Journal of Veterinary Behavior 15 (2016) 66–77

Contents lists available at ScienceDirect

Journal of Veterinary Behavior

journal homepage: www.journalvetbehavior.com

Canine Research

No better than flipping a coin: Reconsidering canine behavior evaluations in animal shelters

Gary J. Patronek [a,*], Janis Bradley [b]

[a] Center for Animals and Public Policy, Cummings School of Veterinary Medicine, Tufts University, North Grafton, Massachusetts
[b] The National Canine Research Council, Amenia, New York

ARTICLE INFO

Article history:
Received 12 May 2016
Received in revised form
27 July 2016
Accepted 4 August 2016
Available online 13 August 2016

Keywords:
animal shelter
dog behavior evaluation
aggression
dog personality
sensitivity
predictive value

ABSTRACT

Use of behavior evaluations for shelter dogs has progressed despite their lack of scientific validation as reliable diagnostic tools. Yet results of these evaluations are often used to make life-and-death decisions. Despite acknowledging the significant limitations of evaluations, most authors suggest that the solution is to continue to attempt to remedy deficiencies. We take a contrary position and use existing data and principles of diagnostic test evaluation to demonstrate that reliably predicting problematic behaviors in future adoptive homes is vanishingly unlikely, even in theory, much less under the logistical constraints of real-world implementation of these evaluations in shelters. We explain why it would be difficult, if not impossible, to calculate robust values for sensitivity and specificity of a shelter canine behavior evaluation as required for any valid diagnostic test. We further explain the consequences of disregarding the effect of prevalence on the predictive value of a positive test (e.g., eliciting biting or warning behavior from the dog in the behavior evaluation). Finally, we mathematically demonstrate why, for any plausible combination of sensitivity, specificity, and prevalence of biting and warning behaviors, a positive test would at best be not much better than flipping a coin, and often be much worse, because many of the dogs who test positive will be false positives. Shelters already screen from adoption obviously dangerous dogs during the intake process. Subsequent provocative testing of the general population of shelter dogs is predicated on an assumption of risk that is far in excess of existing data and relies on assumptions about dog behavior that may not be supportable. We suggest that instead of striving to bring out the worst in dogs in the stressful and transitional environment of a shelter and devoting scarce resources to inherently flawed formal evaluations that do not increase public safety, it may be far better for dogs, shelters, and communities if effort spent on frequently misleading testing was instead spent in maximizing opportunities to interact with dogs in normal and enjoyable ways that mirror what they are expected to do once adopted (e.g., walking, socializing with people, playgroups with other dogs, games, training). In conjunction with a thorough and objective intake history when available, these more natural types of assessment activities will help identify any additional dogs whose behavior may be of concern. Engaging in the normal repertoire of activities familiar to pet dogs has the additional benefit of enriching dogs' lives and minimizing the adverse effect of being relinquished and confined to a shelter, will be more indicative of the typical personality and behavior of dogs, and may help make dogs better candidates for adoption.

Introduction

Use of formal canine behavior evaluations in animal shelters as a way to assess propensity for various undesirable behaviors in dogs before making them available for adoption to the public has been going on for more than 2 decades. The first published report of a behavior evaluation of shelter dogs appeared in the literature in 1991 (Van der Borg et al., 1991), and various other instruments have

* Address for reprint requests and correspondence: Gary J. Patronek, Center for Animals and Public Policy, Cummings School of Veterinary Medicine, Tufts University, North Grafton, MA 01536.
E-mail address: gary.patronek@tufts.edu (G.J. Patronek).

http://dx.doi.org/10.1016/j.jveb.2016.08.001

been developed since (Haverbeke et al., 2015). These range from very systematic batteries of tests designed by individuals credentialed in animal behavior, to *ad hoc* procedures developed by shelter staff members, to impromptu combinations of both that have been modified and adapted according to the preferences of different users. Although each evaluation is different, they generally include exposing dogs to a series of provocative stimuli (tests) in a semi-controlled environment to determine whether behaviors such as growling, snarling, snapping, lunging, or biting can be elicited, sometimes along with other behaviors that might prove either problematic or even desirable (e.g., trainability) in an adoptive home. In our experience, the resources required to conduct these evaluations are substantial, and shelters may rely on the results to make life-and-death decisions for dogs, so the consequences are significant for all involved.

The extent of use of formal canine behavior evaluations is unknown, but results from one online convenience sample of mostly small, private sheltering organizations indicated that about 25% of the organizations used one, with most of those (60%) using a test of their own design (D'Arpino et al., 2012). Large, public shelters, however, were very underrepresented in that sample. Although we have no systematic information either on why shelters came to adopt this practice or their current reasons for maintaining it, anecdotal reports among people involved in shelter work suggest that they originally emanated from a desire to protect the public from potentially dangerous dogs. In some cases, this has grown to include making the best match between dogs and adopters or trying to identify behavioral issues that may require attention while in the shelter. Another underlying motivation may be to remove or mitigate some of the emotional stress on shelter staff when confronted with making euthanasia decisions to make space for incoming dogs. In these situations, the behavior evaluation process could provide the appearance of a less-arbitrary, more justifiable rationale than number of days in the shelter or workers' opinions about which dogs would be more attractive to adopters. It is also possible that shelter staff or board members may have been influenced by reports in the medical, veterinary, and behavioral literature in which dog bites frequently are framed as an epidemic (despite declines of about 90% in reports of dog bites from the 1970s through the 2000s) (NCRC, undated). Numerous published reports about reasons for relinquishment of dogs to shelters may also have contributed to an impression that shelter dogs are "damaged goods," somehow markedly different from owned dogs. This would be unfortunate because data indicate that human-related factors such as housing, cost of care and/or veterinary treatment, and family problems are important contributors to relinquishment (Weiss et al., 2015; see Coe et al., 2014 for a comprehensive review). Furthermore, being relinquished for a manageable problem (e.g., housetraining) likely reflects more on the owner's commitment and ability than on the dog. The desire on the part of shelters to avoid liability may also play a role, but the question is one that needs study. (Interestingly, legal experts have not come to an agreement about what effect performing such an evaluation would have on a shelter's liability in the event of a bite. They have, however, identified several strategies to reduce liability, such as being sure that ownership of the dog is transferred at the time of adoption and disclosing any information the shelter has regarding prior behavior [Lutz, 2009]).

How do we begin to evaluate the merits of canine behavior evaluations in shelters as valid diagnostic instruments? The goal of a clinical diagnostic is to determine whether a subject has a particular condition or trait. This is seldom straightforward for any diagnostic test, as there is not always a clear biological "black and white" cutpoint for an individual who is positive or negative for a condition. It is even more challenging for a condition requiring a subjective assessment. A good example is the radiology literature, where studies have shown that agreement about the diagnosis of a physical condition or disease state on a radiograph at a single point in time is far from perfect, even among seasoned specialists working under ideal conditions (e.g., Arealis et al., 2014; Khan et al., 2011; Matsunaga et al., 2009). For a canine behavior evaluation, "diagnosis" would involve ascertaining not only whether a dog did or did not exhibit a behavior of interest on one or more tests in the shelter but also that the behavior, if it occurred, constituted a stable trait that would be expressed in other contexts and that it posed a danger. In the unlikely case that the first of these conditions could achieve reliability, the other two remain entirely speculative.

