Research Report

Perceptual Serial Dependence in a Medical Classification Task

Dana Pietralla

Student number: 7357858

dpietral@smail.uni-koeln.de

Graduate Program

Psychology (research-oriented)

University of Cologne

20.09.2022

Declaration of Originality

I confirm that this work is original and was written by me without further assistance. Appropriate credit has been given where reference has been made to the work of others.

Dana Pietralla Cologne, 20.09.2022

Abstract

Past observations, affect actions and behaviours in the present. On a small scale, this effect is called *perceptual serial dependence*, a sequential effect in which what was previously seen influences what is seen at the moment. By doing this, the visual system recycles previous perceived objects and features instead of processing each momentary image in an independent way, a useful strategy in a variety of events in everyday life. However, this perceptual effect can also have severe consequences in situations concerning everyday visual classification, such as security controls and detection of medical anomalies.

In line of the latter, this work hypothesizes that because of serial dependence, classification ability to detect skin cancer in dermatological images is biased towards the previous image, systematically altering image interpretation. Further hypotheses investigate a more detailed approach within serial dependence. All hypotheses were investigated with a real-world data set provided by a medical AI company and consisted of 758.139 data points of N = 1,137 medical trained users who classified dermatological images.

Data analysis showed that real-world dermatological diagnostics was affected by the observer's past visual experience, such that classification was pulled towards the previous image class seen, hence showing the existence of serial dependence.

The results give insights to mechanisms leading to serial dependence and are important to offer solutions that minimize sources of errors in everyday and work-related visual decision-making.

Keywords: serial dependence, classification, visual perception, heuristic, diagnostic errors

Introduction

Visual Heuristics

Whether it's the sound of a musical instrument, the movement of a car or the smell of freshly brewed coffee - Perception is a key purpose of the brain, which processes all of the sensations we experience every day. A huge proportion of this activity is devoted to vision. One of the most remarkable aspects about vision is the transferral the visual system is capable of, a 3 dimensional world is turned into a 2 dimensional retinal image and decoded into a 3 dimensional perception we are able to understand and interact with (Haber, 1978).

When perceiving billions of stimuli every day we gather evidence on what stimuli are important to pay attention to and how to respond to them in order to receive a positive outcome. We learn that some stimuli and the consequence of their appearance is occurring repeatedly. The human brain has evolved under this premise and developed mental shortcuts that enable time- and energy-saving opportunities. These *heuristics* (Kahneman et al., 1982) display a simple way to navigate through the world without the need of extensive elaboration on each decision to be made. By doing this, humans take advantage of the autocorrelation in the world to respond in the best way possible. Autocorrelation is a characteristic of data which shows the degree of similarity and therefore dependence between the values of the same variables over successive time intervals (Bock, n.d.): If the temperature has been increasing the last days it's reasonable to expect it to be rising tomorrow as well.

Heuristics occur in every area of life where some kind of decision-making is relevant. As an educated guess, they offer a sufficient, if not an optimal solution. But as all approximations, they cannot capture the whole complexity that exists in the world. And while they are usually helpful in enabling an effective way to react to most stimuli in the world they can lead to wrong and deceptive outcomes. Such shortcuts can cause *cognitive biases* and are defined as a repeating misstep in processing, thinking or assessing, causing a systematic pattern of deviation from rationality (Gilovich et al., 2002). If cognitive biases occur the human responds to only a few stimuli available and constructs a false representation that has been – consciously or unconsciously – learned (Mitchell, 1980).

On a perceptual level, cognitive biases lead us to draw wrong conclusions and even may trigger perceptual distortion and inaccurate judgment (Gilovich et al., 2002). Since a huge proportion of our perceptions are visual, a lot of cognitive biases occur due to visual stimuli. Most of those biases need some kind of cognitive effort to respond to and evaluate the visual perceptions. For example, the appearance of a black person needs a negative evaluation to lead to prejudice. These biases are learned.

However, there are visual biases that solely rely on the information that is processed in the visual system and do not need a cognitive evaluation of some kind. These biases are innate. One example for this are visual illusions which distort human's perception of reality and reveal insights on how the human brain organizes and interprets stimulation (Eagleman, 2001).

As heuristics are *usually* helpful because they enable an effective way to react to stimuli in an autocorrelated world, visual illusions are the "by-product" of an *usually* effective strategy. As in all illusions, the misperception that is triggered is the exception, not the rule. Visual illusions are 'controlled images' and provoke a wrong perception that show the limits of the visual system (see Figure 1A for an example). But by doing so, they enable insights on how the visual system works when perceiving and interpreting visual stimuli under normal conditions.

Serial Dependence I: Introduction

Once a perception is created we try to replace ambiguity with certainty to make sense of the things we see by interpreting the most probable outcome we anticipate. If the pattern of stimuli we perceive now have been produced by a certain cause 99% of the time, we can expect that this is the cause behind the current perception as well (Vasiliev, 2020). We bet on the familiar as they are a robust estimate of the underlying cause. Relying on the non-accidental is useful because patterns of any kind tend to repeat themselves (Group et al., 2014; Maloney et al., 2005), a behaviour that has developed over human evolution and is mostly innate.

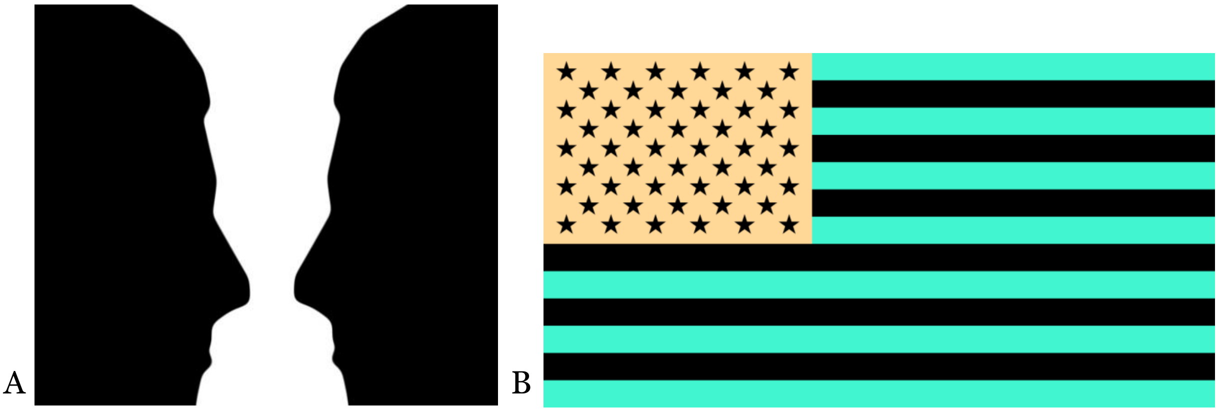
Fischer and Whitney (2014) found out about this process in a very specialized context, a visual heuristic they coined *perceptual serial dependence* (also simply *serial dependence*). This effect describes that both prior and present visual input is used to inform perception at the present moment.

Since the physical world is largely stable over time, serial dependence benefits the visual system: "Physical properties tend to persist over time, making the recent past a good predictor of the present" (Manassi et al., 2021). Because the world is full of autocorrelations, serial dependence offers a practical method to reach a sufficient and immediate visual solution. This effect in the human visual system is found across various cognitive domains such as perception, decision making and memory (Fisher et al., 2020) as well as different features and stimuli such as orientation, position, faces, attractiveness, ambiguous objects and orientation (Manassi & Whitney, 2022).

The perceptual benefit of this effect is that the visual system recycles previously perceived objects and features instead of processing each momentary image in an independent way (Cicchini et al., 2018). A visual heuristic as serial dependence is highly useful to the human regarding two aspects. First, the visual system encounters millions of external and internal sources of noise and irregularities that may disrupt a smooth perception of the world: eye blinks, occlusions, shadows, camouflage and retinal motions. Serial dependence helps the human to perceive a continuous and stable world despite these confounding factors. Second, the autocorrelation of the world makes it useful to perceive stimuli as visual patterns. This makes human beings highly adaptable to recurring situations, even if some aspects vary from one situation to the other. As such, serial dependence is a mechanism for perceptual stability, that is "well suited to meet the delicate balance between the need for sensitivity to change and the need for sensitivity to physical autocorrelations in the visual environment" (Fischer & Whitney, 2014).

Figure 1

Representation of visual phenomena. A. Example of a visual illusion: An ambigious image letting the observer either perceive a vase or two people facing each other. B. Image to induce an afterimage of an American flag.[1]



[1] Stare at the image for 30-60 seconds and then look at a white surface. Blinking a couple of times while looking at the white surface may help you to experience the afterimage. Image source: https://www.illusionsindex.org/ir/negative-afterimages

While it has been shown that that serial dependence cannot be explained by other effects such as priming, explicit memory, expectation or experimental and statistical artefacts (Manassi et al. 2021), certainty does have an impact (Manassi & Whitney, 2022). Perceptual serial dependence is modulated by uncertainty, with increased noise leading to higher serial dependence (Cicchini et al., 2017).

