Document Title:	The Interpretation of Patterned Injuries in Medicolegal Death Investigation
Author(s):	William R. Oliver, M.D.
Document Number:	250570
Date Received:	February 2017
Award Number:	2012-DN-BX-K053

Report Title: Draft Final Report

Award number: 2012-DN-BX-K053

Author: William R. Oliver, MD

Please note that much of this report is excerpted verbatim from articles published (1,2), accepted for publication (4), promulgated as a white paper to the Organization of Scientific Area Committees(3), or submitted for publication(5).

Abstract

This research was originally intended to evaluate the effect of image processing and image quality on interpretability of images of patterned injury of the skin by forensic pathologists. It was designed to consist of three surveys. The first was to be a collection of “classic” images that almost all pathologists were expected to diagnose with high consensus to act as a baseline. The second survey was to consist of degraded images with lesser resolution, poorer composition, etc. to see how the degradation affected diagnostic consensus. The third survey was to present images treated with various enhancement techniques (primarily contrast enhancement) and see if any benefit was associated with that enhancement.

The results of the first survey were surprising. Instead of a high consensus among pathologists, the average consensus for all questions was only 72% (median 74%), with a range from 25% to 100% concurrence. This clearly violated our assumption that there would be uniformly high concordance involving “classic” patterned injuries. The results of this survey were presented at the 2015 AAFS meeting and the paper has been acccepted for publication in the Journal of Forensic Sciences.

We discussed this unexpected finding with our program manager, and decided that in light of the recent interest in human factors in forensic science, it might be interesting to modify the remaining surveys to look at why these discrepancies exist and what the effect of history and context might have. The second survey was then changed to ask the original respondents why they did not answer consensus answers, and the third survey tested the effect of providing history and context.

The second survey has been completed. There are a number of statistically significant differences among respondents regarding why they did not provide a consensus answer, most prominently involving age and experience. Of greater importance, the primary reason that people did not provide consensus answers was the perception of some sort of ambiguity – due to naming issues, lack of specificity, etc., rather than an overt disagreement about the diagnosis. Many of the respondents specifically noted in comments that it was impossible to make a diagnosis to any certainty, even in the presence of a “classic” image, in the absence of history. The results of this survey was presented at the 2016 AAFS meeting, and the paper has been accepted pending revision.

The third survey was deployed in December, 2016. The responses have been analyzed and the results demonstrate the importance of context and history to forensic pathologic diagnosis, at least with respect to photographs of patterned injury of the skin. When provided with history, consensus rose to approximately 98% per question (median value) from a median of 77% for the matching subset of the first survey.

Table of Contents

Executive Summary

Photography has become a very important factor in forensic autopsy performance. A few years ago, when film photography was more common, limited photographs were taken at autopsy. The marginal cost of photographs was high and the quality of a photograph could not be judged until the film was developed and the photograph printed. This led to the widespread practice of limiting photography to documentation of major findings. With the advent of high-quality, inexpensive digital photography, the marginal cost of photographs has become negligible and it is possible to review the results in the autopsy suite. This has resulted in many offices taking many more photographs, and review of photographs has become a more important component of reviewing cases during quality control or as a consultant performing case review.

Another consequence of digital imaging is the easy availability of image processing software, both within cameras and as ancillary packages. While image processing has been used in forensic pathology, the value of image processing methods such as contrast enhancement has been more formally investigated in other specialities such as radiology. For instance, a common method of contrast enhancement called “contrast limited adaptive histogram equalization” was developed in the late 1980s for radiology applications. Investigation into its usefulness determined that it often increased the detection of lesions, but did not aid in their identification.

Three surveys were planned to examine the effect of image quality on evaluation of images of patterned injury of the skin. The first was to be a collection of “classic” images that almost all pathologists were expected to diagnose with high consensus to act as a baseline. The second survey was to consist of degraded images with lesser resolution, poorer composition, etc. to see how the degradation affected diagnostic consensus. The third survey was to present images treated with various enhancement techniques (primarily contrast enhancement) and see if any benefit was associated with that enhancement.

The first survey was composed of a collection of 68 images of patterned injuries of the skin with associated multiple choice questions. Respondents were shown the images three times. The first time they were asked to determine the general class of injury, e.g. sharp vs blunt vs penetrating, etc. This is called a “Tier 1” response in the remainder of the paper. The second time the respondents were asked to determine the kind of injury within the general class. For blunt trauma, for instance, it would be asked if the injury was primarily contusion, laceration, abrasion, etc. This is called a “Tier 2” answer. The third time the respondents were asked to identify the specific object or injury. For a contusion, for instance, the options might include pool cue, brick, or car bumper. This is called a “Tier 3” answer. For each answer, the respondents were asked to provide a numerical score for their subjective degree of confidence on a scale of 1 to 10. This resulted in a total of 408 image-based questions, 14 demographics questions, and 4 administrative screens.

In contrast to expectations, there was a relatively low rate of consensus (average of 74%), with a wide range of the degree of consensus in the individual questions.

After consultation with our Program Manager, the focus of the research changed from investigating image issues to looking at why forensic pathologists differed so greatly in their evaluations. A second survey was constructed and sent to the participants of the first survey to ask why they answered in the manner the did.

The second survey was made up of multilayered questions consisting of a bank of 68 sets of three subquestions. The respondent was only asked about a question if he or she did not provide the consensus response in the first survey (subquestion 1), indicated a low confidence of less than 8 out of 10 (subquestion 2) or indicated that image quality was an issue (subquestion 3). Subquestions 1 and 2 asked why they provided the answer they did, (e.g. “did not know the answer,” “poor image quality,” “naming convention,” “nonspecific lesion,” etc.). Subquestion 3 related to imaging issues such as resolution, dynamic range, etc.

Of the 363 respondents to the original survey, 153 responded to the second survey, and 102 completed it. While the full statistical analysis of the results is still ongoing, the intial results have been enligtening and have been accepted for presentation at the 2016 AAFS meeting. The manuscript will be finished by then. Note that the following p values are still pending review by our statistician.

There were significant differences in why respondents did not answer consensus answers by Tier. Respondents were more likely to invoke nomenclature issues or the presence of multiple injuries in Tier 2, while they were more likely to state that the injury was not specific to the consensus answer in Tier 3.

Many of the specific significant findings are described in the body of the report. Of particular interest from a human factors viewpoint was that relatively few of the discrepancies were due to a belief in a different diagnosis. The vast majority of the differences had to do with issues of ambiguity – nomenclature, the presence of multiple lesions (e.g. abrasions along with lacerations), or a hesitance to make a specific diagnosis in the absence of corroborating history. A number of respondents commented that they simply would not make a definitive diagnosis without a supporting history.

The third survey consisted of 23 images chosen to represent a range of degrees of consensus in the first survey. Similar (but different) examples of the lesions were used, and the options for answers were replicated as much as possible without being nonsensical (e.g. inappropriate gender issues, etc.). In addition to the image and options, a short history and relevant laboratory results were provided (e.g. toxicology). The third survey was closed Jan 18, 2016. Results provide a median consensus of 98% compared to 77% in the first survey. The answers of those who took the first survey and those who took the second survey are not significantly different (so far), suggesting a minimal training bias.

