Accepted Manuscript

Animal Olfactory Detection of Human Diseases: Guidelines and Systematic Review

Timothy L. Edwards, Clare M. Browne, Adee Schoon, Christophe Cox, Alan Poling

PII: S1558-7878(16)30169-1

DOI: 10.1016/j.jveb.2017.05.002

Reference: JVEB 1059

To appear in: *Journal of Veterinary Behavior*

Received Date: 27 October 2016

Revised Date: 9 February 2017

Accepted Date: 1 May 2017

Please cite this article as: Edwards, T.L., Browne, C.M., Schoon, A., Cox, C., Poling, A., Animal Olfactory Detection of Human Diseases: Guidelines and Systematic Review, *Journal of Veterinary Behavior* (2017), doi: 10.1016/j.jveb.2017.05.002.

Animal Olfactory Detection of Human Diseases: Guidelines and Systematic Review

Timothy L. Edwards[a]*, Clare M. Browne[a,b], Adee Schoon[c,d], Christophe Cox[c], and Alan Poling[c,e]

[a] School of Psychology, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

[b] School of Science, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

[c] Anti-Persoonsmijnen Ontmijnende Product Ontwikkeling, Sokoine University of Agriculture, Tiba Road, PO Box 3078, Morogoro, Tanzania

[d] Animal Detection Consultancy, Vorchten, Gelderland, The Netherlands

[e] Department of Psychology, Western Michigan University, Kalamazoo, Michigan 49008-5439, United States

* Corresponding author

Email: edwards@waikato.ac.nz

Address: School of Psychology, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

Abstract

Animal olfactory detection of human diseases has attracted an increasing amount of interest from researchers in recent years. Because of the inconsistent findings reported in this body of research and the complexity of scent detection research, it is difficult to ascertain the potential value of animal detectors in operational diagnostic algorithms. We have outlined key factors associated with successful training and evaluation of animals for operational disease detection and, using these key factors as points for comparison, conducted a systematic review of the research in this area. Studies that were published in peer-reviewed outlets and that described original research evaluating animals for detection of human diseases were included in the review. The majority of relevant studies have assessed dogs as detectors of various forms of cancer. Other researchers have targeted bacteriuria, *Clostridium difficile*, hypoglycemia, and tuberculosis. Nematodes and pouched rats were the only exceptions to canine detectors. Of the 28 studies meeting inclusion criteria, only 9 employed operationally feasible procedures. The most common threat to operational viability was the use of a fixed number of positive samples in each sample run. Most reports included insufficient information for replication or adequate evaluation of the validity of the findings. Therefore, we have made recommendations regarding the type of information that should be included when describing research in this area. The results of this systematic review suggest that animal detectors hold promise for certain diagnostic applications but that additional research evaluating operationally viable systems for olfactory detection of human diseases is necessary.

Keywords: animal behavior; diagnostic technology; discrimination; olfaction; scent detection

Introduction

It is common knowledge that many animals possess and rely heavily upon a highly developed sense of smell when locating food, avoiding predators, finding mates, and navigating their environments. Humans, who have a relatively poor sense of smell, often employ other animals in the detection of targeted substances by training them to make an identifiable response in the presence of volatile compounds that emanate from those substances. Dogs have been trained to locate explosives, landmines, illicit drugs and other contraband, missing persons, disaster victims, and a wide variety of other targets (Browne et al., 2006; Williams & Johnston, 2002). Bees, pigs, mice, rats, and a number of other animals have also been successfully trained to identify targeted substances (Bodyak & Slotnick, 1999; Poling et al., 2010a; Rains et al., 2008; Talou et al., 1990).

Several anecdotal reports of dogs spontaneously showing interest in skin cancer on their owners have been published. Williams and Pembroke (1989) wrote of a patient whose dog persistently sniffed a mole on her leg. The dog's excessive interest in the mole prompted the patient to visit a dermatologist, who identified the mole as a malignant melanoma. Church and Williams (2001) reported a man whose dog constantly sniffed at a patch of eczema on his leg. After excision of the lesion, it was found to be a basal cell carcinoma. Campbell et al. (2013) described a case in which a man's dog persistently licked a lesion behind his right ear, which was later confirmed to be malignant melanoma. In each of these cases, the dog was apparently able to detect and was attracted to volatile organic compounds (VOCs) emanating from the affected area on its owner's skin.

VOCs are organic chemicals with high vapor pressure at typical room temperature, resulting in evaporation or sublimation of the molecules into the air surrounding the source. VOC profiles reliably associated with asthma, several types of cancer, cholera, cystic fibrosis, diabetes mellitus, dental diseases, gut diseases, heart allograft rejection, heart diseases, liver

diseases, pre-eclampsia, renal disease, and tuberculosis (TB), have been identified (Corradi et al., 2010; Dent et al., 2013; Shirasu & Touhara, 2011). Disease-related VOCs may be found in the blood, breath, feces, skin, sputum, sweat, urine, and vaginal secretions of affected individuals. Research investigating the VOCs associated with various human diseases is underway, primarily driven by the goal of developing instrumentation for use in clinical diagnostics that is capable of reliably identifying specific disease-associated VOC marker profiles. Currently, the development of this technology is limited by the prohibitively high cost of the necessary laboratory instrumentation and difficulties in standardizing sample collection and preparation procedures in clinical settings (Sethi et al., 2013).

An increasing number of experimental analyses examining animal detection of human diseases have appeared in the literature since Pickel et al. (2004) reported the high detection accuracy of two dogs trained to detect melanoma. The cumulative number of relevant studies published between 2004 and 2016 are displayed in Figure 1. The steepening gradient in the data path suggests that interest in this topic has increased over time. This body of literature on which Figure 1 is based has been reviewed from various perspectives (Bijland et al., 2013; Boedeker et al., 2012; Dent et al., 2013; Desikan, 2013; Freeman & Vatz, 2015; Jezierski et al., 2015; Johnen et al., 2013; Lippi & Cervellin, 2012; Luque de Castro & Fernandez-Peralbo, 2012; Marcus, 2012; McCulloch et al., 2012; Moser & McCulloch, 2010; Oh et al., 2015; Wells, 2012). Many reviewers and researchers have remarked that critical components of the training and testing procedures in the relevant studies are often unreported or are deficient. Therefore, it is difficult to ascertain the potential of animals as detectors of various human diseases (e.g., Elliker et al., 2014; Jezierski et al., 2015). The purpose of the present article is, first, to suggest required and preferred conditions for training and testing animals for operational disease detection and, second, to evaluate the existing research with respect to these conditions. Our hope is that the guidelines we propose will be useful for researchers,

animal trainers, and medical practitioners who are interested in olfactory detection of human diseases.

Training Conditions

Operant discrimination training, in which indication responses (e.g., barks by a dog) to samples known to be positive for the disease in question are reinforced (rewarded, as by delivery of a preferred food) and responses to samples not known to be positive for the disease are not reinforced, is used to teach animals to detect the disease. Once an animal reliably emits the indication response only in the presence of known-positive samples, samples of unknown status are presented and the animal's response to those samples is recorded. Samples that engender an identification response are considered to be disease-positive according to the animal detector, although additional confirmatory technology is often used to ensure the patient who provided the sample actually has the disease. Responses to samples of unknown status are not reinforced and, to maintain performance, known-positive samples have to be included in the sample array. As in training, responses to such samples are reinforced. Details of training differ widely across studies, but certain aspects of training are of general, and critical, significance.

Required conditions for training

The conditions outlined in this section are necessary for training an animal to reliably indicate the presence of disease-related VOCs in novel samples.

Confirmed positive samples. Ideally, the status of every sample used in training (i.e., samples that are positive for disease as well as those that are negative for disease) is determined with the gold standard or best available diagnostic technology for the targeted disease. However, knowing the true status of "positive" samples that will be used to arrange reinforcement for correct indications is a required condition. Even occasional reinforcement of a positive indication to a disease-negative sample can lead to persistent false indications

(positive indications of disease-negative samples). Intermittent reinforcement generates patterns of behavior that persist even when reinforcement is no longer forthcoming (Angle et al., 2015; Nevin, 1988). Persistent indication of disease-negative samples negatively impacts specificity (proportion of disease-negative samples that are accurately classified as such), negative predictive value (NPV; the number of correct rejections [negative indications] divided by the total number of rejections), and positive predictive value (PPV; the number of correct positive indications divided by the total number of positive indications).

Intermittent schedules of reinforcement for correct indication responses have the desirable effect of preparing the animal for conditions under which correct indication responses cannot be reinforced. Such conditions are inevitable if the animal will be used operationally because an animal detector would provide no additional value in an operational scenario in which the status of all samples is already known. For this reason, knowing the status of all "negative" samples used for training is not a required condition. If a positive sample is incorrectly classified as "negative" and the animal's correct identification response is not reinforced, the animal will learn to continue evaluating the remaining samples, as it would be required to do in an operational scenario. In early stages of training, while the search and indication behaviors are being shaped, a high ratio of reinforcement (i.e., continuous reinforcement) is necessary, but the schedule of reinforcement should be gradually thinned to match the schedule of reinforcement anticipated under testing and operational conditions. Under training conditions in which consequences are provided for correct or incorrect identification of negative samples (e.g., Gordon et al., 2008; Walczak et al., 2012), it is important that the status of "negative" samples is confirmed with the best available diagnostic technology.

Control (negative) samples that are comparable to positive samples except for disease status. If positive samples and control samples have systematic differences other than their

status with respect to the targeted disease, the animal may learn to rely on these additional cues partially or entirely and its true ability to detect disease-positive samples will be obscured. For example, because a high percentage of individuals with lung cancer also smoke, an animal may dismiss all samples from non-smokers as negative, despite the presence of lung cancer in some non-smokers. Matching controls in terms of age, gender, and other factors that are likely to influence outcome measures is standard practice in medical research, but consideration of additional factors is necessary when working with animals.

Control samples collected in a different time period, from a different location, or by a different person from the one who collected the positive samples are likely to emit VOCs irrelevant to the targeted disease (e.g., volatiles associated with a specific clinic) that may function as cues during training. Attempts to prevent cross-contamination between positive and control samples can also lead to systematic differences between the samples. For example, control samples regularly may be prepared first, by a different person, or in a different room. Any systematic discrepancy between sample collection, handling, and preparation procedures for positive and control samples can result in cues that control the indication response instead of the relevant disease-associated volatiles. Under these conditions, the animal may perform well under training and testing conditions but poorly under independent testing or operational conditions. Because many disease detection research projects never reach the stage of independent testing or operational trials, researchers may inadvertently obtain and report spurious results. Therefore, it is essential to ensure that there are no systematic differences, save for disease status, between positive and negative samples used for training and testing.