A large body of science has developed around the principles of developing, assessing, validating, and using diagnostic tests. The formulas and principles for evaluating key attributes of diagnostic tests (sensitivity, specificity, predictive value of a positive test, predictive value of a negative test) are well established and fairly straightforward. However, the process of doing so is complicated and costly, and it is unsurprising that no behavior evaluation for shelter dogs has yet been scientifically validated. Given the resource-constrained environments of animal shelters, and a sincere desire to adopt best practices when possible, promotion and use of behavior evaluations for shelter dogs has progressed well ahead of their scientific validation as a reliable diagnostic tool. Indeed, one of the authors has been involved in efforts to develop, implement, and validate such behavioral tests (Gary J. Patronek) and the other (Janis Bradley) has been involved in administering tests. The limitations of canine behavioral evaluations have been well described, although the tendency is that after conceding these points, most authors suggest that the solution is to attempt to remedy the deficiencies (Rayment et al., 2015; King et al., 2012; Mornement et al., 2010; van der Borg et al., 2010; Diesel et al., 2008; Christensen et al., 2007; Diederich & Giffroy, 2006; Taylor & Mills, 2006).

In this article, we take a contrary position and argue that it might be time to step back and ask a more fundamental question—namely, is it even feasible to develop a canine behavioral evaluation that is sufficiently predictive of certain unwanted behaviors in the future home to justify the cost to shelters and dogs? To address that question, we unpack each of the criteria and assumptions for constructing and validating diagnostic tests and examine some conceptual issues related to canine behavior and conducting these tests in a shelter. We will limit the discussion to the evaluation of behaviors considered as dangerous by the test designers because of the emphasis on provoking warning and biting behaviors and because this is consistently the top, sometimes the only, priority of organizations that use behavior evaluations. Finally, we will explain why eliciting warning and biting behaviors (referred to here as a positive finding or positive test) in particular is no better than flipping a coin in terms of informative value for either improving public safety or justifying euthanasia decisions for dogs and make recommendations for moving forward. The simulations described in this article demonstrate how achieving a result better than simple chance with regard to reliably predicting whether dogs will exhibit growling, snarling, snapping, or biting behavior that becomes problematic in their adoptive homes is vanishingly unlikely, even in theory, much less under the logistical constraints of real-world implementation in shelters.

Key attributes of diagnostic tests

Sensitivity and specificity

Every diagnostic test has 2 inherent characteristics, sensitivity and specificity, that play a major role in determining the

performance or validity of that test in real-world populations of individuals. Sensitivity is the percentage of individuals who are actually positive for the condition whom the test identifies as positive, and specificity indicates how many of those who are actually negative for the condition the test identifies as negative. These characteristics influence the ability to predict an accurate result (either positive or negative) on an evaluation for any given individual dog. Sensitivity and specificity may be calculated after the 4 cells (a, b, c, and d) of a 2 × 2 table (Figure 1) are filled in with the correct values. These concepts are reviewed in depth in standard epidemiology texts and there may well be hundreds of articles explaining them in the scientific literature (perhaps an indication of the extent to which clinicians find them confusing); a very succinct and accessible summary has been published by Akobeng (2007), and much information is available on the Internet.

A key component of calculating sensitivity and specificity involves comparing the test results against a reference standard (e.g., best available confirmatory diagnostic) for the condition in question (in Figure 1, we label this as "Response of the dog post-adoption"). We contend it would be extremely difficult if not impossible to calculate sensitivity and specificity for a behavior evaluation. One obstacle is that this reference standard cannot ethically or practically be fully implemented to determine the rate of either true positives or false positives.

Animal shelters take seriously their responsibility to protect the public, and no one wants to place a dog in a situation where she/he would be a danger to herself/himself or others. Consequently, it is a common practice in shelters for dogs being surrendered with a history of biting or serious attempts to bite to be euthanized (or sometimes placed with a qualified rescue group or sanctuary). The same is usually true for a dog who attempts to bite any of the shelter personnel or is too threatening to be safely handled. Therefore, many dogs suspected of having the condition of interest (i.e., believed to be true positives for biting and/or warning behavior) will have been removed from the testing pool. This issue will become very important later when we consider the critically important influence of the prevalence of problematic behavior on the predictive value of a positive test in the behavior evaluation when applied to the general population of shelter dogs, most of whom do not have a known history of biting or warning behavior. Some shelters do place dogs believed to have manageable behavior problems into adoptive homes, and when that occurs, it is typical that the adopter will be provided management instructions to minimize, if not eliminate, opportunities for that behavior to be triggered or behavior modification instructions to change the dog's responses. This was done, for example, with dogs who had tested positive for food guarding at the Wisconsin Humane Society and were adopted (Mohan-Gibbons et al., 2012). However, such sensible

Shelter behavior evaluation result:	Response of the dog post-adoption[a]	
	Dog has problematic behavior	Dog does not have problematic behavior
Dog tests positive (detect problematic behaviors)	# of *True Positive* dogs (a) (dog tests positive and dog will show problematic behavior)	# of *False Positive* dogs (c) (dog tests positive but dog will not show problematic behavior)
Dog tests negative (do not detect problematic behaviors)	# of *False Negative* dogs (b) (dog tests negative but dog will show problematic behavior)	# of *True Negative* dogs (d) (dog tests negative and dog will not show problematic behavior)

Sensitivity: The ability of the test to correctly identify dogs who have the problematic behavior; it is calculated as the proportion of dogs with the behavior who test positive. Calculated DOWN the first COLUMN [a/(a+b)]

Specificity: The ability of the test to correctly identify dogs who do not have the problematic behavior; it is calculated as the proportion of dogs without the behavior who test negative. Calculated DOWN the second COLUMN [d/(c + d)]

Predictive value of a positive test (positive predictive value**)**: The proportion of dogs who test positive who actually have the problematic behavior. Calculated ACROSS the first ROW [a/(a + c)]. This answers the critically important question, "*If a dog tests positive in the shelter, what is the probability that she or he has the problematic behavior*"?

Predictive value of a negative test (negative predictive value): The proportion of dogs who test negative who are actually free of the problematic behavior. Calculated ACROSS the second row [d/(b + d)]. This answers the critically important question, "*If a dog tests negative in the shelter, what is the probability that she or he does not have the problematic behavior*?"

Figure 1. Standard 2 × 2 table for calculating key attributes of a canine behavior evaluation. [a]*Only for purposes of simplifying the mathematical analysis for this hypothetical scenario, we will assume that problematic behavior in the home can be unambiguously defined and dichotomized into present/absent.* Problems with these assumptions are discussed in the text.

and pragmatic precautions to prevent a problem would also interfere with evaluating test performance in a research context.

