Journal of Veterinary Behavior 35 (2020) 38–44

Contents lists available at ScienceDirect

Journal of Veterinary Behavior

journal homepage: www.journalvetbehavior.com

Canine Research

How dogs learn to detect colon cancer—Optimizing the use of training aids

G. Adee A. Schoon [a,*], Danielle De Jonge [b], Patrick Hilverink [b]

[a] Animal Detection Consultancy, Vorchten, The Netherlands
[b] Royal Dutch Guide Dog Foundation, Amstelveen, The Netherlands

ARTICLE INFO

Article history:
Received 22 March 2019
Received in revised form
16 September 2019
Accepted 2 October 2019
Available online 10 October 2019

Keywords:
dogs
colon cancer
detection
generalization
learning

ABSTRACT

The use of dogs in the detection of cancer is booming. Many dedicated dog trainers have begun to pursue this field, helped by doctors that provide them with samples. Unfortunately, samples are usually from a very limited number of patients, and matched controls are usually lacking. Testing is not always as rigorous as it should be. In this study, training and testing has been integrated to optimally use a limited number of samples. Two groups of dogs (5 and 3) have been trained at the Royal Dutch Guide Dog Foundation (KNGF Geleidehonden) using stool samples in a carousel setup. By routinely testing samples from new patients and controls before including them in the training odor set, valuable information on how the dogs learn has been gathered, especially that pertaining to how they generalize and develop an odor concept. Using such a strategy provides insight into conditions that need to be monitored for the training of dogs to become successful as a diagnostic tool.

Introduction

The detection of diseases such as cancer using dogs or other animals has been the subject of attention since the first story appeared in the Lancet in 1989 (Williams and Pembroke). Since then animals, mainly dogs, regularly appear in both the popular and scientific literature as possible assistants in disease detection. In a number of areas, such as diabetic alert dogs and dogs that assist people that suffer from epilepsy, dogs are being used although results have not yet been scrutinized by scientific research. In other applications, samples are taken from patients with cancer and controls to train dogs and results are being published by behavioral researchers and medical doctors. Usually, they use a technique where samples are offered in a lineup or carousel, and the training protocol used consists of shaping a trained response to the odor of samples from patients with cancer. This technique has been described as "remote scent tracing" when it was applied in the area of mine detection. This technique has been approved for the clearance of air cargo by dogs in the EU, documented for the detection of corrosion under insulation in gas and oil plants (Schoon et al., 2014) and is in operational use for the detection of tuberculosis by African pouched rats in Tanzania and Mozambique by Apopo. Fundamental aspects of protocol, training, and operational have been the topic of discussion (Goldblatt et al., 2011; Goldblatt 2017).

A recent review article (Edwards et al., 2017) summarized publications on the results obtained by biodetectors in medical research, focusing on essential aspects of remote scent tracing technology. Although the results seem promising, the training of animals involves the availability of many samples, both from patients and matched controls, a careful training and testing protocol, and realistic testing using completely new samples from both patients and controls. The review concluded that no study met all essential criteria.

In reality, there is usually only a limited sample set available to train and test dogs. One important aspect that is often overlooked is the necessity to gather samples before patients are diagnosed and treated to prevent any other kind of cue for the animals. This leads to a slow collection process even if a hospital is fully cooperative and all medical ethical requirements are met.

Another aspect that has been overlooked is the learning process of the animals involved. Stimulus generalization, defined as responding to a new (thus untrained) stimulus in the same manner as to a conditioned (thus trained) stimulus, is a crucial element in training detection dogs. To be useful detectors, dogs have to respond to stimuli that they have never been trained on and that

* Address for reprint requests and correspondence: G. Adee A., Schoon, Animal-Detection Consultancy, Vorchter Enkweg 4, 8193KM Vorchten, Netherlands.
E-mail address: adee.schoon@planet.nl (G.A.A. Schoon).

https://doi.org/10.1016/j.jveb.2019.10.006
1558-7878/

may vary in terms of concentration and composition as defined by the organization using such dogs. For example, a drug dog trained on a small amount of marijuana grown in Morocco will be expected to respond in a similar manner to a larger amount that was grown in Algeria. Although both are the same plant, the smell of the two is quite different even to human noses.

In general, dogs are trained with one sample set, and then (ideally) tested double-blind with another sample set from new patients and controls. If successful, this tells you the size of the sample set used was adequate. Dogs are capable of doing this. For example, Wright et al. (2017) trained dogs on a range of accelerants, using 20 targets and 20 controls. The authors then tested the dogs using new, untrained accelerants. During this generalization testing, the dogs responded to the new accelerants significantly better than chance, and the authors concluded that dogs could "assign novel odors to a known category." Unfortunately, no details were given on the type of accelerants used in training, nor on the type or number used in testing. Only half of the dogs reached the testing stage and the response of 2 of the 3 dogs in the testing stage was less accurate than during training.