Within perceptual serial dependence, it is distinguished between different kinds of *tuning*. Each tuning effect refers to a different aspect to which the current stimulus is distorted to. There are three kinds of tuning that are discussed in the following: temporal tuning, feature tuning and spatial tuning.

Temporal Tuning

Temporal tuning refers to the time frame in which serial dependence occurs, with the effect gradually decaying over time. Neurobiologically, this can be explained by the effect of adaptation. When there is no change in stimulation, cells in the brain (including photoreceptors on the retina) stop responding (Shapley & Enroth-Cugell, 1984). The reason for this is that the brain only cares about change, because there is no new information for the brain to process if things stay the same.[2] Studies have shown that perceived stimuli were strongly attracted towards a random object, with a time delta ranging up to 12 seconds prior to the perceived stimuli appeared (Manassi et al., 2019). The research in this work will primarily investigate this temporal tuning effect.

Feature Tuning

[2] In the framework of visual illusions, this can be seen in the Troxler effect (also called Troxler's fading). When one fixates on a particular point for even a short period of time, an unchanging stimulus away from the fixation point will fade away and disappear.

Feature Tuning refers to the similarity between stimuli presented consecutively, meaning that serial dependence occurs only between similar features and not between dissimilar ones. 'Feature' in this sense describes visual characteristics that can be grouped into the same category. For this reason, the investigation of similar, and therefore very difficult to classify stimuli will be exploratively investigated in this work.

Spatial Tuning

Spatial Tuning refers to the position of the stimuli presented, meaning that serial dependence occurs only within a limited spatial window. It is strongest when previous and current objects are presented at the same location and gradually decays as the relative distance between both increases (Manassi et al., 2019). Unlike the other two tuning effects, the spatial tuning effect will not be investigated in this work.

Difference to Other Psychological Effects

In order to clearly distinguish perceptual serial dependence from other existing concepts within cognitive science, seemingly related but clearly distinguishable phenomena are discussed in the following.

Negative Aftereffect. An aftereffect (or afterimage) formally described as *Palinopsia* (Bender et al., 1968) is a common effect when looking at a bright light and finding a dark image of the object remaining in the visual field afterwards (see Figure 1B). This experience is caused by a previously seen stimulus, when the stimulus itself is no longer present. In contrast to serial dependence - being a perceptual effect -, negative aftereffects are seen as a neurocellular effect which occurs because some retinal cells (cones) do not respond to the present stimulation as they have been desensitised by the previous stimulus (Bender et al., 1968). Past research has presumed that the effect of serial dependence is dominant if stimulus is shown a short period of time, while the negative aftereffect is dominant when a stimulus is shown a longer period of time (Manassi et al, 2019).

Priming. In contrast to priming, serial dependence actually alters the perception, creating a kind of visual illusion. It actively changes the *perception* of the current stimulus towards the perception of the previous one, while in classical visual priming, the previous stimulus activates a network, which may facilitate the processing of the current stimulus (Kristjánsson & Campana, 2010).

More precisely, “Priming yields an improvement in reaction time and/or discriminability of a repeated stimulus [while] serial dependence can actively reduce the discriminability of simultaneously presented stimuli by altering their appearance” (Fischer & Whitney, 2014).

Change Blindness. Change blindness is a perceptual phenomenon where individuals fail to notice something different about a visual field from one moment to the next (Driver, 1998). As serial dependence enables the perception of a continuous and stable world despite noise and change, does it work against the ability to detect sudden changes? Scientific results suggest that it does not: “When (unnatural) spontaneous and dramatic changes occur in objects and scenes, they often go unnoticed. Perceptual serial dependence may contribute to change blindness by imposing a stability prior on orientation perception, but, crucially, serial dependence is gated by attention and is therefore not responsible for failures to detect changes resulting from inattention” (Fischer & Whitney, 2014).

Gambler’s Fallacy. As a strategy for (irrational) reasoning the gambler's fallacy occurs when humans erroneously believe that a certain event is less likely to happen in the present or future, because it has already happened frequently in the past - or vice versa (Tversky & Kahneman, 1971). For random events, this line of thinking is incorrect, since past events do not change the probability of reoccurrence in the future. Although it seems that serial dependence contradicts the idea of gambler's fallacy, as it presumes that humans believe that the current event is similar to (and not different from) the previous event, this is not the case. The main difference between both concepts is the level of processing which it affects. Serial dependence occurs on a perceptual level, while the gambler's fallacy occurs on a reasoning level, actively altering human judgement and decision-making.

Serial Dependence II: Relevance

Serial dependence can be a helpful mechanism serving the visual system when interacting with an autocorrelated world. But not every real-world setting contains dependent events, especially not if they are "artificially" created, such as laboratory experiments. These experiments conducted in various scientific disciplines build an artificial world in which specific variables are actively manipulated, while all others remain stable, in order to study the pure effect of this specific variable. In such experimental settings, stimuli are randomly ordered, assuming that trials are treated independently by the observer. Yet, such pure situations hardly exist in our natural environment. Since the brain is used to an autocorrelated world, serial dependence may therefore negatively impact the ability to experimentally measure human performance in these situations (Manassi et al., 2019). The same is true for visual processing: The underlying assumption that the visual system treats the current perceptual experience as independent of the previous one is incorrect. For scientific experiments with visual stimuli this has no drastic consequences, except for the possibility of incorrect results. But in real-world settings, where events are truly independent from each other serial dependence can have high costs.

Examples for such settings can be found in the area of medicine and security. When doctors view x-rays of their patients, these are independent, because each image belongs to a completely different person with completely different impairments and symptoms. When security officers do a baggage control check before a flight and view the content of a bag through a scanner these images are independent because each bag belongs to completely different person with completely different items they are carrying. The basic activities behind both tasks are very similar: Both include visual search, which is an active scanning of the visual environment for a particular object or feature (the target) among other objects or features (the distractors) (Eckstein, 2011). It is a crucial task to clarify elementary questions about our daily safety: "Is there a weapon (target) in this bag?" or "Is this person affected by a tumor (target)?"

However, the ability of visual search of radiologists and security officers, like of all human beings, is highly unreliable, as more than 30–35% of mistakes in radiological screening are considered to reflect interpretation errors (Manning et al., 2005). It has been shown, that serial dependence may be one reason for this as it impairs classification performance towards previously seen content (Manassi et al., 2021).

For temporal tuning, it has been shown, that adaptation to previous history up to 60 seconds can affect the ability of medical discrimination away from the past (as in a negative aftereffect), while adaptation to previous history up to 500 ms can strongly bias the ability of medical discrimination towards the past (Kompaniez-Dunigan et al., 2015). More recent results show that classification performance of x-ray scans has been strongly biased towards previously presented objects even up to 12 seconds in the past (Manassi et al., 2019), as the human visual system strongly expects constancy from one moment to the next – regardless if the classification of medical images represents a non-correlated environment.

For feature tuning, it has been shown that within a visual search task search speed for a target is faster if previous and current targets share the same features (Manassi et al, 2019), indicating that humans perform better if features are more similar to one another. Since the human visual system morphs current perceptions with previous perceptions, and trials thus appear more similar to each other than they actually are, serial dependence could be shown to be most prominent between very similar stimuli (Manassi & Whitney, 2022).

Most important, the source of error which serial dependence may hold is not unavoidable (Manassi et al., 2021), making its investigation in the real-world so important. Existing research in this area show that serial dependence is an explanatory approach for systematic errors in a variety of domains, medical misdiagnosis being one of them. This work complements the current state of research on serial dependence in medicine and addresses several open research questions.

1. Investigating the difference between long adaptation to previous history (causing negative aftereffect) and short adaptation to previous history (causing serial dependence). Previous research has stressed the importance of investigating the specific conditions that determine which of the two opposing biases in visual perception occur (Manassi et al, 2019).

2. While most research on serial dependence has been conducted with an experimental design investigating stimulus adjustment, a method where observers recollect and recreate the stimulus as they have previously encountered it, this work uses a binary forced choice task. Although stimulus adjustment represents a

sophisticated scientific instrument that can provide detailed results, the method of a binary forced choice task represents a realistic scenario found in everyday-life.

