Evaluating the Effect of Intermittent Reinforcement on Concept Learning in Canine Lung cancer Detection

A thesis

Submitted in partial fulfilment

of the requirements for the degree

of

Master of Applied Psychology in Behaviour Analysis

at

The University of Waikato

by

Linguo Ji

March 9, 2026

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2026

Abstract

Dogs have demonstrated the ability to identify a range of human diseases, including lung cancer, through olfactory analysis of biological samples. Although exhaled breath is a non-invasive and accessible sample type, no single volatile organic compound has been reliably identified as a biomarker. Therefore, the process of identifying lung cancer in human implies a process of concept learning, suggesting dogs' detection of lung cancer may be relying on the subjects' ability to identify a complex, highly variable pattern of volatile organic compounds (VOCs) in exhaled breath. Concept learning appears to be a special form of generalization, with underlying mechanisms in common with those responsible for perceptual concept learning in the visual domain. While intermittent reinforcement is commonly recommended in scent-detection training to simulate and prepare for operational conditions where reinforcement is not always possible, its effects on conceptual generalization remain poorly understood. This study investigated the effects of intermittent reinforcement on canine concept learning in a lung cancer detection task. Five dogs were trained using a fully automated 17-segment carousel apparatus with breath samples collected from 348 patients who visited respiratory clinic, with 115 tested positive and 233 tested negative in lung cancer. A single-subject reversal design was employed; the reinforcement rate for correct indications to positive samples was systematically thinned from 100% down to ranges of 80% and 60%. The findings demonstrated that thinning the reinforcement schedule to a minimum of 60% did not exert a significant disruptive effect on the dogs' diagnostic accuracy. In addition, an exploratory probe test also provided preliminary evidence that the dogs could successfully differentiate between lung-originated cancer and non-lung-originated (NLO) cancer samples.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Tim Edwards. Thank you for your support, patience, and guidance throughout this research journey, also for always respecting my opinions and views on certain issues. This project simply would not have gone this far without your mentorship. A special thank you must also go to our technician Rob Bakker. Thank you for keeping the apparatus running smoothly and for your relentless behind-the-scenes problem-solving.

I would like to thank wonderful families of our participant dogs for your extraordinary commitment. Consistently bringing your lovely dogs into the lab week after week, all the way from March to December. To the dogs themselves, Tui, Memphis, Mac, Harlee, Tommy, Ben and Bayley, thank you for your hard work, sharp noses, and endless enthusiasm.

Finally, I want to dedicate a very special note of remembrance to one of our incredible participant dogs, Tommy, who sadly passed away from lymphoma on December 30, 2025. Thank you, Tommy, for your amazing contribution to this lab, you are deeply missed.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

Dogs are renowned for their exceptional olfactory abilities. While this trait has historically been utilized for hunting and tracking, a growing body of empirical evidence suggests that these olfactory abilities can be employed as reliable tools for medical diagnosis (Jendrny et al., 2021). Numerous existing studies have found that dogs can detect various stimuli associated with human diseases, such as viruses, bacteria, and cellular abnormalities (e.g., tumours).

For instance, in the detection of infectious diseases, Guest et al. (2019) demonstrated that dogs could identify malaria parasites by sniffing nylon socks worn by infected children, successfully distinguishing them from those worn by uninfected controls, with sensitivity (the ability to correctly identify positive cases) of 72% and specificity (the ability to correctly identify negative cases) of 99%. In the field of neurodegenerative disease, recent research has shown that dogs can detect the specific odour signature of Parkinson's disease from sebum samples (Rooney et al., 2025). Perhaps the most famous early anecdotal evidence involved a dog persistently sniffing a mole on its owner's leg, which was later confirmed to be a malignant melanoma (Williams & Pembroke, 1989).

There has been a marked acceleration of research activity in this field. Edwards et al. (2017) previously noted a "steepening trajectory" of publications, indicating a significant growing interest in the topic up to 2016. In recent years, this interest has matured from a sole academic topic into a serious, applied field. The COVID-19 pandemic, in particular, prompted formal evaluations by

major global health bodies. Notably, the World Health Organization (WHO, 2021) has held consultations on the feasibility of deploying trained dogs for mass public health screening. Within this expanding field, a significant and promising application is the detection of lung cancer using human breath samples.

1.2 Canine in Lung Cancer Detection

Lung cancer remains the leading cause of cancer-related mortality worldwide and is particularly pronounced in developing nations due to delayed diagnoses and unequal access to healthcare (Barta et al., 2019). Early-stage lung cancer is frequently overlooked due to lack of specific symptoms early on; current screening methods are facing challenges in reliably distinguishing between cancerous and healthy tissue, and these methods also impose invasive procedure or significant radiation exposure on patients (Hackner et al., 2016). There is a critical need for non-invasive, cost-effective, and time-efficient screening alternatives. Over the past two decades, the use of canine olfaction for lung cancer detection has emerged as a promising diagnostic tool (see Table 1).

Table 1

Studies of Canine in lung cancer detection

Author	Sample	Sample Size	Result
McCulloch et al. (2006)	Breath	n = 169	Sensitivity = 99% Specificity = 99%
Ehmann et al. (2012)	Breath	n = 220	Sensitivity = 71% Specificity = 93%
Amundsen et al. (2014)	Breath & Urine	n = 93	Small-cell lung cancer: Sensitivity = 100% Specificity = 33.3%
Rudnicka et al. (2015)	Breath	n = 179	Sensitivity = 85.54% Specificity = 71.84%
Hackner et al. (2016)	Breath	n = 122	Sensitivity = 78.6% Specificity =34.4%
Montes et al. (2017)	Breath	n = 113	Sensitivity = 95% Specificity = 98%
Fischer-Tenhagen et al. (2018)	Breath	n = 60	Sensitivity = 95% Specificity = 60%
Riedlova et al. (2022)	Breath & Blood	n = 216	Sensitivity = 72% Specificity = 94%

Existing studies generally report strong sensitivity with several authors like McCulloch et al. (2006) and Amundsen et al. (2014) reaching 99 - 100%. However, specificity varies dramatically. While McCulloch and Montes et al. (2017) report specificities above 98%, others like Amundsen et al. (2014) and Hackner et al. (2016) found specificities as low as 33-34%. Exhaled breath remains the predominant medium for lung cancer detection, appearing in every study listed above. The rationale for this sample choice is that pathological processes induce metabolic shifts, resulting in the emission of unique volatile organic compounds (VOCs) which may be carried in the exhaled breath (Maidodou et al., 2023).

The detection of lung cancer involves a significant stimulus challenge compared to searching for a discrete substance, such as drug search or bomb sweep. In explosive scent tracing, although all targets and non-targets share some similar “noise” (i.e., irrelevant and variable odorants), the target should always contain an additional odor of the specific explosive compound that the animal has been trained to detect (Sargisson et al., 2010). To date, no single, consistent volatile biomarker has been identified in the breath of lung cancer patients (Ratiu et al., 2021). Instead, the "cancer scent" is characterized as a complex, variable "fingerprint" of multiple VOCs. Consequently, detectors may be responding to a general odor pattern rather than a specific invariant compound, and this reliance on identifying a variable pattern aligns the task more closely with the framework of concept formation rather than simple discrimination (Crawford et al., 2023).

1.3 Concept Formation

1.3.1 Concept Formation from Behavioural Perspective

Concept formation is a construct rooted in cognitive psychology, formalized by Bruner et al. (2017). They defined it as a mental process of rule-based categorization, describing how

humans learn to classify stimuli into functional groups (e.g., distinguishing "chairs" from "non-chairs") based on shared attributes rather than rote memorization. Early behavioural research demonstrated that this ability is not unique to humans. In a foundational study, Herrnstein (1979) trained pigeons to peck at images containing trees and to withhold pecking for images without trees. Eventually, the pigeons successfully discriminated novel pictures of trees they had never exposed before. This generalization to novel exemplars provided empirical evidence that non-human animals are capable of learning abstract categories, responding to the defining features of a stimulus class rather than specific training instances.

Zentall, Galizio, and Critchfield (2008) further refined this framework, shifting the focus from "concepts" as internal mental structures to "concept learning" as a behavioural process. Drawing on early principles of reinforcement (Keller & Schoenfeld, 1950), they illustrated that classification is the result of a history of repetitive differential reinforcement. In this view, an organism learns to generalize within a stimulus class (treating different stimuli as equivalent) while discriminating between classes. This capacity appears to be phylogenetically widespread; studies have showed that species as diverse as honeybees (Apis mellifera) can demonstrate learning of abstract concepts of "sameness" and "difference" (Giurfa et al., 2001), and similar capabilities have been documented in parrots and corvids (Zentall et al., 2008).

1.3.2 Concept learning in lung cancer detection

The detection of lung cancer via exhaled breath represents a specific type of concept learning task: perceptual concept learning. Zentall et al. (2014) discuss the structural requirements of concept learning in terms of intra-class and inter-class similarity. Applying these terms clarifies the challenges faced when training the medical detection dog:

1. Learning Intra-Class Similarity: The dog must generalize the indication response across all members of the "cancer" class. The scent of each positive sample can be affected by patient-specific factors such as diet, gender, and smoking status (Ratiu et al., 2021). The dog must learn that despite this high variability (noise: e.g., sample age, patient age, sex, smoking status, and diet), these samples belong to the same category.
2. Detecting Inter-Class Difference: As mentioned, there is no single biomarker for lung cancer; the "cancer scent" is likely to be a pattern of VOCs. The dog must identify the complex “cancer” VOCs pattern that differentiates the “cancer” class from the "non-cancer" class.
3. Ignoring Inter-Class Similarity: The dog must withhold responding to features that are shared by both classes (e.g., odours associated with diet, gender, smoking status, and general human breath odours or background VOCs from the hospital environment).

Therefore, in summary, to be successful, the dog must learn the perceptual class of the "lung cancer odour pattern" by responding to the common, relevant features from a set of diverse examples, while simultaneously learning to ignore the irrelevant variations that characterize individual samples.

While the acquisition of a “conceptual class” is well-studied, a significant gap exists in the literature regarding the maintenance of such acquisition under thinned reinforcement schedules, particularly in canine lung cancer detection. All of the studies listed above focused exclusively on the acquisition of stimulus control under continuous reinforcement (FR1) and subsequently moved directly to the blind test where reinforcement for unknown-status samples were always absent. As noted in the literature, the specific work requirements of the field environment directly impact detection performance, as the frequency and availability of reinforcement plays an important role

in sustaining motivation (Caldicott et al., 2024). Therefore, this approach failed to account for the resistance to extinction required in applied settings where target prevalence is low and feedback is unavailable. This oversight is critical because the stability of a broad perceptual class, such as a "cancer scent", may be uniquely sensitive to shifts in reinforcement frequency. Understanding this relationship is essential to determine whether a learned concept persists or is disrupted when moved from training to operation conditions.

1.4 Intermittent Reinforcement

1.4.1 The Necessity of Intermittent Reinforcement

Intermittent reinforcement schedules are theoretically advantageous for assuring long-term performance, because they tend to be considered to strengthen the stability of a trained behavior (Hall, 2017) and provide significantly greater resistance to extinction— persistence of the behaviour after it is no longer reinforced. Following on from Crawford et al.'s (2022; 2023; 2025) research on several factors that affect dogs' performance in lung cancer detection, an unpublished investigation into dogs' performance in a blind test was undertaken. Dogs evaluated 17 samples in the blind test with a ratio of 10:7 or 7:10 (unknown sample to known sample), where the correct indications to known-status samples were reinforced and no feedback was given for the responses toward unknown samples. During the blind test, accuracy dropped significantly, which might suggest a disruptive effect from the reinforcement schedule, since reinforcement for correct indications of unknown-status samples was absent during the blind test. Although intermittent reinforcement was introduced during training, the schedule was very minimal and consistent. However, the number of positive samples with unknown status in the blind test was variable and, in some cases, larger than what the dogs had been exposed to previously, resulting in the reinforcement schedule being inconsistent, which could be responsible for the performance drop.

A critical gap exists in the current literature regarding the application of intermittent reinforcement schedules to scent detection with complex targets; it remains unclear whether the withdrawal of consistent feedback affects the stability of these broad perceptual categories. This study investigated the effect of intermittent reinforcement on the performance of dogs trained to detect lung cancer, a concept learning task.

1.4.2 Intermittent Reinforcement & Concept Learning

"…Any plausible account must also explain how categories add and lose members, merge and fracture, share members that may belong to different categories under different circumstances, support the spontaneous transfer of function from one member to another, and so forth."

- *Zentall et al. 2002*

As noted, due to high variability across samples, the members of each class (positive and control) might be highly dynamic due to the reinforcement rate. If intermittent reinforcement does influence concept learning, there are two likely outcomes of implementing an intermittent reinforcement schedule:

1. Rejection Bias ("Giving-up Strategy"): Detectors start to reject more samples, as some correct positive indications are no longer followed by reinforcement. Detectors might begin to learn and generalize this new concept to other positive samples and stop indicating. In this case, some positive targets are "re-categorized" into the control group and are treated as negatives from that point.
2. Indication Bias (Try Everything Strategy): Detectors start to indicate more samples, as some correct positive indications are no longer consistently reinforced. Our training protocol initially employed a high ratio of positive to negative samples to teach dogs

the target odour and indication response. Therefore, detectors were exposed to a situation where indications were mostly reinforced. More negative samples were added into sessions as training progress, so dogs were in fact learning what to reject. Once learned lung cancer features become unreliable due to the intermittent reinforcement schedule, the dogs might start to relearn the task by treating more samples as potential targets.

