The average ICC was calculated for each histologic feature and summarized (Table 3.5).

The evaluation of fibrosis had the highest values of ICC among the pathologists (range of 0.290 to 0.756), suggesting the best consistency for intra-rater agreement with an average ICC of 0.575 and moderate overall agreement. Intra-rater agreement concerning the severity of lymphatic lacunae in biopsies was the lowest of all histologic features measured, with ICC values ranging from -0.156 to 0.639 and an average ICC of 0.231, indicating only poor intra-rater reliability.

3.4.5 Logistic Regression Modelling and Predictive Probabilities

The logistic regression models for the variables, inflammation and fibrosis, with the associated predicted probabilities for histologic severity on Kenney-Doig category both showed a positive relationship between increasing severity of histologic feature and increasing probability

of a more severe Kenney-Doig grade diagnosis (Appendix B: Tables B.23 to B.26). Based on the absence of inflammation alone, there is an estimated 50% probability of assigning a category I and 33.3% probability of assigning category IIA to a given biopsy. Mild inflammation was associated with a 44.5% probability of being graded as a category IIA and a 37.7% probability of being assigned to category IIB. Moderate inflammation revealed a predictive value of 52.7% for category IIB and 29.1% for category III. Biopsies where severe inflammation was present had a 94.6% probability of being assigned to category III.

Modelling for fibrosis revealed similar trends. In the absence of fibrosis, biopsies had a 31.8% probability of being assigned a category I and a 54.1% probability of being assigned to category IIA. The presence of mild fibrosis had relatively equal predictive probabilities, as biopsies had 44.3% and 47.4% probabilities to be placed as either a category IIA or IIB, respectively. Moderate fibrosis was associated with a 58.9% probability of falling within the category IIB and 35.3% probability of being assigned to category III. Finally, the presence of severe fibrosis in a biopsy had a 96.8% probability to be assigned as a category III. Overall, both inflammation and fibrosis showed increasing probability a more severe Kenney-Doig classification when assigned to the more severe descriptive modifiers.

Results for glandular atrophy and lymphatic lacunae did not follow this trend. For glandular atrophy evaluated as either absent, mild, or moderate, the majority of biopsies were predicted to be classified as either category IIA or category IIB (absent: 51.2% for category IIA and 24.6% for category IIB, mild: 33.5% for category IIA and 41.5% for category IIB, moderate: 14.8% for category IIA and 43.7% for category IIB). Severe glandular atrophy was only associated with a 57.1% likelihood of classifying a biopsy as category III.

Similarly, the majority of predicted probabilities for severity of lymphatic lacunae were highest for the two middle Kenney-Doig categories IIA and IIB (absent: 43.3% category IIA and 31.7% category IIB, mild: 33.2% category IIA and 38.8% category IIB, moderate: 8.6% category IIA and 42% category IIB). Assessing the lymphatic lacunae in a biopsy as moderate or severe gave similar predicted probabilities for both category IIB and III (moderate: 42% category IIB and 46.9% category III, severe: 50% category IIB and 50% category III).

3.4.6 Additive Usage of Descriptive Modifiers and Histologic Features

The highly predictive histologic features of inflammation and fibrosis were used to investigate how participating pathologists were interpreting the additive criteria of the Kenney-Doig scale to classify biopsies. Of all 504 observations (8 pathologists x 63 slides), 90 observations were assigned to category IIA when both inflammation and fibrosis were evaluated as at least mild in severity; 72 observations displaying mild inflammation and mild fibrosis, 6 observations displaying mild inflammation and moderate fibrosis, and 12 observations displaying moderate inflammation and mild fibrosis. Conversely, 139 observations were placed in category IIB when both inflammation and fibrosis were evaluated as at least mild in severity; 44 observations displaying mild inflammation and mild fibrosis, 43 observations displaying mild inflammation and moderate fibrosis, and 52 observations displaying moderate inflammation and mild fibrosis. When both moderate inflammation and moderate fibrosis were present, observations were evenly split between category IIB and category III, with 21 observations in each category.

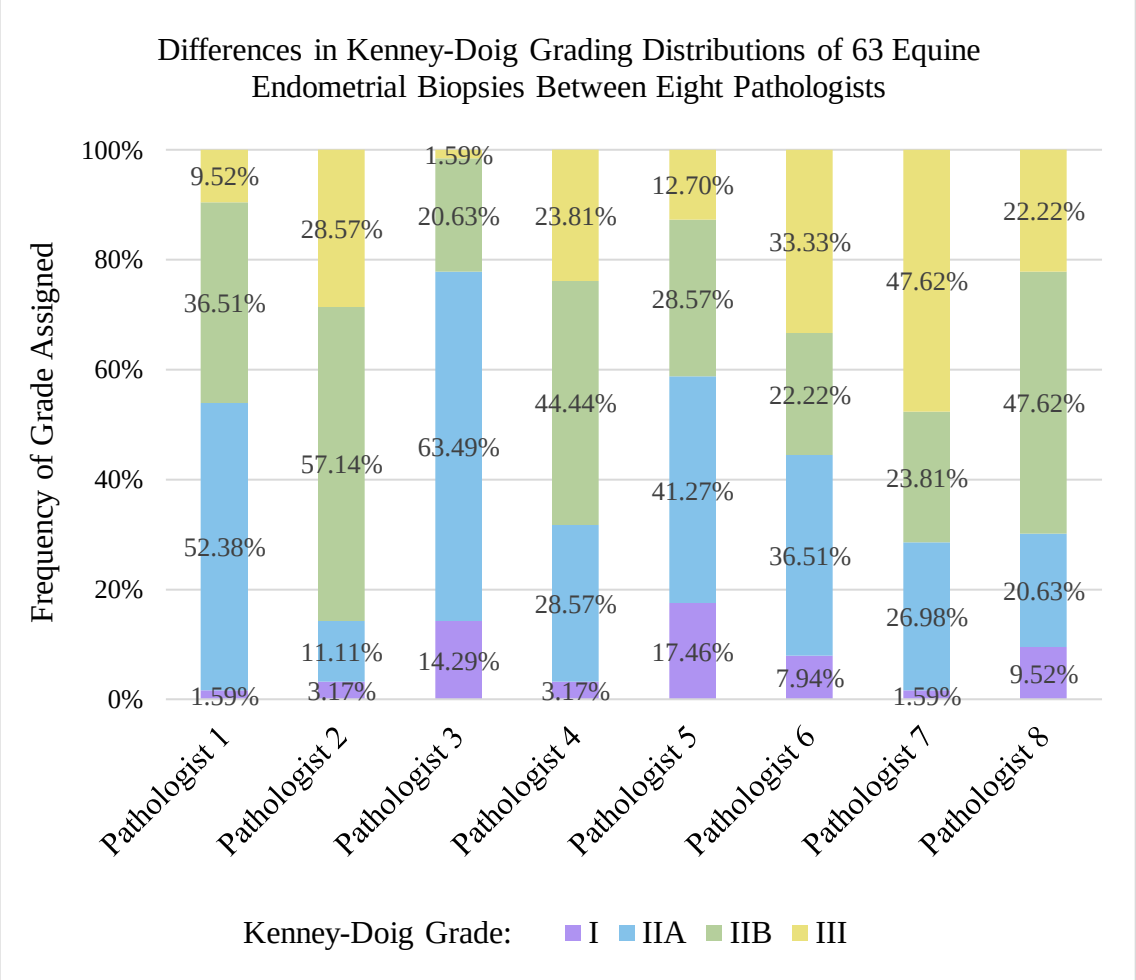

Figure 3.1. Frequency distributions of Kenney-Doig grades assigned by eight different pathologists to the same set of 63 equine endometrial biopsies.

Table 3.3 Unweighted and weighted Light's kappa coefficient measuring the average kappa value of Kenney-Doig grades assigned to 63 endometrial biopsies. Light's Kappa (A): the average kappa value between eight different pathologists, Light's Kappa (B): the average kappa value between five pathologists at different institutions, Light's Kappa (C): the average kappa value between four pathologists within the same institution.

	Unweighted	Weighted
Light's Kappa (A)	0.187	0.359
Light's Kappa (B)	0.143	0.330
Light's Kappa (C)	0.216	0.464

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table 3.4. Summary of the unweighted (U) and weighted (W) Light's Kappa values for the histologic descriptors assigned to the same 63 endometrial biopsies by eight different pathologists. Light's Kappa (A): the average kappa value between eight different pathologists, Light's Kappa (B): the average kappa value between five pathologists at different institutions, Light's Kappa (C): the average kappa value between four pathologists within the same institution.

	Inflammation		Fibrosis		Glandular Atrophy		Lymphatic Lacunae	
	U	W	U	W	U	W	U	W
Light's Kappa (A)	0.183	0.386	0.131	0.316	0.112	0.335	0.120	0.237
Light's Kappa (B)	0.150	0.359	0.113	0.292	0.073	0.359	0.065	0.229
Light's Kappa (C)	0.283	0.482	0.131	0.401	0.083	0.274	0.156	0.268

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table 3.5. Average intraclass correlation coefficients (ICC) measuring intra-rater agreement across eight different pathologists concerning the use of Kenney-Doig grades and descriptive modifiers such as absent, mild, moderate, or severe for grading of the same set of 21 endometrial biopsies. Glandular atrophy was only measured using seven pathologists as one pathologist did not assess this marker.

	Average ICC	Reliability Interpretation
Kenney-Doig Grade	0.553	Moderate
Inflammation	0.551	Moderate
Fibrosis	0.575	Moderate
Glandular Atrophy	0.439	Poor
Lymphatic Lacunae	0.231	Poor

ICC values were interpreted using standards suggested by Koo and Li (2016) where <0.50 is poor, 0.50 – 0.75 is moderate, 0.75-0.90 is good, and >0.90 is excellent agreement.

3.5 Discussion

The Kenney-Doig scale was designed to provide a histopathologic grading system that pathologists could utilize to provide equine clinicians with prognostic information regarding a mare's breeding potential. A histopathologic grading scale is only useful if the system is both valid and reliable, meaning that the prognostic indicators associated with each category must be validated by other experiments, and repeat measurements using the scale must be proven to be consistent between different observers and within the same observer. While some prognostic validity studies have been done on the Kenney-Doig scale, to date this is the only evaluation of inter and intra-rater agreement assessing the reliability of the system to this author's knowledge.

Histopathologic grading scales are commonly used for grading pre-neoplastic and neoplastic lesions in both human and veterinary medicine and have frequently been subject to inter and intra-rater variability studies (Bergeron et al., 1999; Brothwell et al., 2003; de Vet et al., 1990, 1995; Dellon et al., 2010; Grether et al., 1999; Kiupel et al., 2011; Koelink et al., 2018; Malpica et al., 2005; Matos et al., 2012; Munkedal et al., 2016; Northrup, Harmon, et al., 2005; Northrup, Howerth, et al., 2005; Robbins et al., 1995; Scholten et al., 2004; Stenkvist et al., 1983; Warners et al., 2018). Malpica et al. provided a comprehensive summary of inter and intra-rater variability studies concerning multiple systems for grading cervical dysplasia in human medicine from 1956 to 2001 where unweighted kappa values ranged from 0.13 to 0.63 across 13 different studies (Malpica et al., 2005). Other studies identified in the literature examining various histopathologic systems for human oral epitheliomas, breast cancer, and colon cancer biopsy grading have reported 0.24 – 0.70 (Bergeron et al., 1999; Brothwell et al., 2003; Malpica et al., 2005; Scholten et al., 2004; Stenkvist et al., 1983).

While there is no gold standard associated with a minimum accepted value for inter-rater reliability in histopathology, suggestions have been made for a minimum unweighted kappa value of 0.60 (McHugh, 2012). Overall, the inter-rater agreement observed in this study showed an average well below 0.60, only achieving slight to fair agreement between all eight different pathologists when using both unweighted and weighted Light's kappa means. These results were consistently in the lower ranges of observed inter-rater agreement found throughout the literature, suggesting less reliability compared to other existing histopathologic systems used in clinical practice. When examining the pairwise observations for weighted kappa values, only two

observer pairs, both trained and practicing at separate institutions, demonstrated substantial agreement while the majority only exhibited slight to fair agreement. Notably, in the entirety of this slide set, only a single biopsy was consistently given the same category (IIB) grading by all eight pathologists. In contrast, 6/63 slides were assigned all four possible Kenney-Doig grades, meaning that approximately one in every ten slides spanned the width of all possible diagnoses of the scale. Additionally, 3/63 slides were assigned both a category I and a category III by at least two pathologists, though did not span all four possible Kenney-Doig grades.

Interestingly, the average kappa values for pathologists practicing at the same institution were higher than those practicing at different institutions and all eight pathologists as a whole, reaching moderate agreement with the weighted statistic. This finding may not be surprising given the fact that colleagues at the same institution often collaborate during rounds sessions, seek one another's opinions on cases, and frequently train pathology residents under similar curriculum. The four intra-institution pathologists in this study also completed their residency training at the same institute. In contrast, three of the five pathologists included in the inter-institution group were trained at the same center while the remaining two pathologists each trained at different institutions. Similarities in residency training including shared mentorship may have also contributed to the higher levels of agreement observed in the intra-institution group.