3. Although serial dependence has been linked to a variety of different visual search tasks in the real-world (as in the airport's security control or medical diagnosis of x-ray scans) this research investigates serial dependence in a new area: Dermatological diagnostics. Its investigation may lie close to the serial dependence studies investigating the effect in x-ray diagnostic, but there are several distinctions that highlight the importance of this study. First, previous studies on serial dependence in medical diagnostics primarily used computer generated x-ray scans, mimicking the visualization of x-rays. Since dermatological images show actual skin irregularities and cannot be as easily mimicked as a technically designed image (such as x-rays), artificial images are not recommended. Therefore, this study uses real-world dermatological images. Second, a detailed investigation of serial dependence in various medical areas is highly important to understand and subsequently minimize diagnostic errors. Developing strategies to overcome visual biases and strengthening the accuracy of medical diagnoses, is crucial to our health. In radiology, the average error rates are estimated around 3–5% on daily basis, with perceptual errors accounting for 60–80% of the total amount (Funaki et al., 1997; Kim & Mansfield, 2014). Third, next to real-world images this work also uses real-world data from real-world medical experts classifying the images. With this, the current work offers data of a very realistic environment, enabling explicit conclusions on the presence of serial dependence in the real world, and thus the presence of a source for diagnostic errors.

While many existing research works rely on experimental, and thus artificially created settings, this research features a real-world setting regarding the data collection, the subject group and the stimuli, which is not yet found in prior research on serial dependence. The evidence in this work is based on a dataset that adequately captures the versatility and necessary differentiation of the world: a real-world dataset with more than 700.000 diagnostic decisions of more than 1000 people with medical background. This immense dataset allows a fine-grained analysis of serial dependence and thus represents a large strength of the current research, as data sets collected from experimental studies in cognitive science are much smaller.

Research Hypotheses

The aim of the current research is to investigate serial dependence in the area of dermatology. While serial dependence has been broadly investigated in perception, decision making and memory (Manassi & Whitney, 2022), its application in the medical area is of specific interest as it is a possible reason for misdiagnosis. Therefore, the following main hypothesis is set:

H1. *Classification performance on any given current dermatological image is biased towards the previous image.*

In accordance with previous literature on serial dependence discussed, this work will first, further investigate the role of certainty, assuming that high uncertainty will likewise lead to high serial dependence as well as low accuracy and second, further investigate the role of reaction time assuming that fast reaction time will lead to high serial dependence. Therefore, the following subhypotheses are set to investigate serial dependence in the context of temporal tuning (H1.1. to H1.4.) and feature tuning (H1.5.):

H1.1. *Difference between Correct and Incorrect answers: Trials of incorrect answers show higher existence of serial dependence, due to higher uncertainty.*

H1.2. *Difference between the first and the second half of answers per user: Trials of the first half show higher existence of serial dependence, due to a learning effect.*

H1.3. *Difference between Low accuracy and high accuracy users: Trials of low accuracy users show higher existence of serial dependence, due to a higher degree of uncertainty.*

H1.4. *Difference between fast and slow reaction time: Trials with a fast reaction time show higher existence of serial dependence.*

H1.5. *Difference between high difficulty and low difficulty images: Trials with high difficulty images show higher existence of serial dependence, due to a higher degree of uncertainty.*

Furthermore, explanatory analyses will be conducted to further investigate the effect of similarity between stimuli on serial dependence.

Methods

In contrast to most research works in the area of cognitive science, the data collection is not done by the author of this report. Instead, the data used in this research is gathered through *DiagnosUs*, an app developed by *Centaur Labs*, a US medical Artificial Intelligence (AI) company based in Boston, MA.[3] *Centaur Labs* uses the answers given by app users to train an AI algorithm in order to correctly label medical images.

Task Procedure

After downloading the app, users can choose between different tasks. For the dermatological classification task, investigated in this study, users first complete a

[3] See https://www.centaurlabs.com/ for more information.

training session with 10 stimuli. This training explains the procedure of the task and prepares users for the actual classification task, which is identical to the training.

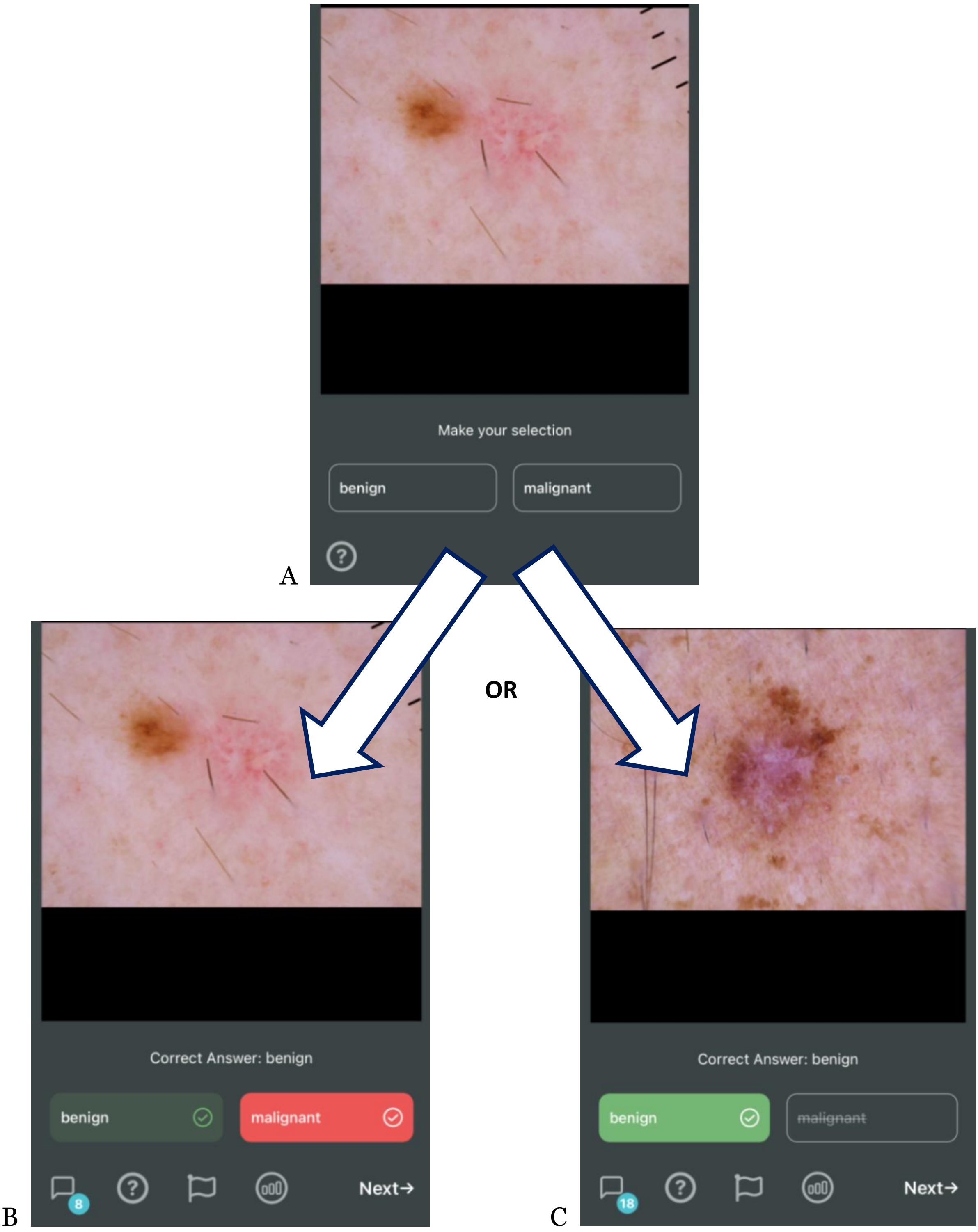
The procedure of the task can be seen in Figure 2. Users are shown a real-world image of a birthmark to be diagnosed. Below the image, they are prompted to choose one of the two possible responses "benign" or "malignant" (Figure 2A). If the answer is incorrect, the correct answer is displayed and marked with a green tick, while the incorrect answer is marked in red (Figure 2B). If the answer is correct, the correct answer is displayed with a green tick, while the incorrect answer is crossed out (Figure 2C). Users can communicate with other users about the image after each decision (lower left corner of Figure 2B and 2C) to foster their learning effect. Afterwards users can move on to the next trial. They can also end the task at any time. Before using the *DiagnosUs* app, users give consent to have *Centaur Labs* use the data they provide through app usage. They receive earnings from a predefined money pool (around US$ 50) for each task they participate in.

Centaur Labs collects 2 million responses per week having gathered over 100 million data points since company foundation. The app is free to download on most smartphone operating systems and contains a variety of medical tasks, mostly classification (e.g. diagnosis of images) or segmentation (e.g. draw around the lesion) tasks. According to *Centaur Labs*, all of the Images used are real-world representations retrieved from either public databases or medical clients. The frequency that binary classification categories are shown are usually balanced, meaning 50% benign and 50% malignant images were used.[4]

[4] This balanced frequency does not exactly match the occurrence in the dataset that was provided to us, which was 57.3% benign to 42.7% malignant (see *Descriptive Results* for more information).