A large number of sample sources (patients and healthy controls). Training an animal to detect disease-positive samples is, in essence, teaching the animal a concept. If only a few exemplars are presented repeatedly during training, the animal will have no trouble

identifying the "positive" samples after a short time, but not necessarily because of what they have in common. Pigeons have been trained to indicate the presence of trees, water, people, and even a specific person in novel pictures, even though there is no single defining feature of any of these concepts (Herrnstein & Loveland, 1964; Herrnstein, Loveland, & Cable, 1976). In these experiments, key pecking was reinforced when a picture containing the target was present and not reinforced when the target was absent. In testing sessions, only novel pictures were presented, and the rate of key pecking in the presence of positive (target-present) and negative (target-absent) pictures was recorded. Herrnstein, Loveland, and Cable used approximately 1000 pictures (half negative and half positive) in training and 800 additional pictures for testing acquisition of each concept. Although these studies did not establish the minimum number of training exemplars required for concept formation, if the researchers had used only a few pictures for training, it is unlikely that the pigeons would have been able to respond accurately to novel pictures during testing.

When training for disease detection, using multiple tissue or fluid samples from the same human is not a solution for training a disease concept for the same reasons that multiple copies of the same photo are not useful when training pigeons the "tree" concept. In one report, dogs trained extensively on one type of 2,4,6-trinitrotoluene (TNT) showed high sensitivity (i.e., hit rate) with the type of TNT used in training but failed to indicate other types of TNT (Goldblatt et al., 2011). Increasing the variation in training samples led to improved generalization. Similar issues were reported in a study on the detection of gunpowder (Oxley & Waggoner, 2009). VOC profiles associated with diseases are much more complex than those associated with TNT and gunpowder. Therefore, researchers must take additional precautions to ensure that training conditions are conducive to concept formation.

A wide variety of positive samples with a single commonality – positive disease status – is required; likewise, a wide variety of negative samples with a single commonality – negative disease status – is required. The minimum size of the sample set will depend on a number of factors including the amount of disease-related VOCs that are available in the headspace of a typical sample and the relative availability of unrelated VOCs. An effective approach to training the targeted disease concept when the number of sample sources is restricted is as follows: (1) Train with a subset of the available sample sources until the animal reaches pre-established performance criteria or until performance is no longer improving. (2) Introduce samples from novel positive and control sources and evaluate performance with these samples prior to any reinforced indications of the positive samples (e.g., on the first presentation of the sample). (3) Include these samples in the training set if performance with the novel samples is low, and repeat these steps until performance with samples from novel sources is reliably high. With this method, progress toward concept formation is regularly evaluated, additional exemplars are systematically added to the training set, and the required number of novel sample sources is minimized.

Preferred conditions for training

The conditions outlined in this section are not necessarily required for successful training of animal detectors of human disease but will typically result in more rapid training or higher performance in subsequent testing and operational use.

Trainer blind to status of all samples. Whenever the trainer has knowledge of the disease-status of the samples in the sample set, the likelihood of cuing is very high, even for well-intentioned trainers. If the reader is not already convinced of the difficulty of abstaining from cuing under non-blind conditions from the story of “Clever Hans,” the research of Lit et al. (2011) may provide the required evidence. The researchers conducted tests with drug and explosives detection dog/handler teams in which no targeted substances were present at any

time, but certain locations and decoys were marked with a paper marker and others were not. The marked locations were statistically significantly more likely to be indicated by the dog/handler team than the unmarked locations, confirming that handler beliefs about the location of targeted substances had a powerful influence on the perceived or actual performance of the dogs. Dogs are very receptive to human communicative cues such as hand signals, body orientation, and the emotional content of verbal commands (Ruffman & Morris-Trainor, 2011; Soproni et al., 2002; Virányi et al., 2004). Due to the insidious effects of cuing on true scent detection performance, this "preferred condition" is a strong candidate for the list of "required conditions." There are, however, a number of training methods that can reduce or eliminate the possibility of cuing.

One such method involves placing the trainer behind a screen positioned in such a way that the trainer can see which samples are being identified as positive but the animal cannot see the trainer. When this strategy is used, the trainer must ensure that no auditory cues are accidentally provided (for example, by repositioning a clicker when the animal is approaching a positive sample). Another method involves positioning an observer with sample status information out of view of the animal and the trainer, who is blind to sample status. Each time the animal identifies a sample as positive, the trainer calls out the sample position, and the observer immediately replies with the sample status so that the trainer can take appropriate action (Elliker et al., 2014; Mahoney et al., 2012). With this type of training, the observer must be careful not to provide any auditory cues. Yet another method is to construct an apparatus that is capable of reliably detecting a positive identification and delivering a reinforcer for identification of samples programmed as positive/reinforcement samples (e.g., Mahoney et al., 2014). When "play" is used as the reinforcer, the automated apparatus can emit a visual or auditory cue, which signals the handler to initiate play, but we are unaware of any studies employing this particular methodology.

Confirmed negative samples. Under typical training conditions, no consequences are scheduled for correct or incorrect identification of negative samples. If, however, "negative" samples are known to be negative, consequences can be provided for correct or incorrect identification of negative samples. For example, Gordon et al. (2008) and Walczak et al. (2012) delivered a mild rebuke when detection dogs incorrectly identified negative samples as positive, which presumably served to punish false indications. Zimmerman and Ferster (1963) found that arranging punishment in the form of time-outs of 10 s – 1 min from the training procedure when an incorrect response occurred in a matching-to-sample task resulted in higher accuracy, which suggests that consequences for incorrect responding can increase accuracy in a discrimination task. Unfortunately, relevant research in the scent-detection field is lacking.

When it comes to evaluating the animal detector's performance, working with confirmed negative samples allows the trainer to determine the precise sensitivity (proportion of true positives identified as positive) and specificity (proportion of true negatives identified as negative) of the detector. When "negative" samples are identified as positive by the animal detector, the apparent specificity will decline, but if some of these samples are actually positive, the true specificity is higher than the apparent specificity. In other words, the animal may appear to be incorrectly identifying negative samples as positive, but some of the identifications are actually correct. Therefore, if the status of all samples is not known, the trainer does not have an accurate picture of the animal's true detection performance and does not have complete data on which to base decisions.

Training samples same as operational samples. If a disease-related VOC has been identified and can be isolated for training and testing purposes, one of the largest challenges in olfactory detection of human diseases – that of sample availability – may be overcome. But training with an isolated compound will not necessarily result in success when testing or

operating with human tissues or fluids. Lazarowski and Dorman (2014) found that training with potassium chlorate was not sufficient to produce generalization to potassium-chlorate-based explosive mixtures for most dogs in their study. Bomers et al. (2012) conducted training with toxigenic *Clostridium difficile* strains on culture plates prior to training with stool samples, but the influence of the initial training with the isolated VOCs on subsequent training with samples obtained directly from patients was not evaluated. Suckling and Sagar (2011) trained honeybees to indicate the presence of chemical compounds that were previously found to be associated with TB, but the bees were not subsequently tested with samples obtained from humans. When training animals for clinical use, it is generally advisable to train under the conditions in which the animal detector will be tested, which should in turn be the same conditions under which the animal will operate.

Positive sample prevalence similar to prevalence of disease in operational setting. Training of animals for any purpose usually proceeds stepwise with simple, fundamental responses trained first, followed by the establishment of more complex responses in more difficult training conditions. Animals trained to detect human diseases should eventually be trained to identify the positive samples in a sample set where the prevalence of positive samples is similar to the prevalence of the disease in the targeted operational scenario. For example, dogs under training for eventual screening of skin samples for skin cancer from a dermatological clinic where six percent of the samples are tested positive for skin cancer should eventually be trained under conditions where 94% of the skin samples are negative. Wolfe et al. (2005) found that the percentage of missed targets increased substantially when the prevalence of targets was reduced in a visual search task with humans (50% prevalence produced 7% miss errors while 1% prevalence produced 30% miss errors). Target prevalence is a critical factor that is likely to impact the probability of hits, false indications, and other key outcome measures in scent detection tasks. Therefore, the terminal prevalence of positive

samples in training should match that of testing, which should match the prevalence anticipated in operations as closely as possible.

Intermittent reinforcement. If identification responses are reinforced every time a positive sample is correctly identified, the response will deteriorate quickly under testing and operational conditions in which the status of some proportion of samples is unknown and, therefore, some correct identifications cannot be reinforced. Behavior that is reinforced intermittently is more persistent when it is no longer reinforced (i.e., it is more resistant to extinction) than behavior that is reinforced each time it occurs (Nevin, 1988; 2012). But, in order to avoid degrading an animal's performance, intermittent reinforcement must be introduced gradually. If the trainer goes from reinforcing every correct response to reinforcing every 20^{th} correct response, the animal will probably not make it to the 20^{th} response before shifting to other behavior that has a higher likelihood of being reinforced. As with positive sample prevalence, the rate of reinforcement should be adjusted gradually to match the rate that is expected to be encountered in testing and operational scenarios. For example, if approximately 25% of the positive samples in an operational scenario will have known status, then the animal should be gradually shifted to a schedule of reinforcement under which only 25% of correct identifications, on average, are reinforced.

Testing Conditions

Required conditions for testing

Testing is usually conducted to determine the sensitivity and specificity of the animal as a detector of the targeted disease after a period of training. Other variables may also be investigated when an animal detector is tested, such as evaluation speed, stamina, and resistance to extinction. For a test to provide a convincing demonstration of an animal's ability to detect human diseases, the following requirements must be met:

Sample sources differ from sources used in training. Samples collected for testing purposes must not come from the same individuals as the samples used for training purposes. If the animal detector is given the opportunity to respond to samples that were taken from the same individuals who provided training samples, it cannot be determined with any degree of certainty that the animal's identification responses are being controlled by disease-related VOCs rather than other volatiles unique to the individual who provided the sample. Using samples that have already been used in training would pose an even larger threat to the validity of the test, as it would be easier still for the animal to identify samples that were correctly identified during training without regard for their disease-status.

Trainer blind to status of all samples and presence of positive samples. Results from testing conducted under non-blind conditions cannot be considered as valid. Not only must the trainer be unaware of the status of individual samples, he or she must also be unaware of the number of positive samples that are present in any set of samples under evaluation (e.g., in each run, with "run" meaning a set of samples simultaneously presented to an animal for evaluation). For example, if an animal is tested with sequential presentations of eight-sample runs comprising one positive and seven control samples, the animal and trainer are afforded significantly more information than would be available under operational conditions in which any number of samples could be positive in any sample set. This problem is further compounded with "forced choice" procedures in which the animal must choose one and only one sample from each run. With forced choice procedures, the trainer often makes a judgment call regarding which sample was selected, and the animal is likely to be influenced by some form of cuing (e.g., the trainer encourages the dog to indicate one of the samples).

Additionally, standard measures of diagnostic accuracy obtained from such procedures are not valid, particularly specificity and negative predictive value (NPV). For example, if a dog evaluated seven samples at a time under a forced choice testing procedure

(in each set of seven samples, one is positive, and the dog must indicate only one sample each run), even if it misses the positive sample each time, it will only be able to indicate one of the six negative samples in each run, therefore establishing a floor of one in six samples incorrectly identified as positive, which translates to a minimum specificity and NPV of 83%. If the dog had instead correctly identified the positive sample in each run, specificity and NPV would be 100% by default. These values are not representative of the animal's performance and do not provide any indication of how it might perform in circumstances where an unknown number of positive and negative samples are present. For accurate calculation of diagnostic accuracy, it is necessary to have a variable number of targets in runs and a protocol that ensures that the animal can make a positive indication to any sample.

