BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Thesis

INTERRATER VARIABILITY BETWEEN LOCAL AND CENTRAL PATHOLOGISTS IN AN INDUSTRY SPONSORED ADJUDICATION PROGRAM

by

ALISON MICHELE OCCHIUTI

B.A., Northwestern University, 2010

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2017

Approved by

First Reader _____

Janice Weinberg, Sc.D.
Director, M.S. in Clinical Investigation Program

Second Reader _____

Lindsay McNair, M.D., M.P.H., M.S.B.
Chief Medical Officer and President, Consulting Services at WIRB-Copernicus Group

Third Reader _____

Glenn Bubley, M.D.
Director of Genitourinary Oncology
Beth Israel Deaconess Medical Center

ACKNOWLEDGMENTS

Thank you to Dr. Janice Weinberg, Dr. Lindsay McNair, Dr. Glenn Bubley, and Stacey Hess Pino for all of their encouragement and support.

INTERRATER VARIABILITY BETWEEN LOCAL AND CENTRAL PATHOLOGISTS IN AN INDUSTRY SPONSORED ADJUDICATION PROGRAM

ALISON MICHELE OCCHIUTI

ABSTRACT

Background: Adjudication is a standardized, objective, and often blinded mechanism designed to assess clinical events with increased accuracy. It is performed by a centralized committee of independent reviewers, who are specialized, expert physicians who have no involvement with either the treatment of study subjects or the trial sponsor. Adjudication can decrease variability and bias in study results and increase the likelihood of correct identification, assessment, and categorization of clinical events such as potential malignancies diagnosed through histopathology. Histopathology is highly variable due to the subjective nature of the assessments.

Thesis: If it is the case that there are clinically significant discrepancies between local and central diagnoses and that central adjudication yields more accurate diagnoses than a local pathologist, then it should be accepted that adjudication ought to be more widely used in clinical trials to assess histopathology-related safety outcomes and endpoints.

Methods and Statistics: This is a retrospective cross-sectional study assessing interrater variability between local and central diagnoses of biopsy samples in a clinical trial setting using kappa scores and percent agreement. Certified Professional Coders (CPC) and central pathologists used the International Classification of Diseases for Oncology revision 3 (ICD-O 3) to codify the local and central assessments to permit comparison.

Three statistical groups (group A: the full dataset, group B: pathology sub-specialty reading groups, and group C: non-melanoma skin cancers versus all other malignancies) were assessed for interrater variability in seven separate analyses: neoplasm versus non-neoplasm (analysis 1), benign versus malignant including non-neoplasms (analysis 2.1), benign versus malignant excluding non-neoplasms (analysis 2.2), discrepancies in morphology and/or behavior including non-neoplasms (analysis 3.1), discrepancies in morphology and/or behavior excluding non-neoplasms (analysis 3.2), all discrepancies leading to differences in treatment (analysis 4.1), and all discrepancies leading to difference in treatment with round 1 matches removed (analysis 4.2).

Results: 602 cases comprised the dataset. Based on kappa scores, there is near perfect agreement between the central and local lab diagnoses in analyses 1, 2.1, and 2.2 in group A (all cases in the dataset). The percent agreement for these analyses is above 90%. The group A (full dataset) kappa score and percent agreement decreased to 0.59 and 68.3%, respectively, in analysis 3.1 (discrepancies in morphology and/or behavior codes, including non-neoplasms). When non-neoplasms were removed (analysis 3.2), the kappa score and percent agreement were 0.52 and 57.0%, respectively. In group C, NMSC had substantial kappa agreement in analyses 1, 2.1, and 2.2, whereas all other malignancies had near perfect kappa agreement. All percent agreements were above 88% and surpassed the minimally acceptable threshold for interrater percent agreement in healthcare (80%). Group B divided the data set into 10 sub-specialty reading groups. Kappa scores ranged from 0.66 (GYN) to 1.00 (lung) in analysis 1; the analysis 1 kappa score for lymphoma was 0.55, but this was not statistically significant. In analysis 2.1,

lung and sarcoma had the highest kappa scores (1.00) and dermatology and GYN had the lowest (0.71). As in analysis 1, the kappa score for lymphoma was 0.55 but was not statistically significant. When non-neoplasms were removed from analysis 2.2, 6 of the 10 sub-groups had kappa scores of 1.00, but all 6 had sample sizes less than 10. Percent agreement ranged from 80 to 100 percent. When all cases were considered regardless of number of rounds of review (analysis 4.1), about 90% of diagnoses would have similar courses of treatment. All sub-groups except sarcoma reached the minimally acceptable agreement rate in healthcare (80%). In the remaining 33% of cases that did not have matching diagnoses in round 1 (analysis 4.2), 34% may have different courses of treatment depending on whether the local or central diagnoses was used. Mid-study updates to the charter and CPC/reviewer manuals and processing of specimens did not have a significant impact on results.

Conclusion: Although there is little discrepancy between local and central pathologists on whether malignancies exist among samples, there is discord regarding specific diagnoses and their associated treatments. Adjudication can assist in decreasing this discordance in order to develop the most specific and accurate safety profile for a compound.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANDA……Accelerated new drug application

CDC……. Centers for Disease Control and Prevention

CEC……Clinical events committees

CEE……. Conjugated equine estrogens

CFR……. Code of Federal Regulations

CI…….....Confidence interval

COPD……. Chronic obstructive pulmonary disease

CPC……. Certified Professional Coder

CRF……. Case report form

CRO……...……...…. Contract Research Organization

DSMB…… Data safety monitoring board

EBV…….……Epstein Barr Virus

eCRF……...Electronic case report form

EMA……European Medicines Agency

ENT……...Ear, nose, and throat

FDA……. Food and Drug Administration

FDAAA……. Food and Drug Administration Amendments Act of 2007

FDAMA……Food and Drug Modernization Act of 1997

GI……. Gastrointestinal

GU……...Genitourinary

GYN……...Gynecology

H&E……. Hematoxylin and Eosin

ICD-10……. International Statistical Classification of Diseases and Related Health Problems, Revision 10

ICD-O 3……. International Classification of Diseases for Oncology Revision 3

ICR……. Independent central review

IHC……. Immunohistochemical

IND……..……. Investigational New Drug Application

LPR…….……..…Local pathology report

MedDRA®…….....……. Medical Dictionary of Regulatory Activities

MI……Myocardial infarction

MOTNAC……...……Manual of Tumor Nomenclature and Coding

MPA……Medroxyprogesterone acetate

NDA……New drug application

NIACC……. National Inter-Observer Agreement in Colorectal Cancer study

PARAGON-B…… Platelet IIb/IIIa Antagonist for the Reduction of Acute Coronary Syndrome Events in a Global Organization Network Trial

PHI…….……Protected Health Information

R&D……Research and development

TRIM……. The Thrombin Inhibition of Myocardial Ischemia Study Group

UK……United Kingdom

WHI……..……. Women's Health Initiative

WHO……. World Health Organization

INTRODUCTION

The goal of pharmaceutical development is to bring effective drugs and treatments to the market through the most cost-effective process. The most efficient clinical trials yield unambiguous data on efficacy and safety. Overwhelming success on both counts permits rapid advancement to the next phase of trials and decreases the probability of repeat studies, while unequivocal failure allows pharmaceutical companies to more quickly reallocate resources to a new and potentially more fruitful project. The rate at which unequivocal data is provided through clinical trials can be increased through independent central review, also known as adjudication, because this practice increases accuracy and decreases bias and variability. The utility of adjudication is well established for assessing efficacy endpoints as well as cardiovascular safety endpoints, such as Major Adverse Cardiac Events (MACE) as defined by the Hicks criteria.[1]

Examples include cardiovascular death, non-fatal myocardial infarction (MI), and non-fatal stroke. Its use in assessing safety endpoints in other therapeutic areas is largely unexplored. It is likely that practicing central adjudication for oncologic safety endpoints can produce similar benefits, facilitating more consistent diagnoses. Adjudication is tightly controlled through training, testing, and data presentation which leads to an increase in reliability and consistency. Central review of tissue samples related to potential malignancy events can corroborate or refute a local diagnosis from a site. This would yield a higher rate of unequivocal results, increasing efficiency for studies regarding oncologic safety in drug development.

Any practice that can improve efficiency in clinical studies is significant, as it could offer tremendous economy for a pharmaceutical industry that faces a failure rate of more than 90%. Based on estimates between 2006 and 2015, the likelihood that a drug in a phase I study will eventually be approved is 9.6% (failure rate of 90.4%). The likelihood that a drug will advance from Phase II to Phase III is only 30.7% (failure rate of 69.3%).[2] In May 2016, the Tufts Center for the Study of Drug Development estimated the cost of bringing a new drug to market (from Phase I through to approval) is estimated at $1.3 billion. Considering research and development (R&D) expenses and the current failure rates mentioned above, the estimate rises to about $4 billion.[3,4]

If central adjudication can introduce significant efficiency in clinical trials with oncologic safety endpoints, this would in turn entail massive savings during pharmaceutical development.

Central Adjudication

Adjudication is a standardized, objective, and often blinded mechanism designed to assess clinical events with increased accuracy. It is performed by a centralized committee of independent reviewers (IRs), who are specialized, expert physicians who have no involvement with either the treatment of study subjects or the trial sponsor.[5] These committees are referred to as an adjudication committee or clinical events committee (CEC). Utilizing complex clinical definitions can lead to heterogenous and often subjective outcomes, and central adjudication is an effective way of curbing this tendency through standardization and objectivity.[6(p695)]

Central adjudication is already widely used, but is by no means ubiquitous. In a meta-analysis of 314 articles from five influential general medical journals covering randomized clinical trials, use of adjudication was mentioned in 105 articles (33.4%).[6] Krumholz-Bahner et al. assessed 35 newly identified molecular entities approved in the United States and 88 drug approvals in the European Union between January 2013 and April 2014. Central adjudication of the primary endpoints was used in 69% of approvals in the United States, and in 41% for the European Union, with most endpoints being efficacy related. Twenty-nine studies used central adjudication to assess efficacy, but only eight studies adjudicated exclusively safety endpoints. Fifteen studies used adjudication to investigate a combination of both.[7]

In the absence of a "truth standard" - a standard believed to give the true state of a patient or true value of a measurement - adjudication can help facilitate increased accuracy in results.[8] The United States Food and Drug Administration (FDA) encourages the use of truth standards in clinical trials to "demonstrate that the results obtained are valid and reliable."[8(p22)] Areas of medicine that have an undefined or nebulous truth standard could benefit from the practice of adjudication as an alternative. Although adjudication does not replace a truth standard, it can streamline data by removing variable interpretations and utilization of definitions which can cloud the dataset.[9] A more homogenous dataset is easier to analyze and makes it more likely that a clear and concise conclusion can be reached.

Limited CEC membership combined with uniform training, data presentation, and application of definitions keep variability low and precision and accuracy high.[9(p265),10]

The entire adjudication process, from identification of a case through independent review, is documented in a charter. The charter includes the following: the process for collecting and processing source materials from the sites, definitions and explanations of terms relevant to making an assessment, minimum requirements for making assessments/how to handle cases that do not meet the minimum requirements, electronic case report form (eCRF) description, committee membership, training, testing requirements, read paradigm, and a bias minimization plan.[11,12]

Adjudication committees are often confused with data safety monitoring boards (DSMBs), but each has a separate role in a clinical trial. The purpose of a DSMB is to ensure ongoing safety of trial participants and continuous validity and scientific integrity of the trial.[10(p112),11(p54)] The DSMB generates periodic risk/benefit assessments and safety reports; the adjudication committee generates independent harmonized assessments of study outcomes. The DSMB is usually unblinded to treatment assignment of participants; the adjudication committee is almost always blinded to this information.[10(p112)]

Bias

"Bias is systematic error that leads to distortion of true treatment effects."[13] This is different from random error or imprecision, which leads to a study outcome that is different from the "truth" due to statistical uncertainties related to obtaining a random sample. Random error is inevitable, but can be addressed retrospectively during statistical analysis. Risk of bias is also inherent in randomized clinical trials and can be mitigated prospectively through a study design that includes adjudication.

Bias in clinical trials includes selection bias, detection bias, performance bias, and attrition bias.[14] Selection bias occurs when individuals are recruited, screened, and/or enrolled into a trial in such a way that systematic differences between study arms are created. Selection bias can also result from enrolling subjects that are not generalizable to the intended study population. Detection bias is a systematic difference in outcome determination and results. Performance bias refers to systematic differences in the type of care and attention provided to study participants, which may give both caregivers and subjects inclinations as to which treatment arm they are assigned. Attrition bias is a systematic difference in study withdrawals from each group. Study withdrawal leads to incomplete or missing information which can alter interpretation of results.[14]

Adjudication can address performance bias and detection bias through maintaining independence from sites and sponsors. Maintaining independence permits the adjudication committee to be more objective in making assessments. Selection bias and attrition bias are controlled for through study design - whether they can be mitigated through adjudication is not addressed in this paper.

The dual role of study investigator and clinical physician leads to potential for detection and performance bias. The goals of clinical practice and clinical investigation are different, which complicates objectivity.[10(p115)] Physicians acting as both caregiver and investigator could have personal and emotional histories with certain patients, which can create difficulty in remaining objective. Adjudication committees are completely separated from patient care, which eliminates this problem.

Adjudication decreases performance bias and resulting detection bias through creating distance between clinical subjects and those responsible for making assessments. Personal preferences can skew investigator assessments of certain outcomes.[5(p56),13(p596)] As stated by KR Cox, "the desire for a successful outcome is felt so strongly in both patients and investigator that objectivity cannot be guaranteed. Both have an emotional stake, overt or occult, in the result. Further, the giving of any treatment…is a strong psychotherapeutic stimulus in itself."[15]

If an unblinded investigator favors one treatment over another, performance bias may lead to detection bias. The investigator may follow those on the preferred treatment more closely, consequently identifying outcomes more often than those in the non-preferred group.[13(p596)] Even blinded investigators with knowledge of all treatment options may believe that a subject is on a particular treatment arm. Subjects perceived to be on the preferred treatment arm could be monitored differently. Even when recording lists of numbers or simple data, mistakes in transcription are often in the direction most personally favored by the investigators.[16] Because adjudication obscures all identifying information between a subject and investigator, it all but eliminates the potential for an emotional connection and personal preference that could lead to performance and detection biases.

Results based on bias undermine the integrity of a study, but adjudication can mitigate the potentially negative effects of performance and detection bias. Adjudication increases the veracity of study results through maintaining independence from study sites, investigators, and subjects. It is important to remove or minimize bias in studies to any

extent possible to maximize the safety and efficacy of treatments predicated upon those studies, and adjudication is a valuable tool in accomplishing this.

Variability

The clinical trial model decreases variability and biasing through randomization and blinding, among other methods, such as following a protocol. Variability measures the spread in a dataset – spread is simply a reference of the level of difference between certain data points or characteristics. Adjudication can further decrease variability through standardized training and monitoring as well as judicious committee membership. The larger and more complicated the clinical trial, the greater the potential for variability. For example, the principal investigator is responsible for creating and managing a study team most capable of accurately implementing study procedures, but variability regarding team size, years of experience, specialization, and division of labor is inevitable.[11(p57)]

Studies with complex outcome definitions or study procedures are at an elevated risk for variability. Examples include medical events based mostly on subjective interpretation due to lack of standardized definitions and/or subject reported information, such as cardiovascular composite outcomes.[17] Events where there is systematic misclassification based on accepted clinical definitions also lead to high variability.[5(p57)] Criteria for medical events may differ between clinical practice and study protocol, and in these instances study personnel could easily confuse definitions while alternating between clinical and investigative duties. Study teams working outside of their areas of expertise could mistakenly identify, fail to identify, or miscategorize events as well. For

example, a rheumatologist acting as principal investigator for a rheumatoid arthritis trial might incorrectly categorize, or even miss, a cardiac event.

An adjudication committee eliminates these issues. Members are selected based on the type of clinical events requiring review and are standardized in review criteria through training and consistent monitoring. Adjudication committee members may be practicing physicians, but they do not have to alternate between seeing study versus non-study patients because they do not see patients enrolled in the study for which they adjudicate.

Variability in study execution can lead to distorted study outcomes, which could under or overestimate true treatment effect. In a simulated tumor growth and tumor growth measurement model, researchers found that variability in assessing tumor size led to attenuation of treatment effect (hazard ratio closer to one) and increased type II error.[18] Several studies have shown that adjudication can change event classification in 20-30% of cases, thereby creating a more well-supported dataset.[6(p699)]

One such study was the second Platelet IIb/IIIa Antagonist for the Reduction of Acute Coronary Syndrome Events in a Global Organization Network Trial (PARAGON-B). Analyses showed that site and central diagnoses of MI disagreed 23% of the time. For 95% of discrepant cases, letters were sent to sites that providing rationale for the central decision. Site investigators then returned the letters either confirming or refuting the central diagnosis; in 80% of cases (307 cases) site investigators came to agree with the central assessments. The remaining 20% (75 cases) were reviewed by a faculty committee of cardiologists. The faculty committee agreed with the site investigator in

only 10 of the 75 cases. The recorded outcome for these 10 cases were determined by the site investigator; in all other cases (372 of 382) the recorded outcome was the outcome determined by the adjudication committee.[19]

In the TRIM trial, the adjudication committee changed the final assessment in 24.3% of site reported events; the total number of endpoints decreased by 11.9%.[17(p774)] The TRITON, RECORD, and PLATO studies of acute coronary syndromes also showed a decrease in total site reported cardiovascular endpoints after adjudication by a central committee. Adjudicated data in the IMPACT II, GUSTO IIb, and PURSUIT acute coronary syndrome trials showed opposite results. In the IMPACT II and GUSTO IIb trials, statistically significant differences between the treatment arms were observed when investigator data was used but not when adjudicated data was used.[5(p58)]

Data safety monitoring boards depend on accurate assessments of endpoints to determine when/if studies should be halted. Regulatory agencies also rely on accurate assessments when determining when/if a drug application should be approved. Adjudication can provide extra assurance that the data used to make such important decisions is reliable. That said, authorities should independently audit studies where the difference between central and local assessments could have a major impact on study outcomes.[17(p776),20]

Safety Monitoring

Adjudication increases the likelihood of correct identification, assessment, and categorization of clinical events essential for accurate data interpretation. The FDA requires sponsors to monitor pre- and post-marketing clinical trials for events indicating

possible safety concerns.[21] Sponsors collect information on adverse events (any untoward medical occurrence associated with the use of a drug in humans, regardless of whether it is considered drug-related) to reveal potential adverse reactions (any adverse event for which there is a reasonable possibility that the drug caused the adverse event).[22] Adverse events and reactions meeting certain criteria are required to be reported to the FDA and may also be required on product labeling. For example, "potential serious risks" and "serious and unexpected suspected adverse reaction[s]" for drugs being tested under an investigational new drug application (IND) must be reported to the FDA within fifteen calendar days of becoming aware of the event.[22]

The Food and Drug Administration Amendments Act of 2007 (FDAAA) authorized the FDA to require (when deemed necessary) post-marketing studies to assess known serious risks, safety signals of serious risks, and identify any unexpected serious risks. The Food and Drug Modernization Act of 1997 (FDAMA) also requires that annual status reports to be submitted to the FDA for drugs approved under a New Drug Application (NDA) or Accelerated New Drug Application (ANDA) for the first time.[23,24] The annual reports must include a summary of new or updated information that might affect safety, efficacy, or labeling of the investigational product. This safety information from various sources including new toxicology data from nonclinical laboratory studies, published clinical trial data, "reports of clinical experience pertinent to safety," and status reports of post-marketing study commitments.[23] Sponsors must continue to submit annual reports until they are notified in writing that post-marketing requirements have been fulfilled or that the sponsor is released from further commitments.

The FDA recommends a systematic approach to clinical trial safety monitoring that depends on accurate recognition and designation of clinical events. Adjudication increases the likelihood of both. Safety determinations can involve multiple endpoints even when some preexisting safety concerns are known. Sponsors must periodically review data in aggregate, collecting data from completed and ongoing studies, to reveal potential causal relationships between events and the intervention under investigation. Examples of areas of concern include: events occurring more frequently in the intervention group versus a control group, and/or a clinically relevant increase in the prevalence of a serious adverse reaction over what is expected based on previous data.[22,25]

Adjudication can supplement safety monitoring plans for studies conducted under NDAs or ANDAs as well as for post-marketing studies. Adjudication data can be presented to DSMBs in addition to or in place of site-generated data. When provided together, DSMBs can compare the two datasets to illuminate study data inconsistencies. Consistent differences in categorization of events between the adjudication committee and a particular site or differences in classification of one type of event between the adjudication committee and all sites could be causes for concern.