Figure 2

User Journey in the DiagnosUs app. A. Users are prompted to classify the shown image as either "benign" or "malignant”. B. Snapshot if the answer given is incorrect. C. Snapshot if the answer given is correct



Sample

The user group of *DiagnosUs* are mostly medical students, with some medical residents. Individual subject information such as age, sex and other that is gathered in scientific experiments is not known as this information is saved in the user profile of *DiagnosUs*, and was not provided by *Centaur Labs*. However, it is known that all users had normal or corrected-to-normal vision. Since the use of the app does not work outside of the United States, users must be located in the U.S. to the time of app usage.

Results

Data Overview

The dataset used in this work is provided by *Centaur Labs*, a Medical AI Company. The Whitney Lab of Perception and Action at the University of California, Berkeley which whom the author has worked with has permission to use this dataset for scientific purposes. The dataset provided had 758,139 data points across 13 variables which were collected between 04. September 2020 and 21. June 2021. Each data point corresponded to one decision of a user, classifying a dermatological image as either benign or malignant. 7 of the 13 variables are of interest to this research paper and define each data point. They are defined as: *User ID* (defining a unique ID for each user of the app), *score* (defining if the answer given has been correct (100) or incorrect (0)), *response submitted at* (defining at what particular time the response of the user was given), *problem appeared at* (defining at what particular time the image appeared on the device of the user), *origin* (defining the image name of the particular image shown), *current correct answer* (defining if the correct answer is either malignant or benign) and *chosen answer* (defining if the answer given is either malignant or benign).

Data Pre-processing

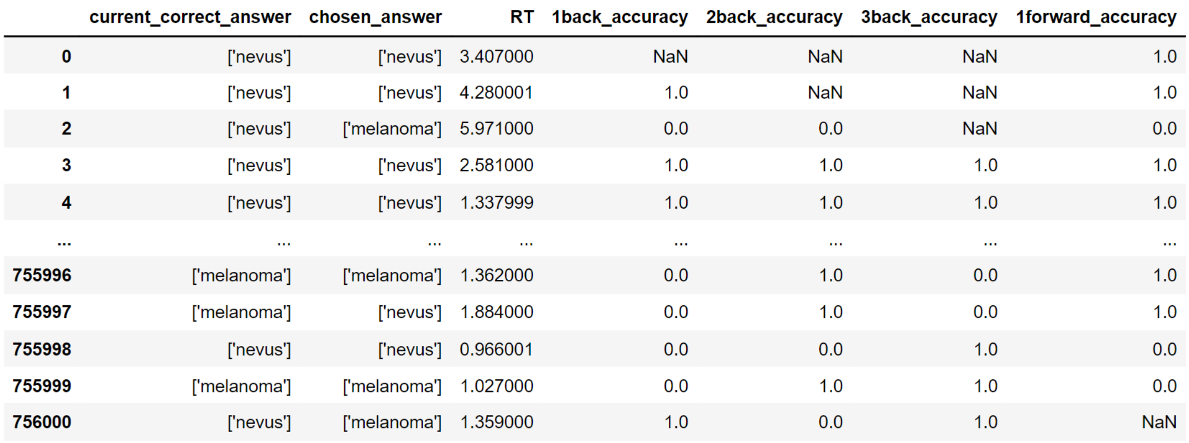
Prior to analyses, the dataset was pre-processed to obtain a clean and prepared dataset for investigation of the before mentioned research hypotheses. All following pre-processing and analysis steps were conducted with Python using Jupyter Notebook (anaconda3). The corresponding code can be found in the appendix.

The following new variables were constructed: *Response time* (defined as the time delta between *response submitted at* and *problem appeared at*, measured in seconds.), *1-back accuracy* (defined as the correspondence between the chosen answer in the current trial and the correct answer in the previous trial), *2-back accuracy* (defined as the correspondence between the chosen answer in the current trial and the correct answer two trials before), *3-back accuracy* (defined as the correspondence between the chosen answer in the current trial and the current correct answer three trials before), *and 1-forward accuracy* (defined as the correspondence between the chosen answer in the current trial and the correct answer one trial ahead[5]). See Figure 3 for a graphical overview of this calculation. For all n-back and 1-forward calculation two values were possible, 1 defining accordance, 0 defining inaccordance. The n-back variables (sometimes also called *n-1, n-2,* etc.) are a common method to investigate serial dependence, while 1-forward serves as a control variable for unrelated biases and potential artifacts and can be directly compared to 1-back (Cicchini et al., 2017; Manassi et al., 2021). Naturally, for each subject the first n trials for n-back and the last trial for 1-forward could not be calculated.

[5] The calculation of the 1-forward accuracy is done to control for unrelated biases and potential artefacts in the analysis that might manifest as spurious serial dependence.

Figure 3

Calculation of n-back: Correspondence between the chosen answer in the current trial and the correct answer one trial (1back_accuracy), two trials (2back_accuracy) and 3 trials (3back_accuracy) before. 1.0 defining accordance, 0.0 defining inaccordance



	current_correct_answer	chosen_answer	RT	1back_accuracy	2back_accuracy	3back_accuracy	1forward_accuracy
0	['nevus']	['nevus']	3.407000	NaN	NaN	NaN	1.0
1	['nevus']	['nevus']	4.280001	1.0	NaN	NaN	1.0
2	['nevus']	['melanoma']	5.971000	0.0	0.0	NaN	0.0
3	['nevus']	['nevus']	2.581000	1.0	1.0	1.0	1.0
4	['nevus']	['nevus']	1.337999	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...	...
755996	['melanoma']	['melanoma']	1.362000	0.0	1.0	0.0	1.0
755997	['melanoma']	['nevus']	1.884000	0.0	1.0	0.0	1.0
755998	['nevus']	['nevus']	0.966001	0.0	0.0	1.0	0.0
755999	['melanoma']	['melanoma']	1.027000	0.0	1.0	1.0	0.0
756000	['nevus']	['melanoma']	1.359000	1.0	0.0	1.0	NaN

Note. Values which could not be calculated are marked with NaN.

The following steps were done to include only valid datapoints in the analyses: First, all datapoints with a larger response time than 3,600s (1 hour) were excluded. As data was collected on a smartphone app, it is assumed that for responses over 1 hour the app was running without users paying attention to it. Second, all remaining data points with a longer response time than three standard deviations of the raw data were excluded, which is a common method to exclude outliers (Miller, 1991). Third, all users with less than 10 trials were excluded to achieve reliable data for the calculation of n-back accuracy. In total, 1,083 data points were excluded due to invalidity.

Descriptive Results

For the analyses 756,001 datapoints from N = 1,137 users were used for further analyses. As the data used is real-world data, users did not all respond to the same number of trials. While the minimum number of trials a user reacted to was 10 (as users completing less than 10 trials were excluded from the analysis), the maximum was 33,739. Results can therefore be calculated on a global level, "Over all datapoints", calculating the mean over all decision trials irrespective of subjects.

Table 1

Descriptive Results "Over all datapoints" and "Over all subjects" each

	Over all datapoints	Over all Subjects
Classified Correct	81.2%	75.7%
Response Time (RT)	1.338 (± 1.977) seconds	1.808 (± 0.793) seconds
1-back accuracy	50.3%	50.2%
2-back accuracy	50.2%	50.2%
3-back accuracy	50.2%	50.3%
1-forward accuracy	50.1%	49.9%

Here each decision trial has the same weight (and each subject a different weight according to the number of trials). Or results can be calculated on a subject level, "Over all Subjects", calculating the mean over all decision trials of each subject and then calculating the grand mean over all subject specific means. Here each subject has the same weight (and each decision trial a different weight), as seen in *Table 1*.

Temporal Tuning Analyses

Since the independence of data cannot be assumed due to the hypothesis of serial dependence, parametric tests are not recommendable. Instead, a non-parametric Wilcoxon signed-rank test (WSR) is conducted. This test investigates whether the central tendencies of two dependent samples are different and is used when the requirements for a t-test for dependent samples are not met.