I. Introduction:

1. Statement of Problem

It has been proposed that the provision of historical data in forensic medicine constitutes a source of cognitive bias, and that physicians should be strictly limited in their access to scene and history data. This is based, in part, on the belief that forensic medicine is a form of pattern matching much like that involved with toomark and shoeprint matching. In contrast, there is a wealth of literature that supports the importance of history, and many cognitive studies that argue against that simplistic view of medical diagnosis. These studies investigate the importance of history in the most simple pattern-matching task in forensic medicine – the evaluation of photographs of patterned injuries of the skin.

2. Literature review

The use of history (and in forensic death investigation, scene information) is an integral part of medical diagnosis. Declining to use history is not an example of removing cognitive bias, but is instead simple malpractice. The literature regarding the importance of history in medical diagnosis is voluminous.

As one author noted about manner determination (4):

The inference of manner is much like the inference of cause of death. One creates the equivalent of a differential diagnosis, ranks and prunes the possibilities, and comes to a conclusion as to which is most likely. The difference is in the degree to which the determination relies on external information. There is often little about a bullet hole that tells one who created it; many wounds are equally consistent with homicide, suicide, or even accident. It is necessary to consider investigational data, scene data, and history.

Some people in the legal community have disparaged the use of these external data as evidence of "cognitive bias" in our determinations. This is no more true for the medical examiner than it is for any other physician who makes a diagnosis integrating history and circumstances with physical examination and laboratory tests. It is the practice of medicine. A physician evaluating a patient with a fever and malaise may try to discover all sorts of things other than what the immediate physical examination provides – travel to other countries where certain disease are endemic, what was recently eaten for dinner, what medications the patient may be on, whether or not other people in the home also have fevers, etc. In light of the recent Ebola outbreak, it is not mere "cognitive bias" to ask about recent travel to Africa when evaluating someone in the emergency room, even if that travel is not a physical finding. In fact, failing to do so would be a lapse in care. The same thing is true with manner determination. It relies on history and circumstances as well as the immediate physical examination.

The importance of history and circumstances in medical diagnosis is sometimes misunderstood by those who are not physicians. In one study of 630 medical cases, for instance, history was the most important component of diagnosis in 56% of cases whereas physical examination was most important in only 17% of cases and laboratory investigation was most important in 23% of cases. In another study of 80 cases, history was most important in 76% of the cases while physical examination was most

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

important in 18% of the cases. These findings have been replicated many times. To dismiss the importance of history and circumstances in medical diagnosis as “cognitive bias” represents a severe misunderstanding of inferential processes in medicine.

Lawson and Daniel, in their discussion of medical inference, provide an excellent description of the importance of history(5):

Additionally, because many diseases produce evolving symptoms and findings over time, obtaining a thorough and time-sequenced history of the patient’s evolving problems is of extreme importance. To obtain a thorough history one needs to keep in mind the documented tendency of physicians to interrupt the patient and cut the history short after the patient tells enough to suggest an initial hypothesis. By analogy, in order to correctly guess the title of a movie, one needs to view as much of it as possible, and replay it enough times until one “gets the picture.” Sometimes the first scene holds the key, and if this is missed, the rest of the plot may be misinterpreted. Hence one needs to roll the movie back to the very beginning. For example: A person is now 20 years old. His asthma started a few years ago. Could it have been caused by psittacosis? Could he be still harboring low-grade psittacosis? Question: Were there birds in the household at that time? Could it be relevant? Was it ever thought of? Could a round of doxycycline help his current asthmatic condition? Unless this part of the script is examined, which it almost never is because we assume someone properly thought of those questions at that time, which they may or may not have done, the true answer may be missed. Hence rolling the entire movie script, reviewing it repeatedly, generating multiple hypotheses, and asking probing questions about hidden parts of the script may be crucial to the thought processes of the expert diagnostician.

Much has been made of the use of confession information and other evidence of low empiric value in the determination of cause and manner of death. Evidence, at least in medical diagnosis, is not a toggle between 100% and zero. Evidence must be weighed and evaluated by the physician. When a patient comes to a doctor, complains of pain, and asks for drugs for relief, it makes a difference in the evaluation of that complaint if the patient has a long history of drug seeking behavior or if he or she is a stoic person who almost never complains of anything. Much of the information that a physician evaluates is not of 100% specificity and sensitivity, even when just considering anatomic findings. That is why it is of paramount importance that the pathologist weigh all available information.   But this is true in all integrative disciplines. David Schum, writing about the interpretation of intelligence data notes(6):

The problem is that most inferences involve processes or variables that are nonindependent in various ways, with genuinely interesting evidential subtleties. A causal assumption of complete independence among identified processes would in most cases invite inferential calamity. So we have no choice but to do our best at capturing what we believe are avenues of probabilistic dependence among processes of concern. To do so, we link nodes representing these processes by various patterns of arcs. I can think of no inference problem, outside the classroom, whose structure is either provided for us of immediately apparent. Constructing a network representation of an inference problem is a purely subjective judgmental task, one likely to result in a different structural pattern by each person who performs it.

The complexity of inference in medical diagnosis is well documented within the cognitive sciences field. This research demonstrates significant differences between medical diagnosis and laboratory investigation.

In spite of the clear importance of history in medical diagnosis, there are those who believe that the integration of history into medical diagnosis constitutes "cognitive bias" and physician access to information should be controlled by gatekeepers. Part of this response, at least in the realm of forensic science, seems to come from the belief that forensic pathology is a rather simple pattern recognition task rather than true medical practice.

Medical diagnosis is not an issue of trivial "pattern recognition," even though pattern recognition of various kinds can be a part of medical diagnosis. In fact, evaluations of medical diagnosis demonstrate that it involves multiple different cognitive strategies, depending on the problem and the context (7). Even in the context of the more classic "pattern recognition" case of cytology evaluation, the model of simple pattern recognition does not apply. As one author notes(8):

When experienced pathologists review slides from a histologic section or cytologic preparation, they do not evaluate the morphologic features in a purely objective, quantifiable manner, adding weights to different features based on prior studies, tallying the scores for the various clinically validated criteria, and finally rendering a numerical probability of likely diagnoses. Such a system,while perhaps desirable, is not reflective of the complex,nonlinear nature of most diseases...

In the context of telemedicine, another author notes (9):

Routine problems that can be solved by simple pattern recognition do not elicit sequential reasoning and, consequently, do not involve opportunistic planning, nor cognitive multitasking. These characteristics are neither present in non-naturalistic (laboratory) situations where the problem-solving environment is artificially contrived...

To view medical diagnosis as simple "pattern matching" or to perceive it as a simple laboratory test may be incorrect.