1.5 This Study

1.5.1 Automated Training

The methodologies employed in prior canine lung cancer detection studies present several operational risks that may compromise diagnostic reliability. For example, Edwards (2019) argued that researcher-mediated procedures, such as the standard sample "lineup," could introduce significant issues and challenges. These include unintentional handler cuing, where the trainer's behavioural repertoire might provide subtle prompts to the dog, and a lack of immediacy and reliability in the manual delivery of reinforcement. To mitigate these factors, dogs in the present study were trained using a fully automated 17-chamber carousel allowing for the precise delivery of contingencies without human influence.

1.5.2 Sample Collection and Arrangement

While previous studies, such as Fischer-Tenhagen et al. (2018), were constrained by limited sample sizes (n = 60), forcing the repeated use of the same samples, researchers like McCulloch et al. (2006) and Riedlova et al. (2022) utilized relatively balanced sample sets (e.g., 115 positive and 101 negative samples) that do not reflect the low-prevalence demographics of a real-world clinical setting. This project addressed these limitations by utilising a significantly larger pool of

samples from 348 unique patients, consisting of 115 positives and 233 negatives. By strictly limiting sample re-use to a maximum of three exposures, we ensured that the dogs were developing a generalised "cancer concept" rather than memorizing individual patient profiles. Furthermore, instead of using fixed positive-to-negative ratios in every session, such as 1:3 (Fischer-Tenhagen et al., 2018), 1:4 (McCulloch et al. 2006; Montes et al., 2017; Hackner et al., 2016), we employed higher ratio arrangements of 6:11 or 5:12. These two arrangements were varied randomly between training days to better simulate the unpredictable nature of a realistic screening environment.

1.5.3 Rejection Training

Most of the studies above did not include training of dogs to reject individual negative samples; instead, dogs were required to indicate one positive samples among all the samples presented, with other samples considered passively rejected once an indication was made. To ensure dogs maintain high specificity while working through a higher volume of control samples, we implemented a two-link behavioural chain. In this sequence, the dog's initial observing response, inserting its nose into the port for evaluation, served as the first link of the chain. If the target was presented, the dog simply maintained its position; once the nose-holding time reached the indication threshold, the reinforcer was automatically delivered. However, in the absence of the target, the second link of the chain was initiated: dogs were required to withdraw their noses from the port and operate the lever to complete an active rejection. Unlike the passive or unstandardized rejection behaviours seen in the Hackner et al. (2016) and Fischer-Tenhagen et al. (2018) protocols, dogs in this study were required to perform an active lever-press to reject a negative sample.

1.5.4 Objective

The primary objective of this study was to investigate how different reinforcement schedules influence the maintenance of performance in a scent-based concept learning task.

Specifically, the study sought to determine whether an intermittent reinforcement schedule could sustain high levels of discrimination accuracy and generalization in a complex medical detection scenario.

The experimental protocol proceeded in two main stages. First, dogs were trained under a continuous reinforcement schedule until stable discrimination accuracy was achieved. Subsequently, the frequency of reinforcement was systematically reduced to examine the effects of intermittent reinforcement on the stability of the concept discrimination and the dogs' ability to generalize to novel samples. The findings were intended to clarify how reinforcement conditions affect complex discrimination learning in scent detection and to inform best practice for operational training and disease detection.

Chapter 2: Method

2.1 Subjects

2.1.1 Recruitment

Subject recruitment prioritised dogs with prior experience operating the automated canine scent-detection apparatus used in this study. Selection criteria also required dogs to demonstrate evidence that dry kibble functioned as reinforcement. Seven dogs were initially included in the study (see Table 2) and all of them had experience working with the same apparatus in past studies. Tui and Ben had participated in the previous lung cancer detection study in 2022. Mac, Memphis, and Bailey participated in another scent detection study also investigating intermittent reinforcement in 2024 in the same lab. Harlee was part of the team investigating dogs' ability to detect koi carp in 2024. All dogs were reported to be working effectively with the apparatus and dry kibbles in the past studies.

Five of the seven dogs completed the entire training and were tested in the intermittent reinforcement phase. Harlee dropped out from the study before the completion of training, due to the change of her owners' working schedule, making the drop-off and pick-up no longer possible. Tommy was retired from the study due to a health condition.

Table 2

Participant dog profiles

Name	Sex	Breed	Cohort
Tui	Male	Border Collie X Huntaway X Kelpie	Morning
Memphis	Male	American Staffy X Mastiff	Morning
Mac	Male	Chocolate Labradors	Morning
Ben	Male	Chocolate Labradors	Afternoon
Bailey	Female	Chocolate Labradors	Afternoon
Harlee	Female	Labrador Retriever	Afternoon
Tommy	Male	Border Collie X Springer Spaniel	Afternoon

Note. All dogs were neutered.

2.1.2 Cohorts

Trainings took place on Monday and Tuesday from 9:00 am to 4:00 pm, which including two cohorts, morning: 9 am – 12 pm; and afternoon: 1 pm – 4 pm (see Table 2). The number of sessions per morning dog was varied from 4 to 5, and 4 to 6 sessions for afternoon dogs.

2.1.3 Ethics Approval

Approval for this experiment was obtained from the University of Waikato Animal Ethics Committees (Protocol #1228).

2.2 Sample

2.2.1 Sample Collection

In this study, breath samples were obtained from patients at one respiratory clinic prior to clinical diagnosis. For each participant, four samples were collected using custom glass tubes

(26 mm OD×2 mm wall thickness, 120 mm length; Duran) containing 1 g of L10Y4 polypropylene fibre (IFG Asota) stabilized between two sections of a cotton wool ball (see Figure 1). To ensure sterility and minimize background contaminants, tubes were either annealed at 540°C or chemically cleaned in 60% nitric acid for 24 hours, followed by a deionized water rinse and oven drying. Patients provided three full exhalations per tube. Following collection, tubes were sealed with LDPE tapered caps (Hi-Q Components) and secured with tape. Each sample was individually sealed in a resealable plastic bag, which were then housed together into one "mother bag" for each respective patient. To maintain procedural consistency, a single nursing team collected all 1,368 samples in a designated consulting room over a four and a half-year period. The nursing team provided demographic information for the patients as soon as information about their diagnostic status became available.

Figure 1

Breath samples utilized in this study

Note. This photograph illustrates a complete breath sample. Exhaled breath is captured within the L10Y4 polypropylene fibre (visible as the shiny material) in the middle, which is stabilised by

cotton wool balls on either side. The glass tube is sealed with red caps on both ends and is placed within the plastic bag shown in the background for storage.

2.2.2 Sample Preservation and Preparation

Samples were maintained at -80°C to -60°C at both clinical site and the laboratory. Demographic analysis was conducted to identify potential confounders, such as gender, smoking status, and cancer subtype. No significant gender differences between lung cancer positive group and control group were observed from the analysis. However, a higher proportion of smokers was noticed in the lung-cancer-positive group. To mitigate the perceptual bias related to tobacco use, the training samples were adjusted by substituting randomly selected control samples from smokers into the control group to match the higher smoking prevalence in the lung cancer positive group.

To prevent cross-contamination, trainers were required to handle all materials (individual bag, mother bag, tube, caps, and tape that was used to seal the tube) associated to samples with nitrile gloves, which were replaced immediately after contact with any patient's sample or experimental materials. Samples were processed on individual metal trays to defrost prior to the sample loading, which took at least 40 min. After each training day, all samples were returned into their original bag and placed back into the freezer; trays and segments that were used to hold samples were cleaned with detergent wash followed by a 1:1 water-isopropanol rinse and then air-dried.

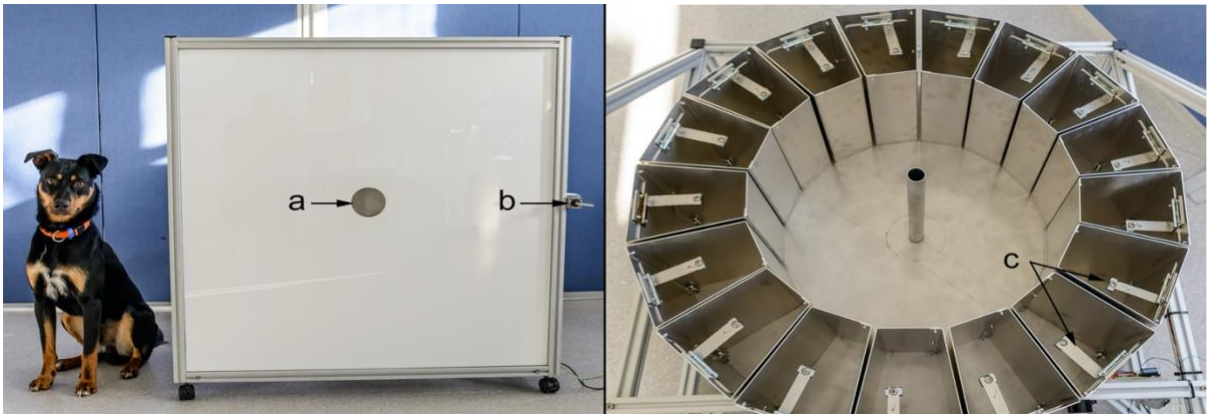
2.3 Apparatus and Operational Contingency

The training took place in 3.2m x 4.3m testing room using an 1 m^3 apparatus containing 17 segments (Edwards, 2019). Dogs engaged with samples by inserting noses into the port on the

front panel, and opening the flap (see Figure 2). A grid of three infrared beam sensors is set up behind the port, these sensors detect the entry of any obejct larger than 15 mm in diameter, such as a dog's nose. An indication is defined as a dog breaking the beam longer than the set indication threshold time, which was adjusted according to individual dogs' performance. The rejection is defined as the dog withdrawing its nose before the nose-holding (beam breaking) time reaches the indication threshold and operating the lever. Any beam break lasted longer than 500 ms was registed as an observation, which is required to make the lever operable.

Figure 2

Apparatus



Note. The left image shows a front view of the apparatus, and the right image provides a bird's-eye view. The carousel was positioned behind the white panel, with segments containing breath samples placed on its surface. The push-to-open flaps (c) of the segments were aligned with the sampling port (a). The dogs inserted their noses into the port and extended further through the flap to evaluate the samples. The lever was only operable once the evaluation time reached 500 ms.

Responses were categorized into four outcomes based on the sample's diagnostic status (positive and negative): Hit (indication to positive sample); Miss (rejection of a positive sample);

False Alarm (FA; indication to a negative sample); and Correct Rejection (CR; rejection of a negative sample). The payoff matrix is shown in Figure. 3.

Figure 3

Payoff Matrix in scent detection tasks

	Indication response	No indication
Lung-cancer Positive	Hit Reinforcement	Miss No reinforcement
Lung-cancer Negative	False Alarm (FA) No reinforcement	Correct Rejection (CR) No Reinforcement

The apparatus utilized synchronized auditory and visual cues to signal both the human trainer and the dogs through the task. A speaker was set up under the carousel providing dogs with additional auditory cues associated with the various states and changes of the apparatus. Strips of LED lights were also installed inside the front panel; different colours were used by human observers to troubleshoot. The carousel advanced to the next sample in the following manner was applied in all four scenarios. The carousel brought the next segment to the port accompanied with a continuous buzz sound. As the next sample is correctly positioned, the buzz sound ceased, followed by two brief, consecutive beep sounds and the apparatus return into the ready state waiting for the dog to re-engage. As a safety precaution, the motor ceases movement if the beam is interrupted (object entry) during carousel rotation. The operational contingencies were managed as follow:

- **Ready state and observation**: The ready state was indicated by two brief consecutive beep sounds. As a dog inserted its nose into the port, the breakage of the beam triggered a continuous beep that persists until dog either make a hit when a positive sample was presented or withdrew its nose from the port.
- **Hit**: When dogs held their nose into the port until the indication threshold time was reached when a positive sample was presented. The automatic feeder located on the opposite side of the training room dispensed reinforcer (dry kibble). Apparatus enters transition procedure (specified above).
- **CR and FA**: If the presented sample was negative, dogs withdrew their noses before the nose-holding time reach the indication threshold time and activated the lever. After a correct rejection, apparatus enters transition procedure. Dogs were allowed to spend unlimited amount of time on evaluating negative samples, once the beam breaks on the negative trials exceeded the set indication threshold time, an FA was recorded. Nothing happens after the FA; dogs must activate the lever eventually to pass on to the next sample.
- **Miss**: We employed error correction for a "miss" situation. If it is a positive sample and the dog withdraws its nose before the indication threshold time is reached, activating the lever, the carousel does not rotate to the next sample, and a continuous buzz sound was emitted until the apparatus returns to ready state. Dogs could not proceed to the next sample until they indicate the missed positive sample, and the reinforcers will be provided once they make a successful indication.

Sensitivity and specificity of dogs' performance were recorded each session, and a general accuracy score was calculated using sensitivity and specificity. They were calculated as follows:

- Sensitivity (Hit Rate) = Number of Hits / Number of Positive Samples;

- Specificity (Correct Rejection Rate) = Number of CRs / Number of Negative Samples;
- Accuracy = (Sensitivity + Specificity) /2.