Consistency of definitions for behavioral end points

Validating any behavioral evaluation would require clear criteria for the problematic behavior being studied (Overall, 2015), and we believe those criteria should also have clinical relevance in the home environment, as opposed to simply showing that the same behavior can be replicated after adoption through additional testing. The word "aggression" commonly appears in discussions of canine behavior and canine behavior evaluations to denote one category of problematic behaviors of concern. A full discussion of what is meant by the term "aggression" is beyond the scope of this article. We have avoided use of the term and place it here in quotes to emphasize that it has been used in so many different contexts that it may have little practical value. The term itself is subject to multiple, sometimes contradictory, definitions even in the behavior literature. One group of authors found different categorizations and descriptions of human-directed aggression in each of 7 articles and a general failure to distinguish between context, motivation of the dog, and emotion (Kikuchi et al., 2014).

We believe there would be general agreement among behaviorists that aggression is a heterogeneous group of postures and actions that are part of the normal behavioral repertoire of the dog, which can occur on a spectrum and vary in frequency and intensity over time, with different stimuli, and in different environments. Behaviors labeled as aggressive in shelters typically include both warning signals (growling, snarling, snapping, and sometimes barking and lunging) and actual biting (both injurious and non-injurious), but sometimes behaviors so labeled are simply neutral or even affiliative, as in the case of the dog described as aggressive because he "climbs" the leash with his mouth in an effort to use it as a tug toy and/or resist its direction, or even the dog who has been severely deprived of opportunities for interspecies or intraspecies interaction and so thrashes around on leash in an effort to close the distance between himself and any person or dog who comes into view. There is evidence from published studies that many of the behaviors elicited during a behavior evaluation that might be deemed to indicate an aggressive temperament are more normal than pathological. For example, Guy et al. (2001) collected information on frequency of certain aggressive behaviors in dogs toward familiar people (e.g., growling, growling/snapping over food or objects, and biting) via a survey from 3,226 dog owners attending 20 general veterinary practices. Many of these responses (18.5%) were for dogs aged <1 year, reflecting high visitation rates for puppies. Behaviors were counted as aggressive even if the owner felt the dog was just growling during play or a bite was during play/deemed accidental. The rate of growling, and/or snarling was 41%, and biting was 15.6%, suggesting that these are normal, common behaviors in the home, and that using these terms to define dangerous behavior would need to be done with considerable qualification of the intensity, frequency, and circumstances. Therefore, we argue that merely showing that a dogs' response to a stimulus (e.g., growling when seeing a stranger approaching in the testing room at a shelter) can be predicted does not necessarily confirm that the dog's behavior is abnormal, nor that the owner will see it as problematic given the circumstances, or that it will present a problem in the future home.

In practice, shelter behavior evaluations define "aggressive" behaviors as having passed the threshold of whatever the specific agency deems as too much for adoptability, under the assumption that the same type and level of behavior would occur in the home following presumptively equivalent stimuli. This threshold on the test can range from a single growl on any one of a battery of tests, to multiple bites to a model or device used for testing, for example, a fake hand used to interfere with a dog while he is eating or a doll used to simulate a child in the shelter. So in practice, the term, "aggressive" is defined more by circumstance and institutional policy than behavioral science, and by itself has little value as a reference standard. With respect to the appropriate reference standard, Sheppard & Mills (2003) point out the problems inherent in a medical model where even normal behavior is pathologized and dichotomized as present/absent, much like an infectious disease or injury that needs to be diagnosed and treated. The utility of this categorical approach in human psychology has been criticized by none other than Dr. Allen Frances, the chair of the DSM-4 Task Force, in his book "Saving Normal: An Insider's Revolt against Out-of-Control Psychiatric Diagnosis, DSM-5, Big Pharma, and the Medicalization of Ordinary Life" (Frances, 2014).

To further complicate matters, unlike diagnosing a physical condition such as a tumor or coronary artery disease, which over the course of hours or days will be comparatively static, dog behavior is extraordinarily plastic and can vary from moment to moment in both frequency and intensity in response to a particular stimulus or to different stimuli. Two stimuli might appear essentially similar to people but be perceived as very different by dogs due to other contextual factors that differ between shelters and homes, or even within the shelter. For example, we are aware of evaluations where dogs presented with another dog on a leash reacted in a way that resulted in them being deemed dog aggressive, but when those same dogs were allowed to interact with other dogs off-leash in a shelter play group, no dog-dog issues were noted. One study of a food aggression test widely used in shelters found low predictability with regard to subsequent food guarding in the home and reported there was little concern on the part of the adopters about whether the behavior occurred or not, as simple management practices such as isolating the dog during feeding could easily prevent a problem (Marder et al., 2013). Another study also found poor predictability with respect to food guarding (Mohan-Gibbons et al., 2012). Few dogs continued to express guarding behavior after 3 months in a home in spite of low owner compliance with protocols recommended for dogs who had shown food guarding. In fact, against the shelter's advice, owners often engaged in the provocative behavior that had elicited guarding in the behavior evaluation (e.g., picking up the food bowl while the dog was eating) without any similar response. Furthermore, dogs the shelter identified as food guarders were returned at a lower rate than the general adopted population, and none were returned for food guarding. In preliminary results of a study of shelters that discontinued food-guarding tests, no difference has been found in adoption, return, length of stay, or live release rates since suspending testing for food guarding (Weiss, 2016). Therefore, determining what types of provocative tests in the stressful, unfamiliar environment of a shelter would be relevant for eliciting a behavior of actual concern in the very different environment of a future home seems extremely problematic.

Single tests versus battery tests

It might be argued that this simple scenario we have discussed of a single provocative test would not accurately reflect sensitivity and specificity in shelter practice, where batteries of individual tests or subtests (potentially with each having different sensitivities and specificities) are used collectively to make a determination about a dog's behavioral tendencies. Battery testing, however, comes with its own trade-offs in sensitivity and specificity. In the typical battery test situation, a number of individual tests are performed sequentially, and in general, the next test is not dependent on the first (i.e., parallel testing). In these circumstances, if the

Table 1
Examples of sensitivity and specificity for some representative medical and behavioral diagnostic tests in people

Test	Reference	Sensitivity	Specificity
Anterior drawer test for cruciate rupture	Jackson et al., 2003	0.48	0.87
Abdominal ultrasound for appendicitis	Terasawa et al., 2004	0.86	0.81
Clinical breast exam for cancer	Bobo et al., 2000	0.59	0.93
Mammography for breast cancer	Banks et al., 2004	0.86	0.96
MRI for breast cancer	Medeiros et al., 2011 (Meta-analysis)	0.90	0.75
Fecal occult blood for colon cancer	Collins et al, 2005	0.24	0.94
Cardiac MRI for coronary artery disease	Schuetz et al., 2010 (Meta-analysis)	0.87	0.70
Cardiac CT for coronary artery disease	Shuetz et al., 2010 (Meta-analysis)	0.97	0.87
Urine culture for urinary tract infection in symptomatic women	Stamm et al., 1982	0.95	0.85
Rapid strep test children	Gurol et al., 2010	0.70	0.98
Rapid strep test adults	Gurol et al., 2010	0.59	0.96
Fasting plasma glucose for diabetes	Bennett et al., 2007	0.56	0.96
Alcoholism questionnaire	Buchsbaum et al., 1991	0.74	0.91
Geriatric depression scale	Pocklington et al., 2016 (Meta-analysis)	0.89	0.77
Lethality screen for domestic violence	Messing et al., 2015	0.92	0.21
Future violent offending	Fazel et al., 2012 (Meta-analysis)	0.92	0.36
Future criminal offending	Fazel et al., 2012 (Meta-analysis)	0.41	0.80

Originally identified from: http://www.getthediagnosis.org/browse.php?mode=dx.
Numbers have been rounded where necessary.

results of the test are combined in an “or” fashion, where a positive result on any test results in a dog being deemed “positive” for the condition, then the overall sensitivity of the evaluation is greater than for any test alone, but the overall specificity will be lower, and therefore more false positives will occur. By contrast, if the results of individual tests are combined in an “and” fashion, where a positive result on several tests is required to declare a dog “positive” for the behavior, then the specificity will be higher for the evaluation overall compared with an individual test, but the sensitivity will be lower. Other concerns arise when a battery of tests are used in a serial fashion. For a full discussion of these issues, see http://radiopaedia.org/articles/sensitivity-and-specificity-of-multiple-tests.