Oncologic Safety

Cardiac safety came into the spotlight in the early 2000s with the downfall of Merck's COX-2 inhibitor Vioxx. Vioxx was approved by the FDA for treatment of acute or chronic pain in 1999. It was removed from the market in 2004 due to increased risk of heart attack after 18 months. It is estimated that eighty-eight thousand Americans had

heart attacks and 38,000 individuals died of cardiovascular related causes after taking Vioxx.[26,27] It should be noted that these deaths were not confirmed to be from Vioxx.

Following the events with Vioxx, the FDA started to demand more stringent cardiac safety monitoring by sponsors. In October 2005, the FDA required Bristol-Myers Squibb and Merck to submit more detailed safety data before approving their new diabetes treatment Pargluva. The previous month, an advisory committee voted eight to one supporting approval of the drug, but the FDA was concerned about data suggesting that drug could double cardiac risk. The FDA also changed its internal practices surrounding safety monitoring. In the summer of 2005, the FDA created the Drug Safety Oversight Board which "advises the [Center for Drug Evaluation and Research] Director on the handling and communicating of important and often emerging drug safety issues."[28] An independent review of FDA post-marketing monitoring procedures was also conducted in late 2005.

Oncologic safety monitoring is less established than cardiac safety monitoring. Drugs may have genotoxic and/or non-genotoxic carcinogenic effects that are not immediately apparent, and many cancers have long minimum latency periods: approximately 2.5 years for thyroid cancer, 4 years for solid cancer, and 11 years for mesothelioma.[29,30] The typical phase 3 clinical trial runs between one and four years, which is an insufficient amount of time to detect a cancer that may have developed during the study.[31]

Long term follow-up and post-marketing safety studies are essential to capturing malignancy events. Adjudication for a multiprotocol drug program can assist in collecting

and assessing this oncology data in a standardized, independent, and consistent manner over time. Long term longitudinal follow up facilitates pattern recognition and can determine both temporal and causal relationships. Data collected over time increases validity and succinct results.[32,33]

Per the United States Centers for Disease Control and Prevention (CDC), chronic diseases are responsible for 70% of deaths each year and treating them accounts for 86% of the United States' health care expenditures.[34] In 2012, the top five therapeutic classes of prescribed drugs were metabolic agents, central nervous system agents, cardiovascular agents, psychotherapeutic agents, and respiratory agents .[35] Diabetes (9.2% of Americans in 2014), chronic pain (30.7% of adult Americans in 2010), symptoms of coronary heart disease (6.0% of adult Americans in 2010), depression (6.7% of adult Americans in 2016), and chronic obstructive pulmonary disease (COPD) (6.3% of adult Americans in 2012) are chronic conditions treated by these drugs.[35,36,37,38,39,40,41] Cancer is a comorbid condition with all the chronic ailments listed above except for chronic pain.[42] Chronic diseases often require long term medication prescription to manage symptoms. The relationship between duration of the condition, time spent on medication treating that condition, and the possible development of cancer requires long term observation and monitoring.

In the early 1990's the Women's Health Initiative (WHI) conducted a long-term hormone therapy study of post-menopausal women and monitored incidence of breast cancer as a primary endpoint. At the time, many post-menopausal women took hormone replacement therapy for extended periods to treat symptoms of menopause. Post-

menopausal women with a uterus received conjugated equine estrogens (CEE) plus medroxyprogesterone acetate (MPA) therapy. Post-menopausal women that had undergone a hysterectomy received CEE only (placebo). After a mean follow up of 5.2 years, the trial ended with 199 cases of breast cancer in the CEE+MPA group and 150 in the CEE only group.[43] Women in the CEE+MPA group had an approximately 4% greater risk of breast abnormalities detected on a mammogram after one year on therapy and an approximately 11% greater risk after five years than women on CEE only. The CEE+MPA group also had significantly more abnormal breast abnormalities detected after one-year post intervention cessation than the CEE only group. After one year however, the differences between the two groups became statistically insignificant. [44,45] The long-term observations in this trial exemplify the complex and unclear nature of the relationship between cancer, intervention, and existing conditions.

Histopathology

Cancer can be detected by blood tests and imaging, but in most cases a biopsy is the only method to obtain a definitive diagnosis. Tissues are examined for underlying pathology or histopathology, which is "the anatomic and physiological deviations from the normal that constitute disease or characterize a particular disease" and the study of these deviations.[46] From a safety perspective, histopathology is especially critical when determining whether a person has a benign or malignant process occurring. The more specific characteristics of the tumor (topography, histology, behavior, and grade) are also essential for determining next steps in terms of treatment.

In the paper "Interobserver agreement in grading of colorectal cancers – findings from a nationwide web-based survey of histopathologists," Ian Chandler et al. commented on the subjective yet important nature of tumor grading: "Tumour grade represents a gestalt of all molecular changes, reflecting aggressiveness, and thereby potentially offering considerable potential to delineate subgroups with differing prognoses."[47] Paul Speight et al. proposes that "grading must impose artificial categories onto what is a diffuse, nonhomogeneous continuum of biological change, with no clear boundaries."[48] The lack of a truth standard in histopathology may lead to complicated, subjective, and variable outcomes. Adjudication can streamline histopathology data through standardization and consistency.

There are several studies which demonstrate high variance of histopathologic interpretations. In the National Inter-Observer Agreement in Colorectal Cancer (NIACC) study (Chandler, et al.), twenty digitized colorectal cancer cases obtained from the Institute of Cancer Research Section of Cancer Genetics repository were uploaded to a dedicated webpage. Each case had one representative hematoxylin and eosin (H&E) stained digitized slide. All United Kingdom (UK) consultant histopathologists in the Royal College of Pathologists database were contacted via email to participate in the survey. A request was made that only pathologists who report gastrointestinal (GI) specimens participate. Participants were instructed to grade the specimens using both a two-grade and three-grade system. The three-grade categorized tumors as "well differentiated, moderately differentiated, and poorly differentiated." The two-grade system combined well and moderately differentiated tumors into low grade leaving

poorly differentiated cases as high grade. One hundred senior pathologists from 59 teaching and district general hospital trusts in the UK (32% of all UK trusts) assessed all twenty slides. After calculating interobserver (also known as interrater) variability using Fleiss's kappa, interobserver agreement was determined to be only fair based upon the Fleiss's kappa categorization scheme.[47(p496)]

Fleiss's kappa is an adaptation of Cohen's kappa coefficient (Cohen's kappa), a statistic measuring interrater variability accounting for the possibility that raters guess on at least some variables due to uncertainty. Cohen's kappa can be used to compare interrater variability between two raters. Fleiss's kappa can be used for three or more raters. Results for either can be interpreted as follows: 0.01–0.20 as no agreement to slight; 0.21–0.40 as fair; 0.41– 0.60 as moderate; 0.61–0.80 as substantial; and 0.81–1.00 as nearly perfect agreement. Confidence intervals should be calculated for the kappa statistic because it is an estimate, not a direct measure.[49]

The overall kappa value for both the three-grade specimen coding system and the two-grade specimen coding system indicated fair interobserver agreement: 0.351 and 0.358, respectively. No confidence intervals were provided but the p-value for both groups was $P<0.0001$, indicating that the results were statistically significant.[47(p496)] Per Chandler, et al.: "This national survey…was prompted by anecdotal experience that there is a great deal of interpersonal variation in how this seemingly straightforward task is performed… [this study] implies that the main difficulty that pathologists face is dividing moderately from poorly differentiated tumours. This is an important distinction from the point of view of patient management, as this is the most clinically relevant division to

make. In addition, in multicenter trials in which patients' tumors are pathologically reported in multiple hospitals, outcome data will not be comparable if the grades allocated to tumours suffer high interobserver variation."[47(p497)]

A substantial or nearly perfect kappa can still be problematic. 765 women at the University of Halle had core biopsies of the breast performed between 2006 and 2008. Only the first biopsy of each woman was considered in this study. Three pathologists specializing in breast cancer reviewed the H&E stained slides and case report form (CRF) for each biopsy. The CRF included information regarding age, localization of biopsy, number of biopsy cores, microcalcification, and a description of the focus. If x-ray images were available those were also provided. Pathologists were instructed to categorize the biopsies per the five-level B-categorization scheme suggested by the European guidelines for quality assurance in mammography screening (B1: normal or uninterpretable; B2: benign; B3: benign but of uncertain biological potential; B4: suspicious of malignancy; and B5: malignant including *in situ* and invasive cancer). Categories B1-B2 usually do not require additional testing unless the biopsy is uninterpretable or determined not to be representative of the lesion, while categories B3-B5 usually require invasive work-up.[50]

Pathologist 1 was the local pathologist from the University of Halle. If needed, Immunohistochemical (IHC) stained slides could be requested to supplement the H&E stained slides already provided. A reference pathologist at the University of Münster reviewed the same materials as Pathologist 1. The reference pathologist could also request additional IHC stained slides. Discrepancies in assessments were resolved over

electronically via digital slide exchange. A third pathologist from Hamburg (pathologist 2) reviewed the same information as pathologist 1, including IHC stained slides.50 (p941)

Interobserver agreement between pathologist 1 and 2 was calculated using Cohen's kappa statistic. The observed kappa based on the five-level categorization was 0.87 (95% confidence interval (CI): 0.84-0.89). The observed kappa based on the two-level categorization (B1-B2 vs B3-B5) was 0.93 (95% CI: 0.91-0.95). Both indicate almost perfect agreement. There were 103 total histopathological diagnosis discrepancies, representing 13.5% of all samples. Forty-nine and a half percent of discrepancies (51 of 103 cases), however, were clinically relevant disagreements (B1-B2 vs B3-B5). In total, 7% of women would have been at risk for negative effects from misdiagnosis. Observed kappa scores of specific histology results were much lower than those for the categorization schemes.[50(p943)] Adjudication of histopathology samples could potentially decrease interobserver variability through decreasing the number of pathologists making assessments, sub-specialization of committee members, and standardized training/monitoring.

Case Study: "Interobserver agreement in dysplasia grading: toward an enhanced gold standard for clinical pathology trials" (Speight, et al.)

A 2015 study conducted by Speight et al. sought to establish a gold standard for clinical pathology trials using adjudication. This trial focused on dysplasia grading of cancers of the lip, oral cavity, and oropharynx, although its results can be extrapolated to other types of cancers.

The adjudication committee was comprised of four senior oral and maxillofacial pathologists (reviewers). Oral scalpel biopsies from 846 patients comprised the dataset: 774 subjects with potentially oral malignant disorders and 72 patients with oral squamous cell carcinomas. The goal of the adjudication committee was to accurately categorize each case into one of seven microscopic diagnostic categories based on the 2005 World Health Organization (WHO) guidelines for cancer and precancer of the oral mucosa. The exact terminology and their microscopic definitions were agreed upon in advance by the adjudication committee. No additional training or calibration was performed. The read paradigm is below.[48(p477)]

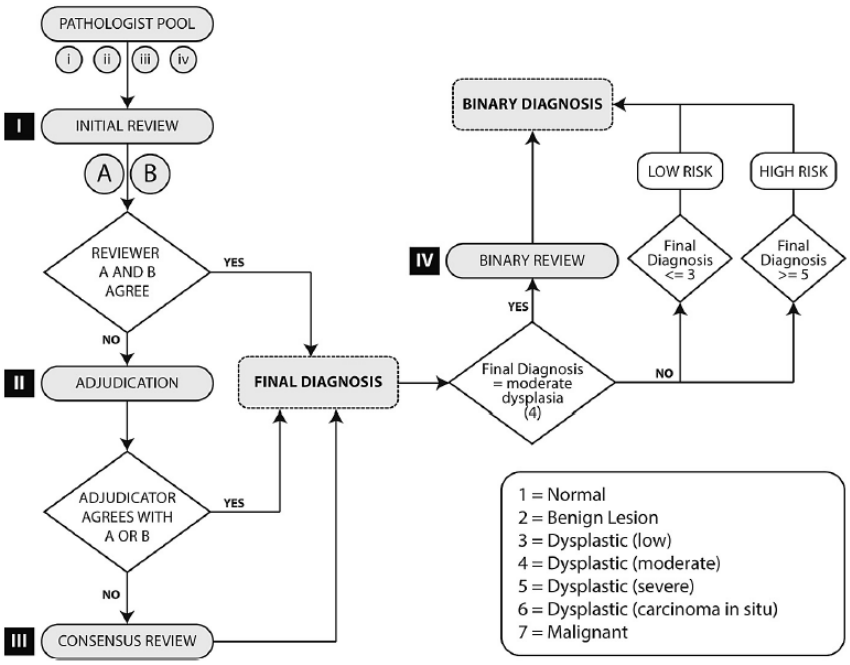

Figure 1:Flow chart illustrating the process for the enhanced gold standard adjudication sequence.[48(p478)]

Two reviewers from different clinical centers assessed each case (reviewer A and reviewer B). The reviewers were also independent from the sites were the biopsies were

originally collected. Each reviewer received a unique set of H&E stained slides (adjacent serial sections approximately 5 μm apart) and the diagnostic slide used for patient management. Reviewers were blinded to all clinical and microscopic findings, impressions, and diagnoses from the sites. They were also blinded to the topographical location of the lesion.[48(p477)]

If reviewer A and reviewer B agreed on a diagnosis, the shared diagnosis was considered final. If reviewer A and reviewer B disagreed, a third reviewer (the adjudicator) reviewed the slides from both reviewer A and B. Only one reviewer acted as adjudicator; selection of the adjudicator was based on expertise. The adjudicator was blinded to the same information as reviewers A and B as well as to their assessments. If the adjudicator's diagnosis was the same as either reviewer A or B, this was considered the final diagnosis. When the adjudicator did not agree with either reviewer A or B, a consensus meeting between all three reviewers (A, B, and the adjudicator) was held. The group reviewed the slides together and was blinded to all previous assessments and site data. All cases classified as moderate dysplasia underwent consensus review to further categorize the samples into high or low risk cases. Kappa scores were calculated for each pair of reviewers and ranged from 0.251 to 0.706. See table from Speight, et al. below.

Table 1: Agreement reviewing pathologists during initial review stage; initial review stage percent agreement and kappa values shown for individual pathologist pairs.[48(p479)]

	Initial Review			
Reviewing pathologists	*N*	*κ (Interpretation)*	*κ 95% CI*	*% Agreement*
i | ii	147	0.706 (Good)	(0.618, 0.793)	81.0%
i | iii	245	0.513 (Moderate)	(0.427, 0.600)	79.6%
ii | iii	115	0.251 (Fair)	(0.126, 0.377)	65.2%
iii | iv	105	0.463 (Moderate)	(0.336, 0.589)	68.6%
ii | iv	234	0.423 (Moderate)	(0.339, 0.508)	62.0%

Reviewers A and B agreed on a diagnosis in 69.9% of cases. An additional 22.8% of cases reached a final diagnosis after adjudication. Only 7.3% of cases required consensus review, after which 100% of cases had a final diagnosis.[48(p479)]

The adjudication model chosen by Speight, et al. was based a social theory on collective decision making postulated by James Surowiecki in his book *The Wisdom of Crowds*. Surowiecki states that a successful crowd wisdom requires diversity of opinion, independence of opinion, decentralization (ability to specialize), and aggregation (mechanism to consider individual judgements to make a collective decision).[51] Speight, et al. suggested that by increasing interobserver agreement, the collective decision of multiple pathologists might lead to more clinically accurate microscopic diagnoses.[48(p481)]

Measuring "correctness" in diagnosing histopathology samples is impossible because the "true" diagnosis is unknown (lack of a truth standard). Using probability theory and overall level of agreement and disagreement between reviewers A and B, Speight et al. also provided a mathematical basis for the chosen adjudication model. The study team calculated probabilities of correct and incorrect diagnoses for six scenarios.

The calculations depended on the following assumptions: "(1) all reviewers had an equal probability of misdiagnosis that was not influenced by the other reviewers; (2) each slide had an equal probability of misdiagnosis; and (3) where 2 reviewers disagree on a particular diagnosis, one was assumed correct and the other was assumed incorrect".[48(p479)] Speight, et. al. recognized the possibility that both reviewers could be incorrect, but the assumption that one was correct and the other was incorrect was necessary for this derivation. The total probability of correct diagnosis for this study was 91%. See the table below for all results. The details of the derivation provided in the supplementary materials of Speight, et. al. can be found in the appendix. The results of Speight's probability calculations are in Table 2.

Table 2: Probability of correct 7-level diagnosis (normal, benign, dysplastic mild, dysplastic moderate, dysplastic severe, dysplastic carcinoma in situ, malignant) with 2 reviewers and use of an adjudicator when the 2 reviewers disagree.[48(p479)]

				Probability		
Probability Scenario	*Reviewer A*	*Reviewer B*	*Adjudicator*	*Equation*	*Probability*	*Final Diagnosis*
1	Correct	Correct	N/A	Pc × Pc	0.664	Correct
2	Wrong	Wrong	N/A	Pw × Pw	0.034	Wrong
3	Correct	Wrong	Correct	Pc × Pw × Pc	0.123	Correct
4	Correct	Wrong	Wrong	Pc × Pw × Pw	0.028	Wrong
5	Wrong	Correct	Correct	Pw × Pc × Pc	0.123	Correct
6	Wrong	Correct	Wrong	Pw × Pc × Pw	0.028	Wrong
Total probability of correct diagnosis					0.91 (91.0%)	
Total probability of wrong diagnosis					0.090 (9.0%)	
Overall total probability					1.00 (100%)	

Note: The values of Pc and Pw represent the probability of a correct and wrong diagnosis, respectively. Even though the model assumptions are a simplification of reality, these probabilities so derived do suggest the substantial gains in correctly diagnosing the lesions that are likely to be achieved through our adjudication process.

Based on the results in the table above, Speight calculated that the probability of correct diagnosis in his study was 91%. The study conducted by Speight, et al. provides a framework onto which further studies of the effectiveness of adjudication in histopathology can be based. The read paradigm is easy to implement and is well-

supported by group psychology. The claim that adjudication is reliable and accurate is substantiated with a convincing probability theory.

Thesis

The aim of this paper is twofold: (1) to demonstrate an overall need for adjudication of histopathology related outcomes in clinical trials through quantifying interrater variability using kappa scores and percent agreement, and (2) to demonstrate that potential malignancies other than oral dysplasia could similarly benefit from adjudication – to include, among others, dermatology, gynecology, genitourinary, and gastroenterology. In reviewing whether adjudication can effectively benefit a variety of pathology sub-specialties, it would need to be established that there is significant interrater variability. To examine whether such variability is present, local diagnoses will be compared with independent diagnoses to find any contrasts and disparities between the results.

The following discrepancies in diagnoses will be assessed: neoplasm vs non-neoplasm; discrepancies in malignant vs benign processes, discrepancies in morphology, and discrepancies in classifications that would lead to different courses of treatment for the subject (cases not similar). These analyses will be performed on the dataset as a whole, within each pathology sub-specialty, and between non-melanoma skin cancers (NMSC) and other malignancies (a clinically significant difference based on patient risk).

If it is the case that there are clinically significant discrepancies between local and central diagnoses and that central adjudication yields more accurate diagnoses than a local pathologist, then it should also be accepted that adjudication ought to be more

widely used in clinical trials to assess histopathology-related safety outcomes and endpoints.

METHODS

This is a retrospective cross-sectional study assessing interrater variability between local and central diagnoses of biopsy samples in a clinical trial setting. Certified Professional Coders (CPC) and central pathologists used the International Classification of Diseases for Oncology revision 3 (ICD-O 3) to codify the local and central assessments to permit comparison. Samples included in this analysis were collected as part of an industry sponsored global adjudication program managed by a contract research organization (CRO). The CRO was responsible for identification of potential malignancy events, collection and processing of specimens, generation of study documents, and independent reviewer selection, training, management, and monitoring. The sponsor approved the potential event identification criteria, the charter, and the electronic case report forms (eCRFs). Specimens included in this dataset were assessed between January 7, 2015 and March 15, 2017.

International Classification of Diseases for Oncology

This study used the International Classification of Diseases for Oncology Version 3, or ICD-O 3, to describe both local and central diagnoses. ICD-O is a standardized coding system developed by the World Health Organization (WHO) to classify site (topography), histology, behavior, and grade of abnormal tissue growths (neoplasms). Developed in 1976, it is used primarily in tumor and cancer registries. ICD-O is based on

the American Cancer Society's Manual of Tumor Nomenclature and Coding (MOTNAC), which was first published in in 1951.[52] The WHO stresses that "ICD-O is a coded nomenclature and not a classification scheme for neoplasms; the listing of terms from different classifications does not represent endorsement of any particular one."[53]

Topography codes are four-character codes; the first character is always "C" and the next three characters are always numbers. The first two numbers indicate site and the number after the decimal indicates a more specific sub-site.