Intra-rater variability presents another problem for reliability concerning histopathologic grading systems. While some differences in opinion between individuals is expected, there is also the issue regarding an individual's own variability on a day-to-day basis. Some comparable studies in the literature found a range of ICCs from 0.53 to as high as 0.92, with some group averages reported between 0.68 to 0.76 for different scoring systems (Christine Bergeron et al., 1999; Koelink et al., 2018). In the present study, pathologists' intra-rater agreement ICCs ranged from 0.116 to 0.774 with a group average of 0.553, a spread of poor to good agreement with the mean indicating moderate agreement. While this wide of spread for intra-rater agreement may relate to the vagueness of the scale, other confounding variables like individual pathologists' experience, training, and caseload of equine endometrial biopsies must be considered. However, when combined with the relatively low inter-agreement, it provides strong evidence that there are inconsistencies inherent with the use of the Kenney-Doig scale itself.

The guidelines Kenney and Doig outlined for their histopathologic grading scale involves describing the severity of multiple features of endometrial histopathology and summating these changes to fall in one of the four categories. Both the quantification of the distribution and severity of these features, and the end summation into a final grade, may be affected by subjective opinion of the observing pathologist. Several different reasons have been postulated by other researchers that could bias an observer and influence how they use any given histopathologic system. The creation of a four-category system may not be robust enough to cover multiple pathologic changes that tend to spread across a continuous scale, making it difficult to accurately quantify them into discreet groups (Cross, 1998; de Vet et al., 1990; McHugh, 2012). The criteria used to define these divisions may also be vague and open to interpretation, especially if the tissue present displays varying severity and distribution of certain pathology resulting in a heterogenous spread of changes versus a more homogenous population (de la Concha-Bermejillo et al., 1982; de Vet et al., 1990; T. Evans et al., 1998; Northrup, Harmon, et al., 2005; Northrup, Howerth, et al., 2005; Ricketts & Alonso, 1991a, 1991b). Additionally, pathologists may disagree on which histologic characteristics may be more important prognostically, thereby affecting how they judge the severity of pathology present and ultimately, how they combine these histologic features into a final diagnostic grade (de Vet et al., 1990; McHugh, 2012).

To further investigate how pathologists use the Kenney-Doig scale, and if disagreements concerning not only the final grade assigned, but also the basic evaluation of histopathologic feature severity and distribution were present, this study also measured inter and intra-rater agreement concerning the use of descriptive modifiers for quantifying the severity of inflammation, fibrosis, glandular atrophy, and lymphatic lacunae for each biopsy.

While the evaluation of inflammation produced slight to moderate inter-rater agreement, intra-rater agreement was highest when evaluating fibrosis, but still only moderate. Both lymphatic lacunae and glandular atrophy produced very low levels of both inter and intra-agreement. There appears not only to be disagreement in the final summating Kenney-Doig grade, but also in the very building blocks of subjective quantification used to reach each diagnosis. These findings were reinforced by the percentage of observations where both mild inflammation and mild fibrosis was evaluated as present in a slide but was graded as a category

IIA when the additive criteria of the Kenney-Doig scale suggests it should be placed within category IIB. While less common, a small portion of observations evaluating both inflammation and fibrosis as moderate were classified as category IIB instead of category III. Pathologists do not appear to agree on exactly what descriptive modifiers of which histologic feature constitute a given Kenney-Doig grade, and even more concerning, they do not appear to always agree with their own evaluations of the same biopsy slide at two different time points.

To try and establish whether some observers may consider certain pathology more prognostically significant than others regarding endometrial biopsy categorization, logistic regression modelling was done to assess the influence of each individual histologic feature measured against the different Kenney-Doig grades given by the eight pathologists. Both inflammation and fibrosis showed a clear trend where increasing severity of descriptive modifiers assigned to either histologic feature resulted in a higher predicted probability for these biopsies to be assigned to increasingly severe Kenney-Doig categories. Notably, the presence of severe inflammation or severe fibrosis alone predicted that biopsies would be assigned to category III 94.6% and 96.8% of the time, respectively. Glandular atrophy and lymphatic lacunae, however, revealed more ambiguous predicted probabilities with the majority of biopsies likely to be classified within category IIA or IIB regardless of being evaluated as either absent, mild, or moderate. Both histologic features showed a 50% or greater likelihood of assigning a category III to the biopsy if evaluated as severe, however neither predicted probability approached the clear majority predicted based on the evaluation of inflammation or fibrosis.

The wide range of predicted probabilities across all four Kenney-Doig categories found in this study for both glandular atrophy and lymphatic lacunae suggest that pathologists are not giving these histologic features the same consideration when assigning a final diagnosis as inflammation or fibrosis. This finding was expected given that both histologic changes are more likely to be attributed to primary disease processes like endometritis or endometrosis. Glandular atrophy and lymphatic lacunae, on the other hand, are commonly considered secondary changes caused by more significant endometrial pathology, and often appear with concomitant evidence of endometrosis in particular (W. Allen, 1992; Hoffmann, Ellenberger, et al., 2009; R. Kenney, 1978; Lehmann et al., 2011; Love, 2011; McCue, 2019, 2019; Schöniger & Schoon, 2020; Stanton et al., 2004). Therefore, using inflammation and fibrosis as more significant drivers in

assessing Kenney-Doig categorization, with glandular atrophy and lymphatic lacunae as additional summating modifiers, seems appropriate. There should be consideration, however, on the relative weighting of inflammation versus fibrosis.

Within the paper outlining the categorization guidelines of the Kenney-Doig system, the authors state that while inflammation is reversible and treatment may result in an improved Kenney-Doig grade, fibrosis is a progressive and irreversible change (R. Kenney & Doig, 1986). The study by Manning that compared endometrial biopsies from PMU mares to subsequent foaling rates also found that when evaluating histologic inflammation and fibrosis, only fibrosis was significantly correlated with foaling outcome (Manning, 2002). Suggestions from this study supported the idea of transient endometritis as a physiologic response to breeding or foaling and that care should be taken when evaluating a biopsy to differentially weight inflammation and fibrosis to decide on a final Kenney-Doig category (Manning, 2002). Based on our predicted probabilities, there does not appear to be significant differences in how pathologists are assigning Kenney-Doig categories based on inflammation and fibrosis.

The results of this study support the hypothesis that pathologists may not agree on what histopathology present in an equine endometrial biopsy constitutes a given Kenney-Doig category and that the grading of biopsies may not be consistently repeatable among all individuals. While none of the participating pathologists' frequency of grades assigned reflected those diagnosed by the original WCVM or PDS pathologist for the selected slide set, the author would like to emphasize that there is no gold standard for assigning Kenney-Doig grades. The differences in frequency distribution do not reflect whether one observer may be considered more "correct" than another, they only serve to further illustrate the variability in inter-rater agreement observed in the study.

While the level of inter-rater agreement found in this study may be low, weighted kappa values were considerably higher than unweighted kappa values for all measures, indicating that agreement improved when accounting for the magnitude of deviation between Kenney-Doig categories. More often than not, pathologists seem to be deviating by only one category in their diagnoses as opposed to multiple. However, given the wide disparity in prognostic indicators associated with each Kenney-Doig grade, even a difference of a single category could have substantial impact on a client's decision whether to pursue breeding the mare in question.

In the end, given the low inter and intra-rater reliability and the significant influence a misplaced Kenney-Doig grade may have on equine clinicians, their clients, and their patients, where does that leave the Kenney-Doig scale? While attempts have been made to improve the system with the addition of more objectively quantifiable variables including various staining and immunohistochemistry techniques, no routinely used methods are being implemented in North America to augment the scale (Aupperle et al., 2004; T. Evans et al., 1998; Hoffmann, Ellenberger, et al., 2009; Lunelli et al., 2013; Mambelli et al., 2014; Minkwitz et al., 2019; Oddsdóttir, 2007; Walter et al., 2001). Researchers in Germany have implemented a modified Kenney-Doig scale to try and clarify the existing guidelines between different grade categories, but this is not commonly used by North American pathologists (H. Schoon et al., 1997). The same group have described new histopathologic features that should arguably be included when evaluating an equine endometrial biopsy, such as endometrial maldifferentation and angiopathies (Hoffmann, Ellenberger, et al., 2009; Lehmann et al., 2011; D. Schoon et al., 1999; H. Schoon et al., 1997, 1999, 2000). These works support the idea that certain histologic features should be differentially weighted and considered when assigning a final prognosis to an equine endometrial biopsy.

Other histopathologic systems that have produced less than desirable inter and intra-rater agreement have been either abandoned or altered, sometimes by splitting the more ambiguous middle categories into either extreme, thereby reducing a multi-category system into only two possible grades (Kiupel et al., 2011; Northrup, Howerth, et al., 2005). Authors have also suggested that general histopathologic grading and staging may be best used in a research atmosphere with less observers, and that those in clinical practice should focus more on the description of the pathology present, integrating those changes and being able to communicate possible etiologies and direct the submitting clinician to the best course of therapy (Cross, 1998; Ishak et al., 1995). These thoughts echo Kenney's final advice concerning the use of his scale, where he described developing an 'epicrisis'; an overarching diagnosis that takes into account not only the Kenney-Doig grade, but also the mare as a whole including her age, general and reproductive history such as maiden status or years barren, her overall physical exam parameters, her reproductive exam parameters such as anatomical appearance, vulvar seal integrity, rectal palpation, ultrasonic and possibly endoscopic visualization of the uterus, as well as cytology and culture results (R. Kenney, 1978; R. Kenney & Doig, 1986). The integration of this information

ultimately by the attending clinician is crucial to able to accurately evaluate the reproductive efficiency of a mare.

While the levels of inter and intra-rater agreement found in this study seem considerably low, there are some limitations that may have affected our results and influenced the inter and intra-observer agreement. Digitized slides have been found to produce almost perfect agreement with their glass slide counter-parts, however differences in computer screen brightness and resolution may have contributed to lower agreement between pathologists (Dellon et al., 2010; Warners et al., 2018). Digital viewing also prohibits the ability to focus through different planes of the slide which may hamper evaluation (Dellon et al., 2010). The sample size used to calculate the ICC's for intra-rater agreement was also relatively small, reflected in the wide confidence intervals obtained. Despite these limitations, this study enabled eight different pathologists across geographically separated institutions to evaluate the same slide set in a convenient and affordable manner. It was appropriately blocked to control for grading and screen-time fatigue. Pathologists were blinded to the original slide diagnosis and were given adequate rest between inter and intra-rater slide sets to control for slide recognition. In sum, this study presents strong evidence that more work is needed to further investigate the reliability of the Kenney-Doig scale and encourages researchers to continue to search for more objectively quantifiable alternatives and/or more pertinent histologic features relating to mare reproductive efficiency. Pathologists are encouraged to communicate these findings to submitting clinicians and integrate their findings in a collaborative fashion to better direct therapy and treatment options.

CHAPTER 4. Concluding Statements

This study aimed to address two different objectives. First, to describe the trend in Kenney-Doig categorization of equine endometrial biopsies given by pathologists at the WCVM/PDS over the twenty-year span between 1998 and 2018. Second, to measure the inter-rater and intra-rater agreement of eight different pathologists' grading of the same set of 63 endometrial biopsies using the Kenney-Doig scale.

The grading trends found at the WCVM/PDS revealed a predominance of the middle-ranked categories IIA and IIB, with relatively small numbers of biopsies assigned to the extreme categories I and III. Six different studies that reported Kenney-Doig grades for at least 150 biopsies were found in the literature and examined to determine if the grading distribution at the WCVM/PDS was repeatable (Kabisch et al., 2019; Kilgenstein et al., 2015; Nambo et al., 2014; Ricketts & Alonso, 1991a; Schilling, 2017; Waelchli, 1990). Significant variation between the Kenney-Doig grading distributions among the six studies and that of WCVM/PDS was found. While the effect of different mare populations between studies regarding parameters such as age, breed, and athletic use could not be excluded, it was hypothesized that some degree of observer variation may be influencing the differences in Kenney-Doig categorization, as has been suggested previously in the literature (de la Concha-Bermejillo et al., 1982; M. Evans et al., 1986; Ricketts & Alonso, 1991a, 1991b; Snider et al., 2011). When examining five separate pathologists' grading distributions at WCVM/PDS where observers see a similar local mare population, there were significant differences between grading tendencies, providing additional support to our hypothesis. To test for the occurrence of observer variability when using the Kenney-Doig scale, inter-agreement and intra-agreement was measured using eight pathologists and the same slide set of endometrial biopsies.

Kenney-Doig categories were recorded for eight different pathologists using the same digital set of 63 endometrial biopsies. Ultimately, only fair to moderate levels of inter-rater agreement were found using two different forms of kappa statistics. While there is no published

gold standard for inter-rater agreement levels concerning the use of histopathologic systems, the levels found in this study fell within the lower end of the range of values reported in the literature for observer variability in other grading systems (Bergeron et al., 1999; Brothwell et al., 2003; Malpica et al., 2005; Scholten et al., 2004; Stenkvist et al., 1983).