Accurate knowledge of sample status. In order to accurately determine the sensitivity and specificity of the animal detector, the disease-status of the samples must be known with a high degree of certainty. This typically involves determining the status of the samples using the "gold standard" diagnostic test, which is the best available test for the disease. Particularly when conducting proof-of-concept testing, for the test results to be convincing to interested parties, the reference technology should be the best available under the circumstances. Follow-up testing of individuals who were found to be disease-negative according to gold standard technology but positive according to animal detectors can provide information regarding the ability of animals to identify diseases in early stages.

Control (negative) samples that are comparable to positive samples except for disease status. For the same reasons provided under Training Conditions, positive and negative samples must not differ in any systematic way. If such differences exist, any test of the animal's disease detection performance will be invalid because the animal may be responding to irrelevant features of the samples.

Sufficiently large number of sample sources. The number of sample sources used in the test should be based on an appropriate sample-size calculation conducted prior to the test. A variety of tools have been developed for this purpose, including formulas (Jones et al., 2003), nomograms (Carley et al., 2005), tables (Flahault et al., 2005), and software designed for conducting power analyses for diagnostic accuracy tests. An inadequate number of samples will result in extremely wide confidence intervals around the obtained accuracy measures, rendering the results of the test meaningless (Hajian-Tilaki, 2011).

Two examples calculated using the R (R Core Team, 2015) power.diagnostic.test function in the MKmisc package (version 0.991) illustrate the range of sample sizes that are necessary to establish estimates of sensitivity. In the first example, the expected sensitivity of the detector is 0.9, and the researcher wishes to establish 95% confidence intervals for sensitivity with a lower limit of no less than 0.75. Samples from 74 disease-positive individuals would be required. In this example, if the animal were to be tested with a disease prevalence of 0.1, samples from 666 disease-negative individuals would also be required. In another example, the expected sensitivity of the detector is 0.8, and the researcher wishes to establish 95% confidence intervals for sensitivity with a lower limit of no less than 0.7. In this case, samples from 215 disease-positive individuals would be required. If testing with a disease prevalence of 0.05, samples from 4,085 disease-negative individuals would also be required. Both of these examples are representative of typical conditions and highlight the importance of conducting appropriate tests when determining the number of samples that must be included in evaluations of animal detectors of human diseases.

Preferred conditions for testing

The conditions described in this section may not be necessary for a test to be considered valid or for a test to produce useful results but are generally desirable for the reasons outlined below.

Mechanical or other objective determination of identification response. With explosives detection and many other scent detection tasks, the animal/trainer pair can be considered a single unit, the animal reacting to the targeted volatiles, and the trainer reacting to the animal's reaction. Nonetheless, the requirement that a trainer identify a response as indicative of the presence of the targeted substance can lead to subjective interpretation of the animal's behavior. While some level of subjectivity can be tolerated in the highly variable conditions encountered in explosives detection and other related work (especially when it involves erring on the side of caution), in a controlled laboratory setting where disease detection work will take place, removing human subjectivity from the equation is advisable for several reasons: the apparent performance of the animal may fluctuate if the trainer is distracted or if one trainer is substituted for another; independent observers may not be able to obtain consistent data from the same test, and; the trainer must make a judgment call on "borderline" identification responses. As with training, an automated testing apparatus may be the best solution, but when construction of such a device is not feasible, an objective definition of an identification response that enables independent observers to obtain high levels of agreement should be developed and used during testing.

Positive sample prevalence similar to prevalence of disease in operational setting. When detection animals are being evaluated for a specific operational application, the prevalence of positive samples in the test should accurately reflect the prevalence in the targeted operational population. If, for example, the animal is tested with a positive sample prevalence of 10%, but the prevalence of the disease in the targeted population is 1%, the information gained from the test will not be sufficient to estimate the performance of the animal detector in the operational setting. Detection performance can and should be expected to vary depending upon the prevalence of positive samples in the sample set (Evans et al., 2013; Wolfe et al., 2005).

Other testing conditions similar to operational conditions. Other testing conditions such as sample quantity and quality (including method of collection, age, and storage conditions), session duration, testing environment, and schedule of reinforcement should be similar to those anticipated in operational conditions. Reinforcement samples may be interspersed among the testing samples if such an arrangement would be operationally feasible, but the prevalence and quality of these samples should be similar to samples expected to be available under operational conditions. Any discrepancy between training and anticipated operational conditions, such as the use of dried rather than fresh blood, may cast doubt on the suitability of the test for evaluating the animal detector's potential for operational use.

Independent evaluation. It is preferable for animal detectors in the testing phase to be independently evaluated by a third party to verify the accuracy of the animal detectors and confirm that preliminary results were not spuriously obtained. Well-trained and well-intentioned researchers can unintentionally build cues into the training process and obtain results that cannot be replicated by others. Independent evaluation of detection animals can take a variety of forms, the simplest of which involves testing of samples that have been provided by a third party without sample classification. The results obtained from the animal detector are submitted to the third party who then reveals the sample classification and the performance of the detector. Other arrangements, such as independent replication of findings across multiple laboratories, can build confidence in the findings and strengthen the case for operational deployment.

Operational Conditions

Required conditions for operations

The conditions that are required for successful operations with animal detectors depend heavily upon the type of disease detection work and the setting in which the work will

take place. Despite this, several overarching requirements can be identified and will be described briefly below.

Ongoing training. Animals are constantly learning and, unless appropriate training is arranged frequently, detection performance will inevitably decline. Ongoing training can only be foregone when identification of positive samples is intrinsically reinforcing, as it appears to be for some dogs that spontaneously show interest in their owners' skin cancer or with nematodes employed as described by Hirotsu et al. (2015), for example. Three possibilities for conducting ongoing training will be described here, although other methods may exist.

The first method involves interspersing known samples (training samples) among samples with unknown status (evaluation samples). Correct identifications of positive training samples are reinforced and correct identifications of positive evaluation samples are not reinforced, essentially maintaining the identification of positive samples under an intermittent schedule of reinforcement. The second method involves alternating between training sample runs and evaluation sample runs. During the runs with training samples, correct positive indications are reinforced, whereas no indications are reinforced during runs with evaluation samples. The third method involves training with samples with known status on some days or sessions and evaluating samples with unknown status on others. Although they have not been systematically compared, the three methods are likely to produce different performance. Therefore, as specified under preferred conditions for testing, the testing conditions should match the operational arrangement.

Regardless of the method of known-sample interspersal, the operational settings should involve a high rate of reinforcement whenever feasible. Sargisson and McLean (2010) found that higher rates of reinforcement (35-75%) produced higher sensitivity in remote explosives tracing with dogs than lower rates of reinforcement (20-30%) without impacting specificity.

Training conditions cannot be discriminable from operational conditions. If an animal can discriminate between training and evaluation conditions or between training and evaluation samples, performance with the evaluation samples will suffer because correct identification of training samples is differentially reinforced (i.e., indication of positive training samples is reliably reinforced while indication of positive evaluation samples is never reinforced). If, for example, the training samples are stored in the freezer and the evaluation samples are not, the animal might be able to discriminate between samples that were frozen, which are reinforced if correctly identified, and those that were not frozen, and to which responses are never reinforced. Therefore, the animal would not respond to samples that were not stored in the freezer. Likewise, if training sessions are conducted in one setting and operational evaluation is conducted in another, the animal's performance in the operational setting will decline.

Regular evaluation of performance with another diagnostic tool. Regular comparison of animal performance with another diagnostic tool enables the trainer to identify changes in performance and take corrective action if necessary. Performance during training or with training samples is not sufficient for performance evaluation because of the potential for discrimination between training and evaluation samples as described earlier. An animal's performance in training can be very good while its performance during operational evaluation is very poor, in which case the trainer must identify and eliminate the discriminable difference between training and operational conditions. As this type of evaluation is likely to impact the cost-effectiveness of animal detection solutions, the frequency of these comparisons can be minimized by adopting and enforcing rigorous standard operating procedures, particularly those associated with sample collection and preparation.

Standard operating procedures. Reliably accurate performance is key to the utility and acceptability of animals as detectors of human disease (Jezierski et al., 2015). The

solution to behavioral variability is standardization of operating procedures. All aspects of ongoing training, housing, feeding, session timing, sample collection, sample preparation, and all other laboratory practice must be clearly defined. After all procedures have been described in detail, arrangements for quality assurance must be put in place to ensure that procedures are being followed as described. Although standard operating procedures cannot eliminate all behavioral variability, well-written procedures with strict adherence can eliminate much of the variability that is within the operator's control. Standard operating procedures should also be developed and followed for training and testing.

Preferred conditions for operations

Consistent sample quality and quantity. A stable flow of both training and evaluation samples makes regular training and operations possible. If training or operations are frequently interrupted, particularly for long periods, retraining and retesting may be necessary before operations can resume. Animal detectors may be able to attain a high level of performance when the quality of samples in an operational scenario varies widely, but if there is a major shift in sample quality, performance is likely to be adversely affected. If, for example, operations are shifted from a clinical setting where samples come from symptomatic individuals to a non-clinical setting, where the samples come from a variety of symptomatic and non-symptomatic individuals, performance should be expected to change because of the shift in sample characteristics and disease prevalence. When shifting to a new operational setting, training and testing should be conducted to ensure that the animal detectors are capable of performing as required in the new setting.

Periodic evaluation of performance relative to gold standard. Frequent evaluation of the animal detector's performance relative to the gold standard can be prohibitively time-consuming and expensive depending upon the current gold standard for the relevant disease, but periodic comparison to the gold standard is highly advisable. Comparison to the gold

standard diagnostic technology can provide the most accurate estimates of the animal detector's sensitivity and specificity (if the sample-size is sufficiently large).

These required and preferred conditions for each stage of the pathway from initial training to operational deployment are undoubtedly incomplete, and there are bound to be exceptions that we have not mentioned. Nonetheless, we hope and expect that consideration of these conditions will be of benefit to people who are interested or involved in olfactory detection of human diseases. With these required and preferred conditions serving as points for comparison, a review of relevant experimental research was conducted.

Method

The PubMed and Web of Science™ databases were searched on 25 August 2016 using combinations of the terms: "animal," "cancer," "canine," "chemota*", "dog," "detect*," "disease," "odour", "odor," "olfact*," "scent," and "smell." Articles that described original research involving animal olfactory detection of human disease using samples collected from human participants were selected for inclusion. In evaluating the performance of the animal detectors, gold standard technology must have been used to confirm the status of positive samples used for testing. Articles describing animal olfactory detection of psychiatric conditions were excluded from the present review because no gold standard diagnostic technology exists for such conditions. Articles describing the use of chemotactic assays with animals for the detection of human diseases were included. Studies examining detection of samples obtained during hypoglycemic episodes were included, although hypoglycemia is not classified as a disease. Additional articles meeting the inclusion criteria that were discovered in a subsequent forward and backward ancestral search of the selected articles were also included. Twenty-eight articles meeting inclusion criteria were discovered.