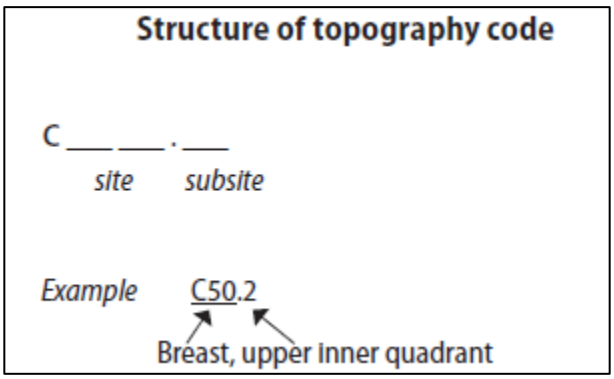

Figure 2: ICD-O Topography Code.[53]

Topography codes run from C00.0 to C80.9. They are closely related to those in the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10).[54] The ICD-O topography list allows for greater site specification for non-malignant tumors and provides topography codes for haematopoietic and reticuloendothelial tumors, which ICD-10 does not cover.[54]

The histology and behavior codes represent the morphology (microscopic features) of a neoplasm. The codes start with the letter "M" and are followed by five

numbers which range from M-8000/0 to M-9992/3. The first four digits are the histology code and the number after the slash is the behavior code. The table below explains possible behavior code values and their definitions (Table 3).

Table 3: ICD-O Behavior Codes.[53]

Code	Definition
/0	Benign
/1	Uncertain, whether benign or malignant Borderline malignancy Low malignant potential Uncertain malignant potential
/2	Carcinoma in situ Intraepithelial Noninfiltrating Noninvasive
/3	Malignant, primary site
/6	Malignant, metastatic site Malignant, secondary site
/9	Malignant, NOS - uncertain whether primary or metastatic site

The second digit after the slash (M0000/0**0**) describes the grade or differentiation of a neoplasm (Table 4). It also represents the immunophenotype designation for lymphomas and leukemias (Table 5).

Table 4: ICD-O Histologic Grading and Differentiation Codes.[53]

Code	Grade	Definition
1	I	Well differentiated Differentiated, NOS
2	II	Moderately differentiated Moderately well differentiated Intermediate differentiation
3	III	Poorly differentiated
4	IV	Undifferentiated Anaplastic
9	Grade or differentiation not determined, not stated or not applicable	

Table 5: ICD-O Immunophenotype Designation for Lymphomas and Leukemias.[53]

Code	Definition
5	T-cell
6	B-cell Pre-B B-precursor
7	Null cell Non T-non B
8	Natural killer (NK) cell
9	Cell type not determined, not stated or not applicable

A complete ICD-O code for one neoplasm will have eleven characters representing topography (four), histology (five), behavior (one), and grade, differentiation, or immunophenotyped (lymphomas and leukemias only) (one).

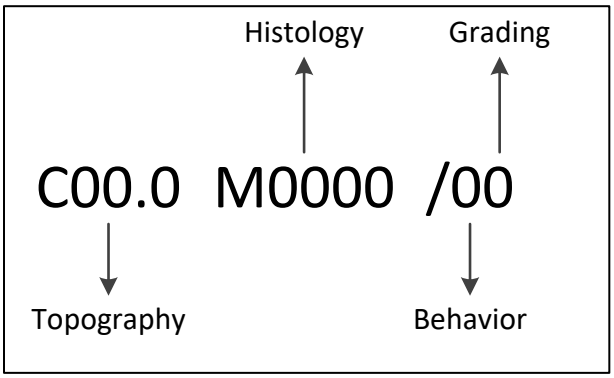

Figure 3: ICD-O Complete Code.[53]

Adjudication Committee and Certified Professional Coders (CPC)

The CRO leveraged an existing relationship with a Boston, Massachusetts medical facility to create the adjudication committee, which independently assessed tissue samples for this study. A senior member of the pathology department, the "lead pathologist," worked with the CRO to create the adjudication committee in November 2014. The original adjudication committee consisted of seven members including the lead pathologist; the committee as of March 2017 had expanded to thirteen members including the lead pathologist. The March 2017 committee had five of the original seven members and eight additional members that were added between February 2016 and November 2016 to keep up with an increasing workload. The lead pathologist left the committee in June 2016; he selected his own replacement as one of the other original six original members. One of the other original members was removed in September 2016 due to declining performance identified during reviewer performance monitoring. In cases where this pathologist had made the final diagnosis, or "authoritative review," a data

integrity assessment was performed. No outliers or data integrity risk was identified. Fourteen pathologists were the authoritative reviewer on a case at least once.

All committee members were United States board certified pathologists licensed to work in the United States. Committee members had a wide variety of educational and professional backgrounds both within and outside of the United States, which contributed to the independent of opinion of each reader. All except one were full time faculty members at the Boston medical facility. The outlier was a pathology fellow selected by the lead pathologist as committee member based on performance. The pathology fellow did not perform any authoritative reviews.

Committee members were divided into sub-specialty reading groups by the lead pathologist. Not all sub-specialty group members were specialists in the field, but per the lead pathologist had enough experience to assess cases in that group. If a case required adjudication (the diagnoses from pathologist 1 and 2 did not match), the "tie-breaker" (adjudicator or PR3) in each sub-specialty group was either a practicing specialist in the field or, if no specialist was available, the lead pathologist or designee selected by the lead pathologist. The lead pathologist also had the authority to reassign committee members to sub-specialty groups as needed.

Two certified professional coders (CPC) were part of the review team, although not committee members. The CPCs received the local pathology reports (LPR) from the sites and coded the local diagnosis for the biopsy per ICD-O 3. The CPCs did not perform any independent reviews for this study.

Pathology fellows selected by the lead pathologist provided auxiliary services when needed, such as identifying stains when they were not specified by the site. Biopsied tissue is usually transparent when put on a slide; staining with certain chemicals assists pathologists with viewing tissue structure and certain cell types.[55] When a local lab report describing multiple biopsies was received and slides were received without labels, the pathology fellows attempted to match the provided slides to the biopsies on the lab report. Pathology fellows, except for one (see above) did not perform any reviews.

All pathology reviewers, pathology fellows, and CPCs were blinded to the identity of the sponsor, investigational product, and protected health information (PHI).

Charter and Independent Review Manuals

The charter outlined the process for collecting and processing source materials from the sites, definitions and explanations of terms relevant to making an assessment, minimum requirements for making assessments/how to handle cases that do not meet the minimum requirements, eCRF description, committee membership, training, testing requirements, read paradigm, and bias minimization plan.[11(p59),12] Pathology reviewers (committee members) were trained on the charter and the independent review manual. The CPCs were trained on the charter and the CPC-specific review manual. Pathology fellows were not required to be trained on the charter or review manuals, but attended an orientation covering the project goals and services required.

The adjudication program had three charter revisions since the initial version came into effect in November 2014. Changes between each charter revision, effective date, and whether re-training of committee members was required is documented in the

table below. Reviewers that were not part of the original adjudication committee were trained on the latest version of the charter at the time that they joined (Table 6).

Table 6: Charter revisions and training requirements.

Version	Changes from previous version	Effective Date	Training required for committee members?
1.0	N/A Initial Release	10-Nov-14	Yes – Initial Release
2.0	• Updated sponsor/CRO roles and responsibilities • Updated committee membership	26-May-15	No – changes do not affect review process
3.0	• Removed draft versions from revision history • Updated committee membership	22-Dec-15	No – changes do not affect review process
4.0	• **Added review process for photomicrographs** • Clarified blinding procedures • Removed reference to sponsor approval for reviewer manuals per latest SOP update • Updated requirement for sponsor signatures on User Requirements to only when sponsor facing changes are made • Added Potential Primary Event notifications as method to receive information regarding potential malignancy events • Added digital pathology process for sites in China • **Clarified processes regarding multiple biopsies received for one malignancy event** • Updated operational workflow diagram • Added histopathology processing as a workflow step (not a process change, just needed to be documented) • Updated assessment workflow diagram • **Added assessment of EBV status for lymphoma cases** • **Removed turnaround time requirements** • Clarified PR3 role • **Added requirement for comments when similarity assessment = no** • **Added definitions for slide image/quality** • **Updated adjudication criteria** • Committee member updates • Updated data management section • Updated close out details • Updated sponsor clinician review process • Removed requirement of U.S. medical licenses and board certifications for reviewers (physicians certified outside of the U.S. permitted)	22-Jun-16	Yes – changes in bold may affect review process

This program had separate manuals for CPCs and independent reviewers. Both included information regarding how to access/navigate the electronic review system and how to complete the eCRF based on the rules in the charter. Both the CPC manual and reviewer manual underwent one revision since the original documents were created. Updates made in both documents in each revision as well as training requirements are outlined below in Table 7. Version 1.0 of both the CPC and reviewer manuals were reviewed and approved by the sponsor. Due to a standard operating procedures (SOP) update at the CRO, Version 2.0 of both documents was not required to be reviewed or approved by the sponsor. Reviewers that were not part of the original adjudication committee were trained on the latest version of the appropriate manual at the time that they joined.

Table 7: Review manuals revisions and training requirements.

Version	Changes from previous version	Effective Date	Training required for committee members?
1.0	N/A Initial Release	15-Dec-14 – CPC Manual 17-Dec-14 – Reviewer Manual	Yes – Initial Release
2.0	• Removed duplicate information already in charter • **Referenced newly added info buttons regarding neoplasm versus non-neoplasm definitions on eCRF** • **Clarified instructions regarding histology and behavior code selection** • **Added instructions for download and assessment of photomicrographs (Reviewer Manual only)** • **Added instructions for review of EBV stained slides for lymphoma cases (Reviewer Manual only)**	16-Aug-16	Yes – changes in bold may affect review process

Histopathology Assessments

Sites were instructed to submit the slides used to make the local diagnosis (diagnostic slides) and/or the block from which the diagnostic slides were cut. Tissue samples were categorized into sub-specialties based on the anatomic location of the biopsy and Medical Dictionary for Regulatory Activities (MedDRA[1]®) preferred term of the potential malignancy event. MedDRA® is a medical terminology dictionary to facilitate international sharing of regulatory information for medical products used by

[1] MedDRA® the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH); MedDRA® trademark is owned by IFPMA on behalf of ICH.

humans.[56] The dictionary is hierarchized, with the preferred term being "a distinct descriptor (single medical concept) for a symptom, sign, disease diagnosis, therapeutic indication, investigation, surgical or medical procedure, and medical social or family history characteristic."[57] The sub-specialty pathology categories were: breast, skin (dermatology), ear/nose/throat (ENT), gastrointestinal tract (GI), gynecology (GYN), genitourinary tract (GU), lung, intradural, lymphoma, sarcoma, and cytology.

Samples were delivered to a sub-specialty committee pathologist by a CRO staff member. The staff member remained with the pathologist until the read was complete to ensure independent was maintained throughout the read and then returned with the slides to the CRO office. Each sample was assessed on a separate eCRF. Pathologists had no information regarding relationships between biopsies, number of biopsies per potential malignancy event, or potential malignancy event MedDRA® terms. They were blinded to the entire LPR, including local diagnosis. Pathologists had access to the following information for each case (if available): biopsy date, biopsy type and details if "other" was selected, anatomic location and details if "other" was selected, number of blocks provided, number of slides provided, and stain types. This is similar to the information provided to the pathologists in Speight's study. Anatomic location was not available to pathologists in Speight's study. The study however was specific to oral cancers so specific location may not have been necessary.[48(p477)] Pathologists were also informed of the CRO provided sample identification number to corroborate against the slide case labels and slide labels to ensure the cases were read on the appropriate eCRFs.

The pathologists were required to assess slide quality and degree of confidence for all cases. Options for slide quality were Good (the technical quality allows appropriate diagnostic interpretation); Fair (the technical quality is not optimal but does not limit the diagnostic interpretation); Poor (the technical quality is sub-optimal and limits the diagnostic interpretation); and Unevaluable (the technical quality is sub-optimal and precludes the diagnostic interpretation). Comments regarding slide quality were optional except when degree of confidence was indicated as low. Pathologists could request additional staining if necessary to make an assessment.

For each sample, pathologists were required to determine whether it was a neoplasm (abnormal tissue growth resulting from uncontrolled cell division (and/or lack of cell death) or non-neoplasm (abnormal tissue growth, reactive, inflammatory, or hamartomatous in origin). ICD-O 3 only codes neoplasms, so samples assessed as non-neoplasm did not require any further evaluation. Neoplastic samples had topography, histology, behavior, and grade or immunophenotype coded per ICD-O 3. Epstein Barr Virus (EBV) status was assessed for samples assessed as lymphoma.

The local diagnosis for each sample was provided to one of two CPCs on a LPR. Each sample was assessed on a separate eCRF, independent of those used by the pathologists. For neoplasms, the CPCs categorized samples as neoplasm or non-neoplasm. If the case was categorized as a neoplasm, the CPC entered topography, histology, behavior, and grade or immunophenotype for the biopsy and associated diagnosis using ICD-O 3. Comments regarding were optional. The CPCs had no access to

the tissue samples and could not request additional information. CPCs did not assess slide quality or provide degree of confidence assessments.

Read Paradigm

In the first round of reviews, each local diagnosis on the provided LPR was categorized and coded (if a neoplasm) by a CPC. All other information on the LPR was used as a reference only. Concurrently, each sample was assessed by a pathologist (PR1) in the sub-specialty reading group to which it pertained based on anatomic location and MedDRA® preferred term. CPCs and pathologists had no interaction with each other at any point in the review process. If the categorization as neoplasm or non-neoplasm and the ICD-O 3 code selection (if neoplasm) by the CPC and the pathologist were identical, the shared assessment was considered the final diagnosis. No further reads occurred.

If the assessments by the CPC and PR1 did not match, the tissue sample was assessed by a second pathologist (PR2) in the same sub-specialty reading group. The PR1 and PR2 diagnoses were compared. If they were the same, no further reads occurred; the central diagnosis decided by PR2 was the authoritative review. PR2 was then provided the LPR with the local diagnosis, the CPC assigned categorization of neoplasm or non-neoplasm, and the CPC selected ICD-O 3 code (neoplasms only). PR2 compared his/her assessment and the CPC assessment to determine whether the diagnoses were similar even though exact categorization and/or codes were different. Similar was defined as similar morphology where the patient management and treatment would be comparable.

If the PR2 and PR1 assessments did not match, then the tissue sample was read by a third pathologist (PR3) in the same sub-specialty reading group. The third pathologist's

assessment was always considered the final and authoritative read. After the final assessment on the sample was made, PR3 was provided the LPR with the local diagnosis, the CPC assigned categorization of neoplasm or non-neoplasm, and the CPC selected ICD-O 3 code (neoplasms only). Like PR2, PR3 indicated whether the CPC coded diagnosis and the final central diagnosis were similar per the definition above. The read paradigm is outlined in Figure 4.

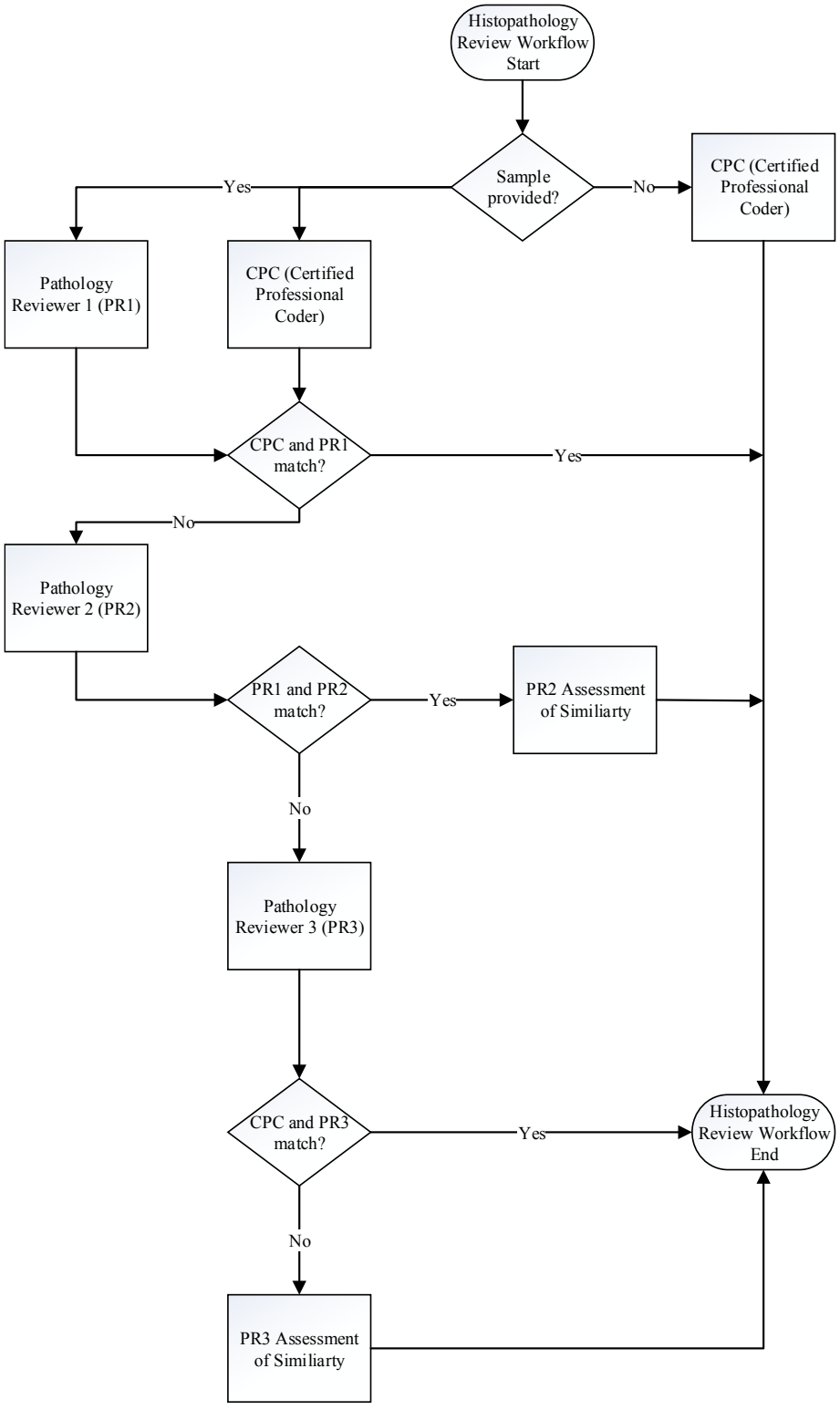

Figure 4: Read Paradigm.

Follow Up Use of Histopathology Results

The final local and central diagnoses for each histopathology sample were provided to a separate oncology adjudication committee. The histopathology information was part of a subject dossier that included clinical information such as consultation reports, discharge summaries, clinical notes, lab reports, and medical history records relating to the potential malignancy event. Histopathology samples representative of the same potential malignancy event were grouped together into one dossier. The oncology adjudication committee then made an overall assessment regarding the event.

The sponsor was interested in assessing differences between two important clinical classes: non-melanoma skin cancers (NMSC) and all other malignant processes. Although both classes represent malignant processes, the significant differences between them are important from a safety perspective. The two most represented non-melanoma skin cancers are basal cell carcinoma and squamous cell carcinoma. Approximately three million people in the United States are diagnosed with non-melanoma skin cancer each year. Only about two thousand people in the United States die from non-melanoma skin cancers each year, making death from these cancers uncommon. The highest risk factor for non-melanoma skin cancers is sun exposure; other risk factors, such as increasing age and light-colored skin, are related to sun exposure. These cancers can be prevented and managed through limiting exposure to sun, skin screenings, and removal of suspicious lesions as soon as they are identified.[58]

Cancers that are not non-melanoma skins cancers tend to be harder to treat, have higher mortality rates, and are more serious safety concerns than non-melanoma skin cancers. For example, melanoma accounts for only about 1% of all skin cancers but has an approximate mortality rate of 11%. The mortality rate for basal and squamous cell carcinomas, which make up 80% of all skin cancers, is less than 0.01%.[58,59] The age-adjusted five-year survival rates for cancers other than non-melanoma skin cancers range from 98% (testicular cancer) to 3% (pancreatic cancer). The average age adjusted five-year survival rate for cancers other than non-melanoma skin cancers is about 54%.[60]

Statistical Analysis

The raw dataset was obtained from the CRO managing the adjudication program. The original file was in Excel format. All sponsor and CRO-specific identifiers and personal information for pathologists and CPCs were blinded before analysis.