Pathologists also graded a second set of slides involving 21 random biopsies that had been previously graded two weeks to twelve months prior from the original 63 slide set, allowing for the measurement of intra-rater agreement regarding Kenney-Doig categories. A wide range of intra-rater agreement was found within the group of eight pathologists, producing an average of only moderate agreement. Similar to the inter-agreement values, those found for the intra-agreement in this study were among the lower values reported in the literature (Bergeron et al., 1999; Koelink et al., 2018).

When grading these endometrial biopsies, all pathologists were also asked to describe the severity of four histologic features that are considered significant in the Kenney-Doig guidelines (R. Kenney, 1978; R. Kenney & Doig, 1986). Pathologists described the inflammation, fibrosis, glandular atrophy, and lymphatic lacunae in each slide as either absent, mild, moderate, or severe. The evaluation of inflammation and fibrosis produced the highest levels of inter and intra-rater agreement respectively, indicating fair and moderate agreement among the group. The evaluation of lymphatic lacunae and glandular atrophy produced the lowest inter and intra-rater agreement, with values ranging between poor and slight agreement. When comparing the use of these descriptive modifiers for inflammation and fibrosis to Kenney-Doig grades, there was evidence that some pathologists may not be consistently using the additive criteria suggested by the scale.

Additionally, logistic regression modelling and predictive probability calculations based off these regression models showed that some histologic features increase the likelihood of assigning certain Kenney-Doig categories more so than others. The evaluation of inflammation and fibrosis were associated with a greater than 90% chance of assigning a category III when evaluated as severe, whereas severe changes in glandular atrophy and lymphatic lacunae were only associated with an approximately 50% chance of assigning a given biopsy to category III. This suggests that pathologists are differentially weighting certain histologic characteristics when assigning a final diagnostic category.

Overall, there was a high prevalence of observer variability among both the final Kenney-Doig grades assigned and the histologic feature evaluations, suggesting that the quantification guidelines laid out by the Kenney-Doig scale are not repeatable between or within observers. There is also evidence of differential use of the evaluation inflammation and fibrosis over the quantification of glandular atrophy and lymphatic lacunae in deciding the final Kenney-Doig category. Given the findings in this study, there are several points to consider in regard to the use of the Kenney-Doig scale for evaluating equine endometrial biopsies.

The histologic features listed by Kenney and Doig as categorization guidelines are lacking new endometrial pathology thought to be associated with fertility (D. Schoon et al., 1999; H. Schoon et al., 1997, 1999; H. Schoon & Schoon, 2003; Snider et al., 2011). While Kenney and Doig have already introduced the concept of differentially weighing certain changes over others as some forms of endometrial pathology is readily treatable, more work needs to be done to investigate the relationship of new features such as angiopathies and endometrial maldifferentiation with foaling outcomes, and to validate the findings of other researchers where fibrosis appears to be a more powerful predictor of fertility than other histologic changes such as inflammation (Manning, 2002). Additionally, there has been a wealth of research in the past two decades investigating the pathogenesis of endometrosis involving both epithelial and stromal protein markers that need to be investigated in regard to foaling outcome and fertility prognosis (Aupperle et al., 2004; Hoffmann, Bazer, et al., 2009; Mambelli et al., 2014; Minkwitz et al., 2019; Walter et al., 2001). Ultimately, the checklist of endometrial pathology used to evaluate a mare's fertility needs to be examined, expanded on, and refined.

However, even if these changes are made to the list of histologic features that need to be considered concerning mare fertility, the guidelines used to quantify these changes need to be redefined. It is evident from both the literature and from the low levels of inter-rater and intra-rater agreement in this study concerning not only the final summated Kenney-Doig grade but also the base quantification of the histologic features, that the original guidelines set for the system are only semi-quantitative and subject to significant observer variation, while categorization is also contingent on certain clinical information that is not always available to the pathologist, leaving the scale unreliable (de la Concha-Bermejillo et al., 1982; M. Evans et al., 1986; Ricketts & Alonso, 1991a, 1991b; Snider et al., 2011). Researchers have used various techniques to try and

lend more objective measures to the existing Kenney-Doig scale, including the use of picrosirius red staining to highlight areas of fibrosis and subsequently measure these areas with varying methods of automated image analysis, and the use of immunohistochemistry to identify certain proteins such as calponin, vimentin, and smooth muscle actin that may act as early indicators of endometrial fibrosis and quantifying these markers with more objective measures such as counting the percentage of positively-stained cells or percentage of glands affected (Aupperle et al., 2004; T. Evans et al., 1998; Hoffmann, Bazer, et al., 2009; Mambelli et al., 2014; Minkwitz et al., 2019; Oddsdóttir, 2007; Walter et al., 2001). While these methods seem promising, further work needs to be done to correlate these methods with some form of reproductive efficiency outcome, ideally with foaling rates from subsequent foaling seasons, as early embryonic loss in the mare is common and delivery of a live foal acts as the best benchmark for reproductive success.

Until this work is done, however, where does that leave today's pathologist when faced with an equine endometrial biopsy? While some may argue that the weighted kappa values in this study showed moderate agreement, and that inconsistency between pathologists concerning a single Kenney-Doig grade may not be a reason for concern, this author would like to highlight the large disparity in percent foaling chance between each Kenney-Doig category. For example, the categorization of a mare in category IIA is associated with a 50 to 80% chance of having a live foal, while category IIB is associated with a 10-50%. This is a significant difference that could influence a client and referring veterinarian when deciding whether to buy a given mare, pursue breeding a current one, or make decisions regarding advanced breeding techniques and uterine therapy. Taking this information into account, immediate steps that may help improve the situation involve better communication and training. The development of a standardized training atlas containing colour photos of case examples including a mare's signalment, history, cytology and culture results, and endometrial features may also be useful for use in diagnostic centers, for theriogenologists, and for field clinicians. Pathologists who find themselves faced with an endometrial biopsy may wish to reach out to a theriogenologist, or to colleagues that are well-versed in the Kenney and Doig system and see a regular caseload of equine endometrial biopsies. These steps may improve interpretation of the additive nature of the Kenney-Doig criteria and thereby reduce observer variation. However, given the low inter-rater and intra-rater agreement concerning the basic severity of histologic features, this may not adequately improve overall

reliability of the scale. Future studies to compare the observer agreement of the Kenney-Doig scale before and after training sessions, and/or the implementation of a training atlas, are needed.

Until then, this author encourages pathologists to communicate with submitting clinicians the potential for issues regarding reliability and validity when using the Kenney-Doig scale given these study results and ongoing research in the field of endometrial pathology. They should report that the most useful clinical information is often obtained when a mare is graded as a category I, showing mild to no endometrial pathology, as all other mares can possibly improve based on treatment and intensively managed breeding techniques. In addition to providing a Kenney-Doig grade and the associated cautions in interpreting that grade, pathologists should comment on the type of histopathology present in a given biopsy, aim to direct therapy based on those findings, and pose questions that may help the clinician further evaluate the mare as a whole. Finally, pathologists are encouraged to contact an equine theriogenologist for case advice and collaboration.

For referring veterinarians, this author encourages equine veterinarians to always read the pathologist's comments in their entirety, to critically evaluate the histologic description, and to never hesitate to call the reporting pathologist with questions or concerns. Advancements in equine reproductive medicine have offered new forms of therapy that were not available when the Kenney-Doig system was devised, therefore, consultation with a theriogenologist for a more accurate prognosis regarding a mare is also encouraged. To rely on the simplified chart supplied by Kenney and Doing over thirty years ago and simply relay the associated foaling percent chance to the client without an accompanying discussion about the underlying endometrial pathology is no longer acceptable, and arguably never was given Kenney's advice for referring veterinarians to develop an overarching 'epicrisis' that integrates all clinical findings concerning the mare with her endometrial biopsy for a more accurate breeding prognosis (R. Kenney, 1978; R. Kenney & Doig, 1986).

In sum, given the outdated histologic criteria and low reliability produced by the Kenney-Doig scale, pathologists, equine theriogenologists, and equine clinicians must work together when interpreting an endometrial biopsy to better evaluate a mare's reproductive prognosis and aid in client decision-making.

REFERENCES

Allen, W. (1992). The diagnosis and handling of early gestational abnormalities in the mare. *Animal Reproduction Science*, *28*(1), 31–38. https://doi.org/10.1016/0378-4320(92)90088-U

Allen, W. R., Brown, L., Wright, M., & Wilsher, S. (2007). Reproductive efficiency of Flatrace and National Hunt Thoroughbred mares and stallions in England. *Equine Veterinary Journal*, *39*(5), 438–445. https://doi.org/10.2746/042516407X1737581

Animal and Plant Health Inspection Service. (2009). *Beef 2007-2008, Part II: Reference of Beef Cow-calf Management Practices in the United States* (Government Study N512.0209; National Animal Health Monitoring System, p. 44). United States Department of Agriculture, Center for Epidemiology and Animal Health.

Assad, N., & Pandey, A. (2015). Different approaches to diagnose uterine pathology in mares: A review. *Theriogenology Insight - An International Journal of Reproduction in All Animals*, *5*(3), 157–182. https://doi.org/10.5958/2277-3371.2015.00018.2

Aupperle, H., Özgen, S., Schoon, H., Schoon, D., Hoppen, H., Sieme, H., & Tannapfel, A. (2000). Cyclical endometrial steroid hormone receptor expression and proliferation intensity in the mare. *Equine Veterinary Journal*, *32*(3), 228–232. https://doi.org/10.2746/042516400776563554

Aupperle, H., Schoon, D., & Schoon, H. (2004). Physiological and pathological expression of intermediate filaments in the equine endometrium. *Research in Veterinary Science*, *76*(3), 249–255. https://doi.org/10.1016/j.rvsc.2003.11.003

Back, W., & Clayton, H. (2013). *Equine Locomotion—E-Book*. Elsevier Health Sciences.

Bergeron, C., Nogales, F., Masseroli, M., Abeler, V., Duvillard, P., Muller-Holzner, E., Pickartz, H., & Wells, M. (1999). A multicentric European study testing the reproducibility of the WHO classification of endometrial hyperplasia with a proposal of a simplified working classification for biopsy and curettage specimens. *The American Journal of Surgical Pathology*, *23*(9), 1102.

Bosh, K., Powell, D., Shelton, B., & Zent, W. (2009). Reproductive performance measures among Thoroughbred mares in central Kentucky, during the 2004 mating season. *Equine Veterinary Journal*, *41*(9), 883–888. https://doi.org/10.2746/042516409X456068

Brook, D., & Frankel, K. (1987). Electrocoagulative removal of endometrial cysts in the mare. *Journal of Equine Veterinary Science*, *7*(2), 77–81. https://doi.org/10.1016/S0737-0806(87)80035-7

Brothwell, D., Lewis, D., Bradley, G., Leong, I., Jordan, R., Mock, D., & Leake, J. (2003). Observer agreement in the grading of oral epithelial dysplasia. *Community Dentistry and Oral Epidemiology*, *31*(4), 300–305. https://doi.org/10.1034/j.1600-0528.2003.00013.x

Buczkowska, J., Kozdrowski, R., Nowak, M., Raś, A., & Mrowiec, J. (2014). Endometrosis – significance for horse reproduction, pathogenesis, diagnosis, and proposed therapeutic methods. *Polish Journal of Veterinary Sciences*, *17*(3). https://doi.org/10.2478/pjvs-2014-0083

Buczkowska, J., Kozdrowski, R., Nowak, M., Raś, A., Staroniewicz, Z., & Siemieniuch, M. (2014). Comparison of the biopsy and cytobrush techniques for diagnosis of subclinical endometritis in mares. *Reproductive Biology and Endocrinology*, *12*(1), 27. https://doi.org/10.1186/1477-7827-12-27

Bujang, M., & Baharum, N. (2017). Guidelines of the minimum sample size requirements for Cohen's Kappa. *Epidemiology Biostatistics and Publich Health*, *14*(2), e12267-1 to e12267-10.

Bukowski, J., & Aiello, S. (n.d.). *Breeding and Reproduction of Horses*. Merck Veterinary Manual. Retrieved July 6, 2018, from https://www.merckvetmanual.com/horse-owners/routine-care-and-breeding-of-horses/breeding-and-reproduction-of-horses

Burger, D., Wohlfender, F., & Imboden, I. (2008). Managing a mare for breeding and sport. *Pferdeheilkunde*, *24*(1), 102–107. https://doi.org/10.21836/PEM20080122

Canisso, I., Pinn, T., Gerdin, J., Ollivett, T., Buckles, E., Schweizer, C., & Ainsworth, D. (2013). B-cell multicentric lymphoma as a probable cause of abortion in a Quarter horse broodmare. *The Canadian Veterinary Journal*, *54*(3), 288–291.

Cartwright, D. (1956). A rapid non-parametric estimate of multi-judge reliability. *Psychometrika*, *21*(1), 17–29.