Three tables were prepared for organization and analysis of the findings. In Table 1, basic information about the studies was compiled. In Table 2, key features of training as outlined in the required and preferred conditions for training provided in the Introduction were organized. Categories in Table 3 were chosen according to the required conditions for testing provided in the Introduction. A final column in Table 3 was used to classify the testing conditions in each study as feasible for operations and, therefore, valid evaluations of the capabilities of disease detection animals, or not, based on the other items in the table. One coder compiled the information reported in Tables 1 through 3. A second coder independently evaluated 7 of the 28 studies (25%) and checked the information obtained from the remaining 21 studies for accuracy. Using the independently coded studies, a measure of inter-coder agreement was obtained by dividing the number of fields in the tables in which agreement with the original coder was obtained by the total number of fields. This resulted in an agreement percentage of 73%. Coders discussed inconsistencies in data extracted from each of the studies and updated the tables after reaching consensus.

Results

A brief summary of key information from each of the studies is provided in Table 1, including the disease targeted for detection, the type of sample, the animal detector, and the sensitivity and specificity obtained in the study. Cancer detection has clearly received the most attention, with 20 of the 27 studies targeting one or more cancers. Of the remaining seven studies, four have targeted TB, two hypoglycemia, one bacteriuria, and one *Clostridium difficile*. Urine samples have been used in 11 studies, breath samples in 6, sputum in 4, tissue in 3, stool in 2, blood in 2, and live patients in 1. The only exceptions to canine detectors in the reviewed studies were pouched rats in four studies and nematodes in another. The obtained sensitivity and specificity varied widely, ranging from perfect to

chance performance, with considerable variation even among studies examining the same disease, sample, and detector. A fixed number of positive samples were present during test runs in 19 studies, 14 of which used forced choice procedures.

Table 1. Study summary

Year	1st Author	Disease	Sample	Detector	Mean Sensitivity	Mean Specificity
2004	Pickel	Cancer: melanoma	Tissue	Dog (*Canis familiaris*) (2)	Samples in run: 100% Samples planted on healthy volunteers: 100% Actual patients: 80%	FC[a]
2004	Willis	Cancer: bladder	Urine	Dog (6)	Wet urine: 50% Dry urine: 22% Overall: 41%	FC[a]
2006	McCulloch	Cancer: lung, breast	Breath	Dog (5)	Lung: 99% Breast: 88%	FC[a] (lung: 99%; breast: 98% reported)
2008	Gordon	Cancer: breast, prostate	Urine	Dog (10; bc: 6, pc: 4)	Breast: 22% Prostate: 18%	FC[a]
2008	Horvath	Cancer: ovarian	Tissue	Dog (1)	100%	98% (fixed number of positives in each run)
2009	Weetjens	Tuberculosis	Sputum	Pouched rat (*Cricetomys ansorgei*) (20 trained, 2 tested[b])	73%	93%
2010	Horvath	Cancer: ovarian	Tissue, blood plasma	Dog (2)	Tissue: 100% Blood: 100%	FC[a] (tissue: 95%; blood: 98% reported)
2010	Willis	Cancer: bladder	Urine	Dog (4)	64%	3 control groups - healthy: 89%, disease: 83%, urological disease: 61% (fixed number of positives in each run)
2011	Cornu	Cancer: prostate	Urine	Dog (1)	91%	FC[a] (91% reported)
2011	Mgode	Tuberculosis	Sputum	Pouched rat (10)	Mean: UC 2/10 cutoff: 80%	Mean: UC 2/10 cutoff: 72%
2011	Sonoda	Cancer: colorectal	Breath, stool	Dog (1)	Breath: 91% Stool: 97%	FC[a] (breath: 99%; stool: 99% reported)
2012	Bomers	*Clostridium difficile*	Stool, patient	Dog (1)	Stool: 100% Patient: 83%	Stool:100% Patient: 98% (fixed number of positives in each run)
2012	Buszewski	Cancer: lung	Breath	Dog (UC)	82%	FC[a] (82% reported)
2012	Ehmann	Cancer: lung	Breath	Dog (4)	71%	FC[a] (93% reported)
2012	Mahoney	Tuberculosis	Sputum	Pouched rat (10)	Mean: 68%	Mean: 87%

					2/10 cutoff[c]: 81% 3/10 cutoff[c]: 76%	2/10 cutoff[c]: 76% 3/10 cutoff[c]: 82%
2012	Walczak	Cancer: breast, lung, melanoma	Breath	Dog (6 trained, 3 tested)	37% (all cancer types tested together)	FC[a]
2013	Dehlinger	Hypoglycemia	Skin swab	Dog (3)	Mean: 56%	Mean 53%
2013	Horvath	Cancer: ovarian	Blood plasma	Dog (2)	Series 1 (during chemotherapy): 97% Series 2 (3 month follow up): 70% Series 2 (6 month follow up): 80%	Series 1: 99%; series 2 (3 month follow up): 95%; series 3 (6 month follow up): 92% reported (fixed number of positives in each run)
2014	Amundsen	Cancer: lung	Breath, urine (training with tissue)	Dog (4)	Breath test 1: 65% Breath test 2: 56% Urine test 1: 74% Urine test 2: 64%	Breath test 1: 8% Breath test 2: 33% Urine test 1: 25% Urine test 2: 29%
2014	Elliker	Cancer: prostate	Urine	Dog (10 trained, 2 tested)	19%	FC[a] (73% reported)
2014	Rudnicka	Cancer: lung	Breath	Dog (2)	86%	FC[a] (72% reported)
2015	Hardin	Hypoglycemia	Perspiration & breath (combined)	Dog (6)	78%	FC[a] (96% reported)
2015	Hirotsu	Cancer: oesophageal, gastric, colorectal, breast, pancreatic, bile duct, prostate	Urine (preliminary testing with cell cultures, tissue, blood serum, urine)	Nematode (*Caenorhabditis elegans*)	96%	95%
2015	Reither	Tuberculosis	Sputum	Pouched rat (7)	Mean: 41% 1/7 cutoff: 72% 2/7 cutoff: 57%	Mean: 87% 1/7 cutoff: 59% 2/7 cutoff: 81%
2015	Taverna, Tidu, Grizzi, Torri, et al.	Cancer: prostate	Urine	Dog (2)	99%	98%
2015	Taverna, Tidu, Grizzi, Stork, et al.	Cancer: prostate	Urine	Dog (2)	Pre-operative: 100%	UC
2015	Urbanova	Cancer: prostate	Urine	Dog (1)	94%	FC[a] (92% reported)

2016	Maurer	Bacteriuria	Urine	Dog (5)	*Escherichia coli*: 99.6% *Enterococcus*: 100% *Klebsiella*: 100% *Staphylococcus aureus*: 100%	*Escherichia coli*: 91.5% *Enterococcus*: 93.9% *Klebsiella*: 95.1% *Staphylococcus aureus*: 96.3% (fixed number of positives in each run)

Note. FC = forced choice; UC = unclear.

[a] With forced choice procedures, specificity calculations are invalid.

[b] Comparison to culture (gold standard) in Experiment 1 only.

[c] Group criterion in which samples indicated by 2 of 10 or 3 of 10 rats are considered as indicated (see Mahoney et al. (2012) for additional cutoff values).

Table 2 summarizes key aspects of the training conditions in each of the studies. In six studies and in one condition of a seventh study, all training samples were evaluated with gold standard technology. In 12 studies and in one condition in a thirteenth study, at least some of the control samples were obtained from healthy but untested individuals. In three studies, all samples were evaluated with microscopy, which is highly specific but not sensitive for diagnosis of TB. For four studies, no information regarding testing of training samples was provided. Training procedures were not reported by Hirotsu et al. (2015) because training was not required for their approach to cancer detection involving nematodes. Therefore, this study is excluded from the following description of training conditions.

Table 2. Training conditions

Year	1st Author	Known sample status (positive and negative tested with gold standard technology)	Controls comparable to positive samples	Number of sample sources		Handler blind to sample status	Training comparable to testing	Prevalence of positives in sample set	
				Pos.	Cont.			Run size (no. of positives)	% pos.
2004	Pickel	Y	N (non-tissue controls)	tm: UC tsp: 1	N/A	N	tm: N (retrieval training) tsp: Y	tm: UC (1) tsp: UC (1)	UC
2004	Willis	N (untested controls)	Y	27	54	N	Y	7 (1)	14%
2006	McCulloch	N (untested controls)	N (only healthy controls)	lc: 27 bc: 25	66	Y	Y	5 (1)	20%
2008	Gordon	N (untested controls)	Y	bc: 53 pc: 46	134 120	Y	Y	7 (1)	14%
2008	Horvath	Y	N (pos. and cont. in separate rooms)	31	UC	N	Y	10 (2)	20%
2009	Weetjens[a]	N (microscopy not gold standard)	Y	UC	UC	Y	Y	10 (UC)	5-20%
2010	Horvath	ts: Y bls: N (untested controls)	ts: Y bls: N (younger controls, some males)	UC	UC	N	Y	6 (1)	17%
2010	Willis	N (untested controls)	Y	UC	UC	N	Y	7 (0-1+)	UC
2011	Cornu	Y	Y	26	16	N	N (different run size)	2 (1)	50%
2011	Mgode[b]	N (microscopy not gold standard)	Y	UC	UC	Y	Y	10 (UC)	5-20%
2011	Sonoda	UC	UC	UC	UC	N	Y	5 (1)	20%
2012	Bomers	Y	Y	UC	UC	N	N (different sample types)	UC	UC
2012	Buszewski	N (untested controls)	N (only healthy controls)	UC	UC	UC	Y	5 (1)	20%
2012	Ehmann	N (untested controls)	Y	35	60	UC	Y	UC	UC
2012	Mahoney[a]	N (microscopy not gold standard)	Y	UC	UC	Y	Y	10 (UC)	5-20%
2012	Walczak	N (untested controls)	N (only healthy controls)	bc: 57 m: 45 lc: 118	305	Y	Y	5 (1)	20%

2013	Dehlinger	UC	UC	UC	UC	UC	UC	UC	UC
2013	Horvath[c]	UC	UC	UC	UC	N	N (different run size)	4-10 (0-3)	UC
2014	Amundsen	N (untested controls)	UC	ts: 1 bs: UC us: UC	20	UC	UC	6 (0-6)	50%
2014	Elliker	N (untested controls)	Y	50	67	Y	Y	4 (1)	25%
2014	Rudnicka[d]	N (untested controls)	N (only healthy controls)	bc:57 m: 45 lc: 118	305	Y	Y	5 (1)	20%
2015	Hardin	Y	Y	UC[e]	UC	Y (no handler present)	Y	7 (1)	14%
2015	Horitsu	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2015	Reither[f]	Y (196 known samples prior to test)	Y	UC	UC	Y	Y	10 (UC)	5-20%
2015	Taverna, Tidu, Grizzi, Torri, et al.	N (untested controls)	Y (but 23% of control samples from females)	200	230	Y	Y	6 (0-6)	UC
2015	Taverna, Tidu, Grizzi, Stork, et al.	N (untested controls)	Y (but 23% of control samples from females)	200	230	Y	Y	6 (0-6)	UC
2015	Urbanova	Y	Y	UC	UC	UC	Y	3 (1)	33%
2016	Maurer	Y	Y	UC	UC	N	Y	5 (1)	20%

Note. bc = breast cancer; bs = breath sample; bls = blood sample; lc = lung cancer; m = melanoma; N/A = not applicable; pc = prostate cancer; ss = stool sample; tm = tissue mixture; ts = tissue sample; tsp = tissue sample planted on person; UC = unclear; us = urine sample.