Local and central diagnoses of the cases were compared to assess interrater variability. The table below outline the categories for inter-read comparison (Table 8).

Table 8: Analysis categories for interrater variability comparison.

Analysis	Description
1	Neoplasm versus Non-Neoplasm
2	Benign versus Malignant
3	Discrepancies in Morphology
4	All discrepancies leading to differences in treatment (cases NOT similar), including those identified in categories 1, 2, and 3

Knowing whether a sample is a neoplasm or non-neoplasm is the first step to determining whether a malignancy is possible (analysis 1). Neoplasms can become malignant whereas

non-neoplasms usually do not. From a safety perspective, analysis 2 (benign versus malignant) is the most important. Not only could misclassification potentially lead to an inaccurate risk profile for the drug under investigation, but could result in serious and often life-threatening consequences for subjects. Analysis 2 will be performed under two sets of circumstances; in the first (2.1), all cases in the dataset will be taken into consideration. In the second (2.2), all non-neoplasms will be removed. Under 2.1, non-neoplasms will be considered benign. Under both 2.1 and 2.2, behavior codes of /0 and /1 will be considered benign; all other behavior codes will be considered malignant. The effects of discrepancies in morphology (analysis 3) could be negligible if the subject would be treated the same regardless of whether the local or central diagnosis was used. However, knowing more specific information about potential malignancies can elucidate areas of safety concern that require further and/or more specialized subject monitoring. Otherwise the differences could be significant. Analysis 3 will be performed under two sets of circumstances; in the first (3.1), all cases in the dataset will be taken into consideration. In the second (3.2), all non-neoplasms will be removed. Analysis 4 will capture all cases where the authoritative pathologist indicated that the local and central diagnoses were not similar (i.e. different courses of treatment would be taken for one versus the other). Analysis 4 will be performed under two sets of circumstances. In the first (4.1), all cases in the dataset will be included (where a similarity assessment did not occur (when the local ICD-O code matched with the ICD-O code selected by PR1) will be considered similar). In the second (4.2), all round 1 agreements will be removed and

only cases that had a similarity assessment (cases that were assessed by PR2 and/or PR3) will be analyzed.

The statistical sub-groups are in Table 9. Each sub-group was analyzed per all categories in Table 8. The statistical sub-groups were designed to illuminate potential differences in types of discrepancies seen in each.

Table 9: Statistical Sub-Groups.

Group	Description
A	Overall dataset (all cases)
B	Individual Pathology Sub-Specialty Reading Groups (breast, skin (dermatology), ear/nose/throat (ENT), GI, gynecology (GYN), GU, lung, intradural, lymphoma, sarcoma, and cytology)
C	Non-melanoma skin cancers (NMSC) versus all other malignant processes (non-NMSC or "other")

Interrater variability will be measured using Cohen's kappa coefficient/statistic (referred to as "kappa" or "kappa score" in later sections) and percent agreement. Cohen's kappa is based on the chi-square table in statistics and measures interrater variability between two raters while accounting for the possibility that raters guess on at least some variables due to uncertainty. Percent agreement is easier to calculate than Cohen's kappa (number of differences between local and central diagnoses divided by the sum of the cases in each analysis), but the possibility of guesswork is not taken into account.[49(p3)] Many references recommend 80% agreement as the most common minimum acceptable percent agreement when used in a healthcare setting. Percent agreement under 80% could be indicative of inconsistencies in assessments. The formula for Cohen's kappa is kappa = $(p_0 - p_e)/(1-p_e)$ where p_0 is the relative observed agreement

among raters and p_e is the hypothetical probability of chance agreement. [49(p5),61] Results range between 0.01 and 1.00 with 0.01–0.20 as no agreement to slight; 0.21–0.40 as fair; 0.41– 0.60 as moderate; 0.61–0.80 as substantial; and 0.81–1.00 as nearly perfect agreement. Confidence intervals should be calculated for the kappa statistic because it is an estimate, not a direct measure.[49(p7)] For this study a 95% confidence interval will be used. P-values will also be noted ($p<0.05$ considered statistically significant). Statistics were generated using R x64 3.3.1.

RESULTS

The sample size flow chart is in Figure 5. The original dataset had 1,063 cases assessed between January 7, 2015 and March 15, 2017. Four hundred two cases did not have any tissue samples provided, leaving 661 with biopsy samples available for central assessment. Photomicrographs were substituted for physical tissue samples in one case; this case was removed from the analysis. All cases with either poor or unevaluable slide quality were omitted from the dataset (41 cases). Cases for which additional data was requested were also removed (17 cases). The final dataset included 602 cases.

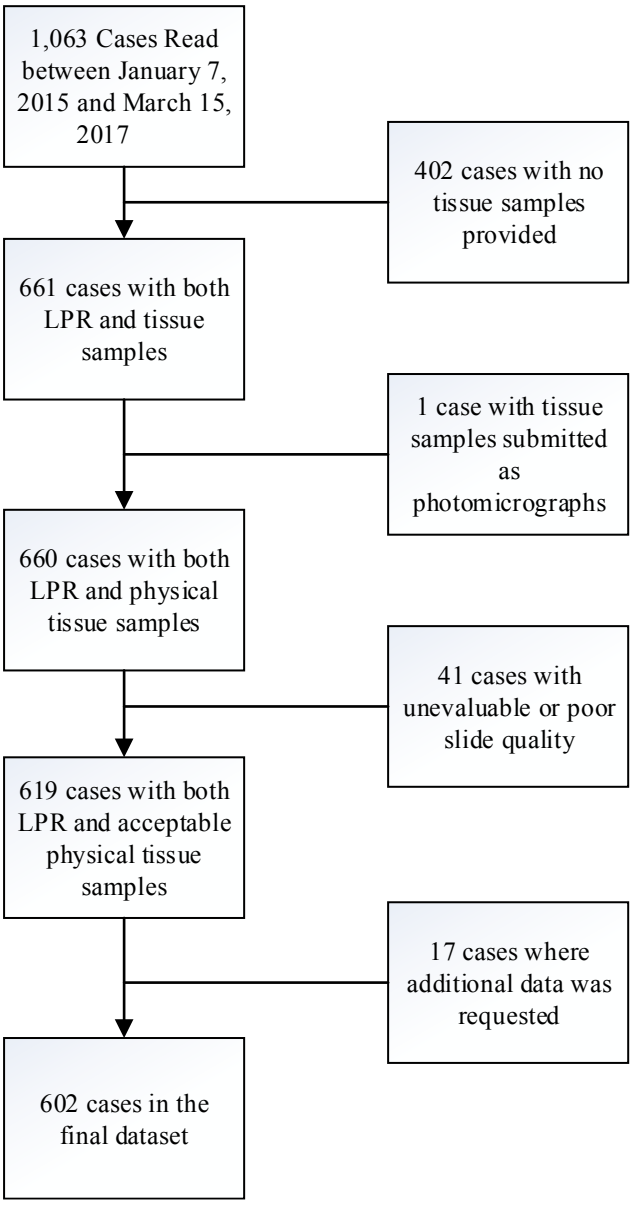

Figure 5: Study Population Flow Chart.

The analyses performed and their associated identifiers are outlined in Table 10.

Table 10: Analysis Legend.

Analysis Code	Description
Analysis 1	Neoplasm versus Non-Neoplasm
Analysis 2.1	Benign versus Malignant (all cases)
Analysis 2.2	Benign versus Malignant (cases assessed as non-neoplasm by the site or authoritative central pathologist removed)
Analysis 3.1	Discrepancies in Morphology (all cases)
Analysis 3.2	Discrepancies in Morphology (cases assessed as non-neoplasm by the site or authoritative central pathologist removed)
Analysis 4.1	All discrepancies leading to differences in treatment (cases NOT similar), including those identified in categories 1, 2, and 3 (all cases)
Analysis 4.2	All discrepancies leading to differences in treatment (cases NOT similar), including those identified in categories 1, 2, and 3 (cases where site and central diagnosis matched in the first round of reviews removed)

Results by Statistical Sub-Group

The interrater variability results for all groups are below. Each table for analyses 1, 2.1, 2.2, and 3 will include the kappa score, 95% confidence interval and p-value for the kappa score, kappa agreement category, percent agreement, and the sample size (N). The kappa category qualitatively describes the level of agreement for each analysis. Results from analysis 4 will be addressed separately.

Group A interrater variability results are for analyses 1, 2.1, 2.2, 3.1, and 3.2 in Table 11. Group A includes all cases in the dataset.

Table 11: Interrater variability results - Group A.[2]

	Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Analysis 1	0.81	0.76-0.85	P<0.001	near perfect	90.4	602
Analysis 2.1	0.84	0.79-0.88	P<0.001	near perfect	91.9	602
Analysis 2.2	0.85	0.75-0.95	P<0.001	near perfect	97.1	309
Analysis 3.1	0.59	0.55-0.64	P<0.001	moderate	68.3	602
Analysis 3.2	0.52	0.47-0.58	P<0.001	moderate	57.0	309

All results were statistically significant based on confidence intervals and p-values. Analyses 1, 2.1 and 2.2 had near perfect kappa agreement with all percent agreements over 90%. Analyses 3.1 and 3.2 had moderate kappa agreement with percent agreements of 68.3% and 57.0%, respectively.

When all cases were considered (analysis 4.1), 32% of local diagnoses did not match the final central diagnosis and the authoritative central pathologist considered 11% of diagnoses not similar. When cases with matching diagnoses in the first round of reviewers were removed (analysis 4.2), 95% of local diagnoses did not match the final central diagnosis and 34% of diagnoses were not considered similar. Due to an application error, ten cases had a similarity assessment incorrectly triggered after the authoritative read even though the local and final central diagnosis were the same. These

[2] Analysis 1: Overall dataset: neoplasm vs non-neoplasm, Analysis 2.1: benign vs malignant (all cases), Analysis 2.2: benign vs malignant (non-neoplasms removed), Analysis 3.1: discrepancies in morphology (all cases), Analysis 3.2: discrepancies in morphology (non-neoplasms removed)

account for the 5% of cases that had matching local and central diagnoses that underwent a similarity assessment.

Group B consists of all cases in group A, but separated by sub-specialty reading groups (sub-groups). Note that 4 cases were removed from the analysis because they could not be clearly categorized into the main sub-specialty reading groups. Tables 12 (analysis 1), 13 (analyses 2.1 and 2.2), and 14 (analyses 3.1 and 3.2) include the interrater variability results for each sub-group. For sub-groups in **bold**, the upper boundary of the confidence interval exceeded 1.00. Sub-groups in *italics* were not statistically significant based on the confidence interval and p value. In all tables, dermatology was the largest sub-group. In analyses 1 and 2, all categories except lymphoma had near perfect or substantial kappa agreement and percent agreement was above 80% for all categories. In table 14, the range of kappa values and percent agreements were wider than in previous analyses.

Table 12: Interrater variability results - Group B Analysis 1.[3]

	Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Breast	0.93	0.8-1.00	5E-07	near perfect	96.6	29
Cytology	0.77	0.35-1.00	0.0177	substantial	88.9	9
Dermatology	0.69	0.59-.79	p<0.001	substantial	87.4	253
ENT	0.76	0.44-1.00	0.0018	substantial	88.2	17
GI	0.8	0.68-0.92	6.66E-16	substantial	90.9	99
GU	0.92	0.85-0.99	p<0.001	near perfect	96	126
Gynecology	0.66	0.42-0.91	6.16E-06	substantial	83.3	36
Lung	1	1.00-1.00	0.000911	near perfect	100	11
Lymphoma	0.55	0.04-1.00	0.0528	moderate	80	10
Sarcoma	1	1.00-1.00	0.00468	near perfect	100	8

[3] Individual Pathology Sub-Specialty Reading Groups: neoplasm vs non-neoplasm

Table 13: Interrater variability results - Group B Analyses 2.1 and 2.2.[4]

Analysis 2.1							Analysis 2.2						
	Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N		Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Breast	**0.86**	**0.68-1.00**	**2.74E-06**	**near perfect**	**93.1**	**29**	Breast[A]	N/A	N/A	N/A	near perfect	92.9	14
Cytology	**0.77**	**0.35-1.00**	**0.0177**	**substantial**	**88.9**	**9**	Cytology*	1	1.00-1.00	p<0.001	near perfect	100	5
Dermatology	0.71	0.62-0.80	p<0.001	substantial	87.4	253	Dermatology	0.61	0.32-0.89	2.66E-15	substantial	96.4	167
ENT	**0.88**	**0.64-1.00**	**0.00027**	**near perfect**	**94.1**	**17**	ENT	1	1.00-1.00	0.0027	near perfect	100	9
GI	**0.93**	**0.83-1.00**	**p<0.001**	**near perfect**	**98**	**99**	**GI**	**0.93**	**0.80-1.00**	**3.05E-07**	**near perfect**	**96.7**	**30**
GU	0.92	0.85-0.99	p<0.001	near perfect	96	126	GU	1	1.00-1.00	9.24E-13	near perfect	100	51
GYN	0.71	0.44-0.97	1.90E-05	substantial	88.9	36	**GYN**	**0.84**	**0.55-1.00**	**2.08E-03**	**near perfect**	**92.3**	**13**
Lung	1	1.00-1.00	0.00091	near perfect	100	11	Lung*	1	1.00-1.00	p<0.001	near perfect	100	8
Lymphoma	**0.55**	**0.04-1.00**	**0.0528**	**moderate**	**80**	**10**	Lymphoma*	1	1.00-1.00	p<0.001	near perfect	100	2
Sarcoma	1	1.00-1.00	0.00468	near perfect	100	8	Sarcoma*	1	1.00-1.00	p<0.001	near perfect	100	7

[A] kappa score could not be calculated using R. Kappa assumed to be at least 0.81 (near perfect) based on percent agreement

* kappa score could not be calculated using R. Kappa score of 1.00 with 95% confidence interval of 1.00-1.00 and p value p<0.001 assigned based on values taken from other analyses with 100% agreement

Table 14: Interrater variability results - Group B Analyses 3.1 and 3.2.[5]

Analysis 3.1							Analysis 3.2						
	Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N		Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Breast	0.69	0.50-0.88	2.43E-10	substantial	79.3	29	Breast	0.45	0.17-0.73	0.00329	moderate	64.3	14
Cytology	0.53	0.20-0.85	0.00459	moderate	66.7	9	*Cytology*	*0.23*	*-0.05-0.51*	*0.171*	*fair*	*60*	*5*
Dermatology	0.45	0.38-0.52	p<0.001	moderate	53.4	253	Dermatology	0.39	0.31-0.48	p<0.001	fair	48.5	167
ENT	0.48	0.25-0.72	3.60E-07	moderate	58.8	17	ENT	0.37	0.10-0.63	0.000461	fair	44.4	9
GI	0.55	0.43-0.66	p<0.001	moderate	74.7	99	GI	0.33	0.16-0.49	4.71E-06	fair	46.7	30
GU	0.88	0.80-0.96	p<0.001	near perfect	93.7	126	GU	0.73	0.52-0.94	p<0.001	substantial	94.1	51
GYN	0.47	0.28-0.66	6.66E-12	moderate	63.9	36	GYN	0.42	0.17-0.68	5.21E-10	moderate	46.2	13
Lung	0.77	0.52-1.00	3.93E-08	substantial	81.8	11	**Lung**	**0.67**	**0.35-1.00**	**0.000102**	**substantial**	**75**	**8**
Lymphoma	0.41	0.02-0.81	5.05E-03	moderate	70	10	Lymphoma*	N/A	N/A	1.57E-01	N/A	50	2
Sarcoma	0.33	0.04-0.63	4.46E-05	fair	37.5	8	*Sarcoma*	*0.24*	*-0.031-0.51*	*0.00361*	*fair*	*28.6*	*7*

* A kappa score could not be calculated using R because the sample size was too small

[4] Individual Pathology Sub-Specialty Reading Groups: Analysis 2.1: benign vs malignant (all cases), Analysis 2.2: benign vs malignant (non-neoplasms removed)

[5] Individual Pathology Sub-Specialty Reading Groups: Analysis 3.1: discrepancies in morphology (all cases), Analysis 3.2: discrepancies in morphology (non-neoplasms removed)

Discrepancies in diagnoses for all sub-groups in category B are summarized in Table 15, including those that would lead to differences in treatment. The same 4 cases removed in the analyses 1, 2, and 3were removed for the analyses below because they could not be clearly categorized into the main sub-specialty reading groups. The percentage of dissimilar diagnoses leading to treatment differences varied by sub-group, however it is worth nothing that the ENT and lung sub-groups had no discrepancies that would lead to any treatment differences.

Table 15: Discrepancies in diagnosis and treatment - Group B Analyses 4.1 and 4.2.[6]

		Percent Mismatch between Local and Central Diagnosis	Percent Dissimilar leading to different treatments	N
Analysis 4.1	Breast	21	7	29
	Cytology	33	11	9
	Dermatology	47	17	253
	ENT	41	0	17
	GI	25	5	99
	GU	6	5	126
	GYN	36	19	36
	Lung	18	0	11
	Lymphoma	30	20	10
	Sarcoma	62	25	8
Analysis 4.2	Breast	86	29	7
	Cytology	100	33	3
	Dermatology	97	35	122
	ENT	100	0	7
	GI	93	19	27
	GU	89	67	9
	GYN	93	50	14
	Lung	100	0	2
	Lymphoma	100	67	3
	Sarcoma	100	40	5

[6] Individual Pathology Sub-Specialty Reading Groups: Analysis 4.1: All discrepancies leading to differences treatment (cases NOT similar), including those identified in categories 1, 2, and 3, Analysis 4.2: All discrepancies leading to differences treatment (cases NOT similar), including those identified in categories 1, 2, and 3 (round 1 agreements removed)

Thirty-two percent of local diagnoses did not match the final central diagnosis. 11% of local and central diagnoses were not similar when all cases in group B were considered (4.1). When cases with matching diagnoses in the first round of reviews were removed (4.2), 95% of local diagnoses did not match the final central diagnosis and 34% were considered not similar. The averaged results for both analyses 4.1 and 4.2 for group B are logically the same as those for group A, as group A is comprised of all sub-groups.

The group C (NMSC vs other malignancies (non-NMSC)) interrater variability results for analyses 1, 2.1, 2.2, 3.1, and 3.2 are presented in Table 16. Note that 2 cases were not included in the analysis because both non-melanoma skin cancer and melanoma were provided as diagnoses in at least one round during the review process. Almost all categories had near perfect or substantial kappa agreement except in analysis 3.2.

Table 16: Interrater variability results - Group C.[7]

		Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Analysis 1	NMSC	0.72	0.62-0.82	p<0.001	substantial	88.8	240
	Other	0.83	0.77-0.89	p<0.001	near perfect	91.7	360
Analysis 2.1	NMSC	0.74	0.65-0.83	p<0.001	substantial	88.8	240
	Other	0.87	0.82-0.93	p<0.001	near perfect	94.2	360
Analysis 2.2	NMSC	0.65	0.34-0.97	p<0.001	substantial	97.5	160
	Other	0.87	0.76-0.97	p<0.001	near perfect	95.9	148
Analysis 3.1	NMSC	0.46	0.39-0.53	p<0.001	moderate	54.6	240
	Other	0.66	0.61-0.72	p<0.001	substantial	77.5	360
Analysis 3.2	NMSC	0.39	0.30-0.48	p<0.001	fair	48.8	160
	Other	0.57	0.49-0.65	p<0.001	moderate	65.5	148

In analysis 4.1, when all NMSC cases were considered, 45% of local diagnoses did not match the final central diagnosis and 15% of local and central diagnoses were not considered similar. When cases with matching diagnoses in the first round of reviewers were removed (4.2), 96% of local diagnoses did not match the final central diagnosis and 33% were considered not similar. 22% of local diagnoses did not match the final central diagnosis and the authoritative central pathologist did not consider 9% similar for non-NMSC in analysis 4.1. When non-NMSC cases with matching diagnoses in the first round of reviewers were removed (4.2), 93% of local diagnoses did not match the final central diagnosis and 36% were considered not similar.

[7] Non-melanoma skin cancers versus all other malignant processes (non-NMSC): Analysis 1: neoplasm vs non-neoplasm, Analysis 2.1: benign vs malignant (all cases), Analysis 2.2: benign vs malignant (non-neoplasms removed), Analysis 3.1: discrepancies in morphology (all cases), Analysis 3.2: discrepancies in morphology(non-neoplasms removed)

Results by Analysis

Figure 6 summarizes the results of analysis 1 (separating specimens between neoplasms and non-neoplasms) for all groups. The average kappa score (0.80) and percent agreement (90.9%) were the same for group A as well as when the kappa scores and percent agreements for all sub-groups when averaged together. All percent agreements for analysis 1 surpassed the minimally acceptable agreement rate in a healthcare setting of 80%.