In summary, canine behavior evaluations lack an essential component of any valid diagnostic test because key attributes of test validity (sensitivity and specificity) have not been, and likely cannot be, calculated in the context of a research situation in real shelters and adoptive homes. Furthermore, there is neither consensus nor confirmatory research on the specific behaviors elicited during a provocative test in a shelter, the relevant intensity of those behaviors, or the frequency of those behaviors in the various subtests that would be considered indicative of a potentially dangerous dog. One would expect that these deficiencies alone would be sufficient to dispel any notion that canine behavior evaluations can be scientifically validated for use on shelter dogs, but for purposes of continuing this hypothetical scenario, we will assume these hurdles have been overcome.

Assessing potential predictive value of behavioral evaluations

As Akobeng (2007) emphasizes, while sensitivity and specificity are critical for determining the validity of a diagnostic test or risk assessment instrument, they do not have any practical value in conveying the likelihood of a particular individual having a particular diagnosis or engaging in a particular behavior. That clinically relevant information is provided by two other attributes of a diagnostic test: the positive predictive value and the negative predictive value, also calculated from the 4 cells in the 2 × 2 table (Figure 1).

To continue with our hypothetical scenario (and accepting the unrealistic assumption that sensitivity and specificity can in fact be calculated in shelters participating in behavior evaluation research), to get the point where the predictive value of the evaluations can be calculated, we first need to establish plausible ranges of sensitivity and specificity, recognizing that they must fall between 0.0 and 1.00 (0 and 100%). In Table 1, we summarize a sample of published values for human medical and behavioral diagnostic tests. An extensive list is available online (http://www.getthediagnosis.org/browse.php?mode=dx). As these data and other readily available resources indicate, even within a specific medical condition, there can be a considerable range for both sensitivity and specificity of diagnostic tests in different studies.

For our hypothetical scenario, we have decided to draw on sensitivity and specificity values published for behavioral risk assessment in humans (as opposed to say a test for cancer or infectious disease, which involves comparatively less subjectivity in its administration and interpretation). In particular, a large meta-analysis of risk assessments of people summarized 37 individual studies in 24,827 people from 13 countries and reported overall values of 92% and 36%, for sensitivity and specificity, respectively, for predicting violent offending and 41% and 80%, respectively, for predicting criminal offending (Fazel et al., 2012; Table 1). (This meta-analysis also illustrates the level of replication [i.e, number of different studies in different populations] needed to derive reasonably robust values for these key test parameters. Even if a single solid study of a canine behavior evaluation was published, that study must be replicated in different shelters to assert it was in any way generalizable. Given the very limited resources for animal shelter studies, the notion of sufficient replication strikes us as highly unlikely.) We will use these in our simulation.

Students as well as clinicians often confuse the probabilities for sensitivity/specificity and positive/negative predictive value since all are calculated from the same 2 × 2 table, and the distinctions may indeed seem subtle. Nevertheless, the proportion of dogs with problematic behavior who test positive on a behavioral evaluation (sensitivity) is very different from the proportion of dogs who test positive who also actually have problematic behavior (predictive

value of a positive test), as is the proportion of dogs without problematic behavior who test negative (specificity) different from the proportion of dogs with negative tests who will not exhibit problematic behavior (predictive value of a negative test). This situation can be visualized by examining the direction of the calculations in Figure 1—which differs by going either down columns (sensitivity and specificity) or across rows (predictive values).

The problem posed by prevalence

It is critical to appreciate that the predictive value of any diagnostic test is strongly influenced by prevalence of the condition in question in the population being evaluated. "Prevalence is what connects the validity of the test, as determined by sensitivity and specificity, with conditions in the real world." A corollary of this is that in a low prevalence situation, the predictive value of a negative test will tend to be very high (few false negatives), whereas the predictive value of a positive test will be low (i.e., there will be many false-positive results).

What might we use to estimate plausible values for prevalence of problematic behaviors related to aggression in the population of shelter dogs? One measure of interest to shelters as well as adopters would be biting humans. Dog bite statistics are extremely variable across the United States and suffer from a variety of problems with reporting and definition, which make generalizable estimates difficult to come by (Devadas et al., 2013). However, for this exercise, we will start with the highest numbers ever reported in the general US population: 4.7 million persons estimated to have been bitten by dogs in 1994 and 4.5 million in 2001-2003, based on samples from the Injury Control and Risk Survey (ICARIS)-1 (Sacks et al., 1996) and ICARIS-2 (Gilchrist et al., 2008) telephone surveys, respectively, done by the Centers for Disease Control (CDC). In those surveys, respondents were simply asked, "In the past 12 months, has anyone in your household been bitten by a dog?" No criteria for what constituted a bite was provided; so presumably, this must include both reported and unreported bites, many of which were inevitably trivial in nature (e.g., no medical treatment required even if assessment was sought, self-treated bites, etc.) and/or bites that were likely accidental. It should be noted that the prevalence reported by other sources for reported bites or medically attended bites is substantially (80%-90%) lower (Patronek & Slavinsky, 2009). Another widely used statistic would be the estimated number of medically attended bites (799,701 in ICARIS-1 [Sacks et al., 1996] and 885,000 in ICARIS-2 [Gilchrist et al., 2008]). Those estimates would also include some bites evaluated by physicians due to concern about infection or rabies rather than injury *per se*, and which required minimal to no treatment, but again we will use that estimate without qualification. Using those CDC-ICARIS estimates of prevalence for all bites, a population of ~52 million dogs in the United States in 1991 and ~63 million in 2001 (Wise et al., 2002), and conservatively assuming that each bite represents a different dog, would mean that <9% of dogs bite a person at any level of severity or concern in a given year. Another way to look at these numbers of course is to conclude that >90% of dogs did not bite anyone in a given year and that <1.5% of dogs inflicted a bite for which medical assessment was sought, regardless of whether treatment was actually necessary. The proportion actually requiring medical treatment or hospitalization would be much lower.