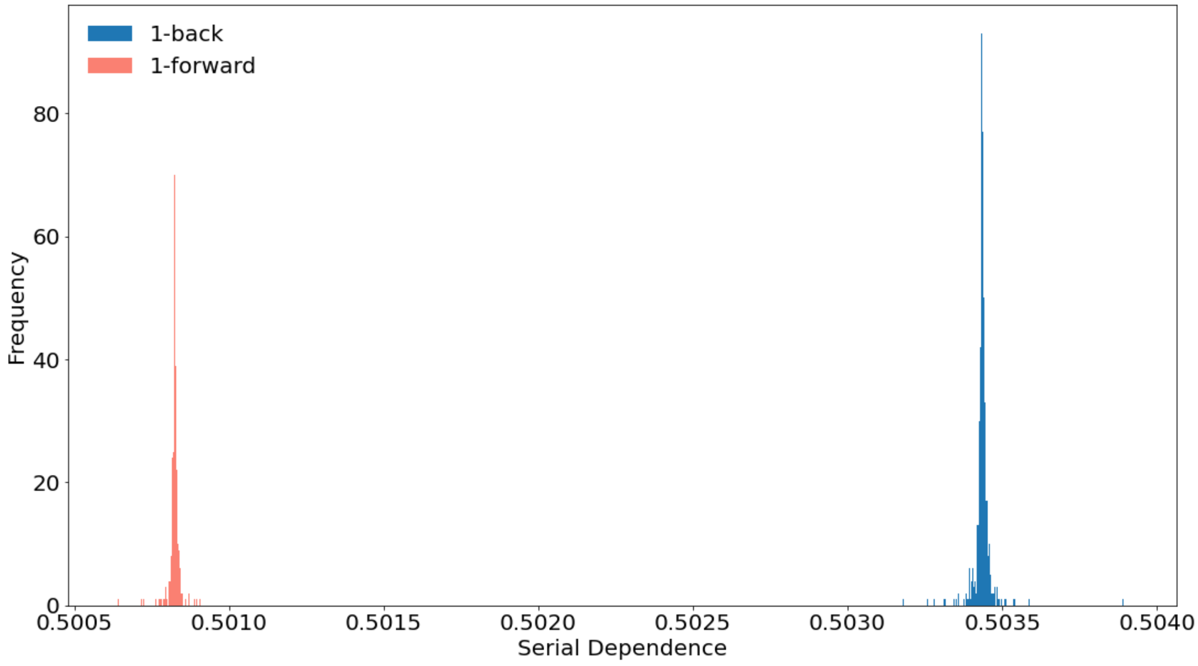
To investigate H1., assuming that classification performance on any given current dermatological image is biased towards the previous image, it is tested whether the descriptive differences between 1-back and 1-forward over all data points are significant. The data investigated in this hypothesis are dependent, as the only difference between 1-back and 1-forward is the direction of cause (1-back indicating the correspondence between the current and previous trial and 1-forward indicating the correspondence between the current and next trial), with 1-back values eventually capturing additional serial dependence.

The Wilcoxon test yielded a significant result: Differences in percentage between 1-back and 1-forward are significantly different from another, meaning that classification performance on any given current dermatological image is biased towards the previous image ($W=0.0$, $p < .001$). Figure 4 shows this difference with a Histogram using *Leave-one-out cross validation (LOO-CV)*, a popular resampling method in machine learning to estimate and depict the accuracy of a statistical model (Wong, 2015). In the current work, this method does not just show the robustness of the average value for 1-back and 1-forward but also reveals that there is no specific individual or subgroup that carries this serial dependence effect, as each distribution appears very narrow. It furthermore illustrates that values of both samples are highly unlikely to come from the same distribution, as both distributions lie – relatively far away from each other and no overlap in values can be seen.

As most hypotheses in this work rely on paired data, a test for dependent samples is needed, such as the WSR calculated for H1. However, this test also requires a balanced dataset for calculation (the same sample size for both to be tested

Figure 4

Histogram using Leave-one-out cross validation (LOO-CV): The average 1-back (blue) and 1-forward (red) value is calculated with one subject left out each time. This results in 1173 values each showing the "true mean" of the given variable and investigating whether outliers are carrying the effect



variables must be equal). While this is true for H1, this is not the case for e.g. H1.5., investigating differences between high difficulty images (91.633 images in dataset) and low difficulty images (66.2318 images in dataset). However, since all hypotheses do not necessarily investigate subject-based differences (as done by dependent tests), but group-based differences (as done by independent tests) a Mann-Whitney-U-Test, examining whether the median of two samples are different will be conducted. This variation in analysis is acceptable, if it is taken into account when interpreting the results. Therefore, to further investigate 1-back serial dependence in the dataset, hypotheses H1.1. to H1.5. are investigated with a Mann-Whitney-U-Test[6]. This is done by first calculating the 1-back accuracy per subject and hence comparing the median value of the respective groups.

H1.1. investigated differences between correct and incorrect answers, assuming that trials of incorrect answers show higher existence of serial dependence. The Mann-Whitney-U-Test yielded a significant result, but in the other direction than assumed in the hypothesis: Correct trials show higher serial dependence than incorrect trials (*U* =682,136*, p = 0.017*). The median 1-back serial dependence value for incorrect trials resulted in 50.0% (correct trials: 50.4%), indicating that incorrect trials show lower 1-back serial dependence than calculated over all data points (50.3%) and over all subjects (50.2%; see *Table 1*).

H1.2. investigated differences between the first and the second half of answers per user, assuming that trials of the first half show higher existence of serial dependence. The Mann-Whitney-U-Test yielded a non-significant result: Trials of the first half do not differ from trials of the second half in serial dependence (*U = 628,596, p= 0.256*). The median 1-back serial dependence value for the first half resulted in 50.0% (second half: 50.4%), indicating that the first half shows lower 1-back serial dependence than calculated over all data points (50.3%) and over all subjects (50.2%; see *Table 1*).

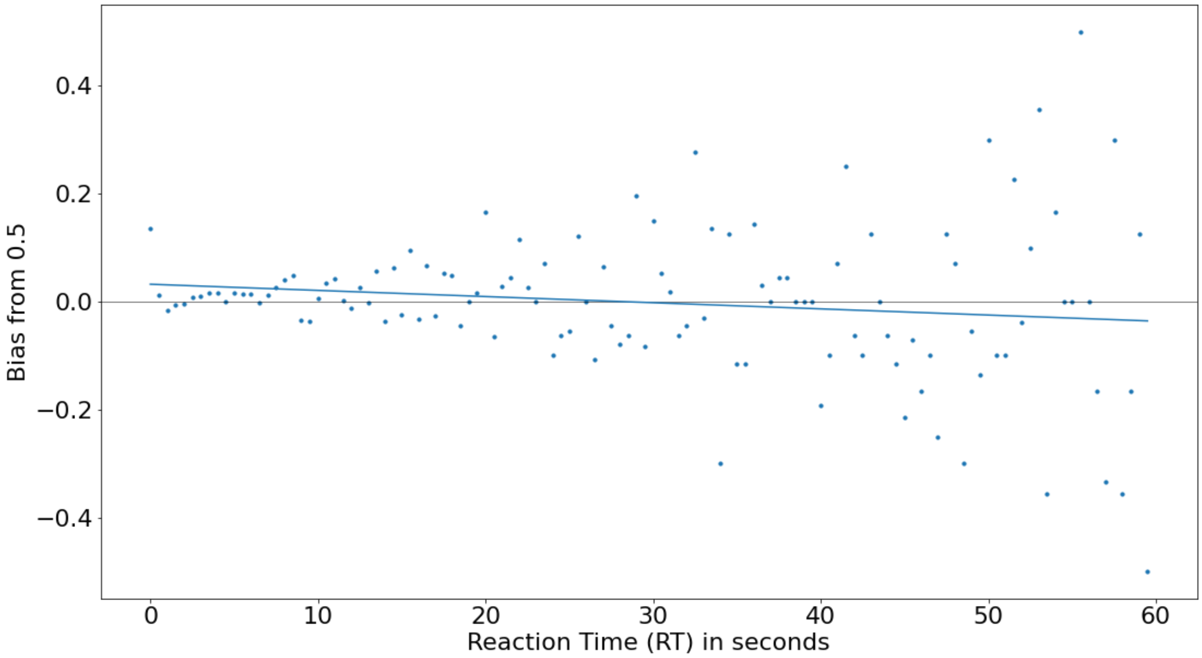
H1.3. investigated differences between low accuracy and high accuracy users, assuming that trials of low accuracy users show higher serial dependence. The Mann-

[6] Analysis results will show a very large test statistic value which is a normal consequence of the huge data set, as U is related to the sample sizes (Conroy, 2018).

Whitney-U-Test yielded a non-significant result: trials of low accuracy users do not differ from trials of high accuracy users in serial dependence (*U = 158,761, p= 0.719*). The median 1-back serial dependence value for low accuracy users resulted 50.4% (high accuracy: 50.1%), indicating that low accuracy users show marginally higher 1-back serial dependence than calculated over all data points (50.3%) and over all subjects (50.2%; see *Table 1*).

Figure 5

Relation between Reaction Time (RT) in seconds (x-axis) and 1-back serial dependence (y-axis): Trials with a fast RT show higher existence of serial dependence than trials with a slow RT. Blue line shows the fit of a linear regression. RT values are binned for each 0.5 seconds



Note. Binned values over 60 seconds could not be calculated due to an insufficient number of values.