3. Hypothesis

The hypothesis of this research is that, like all other medical specialties, the integration of history is fundamental to the evaluation of patterned injuries of the skin. While this was not the original intent of this study, the revised prediction is that forensic pathologist performance on the evaluation of photographs of patterned injuries of the skin is exquisitely reliant on historical information. Even in the case of "classic" patterned injuries, there exists significant ambiguity in forensic pathologist evaluations in the absence of history, and physicians require, and rely on, history and contextual information to resolve these ambiguities.

II. Methods

First survey:

Structure

The first survey was composed of a collection of 68 images of patterned injuries of the skin with associated multiple choice questions. Respondents were shown the images three times. The first time they were asked to determine the general class of injury, e.g. sharp vs blunt vs penetrating, etc. This is called a “Tier 1” response. The second time the respondents were asked to determine the kind of injury within the general class. For blunt trauma, for instance, it would be asked if the injury was primarily contusion, laceration, abrasion, etc. This is called a “Tier 2” answer. The third time the respondents were asked to identify the specific object or injury. For a contusion, for instance, the options might include pool cue, brick, or car bumper. This is called a “Tier 3” answer. For each answer, the respondents were asked to provide a numerical score for their subjective degree of confidence on a scale of 1 to 10. This resulted in a total of 408 image-based questions, 14 demographics questions, and 4 administrative screens.

Responses to all questions were not required. The survey was not time-limited, and respondents could stop and return to the survey.   Results were collected from November 21, 2013 until Feb 21, 2014. The time taken for respondents to complete the survey from time to logon to completion varied from just under two hours to approximately 85 days, with most responses taking between two and three hours.

The survey was web-based, using Qualtrics survey software. No measurement or determination of viewing conditions was obtained; the study design was to replicate how consultation images would be viewed in an office, and it was assumed that those, or similar, conditions would be used for the survey.

Selection of images

An expert panel of 12 senior forensic pathologists from across the United States volunteered to help provide and select images. In addition, a request for “classic” patterned injuries with diagnosis was promulgated through forensic pathology mailinglists. From these, approximately 150 images were obtained, and 68 were judged to be of sufficient quality and uniqueness for use in the survey by consensus of the expert panel. The list of images is given in Table 1.

Because these images were largely taken from teaching sets and private collections, it was not possible to independently determine the “ground truth” of the diagnosis. Accordingly, this is a study of consensus of diagnosis rather than correctness.

Selection of participants

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Participants were recruited through broadcast email to the membership of the National Association of Medical Examiners.

Protection of human rights

The survey and the experimental protocol were reviewed by the Institutional Review Board of Brody School of Medicine at East Carolina University as well as by the National Institute of Justice for anonymization of the images and protection of participant's identities. Participant identification is retained through the series of surveys for the purpose of troubleshooting emails, logistics, and follow-up, but will be discarded and the responses anonymized upon completion of the project.

Statistical analysis

Statistical analysis was done using the “R” statistical package as well as Excel. Excel was used for initial viewing of data and early simple statistical triage. For the descriptive statistics in this manuscript, simple descriptive statistics were performed using the “describe,” “summary,” and “stat.desc” functions. The “summary” function is found in the “psych” library, and “stat.desc” is found in the “pastecs” library. Simple correlations for continuous data (e.g. average confidence or percent consensus) were performed using the “cor.test” function. Correlation of dichotomous results with continuous results was performed using the “biserial” function of the “psych” library.

Data cleaning

Upon review of the responses, it was noted that a number of participants chose the option “other” as a diagnosis and then inserted a diagnosis that was equivalent to one of the other options. In particular, a number of participants insisted on adding more specific diagnoses at Tier 1; rather than choosing “blunt trauma” a participant would choose “other” and write in “blunt trauma with a baseball bat.” These were recoded as “blunt trauma.” Other, similar answers were similarly recoded. One Tier 3 question was dropped because the wrong answer set was published in the survey, and an appropriate diagnosis was not offered as a choice. In the demographics section, some international degrees and certifications were recoded as the nearest similar certification for the United States system.

Second Survey

Structure

The second survey consisted of the 68 images of the first survey. Each question queried a database of answers to the first survey. If the respondent did not provide the consensus response to the question, he or she was asked why, with the options of: the consensus answer is incorrect, the answer is equivalent to the consensus answer, there were multiple lesions and the respondent answered about a different one, the lesion was not specific and the respondent considered other possibilities, the lesion might be a consensus answer but is an atypical presentation, the image quality was insufficient for diagnosis, the respondent was not familiar with the lesion and did not know what it was, the respondent knew the answer but was uncertain about naming conventions used in the survey, or other (write in).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

If the respondent indicated uncertainty of less than 8 out of 10, he or she was asked why, with the options of : poor image quality, not a characteristic example, lesion was not specific, uncertainty about naming convention, uncertain about what it was, set the value incorrectly by accident, and other (write in). This was presented as a matrix to allow a different response for each Tier.

If the respondent answered poor image quality to either of the above questions, a third question was presented that asked what was wrong with the image, with the options of too blury, the lesion was too small, it was too dark, it was too light, it needed a scale, the colors were wrong, insufficient contrast, too much contrast, and other (write in).

Participants were offered the option of viewing the consensus answers for all questions, even if they answered the consensus answer with high confidence.

Selection of Participants

Invitations were emailed to all participants of the first survey.

Protection of Human Rights

The same conditions were applied as for the first survey.

Data cleaning

Many of the “other” statements were obvious restatements of one of the other options, e.g. “other – the image was too dark to see details.” These were recoded to the appropriate option. Some respondents were unable to finish the survey and requested a second link. The answers to these two partial survey responses were merged as a post processing task. Demographics were recoded as in the first survey.

Statistical Analysis

Statistical analysis iwas performed as for the first survey, though almost all evaluations use the chi squared statistic.

Third survey

Structure of the survey

The survey was composed of a collection of 31 patterned injuries corresponding to a subset of of those in the first survey (Table 1). These images were not the same images as those in the first survey, but were of similar injuries. The respondents were presented with the image and a multiple choice question asking the the most likely diagnosis. The choices were identical to the same question for the corresponding most specific diagnosis in the first survey (denoted “Tier 3” in the first survey), with the exception of minor modifications when the results would be nonsensical (e.g. wording specific for the wrong sex, for instance). A follow-up question asked how confident the respondent was on a scale of 1 to 10.

Responses to all questions were not required. The survey was not time-limited, and respondents could stop and return to the survey. Results were collected from 24 DEC 2015 to 22 JAN 2016. The time taken for respondents to complete the survey varied between 24 and 179 minutes, ignoring those who simply opened the survey but did not complete it.

The survey was web-based, using Qualtrics survey software hosted by East Carolina University. No measurement or determination of viewing conditions was obtained; the study design was to replicate how consultation images would be viewed in an office, and it was assumed that those, or similar, conditions would be used for the survey.

Selection of Images:

An expert panel of 10 senior forensic pathologists from across the United States volunteered to help provide and select images. In contrast to the first survey, where any patterned injury with diagnosis was requested, only those matching diagnoses from the first survey were requested. Approximately 50 images were obtained and added to the pool of approximately 150 images from the first survey. From these 32 images were chosen on the basis of matching the diagnosis of an image in the first survey, classical presentation and quality. Because of a programming error, one image was inadvertently dropped from the survey when it was deployed.