It is unknown how much variation in stimuli is required for the animals to become operationally reliable detectors, but it is obvious that it is impossible to train them all. From explosive detection training, there are several studies that demonstrate that training on a limited number of targets limits stimulus generalization: while responding adequately to the "known" targets, they miss "unfamiliar" targets (Oxley and Waggoner, 2009; Goldblatt et al., 2011; Kranz et al., 2014; Lazarowski and Dorman, 2014). Determining how much variation is necessary to ensure stimulus generalization is important because explosive detection dogs are relied on for our security.

In a structured search setup with defined sample units, it was also found that dogs spend less sniffing time on true negative samples (Concha et al., 2014) in comparison to other choices. Learning what not to respond to is also part of the learning process, as has shown to be the case with rats in odor discrimination learning (Lu et al., 1993).

In 2013, KNGF Geleidehonden began a collaboration with the Amsterdam University Hospital VUmc for their Medical Detection Dog program. VUmc provided a set of stool samples of patients with colorectal cancer (CRC) and controls to be used in a pilot study. The number of samples was insufficient to conduct both training and testing, so the focus of the study was to use the samples in an optimal manner to study the learning behavior of the dogs, following a scheme inspired by Oldenburg et al. (2016). In this study, a schedule of training on one target followed by generalization testing using a new target was proposed. If the test failed, the testing target became a second training target. When the dog had learned both targets, a new generalization test using a third new target would be conducted. This rhythm of training and generalization testing, as proposed by Oldenburg, was chosen in this study because it would provide information on how well the targets had been generalized into the "colon cancer" group. After a first group of dogs was trained, a second smaller group of dogs was trained in 2016, in a similar manner using the same material to evaluate the robustness of the training method. This article presents the results of this project.

Material and methods

Dogs and handlers

Two groups of dogs (5 and 3) were used for the pilot project (see Table 1). The dogs were all kept in family homes and were present at the training location 2-3 times per week, except when circumstances such as illness prevented this. The dogs were given the normal high-quality care all KNGF Geleidehonden dogs are given (guidelines in Dutch: https://geleidehond.nl/pagina/over-ons/onze-organisatie/dierenwelzijn).

The dogs were handled in rotation by 2 handlers. These handlers both worked at the KNGF and had over 25 years of experience in training dogs in various fields including tracking, scent discrimination, and search and rescue. They trained the medical detection dogs as a part of their jobs at KNGF. Later, other assistants were added to the project to assist in sample preparation and documentation.

Patients and sample preparation

The stool samples that were used in this pilot study had been collected from patients at the VUmc in Amsterdam during 2007-2011. They were collected before colonoscopy or cancer treatment, homogenized in stabilization buffer (property of Exact Sciences), and stored at −80°C. At the onset of the study, the only information available was the age and gender of the patients, the disease status (CRC or control), and the Union for International Cancer Control stage of the cancer (0-4). Samples from patients with other colon problems such as benign tumors, Crohn's disease, or celiac disease were not included in this study. Later in the project, fecal occult blood scores were obtained from approximately half the patients and controls. The patients with CRC and the controls were not fully age and gender matched, as can be seen in Table 2.

The homogenized buffered stool samples were thawed and aliquoted in approximately 30 portions per patient into 1 mL Eppendorf tubes. These samples were then transported to the training location and stored at −40°C. Samples that were necessary for training were taken out of the freezer, opened, and placed in closed 30 mL plastic containers where they were thawed for 1 hour at room temperature. Scent samples were prepared by placing ca. 7 cm^2 pieces of cotton wool (1/4 of cotton facial cleaning pads with a diameter of 6 cm) into the containers containing the thawed stool sample: one piece of cotton wool in each capped container. These were kept there for 15 minutes, and then the stool samples were removed from the container and refrozen. The containers with the pieces of cotton were stored in the refrigerator and used for training the dogs the next day. The samples were taken out a few hours

Table 1
Overview of dogs

Group	Dog	Breed	M/F	Born	Remarks
1	Catoo	Nova Scotia duck tolling retriever	F	Nov. 2011	Privately owned
1	Glenny	Labrador × golden retriever	F	May 2012	KNGF breeding program (rejected as guide dog)
1	Mintha	English cocker spaniel	F	June 2013	Working dog line
1	Moo	English cocker spaniel	F	June 2013	Working dog line
1	Zorah	Labrador × golden retriever	F	March 2013	KNGF breeding program (rejected as guide dog)
2	Naomi	Labrador × golden Retriever	F	December 2014	KNGF breeding program (rejected as guide dog)
2	Nilson	English cocker spaniel	M	July 2016	Working dog line
2	Robin	Belgian Malinois	F	December 2016	Working dog line