2.4 Procedure

2.4.1 Training

As all participant dogs had previously been exposed to the apparatus, only a brief apparatus-use training was conducted for the first two weeks of training. A total of 10 samples were included initially with the ratio of 4:6 (positive to negative), indication thresholds were set as 1000 ms for all dogs. Each week, one positive and negative were added until a total of 17 samples with ratio of 6:11 (Positive to Negative) was reached and then arrangements were varied between 6:11 and 5:12 randomly for each training day. The proportion difference of smokers among positive and negative samples were controlled within 25%. Both sample selection and sample order were randomized each training day. The same set of samples were used throughout a training day, and each dog completed an average of 5 to 6 sessions per day with these samples. The indication threshold time was adjusted by 500 ms if the dogs' performance met the predetermined criteria for this change. The indication threshold time was increased by 500 ms if a dog's sensitivity remained at 100%, and the specificity was below 50% for three consecutive sessions. On the other hand, the indication threshold time was decreased by 500 ms if a dog's sensitivity remained below 66.67% and the specificity was above 50% for three consecutive sessions.

Due to the sample degradation and the limited amount of samples collected, each sample was used up to three times (Crawford et al., 2023). During the training phase, a dramatically increased accuracy, specificity in particular, was observed on the training session 276 among all

dogs. Upon investigation, trainer identified a potential confounding variable related to the chronological age of samples. Consequently, all samples collected from patients prior to ID 1001 were excluded from that point onward.

A Weibull cumulative function was utilized to aid us making decisions whether dogs' performance have reached asymptote and ready to be transitioned from training phase to testing phase. Asymptote is a state in a response acquisition where the subjects stop learning, considered as the completion of acquisition, and the Weibull function was reported as a useful quantitative approach to determine asymptote (Crawford et al., 2022).

2.4.2 Intermittent Reinforcement

Dogs were transitioned into the intermittent reinforcement stage once their performance was considered to be at asymptote. A reversal design was implemented for all dogs, with each dog exposed to one of two reversal designs, A-B-A-C and A-B-C-A. In Intermittent reinforcement (IR) phase, sessions were arranged with 6 positive and 11 negative samples or 5 positive and 12 negative samples, which was the same as the training arrangement. In addition, positive samples that were not associated with reinforcement were programmed as negative in the software; therefore, dogs needed to press lever to advance the apparatus. For these samples, a lever-press after nose-holding time reaching the indication threshold was recorded as a Hit or a CR if they pressed the lever before the indication threshold was reached.

Baseline data (A) were established using the mean accuracy from the final five training sessions under a continuous reinforcement schedule. Following the first baseline phase (A), dogs were transitioned to intermittent reinforcement phase 1 (B; IR1), where the reinforcement rate for correct positive indications was reduced to 80% (4 out of 5 positive samples were associated with

reinforcement) or 83% (5 out of 6 positive samples were associated with reinforcement) for five sessions. The subsequent experimental trajectory was determined by each individual's performance in IR1 (B):

1. A-B-A-C: for dogs who exhibiting a significant performance decline (>10% below first baseline) were placed back to second baseline phase (A) for five sessions to recover performance before proceeding to the more challenging IR2 Phase (C).
2. A-B-C-A: Dogs maintaining stable performance (fluctuating within a 10% range of first baseline) were transitioned to IR2 (C), then returned to another baseline phase (A).

In IR2 (C), the reinforcement rate was further decreased to 60 (3 out of 5 positive samples were associated with reinforcement) or 66.7% (4 out of 6 positive samples were associated with reinforcement).

2.4.3 Probe Test

Breath samples from non-lung-originated (NLO) cancer patients, individuals with primary malignancies located outside the lungs, were excluded during training and testing phases. This exploratory probe test was prompted by investigating how dogs trained exclusively on lung-originated cancer samples (hereafter referred as lung cancer positives) respond to NLO sample.

All protocols were the same as training phase, only sample arrangements were slightly different. Dogs were placed in sessions containing either four or five probe samples interspersed with standard positive and negative breath samples. Two arrangements were employed:

Arrangement A: 5 positives, 8 negatives and 4 probes.

Arrangement B: 6 positives, 7 negatives and 4 probes.

Probe samples were programmed as "negatives" in the software system. Also, to prevent dogs learning probes as negative as the result of being exposed to such contingencies for too many sessions, only two sessions were arranged for each training day, and the testing was conducted for only two days. This test allowed for an evaluation of whether trained lung cancer detection dogs' responses were specific to lung-origin cancer profile or if they generalized to common cancer-associated VOCs.

2.5 Data Analysis

As previously noted, individual breath samples were used up to three times, also the same sample array was repeatedly presented across sessions within a given training day. These two types of repeated exposure would likely lead to better performance from them in the later encounters, either within a training day or across training days. However, the core objective of this task was to train dogs to generalise the learned lung-cancer concept to novel stimuli. Therefore, to accurately evaluate dogs' target acquisition, only their initial responses toward samples from novel patients were analysed.

These were fitted in Weibull cumulative function. In this model, the starting value (g) was manually fixed at 0.5 (representing chance level performance). The slope (k) and inflection point (b) were allowed to vary as free parameters. This function was employed to forecast the performance plateau where further training yielded no significant gains in accuracy. Once dogs' performances were statistically confirmed to have reached asymptote, they were transitioned to the intermittent reinforcement (IR) phase. However, the result of this analysis did not provide a clear indication of when dogs should be transitioned to the IR phase, so the decision was made based on a visual analysis of the data.

Comparison of performance between the acquisition phase and different IR phases were conducted using Linear Mixed-Effects Models (LMMs). LMMs were selected to account for the hierarchical structure of the data and the non-independence of repeated measures derived from the same dogs. The experimental phase was entered as a fixed effect, while Dog ID was included as a random effect to control for individual variability in baseline performance. We used models to compare data from each intermittent reinforcement phase to other phases.

To ensure the most accurate assessment of the dogs' spontaneous response to NLO odours without the influence of within-session feedback, data analysis focused primarily on the first session of each testing day. Each dog's indication rates on lung cancer positives, lung cancer negatives and NLO were collected and a one-way repeated measures analysis of variance (ANOVA) was conducted to assess the statistical significance of differences in indication rates between sample types within subjects. Significant main effects were investigated using post-hoc pairwise paired t-tests. To control for the family-wise error rate associated with multiple comparisons, all *p*-values were adjusted using the Bonferroni correction. Statistical significance was set at $\alpha = .05$.

Additionally, a Pearson correlation analysis was employed to investigate the relationship between sample age and indication rate.

All statistical analyses were conducted using the R statistical software (Version 4.5.2) using "lme4" (Version 1.1.38; Bates et al., 2015) and "rstatix" (Version 0.7.3; Kassambara, 2025) packages.

Chapter 3: Result

3.1 Acquisition Phase

3.1.1 Training

This study is a direct continuation of a previous lung cancer detection project, which formally concluded at session 235. Therefore, the current training phase start from session 236 to 287. General information regarding total training sessions completed by each dog is presented in Table 3, including two initial apparatus training sessions. Individual performance was recorded using subjects' first response to new samples from novel patients; they are detailed in Table 4. Results showed that individual accuracy ((sensitivity + specificity)/2) ranged from 58.16% to 65.65%, with Mac achieving the most balanced performance between sensitivity and specificity. The cohort exhibited high sensitivity (M = 89.27%, SD = 2.73), indicating a strong ability to correctly identify positive samples. However, specificity was notably lower (M = 33.46%, SD = 7.22), reflecting a general indication bias on all samples during the training phase.

Table 3

Total training sessions completed. (Including two initial sessions for apparatus training; session 236-287)

Dog	Number of training sessions	Mean number of sessions per day	Number of half days completed
Tui	224	4.77	47
Memphis	242	4.84	50
Mac	227	4.83	47
Ben	231	4.91	47
Bailey	228	4.85	47

Table 4

Individual performance of detection dogs during training phase (session 238 - 287)

Dog	Training Sessions	Sensitivity Mean (*SD*)	Specificity Mean (*SD*)	Accuracy Mean (*SD*)
Tui	45	87.78% (26.45)	28.54% (29.00)	58.16% (19.49)
Memphis	48	93.75% (19.64)	23.90% (26.76)	58.83% (14.75)
Mac	45	89.26% (25.41)	42.04% (24.54)	65.65% (16.21)
Ben	45	86.67% (24.77)	37.09% (31.60)	61.88% (20.54)
Bayley	45	88.89% (23.57)	35.71% (29.87)	62.30% (15.30)
Mean	45.6	89.27% (2.73)	33.46% (7.22)	61.36% (3.03)

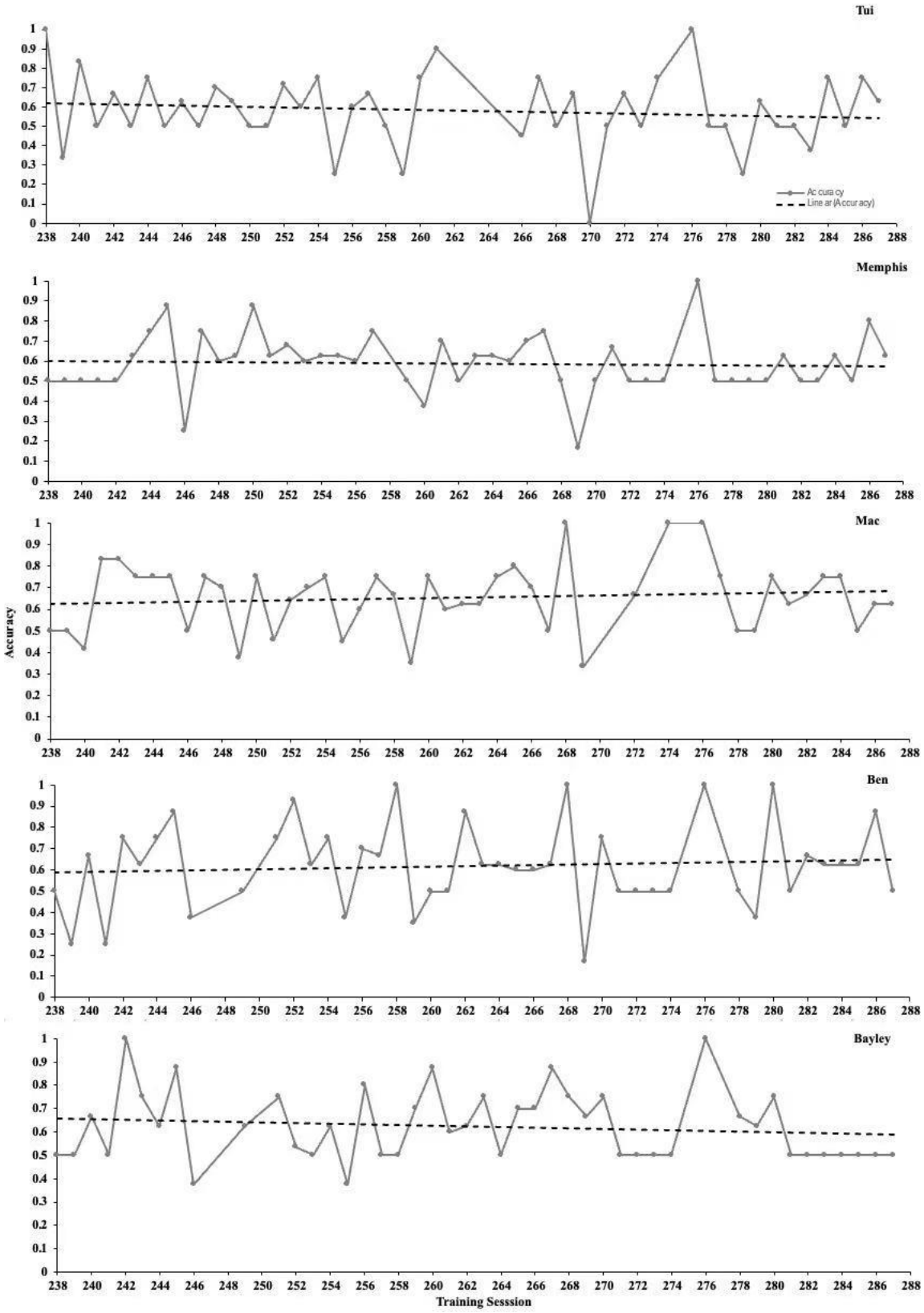
Note. SD = Standard Deviation; Accuracy Mean = (Sensitivity + Specificity) / 2

3.1.2 Transition from Training to Intermittent Reinforcement Phase

Figure 4 displays the mean accuracy ([sensitivity + specificity]/2) of the cohort across the training phase. Grey paths and markers represent the mean accuracy; the black dashed line is a fitted trendline, calculated using ordinary least squares linear regression (OLS), plotted out the overall trajectory across the whole training phase. Due to only hits being reinforced, dogs tended to indicate all samples encountered at the beginning of the training phase. Therefore, mean accuracy of 50% (sensitivity = 100%, specificity = 0%) was expected as initial performance for most of the dogs. For this reason, the lower asymptote "γ" in the Weibull function was manually fixed at 0.5 for the analysis. However, Tui had previously participated in the lung cancer detection project two years prior and displayed nearly perfect performance initially; subsequently, the trend line headed downward as performance stabilised. Overall, the data demonstrates high variability throughout training; no significant upward acquisition trend can be observed in the graph. This visual evidence corresponds with the findings of the Weibull analysis (see Table 5).

Figure 4

Dogs' performance across training phase (Session 238 - 287).