[a] Training information obtained from Poling et al. (2011).

[b] Training information obtained from Weetjens et al. (2009).

[c] Training information obtained from Horvath et al. (2008) and Horvath et al. (2010).

[d] Training information obtained from Walczak et al. (2012).

[e] Positive and negative samples obtained from same individuals.

[f] Training information obtained from Poling et al. (2011), Poling et al. (2010b), and Weetjens et al. (2009).

In 16 studies and in one condition of a seventeenth, the set of control samples appears to have been comparable to the positive samples with respect to sample quality and patient characteristics. But, the possibility remains that unreported differences between positive and control sample sets may have existed in any of these experiments. For five studies, insufficient detail was provided to make the determination. Pickel et al. (2004) used medical supplies, such as bandages and tape, but no tissue samples as controls when training dogs to detect tissue samples with melanoma. Horvath et al. (2008) specified that positive and control samples were prepared in separate rooms, which may have introduced additional cues associated with positive and negative samples. In the remaining 6 studies and in one condition in a seventh study, all of the control samples came from individuals who were asymptomatic or differed from the target population in other ways.

For studies and conditions in which relevant information was provided, the number of sample sources ranged from 1 to 200 (*Median* = 45) for positive sample sources and from 16 to 305 (*Median* = 94) for control sample sources. The authors did not specify the number of individuals from which training samples were collected in at least one condition in 17 of the reviewed studies.

The handler was blind to the status of individual samples (at least during the final stage of training) in 12 of the studies. In 10 studies, the handler was not blind to sample status during training. In five studies this information was unclear or unspecified.

In 21 of the 27 studies, the training conditions corresponded with the testing conditions, as reported. Pickel et al. (2004) trained using search and retrieval trials with tissue samples but conducted testing by presenting the tissue samples among control

samples, planting samples on a healthy volunteer, and allowing the dogs to search patients with melanoma. Bomers et al. (2012) tested, but did not train, with live patients. In two studies, the number of samples in each run differed between training and testing (Cornu et al., 2011; Horvath et al., 2013). Two reports provided insufficient information regarding the similarity of training and testing conditions (Amundsen et al., 2014; Dehlinger et al., 2013).

Sample runs contained a fixed number of samples in 22 of the studies. The size of these runs ranged from 2 to 10 (*Median* = 6). Sample runs in 13 of these 22 studies contained one positive sample, and in another study contained two positive samples (Horvath et al., 2008). In 4 of these 22 studies, each run contained a variable number of positive samples and, in the remaining 4 studies, the number of positive samples in each run was not specified. The authors did not specify the number of samples in training runs for four of the studies. Pickel et al. (2004) reported that a single positive sample was present in each run but did not specify the run size, while the other three studies did not report the run size or the number of positive samples in each run. Horvath et al. (2013) employed training runs of varying sizes (4-10) with a varying number of positive samples (0-3). Authors reported or we were able to calculate the prevalence of positive samples in the training set for 19 of the 27 studies. In these studies, prevalence ranged from 5% to 50% (*Median* = 20%).

Table 3 summarizes critical aspects of the testing conditions reported in each of the studies as described under the required conditions for testing provided in the Introduction. For the 20 studies in which sufficient information was provided to make the determination, individuals who provided samples for training differed from those

individuals who provided samples for testing of the animal detectors, with the exception of the planted tissue samples used in one study (Pickel et al., 2004). In the remaining seven studies (excluding Hirotsu et al., 2015), this information was unspecified or unclear.

Table 3. Testing conditions

Year	1st Author	Sample sources (individuals) different from training	Known sample status (positive and negative tested with gold standard technology)	Controls comparable to positive samples	Number of sample sources		Prevalence of positives in sample set		Handler blind to sample status and presence of positive sample in sample set	Testing conditions feasible for operations
					Pos.	Cont.	Run size (no. of positives)	% pos.		
2004	Pickel	ts: Y tsp: N pa: Y	N/A (non-tissue controls)	ts, tsp: N (non-tissue controls) pa: Y	ts: 1 tsp: 1 pa: 7	N/A (non-tissue controls)	ts: 10 (1) tsp: UC[a] pa: 8-30 (1)	ts: 10% tsp: UC pa: 3-13%	status: Y presence: ts: N tsp: Y pa: Y	ts: N (FC[b]) tsp: N (planting tissue on volunteers) pa: UC (bandages over suspected and non-suspected areas)
2004	Willis	Y	N (untested controls)	Y	9	54	7 (1)	14%	status: Y presence: N	N: FC[b]
2006	McCulloch	Y	N (untested controls)	N (only healthy controls)	lc: 28 bc: 6	17	5 (1)	20%	status: Y presence: N	N: FC[b]
2008	Gordon	Y	N (untested controls)	Y	bc: 9 pc: 11	bc: 54 pc: 66	7 (1)	14%	status: Y presence: N	N: FC[b]
2008	Horvath	UC if from other individuals	Y	N (pos. and cont. in separate rooms)	20	UC	10 (2)	20%	status: Y presence: N	N: fixed number of positives in each run, ts
2009	Weetjens	Y	Y	Y	UC	UC	10 (UC)	8%	status: Y presence: Y	Y (if variable number of positives in runs)
2010	Horvath	UC if from other individuals	ts: Y bls: N (untested controls)	ts: Y bls: N (younger controls, some males)	UC	UC	6 (1)	17%	status: Y presence: N	N: FC[b]
2010	Willis	Y	N (untested controls)	Y	30	180	7 (1)	14%	status: Y presence: N	N: fixed number of positives in testing each run

2011	Cornu	Y	Y	Y	33	33	6 (1)	17%	status: Y presence: N	N: FC[b]
2011	Mgode	Y	Y	Y	56	228	10 (UC)[c]	UC	status: Y presence: Y	Y (if variable number of positives in runs)
2011	Sonoda	Y	Y	Y	bs: 33 ss: 37	bs: 132 ss: 148	5 (1)	20%	status: Y presence: N	N: FC[b], ss collected during colonoscopy
2012	Bomers	UC if from other individuals	ss: Y pa: N (untested controls)	Y	ss: 50 pa: 30	ss: 50 pa: 270	ss: 1 (0-1) pa: 10 (1)	ss: 50% pa: 10%	status: Y presence: ss: Y pa: N	ss: Y pa: N (fixed number of positives in each run)
2012	Buszewski	UC	N (untested controls)	N (only healthy controls)	UC	UC	5 (1)	20%	status: Y presence: N	N: FC[b]
2012	Ehmann	Y	N (untested controls)	Y	25	100	5 (1)	20%	status: Y presence: N	N: FC[b]
2012	Mahoney	Y	Y	Y	81	409	10 (UC)	18%	status: Y presence: Y	Y (if variable number of positives in runs)
2012	Walczak	UC if from other individuals	N (untested controls)	N (only healthy controls)	29	UC	5 (1)	20%	status: Y presence: N	N: FC[b]
2013	Dehlinger	Y	Y	Y	3[d]	3[d]	1 (0-1)	50%	status: Y presence: Y	Y (but high prevalence)
2013	Horvath	Y	N (untested controls)	N (only healthy controls)	Series 1: 42 Series 2 (3 month follow up): 10 Series 2 (6 month follow up): 10	Series 1: 210 Series 2 (3 month follow up): 50 Series 2 (6 month follow up): 50	7 (1 positive; 1 reference)	29%	status: Y presence: N	N: fixed number of positives in each run
2014	Amundsen	Y	Y	Y	UC	UC	6 (0-6)	50%	status: Y presence: Y	Y (but high prevalence)
2014	Elliker	Y	N (untested controls)	Y	Test 1: 15 Test 2 & 3: 16	Test 1: 45 Test 2 & 3: 48	4 (1)	25%	status: Y presence: N	N: FC[b]

2014	Rudnicka	Y	N (untested controls)	Y	108	145	5 (1)	20%	status: Y presence: N	N: FC[b]
2015	Hardin	UC	Y	Y	UC[d]	UC[d]	7 (1)	14%	N/A (handler not present)	N: FC[b]
2015	Hirotsu	N/A	N (untested controls)	Y	24	218	1 (0-1)	N/A[e]	status: Y[f] presence: Y	Y
2015	Reither	Y	Y	Y	109	360	10 (UC)	UC	status: Y presence: Y	Y (if variable number of positives in runs)
2015	Taverna, Tidu, Grizzi, Torri, et al.	Y	N (untested controls)	Y (but 23% of control samples from females)	362	540	6 (0-6)	40%	status: Y presence: Y	Y
2015	Taverna, Tidu, Grizzi, Stork, et al.	Y	UC (no control information)	UC	114	0	UC	UC	UC	UC
2015	Urbanova	UC	Y	Y	45	25	3 (1)	33%	status: Y presence: N	N: FC[b]
2016	Maurer	Y	Y	Y	231 (191 *E coli*; 11 *S. aureus*; 10 *Enterococcus*; 19 *Klebsiella*)	456	5(0-1)[g]	20%	status: Y presence: UC[g]	N: fixed number of positives in each run[g]

Note. bc = breast cancer; bs = breath sample; FC = forced choice; lc = lung cancer; pa = patients, pc = prostate cancer; ss = stool sample; ts = tissue samples; tsp = tissue sample planted on person; UC = unclear.

[a] 10-11 stimuli reported on target-present test trials but number of stimuli present on target-absent trials unclear (only results from target-present trials were reported).

[b] With FC procedures, the detection task involves locating a fixed number of positive samples in a run of fixed size. Such conditions cannot be arranged in an operational scenario.

[c] From Weetjens et al. (2009).

[d] Positive and negative samples collected from the same sources.

[e] Individual worms did not evaluate all samples.

[f] Automated chemotaxis assay (Bargmann et al., 1993; Yoshida et al., 2012).

[g] Testing included some runs (unspecified number) in which only negative samples were available, so runs contained either zero or one positive samples.

Researchers used positive and negative samples tested with gold standard diagnostic technology in 12 studies, and in one condition in two other studies. In 12 studies and one condition in two more studies, positive samples were tested with gold standard technology but some proportion of control samples were not tested. These control samples were typically obtained from healthy volunteers. Pickel et al. (2004) used non-biological controls such as gauze and latex gloves. Taverna et al. (2015b) provided no information indicating that control samples were used.

Based on the information provided, positive and control samples used for testing appear to have been comparable, with the following exceptions. Pickel et al. (2004) used non-tissue controls in two conditions. In four studies, only healthy controls were used, and in another study the controls were significantly younger than individuals providing positive samples. Taverna et al. (2015b) provided no information about control samples.