Table 2
Overview of cancer patients and controls

	N	Average age (±SD)	% Male	FOB scores			
				<25		>1000	n/a
Patients	10	70.2 (±9.0)	60%	2	2	4	2
Controls	60	57.4 (±13.8)	39%	26	3	0	31

FOB, fecal occult blood; n/a, not available.

before training and 15 minutes before training the caps were taken off. All handling of stool and scent samples was conducted while wearing thin polyethylene gloves, using pincers and taking care to prevent (cross) contamination. Stool samples from patients with cancer were reused up to 6 times, from controls up to 8 times.

Training

The training of the dogs was conducted in a manner very similar to those used for the detection of corrosion under insulation (Schoon et al., 2014) and consisted of the following phases:

1) Pretraining outside carousel room, where the dogs were trained to search in different locations and to perform a passive (sit/lie down) indication for progressively smaller pieces of Kong (a dog's toy) and in a gradually more structures set up using containers, using successive approximation, and positive reinforcement with a full Kong, another toy, or food;
2) Kong discrimination phase in a line of containers that gradually also included stand-alone arms of the carousel, where pieces of Kong were replaced with scent samples of Kong made on 1/4 pieces of cotton facial cleaning pads, gradually introducing scent samples from distracters, and balancing the reward for the individual dogs to maintain motivation and prevent overexcitation;
3) Pretraining inside the carousel room, where the dogs were trained to systematically search the carousel in an anticlockwise manner, detect a sliver of Kong placed inside a container in the carousel among scent samples from distracters, perform a 2- to 3-sec passive indication, not make false alarms, and perform a number of zero runs (without Kong in the carousel) within a training session;
4) Transition phase, using scent samples from stool of 1 patient with cancer (CRC stool odor) and 10 controls, the samples from the patient with cancer initially combined with a sliver of Kong and fading this out, dogs still working the carousel systematically, performing a passive indication, not making false alarms, and being able to perform a number of zero runs. The two groups were trained using a different patient with cancer and different controls in this phase;
5) Full-scale training following the routine described in the following; handler usually blind to odor placement, gradual introduction of samples from new patient with cancer and controls in test sessions; after introduction samples from this person (cancer patient or control) were added to the general set of training odors. Group 1 was introduced to new odors gradually over a period of 99 sessions (average 0.6 per session), group 2 much more quickly over a period of 30 sessions (average 2 per session). Occasionally during this period, additional training sessions were conducted to solve particular training issues. This included repeating a choice to give the dog an additional reward opportunity, or repeat a run without a target in it to extinguish a false alarm, reinforcing correct rejection of all control samples (described in the protocol). No new odors were added during these sessions. Another aspect of such training was training the "praise off" procedure as a form of variable reinforcement. This was performed by calling the dog away from a correct response/true positive ("hit") without its usual reward, taking it back to the waiting room, asking it for a simple other behavior and rewarding that verbally or with a piece of food.

Protocol

The layout of the room is illustrated in Figure 1, a picture of the 8-arm carousel is given in Figure 2. The observer could communicate with the handler in the waiting room through a microphone.

A training session consisted of 4-5 runs that each contained 8 samples. In the full-scale training (step 5 above), there were 3-4 CRC stool odor samples, the others were control stool odor samples. Each run could contain 0, 1, or 2 CRC stool odor samples. Each run was conducted for each of the dogs as follows:

- Dog and handler enter the waiting room together through separate door;
- Dog waits in the waiting room (where a hatch opens into the carousel room);
- Handler moves from the waiting room to the carousel room and stands behind a one-way screen;
- Handler opens the hatch and the dog comes in;
- Dog searches the carousel anticlockwise.

One of five different things could then happen:

- Hit (true positive): the dog correctly indicates a (known) CRC stool odor sample;
- False positive (FP): the dog incorrectly indicates a (known) control sample;
- Correct rejection (true negative): the dog smells all the samples properly and correctly does not indicate any of them;
- Miss (false negative): the dog smells all the samples but does not indicate a known CRC stool odor sample;
- Unknown hit: the dog hits a sample of unknown status (double-blind evaluation).