H1.4. investigated differences between fast and slow reaction time, assuming that trials with a fast reaction time show higher serial dependence. The Mann-Whitney-U-Test yielded a significant result: Trials with a fast reaction time show higher serial dependence than trials with a slow reaction time (*U =353,707, p < .001*). The median 1-back serial dependence value for trials with a fast reaction time resulted 54.5% (slow reaction time: 48.7%), indicating that trials with a fast reaction time show clearly higher serial dependence than calculated over all data points (50.3%) and over all subjects (50.2%; see *Table 1*). Figure 5 shows this significant dependence indicating that users with fast RT (left) show a higher bias from random (0.5) and hence higher serial dependence than users with slow RT (right).

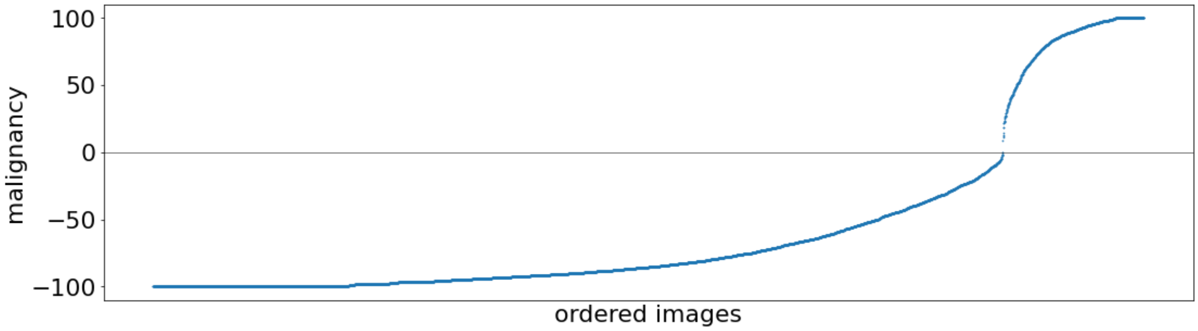
H1.5. investigated differences between high difficulty and low difficulty images, assuming that trials with high difficulty images show higher serial dependence. The Mann-Whitney-U-Test yielded a non-significant result: Trials with high difficulty images do not differ from trials with low difficulty images in serial dependence (*U* = 612,423*, p= 0.164*). The median 1-back serial dependence value for high difficulty resulted 50.0% (low difficulty: 50.3%), indicating that trials with high difficulty images show lower 1-back serial dependence than calculated over all data points (50.3%) and over all subjects (50.2%; see *Table 1*).

Feature Tuning Analyses

To exploratively investigate feature tuning and the effect of similarity within the serial dependence effect, the following steps were done: First, images were ranked according to the malignancy (Figure 6). Values of 100 indicate that all users have classified a given image as malignant, values of -100 indicate that all users have classified it as benign. Consequently, images with a more equal proportion of benign and malignancy classifications tend towards malignancy value of 0. These images are very similar and therefore very difficult.

Figure 6

Overview of all 7,798 (6,688 benign, 1,110 malignant) unique images used, sorted by their malignancy value (-100: classified as benign by all users, 100: classified as malignant by all users)

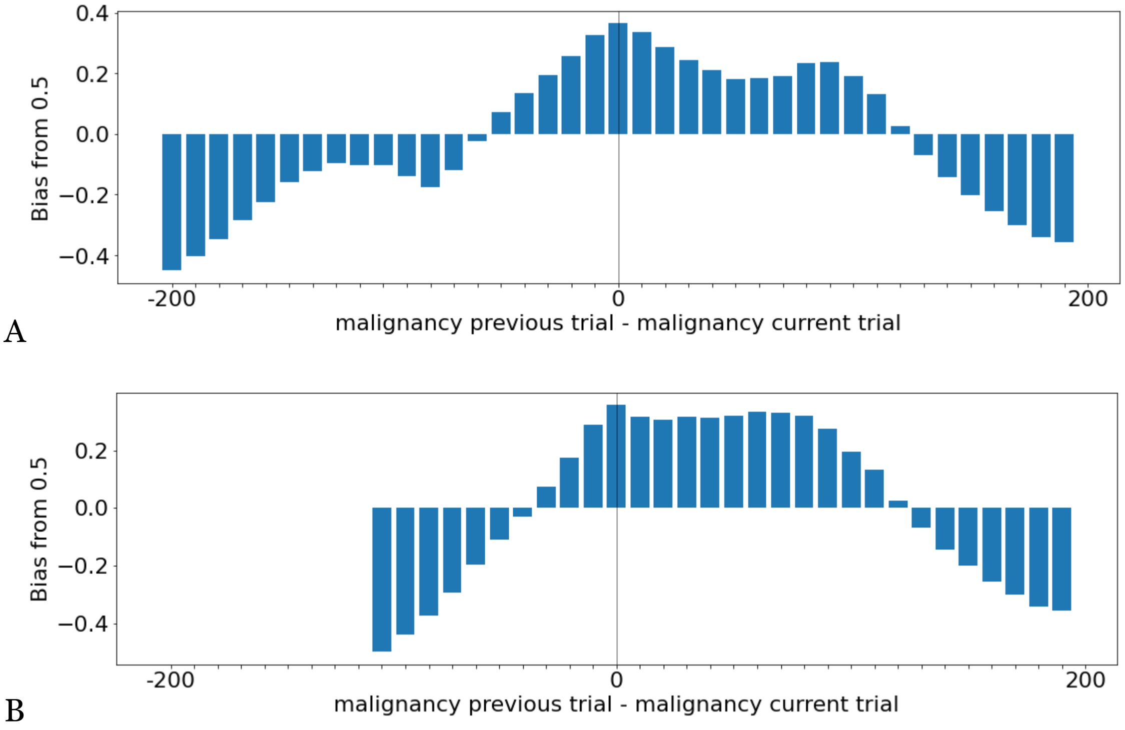


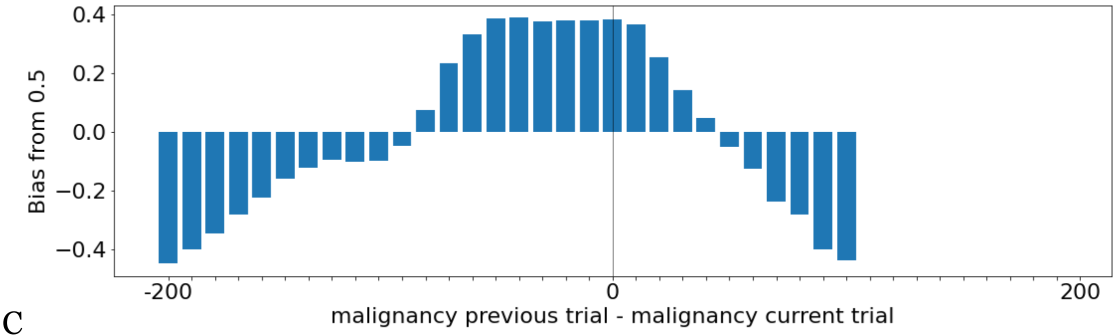
With regard to their true label, 6,688 unique images were benign, 1,110 unique images were malignant. Since it was aimed to achieve an equal frequency of images shown to users, malignant images were presented more often (291 times on average) than benign images (65 times on average).

Subsequently, serial dependence was calculated for different levels of similarity (binned for 10 malignancy units each), as well as different scenarios (for all trials, for benign images in current trial only and for malignant images in current trial only; see Figure 7). Overall, serial dependence is high if the malignancy value of the current and previous image is similar (difference is close to zero, Figure 7A). This could be found for both benign and malignant images *(*Figure 7B and 7C). This supports feature tuning, describing that serial dependence occurs only between similar features and not between dissimilar ones. However, note that if the current image has approximately the same malignancy value as the previous, it does most likely belong to the same category as the previous, while if the current image has a

Figure 7

Bias towards previous image (y-axis) for different levels of similarity (difference of malignancy between previous and current image; x-axis). A. Calculation for all trails. B. Calculation for benign images in current trial only. C. Calculation for malignant images in current trial only. X-axis value of -200 indicates that the previous image was very benign, value of 200 indicates that the previous image was very malignant. Y-axis value of -0.4 indicates a bias away from the previous image, value of 0.4 indicates a bias towards the previous image (serial dependence)





C

Note. Difference in malignancy values are binned for each 10 units. Values for -200 to -100 in Figure 7B and 100 to 200 in Figure 7C could not be calculated due to an insufficient number of values.

totally different malignancy value than the previous it does most likely belong to the opposite category than the previous. This logical consequence must be taken into consideration when interpreting the results.