Causey, R. (2006). Making sense of equine uterine infections: The many faces of physical clearance. *The Veterinary Journal*, *172*(3), 405–421. https://doi.org/10.1016/j.tvjl.2005.08.005

Claes, A., Ball, B., Liu, I., Vaughan, B., Highland, M., & Brown, J. (2015). Uterine B cell lymphoma in a mare. *Equine Veterinary Education*, *27*(7), e5–e8. https://doi.org/10.1111/j.2042-3292.2012.00431.x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Cothran, E., MacCluer, J., Weitkamp, L., Pfennig, D., & Boyce, A. (1984). Inbreeding and reproductive performance in Standardbred horses. *Journal of Heredity*, *75*(3), 220–224. https://doi.org/10.1093/oxfordjournals.jhered.a109916

Cross, S. (1996). Kappa statistics as indicators of quality assurance in histopathology and cytopathology. *Journal of Clinical Pathology*, *49*(7), 597–599. https://doi.org/10.1136/jcp.49.7.597

Cross, S. (1998). Grading and scoring in histopathology. *Histopathology*, *33*, 99–106.

de la Concha-Bermejillo, A., MVZ, MS, & Kennedy, P. (1982). Prognostic value of endometrial biopsy in the mare: A retrospective analysis. *Journal of Clinical Epidemiology*, *181*(7), 680–681.

de Vet, H., Knipschild, P., Schouten, H., Koudstaal, J., Kwee, W., Willebrand, D., Sturmans, F., & Arends, J. (1990). Interobserver variation in histopathological grading of cervical dysplasia. *Journal of Clinical Epidemiology*, *43*(12), 1395–1398. https://doi.org/10.1016/0895-4356(90)90107-Z

de Vet, H., Koudstaal, J., Kwee, W., Willebrand, D., & Arends, J. (1995). Efforts to improve interobserver agreement in histopathological grading. *Journal of Clinical Epidemiology*, *48*(7), 869–873.

Dellon, E., Fritchie, K., Rubinas, T., Woosley, J., & Shaheen, N. (2010). Inter- and intraobserver reliability and validation of a new method for determination of eosinophil counts in patients with esophageal eosinophilia. *Digestive Diseases and Sciences*, *55*(7), 1940–1949. https://doi.org/10.1007/s10620-009-1005-z

Desmet, V. J., Gerber, M., Hoofnagle, J. H., Manns, M., & Scheuer, P. J. (1994). Classification of chronic hepatitis: Diagnosis, grading and staging. *Hepatology*, *19*(6), 1513–1520. https://doi.org/10.1002/hep.1840190629

Dini, P., Bartels, T., Revah, I., Claes, A., Stout, T., & Daels, P. (2020). A retrospective study on semen quality parameters from four different Dutch horse breeds with different levels of inbreeding. *Theriogenology*, *157*, 18–23. https://doi.org/10.1016/j.theriogenology.2020.07.017

Doig, P., McKnight, J., & Miller, R. (1981). The use of endometrial biopsy in the infertile mare. *Canadian Veterinary Journal*, *22*, 72–76.

Eilts, B., Scholl, D., Paccamonti, D., Causey, R., Klimczak, J., & Corley, J. (1995). Prevalence of endometrial cysts and their effect on fertility. *Biology of Reproduction*, *52*(monograph_series1), 527–532. https://doi.org/10.1093/biolreprod/52.monograph_series1.527

Ellenberger, C., Aupperle, H., Bartmann, C., Hoppen, H., Schoon, D., & Schoon, H. (2002). Endometrial maldifferentiation caused by ovarian disorders in the mare—Morphological and immunohistochemical studies. *Theriogenology*, *58*(2), 499–502. https://doi.org/10.1016/S0093-691X(02)00817-8

Esteller-Vico, A., Liu, I., Vaughan, B., Steffey, E., & Brosnan, R. (2015). Effects of vascular elastosis on uterine blood flow and perfusion in anesthetized mares. *Theriogenology*, *83*(6), 988–994. https://doi.org/10.1016/j.theriogenology.2014.11.032

Evans, M., Hamer, J., Gason, L., Graham, C., Asbury, A., & Irvine, C. (1986). Clearance of bacteria and non-antigenic markers following intra-uterine inoculation into maiden mares: Effect of steroid hormone environment. *Theriogenology*, *26*(1), 37–50. https://doi.org/10.1016/0093-691X(86)90110-X

Evans, T., Miller, M., Ganjam, V., Niswender, K., Ellersieck, M., Krause, W., & Youngquist, R. (1998). Morphometric analysis of endometrial periglandular fibrosis in mares. *American Journal of Veterinary Research*, *59*(10), 1209–1214.

Fadare, O., Parkash, V., Dupont, W., Acs, G., Atkins, K., Irving, J., Pirog, E., Quade, B., Quddus, M., Rabban III, J., Vang, R., & Hecht, J. (2013). The diagnosis of endometrial carcinomas with clear cells by gynecologic pathologists: An assessment of interobserver variability and associated morphologic features. *Yearbook of Pathology and Laboratory Medicine*, *36*(8), 162–165. https://doi.org/10.1016/j.ypat.2012.11.084

Ferreira-Dias, G., & King, S. (1994). Morphologic characteristics of equine endometrium classified as Kenney categoreis I, II, and III, using light and scanning electron microscopy. *American Journal of Veterinary Research*, *55*(8), 1060–1065.

Flores, J., Rodríguez, A., Sánchez, J., Gómez-Cuétara, C., & Ramiro, F. (1995). Endometrosis in mares: Incidence of histopathological alterations. *Reproduction in Domestic Animals*, *30*(2), 61–65. https://doi.org/10.1111/j.1439-0531.1995.tb00606.x

Freeman, S., England, G., Bjornson, S., & Smith, R. (1997). Uterine T cell lymphoma in a mare with multicentric involvement. *The Veterinary Record*, *141*, 391–393. https://doi.org/10.1136/vr.141.15.391

Geboes, K. (2000). A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut*, *47*(3), 404–409. https://doi.org/10.1136/gut.47.3.404

Gerstenberg, C., Allen, W., & Stewart, F. (1999). Cell proliferation patternsin the equine endometrium throughout the non-pregnant reproductive cycle. *Journal of Reproduction and Fertility*, *116*, 167–175.

Govaere, J., Maes, S., Saey, V., Blancke, W., Hoogewijs, M., Deschauwer, C., Smits, K., Roels, K., Vercauteren, G., & de Kruif, A. (2011). Uterine fibrosarcoma in a Warmblood mare. *Reproduction in Domestic Animals*, *46*(3), 564–566. https://doi.org/10.1111/j.1439-0531.2010.01694.x

Grether, J., Eaton, A., Redline, R., Bendon, R., Benirschke, K., & Nelson, K. (1999). Reliability of placental histology using archived specimens. *Paediatric and Perinatal Epidemiology*, *13*(4), 489–495. https://doi.org/10.1046/j.1365-3016.1999.00214.x

Grimm, A., Schoon, H., & Schöniger, S. (2017). Histopathological features of endometritis eosinophilica in mares. *Histology and Histopathology*, *32*(11), 1161–1173. https://doi.org/10.14670/HH-11-872

Grüninger, B., Schoon, H., Schoon, D., Menger, S., & Klug, E. (1998). Incidence and morphology of endometrial angiopathies in mares in relationship to age and parity. *Journal of Comparative Pathology*, *119*(3), 293–309. https://doi.org/10.1016/S0021-9975(98)80051-0

Gunson, D., Gillette, D., Beech, J., & Orsini, J. (1980). Endometrial adenocarcinoma in a mare. *Veterinary Pathology*, *17*(6), 776–780. https://doi.org/10.1177/030098588001700615

Häfner, I., Schoon, H., Schoon, D., & Aupperle, H. (2001). Disorders of differentiation in the equine endometrium – light microscopic and immunohistological studies. *Pferdeheilkunde*, *17*(2), 103–110. https://doi.org/10.21836/PEM20010201

Hallgren, K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hanada, M., Maeda, Y., & Oikawa, M. (2014). Histopathological characteristics of endometrosis in Thoroughbred mares in Japan: Results from 50 necropsy cases. *Journal of Equine Science*, *25*(2), 45–52. https://doi.org/10.1294/jes.25.45

Hanlon, D., Stevenson, M., Evans, M., & Firth, E. (2012). Reproductive performance of Thoroughbred mares in the Waikato region of New Zealand: 1. Descriptive analyses. *New*

Zealand Veterinary Journal, *60*(6), 329–334. https://doi.org/10.1080/00480169.2012.693039

Held, J., & Rohrbach, B. (1991). Clinical significance of uterine biopsy in the maiden and non-maiden mare. *Journal of Reproduction and Fertility*, *Supplement 44*, 698–699.

Hoffmann, C., Bazer, F., Klug, J., Aupperle, H., Ellenberger, C., & Schoon, H. (2009). Immunohistochemical and histochemical identification of proteins and carbohydrates in the equine endometrium: Expression patterns for mares suffering from endometrosis. *Theriogenology*, *71*(2), 264–274. https://doi.org/10.1016/j.theriogenology.2008.07.008

Hoffmann, C., Ellenberger, C., Mattos, R., Aupperle, H., Dhein, S., Stief, B., & Schoon, H. (2009). The equine endometrosis: New insights into the pathogenesis. *Animal Reproduction Science*, *111*(2–4), 261–278. https://doi.org/10.1016/j.anireprosci.2008.03.019

Hurtgen, J. (2006). Pathogenesis and treatment of endometritis in the mare: A review. *Theriogenology*, *66*(3), 560–566. https://doi.org/10.1016/j.theriogenology.2006.04.006

Ishak, K., Baptista, A., Bianchi, L., Callea, F., De Groote, J., Gudat, F., Denk, H., Desmet, V., Korb, G., MacSween, R., Phillips, M., Portmann, B., Poulsen, H., Scheuer, P., Schmid, M., & Thaler, H. (1995). Histological grading and staging of chronic hepatitis. *Journal of Hepatology*, *22*(6), 696–699. https://doi.org/10.1016/0168-8278(95)80226-6

Kabisch, J., Klose, K., & Schoon, H. (2019). Endometrial biopsies of old mares – What to expect?! *Pferdeheilkunde Equine Medicine*, *35*(3), 211–219. https://doi.org/10.21836/PEM20190302

Keller, A., Neves, A., Aupperle, H., Steiger, K., Garbade, P., Schoon, H., Klug, E., & Mattos, R. (2006). Repetitive experimental bacterial infections do not affect the degree of uterine

degeneration in the mare. *Animal Reproduction Science*, *94*(1), 276–279. https://doi.org/10.1016/j.anireprosci.2006.04.012

Kenney, R. (1978). Cyclic and pathologic changes of the mare endometrium as detected by biopsy, with a note on early embryonic death. *Journal of American Veterinary Medical Association*, *172*(3), 241–262.

Kenney, R., & Doig, P. (1986). Equine endometrial biopsy. In D. Morrow (Ed.), *Current Therapy in Theriogenology* (2nd ed., pp. 723–729). WB Saunders.

Kenney, R. M. (1978). Clinical aspects of endometrial biopsy in fertility evaluation of the mare. *Proceedings - Annual Convention of the American Association of Equine Practitioners (USA)*. https://agris.fao.org/agris-search/search.do?recordID=US7804888

Kilgenstein, H., Schöniger, S., Schoon, D., & Schoon, H. (2015). Microscopic examination of endometrial biopsies of retired sports mares: An explanation for the clinically observed subfertility? *Research in Veterinary Science*, *99*, 171–179. https://doi.org/10.1016/j.rvsc.2015.01.005

Killisch, R., Böttcher, D., Theuß, T., Edzards, H., Martinsson, G., Einspanier, A., Gottschalk, J., & Schoon, H. (2017). Seasonal or pathological findings? Morphofunctional characteristics of the equine endometrium during the autumn and spring transition. *Reproduction in Domestic Animals*, *52*(6), 1011–1018. https://doi.org/10.1111/rda.13016

Kiupel, M., Webster, J., Bailey, K., Best, S., DeLay, J., Detrisac, C., Fitzgerald, S., Gamble, D., Ginn, P., Goldschmidt, M., Hendrick, M., Howerth, E., Janovitz, E., Langohr, I., Lenz, S., Lipscomb, T., Miller, M., Misdorp, W., Moroff, S., … Miller, R. (2011). Proposal of a 2-tier histologic grading system for canine cutaneous mast cell tumors to more accurately predict biological behavior. *Veterinary Pathology*, *48*(1), 147–155. https://doi.org/10.1177/0300985810386469

Klug, E., Bartmann, C., Schöning, A., Schoon, D., & Schoon, H. (1997). Influence of longterm progestin application on sexual cycle and endometrial structures and functions of the mare: *Pferdeheilkunde*, *13*(5), 490–498. https://doi.org/10.21836/PEM19970511

Koelink, P., Wildenberg, M., Stitt, L., Feagan, B., Koldijk, M., van 't Wout, A., Atreya, R., Vieth, M., Brandse, J., Duijst, S., te Velde, A., D'Haens, G., Levesque, B., & van den Brink, G. (2018). Development of reliable, valid and responsive scoring systems for endoscopy and histology in animal models for inflammatory bowel disease. *Journal of Crohn's and Colitis*, *12*(7), 794–803. https://doi.org/10.1093/ecco-jcc/jjy035

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Langley, fa. (1978). Quality control in histopathology and diagnostic cytology. *Histopathology*, *2*(1), 3–18. https://doi.org/10.1111/j.1365-2559.1978.tb01689.x

LeBlanc, M., Magsig, J., & Stromberg, A. (2007). Use of a low-volume uterine flush for diagnosing endometritis in chronically infertile mares. *Theriogenology*, *68*(3), 403–412. https://doi.org/10.1016/j.theriogenology.2007.04.038

Lee, K., & Nelson, C. (2012). New insights into the regulation of epithelial–mesenchymal transition and tissue fibrosis. In K. Jeon (Ed.), *International Review of Cell and Molecular Biology* (Vol. 294, pp. 171–221). Academic Press. https://doi.org/10.1016/B978-0-12-394305-7.00004-5

Lehmann, J., Ellenberger, C., Hoffmann, C., Bazer, F., Klug, J., Allen, W., Sieme, H., & Schoon, H. (2011). Morpho-functional studies regarding the fertility prognosis of mares suffering

from equine endometrosis. *Theriogenology*, *76*(7), 1326–1336. https://doi.org/10.1016/j.theriogenology.2011.06.001

Light, R. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365–377.