For a prevalence of ~9% of dogs biting (derived from the CDC-ICARIS surveys) and a test with a sensitivity of 92% and a specificity of 36% (values for predicting risk of future violent offending from the meta-analysis in Table 1, with risk of biting substituted for risk of future violent offending), the positive predictive value of such a test would be only ~12%, meaning that ~88% of dogs identified as likely to bite on the evaluation would be false positives! For the lower prevalence of ~1.3% of dogs associated with medically attended bites (interpolating medically attended bites to be ~842,000 in 2001), and using 63 million for the size of the US dog population, the predictive value of identifying potentially only more serious bites would be ~2%, meaning that almost all (98%) dogs identified by the test as likely to bite in the future would be false positives (see http://vassarstats.net/clin2.html). Alternatively, using test parameters for risk of future criminal offending (sensitivity 41% and specificity 80%) and a prevalence of ~9%, the predictive value of identifying a dog as exhibiting biting or warning behavior in the behavior evaluation would be slightly better (~17%) but still hardly useful, with ~83% of positive tests being false positives for future behavior.

It might be argued that prevalence estimates of biting from random community samples used in the CDC-ICARIS surveys are either unrealistically low for use in shelter dog populations or that behaviors other than those resulting in an actual bite should be screened for. If increasing public safety is the ultimate goal, this argument is difficult to defend because many more dogs express warning signals than actually bite. But we will put this aside and say for the moment that there is also value in predicting which dogs will express warning behaviors in homes, and so it is essential to know or estimate the prevalence of these behaviors. To answer that question, we need to determine what would be a plausible upper limit on prevalence of biting/warning/attempting to bite combined in the tested shelter population. Here we provide 4 independent estimates.

A national telephone survey sponsored by the American Society for the Prevention of Cruelty to Animals suggests that about 16% of dogs who are rehomed, whether to a friend, family member, veterinarian, or shelter/rescue, are rehomed due to owner-perceived aggression (no definition provided). In other words, among 391 owners who rehomed one or more dogs during the previous 5 years, 46% indicated this was due to a pet-related problem and 35% of these described it as due to aggression, for an overall value of ~16% in rehomed dogs (i.e., 35% × 46% = 16%) (Weiss et al., 2015). We cannot know, of course, how many of those were actually misinterpretations of play or greeting behavior, but for our purposes here, we will take them at face value because they were problematic for the owners. To investigate other support for that estimate (~16%), we examined data from the Regional Shelter Relinquishment Study sponsored by the National Council on Pet Population Study and Policy (1995-1996), which reported that at least 1 behavioral reason (out of a possible of 5 behavioral reasons) was listed for 1984 dogs relinquished to 12 US animal shelters (Salman et al., 2000). For 379 dogs for whom behavior was listed as the only reason for relinquishment, biting was listed as the most common reason (22.2%), with 17.4% listing aggression to people and 11.3% listing aggression to animals. Similarly, for 422 people listing mixed reasons for relinquishments, 9.7% listed biting and 12.1% aggression to people. Taking the most conservative position (assuming that each reason was mutually exclusive, which they were not), those data would imply that about 284/1,984 dogs (14.3%) were relinquished for biting or other kinds of aggression. Allowing for the nonmutually exclusive nature of the reasons in that study, the true proportion of dogs relinquished for these problems was likely lower. For a third source of data, Bollen & Horowitz (2008) report owner-provided behavioral history for 1,911 dogs relinquished to an open-admission shelter for whom history in the prior home was provided. Of these, 217 (11.3%) were categorized as having a positive aggression history because their owners reported that they had "growled, snarled, lunged, snapped, or bitten" in response to "strangers or visitors" or "being approached while eating" or "having possessions taken away" or "being removed from furniture" or "being handled." Finally, a convenience sample survey of 3,897 dog owners in the United

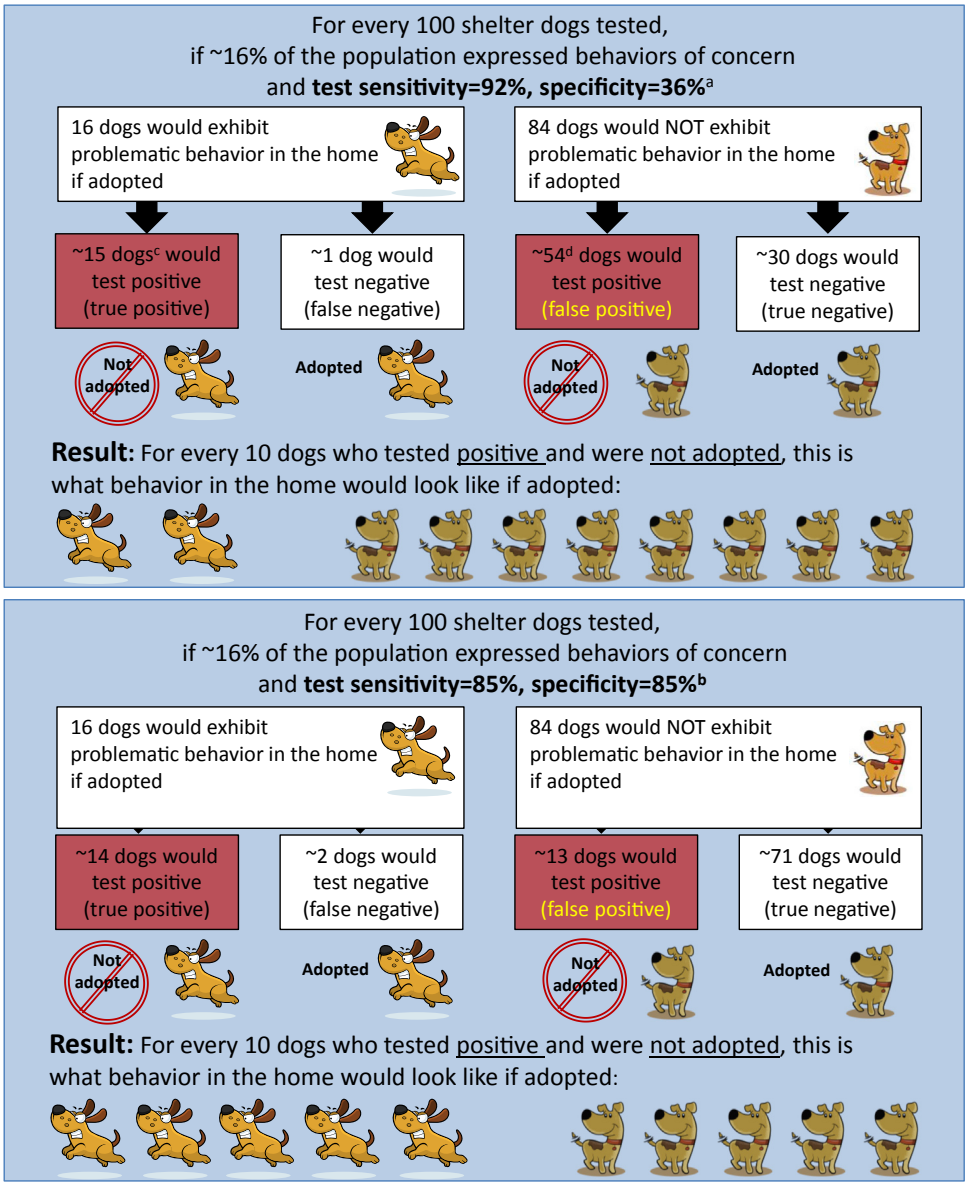

Figure 2. Results of a behavior evaluation using realistic values[a] (top) and unrealistically optimistic values[b] (bottom) for key attributes of a diagnostic test. [a]Values used are those from a meta-analysis of instruments predicting violent offending in people (Fazel et al., 2012). [b]This combination of values exceeds what is commonly reported for many validated human diagnostic tests, which usually involve trade-offs between sensitivity and specificity. [c]Numbers listed are approximate due to rounding of fractions. [d]With a positive predictive value = 22%, 78% (n = 54) of the 69 dogs testing positive will be false positives.