The number of positive sample sources used for testing ranged from 1 to 362 (*Median* = 29) in the 22 studies reporting this information. The number of negative sample sources ranged from 0 to 540 (*Median* = 116) in the 18 studies with relevant information. Studies involving detection of samples obtained during hypoglycemic episodes were not included in this summary because control and positive sample sources are the same and a large number of sources is not necessary for an adequate test of a detector's performance with this application. Authors did not specify the number of sample sources or this information was unclear in five and seven studies for positive and negative samples, respectively.

Runs consisted of a fixed number of samples in 26 studies and in one condition of another study. In at least one condition in three of these studies, a single sample at a time

was presented to the animal for evaluation (Bomers et al., 2012; Dehlinger et al., 2013; Hirotsu et al., 2015). In the remaining studies and conditions, run size ranged from 3 to 10 (*Median* = 6). In at least one condition of 16 studies, each run included one positive sample. Maurer et al. (Maurer et al., 2016) included one positive sample in each run but also arranged an unspecified number of runs with no positive samples. Horvath et al. (2008) and Horvath et al. (2013) included two positive samples in each run. In at least one condition in nine studies, runs included a variable number of positive samples. Taverna et al. (2015b) included no information on the composition or size of runs. In the 24 studies with sufficient information to calculate positive sample prevalence, prevalence ranged from 8% to 50% (*Median* = 20%). These prevalence values differ from prevalence as calculated from the number of positive and control sample sources, 10% to 64% (Median = 20%). In 10 of the 21 study conditions with sufficient information for comparison, sample prevalence and source prevalence did not match, indicating that control or positive samples from the same source were presented to the detector a disproportionate number of times.

Individuals working with the detector did not know the status of individual samples in all 26 of the studies in which the relevant information was reported. In 15 studies and at least one condition in two additional studies, handlers were aware of the number of positive samples in each run. Maurer et al. (2016) did not specify if handlers knew which runs contained no positive samples.

Based on the criteria outlined under the required conditions for testing provided in the Introduction, we consider the methods used in the 16 studies and two conditions in two other studies that employed forced choice procedures or used a fixed number of

positive samples in each run unsuitable for evaluating the operational viability of disease detection animals. Maurer et al. (2016) included runs with no positive samples during testing, but this arrangement does not allow for multiple positive samples in a single run and is not operationally feasible. Pickel et al. (2004) planted tissue samples on volunteers in one condition, which does not appear to be an operationally feasible method. In another condition, they put bandages on suspected and non-suspected areas, but it is not clear if this is an operationally feasible approach to melanoma detection. Horvath (2008) used tissue samples for detection of ovarian cancer and Sonoda (2011) used stool samples that were obtained during colonoscopy for colorectal cancer detection; these samples are collected through invasive procedures, the same procedures that are required for application of gold standard diagnostic analysis and, therefore, it is unlikely that any advantage would be gained by using an animal detector in such applications. In eight studies and one condition of a ninth, the testing conditions appear to be feasible for relevant operational disease detection and, therefore, provide convincing evidence regarding the suitability of animals for the detection of human diseases.

Discussion

In this review of 28 studies evaluating the use of animal olfaction for detection of human diseases, 9 studies used training and testing protocols that appear to be operationally viable (see shaded rows in Table 1). The primary reason for a lack of operational viability in the remaining studies was the presentation of a fixed number of positive samples in runs, which was the case in 19 of the reviewed studies, 14 of which employed forced choice procedures. In these studies, the animal detectors were not

evaluated under conditions that can be arranged in operational settings, and standard accuracy measures such as specificity are not informative.

Although the majority of studies reviewed herein evaluated dogs as cancer detectors, only two such studies employed procedures that might work in an operational scenario. Amundsen et al. (2014) evaluated dogs as lung cancer detectors using breath and urine samples, and Taverna et al. (2015a) examined dogs' ability to detect prostate cancer in urine samples. Although operationally feasible methods appear to have been employed in both studies, several key aspects of training and testing conditions were unspecified or unclear. Bomers et al. (2012) evaluated a method of *C. difficile* detection in stool samples that appears to be operationally feasible, but did not specify some key details of training and testing, such as whether or not testing sample sources differed from training sample sources. Hirotsu et al. (2015) used a chemotaxis assay to evaluate nematodes' natural propensity to move toward urine samples from individuals with a variety of cancers, finding high sensitivity and specificity. This appears to be an operationally viable approach to cancer detection.

In the four studies evaluating pouched rats as detectors of TB, the research was carried out using methods that appear to be operationally viable, although the distribution of positive samples throughout runs was not specified in any of the studies (Mahoney et al., 2012; Mgode et al., 2012; Reither et al., 2015; Weetjens et al., 2009). The same research group has reported results from operational deployment of pouched rats in diagnostic algorithms for TB detection in Tanzania and Mozambique, suggesting that pouched rats serve a valuable function in this role (Edwards et al., 2016; Mahoney et al., 2011; Poling et al., 2010b).

The accuracy of animal detectors of human diseases, as reported in this body of literature, varies widely. In Tables 2 and 3, we summarized factors that we consider particularly important to successful training and evaluation of disease detection animals. Although a quantitative analysis examining the influence of each of these factors on accuracy measures would be valuable, current studies have too little in common for such an analysis to produce meaningful results. Moreover, relevant information was often unreported or undecipherable.

Many researchers used samples from some untested controls during training and testing. Although we described "accurate knowledge of sample status" as a required condition for testing, in most cases these untested controls were healthy individuals in which the probability of the relevant targeted diseases would be extremely minimal. Moreover, testing these individuals with the relevant gold standard diagnostic technology, for example by biopsy, would not be possible in many cases. This sample source arrangement is only a problem when the set of control sample sources consists exclusively, or almost exclusively, of healthy controls, in which case the detector may simply learn to indicate samples from individuals with health abnormalities. Such conditions were arranged in four of the reviewed studies (Buszewski et al., 2012; Horvath et al., 2013; McCulloch et al., 2006; Walczak et al., 2012).

One of the biggest shortcomings of the present body of literature is that much of the information related to sample characteristics, training methods, and evaluation procedures was not specified clearly or at all, although some reports were very thorough (e.g., Hirotsu et al., 2015; McCulloch et al., 2006). We recommend that authors of future studies in this area of research refer to Tables 1 – 3 as a guideline regarding the

fundamental information that should be included when reporting results. Additionally, when space limitations prevent authors from providing sufficient detail to allow others to replicate their training and evaluation procedures, they should consider making standard operating procedures for training and evaluation available as supplemental information. As others have pointed out, there are no standards for training disease detection animals (e.g., Jezierski et al., 2015; Oh et al., 2015; Walczak et al., 2012), and the practice of sharing procedures may be instrumental in the development of such standards.

One of the largest threats to the validity of the results from each of these studies is the possibility that systematic differences, other than disease status, between positive and control samples in both training and testing phases were responsible for the obtained results. An animal trained under such conditions may appear to be a competent detector of the targeted disease but is in fact a competent detector of an irrelevant sample characteristic. Although the information that was provided suggests that positive and control samples were comparable in at least one condition of 22 of the studies, unreported differences could have been present in any of the studies in this review. Replications of these studies, particularly by different research groups, would add significantly to the validity of the findings.

The small number of positive and control sample sources used in the evaluation phases of most studies resulted in imprecise estimates of detection accuracy. Using the median positive and control sample source numbers from testing phases and assuming that the detector responded correctly to 23 of the 29 positives and 90 of the 116 controls, we have calculated 95% confidence intervals for sensitivity (60%, 92%) and specificity

(69%, 85%), using MedCalc Software (Ostend, Belgium). In this representative example, sensitivity has not been established with great precision.

Some research involving olfactory detection of diseases is not aimed at evaluating an operationally viable product. For example, Horvath et al. (2008) were primarily interested in determining whether dogs in previous cancer-detection studies were responding to VOCs associated directly with cancer or to other associated odors, such as metabolic products. This research question shaped their methodology such that it was not suitable for evaluation of an operationally viable detection technology. Research with animal detectors might also play a key role in the development of electronic, volatile-detection technology; the operational viability of procedures in this area of research may be unimportant. In the present review, however, our aim was to evaluate the potential of animal detectors for operational diagnosis of human diseases.

Olfactory detection of diseases holds the promise of rapid, non-invasive, and cost-effective diagnosis. At first glance, there appears to be substantial evidence supporting the utility of animal detectors in this role. Unfortunately, we have found significant limitations associated with this body of research and, with the exception of one operationally active research group, animals have not been employed for diagnostic purposes in large-scale operational settings. Considering the results obtained using methods that we have classified as operationally viable in the present review, we can only make tentative conclusions about the potential of animals in this role. Under the right circumstances, animals appear to be capable of precisely discriminating between samples obtained from disease-positive and disease-negative individuals. Programs of research designed to elucidate the specific conditions that engender consistently high accuracy in

animal detectors are needed. A major barrier preventing deployment of animal detection technology in the medical field is that many of the procedures used to evaluate animal detectors do not translate directly to operations. Therefore, additional research evaluating the effectiveness of operationally viable procedures for olfactory detection of diseases is also required.

Conflict of Interest

Edwards is a former employee of Anti-Persoonsmijnen Ontmijnende Product Ontwikkeling (APOPO), and Schoon, Cox, and Poling are currently affiliated with APOPO; some publications produced by this organization are reviewed in this article. No other conflicts of interest are declared.

Authorship

The idea for this paper was conceived by Edwards, Schoon, Cox, and Poling. The review information and data were obtained and analyzed by Edwards and Browne. All authors contributed to the writing of this paper and have approved the final article.

References

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 Oncol. 53*, 307-315. doi:10.3109/0284186x.2013.819996

Angle, T. C., Passler, T., Waggoner, P. L., Fischer, T. D., Rogers, B., Galik, P. K., Maxwell, H. S., 2015. Real-time detection of a virus using detection dogs. *Front. Vet. Sci. 2*, 79. doi:10.3389/fvets.2015.00079

Bargmann, C. I., Hartwieg, E., Horvitz, H. R., 1993. Odorant-selective genes and neurons mediate olfaction in C-elegans. *Cell 74*, 515-527. doi:10.1016/0092-8674(93)80053-h

Bijland, L. R., Bomers, M. K., Smulders, Y. M., 2013. Smelling the diagnosis: A review on the use of scent in diagnosing disease. *Neth. J. Med. 71*, 300-307.

Bodyak, N., Slotnick, B., 1999. Performance of mice in an automated olfactometer: Odor detection, discrimination and odor memory. *Chem. Senses 24*, 637-645. doi:10.1093/chemse/24.6.637

Boedeker, E., Friedel, G., Walles, T., 2012. Sniffer dogs as part of a bimodal bionic research approach to develop a lung cancer screening. *Interact. Cardiovasc. Thorac. Surg. 14*, 511-515. doi:10.1093/icvts/ivr070

Bomers, M. K., van Agtmael, M. A., Luik, H., van Veen, M. C., Vandenbroucke-Grauls, C. M. J. E., Smulders, Y. M., 2012. Using a dog's superior olfactory sensitivity to identify *Clostridium difficile* in stools and patients: Proof of principle study. *Br. Med. J. 345*, 8. doi:10.1136/bmj.e7396

Browne, C., Stafford, K., Fordham, R., 2006. The use of scent-detection dogs. *Ir. Vet. J. 59*, 97-104.