The handler observed the dog through a one-way window in the screen in the carousel room and remained behind the screen until

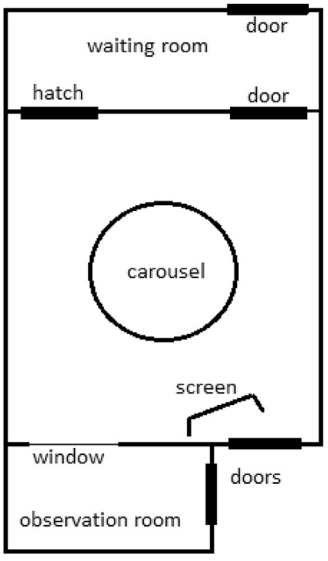

Figure 1. Setup of experimental facility.

Figure 2. The carousel.

the dog had made a choice or had smelled all the samples properly. An observer also watched the dog through a one-way window of the observation room and signaled the handler with lights when the dog indicated or when the dog had made a correct rejection. The handler followed up on this signal. The signals and the follow-up are described in the following:

- Hit: the observer gives a green light, the handler gives the clicker word "yes," comes out, and rewards the dog. Then the two go back to the waiting room where rewarding continues and wait for the signal to start again;
- FP: the observer gives a red light; the handler takes the dog back to the waiting room. There they wait for the signal to start again;
- Correct rejection: after the dog has smelled all samples, the handler terminates the visit by calling the dog away from the carousel and moving from behind the screen. The observer gives a combined light signal. The handler rewards the dog; the two go back to the waiting room where rewarding continues and wait for the signal to start again;
- Miss: after the dog has smelled all samples, the handler terminates the visit by calling the dog away from the carousel and moving from behind the screen. The observer gives a red light to indicate the miss. The handler takes the dog back to the waiting room where they wait for the signal to start again;
- Unknown hit: obviously, an unknown hit cannot be rewarded. But not receiving an expected reward can be highly frustrating. To keep stress from building up and affecting motivation and performance, the following "praise off" procedure was followed. After hitting a sample of unknown status, the observer would signal this with an orange light. The handler takes the dog back to the waiting room, asks the dog for a simple behavior that is rewarded, and waits for the signal to start again.

A visit ended when the handler returned to the waiting room with the dog. The assistant then entered the carousel room through the other door. If one sample had been hit but not all samples had been smelled (such as after a hit, a false alarm, or an unknown hit), the assistant removed that particular sample and rotated the carousel for the next visit of the dog. If all samples had been smelled and none responded to (as after a correct rejection or a miss), the assistant removed all samples from the carousel and put in the set of samples from the next run for that dog.

In the full-scale training, training sessions were designed using a custom-designed computer program that allowed for manual selection of the stool samples to be used. Stool samples from 2-4 different patients with cancer were used, and maximum 2 samples from each control, taking care to balance the number of times stool samples were reused. The computer program randomized the placement of these samples into 4-5 runs of 8 samples each. The lab assistant prepared scent samples from these stool samples as described previously.

All scent samples in a training session were analyzed by one dog first before moving on to the next dog. Because the position of the targets was randomized over the arms, any arm could contain a target or control sample at any time. The dogs worked in a random rotation schedule, and the arms of the carousel were cleaned using a paper towel and clean water between dogs. The computer program was also used for data entry (all hits of the dogs were recorded), and data were exported to Microsoft Excel for further analysis. Test sessions (where samples from new patients with cancer or new controls were introduced) were prepared and conducted in the same manner.

Results

The data collected in phase 5, the full-scale training and testing, are presented here. In this phase, the dogs were no longer given any other cue than the odor of the stool samples because the odor of the Kong had been completely faded. Data from training sessions that deviated from the protocol to address training issues were not included.

In Table 3, the average results of the individual dogs are given for all the training/testing sessions that were conducted: 99 for group 1 and 30 for group 2. They are divided into results obtained on familiar patients with cancer/controls, and results obtained on new patients with cancer and controls at their introduction in test sessions. Not all dogs participated in all the training sessions because

Table 3
Average results per dog on familiar and new samples from patients with cancer and controls in percentages

	Patients with cancer						Controls					
	Familiar			New (test)			Familiar			New (test)		
	N	Hit	Hit rate (%)	N	Hit	HR(%)	N	FP	FP rate (%)	N	FP	FP rate (%)
Catoo	331	256	77.3	17	12	70.6	2363	62	2.6	94	9	9.6
Glenny	366	318	86.9	17	15	88.2	2608	95	3.6	94	14	14.9
Mintha	370	322	87.0	17	15	88.2	2636	130	4.9	94	19	20.2
Moo	360	295	81.9	17	13	76.5	2567	110	4.3	94	14	14.9
Zorah	364	304	83.5	17	14	82.4	2595	100	3.9	94	15	16.0
Naomi	89	73	82.0	18	14	77.8	751	31	4.1	102	5	4.9
Nilson	89	74	83.1	18	14	77.8	751	19	2.5	102	7	6.9
Robin	89	81	91.0	18	14	77.8	751	31	4.1	102	5	4.9

FP, false positive; HR, hit rate.