Lopez, C., Ciccarelli, M., Gold, J., & Tibary, A. (2018). Uterine adenocarcinoma in Quarter Horse mare. *Equine Veterinary Education*, *30*(12), 640–644. https://doi.org/10.1111/eve.12795

Love, C. (2011). Techniques in reproductive examination: Endometrial biopsy. In A. McKinnon, E. Squires, W. Vaala, & D. Varner (Eds.), *Equine Reproduction* (2nd ed., Vol. 2, pp. 1929–1939). Wiley-Blackwell.

Lucas, Z., Raeside, J., & Betteridge, K. (1991). Non-invasive assessment of the incidences of pregnancy and pregnancy loss in the feral horses of Sable Island. *Journal of Reproduction and Fertility. Supplement*, *44*, 479–488.

Lunelli, D., Cirio, S., Leite, S., Camargo, C., & Kozicki, L. (2013). Collagen types in relation to expression of estradiol and progesterone receptors in equine endometrial fibrosis. *Advances in Bioscience and Biotechnology*, *04*(04), 599–605. https://doi.org/10.4236/abb.2013.44078

Mahon, G., & Cunningham, E. (1982). Inbreeding and the inheritance of fertility in the Thoroughbred mare. *Livestock Production Science*, *9*(6), 743–754. https://doi.org/10.1016/0301-6226(82)90021-5

Malpica, A., Matisic, J., Niekirk, D., Crum, C., Staerkel, G., Yamal, J.-M., Guillaud, M., Cox, D., Atkinson, E., Adler-Storthz, K., Poulin, N., MacAulay, C., & Follen, M. (2005). Kappa statistics to measure interrater and intrarater agreement for 1790 cervical biopsy specimens among twelve pathologists: Qualitative histopathologic analysis and

methodologic issues. *Gynecologic Oncology*, *99*(3), S38–S52. https://doi.org/10.1016/j.ygyno.2005.07.040

Mambelli, L., Mattos, R., Winter, G., Madeiro, D., Morais, B., Malschitzky, E., Miglino, M., Kerkis, A., & Kerkis, I. (2014). Changes in expression pattern of selected endometrial proteins following mesenchymal stem cells infusion in mares with endometrosis. *PLOS ONE*, *9*(6), e97889. https://doi.org/10.1371/journal.pone.0097889

Manning, S. (2002). *A prospective study of components of the breeding soundness evaluation, including endometrial biopsy, in mares selected for fertility.* University of Saskatchewan.

Matos, A., Baptista, C., Gärtner, M., & Rutteman, G. (2012). Prognostic studies of canine and feline mammary tumours: The need for standardized procedures. *The Veterinary Journal*, *193*(1), 24–31. https://doi.org/10.1016/j.tvjl.2011.12.019

McCue, P. (2019). *Clinical Equine Reproduction: Anatomy, Physiology, Pathology, and Breeding Management* (Vol. 1). Colorado State University.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Minkwitz, C., Schoon, H., Zhang, Q., & Schöniger, S. (2019). Plasticity of endometrial epithelial and stromal cells—A new approach towards the pathogenesis of equine endometrosis. *Reproduction in Domestic Animals*, *54*(6), 835–845. https://doi.org/10.1111/rda.13431

Morris, J. (1994). Information and observer disagreement in histopathology. *Histopathology*, *25*(2), 123–128. https://doi.org/10.1111/j.1365-2559.1994.tb01567.x

Mosli, M., Feagan, B., Zou, G., Sandborn, W., D'Haens, G., Khanna, R., Behling, C., Kaplan, K., Driman, D., Shackelton, L., Baker, K., MacDonald, J., Vandervoort, M., Samaan, M.,

Geboes, K., Valasek, M., Pai, R., Langner, C., Riddell, R., … Levesque, B. (2015). Reproducibility of histological assessments of disease activity in UC. *Gut*, *64*(11), 1765–1773. https://doi.org/10.1136/gutjnl-2014-307536

Müller-Unterberg, M., Wallmann, S., & Distl, O. (2017). Effects of inbreeding and other systematic effects on fertility of Black Forest Draught horses in Germany. *Acta Veterinaria Scandinavica*, *59*(1), 70. https://doi.org/10.1186/s13028-017-0338-4

Munkedal, D., Laurberg, S., Hagemann-Madsen, R., Stribolt, K., Krag, S., Quirke, P., & West, N. (2016). Significant individual variation between pathologists in the evaluation of colon cancer specimens after complete mesocolic excision. *Diseases of the Colon & Rectum*, *59*(10), 953–961. https://doi.org/10.1097/DCR.0000000000000671

Nambo, Y., Urayama, S., Ito, K., Shikichi, M., Orino, K., Watanabe, K., Ono, M., Ohtaki, T., & Tsumagari, S. (2014). Influence of age and endometrial biopsy score on the expression of lactoferrin in the uterus of mares. *Journal of Equine Veterinary Science*, *34*(1), 142. https://doi.org/10.1016/j.jevs.2013.10.097

Nath, L., Anderson, G., & McKinnon, A. (2010). Reproductive efficiency of Thoroughbred and Standardbred horses in north-east Victoria. *Australian Veterinary Journal*, *88*(5), 169–175. https://doi.org/10.1111/j.1751-0813.2010.00565.x

Nielsen, J., Nielsen, H., & Petersen, M. (2012). Diagnosis of equine endometritis- Microbiology, cytology and histology of endometrial biopsies and the correlation to fertility. *Pferdeheilkunde*, *28*(1), 8–13. https://doi.org/10.21836/PEM20120102

Northrup, N., Harmon, B., Gieger, T., Brown, C., Carmichael, K., Garcia, A., Latimer, K., Munday, J., Rakich, P., Richey, L., Stedman, N., Cheng, A., & Howerth, E. (2005). Variation among pathologists in histologic grading of canine cutaneous mast cell tumors.

Journal of Veterinary Diagnostic Investigation, *17*(3), 245–248. https://doi.org/10.1177/104063870501700305

Northrup, N., Howerth, E., Harmon, B., Brown, C., Carmicheal, K., Garcia, A., Latimer, K., Munday, J., Rakich, P., Richey, L., Stedman, N., & Gieger, T. (2005). Variation among pathologists in the histologic grading of canine cutaneous mast cell tumors with uniform use of a single grading reference. *Journal of Veterinary Diagnostic Investigation*, *17*(6), 561–564. https://doi.org/10.1177/104063870501700606

Oddsdóttir, C. (2007). *Development of endometrial fibrosis in the mare: Factors involved in tissue remodelling and collagen deposition* [Doctor of Philosphy]. University of Edinburgh.

OLYMPUS. (2020). *OlyVIA Net Image Server* (Build: 13778 (v1.0.6c37)) [Computer software].

Overbeck, W., Witte, T., & Heuwieser, W. (2011). Comparison of three diagnostic methods to identify subclinical endometritis in mares. *Theriogenology*, *75*(7), 1311–1318. https://doi.org/10.1016/j.theriogenology.2010.12.002

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.0) [Computer software]. R Foundation for Statistical Computing. http://R-project.org/

Ricketts, S., & Alonso, S. (1991a). Assessment of the breeding prognosis of mares using paired endometrial biopsy techniques. *Equine Veterinary Journal*, *23*(3), 185–188.

Ricketts, S., & Alonso, S. (1991b). The effect of age and parity on the development of equine chronic endometrial disease. *Equine Veterinary Journal*, *23*(3), 189–192.

Ricketts, S., & Barrelet, A. (1997). A retrospective review of the histopathological features seen in a series of 4241 endometrial biopsy samples collected from UK Thoroughbred mares over a 25 year period. *Pferdeheilkunde*, *13*(5), 525–530.

Riddle, W., LeBlanc, M., & Stromberg, A. (2007). Relationships between uterine culture, cytology and pregnancy rates in a Thoroughbred practice. *Theriogenology*, *68*(3), 395–402. https://doi.org/10.1016/j.theriogenology.2007.05.050

Robbins, P., Pinder, S., de Klerk, N., Dawkins, H., Harvey, J., Sterrett, G., Ellis, I., & Elston, C. (1995). Histological grading of breast carcinomas: A study of interobserver agreement. *Human Pathology*, *26*(8), 873–879. https://doi.org/10.1016/0046-8177(95)90010-1

Scheuer, P. (1997). Chronic hepatitis: What is activity and how should it be assessed? *Histopathology*, *30*(2), 103–105. https://doi.org/10.1046/j.1365-2559.1997.d01-588.x

Schilling, A. (2017). *Die endometriumbiopsie bei der stute- eine analyse der histologischen befunde zwischen 1992- 2012 am Leipziger Institut für Veterinär- Pathologie [The endometrial biopsy in the mare—An analysis of the histological findings between 1992-2012 at the Leipzig Institute for Veterinary Pathology]* [University of Leipzig].

Schlafer, D. (2007). Equine endometrial biopsy: Enhancement of clinical value by more extensive histopathology and application of new diagnostic techniques? *Theriogenology*, *68*(3), 413–422. https://doi.org/10.1016/j.theriogenology.2007.04.040

Schmitt-Gräff, A., Desmoulière, A., Gabbiani, G., Schmitt-Gräff, A., & Desmoulière, A. (1994). Heterogeneity of myofibroblast phenotypic features: An example of fibroblastic cell plasticity. *Virchows Archiv*, *425*(1), 3–24. https://doi.org/10.1007/BF00193944

Scholten, A., Smit, V., Beerman, H., van Putten, W., & Creutzberg, C. (2004). Prognostic significance and interobserver variability of histologic grading systems for endometrial carcinoma. *Cancer*, *100*(4), 764–772. https://doi.org/10.1002/cncr.20040

Schöniger, S., & Schoon, H. (2020). The healthy and diseased equine endometrium: A review of morphological features and molecular analyses. *Animals*, *10*(4), 625. https://doi.org/10.3390/ani10040625

Schoon, D., Schoon, H., & Klug, E. (1999). Angioses in the equine endometrium—Pathogenesis and clinical correlations. *Pferdeheilkunde*, *15*(6), 541–546.

Schoon, H., & Schoon, D. (2003). The Category I mare (Kenney and Doig 1986): Expected foaling rate 80-90%—Fact or fiction? *Pferdeheilkunde*, *19*(6), 698–701.

Schoon, H., Schoon, D., & Klug, E. (1992). Endometrial biopsies as an ancillary aid in diagnosis and prognosis of subfertility in the mare. *Pferdeheilkunde*, *8*(6), 355–362.

Schoon, H., Schoon, D., & Klug, E. (1997). The endometrial biopsy in the mare with regard to clinical correlations. *Pferdeheilkunde*, *13*(5), 453–464.

Schoon, H., Schoon, D., Wiegandt, I., Bartmann, C., & Aupperle, H. (1999). Endometrial maldifferentiation—A clinically significant diagnosis in equine reproduction? *Pferdeheilkunde*, *15*(6), 555–559.

Schoon, H., Wiegandt, I., Schoon, D., Aupperle, H., & Bartmann, C. (2000). Functional disturbances in the endometrium of barren mares a histological and immunohistological study. *Journal of Reproduction and Fertility, Supplement 56*, 381–391.

Scoggin, C. (2015). Not just a number: Effect of age on fertility, pregnancy and offspring vigour in thoroughbred brood-mares. *Reproduction, Fertility and Development*, *27*(6), 872. https://doi.org/10.1071/RD14390

Scott, C. (2020). A review of fungal endometritis in the mare. *Equine Veterinary Education*, *32*(8), 444–448. https://doi.org/10.1111/eve.13010

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19*(3), 321–325.