Tui
Memphis
Mac
Ben
Bayley
Accuracy
Training Sesssion
Ac cura cy
Line ar (Accur acy)

As shown in **Table 5**, the estimated asymptote for the whole cohort ranged from 61% to 66.8%. The inflection points for Tui, Memphis, Mac, and Bayley occurred at Session 1. Statistically speaking, this result indicated that all dogs' performance reached the plateau at the beginning of training. Only one subject, Ben, exhibited a brief acquisition period, stabilised at Session 8. Therefore, this analysis provided very little information to us on whether dogs were ready to be transitioned to experimental phase.

Table 5

Weibull analysis for performance in training phase

Dogs	Lower Asymptote (γ)	Upper Asymptote (α)	Slope (β)	Plateau Point (Session)
Tui	.50	.616	21.10	1
Memphis	.50	.616	21.10	1
Mac	.50	.668	3.43	1
Ben	.50	.637	4.74	8
Bayley	.50	.629	2.50	1

Note. A four-parameter Weibull function was fitted to daily diagnostic accuracy. The Upper Asymptote (α) represents the subject's estimated maximum potential accuracy for this phase. The Slope (β) indicates the rate of learning. The Plateau Point indicates the session number in which the acquisition stops for this subject.

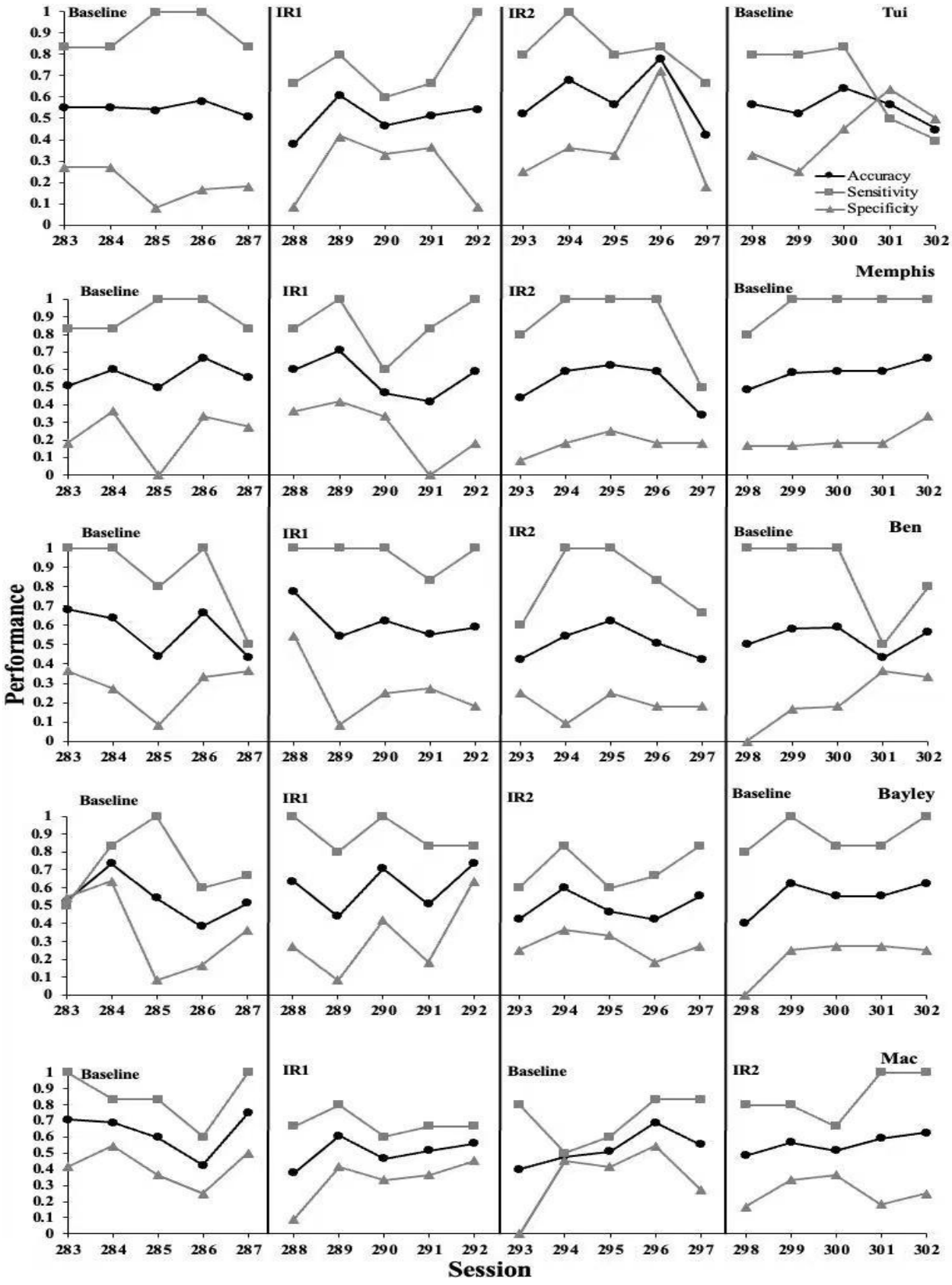
3.2 Intermittent Reinforcement

The effect of varying reinforcement schedules on lung cancer detection was evaluated using a reversal design. Individual performance across the baseline, IR1, and IR2 phases is presented in Figure 5. Black data paths display the mean accuracy of dogs' responses, grey paths

with square markers represent the sensitivity, and the grey paths with triangle markers represent the specificity. Visual inspection of the individual data paths indicates that all five dogs maintained their baseline levels of accuracy as the thinner reinforcement schedules were applied. Minor fluctuations were observed in Tui's second baseline performance, where sensitivity fell below specificity for the last two sessions, but, overall, no immediate or sustained drops in performance were observed following the transition from Baseline to the two variable ratio schedules (IR1 and IR2).

Figure 5

Performance across different reinforcement conditions

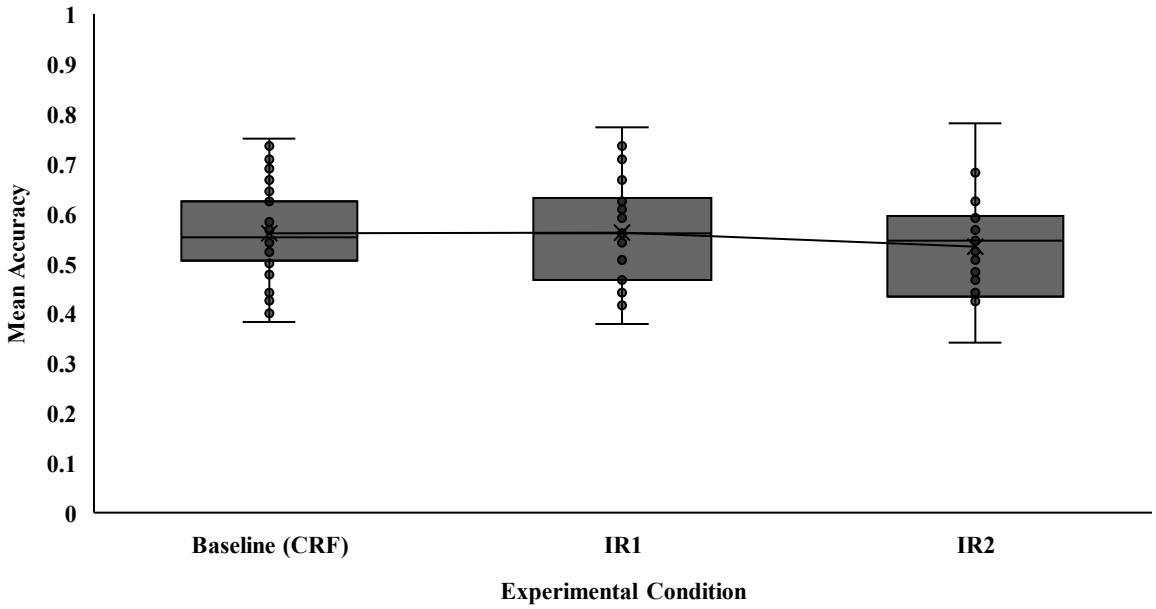

Baseline
IR1
IR2
Baseline
Tui
Accuracy
Sensitivity
Specificity
Memphis
Ben
Bayley
Mac
Performance
Session

Note. Two different designs were employed: A-B-C-A (Tui, Memphis, Ben, and Bayley) and A-B-A-C (Mac). Despite the schedule being thinned from 100% to as low as 60%, visual analysis reveals that mean accuracy remained stable.

Figure 6 summarizes the distribution of the entire cohort's accuracy scores across four conditions. The group median accuracy remained consistent across all phases, a slight drop is shown in IR2, but still with overlapping interquartile ranges. A Linear Mixed Model (LMM) was used to evaluate the effect of reinforcement schedule on detection accuracy. To account for variations in the chronological order of phase across subjects, A-B-C-A and A-B-A-C, two baseline sessions were pooled into a single control condition. Therefore, the model included "Phase" (Baseline, IR1, IR2) as a fixed effect and individual dogs as a random effect to account for repeated measures. The analysis revealed no statically significant main effect of the experimental phase on accuracy, $F(2,97) = .748$, $p = .476$. These findings suggest that dogs' detection performance in this task was not significantly disrupted as reinforcement rate ranged between 60% - 100%.

Figure 6

Cohort's accuracy across three different reinforcement schedules.



Note. The box plots illustrate the distribution of the cohort's accuracy across the pooled Baseline, IR1 and IR2 phases. The horizontal line within each box represents the median, while the upper and lower boundaries of the box indicate the interquartile range (IQR; 25th to 75th percentiles). The error bars (whiskers) extend to the lowest and highest values within 1.5 times the IQR from the lower and upper quartiles.

3.3 Probe test

A one-way repeated measures ANOVA was conducted to compare individual indication rates across the three sample types. There was a significant effect of sample type, F (2,8) = 19.00, p < .001. Each dog's performance data is detailed in Table 6, Tui, Memphis, Mac and Ben demonstrated a clear discrimination pattern, where indication rates for NLO samples dropped to levels comparable to negative controls (54.5% - 63.6%). However, Bayley demonstrated a high

indication rate (81.8%) for NLO samples, divergent from the cohort trend for probes, similar to her response to positive samples (84.0%).

To control for the family-wise error rate associated with multiple comparisons, all p-values were adjusted using the Bonferroni correction (see Table 7). For the cohort as a whole, their indication rates on NLO and negative samples do not differ from each other ($p = 1.000$).

Table 6

Indication Rates (%) Across Sample Types During Probe Testing

Dog	Lung Cancer positive	Lung cancer negative	NLO
Tui	77.3	64.5	63.6
Memphis	90.7	71.7	54.5
Mac	85.3	57.2	54.5
Ben	77.3	54.2	54.5
Bayley	84.0	57.8	81.8
Group Mean	82.9	61.1	61.8

Note. Data represents the percentage of samples where the dog made a positive indication (not accuracy). "NLO" refers to Non-Lung Originated cancers (probe samples).

Table 7

Bonferroni-Corrected Pairwise Comparisons of Indication Rates Between Sample Types.

Comparison	Mean difference	Standard Error	*p*-value
Positive vs. Negatives	+21.8	3.24	.002***
Positive vs. NLO	+21.8	4.88	.018**
NLO vs. Negative	+.7	5.92	1.000 (ns)

Note. ns = not statistically significant (p > .05).

3.4 Sample age

3.4.1 Rejection Bias towards Older Sample

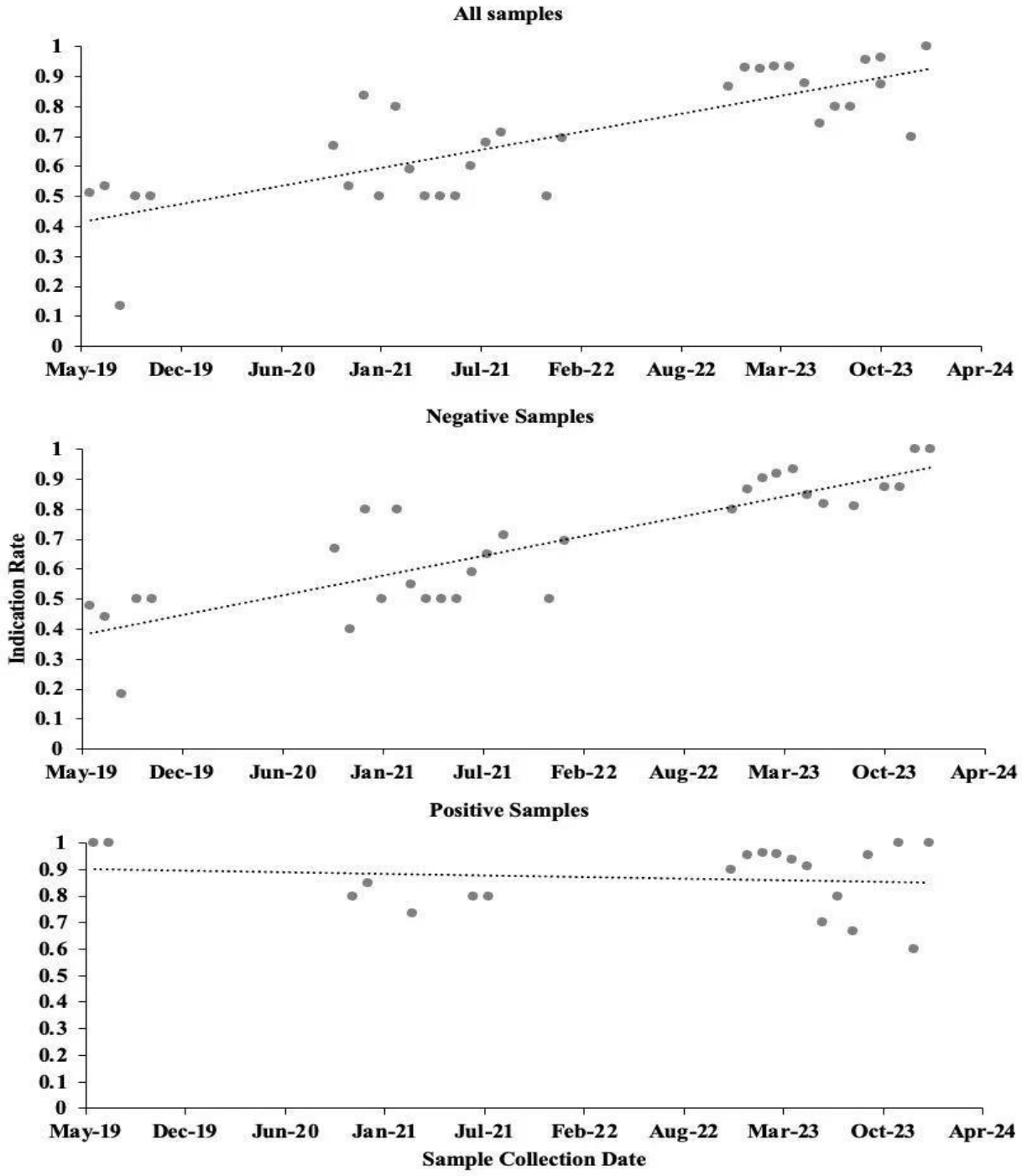
This analysis was prompted by a significantly increased specificity observed during training session 276. The trainer noted that the negative samples used in that session were predominantly older (collected earlier) than positive samples, leading to the hypothesis that VOC characteristics associated with sample age might be acting as discriminative stimuli. Indication rate and sample collection dates are plotted (see Figure 7). A visual analysis reveals that upward trendlines are observed among all samples and negative samples, suggesting that dogs tended to make positive indications more frequently to relatively fresh samples.