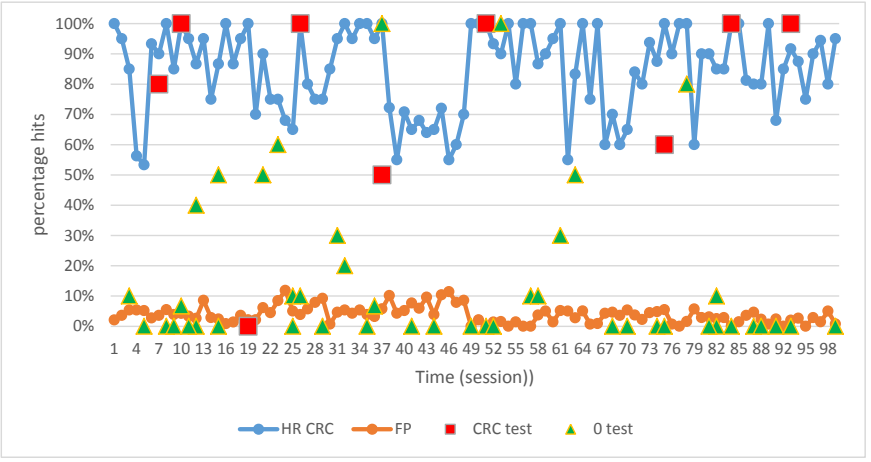

Figure 3. Training and testing results averaged over the 5 dogs in group 1 in time (sessions). HR CRC is hit rate on samples of familiar patients with cancer; FP is false positive rate on samples of familiar controls; CRC test is hit rate on samples of new patients with cancer, 0 test is FP rate on samples of new controls. HR, hit rate; CRC, colorectal cancer; FP, false positive.

of holidays, illnesses, or other logistic problems, but they all participated in the test sessions. The average hit rate (HR) on new patients with cancer did not differ significantly from the overall HR (χ^2, ns) for either group. The false alarm rate on new controls was significantly higher (χ^2, $P < 0.001$) for group 1 but not for group 2. Group 2 was being trained with stool samples that had also been used by group 1 and had been thawed and refrozen several times. This did not have an overall negative effect on their results within the limits of this study.

Looking at the results in time provides a view into the learning process. In Figures 3 and 4, the average results of the two groups of dogs per session on familiar samples are presented in two solid lines: HR on patients with cancer and FP rate on controls per training session. The HR is based on 2-4 colon cancer scent samples per training session where each of these was smelled (at least once) by all the dogs in a group. It follows that for group 1 (5 dogs), the HR is based on 10-20 sniffs per session, and for the second group (3 dogs) on 6 to 12 sniffs. Because these numbers are small, the HR fluctuates much more than the FP rate. The FP rate is based on >100 sniffs per session for group 1 and >60 sniffs for group 2.

Two samples from a new patient with cancer or from new controls were introduced in test sessions. The results of the dogs on these new odors are illustrated as separate points in Figures 3 and 4: for group 1, the first new patient with cancer was introduced in the seventh session, for group 2, in the third session. For group 1, the first two new patient with cancer led to direct hits, the third new patient with cancer was missed by all dogs, but the fourth was hit again. For group 2, the first new patient with cancer was only hit half the time, the 7th not at all but >80% of sniffs on samples from all other new patient with cancer were hit. Aside from the one patient with cancer missed late by group 2, there is an increasing detection rate on new patient with cancer in time.

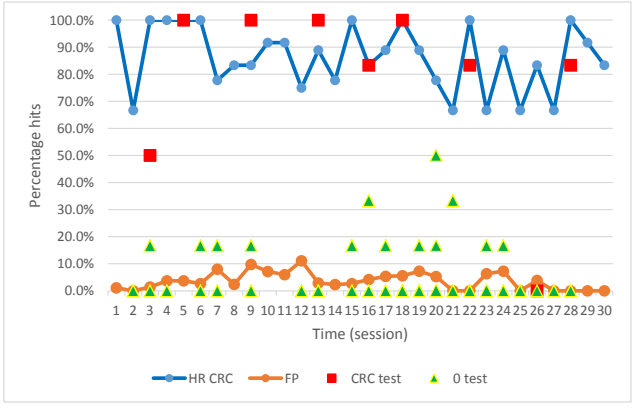

Figure 4. Training and testing results averaged over the 3 dogs in group 2 in time (sessions). HR CRC is hit rate on samples of familiar patients with cancer; FP is false positive rate on samples of familiar controls; CRC test is hit rate on samples of new patients with cancer, 0 test is FP rate on samples of new controls (a number of these points overlap at 0% FP). HR, hit rate; CRC, colorectal cancer; FP, false positive.