As with the first survey, these images were largely taken from teaching sets and private collections, it was not possible to independently determine the “ground truth” of the diagnosis. Accordingly, this is a study of consensus of diagnosis rather than correctness. The list of diagnoses are presented in Table 1.

Recomputation of selected first survey results

The Tier 3 responses from the first survey corresponding only to those images with matching diagnoses were selected and reanalyzed. . The comparisons in this article refer to this matching subset rather than the entire first survey results.

Protection of human rights

The survey and the experimental protocol were reviewed by the Institutional Review Board of Brody School of Medicine at East Carolina University as well as by the National Institute of Justice for anonymization of the images and protection of participant's identities. Participant identification is retained through the series of surveys for the purpose of troubleshooting emails, logistics, and follow-up, but will be discarded and the responses anonymized upon completion of the project.

Data cleaning:

Data cleaning was performed in the same manner as with the first survey (1).

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

Statistical analysis

Statistical analysis was done using the “R” open source statistical package as well as Microsoft Excel, a spreadsheet package as described for the first survey.

III. Results

Summary:

The first survey demonstrated as surprising heterogeneity in consensus, with a median consensus of 74%. There was a high correlation between confidence and degree of consensus at the question level, but only mild correlation at the individual level. Older pathologists and those who were actively performing autopsies were more likely to provide consensus responses.

The second survey demonstrated that most of the non-consensus answers did not represent actual disagreement about what the answer was, but instead represented various expressions of ambiguity – that the lesion could have been the consensus but was not really specific, that there were issues with naming conventions, etc. There were also significant issues with the graphical user interface, particularly with using the slider than indicated the degree of confidence in the answer. There were a large number of significant differences in responses by demographic groups (the statistical analysis on this is not quite complete). For instance, men were more likely than women to indicate that the consensus answer was wrong, and their answer was right (p=0.01). Women were much more likely to indicate that a lesion was nonspecific (p=2x10-8), were also more likely to simply say they didn't know the answer (p=1.5x10-15), and were more likely to indicate a problem with naming convention (p=0.04). Younger pathologists were more likely to say they did not know the answer, middle aged pathologists were much more likely to invoke nonspecificity, and older pathologists were more likely to indicate that the consensus answer was wrong. Older pathologists were also more likely to invoke issues of ambiguity, though at less significant level than middle aged pathologists. The full statisical analysis is not yet complete, but will be complete by the end of the grant period.

The third survey is still ongoing. Preliminary results demonstrate a profound effect of history. The median degree of consensus changed from 74% in the first survey to 98% in the third survey, looking at the 153 responses so far.

Specific statistical findings:

First survey:

Approximately 1115 emails were sent to the NAME membership. While this survey was oriented towards forensic pathologists, non-pathologist members were welcome to participate. Approximately 363 surveys were started and 210 surveys completed. This response rate is characteristic of NAME surveys.

Participants:

Approximately 1115 emails were sent to the NAME membership. While this survey was oriented towards forensic pathologists, non-pathologist members were welcome to participate. Approximately 363 surveys were started and 210 surveys completed. This response rate is characteristic of NAME surveys.

Demographics:

Of the 210 respondents who completed the survey, 78 were female (37%) an 131 were male (63%). The average age of respondents was 51.8 years (range 27-81). The average years of experience was 16.7 (range 0-46), skewed towards fewer years (10th percentile was 2.0 years, 25th percentile 15 years, 75th percentile 25 years, 90th percentile 34 years).

Of the respondents, 201 were medical doctors of whom 9 also held PhDs, 1 MBA/MPA, and 4 JD. Two participants held DDS degrees, two held terminal MS degrees, three held terminal BS degrees, and 2 held terminal AS degrees.

Two hundred of the respondents performed autopsies on a regular basis. Seventy did death investigation on a regular basis. Fifty-seven peformed administrative duties on a regular basis.

Sixty-four respondents were certified in Anatomic Pathology only, 125 were certified in combined Anatomic and Clinical Pathology, and 183 were board-certified Forensic Pathologists. Twenty respondents also held other subspecialty certifications (most commonly neuropathology). Twelve respondents were ABMDI certified investigators.

Eighty-five (40%) were university-affiliated, and 54 (25%) were hospital-affiliated.

The majority of respondents worked in jurisdictions of over 500,000 people, with 56 working in jurisdictions of 500,000 to 999,999 people and 88 working in jurisdictions of over one million people. Only 8 worked in jurisdictions of less than 50,000 people. One hundred eleven respondents worked as staff, while 55 were Chiefs or Deputy Chiefs. Twenty-five were consultants and nine were retired. One hundred fifteen respondents were Fellows in the National Association of Medical Examiners, and 60 held Member status. Eleven held emeritus status.

Consensus answer versus supplied diagnosis:

While ground truth was not independently verified, the accepted diagnosis was obtained with the teaching cases and cases known to the person supplying the image. In general, the consensus answer was the provided diagnosis. There were rare discrepancies, however. At the Tier 1 level, a healing cigarette burn was identified as a “penetrating injury” rather than a thermal injury, though at Tier 2 and Tier 3, the supplied and consensus answers were the same. Similarly, the consensus answer for a congenital dermal melanocytosis was “blunt trauma” at Tier 1, though higher tier answers were correct.

Further discrepancies reflected nomenclature issues. In the instructions for the survey, the respondents were instructed to code animal bites as “penetrating.” However, the consensus answer for one dog bite was “blunt trauma,” with some participants adding the notation that they refused to call a dog bite penetrating because they were trained otherwise.

With respect to Tier 2 answers, a chopping wound with an machete had a consensus answer of “incised,” rather than “chop.” Two questions had “I don't know” as consensus answers – a TASER dart mark partially obscured with bleeding in a patient with disseminated intravascular coagulation, and an image of recent skin popping marks.

With respect to Tier 3, an image of ice pick stab wounds (without a scale) had a consensus answer of “double-edged dagger.” Two images had that the consensus answer of “It's specific and I could match it against an exemplar, but I don't know” included a bludgeoning with a roofing hammer and a baseball bat blow to the head that left a logo imprint. The TASER mark and skin popping images were coded as “I don't know” in Tier 3 as in Tier 2.

Descriptive statistics:

It was expected that overall consensus would be highest at tier 1, intermediate at Tier 2, and lowest at Tier 3. This was not the case. Consensus at Tier 1 averaged 0.77, Tier 2 averaged 0.68, and Tier 3 averaged 0.72, with Tier 2 having significantly lower consensus than either Tier 1 or Tier 3. In contrast, average confidence did steadily decrease, being 8.42 for Tier 1, 8.27 for Tier 2, and 7.92 for Tier 3 on a scale of 1-10.