A Pearson correlation coefficient was conducted to investigate the relationship between dogs' indication rate and the sample collection date (see Table 8). For positive samples, there was no statistically significant correlation between collection date and indication rate ($r = .04$, $p = .73$), confirming that the dogs' sensitivity to the target odour remained stable regardless of whether the samples were fresh or aged. In contrast, a significant moderate positive correlation was observed for negative samples ($r = .40$, $p < .001$). This confirms the trainer's observation: older samples

(collected earlier) were associated with significantly lower indication rates (higher specificity), whereas newer samples elicited significantly more indications among negative samples.

Figure 7

Dogs' indication rate on samples collected at different time points.



Note. The graph illustrates the relationship between the sample collection timepoints and the cohort's indication rate. Each data point represents the percentage of total responses that were

positive indications for samples collected at that specific timepoint. An ordinary least squares (OLS) regression line is fitted to the data to illustrate the linear trend.

Table 8

Pearson Correlations (r) between sample age and indication rates across sample types.

Sample Type	Pearson's *r*	*p*-value	95% CI
Negative	.40***	<.001	[.25, .52]
Positive	.04	.73(ns)	[-.19, .27]
All samples	.41***	<.001	[.30, .51]

Note. CI = confidence interval. ***$p < .001$.

3.4.2 Post-adjustment Analysis

Samples from patient 1000 and earlier were excluded from the sample pool after a significant positive correlation was found between indication rate and overall sample age. A subsequent Pearson correlation coefficient was conducted at the end of study, between dogs' indication rate and sample age of those from the adjusted sample pool (see Table 9). The results showed that a significant positive correlation persisted across all remaining samples ($r = .62$, $p < .001$), as well as within the negative samples subset ($r = .58$, $p = .002$). However, there was no significant correlation between the indication rate and sample age in positive samples subset ($r = -.58$, $p = .059$). As illustrated in Figure 8, the connected data points reveal a distinctive, "step-like" increase between January 2022 and December 2022.

Table 9

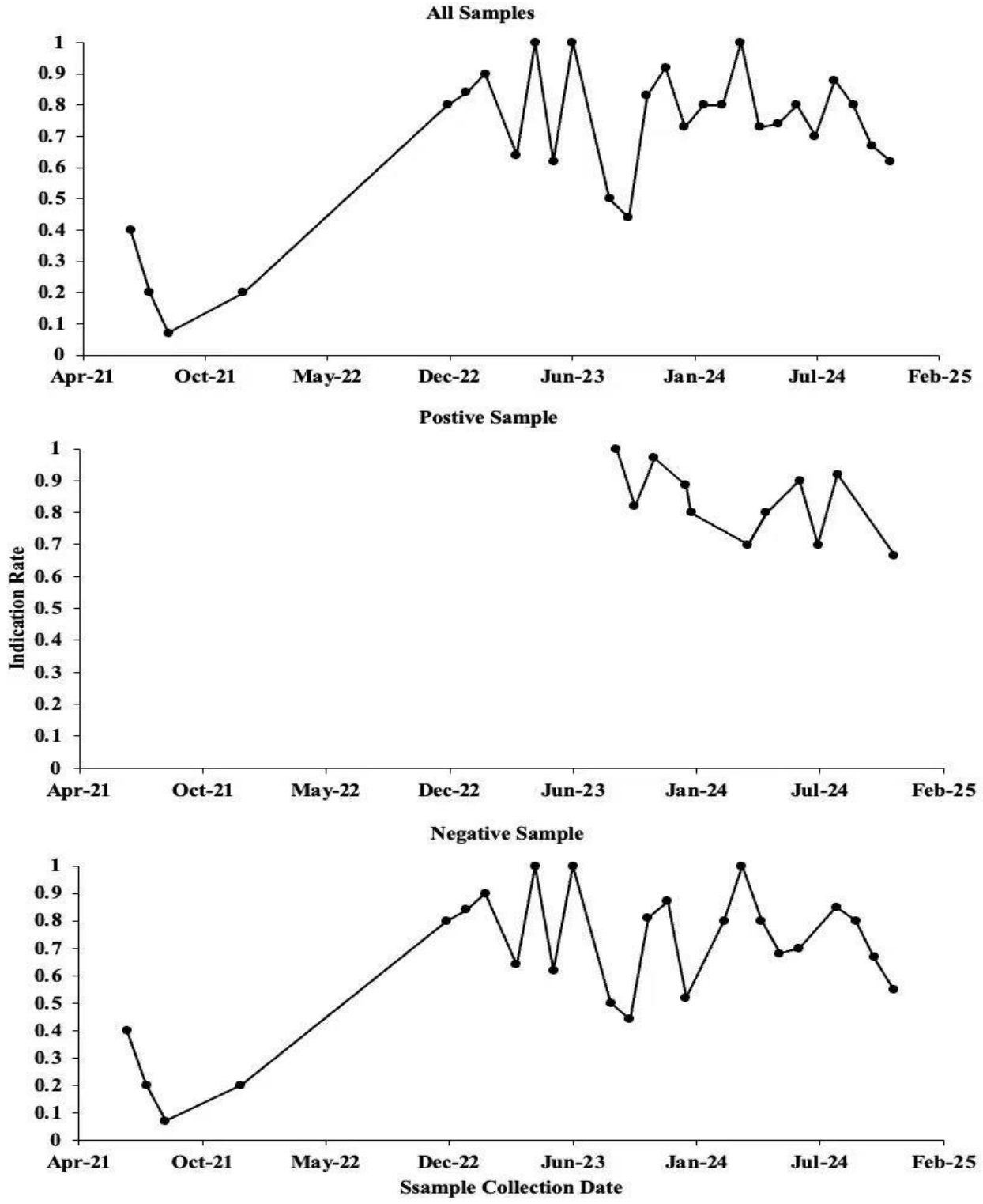
Pearson Correlations (r) between sample age and indication rates across sample types after the adjustment of sample pool

Sample Type	Pearson's *r*	*p*-value	95% CI
Negative	.58**	.002	[.24, .79]
Positive	-.58	.059	[-.87, .02]
All samples	.62***	<.001	[.31, .81]

Note. CI = confidence interval. ** $p < .01$. ***$p < .001$.

Figure 8

Dogs' indication rate on samples collected at different time points after the adjustment of sample pool



Note. The relationship between the sample collection timepoints and the cohort's indication

rate were plotted in scatter points with lines. A “step-like” increase can be observed between Jan - 22 and Dec - 22.

Chapter 4: Discussion

4.1 Intermittent Reinforcement

Results showed no significant differences in dogs' performance across conditions with different reinforcement schedules. It was initially hypothesised that introducing IR schedule might alter the lung cancer concept learned by dogs during training phase. Specifically, it was theorised that withholding reinforcement on correct indications might induce either a rejection bias ("giving-up" strategy) or an indication bias ("try everything" strategy).

As mentioned before, during the IR phase, positive samples that were not associated with reinforcement were programmed as negative controls to better mirror a real-life screening setting. Consequently, the indication responses toward those samples were placed under extinction, and dogs were required to make a rejection response by activating the lever to advance the carousel. It was thought that rejection behaviour toward negative controls was maintained by the conditioned reinforcer of sequence continuation (rotating carousel that brings the next potential target that could yield kibble). Therefore, theoretically, as thinner reinforcement schedules were implemented, dogs could have begun to "categorise" those unreinforced positive samples as non-targets. A rejection bias (giving up strategy) could be observed as a result. Interestingly, some observations within individual testing days, which typically comprised 5 to 6 sessions, revealed instances of dogs beginning to reject the specific positive samples in later sessions that had gone unreinforced in earlier sessions. However, these learnings within testing day did not generalise across testing days; dogs resumed their indication behaviour when they encountered other positive samples on subsequent testing days. This failure to generalise the rejection behaviour, or not indicating, towards other positive samples is likely due to dogs having gone through extended training and learned a highly resilient target concept. A study examining concept learning in pigeons described

by Wasserman and Bhatt (1992) found that while it took pigeons longer to acquire the categories with 12 exemplars, their ability to generalise to novel stimuli was significantly higher than those trained with only 1 or 2 exemplars. In the current study, the dogs had been exposed to hundreds of novel target exemplars; this large volume of exemplar training likely resulted in the learning of a robust and deeply entrenched concept in dogs. Behavioural momentum theory further explains why the indication behaviours were so resistant to extinction during thinned schedules. This theory suggested that behaviour with a rich, dense history of reinforcement acquires significant "mass" or momentum, making the trained behaviour highly resistant to change when faced with environmental disruptions, such as thinner schedules and extinction (Nevin, 2012). In the present study, dogs' indication behaviour toward positive samples had been heavily, consistently and exclusively reinforced with food across large arrays of unique target samples during training phase. Consequently, the disruptive factor of encountering a few unreinforced positive samples across only 5 sessions in each condition did not significantly impact the learned concept and corresponding response.

On the other hand, extreme indication responses ("try everything" strategy) were also not observed. The absence of "try everything" could suggest that the reinforcement schedule thinned from 100% to ranges of 80% and 60% was not severe enough to cause "extinction bursts" which might indicate that such reinforcement densities are still sufficient to maintain the integrity of the learned concept and unable to disrupt trained behaviours. An extinction burst describes a temporary increase in the frequency, duration, or intensity of a behavior when the reinforcement for that behaviour is removed (Cooper et al., 2020). This phenomenon was firstly documented in rats (Jones & Skinner, 1939), where researchers observed a rapid spike in the trained behaviour at the onset of extinction. In the current study, it was hypothesised that thinned reinforcement rates

might gradually shift indication behaviour under extinction, potentially akin to an extinction burst as the schedule became increasingly sever. The data did not support this outcome either.

However, it must be acknowledged that due to time constraints, the testing phase for this study was shorter than the team had planned; only 5 sessions were conducted for each IR condition. Furthermore, the schedule was only thinned to 60%, which was not sufficient to alter hit rates in either direction, an increases in hit rates may have also been unlikely due to a ceiling effect; current level of reinforcement schedule was probably not severe enough to pull their accuracy off the baseline level. Therefore, it remains highly possible that a more extreme schedule (50% or lower), or prolonged exposure to an intermittent reinforcement schedule over a much longer period might eventually cause a change in accuracy Overall, the current results demonstrate that a reinforcement rate fluctuating between 60% to 100% does not exert an immediate effect on performance in scent-based concept learning in lung cancer detection tasks, and such schedules could potentially sustain baseline levels of discrimination accuracy.

4.2 Training Approach

Previous training protocols in canine lung cancer detection studies have typically placed dogs in a free-operant environment. In these continuous-search setups, dogs were required to engage with multiple samples within a single trial, making active indications toward positive samples while passively rejecting non-targets by making indications. In contrast, the current study utilized a discrete-trial design that required the dogs to make an active, operant response to every sample presented. This active rejection requirement was implemented to ensure a more robust and quantifiable measure of specificity.

Because of this methodological shift, understanding the operant mechanisms of the training protocol is critical. In the current design, the nose-holding duration threshold for an indication imposes a significant response requirement and is considered as one of the most critical elements of the task that can lead to discrimination between positive and negative samples. The threshold quantifies the effort required to make a positive indication; Rasing the threshold increases the time and effort spent on nose-holding required for dogs to make an indication, and because this time and effort is never followed by reinforcement when a negative sample is present, the dogs eventually learn to reject these samples. During preliminary training, the threshold was set at 2000 ms, since only hits were reinforced in our protocol, dogs were indicating all samples encountered at this stage. Consequently, an indication bias was developed in dogs from the beginning, and dogs were essentially trained to reject negative samples from this point, while maintaining indication responses on positive samples. Therefore, elevated FA rates and low CR rates emerged as the primary barriers to high discrimination accuracy in this task. At least two potential training approaches could address the issue: reinforcing both Hits and CRs; increasing the indication threshold.