Sertich, P. (n.d.). *Breeding Soundness Examination of the Mare—Management and Nutrition*. Merck Veterinary Manual. Retrieved July 3, 2018, from

https://www.merckvetmanual.com/management-and-nutrition/management-of-reproduction-horses/breeding-soundness-examination-of-the-mare

Shideler, R., McChesney, A., Voss, J., & Squires, E. (1982). Relationship of endometrial biopsy and other management factors on fertility of broodmares. *Journal of Equine Veterinary Science*, *2*(1), 5–10. https://doi.org/10.1016/S0737-0806(82)80053-1

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Sikora, M., Nowak, M., Racheniuk, H., Wojtysiak, K., & Kozdrowski, R. (2017). Reliability of histopathological examination and immunohistochemistry of a single biopsy for evaluation of endometrial health in Icelandic mares. *Folia Histochemica et Cytobiologica*, *55*(3), 168–175. https://doi.org/10.5603/FHC.a2017.0017

Silcocks, P. (1983). Measuring repeatability and validity of histological diagnosis—A brief review with some practical examples. *Journal of Clinical Pathology*, *36*(11), 1269–1275. https://doi.org/10.1136/jcp.36.11.1269

Snider, T., Sepoy, C., & Holyoak, G. (2011). Equine endometrial biopsy reviewed: Observation, interpretation, and application of histopathologic data. *Theriogenology*, *75*(9), 1567–1581. https://doi.org/10.1016/j.theriogenology.2010.12.013

Stanton, M., Steiner, J., & Pugh, D. (2004). Endometrial cysts in the mare. *Journal of Equine Veterinary Science*, *24*(1), 14–19. https://doi.org/10.1016/j.jevs.2003.12.003

Stenkvist, B., Bengtsson, E., Eriksson, O., Jarkrans, T., Nordin, B., & Westman-Naeser, S. (1983). Histopathological systems of breast cancer classification: Reproducibility and clinical significance. *Journal of Clinical Pathology*, *36*(4), 392–398. https://doi.org/10.1136/jcp.36.4.392

Thomas, G., Dixon, M., Smeeton, N., & Williams, N. (1983). Observer variation in the histological grading of rectal carcinoma. *Journal of Clinical Pathology*, *36*(4), 385–391. https://doi.org/10.1136/jcp.36.4.385

Thompson, R., Armién, A., Rasmussen, J., & Wolf, T. (2014). Uterine adenocarcinoma in a Przewalksi's wild horse (Equus ferus przewalksi). *Journal of Zoo and Wildlife Medicine*, *45*(2), 441–445.

van Buiten, A., Westers, P., & Colenbrander, B. (2003). Male, female and management risk factors for non-return to service in Dutch mares. *Preventive Veterinary Medicine*, *61*(1), 17–26. https://doi.org/10.1016/S0167-5877(03)00128-4

Waelchli, R. (1990). Endometrial biopsy in mares under nonuniform breeding management conditions: Prognostic value and relationship with age. *Canadian Veterinary Journal*, *31*, 379–384.

Walter, I., Handler, J., Reifinger, M., & Aurich, C. (2001). Association of endometrosis in horses with differentiation of periglandular myofibroblasts and changes of extracellular matrix proteins. *Reproduction, 121*, 581–586.

Walter, I., Helmreich, M., Handler, J., & Aurich, C. (2003). Mineralised deposits in the uterine glands of mares with chronic endometrial degeneration. *Veterinary Record*, *153*(23), 708–710. https://doi.org/10.1136/vr.153.23.708

Warners, M., Ambarus, C., Bredenoord, A., Verheij, J., Lauwers, G., Walsh, J., Katzka, D., Nelson, S., van Viegen, T., Furuta, G., Gupta, S., Stitt, L., Zou, G., Parker, C., Shackelton, L., D`Haens, G., Sandborn, W., Dellon, E., Feagan, B., … Pai, R. (2018). Reliability of histologic assessment in patients with eosinophilic oesophagitis. *Alimentary Pharmacology & Therapeutics*, *47*(7), 940–950. https://doi.org/10.1111/apt.14559

Westin, J., Lagging, L., Wejstål, R., Norkrans, G., & Dhillon, A. (1999). Interobserver study of liver histopathology using the Ishak score in patients with chronic hepatitis C virus infection. *Liver*, *19*(3), 183–187. https://doi.org/10.1111/j.1478-3231.1999.tb00033.x

Wolfe, M., Ellis, L., & MacMullen, R. (1989). Reproductive rates of feral horses and burros. *The Journal of Wildlife Management*, *53*(4), 916–924. https://doi.org/10.2307/3809588

Woodward, E., Christoffersen, M., Campos, J., Squires, E., & Troedsson, M. (2012). Susceptibility to persistent breeding-induced endometritis in the mare: Relationship to endometrial biopsy score and age, and variations between seasons. *Theriogenology*, *78*(3), 495–501. https://doi.org/10.1016/j.theriogenology.2012.02.028

APPENDIX A: Supplemental Materials and Methods

Table A.1. Example of questions associated with each endometrial biopsy slide given to pathologists for evaluation via Survey Monkey software (San Mateo, California, USA) (http://www.surveymonkey.com).

Question	Response (Must select one of four options)			
Please describe the inflammation.	Absent	Mild	Moderate	Severe
Please describe the fibrosis.	Absent	Mild	Moderate	Severe
Please describe the lymphatic lacunae.	Absent	Mild	Moderate	Severe
Please describe the glandular atrophy.	Absent	Mild	Moderate	Severe
Please grade the biopsy using one of the Kenney-Doig categories.	I	IIA	IIB	III
Optional additional comments				

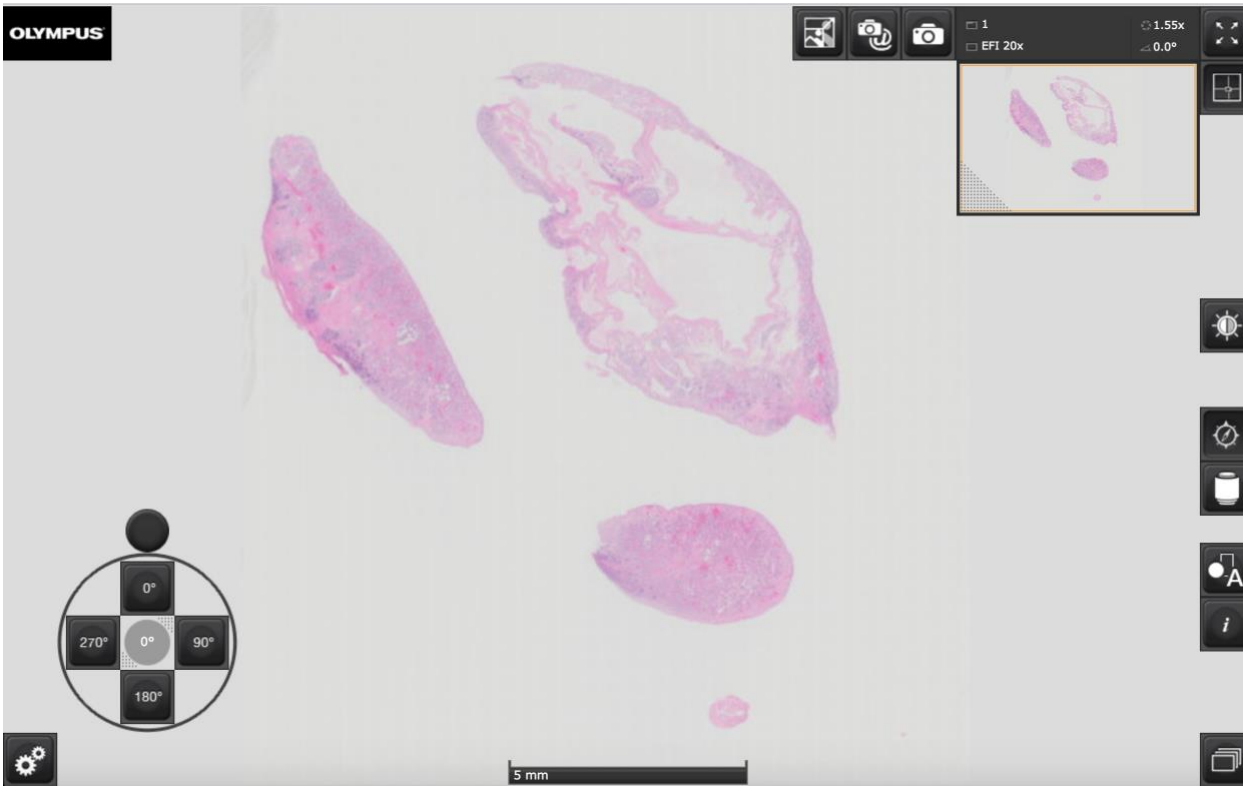

Figure A.1. Example of the OlyVIA web viewer window used by observers to evaluate slides (OLYMPUS, 2020). A hyperlink for each slide was embedded in the Survey Monkey software to open to the corresponding viewing window in the OlyVIA server.

APPENDIX B: Supplemental Results

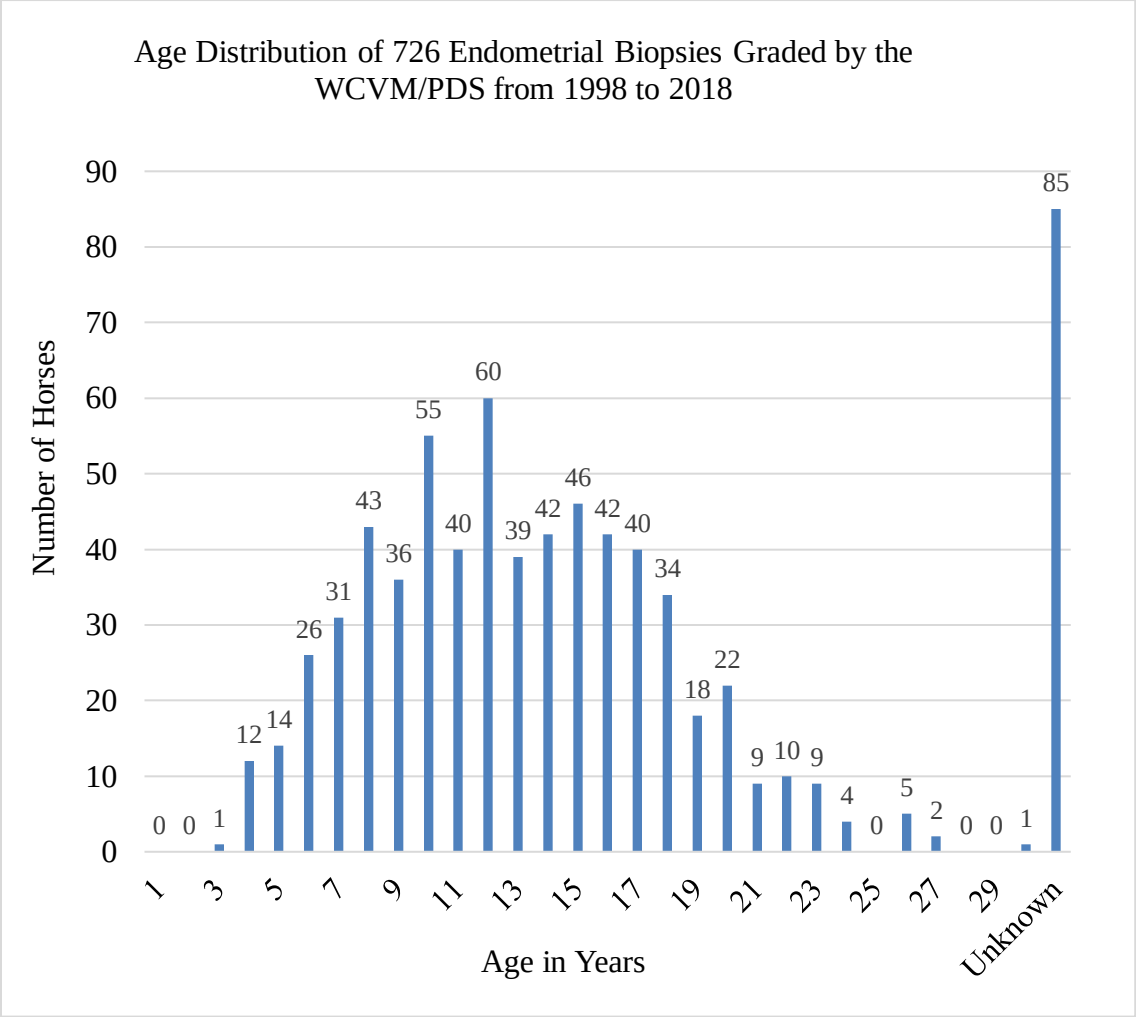

Figure B.1. Distribution of age categories within the 726 endometrial biopsy database collected from the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS) from 1998 to 2018.