Kingdom suggests a prevalence rate similar to the 11% and 16% cited previously for dogs relinquished to shelters among owned dogs. For each question framed along the lines of "does your dog…?," owners were asked about currently occurring behavior, behavior which had but no longer occurred, behavior that had ever occurred, and also whether they considered the behavior a problem. Among these owners, 579 (14.8%) reported a history (ever having occurred) of barking, lunging, growling, or biting family members, and/or unfamiliar people entering the home, and/or encountered outside the home (Casey et al., 2014). This was a lifetime prevalence of ever occurring, and the questions were not mutually exclusive, so the number 14.8% must include some owners reporting >1 type of aggressive behavior ever occurring. Indeed, 200 of these responses indicated that they occurred in the past but were not presently a problem, which further underscores the plasticity and circumstantial nature of dog behavior.

All of these independent sources of data suggest that a prevalence of ~16% is a plausible starting point for these types of problematic behavior in dogs relinquished to shelters. And given that some unknown portion of dogs surrendered with a history of biting or aggressive behavior will typically be euthanized or otherwise removed from the pool of dogs undergoing formal behavior evaluation, we believe it is also a conservative estimate for prevalence in the evaluated population (in the sense that a larger number will maximize the predictive value of a positive test and put the results of a dog behavior evaluation in a more favorable position than a smaller estimate of prevalence) and one we will use in our hypothetical scenario.

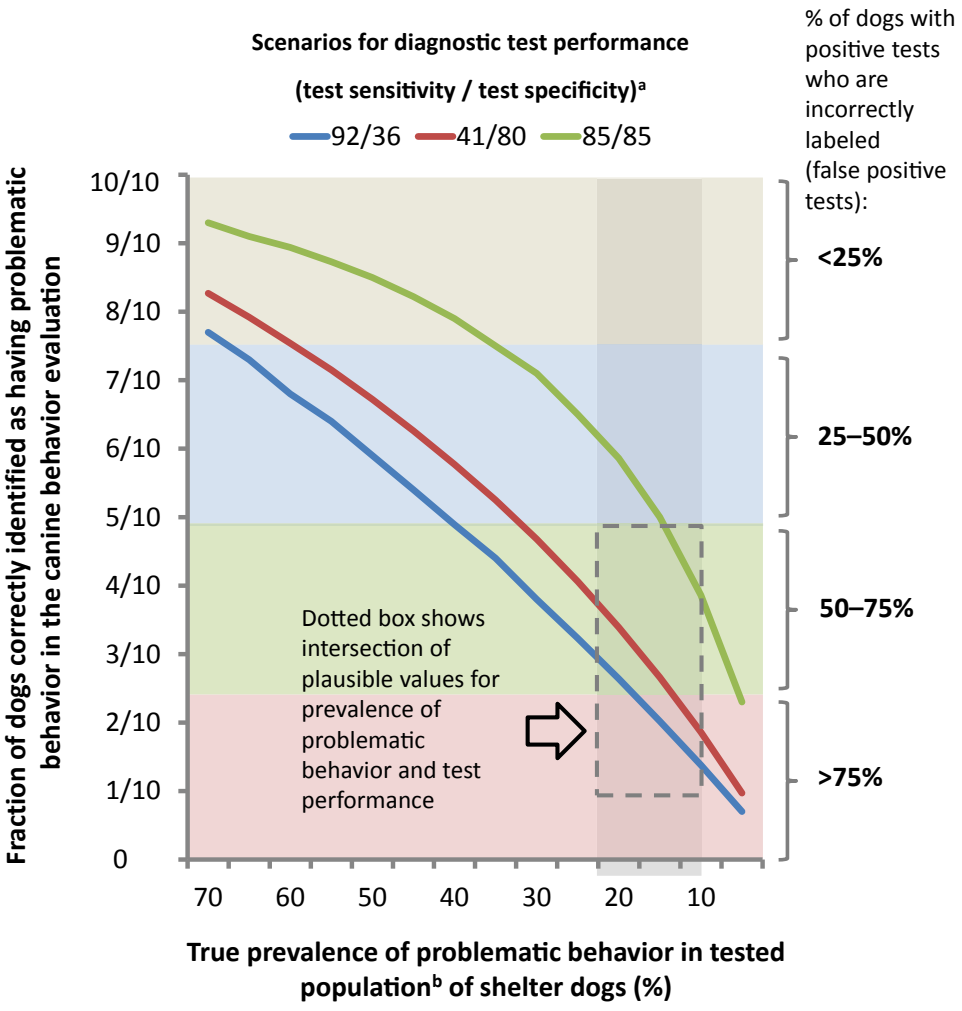

Figure 3. Relationship between prevalence of problematic behavior and predictive value of behavioral evaluation as diagnostic tests, using tests with different levels of sensitivity and specificity. [a]Values for sensitivity and specificity are from a meta-analysis of instruments for predicting human violent offending (92/36, respectively) and criminal offending (41/80, respectively) (Fazel et al., 2012), and a hypothetical very optimistic scenario with high sensitivity and high specificity (85/85, respectively). [b]Tested population excludes dogs screened out at intake. Shaded column indicates the most plausible range for prevalence of problematic behavior in tested shelter dogs. The intersection of plausible values for prevalence of problematic behavior and test performance occurs at point (dotted line box) where from half to >three quarters of dogs exhibiting problematic behavior in response to the provocative tests would be incorrectly labeled (i.e., false positives) because they would not show the problematic behavior in the future home.

Completing the simulation—the problem of false-positive results

Now that we have identified plausible starting values for test sensitivity and specificity, as well as prevalence of problematic behaviors related to aggression, we can complete the simulation. If a behavioral evaluation with a sensitivity of 92% and specificity of 36% (again, same values for predicting future violent offending in people in Table 1) was performed in a population of dogs with a baseline prevalence of problematic behavior of ~16%, the predictive value of a positive test would be at best ~22%, meaning that 78% of dogs testing positive would be false positives. Using the values for any criminal offending (41% and 80%, respectively) from the human risk assessment meta-analysis yields a positive predictive value of 28%, meaning that 72% of positive tests will be false positives. These scenarios are explained in Figure 2.