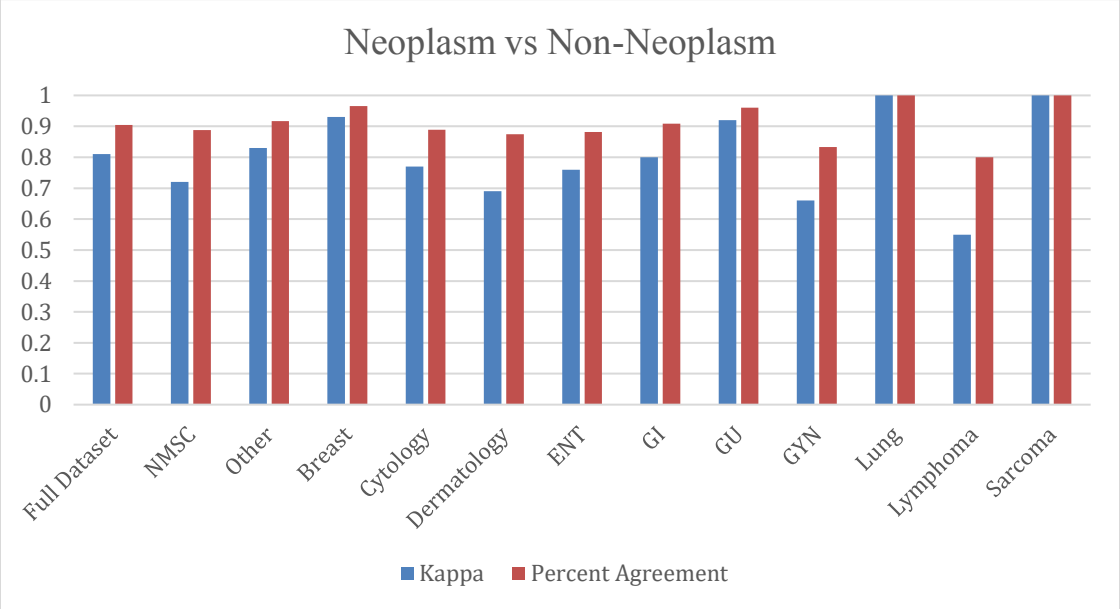

Figure 6: Analysis 1 results for all groups.

The kappa category distribution is in Figure 7.

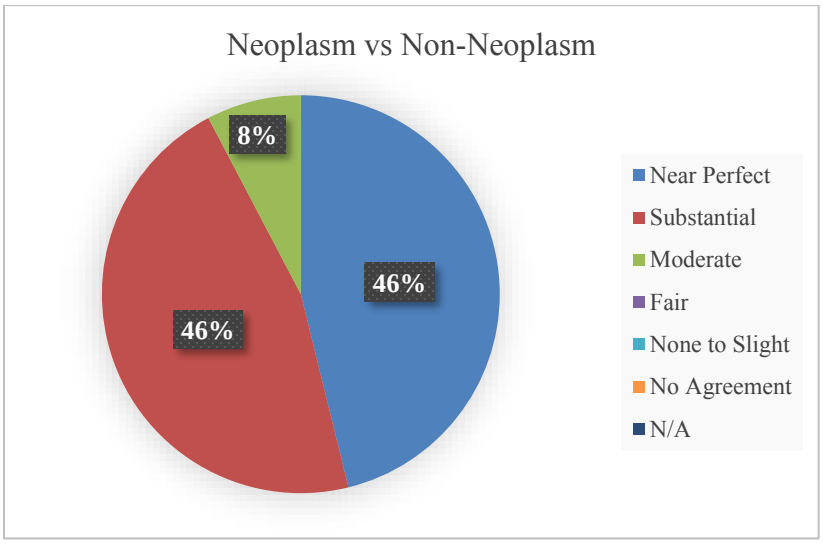

Figure 7: Analysis 1 - Kappa categories.

Ninety two percent of sub-groups (all except lymphoma) had either near perfect or substantial agreement; the lymphoma sub-group had moderate agreement.

Analysis 2 assessed interrater variability when determining whether specimens were benign or malignant. This distinction is crucial as misdiagnosis could result in inaccurate risk profiles and/or grave health consequences for subjects. Figure 8 shows the differences in kappa scores when non-neoplasms were (2.1) and were not (2.2) included in the analysis.

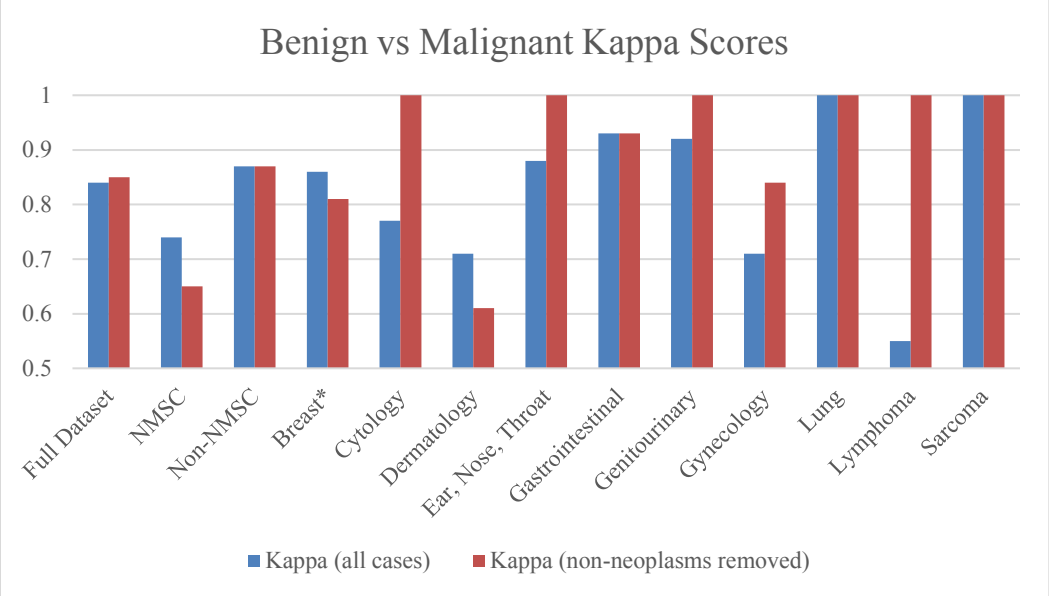

Figure 8: Analyses 2.1 and 2.2 - Benign vs malignant kappa scores for all groups.

A kappa score for the sub-category "breast" could not be calculated using R when non-neoplasms were removed. The value is assumed to be at least 0.81 (near perfect agreement) based on percent agreement; this is the value used in the figure above.

The group A kappa score when including and not including non-neoplasms was 0.84 and 0.85, respectively. The average kappa score among all sub-categories when including non-neoplasms was 0.83; when excluding non-neoplasms, it was 0.89. All kappa scores indicated at least moderate agreement and most displayed substantial to near perfect agreement.

Figure 9 shows percent agreement among raters when assessing whether a sample was benign or malignant. All analysis groups were included.

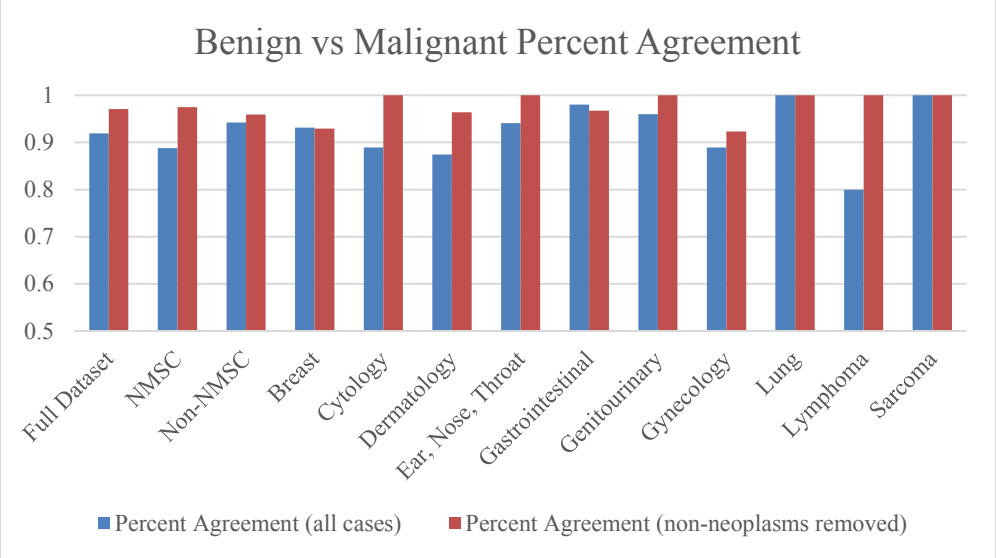

Figure 9: Analyses 2.1 and 2.2 - Benign vs malignant percent agreement for all groups.

Average percent agreement among sub-groups when including and not including non-neoplasms was 92.4% and 97.6%, respectively. Both were within 1.5% of the group A values (91.9% and 97.1%, respectively). Raters reached the suggested minimally acceptable percent agreement in healthcare (80%) in all sub-categories.

Kappa scores for discrepancies in morphology codes when non-neoplasms were (analysis 3.1) and were not (analysis 3.2) included are in Figure 10.

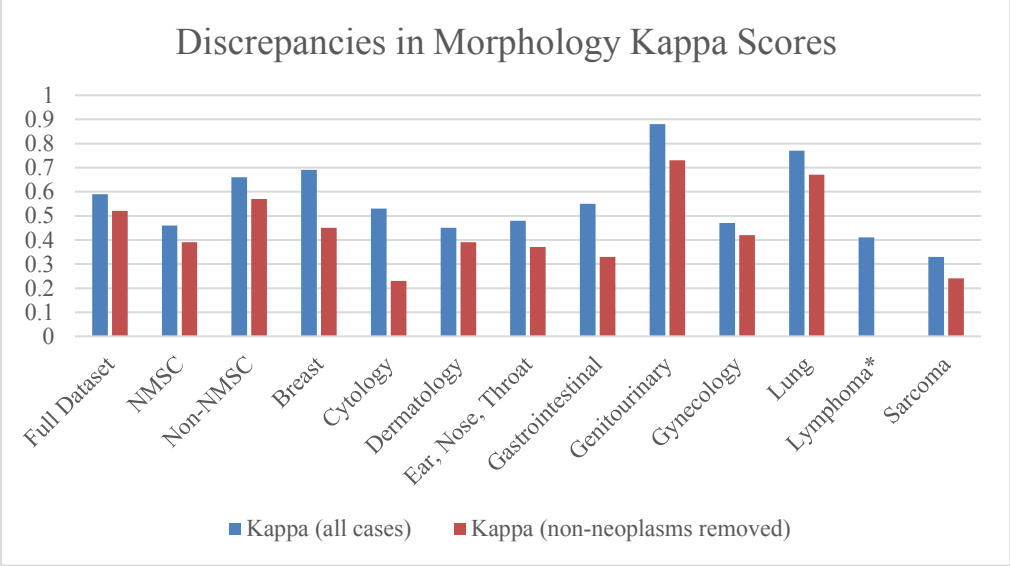

Figure 10: Analyses 3.1 and 3.2 - Discrepancies in morphology kappa scores for all groups.

A kappa score for sub-category lymphoma cases could not be calculated using R when non-neoplasms were not included due to sample size.

Kappa scores for this analysis were considerably lower than those in previous analyses. The group A kappa score was 0.59 when including non-neoplasms and 0.52 when they were removed. The average kappa score for all sub-categories when including non-neoplasms was 0.56; when excluding non-neoplasms, it was 0.44.

Figure 11 shows percent agreement in morphology when non-neoplasms were and were not included.

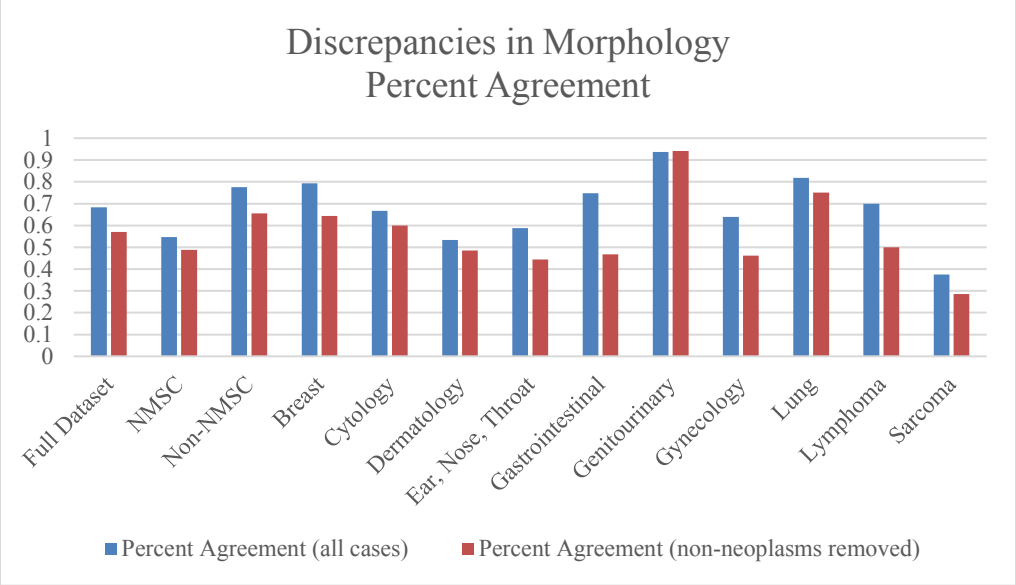

Figure 11: Analyses 3.1 and 3.2 - Discrepancies in morphology percent agreement for all groups.

The average percent agreement was 67.7% when including non-neoplasms and 56.1% when excluding non-neoplasms.

Kappa agreement categories for analyses 2.1, 2.2, 3.1, and 3.2 are presented in Figure 12. When considering category A and all sub-groups in categories B and C, there were 13 separate analysis groups (full data set, breast, cytology, dermatology, ENT, GI, GU, GYN, lung, lymphoma, sarcoma, NMSC, and non-NMSC malignancies); therefore, percentages were calculated based on a total of thirteen groups.

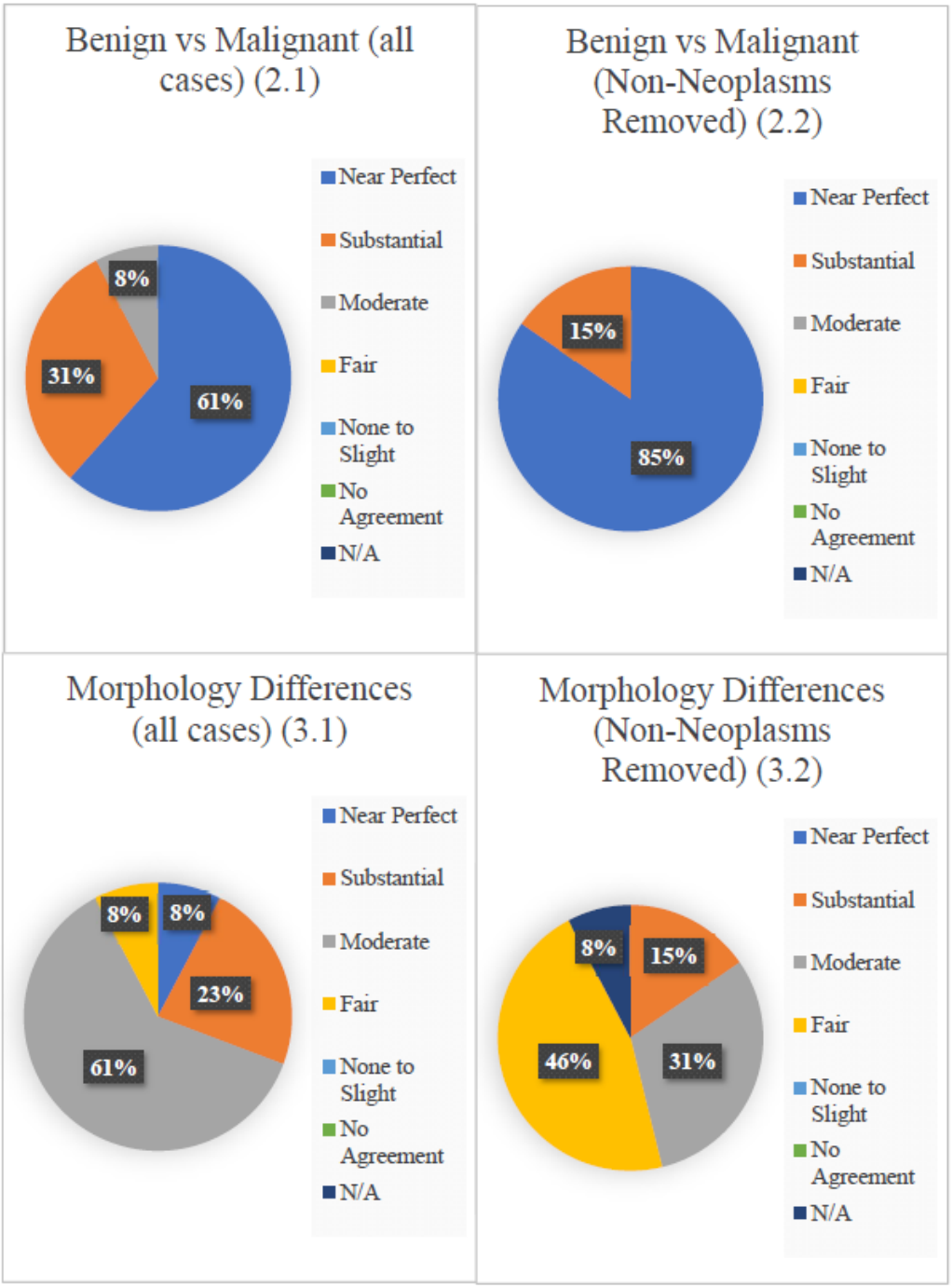

Figure 12: Kappa categories for analyses 2.1, 2.2, 3.1, 3.2.[8]

[8] Analysis 2.1: benign vs malignant (all cases), Analysis 2.2: benign vs malignant (non-neoplasms removed), Analysis 3.1: discrepancies in morphology (all cases), Analysis 3.2: discrepancies in morphology(non-neoplasms removed)

Analysis 2.1 had a similar kappa category distribution as analysis 1 (92% at either near perfect or substantial agreement and 8% at moderate agreement (lymphoma). In analysis 2.1 however there were more cases that showed near perfect agreement versus substantial agreement when in analysis 1 there were equal numbers of cases in each category. When non-neoplasms were removed (2.2), 85% of cases (all except NMSC and dermatology) had near perfect agreement and the remaining 15% had substantial agreement.

Analyses 3.1 and 3.2 had lower levels agreement overall. When including all cases, 8% had near perfect agreement, 23% had substantial agreement, 61% had moderate agreement, and 8% had fair agreement. When non-neoplasms were removed, 15% had substantial agreement, 31% had moderate agreement, 46% had fair agreement, and 8% (lymphoma) did not have a kappa score calculated and therefore could not be categorized.

Kappa scores and percent agreement quantify how concordant two raters are, but cannot assess whether there is a treatment difference between different diagnoses. The discrepancies in diagnoses when including all cases are in Figure 13. This includes cases that did not have a comparison between the local and central diagnosis performed because they were an exact match in round 1 of reviews; these cases were considered similar for this analysis.

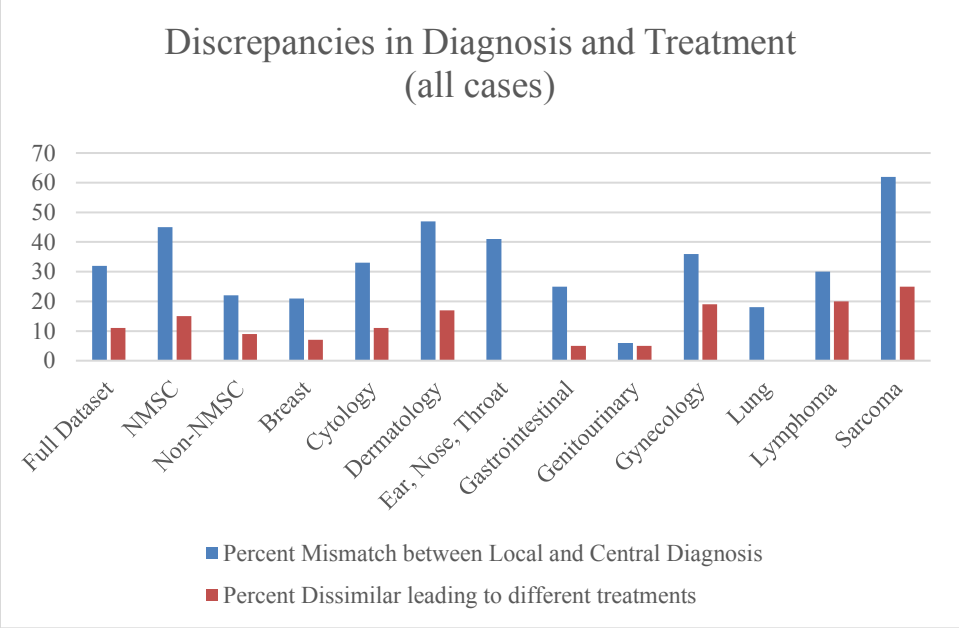

Figure 13: Analysis 4.1 - Discrepancies in diagnoses for all groups.