The difference in consensus for each category in each tier is shown in Table 2. As shown in Table 3, the only category that showed a significant difference was blunt trauma when moving from Tier 1 to Tier 2. This drove the entire dataset into significant range, and represents the reason that Tier 2 consensus is lower than Tier 3. All other categories except electrical injury showed the opposite trend. Electrical injury had a greater drop, but consisted of only two questions, and was thus not statistically significant. When comparing Tier 1 with Tier 3, there were significant changes in the categories of “natural disease” and “electrical.” The natural disease result represents the ambiguity of the photograph of senile purpura, which many classified as “blunt trauma” in Tier 1, often with the comment similar to “senile purpura requires some trauma.” When comparing Tier 2 with Tier 3, the only significant change was in the “penetrating” injuries. This was largely represented by the fact that many respondents mistook an ice-pick injury for a double-edged knife wound, and some disagreement in the interpretation of ambiguous animal bites.

The 10 images with the highest consensus scores were images of a shotgun wound, hesitation marks, an intermediate range gunshot wound, a hilt mark, defense incised wounds, a distant gunshot wound, a ligature mark, drag marks, a muzzle imprint, and a Lichtenberg figure. The 10 images with the lowest consensus were images of a stabbing with an ice pick, a healing cigarette burn on the foot of an infant, claw marks from a vulture on a torso, a “Bic” cigarette lighter burn mark, recent skin popping, a baseball bat bludgeoning with a logo skin imprint, defibrillator paddle marks (consensus was divided between “electric” versus “abrasion” at Tier 1), TASER marks (low consensus at Tiers 1 and 2, but high consensus at Tier 3), and congenital dermal melanocytosis (low consensus at Tier 1, but high consensus at Tiers 2 and 3).

The results of the first survey were surprising. Instead of a high consensus among pathologists, the average consensus for all questions was only 72% (median 74%), with a range from 25% to 100% concurrence. This clearly violated our assumption that there would be uniformly high concordance involving “classic” patterned injuries. The results of this survey were presented at the 2015 AAFS meeting and the paper has been accepted for publication in the Journal of Forensic Sciences.

There was a very high correlation between the average percent consensus with average confidence for each questions (r = 0.86, 0.86, 0.85, 0.86, for Tiers 1,2,3 and combined respectively, p<2.2 x 10-16 for each). See Figure 1. However, the correlation between confidence and consensus was low at an individual participant level using (r=0.26, p=0.0002 Tier 1, r=0.12,p=0.07 Tier 2, r=0.15, p= 0.05 Tier 3, r=0.17, p=0.01 combined). See Figure 2.

The less confident a respondent was overall, the more accurate the correlation between confidence and consensus , though the correlation was poor (r= -0.16, p=0.02 for combined tiers). Conversely, the higher the overall percent consensus answers for an individual, the higher the correlation between consensus and confidence (r=0.22, p= 0.0001 for combined tiers). Thus, people who had a higher percentage of consensus answers were less confident in general, but more accurate in their assessment of whether or not their answer was “right.” See Figures 3 and 4.

To look at how many people it would take to achieve a higher correlation between consensus and confidence compared to individual results, the participant pool was sampled without replacement into groups of three to 70 people, and their aggregate correlation between consensus and confidence was measured. For each group size, 300 group samples were created. The decrease in standard deviation followed the expected power law pattern (Figure 5).There was no signficant difference between the responses of men and women. There was a mild negative correlation with age (r=-0.22, p=0.003 for combined tiers). With respect to other demographics, there was significant difference between the degree of consensus among those who are actively performing autopsies versus those who are not (0.73 versus 0.63, p=0.000006 for combined tiers), and the difference in confidence was nonsignificantly higher (8.2 v 7.7, p=0.08 for combined tiers). Those who performed administrative tasks did not display a difference in average rate of consensus but were significantly more confident (0.71 v 0.72 consensus, p=0.55, 8.7 v 8.3 confidence, p=0.03 for combined tiers.).

Physicians were more likely to provide a consensus answer (0.72 v 0.62, p=0.000004 for combined tiers), but showed a non-significant difference in confidence (8.2 vs 7.7, p=0.13). People with PhD or JD degrees , and private consultants, were not significantly likely to provide more or less consensus answers, but were significantly more confident of their responses. Because all but a few of the physician respondents were certified in Forensic Pathology (or international equivalent), no meaningful comparison between certifications could be done. Pathologists practicing in large jurisdictions were more likely to provide consensus answers (0.73 v 0.70, p=0.01), but were not different with respect to confidence. Pathologists practicing in jurisdictions of less than 5000 were less likely to provide consensus answers (0.64 v 0.72, p=0.01), but were not different with respect to confidence. Pathologists working in lay coroner and forensic pathology coroner systems were more likely to provide consensus results (two sample t-test, lay coroner v other and FP coroner v other), but this difference disappeared with controlled for jurisdiction size. University or hospital affiliation did not provide any significant differences.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Staff members were more likely to provide consensus answers than Chiefs or Division Chiefs (0.73 v 0.69, p=0.008), but showed no difference in confidence.

Affiliate members of NAME (who are not physicians) were less likely to provide consensus answers (0.67 v 0.71 p=0.04) but showed no difference in confidence. The same was true of Emeritus Fellows (0.65 v 0.72 p=0.009). Conversely, Fellows and Members were more likely to provide consensus answers.

Second Survey:

Of the 363 respondents to the original survey, 153 responded to the second survey, and 102 completed it. While the full statistical analysis of the results is still ongoing, the intial results have been enligtening and have been accepted for presentation at the 2016 AAFS meeting. The manuscript will be finished by then. Note that the following p values are still pending review by our statistician.

There were significant differences in why respondents did not answer consensus answers by Tier. Respondents were more likely to invoke nomenclature issues or the presence of multiple injuries in Tier 2, while they were more likely to state that the injury was not specific to the consensus answer in Tier 3.

Many of the specific significant findings are described in the body of the report. Of particular interest from a human factors viewpoint was that relatively few of the discrepancies were due to a belief in a different diagnosis. The vast majority of the differences had to do with issues of ambiguity – nomenclature, the presence of multiple lesions (e.g. abrasions along with lacerations), or a hesitance to make a specific diagnosis in the absence of corroborating non-anatomic evidence. A number of respondents commented that they simply would not make a defnitive diagnosis without a supporting history.

Unlike the first survey, there were significant gender differences. Men were more likely than women to indicate that the consensus answer was wrong, and their answer was right (p=0.01). Women were much more likely to indicate that a lesion was nonspecific (p=2x10-8), were also more likely to simply say they didn't know the answer (p=1.5x10-15), and were more likely to indicate a problem with naming convention (p=0.04). Men were somewhat more likely to invoke image quality (p=0.04). The results were similar when looking at reasons for lower confidence, though to a much lesser degree. Women were more likely to indicate that image quality was poor because the lesion was too small in the frame (p=0.03).

Age and experience made large differences in why people did not answer the consensus answer. Younger pathologists (age < 40) were much more likely to simply say they didn't know the answer (p=2x10-16), and were less likely to invoke nonspecificity (p=0.008), atypical presentation (p=0.002). They were less likely to say they simply pushed the wrong button (p=0.03). Middle-aged (41-60 year were much more likely to invoke nonspecificity (p=0.00000018) and were less likely to say the didn't know the answer (p=0.000004). Older pathologists (61+ years) were again more likely to indicate that the consensus was wrong (p=0.03), and were also more likely to invoke equivalency, the presence of multiple lesions, naming convention issues, and nonspecificity. They were much more likely to have considered the presentation atypical (p=0.00001), and were much less likely to say they just didn't know what it was (p=3x10-11).