Furthermore, Figure 7B and 7C do not distribute evenly around 0 but show a clear tilt to the right and left side, respectively. It is assumed this may be due to an unbalanced amount of benign and malignant images shown. Future research must further investigate this matter.

Discussion

The structure of the visual system and the process from perceiving a visual stimulus to understanding and acting upon it is complex. The visual system uses visual heuristics, such as serial dependence, to save energy and time. In general, a useful procedure to find one's way in an autocorrelated world, we interact with day to day. But at the same time a source of error in areas where events are fully independent, such as medical diagnostics (Manassi et al. 2021). Since biases of any kind are detrimental to guarantee a suitable medical treatment (Blumenthal-Barby & Krieger, 2015) it is crucial to investigate these kinds of effects and offer strategies to minimize them. This work has investigated temporal tuning and feature tuning within the serial dependence effect in dermatological image classification.

Interpretation of Results

Descriptive results show marginal effects of 1-back, 2-back and 3-back compared to 1-forward, supporting the assumption of serial dependence. An accuracy of 81.2% also showed that users were very skilled.

The main hypothesis of this research H1., assumed that classification performance on any dermatological image is biased towards the previous image. A Wilcoxon signed-rank test was conducted to investigate differences between 1-back and 1-forward over all data points showing a significant result. This indicates that the descriptive differences in percentage between 1-back and 1-forward (see *Table 1*) are significantly different from another, and serial dependence is therefore present in the current dataset.

For the five subhypotheses H1.1. to H1.5. two yielded significant results, however, with one (H1.1. incorrect vs. correct) showing significant results in the opposite direction of assumption. This indicates that the data investigated in this work shows patterns that contradict the postulated hypotheses and previous research results regarding serial dependence. More precise, correct trials showed higher serial dependence than incorrect. H1.4. (fast reaction time vs. slow reaction time) yielded a significant result according to the hypothesis, indicating that trials with a fast reaction time showed higher existence of serial dependence than trials with a slow reaction time. Since the Mann-Whitney-U-Test was used, the slow and fast reaction time groups were treated as independent instead of dependent groups (see *Results* for a justification of this choice). Therefore this significant result can only be interpreted on a group level, not on a subject-specific level. Real-world implication of this result that offer strategies for reducing serial dependence can be found at a later point in this work.

Over all, the results show serial dependence in a realistic dataset with more than 700.000 diagnostic decisions of more than 1000 people with medical background. A more detailed analysis of serial dependence regarding attributes such as difficulty, reaction time and accuracy showed mixed results, sometimes even indicating contrary results to theory-driven hypotheses.

For temporal tuning, referring to the time frame in which serial dependence occurs, with serial dependence gradually decaying over time, this work offers clear results: Both a general investigation of 1-back vs. 1-forward as well as a fine-grained investigation of reaction time on serial dependence was significant. On a descriptive level, this can be also seen by higher values for 1-back, as compared to 2-back or 3-back. As there are a variety of options on how past stimuli can influence decision making in the present moment, future research can include calculating a *time series forecasting*, an analysis model to predict future choices based on previously observed labels (Chatfield, 2000), such as the past five images seen.

For feature tuning, referring to the similarity between stimuli presented consecutively, meaning that serial dependence occurs only between similar features and not between dissimilar ones, the data shows mixed results. While a visual analysis showed high serial dependence if the malignancy value of the current and previous image is similar (see Figure 8), this also is a logical consequence if current and previous image belong to the same category. Future research can more specifically investigate this dependence, e.g. by only taking trials into account where the category of the current image is different from the category of the previous image. Furthermore, advancing the current feature analysis to investigate how the similarity with previous affects serial dependence would lead to more detailed insights on how to minimize the occurrence of feature tuning within serial dependence.

Strategies for Overcoming Serial Dependence

Due to the benefit of a realistic dataset, this study offers opportunities to give advice on how to minimize and eventually eliminate serial dependence in medical diagnostics.

Mainly, pay full attention to the visual stimuli at hand. Manassi et al. (2019) have found that guidance plays a crucial role: "by focusing their attention on spotting tumors in x-ray images, radiologists were shown to miss objects as salient as gorillas" (Manassi et al, 2019). Especially for highly unlikely targets, this is hugely important, as the prevalence of target occurrence plays an important role in visual search, since target misses strongly increase with decreasing target prevalence (Wolfe et al., 2007). In the before mentioned tasks found in medicine and security, malignant tumors (target) for radiologists or weapons (target) during baggage control are rarely encountered. In this case, it has been found that serial dependencies may systematically bias human perception towards more frequent objects. Consequently, recognizing rare targets is even more difficult (Manassi et al, 2019). Therefore, attention must be paid to the true base rate of the specific classification task in the real-world. In the area of dermatology, a hint of the true base rate of malignant images can be retrieved from the number of unique images that are used in the app. In total, 85.8% of the images are benign and 14.2% malignant. Since the images used are real-world images retrieved mainly from data-bases, the true base-rate of malignancy is estimated to be around 10 - 15%, and therefore lies below the rate of malignant images occurring in the classification task.

Taking time to correctly classify each visual percept is crucial. As results in this work have shown, serial dependence is most likely to occur when reaction time was below 30 seconds (see Figure 5). At least this amount of time should be used to make each decision on visual classification. After all, these decisions may be a crucial task in certain jobs to clarify very important questions about our health and safety.

Strengths and Limitations of the Study

The present work has a variety of strengths, while, at the same time, limitations and possible improvements for further research studies must be taken into account.

First, the biggest strength of this study is the use of a large dataset. With over 700.000 data points the analyses give robust results, and allows for the analysis of subgroups within the dataset still offering reliable results. In usual datasets of psychological studies, this is not possible due to a much smaller sample.

Second, the data used actually represents the real-world, in multiple ways. The data is not gathered in an artificial experimental setting but in a “natural” environment of users; their smartphone. The images used are not artificially created stimuli mimicking a certain representation, such as simulated lesions (Manassi et al, 2019). Instead, they are retrieved from public databases and represent actual dermatological images from benign and malignant birthmarks. At last, the sample of observers are actual medical professionals. While most psychological studies use student samples leading to questionable conclusions for the real-world (Mullinix et al., 2015), the app is used by a highly specified and qualified group of medical students and medical professionals. This has a direct influence on the interpretation of the results: Since this study investigates serial dependence as a possible cause of misdiagnosis, the realistic conditions can be directly applied to the conditions in everyday medical practice.

Third, the scientific cooperation with Centaur Labs shows the positive impact that can be achieved if corporates and science labs collaborate. Corporates gain through the scientific analysis and publications for both scientific and societal purposes while science gains through the access to a dataset providing valuable scientific insights with highly generalizable inferences.

Fourth and finally, this study pinpoints the real-world relevance a scientific work can have: Scientific evidence that informs and shows deeper insights on how crucial errors in everyday life - such as medical misdiagnosis - occur and what can be done to minimize them.

A limitation of this work, which goes hand in hand with the benefit of using external data, is the impossibility to co-design the task setup. An accurate procedure of the task is important to obtain clean, usable data (Smucker et al., 2018). Since neither the setup of the task nor the data gathering was done or supervised by the author of this research, it can only be relied on the data provided by Centaur Labs. This is particularly evident in two areas: The environmental conditions and the statistical conditions.

First, under normal experimentally conducted studies, environmental conditions are kept constant to minimize confounding variables and contextual effects (Smucker et al., 2018). However, there is no guarantee of the same conditions when obtaining data in this dataset. Both external conditions (e.g., light) and internal conditions of the user (e.g., motivation, fatigue) can be completely different between users, but also within the data of one user, weakening the robustness of their data.

Second, under normal experimental conditions, statistical symmetry is assured in order to keep the analysis of data and their interpretation as straightforward as possible. For example, subjects are divided equally between two experimental groups, or they all complete the same number of trials.[7] This allows for balanced data which is crucial for numerous statistical analyses (Ince et al., 2022) and has also complicated the analyses done in this work. Statistical symmetry is not present for the data at hand. With 57.3% benign and 42.7% malignant images, the occurrence of each class did not match 50.0% which brings difficulties to the analyses frequently used for serial dependence (as in Manassi et al., 2021). This may also be the reason for marginal effects above 50.0 for 1-forward values, as behavioural effects from users can be excluded.

Third, a 57:43 ratio of occurrence of each class does not match the base rate in the real world either, making it difficult to interpret the results, as the medical professionals undergoing the task might unconsciously behave as in “realistic” base rates occurring in the real world, estimated to be around 10-15% for malignant images.

Fourth and finally, the number of images users classified varied greatly: The minimum number was 10 (as all users completing less than 10 trials were excluded), the maximum was 33.739. With this, each user in the dataset contributed with a different extent to the average value of serial dependence for the whole dataset, which is the reason why descriptive results can either be calculated for ”Over all datapoints” or “Over all subjects”.