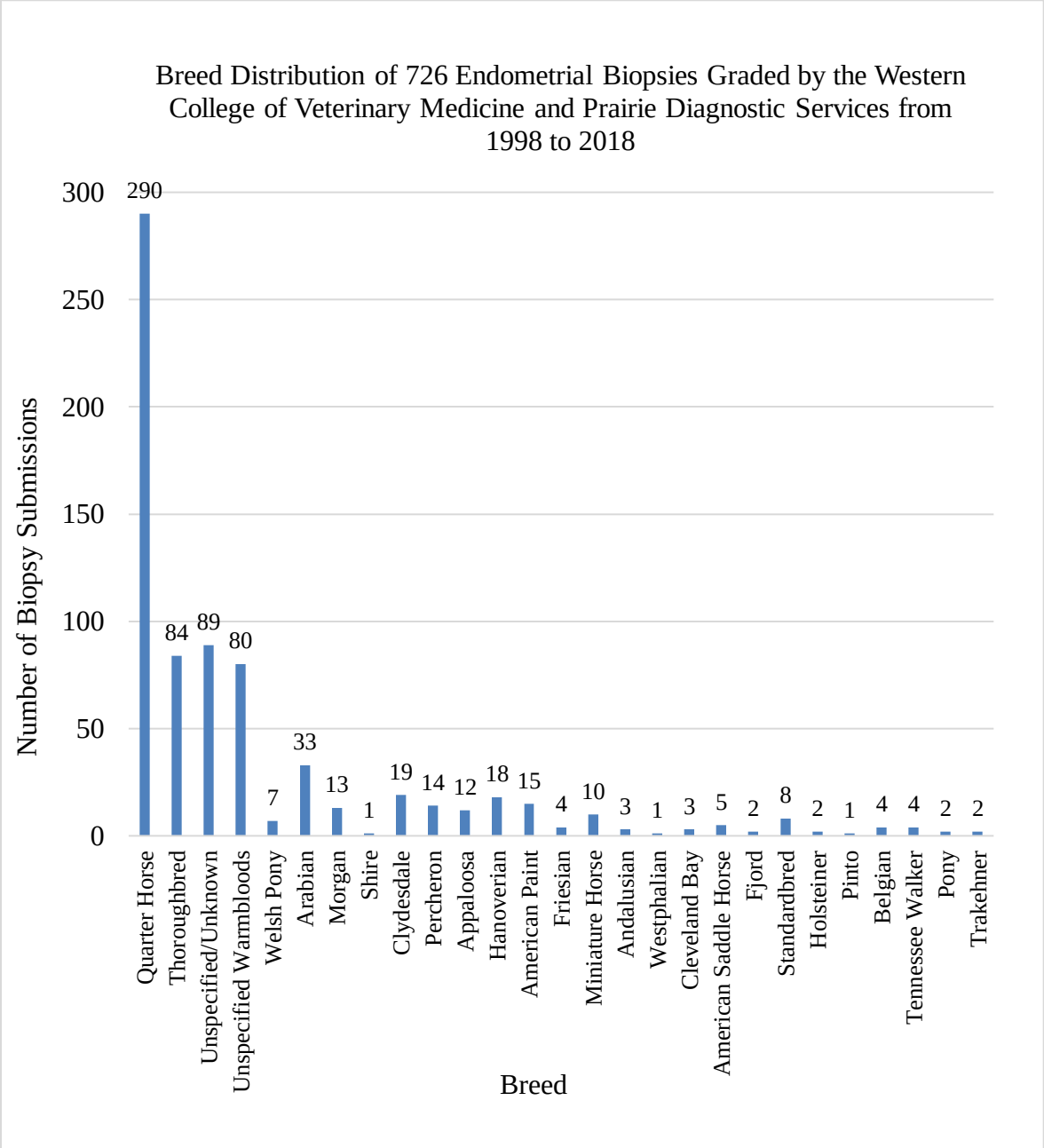

Figure B.2. Distribution of different breeds within the 726 endometrial biopsy database collected from the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS) from 1998 to 2018.

Table B.1. P-values* for Fisher's exact test pairwise comparison of five individual pathologists' Kenney-Doig grade distributions found at the Western College of Veterinary Medicine and Prairie Diagnostic Services. Significant differences are in **bold**.

Pathologist	A	B	C	D	E
A	X				
B	**<0.001**	X			
C	0.006	**<0.001**	X		
D	**<0.001**	**<0.001**	**<0.001**	X	
E	0.065	**<0.001**	**<0.001**	**<0.001**	X

*Bonferroni adjusted p-value significance threshold set at 0.005

Table B.2. Chi-square results comparing the Kenney-Doig grade distribution found at the Western College of Veterinary Medicine (WCVM) and Prairie Diagnostic Services (PDS) against six studies identified in the literature. Significant differences are in **bold**.

Study from the Literature vs WCVM/PDS	X^2 value (degrees of freedom)	P-value*
Ricketts and Alonso	252.41 (3)	**<0.001**
Waelchli	244.8 (3)	**<0.001**
Nambo et al.	117.72 (3)	**<0.001**
Kilgenstein et al.	143.1 (3)	**<0.001**
Kabisch et al.	321.47 (3)	**<0.001**
Schilling	258.04 (3)	**<0.001**

*Bonferroni adjusted p-value significance threshold set at 0.008

Table B.3. Frequency distributions for eight pathologists' Kenney-Doig grades assigned to the same set of 63 endometrial biopsies.

Pathologist	Frequency of Kenney-Doig Grades Assigned			
	I	IIA	IIB	III
1	1.59%	52.38%	36.51%	9.52%
2	3.17%	11.11%	57.14%	28.57%
3	14.29%	63.49%	20.63%	1.59%
4	3.17%	28.57%	44.44%	23.81%
5	17.46%	41.27%	28.57%	12.70%
6	7.94%	36.51%	22.22%	33.33%
7	1.59%	26.98%	23.81%	47.62%
8	9.52%	20.63%	47.62%	22.22%

Table B.4. Unweighted (*italics*) and weighted (**bold**) Cohen's kappa coefficient measuring inter-rater agreement between eight pathologists' Kenney-Doig grades assigned to the same set of 63 endometrial biopsies. Inter-institution group comparisons in green, intra-institution comparisons in blue.

Pathologist	1	2	3	4	5	6	7	8
1	X	**0.224***	**0.082**	**0.484***	**0.465***	**0.281***	**0.212***	**0.308***
2	*0.100*	X	**0.087**	**0.497***	**0.374***	**0.468***	**0.338***	**0.375***
3	*-0.052*	*-0.009*	X	**0.184***	**0.213**	**0.187***	**0.094**	**0.093**
4	*0.260**	*0.287**	*-0.016*	X	**0.519***	**0.638***	**0.513***	**0.620***
5	*0.210**	*0.174**	*0.318**	*0.161**	X	**0.542***	**0.405***	**0.452***
6	*0.109*	*0.287**	*0.145**	*0.378**	*0.195**	X	**0.471***	**0.540***
7	*0.230**	*0.158**	*0.031*	*0.300**	*0.120*	*0.216**	X	**0.376***
8	*0.133*	*0.229**	*0.031*	*0.458**	*0.265**	*0.333**	*0.166**	X

*p-value < 0.05

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table B.5. Percent agreement measuring inter-rater agreement between eight pathologists' Kenney-Doig grades assigned to the same set of 63 endometrial biopsies. Green values represent those between pathologists at different institutions while blue values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X							
2	36.5%	X						
3	38.1%	19.0%	X					
4	50.8%	54.0%	27.0%	X				
5	47.6%	38.1%	55.6%	39.7%	X			
6	38.1%	47.6%	39.7%	55.6%	41.3%	X		
7	44.4%	41.3%	25.4%	50.8%	33.3%	46.0%	X	
8	39.7%	50.8%	27.0%	63.5%	46.0%	50.8%	39.7%	X

Table B.6. Frequency distributions for eight pathologists' evaluation of histologic inflammation as either absent, mild, moderate or severe descriptors assigned to the same set of 63 endometrial biopsies.

Pathologist	Frequency of Descriptor Assigned			
	Absent	Mild	Moderate	Severe
1	68.25%	28.57%	3.17%	0.00%
2	14.29%	50.79%	23.81%	11.11%
3	26.98%	63.49%	9.52%	0.00%
4	1.59%	34.92%	50.79%	12.70%
5	14.29%	44.44%	33.33%	7.94%
6	9.52%	60.32%	23.81%	6.35%
7	12.79%	53.97%	28.57%	4.76%
8	4.76%	36.51%	46.03%	12.70%

Table B.7. Frequency distributions for eight pathologists' evaluation of histologic fibrosis as either absent, mild, moderate or severe descriptors assigned to the same set of 63 endometrial biopsies.

Pathologist	Frequency of Descriptor Assigned			
	Absent	Mild	Moderate	Severe
1	4.76%	60.32%	31.75%	3.17%
2	3.17%	58.73%	25.40%	12.70%
3	52.38%	38.10%	9.52%	0.00%
4	38.10%	39.68%	15.87%	6.35%
5	31.75%	38.10%	23.81%	6.35%
6	15.87%	41.27%	23.81%	19.05%
7	9.52%	26.98%	17.46%	46.03%
8	31.75%	39.68%	23.81%	4.76%

Table B.8. Frequency distributions for seven pathologists' evaluation of histologic glandular atrophy as either absent, mild, moderate or severe descriptors assigned to the same set of 63 endometrial biopsies. One pathologist did not assess this marker therefore data is not included.

Pathologist	Frequency of Descriptor Assigned			
	Absent	Mild	Moderate	Severe
A	7.93%	49.21%	41.27%	1.59%
B	7.94%	33.33%	46.03%	12.70%
C	22.22%	44.44%	30.16%	3.17%
D	26.98%	41.27%	26.98%	4.76%
E	61.90%	17.46%	11.11%	9.52%
F	20.63%	34.92%	39.68%	4.76%
G	33.33%	39.68%	19.05%	7.94%

Table B.9. Frequency distributions for eight pathologists' evaluation of histologic lymphatic lacunae as either absent, mild, moderate or severe descriptors assigned to the same set of 63 endometrial biopsies.

Pathologist	Frequency of Descriptor Assigned			
	Absent	Mild	Moderate	Severe
1	14.29%	68.25%	14.29%	3.17%
2	6.35%	50.79%	36.51%	6.35%
3	49.21%	34.92%	15.87%	0.00%
4	77.78%	12.70%	9.52%	0.00%
5	23.81%	58.73%	12.70%	4.76%
6	55.56%	33.33%	11.11%	0.00%
7	47.62%	23.81%	20.63%	7.94%
8	34.92%	34.92%	23.81%	6.35%

Table B.10. Unweighted (*italics*) Cohen's kappa coefficient and weighted (**bold**) measuring inter-rater agreement between eight pathologists' evaluation of histologic inflammation as either absent, mild, moderate or severe assigned to the same set of 63 endometrial biopsies. Green kappa values represent those between pathologists at different institutions while blue kappa values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X	**0.438***	**0.100**	**0.398***	**0.340***	**0.433***	**0.470***	**0.209**
2	*0.181**	X	**0.183**	**0.530***	**0.614***	**0.680***	**0.574***	**0.351***
3	*0.088*	*0.15**	X	**0.157***	**0.170**	**0.181**	**0.144**	**-0.033**
4	*0.171**	*0.259**	*-0.050*	X	**0.647***	**0.478***	**0.431***	**0.412***
5	*0.0194*	*0.332**	*0.044*	*0.306**	X	**0.667***	**0.479***	**0.409***
6	*0.204**	*0.433**	*0.135*	*0.181**	*0.374**	X	**0.533***	**0.451***
7	*0.277**	*0.299**	*0.068*	*0.204**	*0.0866*	*0.248**	X	**0.355***
8	*0.071*	*0.118*	*-0.071*	*0.285**	*0.287**	*0.252**	*0.182**	X

* p-value <0.05

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table B.11. Unweighted (*italics*) Cohen's kappa coefficient and weighted (**bold**) measuring inter-rater agreement between eight pathologists' evaluation of histologic fibrosis as either absent, mild, moderate or severe assigned to the same set of 63 endometrial biopsies. Green kappa values represent those between pathologists at different institutions while blue kappa values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X	**0.197**	**0.114**	**0.188**	**0.250***	**0.339***	**0.164**	**0.172**
2	*0.036*	X	**0.120**	**0.527***	**0.338***	**0.612***	**0.228***	**0.344***
3	*0.068*	*0.007*	X	**0.273***	**0.179**	**0.094**	**0.044**	**0.148**
4	*0.011*	*0.146**	*0.124*	X	**0.616***	**0.603***	**0.236***	**0.640***
5	*0.157**	*0.135**	*-0.001*	*0.445**	X	**0.423***	**0.348***	**0.530***
6	*0.089*	*0.406**	*0.095*	*0.213**	*0.123*	X	**0.392***	**0.479***
7	*0.126**	*0.029*	*-0.013*	*0.027*	*0.103*	*0.126*	X	**0.236***
8	*0.050*	*0.149**	*0.134*	*0.440**	*0.331**	*0.074*	*0.027*	X

*p-value<0.05

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table B.12. Unweighted (*italics*) Cohen's kappa coefficient and weighted (**bold**) measuring inter-rater agreement between seven pathologists' evaluation of histologic glandular atrophy as either absent, mild, moderate or severe assigned to the same set of 63 endometrial biopsies. Green kappa values represent those between pathologists at different institutions while blue kappa values represent those between pathologists at the same institution. One pathologist did not assess this marker and was excluded.