While reinforcing CRs appears to be an intuitive approach to address the issue, findings from Voss, McCarthy, and Davison et al. (1993) suggested otherwise. Researchers conducted two experiments comparing standard signal-detection procedures against an analogue "prey-detection" model using pigeons. Pigeons were trained to peck the left key (indication) when the target stimulus was presented, and the right key (rejection) when the control stimulus was presented. In procedure A, both Hits and CRs were reinforced with food, while both Misses and FAs produced a timeout period ranging from 3s to 120s (adjusted by trainers). In procedure B, which closely mirrors the procedures used in this study, only hits were reinforced with food. They found that

discrimination accuracy was significantly higher in procedure B; however, this heightened accuracy was accompanied by a response bias toward the left key (positive indication). Furthermore, from an applied perspective, the epidemiological nature of lung cancer suggest that positive cases will be far less frequent than negative cases in real-world screening environments. If CRs were also reinforced with food, the low prevalence of targets would likely cause the dogs to develop a strong rejection bias, undermining the screening tool's sensitivity. Therefore, this study utilized an asymmetrical reinforcement matrix during training and testing, wherein Hits were explicitly reinforced with a primary reinforcer (food), while CRs were not. However, CRs were not subjected to true extinction, instead, the operant response to a negative sample, pressing lever to advance the automated carousel, was maintained by a conditioned reinforcer: the sequence continuation that provided the next potential opportunity to encounter a target stimulus (positive sample) that would produce primary reinforcer (food). This design is firmly supported by the differential-outcomes effect (Trapold, 1970). Trapold demonstrated that discrimination tasks are acquired faster and maintained with higher accuracy when the two correct responses yield discriminably different reinforcing events, as the distinct outcome "expectations" provide additional discriminative stimuli. By establishing a differential outcome, primary reinforcement for Hits versus conditioned reinforcement for CRs, the current protocol maximized stimulus control. Managing response effort appeared to be a practical and effective approach in such training protocol. However, a strong bias toward positive response was observed across the whole cohort, which does align with the observation from Voss et al. (1993).

Adjusting the indication threshold was considered as an effective mechanism to improve dogs' discrimination performance, as the cost of making a FA was systematically increased. The necessity of increasing the indication threshold to improve specificity was also supported by recent

findings from Edwards et al. (2022). In their first experiment, five dogs were trained using the identical 17-segment carousel apparatus employed in the current study; the contingencies for indication and rejection were also the same, though the dogs were trained and tested with different stimuli: amyl acetate as the target and deionised water as negative controls. Starting at 4000 ms, the researcher systematically increased the indication threshold in 500 ms increments, eventually reaching 13000 ms. The study found that systematically increasing the response effort significantly reduced the FA rate and correspondingly increased the CR rate. At the initial 4000 ms threshold, the dogs exhibited a low CR rate ($M = 0.67$); however, as the indication threshold increased, dogs' CR rate rapidly increased and stabilised at a high level from the 6500 ms mark onward, and the optimal accuracy was achieved at 8000 ms. The Hit rate began to decrease at the 8000 ms mark, with CR rates remaining higher than Hit rates thereafter. Consequently, during the initial training phase of the current study, a threshold cap of 7000 ms - 8000 ms was agreed, with the intention of ceasing any further indication threshold extensions once 8000 ms is reached, preventing the jeopardization of sensitivity. However, it is important to note that Edwards et al. (2022) was conducted within the context of simple discrimination task, which likely contributed to a ceiling effect, whereby all dogs achieved near-perfect mean accuracy rapidly. In contrast, the current study required the dogs to learn a complex concept lung cancer VOCs. Given the significantly greater complexity of this discrimination task, it was later considered that the dogs might require greater response effort to achieve better performance. Therefore, the research team eventually decided to remove the 8000 ms threshold cap, allowing for further increases to the threshold.

4.3 Transitioning from Training to Testing

Crawford et al., (2022) has demonstrated that the Weibull cumulative distribution function can be an effective statistical tool for quantifying dogs' process on the target acquisition, providing

an objective metric to help trainers determine when dogs have reached asymptote and are ready to be transitioned to the testing phases. However, applying the Weibull function to the training data in the current study did not yield valuable insights regarding phase transitions. The model suggested that all dogs had reached asymptote almost immediately after training started without a significant acquisition period, which was possibly due to the high variability in their performance. "Sample age" could possibly be responsible for such variability, which provided dogs additional discriminative stimuli. Consequently, dramatically increased CR rates were achieved in some of the early sessions, causing an artifactually high mean accuracy in the early training stage, so that the function could not fit to a clear acquisition curve with such high starting point in accuracy. Therefore, while the Weibull function is a valuable analytical tool for stable learning data, it may lack reliability when applied to data sets with high variability.

4.4 Samples

4.4.1 Sample Collection

Sample collection plays an important role in stimulus control. To guarantee that our detectors will be learning the cancer concept during training, it is crucial to control for the unnecessary noise that might function as the confound variable, for example, the fingerprint of the collection sites. McCulloch et al. (2006) had remarkable results of sensitivity = 99% and specificity = 99%; however, the breath samples were collected from three different facilities. Different facilities have different odour profiles, and minor differences in staff and how they collect samples across sites can produce discrepancies in the VOC profiles of the samples. Moreover, if positive and negative sample collection is balanced across sites, this is not necessarily problematic; however, if most (or all) positive samples come from specific sites, dogs can use the site-specific VOC profile to identify the "targets" without learning the cancer profile. Therefore, even if the

research team provided identical collection handbook to all facilities, the samples obtained may still differ in the end, which could be responsible for the 'surprisingly' promising result. This factor was strictly controlled for in the present study with all samples collected by the same team and the same site since beginning; however, we may have encountered a similar issue a similar issue to some extent.

4.4.2 Sample Age

An additional finding from the current study suggests that dogs might be able to detect sample age. Although all samples were stored at -60 or -80, which significantly slows degradation, the process cannot be halted completely. During the training phase, an elevated specificity on session 276 was noticed by training; A subsequent review of the sample arrangement for that session and proceeding sessions revealed that most of the rejected negative samples employed that day were collected from patient 1000 and earlier, whereas several of the negative samples collected after this point were indicated by dogs. It was hypothesised that a rejection bias toward older samples and tendency to indicate fresher samples were likely developed in dogs. A statistical analysis had found a significant trend in samples utilised before session 276, specifically, across all samples (n = 224) and the negative sample subset (n = 155) but was absent within the positive sample subset (n = 69). To address this, samples collected from patient 1000 and earlier were subsequently excluded from the active sample pool. However, a post-adjustment analysis conducted at the conclusion of the study on samples utilised after session 276 suggested significant correlations between indication rate and sample age persisted across all samples (n = 139) and the negative sample subset (n = 88), while it remained non-significant across positive subset (n = 51).

This age disparity arose because samples were delivered from clinics in discrete, sequential batches. Due to the natural epidemiological proportion between positive and negative cases the

balanced ratio of positive and negative samples used to train and evaluate the dogs did not match the proportion of positive and negative samples collected from the clinic. Therefore, the positive samples within a given batch were depleted much faster than the negative samples; consequently, by the time a subsequent batch arrived, newer positive samples were introduced into arrangements, while the negative samples from previous batch were retained. This differential depletion rate became stronger as new batches arrived. By the initiation of the current study, a disproportion age between positives and negatives existed in the sample storage, which had predominantly consisted of newer positive samples collected after Dec-2022, and negative samples collected before Jan-2022. Two potential explanations may account for this observation:

1. **Time-dependent VOCs degradation and storage fidelity**: One possible cause is that the extended passage of time might have created an unintended "age tag", introducing a distinguishable extraneous variable between positive and negative samples. The inherent, time-dependent degradation of the samples' chemical profiles likely generated a distinct olfactory difference. Exhaled breath contains highly volatile organic compounds (VOCs) that naturally dissipate and break down over time (Shirasu & Touhara, 2011). While ultra-low temperature storage slows this process (Harshman et al., 2016), the fidelity of breath sample preservation remains heavily dictated by the total storage duration and the natural permeability of the sampling materials (Goss, 2019). Over years of storage, the continuous degradation of the VOCs in breath samples, perhaps also slow permeation of background VOCs might together alter the baseline olfactory signature of the substrate. It was likely that dogs were able to detect this quantitative or qualitative shift in the VOC profile. Also, given the situation in this specific study, the two classes of samples used in the initial training phase were collected from distinctive time spans, with negative samples relatively

older samples and indicate on fresher samples, which consequently resulted in an indication bias toward fresher samples.

2. **Unreported methodological shifts during the collection gap**: another possible factor involves unrecorded procedural or environmental changes at the clinical collection sites or team during the 10-month cessation of sample collection between Jan-2022 and Dec-2022. During this gap, changes may have occurred regarding clinic personnel, the breath collection material storage, sample transportation and storage protocols, or the ambient hospital environment (e.g., different sterilization chemicals). Although the clinical researchers endeavoured to maintain consistent procedures throughout the course of the study, even minor changes could introduce novel background VOCs into the samples collected upon the resumption of the project in late 2022. It is plausible that the dogs detected a distinct a categorical shift in the background VOC profile of the post-gap samples, which they subsequently learned to discriminate. In the post-adjustment analysis, it was noticed that dogs had different response pattern separated by this collection gap, which was also found in the pre-adjustment analysis (before session 276), where dogs tended to reject samples collection before Jan-2022 and indicate samples collected after Dec-2022. This might indicate a possible change had occurred to the sample collection team or sites during this 10-months gap.

However, it is worth noting that despite there being a significant indication pattern in their responses regarding sample age, better-than-chance accuracy was still achieved after the sample pool was adjusted in the last part of the experiment. According to the descriptive table presented in NLO analysis (Table 7), there was a significant difference between their indication responses toward positive and negative samples. While since sample age did unintentionally correlate with

the reinforcement contingencies during the initial training phases, it's possible that sample did disrupt discrimination. Such an outcome indicates that the concept learning still took place, but it may have been disrupted to some extent by this additional discrimination learning based on sample age, therefore, we could probably have seen higher performance in the absence of this unintended stimulus control; however, the present study was not designed to explicitly evaluate this possibility. While such stimulus control issues might heavily jeopardise the learning outcome in a simple discrimination task, the complex concept-learning task evaluated in the present study appears to be more resilient. This discrepancy suggests that the underlying mechanisms governing stimulus control may differ fundamentally between simple discriminations and complex concept learning, which need further investigations in future research.

4.5 NLO Probe Test

During the exploratory probe test, analysis revealed that the whole cohort's response toward NLO (non-lung-origin) cancer samples was significantly different from their responses to positive samples ($p < .05$), yet not statistically different from their responses to negative controls ($p = 1.000$). As for individual performances, Tui, Memphis, Mac and Ben were responding to NLOs at the same level as they responded to negative controls. The divergent performance of Bayley, who maintained a high indication rate on NLO samples (81.8%) comparable to her positive hit rate (84.0%), suggests that her indication responses may have come under the control of different volatile organic compound (VOC) patterns, perhaps a pattern that corresponded with a more general cancer profile. Despite this individual outlier, the overall cohort data provide preliminary evidence that dogs trained exclusively on lung originated positive samples came under precise stimulus control of lung-originated cancer volatiles and were able to discriminate the target disease and NLO cancer.

4.6 Future Study

Findings from the current study were informative, however, they also raised critical questions for future investigation. First, while this study investigated intermittent reinforcement (IR) schedules ranging from 60% to 100%, these rates remain relatively dense compared to what would likely be encountered in an applied clinical screening environment. In a real-world diagnostic setting, known positive samples must be systematically interspersed within screening arrays, reinforcing the hits to these samples, to prevent extinction of the indication response. However, because the prevalence of actual positive patients' cases in a screening population will be unknown and likely far less frequent than negative cases, dogs sometimes might have to experience extremely low reinforcement rates in such situations. Future studies should investigate the effects of more extremely thinned schedules of reinforcement (50% or lower) on dogs' detection accuracy, response bias, and resistance to extinction over prolonged exposure to such schedules.

In addition, establishing precise stimulus control is crucial in olfactory discrimination tasks. The current study inadvertently introduced sample age as an extraneous variable, which significantly impacted discrimination performance in the early stages of training. Although existing literature suggests that storing breath samples at ultra-low temperatures (e.g. –60^{o}C to -80^{o}C) preserves their viability for extended periods, the present findings demonstrate that detection dogs can still detect subtle time-dependent olfactory degradation between chronologically diverse samples. To prevent the situation, future studies must implement stricter temporal controls, including frequent renewal of the sample inventory and the disposal of older samples to ensure that stimulus control remains exclusively tied to the target.

The results of the exploratory probe test demonstrated that the cohort largely treated NLO cancer samples as negative controls—present a promising avenue for multidisciplinary research. Future studies could conduct chemical analyses (e.g., via gas chromatography-mass spectrometry) to compare the specific VOC profiles of breath from lung-originated cancers against those of NLO cancer. Identifying the precise chemical differentiators that the dogs use to make this behavioural distinction could provide profound insights into advanced medical diagnostics and help refine future target-acquisition training for detection animals.