On average, 32% of local diagnoses did not match the final central diagnosis; 11% of cases had a central and a local diagnosis that could lead to treatment differences (not similar). Only lymphoma (20% discrepant) and sarcoma (25% discrepant) did not meet the minimally acceptable threshold in healthcare of 80% agreement. The median percent disagreement was 11%.

Figure 14 includes only cases that did not have matching diagnoses in round 1 of reviews and therefore had an assessment of similarity between diagnoses performed by the authoritative reviewer during either Round 2 or 3 of the read process.

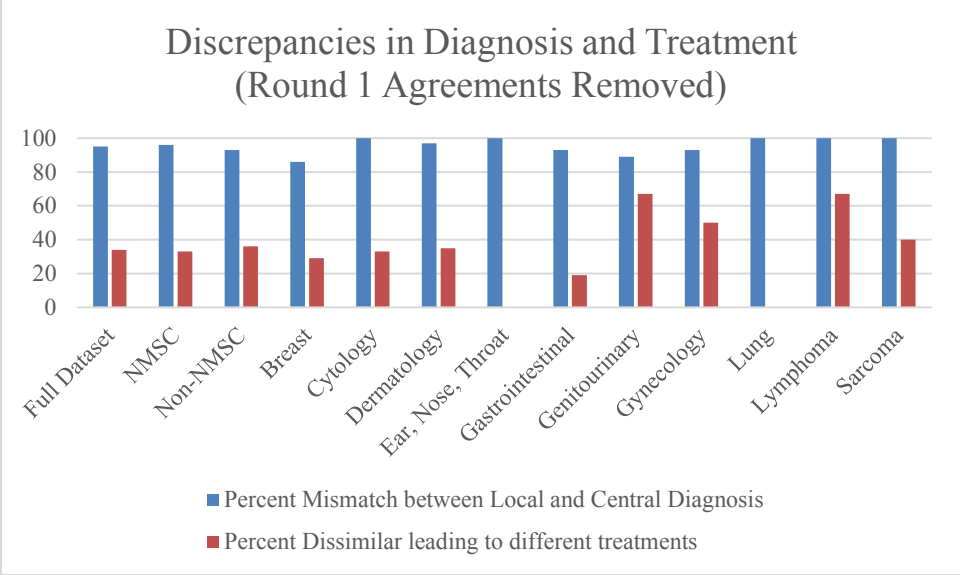

Figure 14: Analysis 4.2 - Discrepancies in diagnoses for all groups.

Almost all sub-groups had near 100% mismatch between local and central diagnoses. Due to an application error, ten cases had a similarity assessment incorrectly triggered after the authoritative read even though the local and final central diagnosis were the same. Sixty six percent of cases on average had similar assessments, which does not meet the minimally acceptable threshold of agreement in healthcare. The maximum percent disagreement was 67% (GU and lymphoma) and the minimum percent disagreement was 0% (ENT). The median percent disagreement was 34%.

DISCUSSION

Analysis 1 included all cases in the dataset. Analyses 2 and 3 were conducted under two sets of circumstances: including (analyses 2.1 and 3.1) and excluding (analyses 2.2. and 3.2) non-neoplasms. ICD-O 3 coding only categorizes neoplasms; depending on the behavior code the specimen is either benign or malignant. This is only a general rule

however, as some cases are borderline between benign and malignant. These cases may require additional workup to better assess their behavior. Tissue growths categorized as *in situ* are examples of borderline cases; they are along the continuum of morphological change between dysplasia and invasive cancers.[62] This level of detail was not taken into account for this study. Performing analyses with and without non-neoplasms permits insight into how the neoplasm versus non-neoplasm agreement rate affects other agreements at more detailed levels of categorization. For this reason, analyses 2.2 and 3.2 should be more representative of the agreement rate for differentiating between benign and malignant samples and matching morphology codes, respectively.

Analyses 1, 2.1, and 2.2 Discussion[9]

Based on kappa scores, there is near perfect agreement between the central and local lab diagnoses in analyses 1 (neoplasm versus non-neoplasm), 2.1 (benign versus malignant including non-neoplasms), and 2.2 (benign vs malignant excluding non-neoplasms) in group A (all cases in the dataset). The percent agreement for these analyses is above 90% as well. The high level of interrater agreement for this analysis is expected as it is the most general categorization of specimens. The results instill confidence that there are few misdiagnoses at this level.

Differences in levels of agreement appear when the data set is separated into sub-groups (NMSC vs all other malignancies (group C) and by sub-specialty pathology reading group (group B). In group C, NMSC had substantial kappa agreement in

[9] Analysis 1: all cases, Analysis 2.1: benign vs malignant (all cases), Analysis 2.2: benign vs malignant (non-neoplasms removed)

analyses 1, 2.1, and 2.2, whereas all other malignancies had near perfect kappa agreement. All percent agreements were above 88% and surpassed the minimally acceptable threshold for interrater percent agreement in healthcare (80%).

Group B divided the data set into 10 sub-specialty reading groups. Kappa scores ranged from 0.55 (lymphoma) to 1.00 (lung) in analysis 1. In analysis 2.1, lung and sarcoma had the highest kappa scores (1.00) and dermatology and GYN had the lowest (0.71). The kappa score for lymphoma was 0.55 but was not statistically significant. When non-neoplasms were removed in analysis 2.2, 6 of the 10 sub-groups had kappa scores of 1.00, but all 6 had sample sizes less than 10. Percent agreement ranged from 80 to 100 percent across analyses 1, 2.1, and 2.2.

The sample sizes for most sub-groups in group B were significantly smaller when non-neoplasms were removed, therefore estimated kappa scores for analysis 2.2 were less certain than those for analyses 1 or 2.1. Averaging many kappa scores generated from smaller samples may not be as accurate a representation of the true population value as a kappa score obtained from one large sample. This may be one reason why the average sub-category kappa score for analysis 2.2 and the corresponding group A kappa score differ (0.85 and 0.89, respectively).

In analysis 2.1 and 2.2, most kappa scores for all sub-groups in groups B and C remained stable or increased when non-neoplasms were removed. This implies that the distinction between neoplasm and non-neoplasm is equally or more variable than the benign versus malignant categorization. The kappa scores for NMSC, non-NMSC

malignancies, and GI cases remained stable or decreased when non-neoplasms were removed, indicating that the distinction between neoplasm and non-neoplasm is equally or less variable than distinguishing benign versus malignant. The kappa score for breast should not be considered because the value was estimated based on percent agreement. The NMSC, dermatology, and non-NMSC sub-groups had the largest sample sizes in all analyses. Assuming that the larger sample size the more accurately the estimate reflects the true population value, the behavior of the kappa scores for the NMSC, dermatology, and non-NMSC sub-groups may be most representative of the relationship between including and excluding non-neoplasms and sub-group kappa scores. Had the sample sizes for the other sub-groups been larger, their kappa scores might have behaved similarly.

When dividing the dataset into sub-groups, based on kappa scores it is unclear whether raters are more or less concordant in categorizing specimens as benign or malignant when non-neoplasms are removed from the analysis. Evaluating only sub-groups with sample sizes over 100 in analyses 2.1 and 2.2, raters were less concordant when non-neoplasms were removed; otherwise, reviewers were more concordant.

Except for breast and GI, percent agreement in groups B and C was higher when non-neoplasms were removed from the analysis. This is the same result in the overall dataset calculation. The percent agreement suggests that the distinction between neoplasm and non-neoplasm is more variable among raters than whether a sample is benign or malignant.

Analyses 3.1 and 3.2 Discussion[10]

The group A (full dataset) kappa score and percent agreement for discrepancies in morphology were 0.59 and 68.3%, respectively. When non-neoplasms were removed (analysis 3.2), the kappa score and percent agreement were 0.52 and 57.0%, respectively. The average kappa score for all sub-groups (groups B and C) when non-neoplasms were included was 0.56; when they were not included it was 0.44. The average percent agreement among sub-groups when including and not including non-neoplasms was 67.7% and 56.0%, respectively. As with earlier analyses, the difference between the group A and sub-group averages for kappa scores and percent agreement may be due to the averaging of kappa scores from smaller samples versus obtaining one kappa score from a larger data set. Lower agreement rates in analyses 3.1 and 3.2 are expected: instead of categorizing specimens into one or two categories (neoplasm versus non-neoplasms and/or benign versus malignant), raters had thousands of ICD-O code combinations to choose from. Overall, kappa scores and percent agreements were higher when non-neoplasms were included, which implies that differentiating neoplasms from non-neoplasm is more consistent among raters than assigning specific diagnoses to samples.

[10] Analysis 3.1: discrepancies in morphology (all cases), Analysis 3.2: discrepancies in morphology (non-neoplasms removed)

Analyses 4.1 and 4.2 Discussion[11]

Diagnoses may be discrepant based on general categorization and more specific ICD-O codes, but the course of treatment to the subject could be the same regardless. In Speight's case study, reviewer pairs agreed on a diagnosis in 69.9% of cases in the first round of reviews.[48(p479)] Comparably, in this study, two-thirds of all cases had matching local and central diagnoses in the first round of reviews. The remaining 33% of cases had diagnosis discrepancies during round 1 or rounds 1 and 2. Like analyses 2.1, 2.2, 3.1, and 3.2, analysis 4 (discrepancies leading to differences in treatment) was conducted under two scenarios. The first (4.1) included all cases regardless of whether a similarity assessment needed to be performed. The second excluded cases where a similarity assessment did not occur because there was an exact ICD-O code match in the first round of reviews (4.2). The separate analyses tier cases into less challenging and more challenging and assumes that cases requiring additional review after round 1 are more challenging than those that do not. Differences in agreement between the two levels can elucidate where discrepancies exist and their implications.

When all cases were considered regardless of number of rounds of review (analysis 4.1), about 90% of diagnoses would have similar courses of treatment. All sub-groups except sarcoma reached the minimally acceptable agreement rate in healthcare (80%). Likewise, in Speight's case study, 92.7% of cases reached a final diagnosis after a

[11] Analysis 4.1: All discrepancies leading to differences treatment (cases NOT similar), including those identified in categories 1, 2, and 3, Analysis 4.2: All discrepancies leading to differences treatment (cases NOT similar), including those identified in categories 1, 2, and 3 (cases where LPR and central diagnosis matched in the first round of reviews removed)

third adjudicator looked at cases with discrepant assessments in the first round of reviews.[48(p479)] In the remaining 33% of cases that did not have matching diagnoses in round 1 (analysis 4.2), 34% may have different courses of treatment depending on whether the local or central diagnoses was used. Details regarding percentage of cases without matching diagnoses in round 1 per sub-group are in Table 17.

Table 17: Cases with round 2 and/or round 2 and 3 reviews per sub-group.

Group	Percent of Cases with Round 2 and/or 3 Reviews
All cases	33.39
NMSC	47.08
Non-NMSC	24.17
Breast	24.14
Cytology	33.33
Dermatology	48.22
ENT	41.18
GI	27.27
GU	7.14
GYN	38.89
Lung	18.18
Lymphoma	30.00
Sarcoma	62.50

No sub-groups except GI, lung, and ENT met the minimally acceptable threshold of agreement (80%) in analysis 4.2. The local and central diagnoses for all ENT and lung cases that went past round 1 of reviews were considered similar (100% agreement); 81% of GI cases were considered similar. GU and lymphoma had the highest levels of disagreement at 67%.

Although non-melanoma skin cancers and all other malignant processes both represent malignant processes, the significant differences between them are important from a safety perspective. Cancers that are not non-melanoma skins cancers tend to be harder to treat, have higher mortality rates, and are more serious safety concerns that non-melanoma skin cancers. The risk of dissimilar diagnoses that could potentially lead to different treatments was comparable for both NMSC and non-NMSC malignancies, despite kappa scores and percent agreements for non-NMSC malignancies being higher in all analyses except analysis 2.2 (difference of less than 2%). NMSC had a slightly higher risk of dissimilar diagnoses in analysis 4.1, but when looking at challenging cases only, NMSC had a 33% risk and other malignancies had a 36% risk of dissimilar diagnoses. Equal attention needs to be paid to both sub-groups, as neither category seems to significantly increase risk of incorrect treatment.

The adjudication model used in Speight's study and in this study differ slightly. In Speight's study, in the case of discrepancy between reviewers A and B, the case went to an adjudicator. If a final diagnosis could not be decided, the case went to a consensus review. In this study, in the case of discrepancy between reviewers A (site pathologist) and B (central pathologist 1 or PR1), the case went an additional central pathologist (PR2). If PR2 and PR1 do not agree on a diagnosis, then the case went to a final central pathologist (adjudicator or PR3). PR3 in this study replaced the consensus review in Speight's study. This study also had a similarity assessment component. The results in both studies however were comparable: the percent of cases with similar (as defined in this study's charter) diagnoses increased with additional rounds of reviews.

Special Considerations

Charter and CPC/Reviewer Manual Updates

Although the charter, CPC manual, and reviewer manual were updated at various points throughout the study, these changes should have had minimal to no impact on the diagnosing of specimens. Tables 6 (charter) and 7 (reviewer/CPC manuals) indicate the document updates that may affect assessments in bold. These updates were further assessed to determine whether they may have had an impact on the analyses performed in this study. The only charter update that may have changed how cases were being assessed mid-study was the addition of clear definitions of slide quality. This change was made in revision 4.0. Seventy-seven percent of cases were read before revision 4.0 of the charter was made effective. Reviewers may have categorized slides differently before and after the update. Slides with poor quality were excluded from the analysis, so miscategorization of slide quality could have led to improper inclusion or exclusion of samples. Based on the conservative approach taken by this adjudication committee, it is likely that a reviewer would downgrade slide quality versus upgrade; the risk is low that slides with poor quality were categorized as having fair or good quality. Therefore, it is more likely that cases with acceptable quality were excluded than that cases with poor quality were included. Under this assumption, little to no data included in this analysis could have been generated from slides of low quality.

Two updates made in revision 2.0 to the reviewer and coding manuals may have affected how specimens were read. Eighty-two percent of cases were read before revision 2.0 of the reviewer and CPC manuals was effective. The first update referenced newly

added info buttons regarding the definitions of neoplasm and non-neoplasm; the second update was the addition of clarifying instructions regarding histology and behavior code selection. Prior to the introduction of the info buttons, reviewers/CPCs defined specimens as neoplasms or non-neoplasms based on their own interpretation of the terms. After the introduction of the info buttons, reviewers/CPCs could view the standardized definitions being used across the program before making an assessment. The risk of incorrect identification of a specimen as a neoplasm or non-neoplasm previous to the inclusion of the info buttons is low; understanding the difference between the two terms is part of normal coding and pathology education and practice.

Additional instructions that may have impacted assessments were added to the CPC and reviewer manuals in revision 2.0. The instructions did not change the review process but instead captured information which had already been imparted to the CPC and review teams over email and verbal communications. The instructions were to 1) select the most specific histology and behavior code and avoid general/non-specific codes unless no other code corresponded to the assessment and 2) if no code adequately described the diagnosis, the most appropriate code should be chosen and comments should be provided. These updates should not have had a significant impact on assessments even though they are directly pertinent. The introduction of instruction 2 most likely did not affect the actual selection of a code because reviewers were always required to choose a morphology code if they categorized the specimen as a neoplasm. The only difference in cases assessed before and after the reviewer/CPC manual update would have been the requirement of adding a comment. Implementation of instruction 1

could have increased or decreased discrepancies. Discrepancies between general and specific codes may have decreased after the change, but this did not ensure that the more specific codes that were chosen were not also discrepant.

Processing specimens

In Speight's study, reviewers A and B reviewed different sets of slides during the first round of review. Understanding how assessing different sets of slides for the same sample affects interrater variability is important, because assessments made on different sets of slides could increase the likelihood of discrepancy. Although it is possible that in any case the slides assessed by the site and central lab were not the same, in processed cases, the slides used to make the diagnosis by the site and the central lab were guaranteed to be different. Table 18 shows the number of specimens processed (either slides were cut and stained from blocks provided by the site or unstained slides provided by the site were stained) at the CRO-contracted (central) lab in each reading sub-group. Overall, 10.13% of cases in the dataset required processing. The dermatology and GI sub-specialties were taken as examples for this analysis because they had the most processed cases.

Table 18: Number of cases processed at central lab per sub-group.

Sub-Specialty	Number of Cases Processed
Breast	6
Dermatology	24
GI	12
GU	8
GYN	6
Lung	1
Lymphoma	2
OTHER	1
Sarcoma	1
Grand Total	61

For analyses 2.1 and 3.1, processed cases constituted 9.5% of the dermatology sub-group and 12.1% of the GI sub-group. When non-neoplasms were excluded (analyses 2.2 and 3.2), processed cases made up 9.0% of dermatology sub-group and 16.7% of the GI sub-group.

The kappa scores and percent agreements for analyses 2.1 and 2.2 separated by sub-group (dermatology or GI) and processing status (processed or not processed) are in Table 19. For sub-groups in **bold**, the upper boundary of the confidence interval exceeded 1.00.

Table 19: Interrater Variability Results Based on Processing (GI and Dermatology) – Analyses 2.1 and 2.2.[12]

Analysis 2.1							Analysis 2.2						
	Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N		Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Dermatology Processed	**0.91**	**0.75-1.00**	**7.12E-0.6**	**Almost Perfect**	**95.8**	**24**	Dermatology Processed[A]	N/A	N/A	N/A	Near Perfect	93.3	15
Dermatology Not Processed	0.69	0.59-0.79	p<0.001	Substantial	86.5	229	Dermatology Not Processed	0.65	0.37-0.93	2.22E-16	Substantial	96.7	152
GI Processed	1.00	1.00-1.00	0.000532	Almost Perfect	100	12	GI Processed	1.00	1.00-1.00	0.0253	Almost Perfect	100.0	5
GI Not Processed	**0.95**	**0.86-1.00**	**p<0.001**	**Almost Perfect**	**98.8**	**86**	GI Not Processed	1.00	1.00-1.00	9.63E-07	Almost Perfect	100.0	24
							[A] kappa score could be calculated using R. Kappa assumed to be at least 0.81 (near perfect) based on percent agreement						

GI case outcomes were less variable than those for dermatology regardless of processing status and analysis. Processed dermatology cases had less variable outcomes than the non-processed dermatology cases in both analyses. Based on these results, processing does not seem to have a significant impact on kappa scores when distinguishing between benign and malignant samples (analyses 2.1 and 2.2).

The relationship between processing and assigning specific morphologies is less clear. The kappa scores and percent agreements for analyses 3.1 and 3.2 separated by sub-group (dermatology or GI) and processing status (processed or not processed) are in Table 20. For sub-groups in **bold**, the upper boundary of the confidence interval exceeded 1.00. Sub-groups in italics were not statistically significant based on the confidence interval and p-value.

[12] Individual Pathology Sub-Specialty Reading Groups: Analysis 2.1: benign vs malignant (all cases), Analysis 2.2: benign vs malignant (non-neoplasms removed)

Table 20: Interrater Variability Results Based on Processing (GI and Dermatology) – Analyses 3.1 and 3.2.[13]

Analysis 3.1							Analysis 3.2						
	Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N		Kappa	95% Confidence Interval	P-value	Kappa Category	Percent Agreement	N
Dermatology Processed	0.32	0.15-0.49	5.39E-06	Fair	41.7%	24	*Dermatology Processed*	*0.12*	*-0.0054-0.29*	*1.04E-01*	*None to Slight*	*20.0%*	*15*
Dermatology Not Processed	0.46	0.39-0.54	p<0.001	Moderate	54.6%	229	Dermatology Not Processed	0.42	0.33-0.51	p<0.001	Moderate	51.3%	152
GI Processed	0.62	0.30-0.95	0.000403	Substantial	75.0%	12	***GI Processed***	***0.58***	***-0.052-1.00***	***0.0504***	***Moderate***	***80.0%***	***5***
GI Not Processed	0.53	0.41-0.66	p<0.001	Moderate	75.0%	86	GI Not Processed	0.29	0.11-0.47	1.61E-04	Fair	41.7%	24

All kappa scores decreased from analyses 2.1 and 2.2 to analyses 3.1 and 3.2. The kappa scores for both GI and dermatology processed cases in analysis 3.2 were not statistically significant. Similarly to categorizing specimens between benign and malignant, when non-neoplasms were included (analysis 3.1), all GI cases regardless of processing status had higher kappa scores (0.62 processed and 0.53 not processed) than those for dermatology (0.32 processed and 0.46 not processed). Even though the GI sub-group had a higher percentage of processed cases, again the outcomes between the central and local labs were less variable than outcomes in the dermatology sub-group.