Image quality was not significant at any age when asked why the respondent did not give the consensus answer, though almost rose to signficance with the oldest group (p=0.07), where the older pathologists were less likely to invoke it as an issue (73 responses, with an expected value of 88). In contrast, young pathologists were much more likely to invoke image quality as a cause for low confidence (p=0.0002). They were also significantly less likely to invoke specificity issues. Middle-aged pathologists were more likely to invoke specificity, and more likely to have accidentally incorrectly set the slider widget (which defaulted to a value of 5 out of 10).

Younger pathologists were more likely to indicate that the lesion was too small in the image, and older middle aged pathologists were more likely to say it was blurry.

Those who performed autopsies as a primary function were more likely to indicate that the consensus was wrong when they did not provide the consensus answer compared to those who did not perform autopsies, and were much more likely to have considered the injury nonspecific (p=1x10-11). They were also more likely to invoke image quality issues (p=0.006). They were much less likely to indicate that they simply didn't know the answer (p=2x10-16). The results were similar for confidence. Statistical evaluation of specific image quality issues was not possible due to the small number of people who invoked it. People who performed investigations were more likely to invoke nospecificity, and were much less likely to indicate that they did not know the answer (p=2x10-16), and somewhat less likely to invoke image issues (p=0.02). Administrators were more likely to indicate that their answer was equivalent to the consensus answer (p=0.003), and were markedly more likely to invoke issues of specificity (p=3x10-15). Administrators were also less likely to indicate they did not know the answer.

People with terminal Associate's degree (nonphysicians) were much more likely to simply not know the answer (p=0.0000004). The number of people with a terminal Bachelor's degree were too small for statistical evaluation. Those with terminal Master's degree were also more likely to indicate they simply didn't know the answer (p<2x10-16). Physicians were more likely to indicate the lesion was nonspecific, invoke naming issues, and much less likely to indicate they did not know the answer (p<2x10-16). Those with a PhD in addition to MD were less likely to indicate their answer was the equivalent of the consensus answer (p=0.03), and were more likely to invoke naming issues (p=0.00001); they were also more likely to have hit the wrong button (p=0.03). Those with JD degrees were more likely to indicate a naming issue (p=0.05), and were less likely to indicate their answer was equivalent to the consensus answer (p=0.02). They were more likely to invoke the presence of multiple lesions (p=0.000000001).

Issues of confidence provided similar results. Image quality results provided too few responses for good statistical evaluation.

Those physicians with AP only certification were more likely to indicate they did not know the answer (p=3x10-9), and atypical presentation, and were less likely to invoke the presence of multiple lesions. Those with AP/CP certification were more likely to say that the consensus was wrong (p=0.009), much less likely to say they didn't know the answer (p=0.000006), and less likely to invoke image quality. Those with FP certification were more likely to say the consensus answer was wrong (p=0.01), invoke the presence of multiple lesions or nonspecificity. Those with other certifications were markedly more likely to say their answer was equivalent to the consensus answer (p=0.00009), multiple lesions (p=0.004), atypical presentation (p=0.00005), and less likely to invoke naming convention (p=0.0004).

Physicians who were also ABMDI certified were more likely to indicate the consensus was wrong (p=0.0003), and nonspecificity, but were less likely to invoke naming convention issues (p=9x10-13).

Those who worked in ME offices were more likely to indicate that the consensus answer was wrong (p=0.000004), and less likely to indicate that their answer was equivalent to the consensus answer (p=0.00017), that the lesion was atypical, naming convention or that they simply didn't know the answer (p=2.6x10-9). They were much less likely to have user interface issues. Those who worked in FP Coroner offices were more likely to say their answer was equivalent to the consensus answer. Those who worked in Physician Coroner offices were more likely to indicate that their answer was equivalent, and much more likely to say they didn't know the answer (p=0.00005). People who worked in Lay Coroner offices were less likely to say their answer was equivalent to the consensus answer (p=0.02), invoke multiple lesions, and much less likely to say they didn't know the answer (68 responses, expected value 143, p=2x10-13), jbut were much more likely to say the lesion was nonspecific (p=7x10-15) or atypical (p=0.00008). Those in Lay Coroner offices were more likely to invoke image quality issues (p=0.05). Private consultants were less likely to consider their answer equivalent to the consensus answer (p=0.000003), invoke image quality, indicate the lesion was nonspecific, but were much more likely to have user interface issues (p<2x10-16). Retired respondents were less likely to indicate the consensus answer was wrong, choose poor image quality, maning convention, or invoke nonspecificity, but much more likely to indicate their answer was equivalent to the consensus answer (98 responses, expected value 32, p<2x10-16) or invoke atypical presentation (32 responses, expected value 6, p< 2x10-16).

Third survey:

The primary finding of the third survey is that the inclusion of history increased the degree of consensus per question from a median of 77% to 98%, degree of consensus per participant from 77% tio 94%, confidence per question from 56% to 90%, and confidence per participant from 79% to 90%.

Please see the attached manuscripts for graphics.

V. Conclusions

This study was originally designed to evaluate imaging issues rather than human factor issues. The results of the first survey was very surprising, but suggested that history was of paramount importance in forensic medicine, even in the evaluation of patterned injury photographs.

The results of the first survey, which was constructed with the aim of providing a uniformly high consensus response, was surprising. “Classic” patterned injuries may not be so classic. There was a wide variation in answers, though the reasons for it may be disparate.

The demographic and practice differences in responses should be interpreted with caution. This survey was constructed as a preliminary survey for a series of studies about image quality issues, and the demographic questions were not the primary goal. The target population was forensic pathologists. While the invitation was open to all NAME members, the invitation letter was explicit in targeting forensic pathologists, and it is not surprising that relatively few others responded. Thus, while the practice and demographic findings may be interesting where statistically significant, the absence of a statistically significant difference does not mean that a real difference does not exist. Virtually all of the demographic responses had statistical power below 20 percent.

Forensic pathology is a specialty where knowledge comes from experience, and it is a perceptual discipline where visual cognition is as, or more important than the abstract weighing of evidence or evaluation of probabilities(10). People who are actively performing autopsies, people in large systems, and staff members tend to provide consensus answers more often than those who do only consultations, people in small jurisdictions, and Chiefs and administrators. There is a mild negative correlation with age, but this may represent the change in autopsy workload or increasing adminstrative or consultative responsibilities. Adding more academic degrees or performing consultations increased confidence, but did not change the likelihood of providing a consensus answer.