Finally, data from the current study do not suggest that dogs have immediate applied utility for lung cancer screening. However, the indication thresholds were considered not fully maximized for most dogs. Tui, Memphis, and Bayley had positive indication thresholds set above 10 seconds, and their hit rate and engagement were still not affected, which might suggest that there could be more potential to further increase the threshold and further improve discrimination performance. Future research could try systematically increase these thresholds, as maximizing the response effort associated with an indication would likely further suppress the FA rate and elevate the CR rate, ultimately yielding higher overall accuracy.

Chapter 5: Conclusion

This study was directly prompted by observations from a previous unpublished study conducted in the same lab, which noted an unexplained decline in diagnostic accuracy during blind testing. Researchers hypothesized that the reduction in the reinforcement rate was a possible cause. While canine lung cancer screening fundamentally relies on complex olfactory concept learning, the impact of intermittent reinforcement on maintaining this specific type of task had not been formally investigated. Current findings demonstrate that fluctuating reinforcement rates between 60% and 100% do not exert a disruptive effect on dogs' performance. However, the resilience of these detection dogs' responses under more extreme schedule thinning, below 60%, remains unknown. Investigating these leaner schedules and exploring refinements to the training procedures which might improve their overall accuracy, are the next crucial steps before transitioning detection dogs into applied clinical screening environments.

Reference

Amundsen, T., Sundstrom, S., Buvik, T., Gederaas, O. A., & Haaverstad, R. (2014). Can dogs smell lung cancer? First study using exhaled breath and urine screening in unselected patients with suspected lung cancer. *Acta Oncologica*, *53*(3). https://doi.org/10.3109/0284186X.2013.819996

Barta, J. A., Powell, C. A., & Wisnivesky, J. P. (2019). Global epidemiology of lung cancer. In *Annals of Global Health*(Vol. 85, Issue 1). https://doi.org/10.5334/aogh.2419

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (2017). A study of thinking. In *A Study of Thinking*. https://doi.org/10.4324/9781315083223

Caldicott, L., Pike, T. W., Zulch, H. E., Mills, D. S., Williams, F. J., Elliker, K. R., Hutchings, B., & Wilkinson, A. (2024). Odour generalisation and detection dog training. In Animal Cognition (Vol. 27, Issue 1). https://doi.org/10.1007/s10071-024-01907-0

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). Applied behavior analysis (3rd ed.). In Pearson education, Inc (Vol. 1).

Crawford, M. A., Chang, C. L., Hopping, S., Browne, C. M., & Edwards, T. L. (2023). Influences of breath sample re-use on the accuracy of lung cancer detection dogs. *Journal of Breath Research*, *17*(1). https://doi.org/10.1088/1752-7163/ac9b7f

Crawford, M. A., Perrone, J. A., Browne, C. M., Chang, C. L., Hopping, S., & Edwards, T. L. (2022). Transitioning from training to testing with scent detection animals: Application to

lung cancer detection dogs. *Journal of Veterinary Behavior*, *55–56*. https://doi.org/10.1016/j.jveb.2022.07.004

Edwards, T. L. (2019). Automated canine scent-detection apparatus: Technical description and training outcomes. *Chemical Senses*, *44*(7). https://doi.org/10.1093/chemse/bjz039

Edwards, T. L., Browne, C. M., Schoon, A., Cox, C., & Poling, A. (2017). Animal olfactory detection of human diseases: Guidelines and systematic review. In *Journal of Veterinary Behavior: Clinical Applications and Research* (Vol. 20). https://doi.org/10.1016/j.jveb.2017.05.002

Edwards, T. L., Giezen, C., & Browne, C. M. (2022). Influences of indication response requirement and target prevalence on dogs' performance in a scent-detection task. *Applied Animal Behaviour Science*, *253*. https://doi.org/10.1016/j.applanim.2022.105657

Ehmann, R., Boedeker, E., Friedrich, U., Sagert, J., Dippon, J., Friedel, G., & Walles, T. (2012). Canine scent detection in the diagnosis of lung cancer: Revisiting a puzzling phenomenon. *European Respiratory Journal*, *39*(3). https://doi.org/10.1183/09031936.00051711

Fischer-Tenhagen, C., Johnen, D., Nehls, I., & Becker, R. (2018). A proof of concept: Are detection dogs a useful tool to verify potential biomarkers for lung cancer? *Frontiers in Veterinary Science*, *5*(MAR). https://doi.org/10.3389/fvets.2018.00052

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. v. (2001). The concepts of "sameness" and "difference" in an insect. In Nature (Vol. 410, Issue 6831). https://doi.org/10.1038/35073582

Goss, K. U. (2019). The physical chemistry of odors — Consequences for the work with detection dogs. Forensic Science International, 296. https://doi.org/10.1016/j.forsciint.2019.01.023

Guest, C., Pinder, M., Doggett, M., Squires, C., Affara, M., Kandeh, B., Dewhirst, S., Morant, S. v., D'Alessandro, U., Logan, J. G., & Lindsay, S. W. (2019). Trained dogs identify people with malaria parasites by their odour. In *The Lancet Infectious Diseases* (Vol. 19, Issue 6). https://doi.org/10.1016/S1473-3099(19)30220-8

Hackner, K., Errhalt, P., Mueller, M. R., Speiser, M., Marzluf, B. A., Schulheim, A., Schenk, P., Bilek, J., & Doll, T. (2016). Canine scent detection for the diagnosis of lung cancer in a screening-like situation. *Journal of Breath Research*, *10*(4). https://doi.org/10.1088/1752-7155/10/4/046003

Hall, N. J. (2017). Persistence and resistance to extinction in the domestic dog: Basic research and applications to canine training. In Behavioural Processes (Vol. 141, Issue Part 1). https://doi.org/10.1016/j.beproc.2017.04.001

Harshman, S. W., Mani, N., Geier, B. A., Kwak, J., Shepard, P., Fan, M., Sudberry, G. L., Mayes, R. S., Ott, D. K., Martin, J. A., & Grigsby, C. C. (2016). Storage stability of exhaled breath on Tenax TA. Journal of Breath Research, 10(4). https://doi.org/10.1088/1752-7155/10/4/046008

Herrnstein, R. J. (1979). Acquisition, generalization, and discrimination reversal of a natural concept. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(2). https://doi.org/10.1037/0097-7403.5.2.116

Jendrny, P., Twele, F., Meller, S., Osterhaus, A. D. M. E., Schalke, E., & Volk, H. A. (2021). Canine olfactory detection and its relevance to medical detection. BMC Infectious Diseases, 21, Article 838. https://doi.org/10.1186/s12879-021-06523-8

Jones, F. N., & Skinner, B. F. (1939). The Behavior of Organisms: An Experimental Analysis. The American Journal of Psychology, 52(4). https://doi.org/10.2307/1416495

Kassambara, A. (2025). rstatix: Pipe-friendly framework for basic statistical tests (Version 0.7.3) [Computer software]. https://doi.org/10.32614/CRAN.package.rstatix

Kassambara, A. (2025). rstatix: Pipe-friendly framework for basic statistical tests (Version 0.7.3) [Computer software]. https://doi.org/10.32614/CRAN.package.rstatix

Keller, F. S., & Schoenfeld, W. N. (2007). Principles of psychology: A systematic text in the science of behavior. In Principles of psychology: A systematic text in the science of behavior. https://doi.org/10.1037/11293-000

Maidodou, L., Clarot, I., Leemans, M., Fromantin, I., Marchioni, E., & Steyer, D. (2023). Unraveling the potential of breath and sweat VOC capture devices for human disease detection: a systematic-like review of canine olfaction and GC-MS analysis. Frontiers in Chemistry, 11, Article 1282450. https://www.google.com/search?q=https://doi.org/10.3389/fchem.2023.1282450

McCulloch, M., Jezierski, T., Broffman, M., Hubbard, A., Turner, K., & Janecki, T. (2006). Diagnostic accuracy of canine scent detection in early- and late-stage lung and breast cancers. *Integrative Cancer Therapies*, *5*(1). https://doi.org/10.1177/1534735405285096

Montes, Á. G., López-Rodó, L. M., Rodríguez, I. R., Dequigiovanni, G. S., Segarra, N. V., Sicart, R. M. M., Ferrández, J. H., FiblaAlfara, J. J., & García-Navarro, Á. A. (2017). Lung cancer diagnosis by trained dogs. *European Journal of Cardio-Thoracic Surgery*, *52*(6). https://doi.org/10.1093/ejcts/ezx152

Nevin, J. A. (2012). Resistance to extinction and behavioral momentum. Behavioural Processes, 90(1). https://doi.org/10.1016/j.beproc.2012.02.006

Pearce, J. M., & Redhead, E. S. (1993). The Influence of an Irrelevant Stimulus on Two Discriminations. Journal of Experimental Psychology: Animal Behavior Processes, 19(2). https://doi.org/10.1037/0097-7403.19.2.180

Ratiu, I. A., Ligor, T., Bocos-Bintintan, V., Mayhew, C. A., & Buszewski, B. (2021). Volatile organic compounds in exhaled breath as fingerprints of lung cancer, asthma and COPD. *Journal of Clinical Medicine*, *10*(1). https://doi.org/10.3390/jcm10010032

Riedlova, P., Tavandzis, S., Kana, J., Tobiasova, M., Jasickova, I., & Roubec, J. (2022). Olfactometric diagnosis of lung cancer by canine scent – A double-blinded study. *Complementary Therapies in Medicine*, *64*. https://doi.org/10.1016/j.ctim.2022.102800

Rooney N, Trivedi DK, Sinclair E, et al. Trained dogs can detect the odor of Parkinson's disease. *Journal of Parkinson's Disease*. 2025;0(0). doi:10.1177/1877718X251342485

Rudnicka, J., Walczak, M., Jezierski, T., & Buszewski, B. (2015). IS IT POSSIBLE TO DETECT LUNG CANCER BY TRAINED DOGS? *Health Problems of Civilization*, *2*. https://doi.org/10.5114/hpc.2015.57108

Sargisson, R., & Mclean, I. (2010). The effect of reinforcement rate variations on hits and false alarms in remote explosive scent tracing with dogs. *Journal of ERW and Mine Action*, *14*(3).

Shirasu, M., & Touhara, K. (2011). The scent of disease: Volatile organic compounds of the human body related to disease and disorder. In Journal of Biochemistry (Vol. 150, Issue 3). https://doi.org/10.1093/jb/mvr090

Trapold, M. A. (1970). Are expectancies based upon different positive reinforcing events discriminably different? Learning and Motivation, 1(2). https://doi.org/10.1016/0023-9690(70)90079-2

Voss, P., McCarthy, D., & Davison, M. (1993). Stimulus control and response bias in an analogue prey-detection procedure. Journal of the Experimental Analysis of Behaviour, 60(2). https://doi.org/10.1901/jeab.1993.60-387

Wasserman, E. A., & Bhatt, R. S. (1992). Conceptualization of natural and artificial stimuli by pigeons. In W. K. Honig & J. G. Fetterman (Eds.), Cognitive aspects of stimulus control (pp. 208).

Williams, H., & Pembroke, A. (1989). Sniffer dogs in the melanoma clinic? In *Lancet* (Vol. 1, Issue 8640). https://doi.org/10.1016/s0140-6736(89)92257-5

World Health Organization. (2021). Consultation on the use of trained dogs for screening COVID-19 cases: Geneva, Switzerland, 8th March 2021. https://www.who.int/publications/m/item/consultation-on-the-use-of-trained-dogs-for-screening-covid-19-cases

Zentall, T. R., Galizio, M., & Critchfield, T. S. (2002). CATEGORIZATION, CONCEPT LEARNING, AND BEHAVIOR ANALYSIS: AN INTRODUCTION. Journal of the Experimental Analysis of Behavior, 78(3). https://doi.org/10.1901/jeab.2002.78-237

Zentall, T. R., Wasserman, E. A., & Urcuioli, P. J. (2014). Associative concept learning in animals. *Journal of the Experimental Analysis of Behavior*, *101*(1).

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K. R., & Rattermann, M. J. (2008). Concept Learning in Animals. Comparative Cognition & Behavior Reviews, 3. https://doi.org/10.3819/ccbr.2008.30002

Appendices

Appendix A: "Cancer project planning"

The whole planning process contains three parts:

- **Proposed training schedules:** The goal of this part is to locate specific samples for the following training days. Required sample states will be arranged by Tim; trainers need to find the according samples and make a "catalogue" of the plan.
- **Sample orders**: This is to arrange the selected samples in both small-to-large and random order.
- **Data sheet:** The goal is to fill the planned details (sample order and sample type) into the manual data sheet. This is for the trainers to record training session "live".

Part one: Proposed Training Schedules

1. Open 'Sample Plan' and 'Proposed Training Schedules' Excel files in the "Weekly Planning" folder. Click on the "Schedule" tab (bottom of the page). Sample status arranged by supervisor are on this page. Find the according day, copy the whole column beneath the date. (8-April as an example)

*** The code: The code indicates the current state of the sample and is made up by three parts. Each part represents one dimension of the sample.

- P/N: P stands for positive, and N stands for negative.
- A/B: Since two breath samples were collected from each patient, there are two samples in the sample bag corresponding to each sample number when they were delivered to the lab. The A/B labelling is for us to differentiate the two.
- 0/1/2: how many times the sample has been used.

e.g. NB2: this sample is negative; it is the second sample of this patient; it has been used twice.

2. Go to “proposed training schedule”. Create a table like the one shown in the picture, colour the table with green or orange to distinguish between positive and negative. Green for positive samples on the top and Orange for negative samples on the bottom. Name the page with according to session arrangement number.