Pathologist	A	B	C	D	E	F	G
A	X	**0.373***	**0.481***	**0.178**	**0.306***	**0.485***	**0.358***
B	*0.029*	X	**0.215***	**0.387***	**0.244***	**0.356***	**0.227***
C	*0.254**	*-0.033*	X	**0.238**	**0.460***	**0.499***	**0.573***
D	*0.067*	*0.040*	*0.0812*	X	**0.182**	**0.157**	**0.281***
E	*0.088*	*-0.014*	*0.173**	*0.017*	X	**0.346***	**0.393***
F	*0.313**	*0.158**	*0.204**	*0.081*	*0.147**	X	**0.287***
G	*0.205**	*0.061*	*0.217**	*0.035*	*0.133*	*0.088*	X

*p-value <0.05

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table B.13. Unweighted (*italics*) Cohen's Kappa Coefficient and weighted (**bold**) measuring inter-rater agreement between eight pathologists' evaluation of histologic lymphatic lacunae as either absent, mild, moderate or severe assigned to the same set of 63 endometrial biopsies. Green kappa values represent those between pathologists at different institutions while blue kappa values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X	**0.253***	**0.211***	**0.218***	**0.329***	**0.301***	**0.232***	**0.506***
2	*0.247**	X	**0.020**	**0.184***	**0.087**	**0.257***	**0.304***	**0.369***
3	*-0.021*	*0.026*	X	**0.168**	**-0.009**	**0.021**	**0.154**	**0.113**
4	*-0.006*	*0.031*	*0.089*	X	**0.233***	**0.439***	**0.262***	**0.378***
5	*0.214**	*0.004*	*-0.014*	*0.084*	X	**0.093**	**0.161**	**0.341***
6	*0.039*	*0.012*	*0.116*	*0.353**	*0.107*	X	**0.241***	**0.406***
7	*0.031*	*0.153**	*0.047*	*0.179**	*0.136*	*0.148*	X	**0.363***
8	*0.295**	*0.262**	*0.074*	*0.208**	*0.111*	*0.186**	*0.248**	X

*p-value <0.05

Kappa statistics were interpreted using standards suggested by Landis and Koch (1977) where <0.00 is poor, 0.00 – 0.20 is slight, 0.21 – 0.40 is fair, 0.41 – 0.60 is moderate, 0.61 – 0.80 is substantial, and 0.81 – 1.00 is almost perfect agreement.

Table B.14. Percent agreement measuring inter-rater agreement between eight pathologists' evaluation of histologic inflammation as either absent, mild, moderate, or severe assigned to the same set of 63 endometrial biopsies. Green values represent those between pathologists at different institutions while blue values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X							
2	52.4%	X						
3	50.8%	47.6%	X					
4	49.2%	49.2%	23.8%	X				
5	41.3%	55.6%	38.1%	54.0%	X			
6	58.7%	65.1%	50.8%	46.0%	60.3%	X		
7	60.3%	55.6%	44.4%	47.6%	41.3%	55.6%	X	
8	42.9%	39.7%	23.8%	55.6%	52.4%	50.8%	46.0%	X

Table B.15. Percent agreement measuring inter-rater agreement between eight pathologists' evaluation of histologic fibrosis as either absent, mild, moderate, or severe assigned to the same set of 63 endometrial biopsies. Green values represent those between pathologists at different institutions while blue values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X							
2	46.0%	X						
3	33.3%	27.0%	X					
4	31.7%	39.7%	44.4%	X				
5	42.9%	39.7%	33.3%	61.9%	X			
6	39.7%	60.3%	33.3%	42.9%	36.5%	X		
7	33.3%	28.6%	15.9%	22.2%	28.6%	34.9%	X	
8	36.5%	41.3%	42.9%	61.9%	54.0%	33.3%	22.2%	X

Table B.16. Percent agreement measuring inter-rater agreement between seven pathologists' evaluation of histologic glandular atrophy as either absent, mild, moderate, or severe assigned to the same set of 63 endometrial biopsies. Green values represent those between pathologists at different institutions while blue values represent those between pathologists at the same institution. One pathologist did not assess this marker and was excluded.

Pathologist	A	B	C	D	E	F	G
A	X						
B	38.1%	X					
C	52.4%	28.6%	X				
D	38.1%	31.7%	38.1%	X			
E	25.4%	15.9%	38.1%	28.6%	X		
F	55.6%	42.9%	46.0%	36.5%	34.9%	X	
G	44.4%	30.2%	46.0%	33.3%	39.7%	34.9%	X

Table B.17. Percent agreement measuring inter-rater agreement between eight pathologists' evaluation of histologic lymphatic lacunae as either absent, mild, moderate, or severe assigned to the same set of 63 endometrial biopsies. Green values represent those between pathologists at different institutions while blue values represent those between pathologists at the same institution.

Pathologist	1	2	3	4	5	6	7	8
1	X							
2	55.6%	X						
3	31.7%	28.6%	X					
4	20.6%	17.5%	49.2%	X				
5	57.1%	36.5%	33.3%	33.3%	X			
6	34.9%	25.4%	47.6%	66.7%	41.3%	X		
7	28.6%	34.9%	38.1%	52.4%	38.1%	46.0%	X	
8	52.4%	47.6%	38.1%	47.6%	39.7%	46.0%	47.6%	X

Table B.18. Intraclass correlation coefficients (ICC) and 95% confidence intervals for eight individual pathologists measuring intra-agreement of Kenney-Doig grades made on one set of 21 endometrial biopsy slides graded at two separate time points.

Pathologist	ICC	Confidence Interval	Reliability based on ICC	Reliability based on Confidence Interval
1	0.58*	0.2 < ICC < 0.807	Moderate	Moderate to Good
2	0.524*	0.121 < ICC < 0.776	Moderate	Poor to Good
3	0.116	-0.286 < ICC < 0.501	Poor	Poor to Moderate
4	0.774*	0.521 < ICC < 0.902	Good	Moderate to Excellent
5	0.698*	0.303 < ICC < 0.875	Moderate	Poor to Good
6	0.708*	O.403 < ICC < 0.871	Moderate	Poor to Good
7	0.444*	0.018 < ICC < 0.732	Poor	Poor to Moderate
8	0.583*	0.21 < ICC < 0.807	Moderate	Poor to Good

*p-value < 0.05

ICC values were interpreted using standards suggested by Koo and Li (2016) where <0.50 is poor, 0.50 – 0.75 is moderate, 0.75-0.90 is good, and >0.90 is excellent agreement.

Table B.19. Intraclass correlation coefficients (ICC) and 95% confidence intervals for eight individual pathologists measuring intra-rater agreement for grading of inflammation with descriptive modifiers of either absent, mild, moderate or severe in the same set of 21 endometrial biopsy slides evaluated at two separate time points.

Pathologist	ICC	Confidence Interval	Reliability based on ICC	Reliability based on Confidence Interval
1	0.333*	-0.049 < ICC < 0.649	Poor	Poor to Moderate
2	0.703*	0.4 < ICC < 0.867	Moderate	Moderate to Good
3	0.312	-0.129 < ICC < 0.65	Poor	Poor to Moderate
4	0.742*	0.421 < ICC < 0.891	Moderate	Moderate to Good
5	0.582*	0.208 < ICC < 0.807	Moderate	Poor to Good
6	0.602*	0.235 < ICC < 0.818	Moderate	Poor to Good
7	0.574*	0.192 < ICC < 0.803	Moderate	Poor to Good
8	0.560*	0.185 < ICC < 0.794	Moderate	Poor to Good

*p-value < 0.05

ICC values were interpreted using standards suggested by Koo and Li (2016) where <0.50 is poor, 0.50 – 0.75 is moderate, 0.75-0.90 is good, and >0.90 is excellent agreement.

Table B.20. Intraclass correlation coefficients (ICC) and 95% confidence intervals for eight individual pathologists measuring intra-rater agreement for grading of fibrosis with descriptive modifiers of either absent, mild, moderate or severe in the same set of 21 endometrial biopsy slides evaluated at two separate time points.

Pathologist	ICC	Confidence Interval	Reliability based on ICC	Reliability based on Confidence Interval
1	0.478*	0.059 < ICC < 0.751	Poor	Poor to Good
2	0.560*	0.185 < ICC < 0.794	Moderate	Poor to Good
3	0.290	-0.08 < ICC < 0.615	Poor	Poor to Moderate
4	0.745*	0.467 < ICC < 0.888	Moderate	Poor to Good
5	0.683*	0.375 < ICC < 0.857	Moderate	Moderate to Good
6	0.756*	0.486 < ICC < 0.894	Good	Poor to Good
7	0.451*	0.022 < ICC < 0.736	Poor	Poor to Moderate
8	0.633*	0.285 < ICC < 0.833	Moderate	Poor to Good

*p-value < 0.05

ICC values were interpreted using standards suggested by Koo and Li (2016) where <0.50 is poor, 0.50 – 0.75 is moderate, 0.75-0.90 is good, and >0.90 is excellent agreement.

Table B.21. Intra-class correlation coefficients (ICC) and 95% confidence intervals for eight individual pathologists measuring intra-rater agreement for grading of glandular atrophy with descriptive modifiers of either absent, mild, moderate or severe in the same set of 21 endometrial biopsy slides evaluated at two separate time points. One pathologist did not assess this marker and is excluded.

Pathologist	ICC	Confidence Interval	Reliability based on ICC	Reliability based on Confidence Interval
A	0.423*	0.027 < ICC < 0.712	Poor	Poor to Moderate
B	0.458 *	0.064 < ICC < 0.734	Poor	Poor to Moderate
C	0.802*	0.571 < ICC < 0.915	Good	Moderate to Good
D	0.184	-0.167 < ICC < 0.531	Poor	Poor to Moderate
E	0.355*	-0.067 < ICC < 0.673	Poor	Poor to Moderate
F	0.261	-0.166 < ICC < 0.611	Poor	Poor to Moderate
G	0.592*	0.219 < ICC < 0.813	Moderate	Poor to Good

*p-value < 0.05

ICC values were interpreted using standards suggested by Koo and Li (2016) where <0.50 is poor, 0.50 – 0.75 is moderate, 0.75-0.90 is good, and >0.90 is excellent agreement.

Table B.22. Intra-class correlation coefficients (ICC) and 95% confidence intervals for eight individual pathologists measuring intra-rater agreement for grading of lymphatic lacunae with descriptive modifiers of either absent, mild, moderate or severe in the same set of 21endometrial biopsy slides evaluated at two separate time points.

Pathologist	ICC	Confidence Interval	Reliability based on ICC	Reliability based on Confidence Interval
1	0.116	-0.286 < ICC < 0.501	Poor	Poor to Moderate
2	0.387*	-0.011 < ICC < 0.689	Poor	Poor to Moderate
3	-0.156	-0.574 < ICC < 0.302	Poor	Poor
4	0.639*	0.304 < ICC < 0.835	Moderate	Poor to Good
5	0.154	-0.301 < ICC < 0.546	Poor	Poor to Moderate
6	0.0968	-0.295 < ICC < 0.482	Poor	Poor
7	0.191	-0.254 < ICC < 0.569	Poor	Poor to Moderate
8	0.420*	-0.019 < ICC < 0.718	Poor	Poor to Moderate

*p-value < 0.05

ICC values were interpreted using standards suggested by Koo and Li (2016) where <0.50 is poor, 0.50 – 0.75 is moderate, 0.75-0.90 is good, and >0.90 is excellent agreement.

Table B.23. Predicted probabilities of a biopsy being assigned to a certain Kenney-Doig category based on logistic regression modelling of the influence of evaluating histologic inflammation as either absent, mild, moderate, or severe.

Kenney-Doig Category	Severity of Inflammation			
	Absent	Mild	Moderate	Severe
I	50%	4.6%	0%	0%
IIA	33.3%	44.5%	18.2%	0%
IIB	2.8%	37.7%	52.7%	5.4%
III	13.9%	13.2%	29.1%	94.6%

Table B.24. Predicted probabilities of a biopsy being assigned to a certain Kenney-Doig category based on logistic regression modelling of the influence of evaluating histologic fibrosis as either absent, mild, moderate, or severe.

Kenney-Doig Category	Severity of Fibrosis			
	Absent	Mild	Moderate	Severe
I	31.8%	0.5%	0%	0%
IIA	54.1%	44.3%	5.9%	0%
IIB	12.9%	47.4%	58.8%	3.2%
III	1.2%	7.8%	35.3%	96.8%

Table B.25. Predicted probabilities of a biopsy being assigned to a certain Kenney-Doig category based on logistic regression modelling of the influence of evaluating histologic glandular atrophy as either absent, mild, moderate, or severe.

Kenney-Doig Category	Severity of Glandular Atrophy			
	Absent	Mild	Moderate	Severe
I	14%	6.1%	1.5%	0%
IIA	51.8%	33.5%	14.8%	10.7%
IIB	24.6%	41.5%	43.7%	32.2%
III	9.7%	18.9%	40%	57.1%

Table B.26. Predicted probabilities of a biopsy being assigned to a certain Kenney-Doig category based on logistic regression modelling of the influence of evaluating histologic lymphatic lacunae as either absent, mild, moderate, or severe.

Kenney-Doig Category	Severity of Lymphatic Lacunae			
	Absent	Mild	Moderate	Severe
I	9.2%	6.2%	2.5%	0%
IIA	43.3%	33.2%	8.6%	0%
IIB	31.7%	38.8%	42%	50%
III	15.9%	21.9%	46.9%	50%