The two groups differed in which patient with cancer was used for the initial training and in the sequence of introduction of new patient with cancer. The initial training of the second group was carried out on the patient with cancer that the first group had not hit on at all at first introduction. In Table 4, the average HR of the two groups on new patient with CRC is compared. One patient (CRC-5 in Table 4) caused the most difficulties. At introduction, group 1 hit 50% of the samples, but in later training, dogs in group 1 had problems detecting samples from this patient. For group 2, this patient was introduced very late in the training (7th) and these dogs did not hit any of this person's samples at introduction. No further training was carried out on this patient with group 2. This patient was the youngest of the patients with CRC (54) and had Union for International Cancer Control stage 2 cancer.

The first new control was introduced in the third session for group 1, and in the second session for group 2. In both groups, new controls were not hit on more often initially than the average FP rate on familiar controls but after a number of new odors had been introduced, there was an increased interest in samples from newly introduced controls. For group 1, the first control leading to a higher FP rate was the 7th, for group 2, it was the 18th. This interest led to all samples from some new controls being hit on by all the dogs by group 1 that was introduced to these odors very gradually (average 0.6 introductions/session). In group 2 (average 2 new introductions/session), the interest was less marked. In time, this interest extinguished: the last introductions of controls (last 8 for group 1, last 10 for group 2) did not generate particular interest.

Analyzing the false indications, it appeared that some controls elicited many more FPs than others. In Table 5, the characteristics of "easy" controls (<10% FP rate in training in at least 2 training sessions) is compared with those of "problematic" controls (>30% FP rate in training in at least 2 training sessions). The problematic controls differed significantly in their age from the easy controls for

Table 4
Comparison of HR at introduction for new patients with CRC

	CRC-1	CRC-2	CRC-3	CRC-4	CRC-5	CRC-6	CRC-7	CRC-8
Group 1	100.0%	100.0%	100.0%	80.0%	50.0%	100.0%	100.0%	60.0%
Group 2	100.0%	100.0%	50.0%	100.0%	0.0%	83.3%	83.3%	83.3%

CRC, colorectal cancer; HR, hit rate.

group 1—older ages leading to more problems—but not for group 2 (comparison of means, $P < 0.001$ for group 1, ns for group 2).

Discussion

The broader goal of this pilot study was to see if dogs could be trained to detect colon cancer in stool samples. Because there was limited material, it had to be used strategically to observe the learning process. The chosen strategy will be discussed in the light of two elements: learning to grasp the generalized concept of "colon cancer odor" in stool and responding to it on the one hand and learning to not respond to control odors on the other.

The results show the developing formulation of a "colon cancer odor" concept. Because stool odors will always vary between people, testing stimulus generalization with stool odors from both new patients with colon cancer and controls is necessary to demonstrate this process. Such testing can be performed in different manners. One method consists of training on one positive stimulus and, having reached criterion, introducing the second positive stimulus until criterion, followed by the third positive until the animals respond immediately to a new positive stimulus. Another method was followed by Wright et al., (2017) who trained dogs on 20 different accelerants simultaneously, and then tested them on a group of new ones. Here, a different strategy was used: after introduction of a new target stimulus in a test (the CRC test in Figures 3 and 4), all the target stimuli were presented semi-randomly. We used this strategy and were able to continue reinforcing the dogs on known targets while working on widening their odor concept.

Samples from some new patients with cancer seemed to be more difficult for the dogs than others, but which ones was different for the two groups with the exception of CRC-5, as can be seen in Table 4. This difference may have been the result of the dogs being given their initial training on a different subset. By giving the dogs in group 1 additional trials on "difficult" patients with cancer outside of the normal training sessions (one of the "training issues" mentioned in phase 5 of the training), the dogs were given additional reinforcement on the odor of these patients, but this did not have the desired effect (increased detection rate on these particular patients) so this additional training was not carried out with group 2.

Because there were only a few new patients with CRC available, it is difficult to draw the conclusion that the dogs had successfully formed a generalized CRC odor concept based on this limited set. On average, they did not differ: group 1 detected new patients with CRC, 81.2%, and group 2 responded to 77.8% of the new samples. For group 1, the patient with CRC who was introduced third was not hit at all at introduction, while for group 2, the patient with CRC who was introduced eighth was not hit at all at introduction. These two outliers (different patients for both groups) have a major effect on the trend line in the data: if ignored, the HR in both groups increases over time.