When non-neoplasms were removed however (analysis 3.2), all kappa scores decreased. The pattern in analyses 2.1, 2.2 and 3.1 was not present here. In analysis 3.2, the highest kappa score was for GI processed cases, followed by dermatology cases that were not processed, then GI cases that were not processed, and finally processed dermatology cases. The kappa scores for both processed groups were not statistically

[13] Individual Pathology Sub-Specialty Reading Groups: Analysis 3.1: discrepancies in morphology (all cases), Analysis 3.2: discrepancies in morphology (non-neoplasms removed)

significant, so the ordering above might not be representative of the true order of the population.

Dermatology morphology has far more minutiae and therefore potential for variability than that for GI, which may explain why GI kappa scores were greater than those for dermatology in almost all analyses and categories (conversation with Dr. Glenn Bubley, July 5, 2017). The complexity of dermatology cases is so great, that even when GI slides reviewed at the site and at the central lab were guaranteed to be different, the kappa scores for all GI cases were still higher than those for any dermatology cases. Due to the small sample sizes and uncertainty in the results for processed cases in analysis 3.2, it is difficult to make a conclusion on the effects of processing on assigning specific morphology of neoplasms.

The results for discrepancies in diagnoses and potential treatment differences (analysis 4.1 and 4.2) are in Table 21.

Table 21: Discrepancies in diagnosis and treatment results based on processing (GI and Dermatology) – Analyses 4.1 and 4.2.[14]

		Percent Mismatch between Local and Central Diagnosis	Percent Dissimilar leading to different treatments	N
Analysis 4.1	Dermatology Processed	58	29	24
	Dermatology Not Processed	45	16	229
	GI Processed	25	0	12
	GI Not Processed	25	6	87
Analysis 4.2	Dermatology Processed	82	41	17
	Dermatology Not Processed	99	34	105
	GI Processed	74	0	4
	GI Not Processed	96	22	23

In analysis 4.1, GI processed and not processed cases had the same percentage of mismatched diagnoses between the local and central labs, but non-processed cases had more discrepancies leading to treatment differences. Processed dermatology cases had a higher rate of mismatched diagnoses and discrepancies leading to treatment differences. In analysis 4.2, dermatology and GI cases that were not processed had more discrepant diagnoses than processed cases. Processing led to a larger percentage of treatment differences in the dermatology sub-group, but a smaller percentage of treatment differences in the GI sub-group. This may be due to the lower number of morphology

[14] Individual Pathology Sub-Specialty Reading Groups: Analysis 4.1: All discrepancies leading to differences treatment (cases NOT similar), including those identified in categories 1, 2, and 3, Analysis 4.2: All discrepancies leading to differences treatment (cases NOT similar), including those identified in categories 1, 2, and 3 (cases where LPR and central diagnosis matched in the first round of reviews removed)

code combinations in this data set for GI (28) versus dermatology (71), however this difference may just be due to the higher number of dermatology cases overall.

Kappa score vs Percent Agreement

Kappa scores and percent agreement only provide estimates of concordance between raters. The true rate of discordance is nearly impossible to quantify with the statistical tools currently available. The kappa and percent agreement statistics are not directly comparable. For example, in analysis 2.2, the kappa score for dermatology was 0.61, yet the percent agreement was 96.4%. In a study comparing automated versus human visual detection of abnormalities in biological samples, similar findings were noted. Results showed only moderate agreement between the human and the automated detection, but the percent agreement generated from the same data showed 94.2% agreement.[63] According to the kappa score, there was (barely) substantial agreement, but based on percent agreement there was nearly perfect agreement.

Mary McHugh tries to address the gap between kappa scores and percent agreement by interpreting Cohen's kappa as not only a measure of agreement, but disagreement as well. She extrapolates estimates of reliability of data from the kappa score as an approximate comparator for percent disagreement. Stricter categorization of levels of agreement also better reflect the implications of different levels of disagreement. This interpretation is in Table 22.[49(p4)]

Table 22: McHugh kappa interpretation.[49(p4)]

Value of Kappa (Cohen)	Value of Kappa (McHugh)	Level of Agreement (Cohen)	Level of Agreement (McHugh)	% of Data that are Reliable (McHugh)
0.01-0.20	0.00-0.20	None to Slight	None	0-4%
0.21-0.40	0.21-0.39	Fair	Minimal	4-15%
0.41-0.60	0.40-0.59	Moderate	Weak	15-35%
0.61-0.80	0.60-0.79	Substantial	Moderate	35-63%
	0.80-0.90	Almost Perfect	Strong	64-81%
0.81-1.00	Above 0.90		Almost Perfect	82-100%

In this study, the average kappa score for neoplasm versus non-neoplasm and benign versus malignant categorization could be considered almost perfect based on Cohen's stratification, but only strong based on McHugh's. The average kappa score for discrepancies in morphology codes change from moderate to weak agreement when using McHugh's stratification. The most significant difference between McMugh and Cohen's kappa categorizations is the interpretation of kappa scores above 0.80. Cohen prescribes one level of agreement to scores between 0.81 and 1.00 (almost perfect) where McHugh has two (strong versus almost perfect). Although the difference may seem negligible qualitatively, from a data reliability standpoint the difference may be significant. Strong agreement implies 64-81% data reliability whereas almost perfect agreement implies 82-100% data reliability.

The kappa score and percent agreement both have limitations. The kappa score considers chance agreement: the greater the expected chance agreement, the lower the kappa score. The chance agreement depends on the marginal sums of the chi-square table on which the statistic is based. The expected chance agreement depends on the following

assumptions: raters guess on every item, raters guess at rates similar to the marginal proportions, and raters are entirely independent. These assumptions cannot be confidently verified. Statistical significance of kappa scores is also difficult to ascertain when there is inconsistency with scoring; variability in assessments is exactly what the kappa score measures. Large confidence intervals may also span many levels of agreement which makes it difficult to obtain true meaning from the kappa score itself. Percent agreement assumes that the majority result is correct and the minority result is incorrect; for example, if there is 90% agreement, the values in the 90% are correct and the values in the remaining 10% are not. Percent agreement also assumes that raters make informed and deliberate choices in all assessments. Like kappa score conditions, these premises can also not be verified with confidence.[49(p2-8)]

Neither kappa scores nor percent agreement can completely quantify or explain interrater variability. Understanding the strengths and weaknesses of each can help identify when one might be more applicable than the other. If subjective assessments (i.e. presence/absence of abnormal morphology in a biological specimen) inherently have more guesswork than objective assessments (i.e. lab values), then kappa scores might be more useful in assessing interrater variability for subjective assessments. Percent agreement might be more suitable for objective assessments. In a healthcare setting, it is best practice to consider both statistics.

Limitations and Future Studies

There are two major limitations for this study. The first is small sample sizes. Although the overall sample was substantial (N=602), when the dataset was divided into sub-groups the sample sizes were not large enough to generate results with confidence. As a result, the confidence intervals for many of the kappa scores were very wide, spanning multiple levels of agreement. Per McHugh, sample sizes should never be less than 30 and ideally should exceed 1,000 in order to get the most accurate and dependable statistics.[49(p8)] It is possible that due to small sample sizes the kappa scores and percent agreements generated as part of this study are misrepresentations of the true population averages.

The second limitation is that the information used to make an assessment at the site was most likely not the same as that used by the central pathologists. Except in cases where processing was performed at the central lab, there is no way to confirm that the slides sent to the central lab are the same slides and/or cut from the same tissue block that were used to make the diagnosis at the local lab. Tissue samples on most slides are very small and could easily represent different parts of a lesion/tumor even if they are cut from the same tissue block. For example, one slide could have tissue taken from a tumor margin and show no malignancy; another sample could be taken from the center of the same tumor and show evidence of malignancy. Although processing of blocks at the central lab didn't seem to have a significant effect on kappa scores or percent agreement, this does not imply that if a larger number of specimens were known to be discrepant there wouldn't be an effect on either statistic.

In addition to slides, there is almost certainly a discrepancy in supporting clinical information provided to the local and central labs. In most clinical practices, the pathologists receive clinical histories, impressions, and additional supporting documents from surgeons and other specialists who have also examined the subject. In this study, the central pathologists made assessments in isolation; they only received (if available): biopsy date, biopsy type and details if "other" was selected, anatomic location and details if "other" was selected, number of blocks provided, number of slides provided, and stain types. Without supporting information, the central pathologists may inadvertently create a discrepancy between central and local diagnoses due to lack of context. For example, certain skin lesions can present as multiple disease processes; a lesion can have features of both basal cell and squamous cell carcinoma. It is up to the pathologists' discretion to choose the most appropriate assessment with the information available. With additional clinical history and other information, the central pathologist could make a more confident assessment with less guesswork. Due to this limitation, the current read model does not allow for a "true" comparison between local and central diagnoses. It can be argued, however, that local pathologists have access to too much clinical information, which can skew their assessments. Instead of looking at the sample independently, the local pathologists are primed to look for clinically suspected or suggested pathology indicated on supporting documentation.

It also cannot be assumed that local diagnosis was the result of a single pathologist; it is possible that multiple pathologists collaborated to reach an assessment. If the local diagnosis was agreed upon by more than one pathologist, then, based on the

social and probability theories through which Speight supports adjudication, the site diagnosis may also have a higher probability of being correct than if the diagnosis was made by a single pathologist. If this were the case then it would be hard to determine whether the site or central assessment was more likely to be correct.

Future studies should ensure that each sub-group has sufficient samples to yield more exact results with smaller confidence intervals. All efforts should be made to obtain the same slides that were made to make the local diagnosis for central review. This will require adequate training, close monitoring, and cooperation by sites and the sponsor and may not be truly feasible. The workflow and read paradigm should be updated to permit presentation of select clinical information to the central pathologists. Although local pathologists may have too much supporting documentation, controlling what is available to local pathologists is not feasible. To maintain independence, the final diagnosis and clinical impressions should not be provided to the central pathologist, but medical history and macroscopic descriptions would be sufficient. The medical history provides background information on why the biopsy may have been obtained as well as any pre-existing conditions that may impact the current diagnosis. The macroscopic description details the shape, size, color, texture, and other defining characteristics of the specimen, which can be helpful in determining underlying pathology.

The current analyses could be performed excluding all cases with a behavior code of 1 (uncertain whether benign or malignant - borderline malignancy, low malignant potential, or uncertain malignant potential) in order to more clearly categorize all samples. Additional analyses should be performed, including comparing all independent

reads for each case, not just the authoritative read, to the local lab result and to each other. These analyses would provide more insight into whether there is consistency among reviewers in how they assess cases/agreement of their assessments and the local results.

CONCLUSION

If it is the case that there are clinically significant discrepancies between local and central diagnoses and that, based on Speight and Surowiecki's theories, central adjudication yields more accurate diagnoses than a local pathologist, then it should be accepted that adjudication ought to be more widely used in clinical trials to assess histopathology-related safety outcomes and endpoints. Based on this study's results, risk of inaccurate representation of high level (neoplasm versus non-neoplasm and benign versus malignant) safety and risk for a compound under investigation is low regardless of whether local or central diagnoses are considered. Safety and risk profiles including information about unique pathologies were more variable, and using adjudicated data should be considered. Despite inexact matches, many ICD-O codes can represent similar diseases processes with similar treatments. Understanding whether the lack of agreement stems from small inconsequential coding differences or misdiagnosis is important and can only be determined with confidence by a pathologist, not statistics. Due to small sample sizes, it cannot be confidently stated whether adjudication would have equal benefit among all pathology sub-specialty groups. There is little discord between the local and central pathologists regarding whether malignancies exist among samples. There are, however, significant discrepancies regarding specific morphology ICDO- 3

codes and their associated treatments. Because there is a significant difference between local and central pathologists in assigning diagnoses, adjudication should be used when providing a safety profile for a compound because it is more specific and more accurate.

APPENDIX

Mathematical model for 3 step adjudication process as determined by Speight, et al.

APPENDIX

Supplementary Materials: Appendix 1—Mathematical Model

The purpose of this section is to provide the mathematic basis for the high-quality gold standard whereby a series of steps can be used to improve the overall level of agreement between pathologists. Because the true diagnosis is unknown, it is not possible to determine the true rates of correct diagnosis. However, if a few simplifying assumptions are made, probabilities of correct diagnosis can be calculated. These assumptions include the following: (1) The 2 reviewers and adjudicator all have equal probability of misdiagnosis and this probability is independent of the other reviewer and/or adjudicator; (2) the probability of misdiagnosis is independent of the individual slide (i.e., each slide has an equal probability of misdiagnosis); (3) when the 2 reviewers disagree on a slide's diagnosis, 1 of the 2 reviewers (which one is unknown) is assumed to be correct in diagnosis and the other reviewer is assumed to have an incorrect diagnosis.

PART A: INITIAL ASSUMPTIONS

Let Pw = probability of a reviewer or adjudicator misdiagnosing a particular slide
Pc = probability of a reviewer or adjudicator correctly diagnosing a particular slide

As such,
Pw(Reviewer A) = probability "Reviewer A" misdiagnoses a particular slide
Pc(Reviewer A) = probability "Reviewer A" correctly diagnoses a particular slide
Pw(Slide 1) = probability "Slide 1" is misdiagnosed by a particular reviewer or adjudicator

(1) ***Assumption 1:*** The 2 reviewers and the adjudicator all have an equal probability of misdiagnosis; and each reviewer's (or adjudicator's) probability of misdiagnosing a particular slide is independent of the other reviewer (and/or adjudicator).
Pw(Reviewer A) = Pw(Reviewer B) = Pw(Adjudicator)
Pc(Reviewer A) = Pc(Reviewer B) = Pc(Adjudicator)

(2) ***Assumption 2:*** Assume that this probability of misdiagnosis is independent of a particular slide—that is, the probability of misdiagnosis is the same for every slide.
Pw(Slide 1) = Pw(Slide 2) = Pw(Slide 3)
Pc(Slide 1) = Pc(Slide 2) = Pc(Slide 3)

(3) ***Conclusion 1:*** The diagnosis of a particular slide by a particular reviewer (or adjudicator) has 2 and only 2 mutually exclusive outcomes: Either (i) correct diagnosis or (ii) wrong diagnosis. Thus, the probabilities add up to 1 (i.e., 100%).
$1 = Pc + Pw$
$Pc = 1 - Pw$

PART B: ESTIMATION OF THE PROBABILITY OF MISDIAGNOSIS

(4) (a) Probability of 2 reviewers agreeing on diagnosis: 591/846 = 0.699
(b) Probability of 2 reviewers disagreeing on diagnosis: 255/846 = 0.301

(5) ***Assumption 3:*** When 2 reviewers disagree on a diagnosis, one of the reviewers has the correct diagnosis and the other reviewer has the wrong diagnosis.
The probability of reviewers disagreeing is 0.301 and equals:
(Pw[Reviewer A] • Pc[Reviewer B]) + (Pc[Reviewer A] • Pw[Reviewer B]) = 0.301
Because: Pw(Reviewer A) = Pw(Reviewer B) = Pw; and Pc(Reviewer A) = Pc(Reviewer B) = Pc,
$(Pw \bullet Pc) + (Pw \bullet Pc) = 0.301$
$2 \bullet Pw \bullet Pc = 0.301$
$Pw \bullet Pc = 0.301/2 = 0.151$
Thus, the probability of Reviewer A being wrong and Reviewer B being correct:
$Pw \bullet Pc = Pw \bullet (1 - Pw) = 0.151$
$-Pw^2 + Pw = 0.151$
$-Pw^2 + Pw - 0.151 = 0$
Giving the standard form of the quadratic equation:
$Pw^2 - Pw + 0.151 = 0$
Using the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$Pw = \frac{1 \pm \sqrt{1 - 4 \cdot 0.151}}{2} = (0.185,\ 0.815)$$

If we select from the 2 possible solutions: Pw = 0.185 then Pc = (1 − Pw) = 0.815 can be shown to be correct by substituting back into original equation and calculating the total number of reviewers in agreement:
(Pc • Pc) + (Pw • Pw) = 0.698, which equals 0.699 (given slight rounding error observed in data, Table IV).
Therefore, based on the assumptions listed earlier, the probability of a reviewer or adjudicator misdiagnosing a slide is:
Pw = 0.185, or 18.5%
Likewise, the probability of a reviewer or adjudicator correctly diagnosing a slide is:
$Pc = 1 - Pw$
Or Pc = 0.815, 81.5%

PART C: OVERALL PROBABILITY OF CORRECT DIAGNOSIS GIVEN 2 REVIEWERS AND AN ADJUDICATOR TO SETTLE DIFFERENCES

The 6 probability scenarios in Table II are mutually exclusive and thus their probabilities add to 1 (within rounding error). The Pc is the probability of an individual reviewer or adjudicator for a given slide

making the correct diagnosis; Pw is the probability of a reviewer or adjudicator making a misdiagnosis for a particular slide. Pc was estimated at 0.815 and Pw was estimated at 0.185.

LIST OF JOURNAL ABBREVIATIONS

Am Heart J…….....American Heart Journal

Am J Health Econ……..……American Journal of Health Economics

Am J Orthod Dentofacial Orthop…….…..American Journal of Orthodontics and Dentofacial Orthopedics

Arch Intern Med…….…….....Archives of Internal Medicine

Biochem Med (Zagreb)……...Biochemia Medica

Clin Trials……..……. Clinical Trials (London, England)

Contemp Clin Trials……Contemporary Clinical Trials

Eur Heart J…….……. European Heart Journal

J Clin Epidemiol……...……..Journal of Clinical Epidemoiology

J Pain……..….The Journal of Pain: Official Journal of the American Pain Society

JACC…….……...……. Journal of the American College Cardiology

JAMA……Journal of the American Medical Association

Lancet……...The Lancet

Open Access J Clin Trials……Open Access Journal of Clinical Trials

Oral Sug Oral Med Oral Pathol Oral Radiol……Oral Surgery, Oral Medicine, Oral Pathology, and Oral Radiology

Pharm Stat……..…Pharmaceutical Statistics

Ther Innov Regul Sci……....Therapeutic Innovation and Regulatory Science

Thomb Haemost…….…Thrombosis and Haemostatsis

REFERENCES

[1] Hicks KA, Tcheng JE, et al. 2014 ACC/AHA Key Data Elements and Definitions for Cardiovascular Endpoint Events in Clinical Trials. *JACC*. 2015;66(4):404-469. http://dx.doi.org/10.1016/j.jacc.2014.12.018.

[2] Informa, Amplion, and the Biotechnology Innovation Organization (BIO). Clinical Development Success Rates 2006-2015. https://www.bio.org/sites/default/files/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf. Published June, 2016. Accessed December 11, 2016.

[3] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Am J Health Econ.* 2016;47:20-33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

[4] Herper M. The Truly Staggering Cost of Inventing New Drugs. Forbes. https://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/#426c42944a94. Published February 10, 2012. Accessed October 10, 2016.

[5] Granger CB, Vogel V, Cummings SR, et al. Do we need to adjudicate major clinical events? *Clin Trials.* 2008;5(1):56-60. DOI: 10.1177/1740774507087972.

[6] Dechartres A, Bouton I, Roy C, Ravaud P. Inadequate planning and reporting of adjudication committees in clinical trials: Recommendation proposal. *J Clin Epidemiol.* 2009;62(7):695-702. https://doi.org/10.1016/j.jclinepi.2008.09.011.

[7] Krumholz-Bahner S, Garibbo M, Getz KA, Widler BE. An overview and Analysis Regarding the Use of Adjudication Methods in EU and US Drug Approvals. *Ther Innov Regul Sci.* 2015;49(6):831-839. DOI: 10.1177/2168479015580382.

[8] U.S. Department of Health and Human Services. Guidance for industry developing medical imaging drug and biological products part 3: design, analysis, and interpretation of clinical studies. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071604.pdf. Published June 2004. Accessed December 23, 2016.