It might be argued that the validity of a canine behavior evaluation is likely to be much better than the measures of sensitivity and specificity characterizing human behavioral assessments for predicting future violence or criminal offending. This would be a weak argument in our opinion, given the number of studies and human subjects participating during 30 years of work summarized in the study by Fazel et al. (2012), but again, for purposes of the exercise, let's say that sensitivity and specificity in our canine scenario are a wildly optimistic 85% and 85%; which generally exceeds those reported from rigorously conducted human diagnostic tests. Having both values be high is unusual, as most diagnostic tests typically involve a trade-off between sensitivity and specificity. With the prevalence remaining at 16%, the predictive value of identifying warning and biting on the behavioral evaluation would be only 52% (also explained in Figure 2). "This means that even under unreasonably optimistic conditions favoring the performance of a behavior evaluation, the chance of a dog who tests positive actually being positive is about the same as flipping a coin." It is clear that making a euthanasia decision (or any other type of decision) on the basis of such test results would be nonsensical.

An endless number of simulations is possible here, and the 3 we have used in this article are presented in Figure 3, showing the effect on the predictive value of a positive test for various combinations of test sensitivity, specificity, and prevalence of problematic behavior; however, one might chose to define that latter term. As Figure 3 demonstrates, all of the 3 simulations are associated with a

very high proportion of false-positive results, not only at any plausible prevalence of problematic behavior, but also at implausible levels of behavior that was likely to be unusual or problematic in the home. Simple online calculators are available that allow anyone to perform their own simulations, and we would encourage readers to do so (e.g., see http://vassarstats.net/; http://vassarstats.net/clin2.html). It should also be noted that in all 3 of our hypothetical scenarios, the predictive value of a negative test is very high (>90%), meaning that a dog who tested negative in the shelter behavioral evaluation would likely continue to fail to show problematic behavior in the home. Again, this is unsurprising for a screening test with high sensitivity performed in a population with low prevalence of the condition.

Is there any other evidence to support our analysis? Planta & De Meester (2007) examined performance of a diagnostic test for dog aggression called the Socially Acceptable Behavior Test, which involved 16 individual subtests performed outdoors. Those authors defined aggressive biting during the test as including "aggressive bites or snaps and also aggressive attacks indicated by lunging on the leash which were prevented." The study was conducted in a sample of 330 privately owned dogs, recruited from breeders and behavior consultants. Interestingly, 94/330 (28.5%) of the dogs had actually previously bitten a person at least once. That sample, unlike shelter samples, provided the availability of a reference standard, as they obtained information about the dog's behavior in the home over at least 1 year after the test. We agree it is appropriate to call this a reference standard because the owners were familiar with the dogs, the dogs had been in the familiar environment where presumably normal behavior would be as stable as behavior can be, and the owners had the opportunity to observe the dogs regularly. The prevalence of snapping, biting, or lunging in an attempt to bite an artificial hand or a doll during the test in that sample was high (37.3%), presumably in part due to animal behavioral consultants' clients

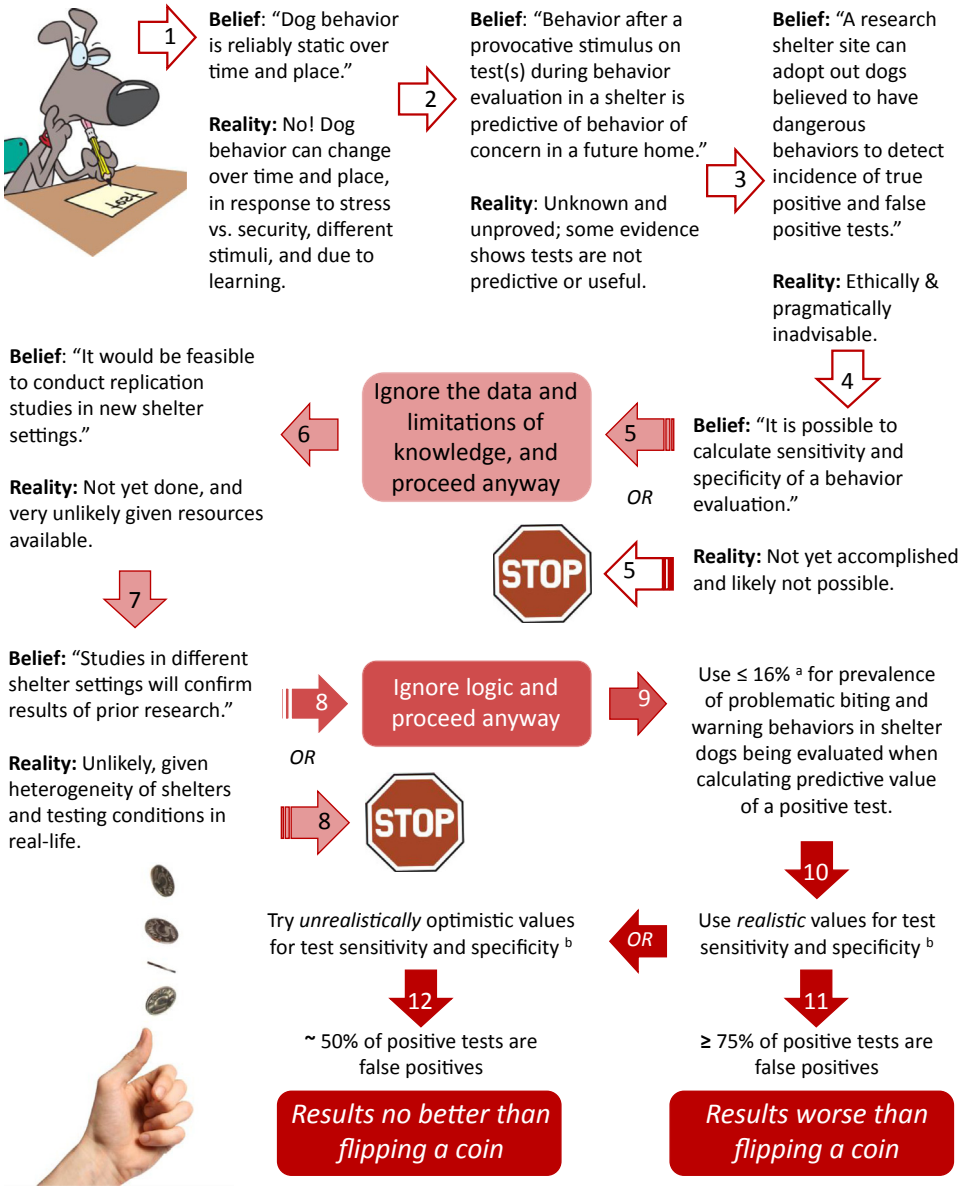

Figure 4. Summary of problems with canine behavior evaluations in animal shelters. [a]See text for details on how value was derived. [b]Assuming a validated test was available for shelters; see Figure 3 for values.

being one of the 2 sources of the sample. Nevertheless, when they used "no aggressive behavior at all" (a very strict interpretation) as the standard on both the test and in the home, there was high sensitivity (84%) and high specificity (81%), but the positive predictive value of the test was 64.2%, again only somewhat better than chance. When only 1 incident of aggressive behavior in a subtest was used as the cutoff for defining a positive test, a slightly more generous value, the sensitivity was 67%, the specificity was 95%, and the positive predictive value was 83%. A logical question that comes up in this scenario where so many dogs had a previous history of biting people is what exactly would be the clinical relevance of a true positive test—all of these dogs were pets living in homes, despite the identification of "aggressive" behavior on the Socially Acceptable Behavior (SAB) test.