Buszewski, B., Ligor, T., Jezierski, T., Wenda-Piesik, A., Walczak, M., Rudnicka, J., 2012. Identification of volatile lung cancer markers by gas chromatography-mass spectrometry: Comparison with discrimination by canines. *Anal. Bioanal. Chem. 404*, 141-146. doi:10.1007/s00216-012-6102-8

Campbell, L. F., Farmery, L., George, S. M. C., Farrant, P. B. J., 2013. Canine olfactory detection of malignant melanoma. *BMJ Case Rep. 2013*. doi:10.1136/bcr-2013-008566

Carley, S., Dosman, S., Jones, S. R., Harrison, M., 2005. Simple nomograms to calculate sample size in diagnostic studies. *Emerg. Med. J. 22*, 180-181. doi:10.1136/emj.2003.011148

Church, J., Williams, H., 2001. Another sniffer dog for the clinic? *Lancet 358*, 930-930. doi:10.1016/s0140-6736(01)06065-2

Cornu, J. N., Cancel-Tassin, G., Ondet, V., Girardet, C., Cussenot, O., 2011. Olfactory detection of prostate cancer by dogs sniffing urine: A step forward in early diagnosis. *Eur. Urol. 59*, 197-201. doi:10.1016/j.eururo.2010.10.006

Corradi, M., Gergelova, P., Mutti, A., 2010. Exhaled volatile organic compounds in nonrespiratory diseases. *European Respiratory Monograph: Vol. 49. Exhaled biomarkers* (pp. 140-151). Plymouth, UK: European Respiratory Society.

Dehlinger, K., Tarnowski, K., House, J. L., Los, E., Hanavan, K., Bustamante, B., Ahmann, A. J., Ward, W. K., 2013. Can trained dogs detect a hypoglycemic scent

in patients with type 1 diabetes? *Diabetes Care 36*, E98-E99. doi:10.2337/dc12-2342

Dent, A. G., Sutedja, T. G., Zimmerman, P. V., 2013. Exhaled breath analysis for lung cancer. *J. Thorac. Dis. 5*, S540-S550. doi:10.3978/j.issn.2072-1439.2013.08.44

Desikan, P., 2013. Rapid diagnosis of infectious diseases: The role of giant African pouched rats, dogs and honeybees. *Indian J. Med. Microbiol. 31*, 114-116. doi:10.4103/0255-0857.115214

Edwards, T. L., Valverde, E., Mulder, C., Cox, C., Poling, A., 2016. Pouched rats as detectors of tuberculosis: comparison to concentrated smear microscopy. *Eur. Resp. J.*, ERJ-00264-2016.

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. *Eur. Resp. J. 39*, 669-676. doi:10.1183/09031936.00051711

Elliker, K. R., Sommerville, B. A., Broom, D. M., Neal, D. E., Armstrong, S., Williams, H. C., 2014. Key considerations for the experimental training and evaluation of cancer odour detection dogs: Lessons learnt from a double-blind, controlled trial of prostate cancer detection. *BMC Urol. 14*, 9. doi:10.1186/1471-2490-14-22

Evans, K. K., Birdwell, R. L., Wolfe, J. M., 2013. If you don't find it often, you often don't find it: Why some cancers are missed in breast cancer screening. *PLoS One 8*, 6. doi:10.1371/journal.pone.0064366

Flahault, A., Cadilhac, M., Thomas, G., 2005. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J. Clin. Epidemiol. 58*, 859-862. doi:10.1016/j.jclinepi.2004.12.009

Freeman, W. D., Vatz, K. A., 2015. The future of health care: Going to the dogs? *Front. Neurol. 6*, 2. doi:10.3389/fneur.2015.00087

Goldblatt, A., Gazit, I., Grinstein, D., Terkel, J., 2011. *The olfactory system and olfaction: Implications for REST*. Retrieved from http://www.gichd.org/fileadmin/GICHD-resources/rec-documents/REST-Nov2011.pdf

Gordon, R. T., Schatz, C. B., Myers, L. J., Kosty, M., Gonczy, C., Kroener, J., Tran, M., Kurtzhals, P., Heath, S., Koziol, J. A., Arthur, N., Gabriel, M., Hemping, J., Hemping, G., Nesbitt, S., Tucker-Clark, L., Zaayer, J., 2008. The use of canines in the detection of human cancers. *J. Altern. Complement Med. 14*, 61-67. doi:10.1089/acm.2006.6408

Hajian-Tilaki, K., 2011. Sample size estimation in epidemiologic studies. *Casp. J. Intern. Med. 2*, 289-298.

Hardin, D. S., Anderson, W., Cattet, J., 2015. Dogs can be successfully trained to alert to hypoglycemia samples from patients with type 1 diabetes. *Diabetes Ther*, 1-9. doi:10.1007/s13300-015-0135-x

Hernstein, R. J., Loveland, D. H., Cable, C., 1976. Natural concepts in pigeons. *J. Exp. Psychol.: Anim. Behav. Process. 2*, 285-302.

Herrnstein, R. J., Loveland, D. H., 1964. Complex visual concept in the pigeon. *Science*, *146*, 49-551.

Hirotsu, T., Sonoda, H., Uozumi, T., Shinden, Y., Mimori, K., Maehara, Y., Ueda, N., Hamakawa, M., 2015. A highly accurate inclusive cancer screening test using *Caenorhabditis elegans* scent detection. *PLoS One 10*, 15. doi:10.1371/journal.pone.0118699

Horvath, G., Andersson, H., Paulsson, G., 2010. Characteristic odour in the blood reveals ovarian carcinoma. *BMC Cancer 10*, 6. doi:10.1186/1471-2407-10-643

Horvath, G., Andersson, H., Nemes, S., 2013. Cancer odor in the blood of ovarian cancer patients: A retrospective study of detection by dogs during treatment, 3 and 6 months afterward. *BMC Cancer 13*, 7. doi:10.1186/1471-2407-13-396

Horvath, G., Jarverud, G. A., Jarverud, S., Horvath, I., 2008. Human ovarian carcinomas detected by specific odor. *Integr. Cancer Ther. 7*, 76-80. doi:10.1177/1534735408319058

Jezierski, T., Walczak, M., Ligor, T., Rudnicka, J., Buszewski, B., 2015. Study of the art: Canine olfaction used for cancer detection on the basis of breath odour. Perspectives and limitations. *J. Breath Res. 9*, 12. doi:10.1088/1752-7155/9/2/027001

Johnen, D., Heuwieser, W., Fischer-Tenhagen, C., 2013. Canine scent detection - Fact or fiction? *Appl. Anim. Behav. Sci. 148*, 201-208. doi:10.1016/j.applanim.2013.09.002

Jones, S. R., Carley, S., Harrison, M., 2003. An introduction to power and sample size estimation. *Emerg. Med. J. 20*, 453-458. doi:10.1136/emj.20.5.453

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. doi:10.1016/j.applanim.2013.11.010

Lippi, G., Cervellin, G., 2012. Canine olfactory detection of cancer versus laboratory testing: Myth or opportunity? *Clin. Chem. Lab. Med. 50*, 435-439. doi:10.1515/cclm.2011.672

Lit, L., Schweitzer, J. B., Oberbauer, A. M., 2011. Handler beliefs affect scent detection dog outcomes. *Anim. Cogn. 14*, 387-394. doi:10.1007/s10071-010-0373-2

Luque de Castro, M. D., Fernandez-Peralbo, M. A., 2012. Analytical methods based on exhaled breath for early detection of lung cancer. *Trac-Trends Anal. Chem. 38*, 13-20. doi:10.1016/j.trac.2012.03.018

Mahoney, A., Edwards, T. L., LaLonde, K., Beyene, N., Cox, C., Weetjens, B. J., Poling, A., 2014. Pouched rats' (*Cricetomys gambianus*) detection of *Salmonella* in horse feces. *J. Vet. Behav.: Clin. Appl. Res. 9*, 124-126. doi:10.1016/j.jveb.2014.02.001

Mahoney, A., Weetjens, B. J., Cox, C., Beyene, N., Reither, K., Makingi, G., Jubitana, M., Kazwala, R., Mfinanga, 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. doi:10.1155/2012/716989

Mahoney, A. M., Weetjens, B. J., Cox, C., Beyene, N., Mgode, G., Jubitana, M., Kuipers, D., Kazwala, R., Mfinanga, G. S., Durgin, A., 2011. Using giant African pouched rats to detect tuberculosis in human sputum samples: 2010 findings. *Pan Afr. Med. J. 9*.

Marcus, D. A., 2012. *Therapy dogs in cancer care*. New York, NY: Springer.

Maurer, M., McCulloch, M., Willey, A. M., Hirsch, W., Dewey, D., 2016. Detection of Bacteriuria by Canine Olfaction. *Open Forum Infect. Dis. 3*, ofw051. doi:10.1093/ofid/ofw051

McCulloch, M., Turner, K., Broffman, M., 2012. Lung cancer detection by canine scent: Will there be a lab in the lab? *Eur. Resp. J. 39*, 511-512. doi:10.1183/09031936.00215511

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. *Integr. Cancer Ther. 5*, 30-39. doi:10.1177/1534735405285096

Mgode, G. F., Weetjens, B. J., Nawrath, T., Cox, C., Jubitana, M., Machang'u, R. S., Cohen-Bacrie, S., Bedotto, M., Drancourt, M., Schulz, S., Kaufmann, S. H. E., 2012. Diagnosis of tuberculosis by trained African giant pouched rats and confounding impact of pathogens and microflora of the respiratory tract. *J. Clin. Microbiol. 50*, 274-280. doi:10.1128/jcm.01199-11

Moser, E., McCulloch, M., 2010. Canine scent detection of human cancers: A review of methods and accuracy. *J. Vet. Behav.: Clin. Appl. Res. 5*, 145-152. doi:10.1016/j.jveb.2010.01.002

Nevin, J. A., 1988. Behavioral momentum and the partial-reinforcement effect. *Psychol. Bull. 103*, 44-56. doi:10.1037/0033-2909.103.1.44

Nevin, J. A., 2012. Resistance to extinction and behavioral momentum. *Behav. Process. 90*, 89-97.

Oh, Y., Lee, Y., Heath, J., Kim, M., 2015. Applications of animal biosensors: A review. *IEEE Sens. J. 15*, 637-645. doi:10.1109/jsen.2014.2358261

Oxley, J. C., Waggoner, L. P., 2009. Detection of explosives by dogs. In Marshall, M., Oxley, J. (Eds.), *Aspects of explosives detection*. Amsterdam, Netherlands: Elsevier, pp. 27-40.