[9] Petersen JL, Haque G, Hellkamp AS, et al. Comparing classifications of death in the Mode Selection Trial: Agreement and disagreement among site investigators and a clinical events committee. *Contemp Clin Trials.* 2006;27(3):260-268. https://doi.org/10.1016/j.cct.2006.02.002.

[10] Walovitch R, Yao B, Chokron P, Le H, Bubley G. Subjective endpoints in clinical trials: the case for blinded independent review. *Open Access J Clin Trials.* 2013;5:111-117. http://dx.doi.org/10.2147/OAJCT.S50283.

[11] Kradjian S, Gutheil J, Baratelle AM, Einstein SG, Kaslow DC. Development of a charter for an endpoint assessment and adjudication committee. *Ther Innov Regul Sci.* 2005;39(1):53-61. DOI: 10.1177/009286150503900107

[12] U.S. Department of Health and Human Services. Guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics. https://www.fda.gov/downloads/Drugs/Guidances/ucm071590.pdf. Published May 2007. Accessed December 25, 2016.

[13] Pandis N. Sources of bias in clinical trials. *Am J Orthod Dentofacial Orthop.* 2011;140(4):595-596. https://doi-org.ezproxy.bu.edu/10.1016/j.ajodo.2011.06.013.

[14] Cochrane Bias Methods Group (BMG). Assessing Risk of Bias in Included Studies. http://methods.cochrane.org/bias/assessing-risk-bias-included-studies. Accessed October 12, 2016.

[15] Cox KR. *Planning Clinical Experiments.* Springfield, IL: Charles C Thomas; 1968.

[16] Wilson EB. *An Introduction to Scientific Research.* New York: McGraw-Hill; 1952.

[17] Naslund U, Grip L, Fischer-Hansen J, Gundersen T, Lehto S, Wallentin L. Impact of an end-point committee in a large multicentre, randomized, placebo-controlled clinical trial: Results with and without the end-point committee's final decision on end-points. *Eur Heart J.* 1999;20(10):771-777. https://doi-org.ezproxy.bu.edu/10.1053/euhj.1998.1351.

[18] Hong S, Schmitt N, Stone A, Denne J. Attenuation of treatment effect due to measurement variability in assessment of progression-free survival. *Pharm Stat.* 2012;11:394-402. http://dx.doi.org/10.1002/pst.1524.

[19] Mahaffey KW, Roe MT, Dyke CK, et al. Misreporting of myocardial infarction end points: Results of adjudication by a central clinical events committee in the PARAGON-B trial. *Am Heart J.* 2002;143(2):242-248. https://doi.org/10.1067/mhj.2002.120145.

[20] Serebruany V, Atar D. Viewpoint: Central adjudication in myocardial infarction in outcome-driven clinical trials – Common patters in TRITON, RECORD, and PLATO? *Thromb Haemost.* 2012;108(3):412-414. DOI: 10.1160/TH12-04-0251.

[21] U.S. Department of Health and Human Services. Review of ongoing investigations. 21 CFR 312.56.

https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.56. Revised April 1, 2016. Accessed January 13, 2017.

[22] U.S. Department of Health and Human Services. IND safety reporting. 21 CFR 312.32. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.32. Revised April 1, 2016. Accessed January 13, 2017.

[23] U.S. Department of Health and Human Services. Other post marketing reports. 21 CFR 314.81. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.81. Revised April 1, 2016. Accessed January 13, 2017.

[24] U.S. Department of Health and Human Services. Food and Drug Modernization Act of 1997. https://www.fda.gov/RegulatoryInformation/Legislation/SignificantAmendmentstotheFDCAct/FDAMA/FullTextofFDAMAlaw/default.htm#SEC. 130. Approved November 21, 1997. Accessed March 30, 2017.

[25] U.S. Department of Health and Human Services. Safety Assessment for IND Safety Reporting Guidance for Industry. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM477584.pdf. Published December, 2015. Accessed February 1, 2017.

[26] National Public Radio. Timeline: The Rise and Fall of Vioxx. http://www.npr.org/templates/story/story.php?storyId=5470430. Published November 10, 2007. Accessed February 2, 2017.

[27] Graham DJ, Campen D, Hui R. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet.* 2005;365(9458):475-481. https://doi.org/10.1016/S0140-6736(05)17864-7.

[28] U.S. Department of Health and Human Services. Drug Safety Oversight Board. https://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm082129.htm. Updated May 5, 2016. Accessed March 1, 2017.

[29] Food and Drug Administration in collaboration with the National Cancer Institute. Meeting to discuss: Methodological Considerations in Evaluation of Cancer as an Adverse Outcome Associated with Use of Non-Oncological Drugs and Biological Products in the Post-Approval Setting. Silver Spring, Maryland. September 10, 2014.

[30] Howard, J. Minimum Latency & Types or Categories of Cancer. World Trade Center Health Program. https://www.cdc.gov/wtc/pdfs/WTCHP-Minimum-Cancer-Latency-PP-01062015.pdf. Published October 17, 2012. Updated January 6, 2015. Accessed March 1, 2017.

[31] U.S Department of Health and Human Services. Step 3: Clinical Research. https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm. Updated May 25, 2017. Accessed March 1, 2017.

[32] OCCUPYTHEORY. Advantages of Disadvantages of Longitudinal Studies. http://occupytheory.org/advantages-of-disadvantages-of-longitudinal-studies/. Published February 18, 2015. Accessed March 13, 2017.

[33] Green Garage: The ECO Friendly Blog. 12 Big Advantages of Disadvantages of Longitudinal Studies. https://greengarageblog.org/12-big-advantages-of-disadvantages-of-longitudinal-studies. Accessed March 1, 2017.

[34] Center for Disease Control and Prevention. Chronic Disease Prevention and Health Promotion. https://www.cdc.gov/chronicdisease/. Accessed March 13, 2017.

[35] Roemer, M. Statistical Brief #468: Expenditures for the Top Five Therapeutic Classes of Outpatient Prescription Drugs, Adults Age 18 and Older, U.S. Civilian Noninstitutionalized Population, 2012. Agency for Healthcare Research and Quality. https://web.archive.org/web/20150906064536/http://meps.ahrq.gov/mepsweb/data_files/publications/st468/stat468.shtml. Published February, 2015. Accessed March 13, 2017.

[36] Drugs.com. Drug Classes. https://www.drugs.com/drug-classes.html. Accessed March 13, 2017.

[37] National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation. 2014 National Diabetes Statistics Report. Diabetes Home. https://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html. Published 2014. Updated May 15, 2015. Accessed March 13, 2017.

[38] Johannes CB, Le TK, Zhou X, Johnston JA, Dworkin RH. The prevalence of chronic pain in United States adults: results of internet based survey. *J Pain.* 2010;11(11):1230-1239. doi: 10.1016/j.jpain.2010.07.002.

[39] Centers for Disease Control and Prevention. Prevalence of Coronary Heart Disease – United States, 2006-2010. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6040a1.htm. Published October 14, 2011. Updated October 14, 2011. Accessed March 13, 2017.

[40] Anxiety and Depression Association of America. Facts & Statistics. https://www.adaa.org/about-adaa/press-room/facts-statistics. Updated August, 2016. Accessed March 13, 2017.

[41] Centers for Disease Control and Prevention. Chronic Obstructive Pulmonary Disease Among Adults – United States, 2011.

https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6146a2.htm. Published November 23, 2012. Updated November 23, 2012. Accessed March 13, 2017.

[42] Rosembaum, EH. Comorbid Disease (Chronic Illnesses). Cancer Supportive Survivorship Care. http://www.cancersupportivecare.com/Survivorship/comorbidity.html. Published August 30, 2009. Updated September 2, 2009. Accessed March 13, 2017.

[43] Chlebowski RT, Hendrix SL, Langer RD, et al. Influence of Estrogen Plus Progestin on Breast Cancer and Mammography in Healthy Postmenopausal Women: The Women's Health Initiative Randomized Trial. *JAMA.* 2003;289(24):3243-3253. doi:10.1001/jama.289.24.3243.

[44] Chlebowski RT, Anderson G, Pettinger M, et al. Estrogen Plus Progestin and Breast Cancer Detection by Means of Mammography and Breast Biopsy. *Arch Intern Med.* 2008;168(4):370-377. DOI: 10.1001/archinternmed.2007.123

[45] Manson JE, Chlebowski RT, Stefanick ML, et al. The Women's Health Initiative Hormone Therapy Trials: Update and Overview of Health Outcomes During the Intervention and Post-Stopping Phases. *JAMA.* 2013;310(13):1353-1368. doi:10.1001/jama.2013.278040.

[46] Merriam-Webster Dictionary. Medical definition of pathology. https://www.merriam-webster.com/dictionary/pathology#medicalDictionary. Accessed April 15, 2017.

[47] Chandler I, Houlston RS. Interobserver agreement in grading of colorectal cancers – findings from a nationwide web-based survey of histopathologists. *Histopathology.* 2008;52:494-499. DOI: 10.1111/j.1365-2559.2008.02976.x.

[48] Speight PM, Abram TJ, Floriano PN, et al. Interobserver agreement in dysplasia grading: toward an enhanced gold standard for clinical pathology trials. *Oral Surg Oral Med Oral Pathol Oral Radiol.* 2015;120:474-482. http://dx.doi.org/10.1016/j.oooo.2015.05.023.

[49] McHugh, ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012;22(3):276-282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/. Published October 15, 2012. Accessed January 7, 2017.

[50] Stang A, Trocchi P, Ruschke K, et al. Factors influencing the agreement on histopathological assessments of breast biopsies among pathologists. *Histopathology.* 2011;59(5):939-949. DOI: 10.1111/j.1365-2559.2011.04032.x

[51] Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* New York: Anchor Books; 2004.

[52] World Health Organization. International Classification of Diseases for Oncology, 3rd Edition (ICD-O 3). http://www.who.int/classifications/icd/adaptations/oncology/en/. Published 2000. Updated October 5, 2015. Accessed April 10, 2017.

[53] Fritz A, Percy C, Jack A, et al. *International Classification of Diseases for Oncology, Third Edition, First Revision.* Geneva: World Health Organization, 2013.

[54] World Health Organization. Classifications. http://www.who.int/classifications/icd/adaptations/oncology/en/. Updated February 3, 2017. Accessed April 19, 2017.

[55] Anderson J. An Introduction to Routine and Special Staining. Leica Biosystems. http://www.leicabiosystems.com/pathologyleaders/an-introduction-to-routine-and-special-staining/. Accessed May 3, 2017.

[56] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Medical Dictionary for Regulatory Activities. http://www.meddra.org/. Accessed February 10, 2017.

[57] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Medical Dictionary for Regulatory Activities: MedDRA Hierarchy. http://www.meddra.org/how-to-use/basics/hierarchy. Accessed February 10, 2017.

[58] American Cancer Society. Basal and Squamous Cell Skin Cancer. https://www.cancer.org/cancer/basal-and-squamous-cell-skin-cancer.html. Accessed June 1, 2017.

[59] American Cancer Society. About Melanoma Skin Cancer. https://www.cancer.org/cancer/melanoma-skin-cancer/about.html. Accessed June 1, 2017.

[60] Cancer Research UK. Cancer survival for common cancers. http://www.cancerresearchuk.org/health-professional/cancer-statistics/survival/common-cancers-compared#heading-Zero. Accessed June 2, 2017.

[61] Andale. Cohen's Kappa Statistic. http://www.statisticshowto.com/cohens-kappa-statistic/. Statistics How To. Published December 8, 2014. Accessed June 3, 2017.

[62] World Health Organization. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) Version for 2010. http://apps.who.int/classifications/icd10/browse/2010/en#/D00-D09. Accessed June 1, 2017.

[63] Simundic AM, Nikolac N, Ivankovic N, Dragica FerenecRuzic D, Magdic B, Kvaternik M, Topic E. Comparison of visual vs. automated detection of lipemic, icteric and hemolyzed specimens: can we rely on a human eye? *Clin Chem Lab Med*. 2009;47:1361–1365. DOI: 10.1515/CCLM.2009.306.

CURRICULUM VITAE

Alison Michele Occhiuti
Born: 1987
Somerville, MA 02144
alison.occhiuti@gmail.com - 781-572-7781

SUMMARY:

Dedicated project manager with seven years of clinical trials experience. Currently managing Endpoint Adjudication program covering five therapeutic areas for a multi-protocol program for an international pharmaceutical company. Previous experience includes medical imaging project management and cross functional management experience with global teams. Broad background in large global Phase II and Phase III trials including expedited breakthrough therapy designation studies. Has participated in sponsor and FDA audits.

Senior Project Manager
WorldCare Clinical - Boston, MA - 2015-11 - Present

- Serve as main contact for Sponsor
- Manage vendors including CROs and supportive services (shipping vendors, etc)
- Establish, write, review, and maintain study documentation
- Manage database procedures, including querying and source data verification
- Participate in site monitoring including timely submission of clinical documents to WorldCare and CRF completion
- Ensure trial compliance with CFR and GCP
- Develop training programs and materials for internal and external project teams
- Ensure adherence to regulatory standards of GCP, GDP, SOPs, and other regulations for all team members and processes including reviewer training, collection and handling of data, data management deliverables, etc
- Provide and create project metrics, reports, and systems to monitor workflow, workload, and progress towards deliverables
- Track budget and finance
- Organize, delegate, and oversee project specific tasks for cross functional internal and external teams
- Troubleshoot and resolve known issues/risks
- Create mitigation and contingency plans to handle potential issues/risk

- Responsible for process streamlining and productivity cross functionally, including identifying resource requirements
- Provide support for budgets and proposals
- Work with senior management on process improvement and reaching company goals
- Operational presenter at medical imaging bid defense meetings

Project Manager
PAREXEL - Billerica, MA - 2013-05 - 2015-11

- Key operational contact between client and Medical Imaging
- Provide management and administration to project teams to keep project on time and within budget
- Resource planning and management
- Revenue recognition and forecasting
- Facilitate collaboration between internal and external stakeholders including application development and validation
- Conduct internal and external team meetings and trainings
- Timeline management
- Attend regular meetings with Associate Director of Imaging Operations to ensure transparency and oversight for all projects
- Maintain 21 CFR Part 11 compliance across project through documentation and guidance
- Contract/Exhibit management
- Member of musculoskeletal strategy team
- Operational presenter at medical imaging bid defense meetings

Imaging Operations Lead in Transition (IOLT)
PAREXEL - Billerica, MA - 2012-05 - 2013-05

- Perform all actions required of an Imaging Operations Lead (see below)
- Part of Associate Project Manager Transition Program
- Attend regular meetings with Associate Directors of Imaging Operations and Senior Manager of Global Operations regarding project management development at PAREXEL
- Work one on one with mentor (current Project Manager) on project management tasks such as document development, interdepartmental interactions, project finances, and software development

Imaging Operations Lead
PAREXEL - Billerica, MA - 2012-02 - 2012-05

- Oversee project team in site qualification steps, processing of imaging examinations, and query management

- Assist project manager in the development and controlling of timelines, deliverables, and project resource requirements
- Develop and edit study related procedure manuals
- Assist project manager in setting up the study specific applications
- Assist project managers with managing the data reconciliation process
- Act as the main point of contact for India Team Leads
- Ensure investigator sites are trained to follow image acquisition technique in accordance with study protocols
- Interact with clients
- Interact with academic centers, independent reviewers / subcontractors as needed
- Point of escalation for investigator sites, clients and / or academic centers during absence of PM / study team.
- Support project manager during kick-off and other client meetings
- Assist Project Manager with delegation of study activities to project team members

Imaging Research Associate
PAREXEL - Billerica, MA - 2011-05 - 2012-02

- Performed quality checks on imaging before images go to central review
- Includes CT, MRI, PET, Bone Scan, X-ray, ECHO, MUGA
  Performed preliminary measurements for MRI/CT Volumetric studies and ECHO and MUGA
- Has experience with SPECT and angiograms as well
- Authored various internal documents including Core Imaging Lab Manuals/Site Operations Manuals for various projects
- Attended client/vendor teleconferences as PAREXEL team representative
- Assigned cases to external radiologists for independent review
- Trained new hires and other team members regarding project specific procedures as well as more general tasks
- Created and shared reports with team members, other vendors, and clients
- Communicated with sites via email, phone, and fax regarding site issues and study progress
- Participant in New Hire Improvement Initiative

Imaging Assistant
PAREXEL - Billerica, MA - 2010-09 - 2011-05

- Tracked and organized incoming imaging exams for multiple projects
- Performed initial quality check on imaging before processing
- Communicated with sites via email, phone, and fax regarding scanning parameters, timely submission, and issues and/or questions regarding imaging exams and site qualification

- Created and shared reports about project progress internally and also with other external vendors and clients
- Lead trainings for team members regarding UPS
- Mentored new hires

Technology Experience
MS Office Suite: Excel, PowerPoint, Word, OneNote, Visio, Project; Outlook; Adobe Acrobat Pro; Sharepoint; Pharmaready; mySignature Book; limited exposure to Inform; RStudio

EDUCATION

Bachelor of Arts, Spanish
Northwestern University, 2010

Master of Science, Clinical Investigation
Boston University School of Medicine, 2017

THERAPEUTIC AREA EXPERTISE:

Indication	Phase	# Patients	# Sites	Countries
Various Autoimmune Indications	I-IV	100->5000 (variable per protocol)	10->200 (variable per protocol)	Global
Breast Cancer	I	117	5	United States, United Kingdom
Metastatic Carcinoma/ Melanoma/Non-Small Cell Lung Carcinoma	I	700+	37	United States, Canada, France, Australia, United Kingdom, Germany
Myelofibrosis/Post essential thrombocythemia mylofibrosis/ Post polycythemia	II	150	60	United States

vera myelofibrosis				
Metastatic Pancreatic Adenocarcinoma	II	244	112	United States, Russia, Germany, Poland
Colorectal Adenocarcinoma	II	265	154	United States, France, Italy, Spain, Germany, Poland
Breast Cancer	II	255	120	United States, Poland, Russia, Australia, France, Brazil Argentina, Czech Republic, Denmark, Israel, Norway, Slovakia, Hungary, Canada, Ukraine, Belgium, Sweden, Spain, Netherlands, Finland
Pancreatic Cancer	II	82	22	China
Hand Osteoarthritis	IIa	120	50	United States, Belgium, France, Netherlands,
Prostate Cancer	III	1800	207	Austria, Belgium, Denmark, Finland, France, Germany, Israel, Lithuania, Netherlands, Poland, Slovakia, Spain, Sweden, United Kingdom
Breast Cancer	III	712	140	United States, Spain, Belgium, Poland, Germany, Russia, France, Italy, United Kingdom, Sweden, Denmark, Thailand, Hungary, Greece
Non-Small Cell Lung Cancer	III	850	225	United States, Austria, Netherlands, Slovenia, Spain, France, Poland, Canada, Russia, Belgium, Germany, United Kingdom, Czech Republic, Romania, Bosnia, Serbia and Montenegro, Croatia, Hungary, Ukraine, Italy

Breast Cancer	III	238	422	United States, Canada, Australia, Belgium, France, Germany, Italy, Netherlands, Portugal, Romania, Russia, Taiwan, Ukraine, South Korea, Mexico, Switzerland, Ireland, Japan, United Kingdom, Turkey
Metastatic Soft Tissue Sarcoma	III	65	500	United States, Spain, Hungary, Belgium, France, Russia, Italy, Germany, Poland, Denmark, Austria, Israel, Australia, Canada
Venous Thromboembolism	III	6000	518	United States, Spain, Croatia, Czech Republic, Finland, Canada, Austria, Belgium, Bulgaria, France, Serbia and Montenegro, Slovakia, Germany, Denmark, Singapore, Hungary, Israel, Italy, Lithuania, Poland, Russia, South Africa, United Kingdom, Australia, India, Estonia, Latvia, Argentina, Brazil, Chile, Peru, Romania, Ukraine
Pediatric Type 2 Diabetes Mellitus l	III	107	172	United States, Belgium, Croatia, Denmark, Germany, Greece, Hungary, Macedonia, Serbia and Montenegro, Spain, Sweden, United Kingdom, India, Israel, Mexico, Russia, Turkey, Canada, Norway, Romania,

Pediatric Type 2 Diabetes Mellitus	III	165	360	United States, Russia, Italy, Thailand, Chile, Israel, Dominican Republic, Costa Rica, Guatemala, Colombia, Mexico, Romania, Lithuania, New Zealand, Latvia, Bulgaria, Malaysia, Spain, Argentina, Hungary, Serbia and Montenegro, Austria, Portugal, Slovakia, Poland, Denmark, Philippines, Australia, Brazil, Germany, South Africa, Sweden, Canada, France
Hip and Knee Osteoarthritis	III	375	80	United States
Nail Psoriasis	III	6	5	United States, Belgium, Germany