The second element that deserves to be discussed is the development of the false alarm rate: the dogs also had to learn to not respond to controls. Training animals in general includes teaching them this, but not much research highlights this interesting phenomenon. An exception is a study on rats by Lu et al. (1993), where learning odor discrimination in a go-no go setup consisted mainly learning to inhibit a response in the no go condition. In this study, a "punishment condition" was necessary to teach the rats to inhibit their response. Training experience with dogs also suggests dogs need to learn to inhibit their response to distracters, but we chose not to use any form of punishment.

It is interesting to see that both groups of dogs went through a phase where they made relatively more FPs. For group 1, this began in session 8 and continued to session 78 (of the 99 training sessions), and in 2 cases (of 45 introductions) all samples from a new control were hit. Group 2 made relatively more FPs in sessions 16-21 (of the 30 sessions), but none of the introductions led to a higher FP rate than 50%. It is a pattern that has been noted in other projects: a phase of increased FP rate, where dogs learn to discriminate between the positive and negative stimuli through trial and error (Schoon and Fjellanger, pers.com). It seems like dogs tend to use the "new" cue for a while in their learning process, as the general trend in FP rate did not increase at all. In group 1, an effort was made to extinguish the FP rate through additional exposure to these controls, but because that did not have the desired effect, this was not continued for group 2.

The testing regime of the two groups differed: while group 1 was usually only presented with samples of one new control in a test session with test sessions spaced out in time, group 2 received samples from 2-5 new controls in test sessions that followed each other more quickly. As a result, group 2 might have learned to ignore the "new" cue more quickly. This is reflected in their similarity of response to new and familiar controls (Table 3). It also explains group 2's lower response to new controls in comparison to group 1 (Table 3). Another possible explanation for the different response to the new controls may lie in a slight change in the rewarding of correct rejections. The trainers had begun rewarding correct rejections a little quicker and more enthusiastically in group 2 compared with group 1 and reported that the behavior of group 2 was subsequently different on a correct rejection or a miss. They felt group 2 was more "active" in such cases, searching out the trainers after having smelled all the samples and not finding a CRC sample, although this is not reflected in a generally lower FP rate.

Comparing the 8 dogs, there seems to be a trade-off between sensitivity and specificity. Dogs with a higher sensitivity on the new patients with CRC also made more false alarms on the newly introduced controls, although regression analysis was not significant ($r = 0.64$, $P = 0.09$). This pattern has been found before (e.g.

Table 5
Characteristics of controls the dogs did (>30% sniffs FP) and did not (<10% sniffs FP) respond to in training and testing

	N	Average age (±SD)	% Male	FOB scores			
				<25		>1000	NA
Group 1							
"Easy" controls	30	51.8 (±11.4)	33.3%	13	1	0	16
"Problematic" controls	8	70.6 (±5.9)	50.0%	7	0	0	1
Group 2							
"Easy" controls	44	55.0 (±13.9)	36.4%	20	2	0	22
"Problematic" controls	3	67.7 (±7.8)	66.7%	3	0	0	0

FP, false positive; FOB, fecal occult blood; NA, not available.

Schoon et al., 2014) and may link into the personality of the dogs. Based on the limited available material, no conclusions can be drawn on the maximum sensitivity or specificity that can be achieved using trained dogs for colon cancer. We did not have a formal testing period using new patients and controls as advised by Edwards et al. (2017) after training because of a lack of adequate material. However, the data show that with increasing variety in both target and control stimuli, the animals generally improve, as was also shown in examples in explosive detection dogs. In the first example (Oxley and Waggoner, 2008), dogs were trained on gunpowder. Training on one, two, or three brands improved the detection rate on novel, untrained brands, except if the brand was based on a different type of powder. In the second example (described by Goldblatt et al., 2011), dogs trained on American flaked TNT improved their detection of foreign TNT after additional training on one foreign brand. In a third example (Lazarowski and Dorman, 2014), dogs trained on pure potassium perchlorate improved their detection of potassium perchlorate in mixtures after having been trained on a number of mixtures.

Another point Edwards et al. (2017) made is that patient and control populations must be matched. In our study, the age of the patients with CRC was significantly higher than that of the controls (comparison of means, $P < 0.01$). And age was found to be a potential confounding factor: the "problematic" controls were older than the "easy" ones (comparison of means: group 1 $P < 0.01$, group 2 ns), and the most problematic patients with cancer (CRC-5) was the youngest in the group (only 54 years old while the average age was over 70 years). Other problematic factors that could play a role are, for example, smoking habits (because they often lead to cancer development) and use of medication (to suppress symptoms that have developed because of being sick). This information was not available, which is another reason for our reluctance to extrapolate sensitivity and specificity findings beyond this article.