Pickel, D., Manucy, G. P., Walker, D. B., Hall, S. B., Walker, J. C., 2004. Evidence for canine olfactory detection of melanoma. *Appl. Anim. Behav. Sci. 89*, 107-116. doi:10.1016/j.applanim.2004.04.008

Poling, A., Weetjens, B. J., Cox, C., Beyene, N. W., Sully, A., 2010a. Using giant African pouched rats (*Cricetomys gambianus*) to detect landmines. *Psychol. Rec. 60*, 715-727.

Poling, A., Weetjens, B., Cox, C., Beyene, N., Durgin, A., Mahoney, A., 2011. Tuberculosis detection by giant African pouched rats. *Behav. Anal. 34*, 47-54.

Poling, A., Weetjens, B. J., Cox, C., Mgode, G., Jubitana, M., Kazwala, R., Mfinanga, G. S., in 't Veld, D. H., 2010b. Short report: Using giant African pouched rats to detect tuberculosis in human sputum samples: 2009 findings. *Am. J. Trop. Med. Hyg. 83*, 1308-1310. doi:10.4269/ajtmh.2010.10-0180

R Core Team, 2015. *R: A language and environment for statistical computing.* R Foundation for Statistical Computer. Vienna, Austria. Retrieved from https://www.R-project.org/

Rains, G. C., Tomberlin, J. K., Kulasiri, D., 2008. Using insect sniffing devices for detection. *Trends Biotechnol. 26*, 288-294. doi:10.1016/j.tibtech.2008.02.007

Reither, K., Jugheli, L., Glass, T. R., Sasamalo, M., Mhimbira, F. A., Weetjens, B. J., Cox, C., Edwards, T. L., Mulder, C., Beyene, N. W., Mahoney, A., 2015. Evaluation of giant African pouched rats for detection of pulmonary tuberculosis in patients from a high-endemic setting. *PLoS One 10*, 13. doi:10.1371/journal.pone.0135877

Rudnicka, J., Walczak, M., Kowalkowski, T., Jezierski, T., Buszewski, B., 2014. Determination of volatile organic compounds as potential markers of lung cancer by gas chromatography-mass spectrometry versus trained dogs. *Sens. Actuator B-Chem. 202*, 615-621. doi:10.1016/j.snb.2014.06.006

Ruffman, T., Morris-Trainor, Z., 2011. Do dogs understand human emotional expressions? *J. Vet. Behav.: Clin. Appl. Res. 6*, 97-98. doi:10.1016/j.jveb.2010.08.009

Sargisson, R., McLean, I., 2010. The effect of reinforcement rate variations on hits and false alarms in remote explosive scent tracing with dogs. *J. ERW Mine Action 14*, 64-68.

Sethi, S., Nanda, R., Chakraborty, T., 2013. Clinical application of volatile organic compound analysis for detecting infectious diseases. *Clin. Microbiol. Rev. 26*, 462-475. doi:10.1128/cmr.00020-13

Shirasu, M., Touhara, K., 2011. The scent of disease: Volatile organic compounds of the human body related to disease and disorder. *J. Biochem. 150*, 257-266. doi:10.1093/jb/mvr090

Sonoda, H., Kohnoe, S., Yamazato, T., Satoh, Y., Morizono, G., Shikata, K., Morita, M., Watanabe, A., Morita, M., Kakeji, Y., Inoue, F., Maehara, Y., 2011. Colorectal

cancer screening with odour material by canine scent detection. *Gut 60*, 814-819. doi:10.1136/gut.2010.218305

Soproni, K., Miklósi, Á., Topál, J., Csányi, V., 2002. Dogs' (*Canis familiaris*) responsiveness to human pointing gestures. *J. Comp. Psychol. 116*, 27-34. doi:10.1037//0735-7036.116.1.27

Suckling, D. M., Sagar, R. L., 2011. Honeybees *Apis mellifera* can detect the scent of *Mycobacterium tuberculosis*. *Tuberc. 91*, 327-328. doi:10.1016/j.tube.2011.04.008

Talou, T., Gaset, A., Delmas, M., Kulifaj, M., Montant, C., 1990. Dimethyl sulfide - The secret for black truffle hunting by animals? *Mycol. Res. 94*, 277-278.

Taverna, G., Tidu, L., Grizzi, F., Torri, V., Mandressi, A., Sardella, P., La Torre, G., Cocciolone, G., Seveso, M., Giusti, G., Hurle, R., Santoro, A., Graziotti, P., 2015a. Olfactory system of highly trained dogs detects prostate cancer in urine samples. *J. Urol. 193*, 1382-1387. doi:10.1016/j.juro.2014.09.099

Taverna, G., Tidu, L., Grizzi, F., Stork, B., Mandressi, A., Seveso, M., Bozzini, G., Sardella, P., Latorre, G., Lughezzani, G., Buffi, N., Casale, P., Fiorini, G., Lazzeri, M., Guazzoni, G., 2015b. The ability of dogs to detect human prostate cancer before and after radical prostatectomy. *EC Vet. Sci. 2*, 47-51.

Urbanova, L., Vylmankova, V., Krisova, S., Pacik, D., Necas, A., 2015. Intensive training technique utilizing the dog's olfactory abilities to diagnose prostate cancer in men. *Acta Vet. BRNO 84*, 77-82. doi:10.2754/avb201585010077

Virányi, Z., Topál, J., Gácsi, M., Miklósi, Á., Csányi, V., 2004. Dogs respond appropriately to cues of humans' attentional focus. *Behav. Process. 66*, 161-172. doi:10.1016/j.beproc.2004.01.012

Walczak, M., Jezierski, T., Gorecka-Bruzda, A., Sobczynska, M., Ensminger, J., 2012. Impact of individual training parameters and manner of taking breath odor samples on the reliability of canines as cancer screeners. *J. Vet. Behav.: Clin. Appl. Res. 7*, 283-294. doi:10.1016/j.jveb.2012.01.001

Weetjens, B. J., Mgode, G. F., Machang'u, R. S., Kazwala, R., Mfinanga, G., Lwilla, F., Cox, C., Jubitana, M., Kanyagha, H., Mtandu, R., Kahwa, A., Mwessongo, J., Makingi, G., Mfaume, S., Van Steenberge, J., Beyene, N. W., Billet, M., Verhagen, R., 2009. African pouched rats for the detection of pulmonary tuberculosis in sputum samples. *Int. J. Tuberc. Lung Dis. 13*, 737-743.

Wells, D. L., 2012. Dogs as a diagnostic tool for ill health in humans. *Altern. Ther. Health Med. 18*, 12-17.

Williams, H., Pembroke, A., 1989. Sniffer dogs in the melanoma clinic. *Lancet 1*, 734-734.

Williams, M., Johnston, J. M., 2002. Training and maintaining the performance of dogs (*Canis familiaris*) on an increasing number of odor discriminations in a controlled setting. *Appl. Anim. Behav. Sci. 78*, 55-65. doi:10.1016/s0168-1591(02)00081-3

Willis, C. M., Britton, L. E., Harris, R., Wallace, J., Guest, C. M., 2010. Volatile organic compounds as biomarkers of bladder cancer: Sensitivity and specificity using trained sniffer dogs. *Cancer Biomark. 8*, 145-153. doi:10.3233/cbm-2011-0208

Willis, C. M., Church, S. M., Guest, C. M., Cook, W. A., McCarthy, N., Bransbury, A. J., Church, M. R. T., Church, J. C. T., 2004. Olfactory detection of human bladder cancer by dogs: Proof of principle study. *Br. Med. J. 329*, 712-714A. doi:10.1136/bmj.329.7468.712

Wolfe, J. M., Horowitz, T. S., Kenner, N. M., 2005. Rare items often missed in visual searches. *Nature 435*, 439-440. doi:10.1038/435439a

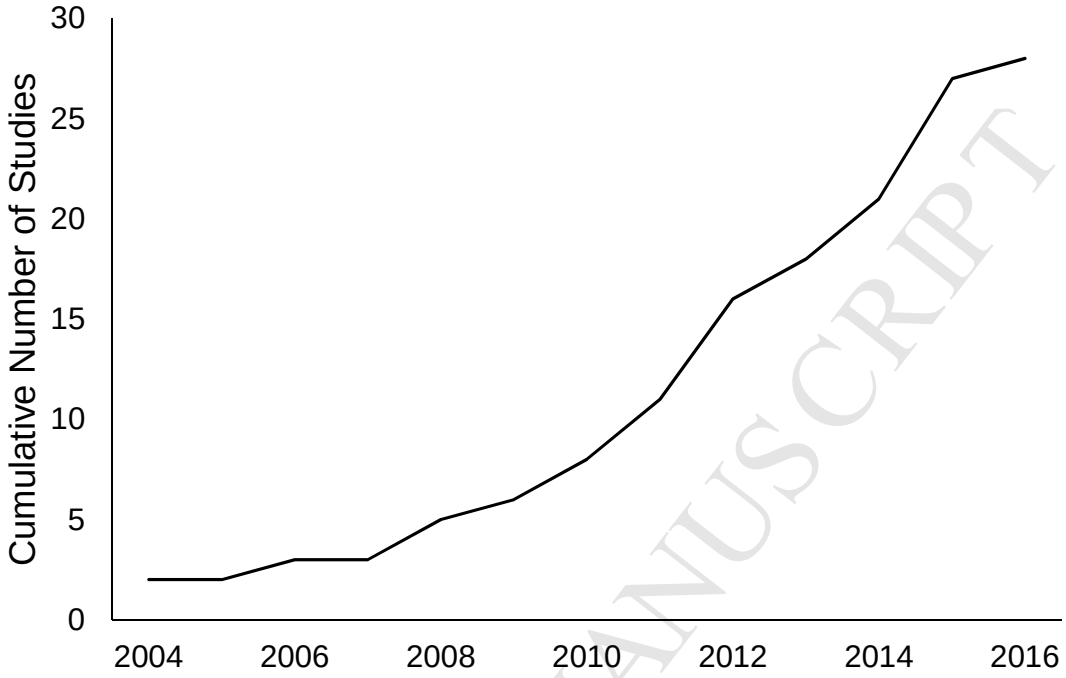
Yoshida, K., Hirotsu, T., Tagawa, T., Oda, S., Wakabayashi, T., Iino, Y., Ishihara, T., 2012. Odour concentration-dependent olfactory preference change in *C. elegans*. *Nat. Commun. 3*, 11. doi:10.1038/ncomms1750

Zimmerman, J., Ferster, C. B., 1963. Intermittent punishment of S^{Δ} responding in matching to sample. *J. Exp. Anal. Behav. 6*, 349-356. doi:10.1901/jeab.1963.6-349

Figure Caption

Figure 1. Cumulative number of studies examining animal olfactory detection of human diseases between 2004 and 2016.

Figure 1



Highlights

- Guidelines for training and testing disease-detection animals are provided
- Using these guidelines as points for comparison the relevant literature is reviewed
- Inconsistent findings have been reported, likely because of procedural differences
- Some commonly used training and testing procedures are problematic
- We make recommendations for research and reporting practices in this area