The issue of cognitive bias and the integration of history into diagnosis is a topic of intense discussion(11). Some authors suggest that history is an integral part of medical diagnosis.(12) Others suggest that knowing history introduces cognitive bias that increases error rate(13). In the first survey, no history was provided, and the diagnoses were made based on the images alone. This caused some consternation among some of the participants. Just as some indicated that they would decline to make a diagnosis of stab wounds without seeing certain prepared visualizations, others indicated in their comments that they were not inclined to make a diagnosis without corroborative history. The survey attempted to avoid this issue by asking for the more likely or probable diagnosis, rather than implying certainty, but the lack of history remained an issue for some participants. One such participant noted:

... [w]e as medical examiners have been taught over and over never to jump to conclusions simply based on first hunches. When we investigate a case, we should never work in a vacuum... What should be promoted is to provide one or two images and then ponder over the possibilities, given a set of facts about the history and circumstances.

Collectively, forensic pathologists had a good sense of confidence versus consensus, but at an individual level the correlation, while positive, was not particularly high. This may be important when considering strategies involving peer review or quality assurance. Asking an individual pathologist how confident he or she is about an individual injury may be as much a question of overall self-confidence as it is real certainty of correctness. This is particularly an issue in light of the finding, albeit small, that the more confident a participant was about his or her answers in general, the less correlation was found between confidence and consensus. The finding that those who had higher average confidence had a lower correlation between confidence and consensus is consistent with the well-known Dunning-Kruger effect where those with less competence are more confident of their work, while those with more competence of less confident (14).

The second survey suggests that most of the disagreements do not represent beliefs in different diagnoses as much as varying degrees of uncertainty and ambiguity in the absence of history. Some pathologists are hesitant to make a firm diagnosis on the basis of “just' an image, and instead rely more

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

heavily on historical and contextual data. Experience was also an important factor, with younger pathologists more likely to simply not know the answer, middle aged pathologists to me more concerned with issues of specificity, and older pathologists being more inclined to believe their answer correct and the consensus response incorrect (which it sometimes was).

The third survey demonstrates the profound effect of historical and contextual data in medical diagnosis. This result is not surprising. The importance of semantic content in visual interpretation is well-known. Evaluating imagery in the context of semantic information utilizes a different part of the brain than evaluating imagery without it(15).

Another author notes(16):

In a number of studies the context provided by a real-world scene has been claimed to have a mandatory, perceptual effect on the identification of individual objects in such a scene. This claim has provided a basis for challenging widely accepted data-driven models of visual perception in order to advocate alternative models with an outspoken top-down character. The present paper offers a review of the evidence to demonstrate that the observed scene-context effects may be the product of post-perceptual and task-dependent guessing strategies. A new research paradigm providing an on-line measure of genuine perceptual effects of context on object identification is proposed. First-fixation durations for objects incidentally fixated during the free exploration of real-world scenes are shown to increase when the objects are improbable in the scene or violate certain aspects of their typical spatial appearance in it. These effects of contextual violations are shown to emerge only at later stages of scene exploration, contrary to the notion of schema-driven scene perception effective from the very first scene fixation. In addition, evidence is reported in support of the existence of a facilitatory component in scene-context effects. This is taken to indicate that the context directly affects the ease of perceptual object processing and does not merely serve as a framework for checking the plausibility of the output of perceptual processes. In other words, adding context, at least visually, enhances memory of the object through the associational cortex.

This context effect is even more pronounced in picture naming, because it involves semantic issues. It turns out that you mix your semantic and visual perceptions when naming objects, and semantic context *becomes* part of the visual process(17):

Object detection and identification are fundamental to human vision, and there is mounting evidence that objects guide the allocation of visual attention. However, the role of objects in tasks involving multiple modalities is less clear. To address this question, we investigate object naming, a task in which participants have to verbally identify objects they see in photorealistic scenes. We report an eye-tracking study that investigates which features (attentional, visual, and linguistic) influence object naming. We find that the amount of visual attention directed toward an object, its position and saliency, along with linguistic factors such as word frequency, animacy, and semantic proximity, significantly influence whether the object will be named or not. We then ask how features from different modalities *are combined during naming, and find significant interactions between saliency and position, saliency and linguistic features, and attention and position. We conclude that when the cognitive system performs tasks such as object naming, it uses input from one modality to constraint or enhance the processing of other modalities rather than processing each input modality independently.*

If history is this important to even the most directly visual parts of medical diagnosis, and data hiding is not a viable option, perhaps there are other options that may work. It may be that providing some small group evaluation of the images of an injury may produce a better of indication of certainty than a statement of degree of certainty of a single practitioner. The decrease in variation in correlation between certainty and consensus decreased according to the standard power law for noise implies that small group peer review would provide almost as much benefit as larger groups, if these findings are repeateable.

Data hiding, a method currently proposed, exacerbates this problem more than it solves it. Because medical diagnosis is primarily ordinal and involves Baconian reasoning, errors derive from the structure of the inferential network constructed by the physician rather than from relatively small errors in the estimates of individual probabilities. Medical diagnosis is ordinal, and is more concerned with cascades of semiquantitative ordering of competing hypotheses. This makes medical diagnosis extraordinarily robust against the kind of influence that gatekeeping attempts of avoid. In fact, studies have indicated that such gatekeeping will increase rather than decrease error.

In one study by Onisko and Druzdzel, examination of medical decision making revealed it to be robust to three orders of magnitude in the error in probability estimation by their models. Instead, the structure of the inferential net was more important. In their examination removing low probability information from their model removed that robustness and increased diagnostic error. They write(18):

O*ur study of the influence of precision of parameters on the diagnostic accuracy of Bayesian networks is inspired by a study performed by Clancey and Cooper [], who probed the sensitivity of the MYCIN expert system [] to the accuracy of its numerical specifications of degrees of belief, certainty factors (CF). CFs are considered an ad hoc measure of uncertainty that does not suffer from the problems encountered in probability, such as the need to add up to 1.0 or the importance of zeroes. Similarly to our result, Clancey and Cooper noticed minimal effect of precision on the performance of MYCIN. However, they attributed it partly to a broad coverage of microorganisms that a possibly incorrectly recommended antibiotic would cover, resulting in a reasonably correct therapy. In case of Bayesian networks we believe that a critical factor may be preservation of ordinal relationships among the parameters. Diagnosis, as interpreted typically in probabilistic context, amounts to finding the most probable hypothesis, which also rests on ordinal relationships among disorders. The results of our experiments touch the foundations of qualitative modeling techniques. As qualitative schemes base their results on approximate or abstracted measures, one might ask whether their performance will match that of quantitative schemes, either in terms of their strength or the correctness of their results. Because our models performed reasonably well, even when every parameeter in the model was equal either to ε or $1 - \varepsilon$, it seems that approximate order of magnitude schemes might offer acceptable* *recommendations, at least if they conform to the basic rules of probability calculus, which is what our models did.*

Our results support another approach, suggested by an anonymous reviewer. One might focus probability elicitation on obtaining verbal probability estimates, such as those on the Likert scale [], covering the categories "very unlikely", "unlikely", "50–50", "likely" and "very likely." This should, of course, be done with much caution, as the meaning of verbal phases typically varies from human to human and is sensitive to context.