3. Paste the copied sample types to the “State” column, using ctrl+shift+v (Windows) or command+shift+v (Mac), so that the copied information from another page can fit the form of this page. Ensure that positive samples are green and negative samples are orange.

4. Return to the “Sample Plan” and click on the “Sample” tab (bottom of the page). The database of all samples is on this page.

5. Find the filter ‘code’. Ensure that all columns are being filtered (each column label should have the small arrow next to it). Click the downward arrow next to CODE, then a list of codes will drop, and all codes are ticked as default.

6. Click on “Select all” box to unselect everything.

7. Click the sample type that is proposed in the plan and click “apply” tab. For example, when searching for an PA0 the filters will look like this:

8. Find the filter ‘randomizer’, which should be at far right. Again, ensure all columns are being filtered. Click the downward arrow next to Randomizer, then click “sort smallest to largest”. Top ones are the randomized samples for the plan.

9. Copy all relevant fields (in blue) for the samples for the plan. Return to the 'Proposed Training Schedules' document and paste the chosen sample into the schedule (ctrl + shift + v for Windows; command + shift + v for Mac). (9999B1 is for example, not real patient.)

*** Repeat step 4 - step 9 for each sample.

10. Pay attention to the "Cancer status" on the right side of table. The proportions of smokers among negative sample and positive sample are displayed here, the "diff" should be within ±25%.

*** This is for training phase, because dogs might discriminate smokers and nonsmokers if disproportional smokers exist in training samples. Manual adjustment is needed if the "diff" falls outside of the ±25%. Either adding or taking out smokers from one side (negative or positive) would help maintain the "diff". Return to the "sample plan", select the code that is needed to be adjusted. Samples are already randomized, only need to find the next "smoker" or "nonsmokers" after the chosen samples.

11. Send proposed plan for that day to supervisor for approval. Once the plan for that day is approved. Update the use count for each sample that will be used for that day. Return to "Sample plan", find the samples that were decided, and plus 1 on the use count. PA0 will turn into PA1 automatically.

12. Once the use counts are updated, it is safe to make plan for another training day. (repeat the step1- step12).

*** Samples are random selected from the grand sample pool using "randomizer" in excel. However, to prevent the same breath sample from being presented on consecutive training days,

a manual review and replacement procedure was enforced. A mandatory "lockout period" of four training days (equivalent to two calendar weeks) was applied to every sample following its use.

For example, if Sample 9999B1 was randomly selected and presented on last Monday, then the next time any samples from patient 9999 could include is next Thursday. If the randomiser selected a sample from that patient during this lockout period, this sample manually excluded from the arrangement and the researcher would then repeat the randomised sampling procedure to find a replacement sample that satisfied the lockout criteria.

Part two: "Sample Orders"

Once all the samples are settled for upcoming training days. Start arranging the sample orders for training sessions.

1. Select the column under "Sample" from approved schedule in the "Proposed training Schedule".

2. Open the 'Weekly Printing' folder. Click on the 'Sample Order' excel file. Go to the 'Programme Iterations' tab. Paste the sample numbers under 'patient' column. (using ctrl + v, or command + v; make sure to paste the corresponding colour into this sheet too)

3. Manually filling the 'type' column. 1 for positive and 2 for negative.

4. Go back to 'Proposed Training Schedule', copy the information under "State" column. Paste them into the 'States' column in 'Programme Iterations'.

5. Create a new tab for this Training Day. Rename with its day and date. (We had Tue 4.8 already; this is for example. Sheet 2 should be renamed as Tue (4.8)).

6. Go back to ‘Programme Iterations’. Click the downward arrow next to the ‘patient’ filter (make sure to filter all three tabs). Click ‘Sort smallest to largest’ and then copy and paste the ordered sample numbers and their ‘state’ information to the ‘Ordered’ and ‘State’ column in the new sheet.

*** ‘Ordered’ is for trainers to find the samples in the freezer. All the samples are sorted in the smallest to largest order across 11 boxes. The location information of the samples is printed on a sheet and sticked onto the wall behind the freezer. The location information is dynamic. As the new sample coming in and the old sampled tossed away, might need to be updated every time new samples arrive to the lab.

*** ‘State’ is for trainer to write up post-it notes on the training day, during the defrosting period. E.g. it says NB1 for ‘0007B2’, then the post-it note should be ‘0007B2 on the top, NB2 in the middle, and 08/04/2025 at the bottom’. (Details seen in ‘Cancer Project’ - ‘SOPs’ - ‘Selecting and Loading Samples’ - ‘Adding post-it notes’)

7. Go back to ‘Programme Iterations’ tab and find the ‘Rand’ filter, click the downward arrow next to it. Click ‘Sort smallest to largest’ and then copy and paste the randomized sample numbers and their types (1/2) to the ’Random’ and ‘Type’ column in the new sheet.

*** ‘Random’ column is the trials order for training sessions. ‘Type’ column indicates whether corresponding sample is positive or negative for lung cancer.

*** Each information needs to be put into config file setting for the apparatus before training.

8. Print them out and put into the binder. Repeat these steps for another training day if there is any.

Part Three: "Data Sheet"

1. Open the 'Weekly Planning' folder. Click on the 'Data Sheets' document. delete any prior sample orders and adjust the date at the top of the page. Put Dog's name and their assigned numbers on the top left corner of the sheet.

2. Go back to 'Sample Order' sheet of the according day. Copy the 'Type' column.

3. Paste to the column, right next to the trail numbers, which should be column E in this picture.

4. There should be two participant dogs in each sheet. Therefore, this is what a 'data sheet' look like in the end.

5. For A0 samples, mark the 1 or 2 with a * (e.g. 1*), and a ^ (e.g. 1^) for B0 samples.

6. Do all the participant dogs and print them out.

Appendix B: Select and Loading Samples

1. Put on latex/nitrile gloves.

2. Line up trays on the table.

3. Open the deep freezer.

4. Pull out the preparation box and place it on top of the freezer.

5. Put on new pair of gloves.

6. Find the first sample in the arrangement.

7. Open the mother bag.

8. Grab the appropriate sample out using thumb and index finger.

9. Place the sample with its mini bag in the first tray.

10. Close the mother bag without using the thumb and index finger.

11. Put the mother bag back into the preparation box.

12. Throw the gloves into the bin.

13. Put on new pair of gloves.

14. Repeat above steps for all samples.

16. Put on new pair of gloves.

17. Write post-it notes for each sample.

18. Put sample numbers on the top, ‘becoming state’ in the middle and the date at the bottom.

19. Stick the post-it notes on the side of the trays without touching trays.

20. Repeat step 17 – step 19 for all samples.

21. Open TimScent files.

22. Open config files for the first dog appears on the screen.

23. Manually update the sample types according to today's plan.

24. Copy and paste them to other config files.

*** Once the defrosting time has reached 40 minutes.

25. Put on latex/nitrile gloves.

26. Open the first sample bag and remove the post-it notes.

27. Hold down one of the sample bungs and take off the tape from one side.

28. Carefully wiggle off the bung.

29. Remove the tape from the other end of the sample.

30. Carefully wiggle off both bungs and place the sample first sample in segment 1.

31. Repeat step 21 – step 26 for all 17 samples.

32. After loading, put on new pair of gloves.

33. Place the lid securely on top of the apparatus.

Appendix C: Sessions

1. Open ISpy on the left monitor.

2. Open TimScent on the right monitor.

3. Make sure sample order and types are updated.

4. Type dog's code into the *Subject* box.

5. Select start.

6. Push *Record All* on Ispy.

7. Open the door to the testing room.

8. Open the crate door and lead the dog to the testing room.

9. Close the door once the dog is in the room.

10. Record the first instance of a *BeamStart* program response as the start time for the session.

11. Put 1 in the Ind. Column, if a positive indication is made (Keeping nose into the port for threshold time).

12. Put 2 in the Ind. Column if a negative indication/rejection is made (lever pressing).

13. Put a 'tick' in the accuracy column if the response matches the sample type.

14. Put a 'cross' in the accuracy column if the response does not match the sample type.

15. Go into the room and prompt, if dogs cannot make a positive indication to positive sample for three times (original trial doesn't count)

16. Record the *ExpEnd* time.

17. Open the door of testing room.

18. Lead the dog back to their crate.

19. Close the door of testing room.

20. Open the event file of the corresponding dog.

21. Record the temperature and humidity.

22. Double check the data sheet. Go back to the event file if anything is missing or unsure.

23. Calculate the hit rate and reject rate.

24. Repeat step 3 – step 23 for all dogs.

Appendix D: Unloading and Samples Storage

1. Put on latex/nitrile gloves.

2. Lift the apparatus lid and place it to the right of the apparatus the underside facing outwards.

3. put on new pair of gloves.

4. Pick up the first sample and push the bungs into each end.

5. Hold the bungs down with your fingers and take the tube over to the table.

6. Wrap the tube with post-it note on the side of trays.

7. Bring the tube to the taping station.

8. Tape down both ends.

9. Add the sample to the sample bag.

10. Squeeze out any excess air before sealing.

11. Put the gloves in the bin.

12. Put on new pair of gloves.

13. Repeat these steps for all 17 samples.

14.Pull out the preparation box and place it on top of the freezer.

15. Put on latex/nitrile gloves.

16. Take out the first mother bag from the box and open it.

17. Add the sample bag into the mother bag using thumb and index finger.

18. Close the mother bag without using the thumb or index finger.

19. Put the mother bag back into the sample box.

20. Put the gloves into the bin.

21. Put on new pair of gloves.

22. Repeat these steps for all 17 samples.

23. Put the box back into the deep freezer.

24. Transfer all trays and put them on the dog scales by the sink, ready to be cleaned.

Appendix E: Cleaning

Cleaning Segments:

1. Put on latex/nitrile gloves/mask and safety goggles.

2. Clean the sink and the right side of sink bench with 70% IPA.

3. Put the plug into the sink.

4. Take a dishwashing tablet and peel, then put into the sink.

5. Fill the sink half full of hot water.

6. let all sides merge in the water

7. Scrub the flap with hands carefully.

8. Submerge the flap side again.

9. Arrange them standing up at the right side of the sink. (Not letting sides touch each other)

10. Put on new pair of gloves.

11. Clean the left side of the sink with 70% IPA.

12. Rinse away the soap around the segments with cold water.

13. Placing them on the left side of the sink.

14. Dry the right side of the sink and cover it with a layer of paper towels.

15. Take the 50% IPA container to the sink and remove the lid.

16. Briefly submerge all sides of each segment in the 50% IPA container.

17. Hold the segment, letting 50% IPA drip back to container.

18. Place them on the right side of the sink.

19. Dry the left side of the sink and cover it with a layer of paper towels.

20. Arrange them standing up at the left side of the sink. (Not letting sides touch each other)

Cleaning Apparatus and Testing Room:

1. Put on latex/nitrile gloves and safety goggles.

2. Bring all the kibble back to the box.

3. Clean the feeder with 70% IPA.

(4. Cleaning the inside of the feeder monthly.)

5. Clean the floor underneath the feed bench with 70% IPA.

6. Clean the apparatus lid 70% IPA.

&. Clean the apparatus base with 70% IPA.

7. Clean the apparatus front board, port, and lever with 70% IPA.

Cleaning Trays:

1. Put on latex/nitrile gloves/mask and safety goggles.

2. Wipe down the top of the deep freezer with 70% IPA.

3. Roll out paper towels to cover the top of the deep freezer.

4. Put on new pair of gloves.

5. Clean the sink and the right side of sink bench with 70% IPA.

6. Put the plug into the sink.

7. add a squirt of dishwashing liquid.

8. Fill the sink half full of hot water.

9. Submerge the trays inside the soapy water.

10. Arrange them upside down on the right side of the sink, creating three stacks of staggered trays.

11. Put on new pair of gloves.

12. Clean the left side of the sink with 70% IPA.

13. Rinse away the soap in the trays with cold water.
14. Placing them on the left side of the sink.
15. Dry the right side of the sink and cover it with a layer of paper towels.
16. Take the 50% IPA container to the sink and remove the lid.
17. Briefly submerge both sides in the 50% IPA container.
18. Hold the trays, letting 50% IPA drip back to container.
19. Place them on the right side of the sink.
20. Bring them to the freezer and align them upside down separately.

Appendix F: Criteria for Changing Threshold

Key words:

1. Threshold: The amount of time a participant dog needs to spend on nose holding in the port to be considered as an indication.

2. Indication: Participant dogs holding their noses in the port for certain amount of time (= threshold set for this dog).

3. Rejection: Participant dogs pull out their noses and press the lever before the threshold time is reached.

4. False Alarm: Participant dogs holding their noses in the port for ≥threshold set for this dog, when it is a negative sample.

5. Miss: Participant dogs pull out their noses and press the lever before the threshold time is reached, when it is a positive sample.

Increasing:

Increase the indication threshold by 500 ms when the hit rate reaches 100% and the correct rejection rate below 50% for three consecutive sessions.

Decreasing:

Decrease the indication threshold by 500 ms if Hit Rate drops below 66.67% but the Correct Rejection rate remains above 50% for 3 consecutive sessions.

Engagement:

- If it took longer than 90 seconds for a dog not interacting with the apparatus at least twice during a session for 3 consecutive sessions. Consider dropping the threshold by 500 ms.

- A “beep” prompt is delivered every 30 seconds during dogs’ disengagement with the apparatus (“debug” – double click “Beep”).

- The trainer should enter the testing room and physically prompt the dog to complete the trial if the dog has been disengaging more than 90 seconds.