To establish the usefulness of dogs in colon cancer detection, further research is necessary to establish sensitivity and specificity of individual dogs and to establish the robustness of the training methodology used here by having more dogs trained up to the same level. Ideally, we would foresee the use of a team of dogs in establishing a canine "positive" diagnosis. Individual dogs, despite their training, can have an off-day as a result of physical and/or mental variables. The onset of a number of diseases, for example, leads to a declining olfactory sensitivity (Myers et al., 1988). A system using multiple dogs and averaging their results, also used by Apopo in their tuberculosis-detecting rats (Mahoney et al., 2012) and suggested for the detection of corrosion (Schoon et al., 2014), seems a sensible solution. Knowing which animals are highly sensitive, and which are very specific, could also contribute to the evaluation of their results. This article demonstrates the use of monitoring learning curves of dogs in training, allowing increased understanding of the choices individual dogs make and appropriate timing of testing and beginning operational work.

Acknowledgments

The project was funded by the following organizations: KNGF Geleidehonden (Royal Dutch Guide Dog Foundation), ASN Bank, the Zabawas Foundation and Foundation W.M. de Hoop. The authors would like to thank Dr. Y. Smulders, Dr. M. Bomers, Dr. N. de Boer, PhD candidate S. Bosch and Dr. B. Carvalho of the "Vrije Universiteit medisch centrum" for providing the samples and for their cooperation in this project.

Ethical considerations

None.

Conflict of interest

The authors declare no conflict of interest.

References

Concha, A., Mills, D.S., Feugier, A., Zulch, H., Guest, C., Harris, R., Pike, T.W., 2014. Using sniffing behavior to differentiate true negative from false negative responses in trained scent-detection dogs. Chem. Senses 39 (9), 749–754.

Edwards, T.L., Brown, C.M., Schoon, A., 2017. Animal olfactory detection of human diseases: guidelines and systematic review. J. Vet. Behav.: Clin. Appl. Res. 20, 59–73.

Goldblatt, A., Gazit, I., Grinstein, D., Terkel, J., 2011. The Olfactory system and olfaction: implications for REST. In: Remote Explosive Scent Tracing, GICHD report. http://www.gichd.org/fileadmin/GICHD-resources/rec-documents/REST-Nov2011.pdf. Accessed October 27, 2019.

Goldblatt, A., 2017. Some Variables Influencing REST (Remote Explosives Training), 10th International Working Dog Conference, 2-6 April 2017, Banff Canada.

Kranz, W.D., Strange, N.A., Goodpaster, J.V., 2014. "Fooling fido"—chemical and behavioral studies of pseudo-explosive canine training aids. Anal. Bioanal. Chem. 406 (30), 7817–7825.

Lazarowski, L., Dorman, D.C., 2014. Explosives detection by military working dogs: olfactory generalization from components to mixtures. Appl. Anim. Behav. Sci. 151, 84–93.

Lu, X.C.M., Slotnick, B.M., Silberberg, A.M., 1993. Odor matching and odor memory in the rat. Physiol. Behav. 53 (4), 795–804.

Mahoney, A., Weetjens, B.J., Cox, C., Beyene, N., Reither, K., Makingi, G., Jubitana, M., Kazwala, R., Mfinga, G.S., Kahwa, A., Durgin, A., Poling, A., 2012. Pouched rats' detection of tuberculosis in human sputum: comparison to culturing and polymerase chain reaction. Tuberc. Res. Treat. 2012, 716989.

Myers, L.J., Nusbaum, K.E., Swango, L.J., Hanrahan, L.N., Sartin, E., 1988. Dysfunction of sense of smell caused by canine parainfluenza virus infection in dogs. Am. J. Vet. Res. 49 (2), 188–190.

Oldenburg Jr., C., Schoon, A., Heitkönig, I.M., 2016. Wildlife detection dog training: a case study on achieving generalization between target odor variations while retaining specificity. J. Vet. Behav.: Clin. Appl. Res. 13, 34–38.

Oxley, J.C., Waggoner, L.P., 2009. Detection of explosives by dogs. In: Marshall, M., Oxley, J.C. (Eds.), Aspects of Explosives Detection. Elsevier, Amsterdam, pp. 27–38.

Schoon, A., Fjellanger, R., Kjeldsen, M., Goss, K.U., 2014. Using dogs to detect hidden corrosion. Appl. Anim. Behav. Sci. 153, 43–52.

Williams, H., Pembroke, A., 1989. Sniffer dogs in the melanoma clinic? Lancet 333 (8640), 734.

Wright, H.F., Wilkinson, A., Croxton, R.S., Graham, D.K., Harding, R.C., Hodkinson, H.L., Keep, B., Cracknell, N.R., Zulch, H.E., 2017. Animals can assign novel odours to a known category. Sci. Rep. 7, 9019.