This quote contains two important concepts that seem to be disregarded in the recommendations supporting data hiding in medical diagnosis. First, it explains why attempting to tag specific probabilities on individual diagnoses is inappropriate, since in reality diagnosis is a semiquantitative ordinal process. Second, because the purpose of data hiding is to take events that have a low probability of being important and impose the value of zero upon that probability, these and other models predict it will degrade the robustness of the inferential system and create, rather than eliminate, error. Worse, as more and more "irrelevant" information is removed, the system degrades exponentially as more data is hidden.

This is supported by other research that suggests that it is the deficiency of information rather than the inclusion of too much information that contributes the most to errors in medical diagnosis. Croskerry, in his well-known evaluation of cognitive bias in medical diagnosis, lists 32 different sources of cognitive errors in medical diagnosis (19). Lawson and Daniel point out that, upon review of the list created by Croskerry(5):

In our view, the large majority of these subconscious tendencies... operate to limit the number of alternatives generated or to limit the adequate consideration and testing of alternatives.

Data hiding is only one of many cognitive forcing strategies that are available. Croskerry, in his discussion of cognitive errors in medical decision making proposes the use of "metacognition training," as a method of reducing cognitive bias (20):

Thus, metacognition generally describes the process of actively stepping back from the pushes and pulls of the immediate situation (de-anchoring), reminding oneself of the limitations and failings of memory, seeing the clinical problem in a wider perspective than that dictated by the obvious presentation (representativeness error), perhaps reminding oneself of specific lapses or failures in the past (availability), and finally activating known cardinal rules or caveats (cognitive forcing strategies). This more dynamic style of decisionmaking fits the naturalistic decisionmaking model and adds strength to the preferred approach, much as Cohen[]has advocated. Also, it appears to create an explicit opportunity for improving transfer of effective decisionmaking across a wide variety of clinical problems...

Deep peer review provides exactly the kind of feedback necessary for this kind of metacognitive training, and can be developed within a formal framework for cognitive debiasing. Metacognitive training has shown mixed results. While training of medical students had limited benefit, its use in dermatopathology showed significant results (21,22). A similar concept of “reflective practice” has also shown some success (23,24,25), as has the idea of “crew resource management” or “cockpit resource management” in other areas where errors can be devastating. The finding that increasing the number of reviewers increases the correlation of certainty and degree of consensus supports this idea.

In conclusion, these studies demonstrate the importance of history in the diagnosis of patterned injury of the skin. Denying history produces significant lack of consensus, primarily due to issues of ambiguity rather than actual differing diagnoses. Providing history increased the degree of consensus by 20 points to near complete agreement. The study also presents evidence that peer review may be a way of both allowing the history needed for diagnosis while providing feedback that diminishes individual effects.

References:

1) Oliver WR, Fang X. Forensic Pathologist Consensus in the Interpretation of Photographs of Patterned Injuries of the Skin. J Forensic Sci. 2016 Jul;61(4):972-8.

2) Oliver WR. Cognitive bias in medicolegal death investigation. Acad Forens Pathol. 2015 5(4):548-560.

3) OSAC-MDI response to Human Factors Resource Committee 4/19/2016

4) Oliver, WR. Reasons for lack of consensus in forensic pathologist interpretation of photographs of patterns of injury of the skin J Forens Sci Accepted pending revision

5) Oliver, WR. Effect of history and conext on forensic pathologist interpretation of photographs of patterned injury of the skin. J Forens Sci Submitted for publication

6) Oliver, WR. Manner determination in forensic pathology. Acad Forens Pathol. 2014 4(4):480-491.

7) Lawson AE, Daniel ES. Inference of clinical diagnostic reasoning and diagnostic error. J Biomedical Informatics. 2011 44:401-412.

8) Schum DA. The evidential foundations of probabilistic reasoning 1 ed. New York:Wiley, John & Sons; 1994.

9) Higgs J, Jones MA, Loftus S. Clinical Reasoning in the Health Professions. 3rd edition. 2008 Butterworth-Heinemann, publisher. 520 pp.

10) Larson AR, Cibas ES, Granter SR, Laga AC. Uncovering bias in the cytologic evaluation of cervical squamous lesions. Am J Clin Pathol 2015 143:143-148.

11) Farand L, Lafrance J-P, Arocha JF. Collaborative problem-solving in telemedicine and evidence interpretation ina complex clinical case. International J. Medical Informatics 1998 51: 153-167.

12) Oliver, WR. Inference in forensic pathology. Acad Forensic Pathol 2011 1(3):254-275

13) Dror, I.E & Charlton, D. Why experts make errors, J. Forensic Identif. 52006 6:600–616

14) Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med. 1992 Feb; 156(2):163-5.

15) Kassin SM, Dror IE, Kukucka J. The forensic confirmation bias: problems, perspectives, and proposed solutions. J Appl Res Mem Cogn. 2013 Mar, 2(1):42–52.

16) Kruger J, Dunning D. Unskilled and Unaware of It: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Personality Soc Psych 1999 77(6):1121-1134.

17) Bar M, Aminoff E. Cortical Analysis of Visual Cortex. Neuron 2003 38(2):347-358.

18) De Graef P, Christiens D, d'Y'dewalie G. Perceptual effects fo scene context on object identification Psychological Research 1990 52(4)317-329

19) Clarke AD, Coco MI, Keller F. The impact of attentional, linguistic, and visual features during object naming. Front Psychol. 2013 Dec 13;4:927.

20) Agnieszka Oniśko and Marek J. Druzdzel. Impact of precision if Bayesian network parameters on accuracy in medical diagnostic systems. Artificial Intelligence in Medicine 2013 57:197-206

21) Croskerry P. Importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003 78:775-780.

22) Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Annals of Emergency Medicine 2003 41(1): 110-121.

23) Feyzi-Behnagh R, Azevedo R, Legowski E, Reitmeyer K, Tseytlin E, Crowley RS. Metacognive scaffolds improve self-judgments of accuracy in a medical intelligent tutory system. Instr Sci. 2014 Mar;42(2):159-181.

24) Sherbino J, Yip S, Dore KL, Siu E, Norman GR. The effectiveness of cognitive forcing strategies to decrease diagnostic error: an exploratory study. Teach Learn Med. 2011 Jan;23(1):78-84

25) Hall KH. Reviewing intuitive decision making and uncertainty: the implications for medical education. Med Educ. 2002;36:216 –224.

26) Mamede S, Schmidt HG, Rikers R. Diagnostic errors and reflective practice in medicine. J Eval Clin Pract. 2007;13:138 –145.

27) Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Med Educ. 2008 May;42(5):468-75

Dissemination:

The results of the first survey were presented at the 2015 American Academy of Forensic Sciences.

The manuscript of the first survey results have been accepted for publication in the Journal of Forensic Sciences.

The results of the second survey were presented at the 2016 American Academy of Forensic Sciences.

The manuscript for the second survey has been submitted to JOFS and has been accepted pending revision.

It is anticipated that the results of the third survey will be submitted for presentation at the 2017 AAFS.

The manuscript of the third survey has been submitted to JOFS.