[7] If subjects decide to end the experiment on their behalf, their data is usually not used in the analysis.

Working with a real-world dataset offers many benefits, but also contains real-world complexity. While data gathered in experimental settings usually show strong effects due to the experimental design that has minimized confounding variables and maximized the effect of variables showing the phenomenon of interest (Smucker et al., 2018), this is not the case for real-world data sets. Real-world data contains noise with underlying patterns being covered up by multidimensional effects. For the investigation of serial dependence in this dataset, the same is true. Due to the extensive size of the dataset there are a variety of analysis options to consider, with each having possible explanations for an existing serial dependence effect. Not all possibilities of statistical analyses can be considered in this work but offer great opportunities for further investigation.

Future Research

Next to ideas that have already been addressed within the discussion of temporal and feature tuning, subsequent future research could include the use of sophisticated models to investigate Computer Vision in the area of medical images to gain further methodological insights. Next to methodological research in the area of medical image generation (Ren et al., 2022) other models could include the extraction of specific image features to predict the class they belong to, such as either benign or malignant (Lu & Weng, 2007).

It could also address different kinds of methods used in the area of cognitive science. These may include eye-tracking to investigate the attentional focus when classifying medical images which gives insights on which features are most important for classification (also improving computer vision) or even cause serial dependence. This investigation may shed light on different questions: Which aspects in an image are giving humans information that is important for a decision? Which features are guiding the attention when making a medical diagnosis? A possible hypothesis could be that experts (such as dermatologists or radiologists) have years of experience and learned which features are specifically important such that they automatically know where to attent to. Conversely, a novice (first year medical students or lay persons) may use an alternative approach such as a heuristic like serial dependence, to come to a conclusion. Furthermore, Fischer and Whitney (2014) have found that serial dependence causes observers to be less accurate in their perception of object properties offering an ideal starting point to further investigate this matter with eye-tracking methods. Also other methods such as EEG or fMRI can be used to investigate neural correlates of the occurrence of serial dependence.

Conclusion

Taken together, the theoretical presentation of previous research on perceptual serial dependence and the statistical conducted analyses on real-world data has given important insights on the circumstances that cause serial dependence as well as methods on how to minimize this source of errors in medical image classification. To minimize this perception bias, results indicate to achieve a more elaborate, slow reaction time in order to minimize serial dependence.

Acknowledgments

The research behind this report would not have been possible without the support of many people: Thanks to the welcoming members of the Whitney Lab at University of California, Berkeley, who made me feel like home during my time at the Bay area. I’m very happy to have worked with you on this research, thank you Professor David Whitney and “Peter” Zhihang Ren.

Thanks also to the University of Cologne for providing me with the financial means to have had the chance to visit UC Berkeley and work on this project, the same is true for the Fulbright Commission.

References

Adams, K. A., & McGuire, E. K. (2022). *Research methods, statistics, and applications*. Sage Publications.

Bender, M. B., Feldman, M., & Sobin, A. J. (1968). Palinopsia. *Brain: a journal of Neurology*, 91(2), 321–338. https://doi.org/10.1093/brain/91.2.321

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making*, *35*(4), 539-557. https://doi.org/10.1177/0272989X14547740

Bock, T. (n.d.). *What is Autocorrelation?*. Display R Blog. Retrieved August 29, 2022, from https://www.displayr.com/autocorrelation/

Chatfield, C. (2000). *Time-series forecasting*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420036206

Cicchini, G. M., Mikellidou, K., & Burr, D. C. (2018). The functional role of serial dependence. *Proceedings of the Royal Society B*, *285*(1890), 20181722. https://doi.org/10.1098/rspb.2018.1722

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017). Serial dependencies act directly on perception. *Journal of vision*, *17*(14), 6-6. https://doi.org/10.1167/17.14.6

Conroy, R. M. (2018). Re: What is the significance of large U value? Retrieved from: https://www.researchgate.net/post/What_is_the_significance_of_large_U_value/5adce4321a5e768dae6cdcb8/citation/download.

Driver, J. (1998). The neuropsychology of spatial attention. *Attention*, 297-340.

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*(12), 920-926. https://doi.org/10.1038/35104092

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of vision*, *11*(5), 14-14.

https://doi.org/10.1167/11.5.14

Fischer, C., Czoschke, S., Peters, B., Rahm, B., Kaiser, J., & Bledowski, C. (2020). Context information supports serial dependence of multiple visual objects across memory episodes. *Nature communications*, *11*(1), 1-11. https://doi.org/10.1038/s41467-020-15874-w

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature neuroscience*, *17*(5), 738-743. https://doi.org/10.1038/nn.3689

Funaki, B., Szymski, G. X., & Rosenblum, J. D. (1997). Significant on-call misses by radiology residents interpreting computed tomographic studies: perception versus cognition. *Emergency Radiology*, *4*(5), 290-294. https://doi.org/10.1007/BF01461735

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Group, T. M. A. D., Fawcett, T. W., Fallenstein, B., Higginson, A. D., Houston, A. I., Mallpress, D. E., ... & McNamara, J. M. (2014). The evolution of decision rules in complex environments. *Trends in cognitive sciences*, *18*(3), 153-161. https://doi.org/10.1016/j.tics.2013.12.012

Haber, R. N. (1978). Visual perception. *Annual review of psychology*. https://doi.org/10.1146/annurev.ps.29.020178.000335

Ince, R. A., Kay, J. W., & Schyns, P. G. (2022). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.05.008

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kim, Y. W., & Mansfield, L. T. (2014). Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. *AJR Am J Roentgenol*, *202*(3), 465-470. https://doi.org/10.2214/AJR.13.11493

Kompaniez-Dunigan, E., Abbey, C. K., Boone, J. M., & Webster, M. A. (2015). Adaptation and visual search in mammographic images. *Attention, Perception, & Psychophysics*, *77*(4), 1081-1087. https://doi.org/10.3758/s13414-015-0841-5

Kristjánsson, Á., & Campana, G. (2010). Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, & Psychophysics*, *72*(1), 5-18. https://doi.org/10.3758/APP.72.1.5

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current biology*, *24*(21), 2569-2574. https://doi.org/10.1016/j.cub.2014.09.025

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), 823-870. https://doi.org/10.1080/01431160600746456

Maloney, L. T., Martello, M. D., Sahm, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences*, *102*(8), 3164 3169. https://doi.org/10.1073/pnas.0407157102

Manassi, M., Ghirardo, C., Canas-Bajo, T., Ren, Z., Prinzmetal, W., & Whitney, D. (2021). Serial dependence in the perceptual judgments of radiologists. *Cognitive research: principles and implications*, *6*(1), 1-13. https://doi.org/10.1186/s41235-021-00331-z

Manassi, M., Kristjánsson, Á., & Whitney, D. (2019). Serial dependence in a simulated clinical visual search task. *Scientific reports*, *9*(1), 1-10.

https://doi.org/10.1038/s41598-019-56315-z

Manassi, M., & Whitney, D. (2022). Illusion of visual stability through active perceptual serial dependence. *Science advances*, *8*(2), eabk2480. 10.1126/sciadv.abk2480

Manning, D. J., Gale, A., & Krupinski, E. A. (2005). Perception research in medical imaging. *The British journal of radiology*, *78*(932), 683-685. https://doi.org/10.1259/bjr/72087985

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, *43*(4), 907 912. https://doi.org/10.1080/14640749108400962

Mitchell, T. M. (1980). *The need for biases in learning generalizations* (pp. 184-191). New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ..

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109-138. https://doi.org/10.1017/XPS.2015.19

Ren, Z., Stella, X. Y., & Whitney, D. (2022). Controllable Medical Image Generation via GAN. *Journal of Perceptual Imaging*, *5*, 1-15. https://doi.org/10.2352/J.Percept.Imaging.2022.5.00050

Shapley, R., & Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. *Progress in retinal research*, *3*, 263-346. https://doi.org/10.1016/0278-4327(84)90011-7

Smucker, B., Krzywinski, M., & Altman, N. (2018). Optimal experimental design. *Nat. Methods*, *15*(8), 559-560. https://doi.org/10.1038/s41592-018-0083-2

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers.

Psychological bulletin, *76*(2), 105. https://doi.org/10.1037/h0031322

Vasiliev, I. R. (2020). Visualization of spatial dependence: an elementary view of spatial autocorrelation. In *Practical handbook of spatial statistics* (pp. 17-30). CRC Press.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of experimental psychology: General*, *136*(4), 623. https://doi.org/10.1037/0096-3445.136.4.623

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839-2846. https://doi.org/10.1016/j.patcog.2015.03.009