We have found only three studies (other than the food guarding follow-up studies mentioned previously) that have attempted to compare the test results of dogs in a shelter to their behavior after adoption. Van der Borg et al. (1991) published an article reporting sensitivity and negative predictive values for a canine behavioral evaluation in shelter dogs, using the reference standard of owner-reported behavior in the new home. However, using the data as presented in their article (their Table 3) and the formulas as they described them to generate sensitivity (82.1%) and negative predictive value (84.8%) in their Table 3, the corresponding predictive value of a positive test would be only 61%, for a false-positive rate of 39%. Alternatively, using the descriptors present in their Table 3, we calculated a false-positive rate of 59%—either way, the results suggest that a positive test was not very useful. The second of the follow-up studies (Valsecchi et al., 2011) did not ask owners about the dog's real-life behavior but rather sent an evaluator out to retest the dog after rehoming; whether the results of the in-home responses to the test stimuli matched those of the shelter test seems irrelevant to the primary issue of whether the dog was considered a good pet or whether his behavior posed a danger in the home or community, which was not evaluated in the study. The third, Christensen et al. (2007), could record only false negatives, as all the positives, both true and false, had been euthanized rather than adopted. This illustrates one of the problems we raised earlier for calculating sensitivity and specificity. The rate of lunging, growling, snarling, snapping, or biting within 13 months of adoption among the dogs who passed the test was 40.9%, which is exactly the rate that Guy et al. (2001) found in a general population of dogs brought to veterinary clinics. It is interesting to note that none of these prospective studies included any report of an injurious bite, although it is unclear whether all of them asked about this.

Our evaluation of behavior evaluations

In summary, for any plausible combination of sensitivity, specificity, and prevalence, a positive test (i.e., eliciting problematic behavior from the dog in the behavior evaluation) is not much better than flipping a coin, and often much worse because, for reasons already explained, many of the dogs who test positive will be false positives (see Figure 4 for a summary of our argument). These results are in line with the conclusions of Fazel et al. (2012), who after an extensive meta-analysis of 73 studies involving risk assessments of 24,827 people concluded "…even after 30 years of development, the view that violence, sexual, or criminal risk [in people] can be predicted in most cases is not evidence-based. This message is important for the general public, media, and some administrations who may have unrealistic expectations of risk prediction for clinicians."

We believe that the mathematical improbability of behavioral evaluations done in a shelter providing reliable information about warning and biting behavior in a future home is sufficient to settle the question regarding the merit of these evaluations. However, beyond the mathematical improbability of success, there are additional significant pragmatic methodological concerns that would need to be addressed before an evaluation could be validated. Even if we believe we have standardized the stimuli used in a provocative test in the shelter (something we suspect is highly unlikely in the real-world diversity of animal shelters), it will never be possible to fully account for the emotions associated with a dog being abandoned and losing the familiarity, safety, and security of her home. It is our opinion that this is a fundamentally insurmountable limitation as well. Dogs in shelters may well act in a self-defensive way because of the fear and uncertainty associated with that environment. Behaviors such as growling, snarling, snapping, and biting are highly situation dependent and can be elicited by an almost infinite possible number of stimuli. Moreover, these behaviors, like all behaviors, are subject to learning, and a cognitively complex being like a dog is constantly processing new information that affects which stimuli he perceives as safe or dangerous. It seems overly optimistic to believe that the behavior of a dog during individual provocative tests in a behavior evaluation, whether due to exacerbation or suppression of normal tendencies, would consistently and reliably predict what would occur in an entirely different setting and how it might change over time.

The take-home message is that shelters have almost certainly, despite the best of intentions, placed undue faith in a diagnostic process that was not fully understood and that has not been scientifically established as being valid or suitably predictive for positive findings to make critical decisions about dogs. Our analysis shows that substituting plausible values for those unknown attributes would mean that a behavior evaluation conducted in the typical population of shelter dogs would result in a high proportion of dogs who test positive for warning or biting behavior being incorrectly labeled for future behavior, that is, be false positives, and potentially denied the opportunity for adoption. Substituting very optimistic instead of plausible values for sensitivity and specificity improves this picture but still produces results that are no better than flipping a coin. The explanation for this is simple: the likely prevalence of seriously problematic behavior (meaning a degree and/or frequency of biting or warning behavior in response to stimuli that would make them unsafe in a community setting) in shelter dogs is in general simply too low (particularly when overtly dangerous dogs are removed from the adoption pool at intake or before formal behavior evaluation) to render the results of a positive test much more informative than chance. In the case of a positive test, it is much more likely that the test has failed the dogs, rather than the dogs failing the test.

If not behavioral evaluations, then what?

We are not suggesting that shelters should abandon efforts to make reasonable attempts to place only behaviorally sound dogs in the community, and we recognize that there are cases where the correct decision about adoption is not straightforward. Nevertheless, given our analysis, the solution is not to settle for a false sense of security or take emotional comfort by relying on a flawed diagnostic, which in a low prevalence situation and in the real-life environment of an animal shelter is likely to unfairly label and potentially condemn dogs who exhibit behaviors deemed problematic during the test. Perhaps with a different balance between enthusiasm and critical thinking, we would have recognized long ago how unlikely it was that it would be possible to accurately

predict future behavior in a new and unknown environment with a test conducted in subjects whose behavior is likely influenced by the emotions associated with abandonment, stress, and fear in an unfamiliar environment. A mathematical simulation now confirms this.

So the question becomes, what exactly is necessary and responsible in a shelter, and how should scarce resources be spent? Nothing in the prevalence estimates we reviewed suggest that overall, dogs who come to spend time in a shelter (and are not screened out based on history or behavior at intake or shortly thereafter) are dramatically more or less inclined toward problematic warning or biting behavior than are pet dogs in general. Consider the implications of the remarkable reductions in euthanasia in communities across the country who have chosen to focus their efforts on saving lives, for example, an open-admission shelter in Hillsborough County, FL, which reduced its euthanasia rate from 49% in 2010 (http://www.hillsboroughcounty.org/DocumentCenter/View/12696) to 13% in 2015 (http://www.hillsboroughcounty.org/DocumentCenter/View/18075). In that shelter, the number of dogs euthanized for all reasons (including health) was actually lower (13%) than the prevalence of problematic behavior we used in our hypothetical scenario (16%). Shelters across the United States have engaged in concerted efforts to remove barriers to adoption and decrease the number of dogs euthanized, and overall, there is no indication this has compromised public safety. It is also highly unlikely that the improvement in adoption rate has come about because of a marked improvement in the behavior profile of admitted dogs. That leaves us with the conclusion that many stated problematic behaviors during behavior evaluations may not be so problematic after all in the future home.

The simplest solution may well be the most reasonable: to collect behavior histories on dogs at the time of relinquishment whenever possible and attempt to verify any serious incidents reported and to designate as ineligible for adoption dogs who inflict injurious bites or are too threatening to handle in the shelter. Subsequent provocative behavioral testing of the general population in shelters seems predicated on an assumption of risk from dogs that is far in excess of the data and on assumptions about canine behavior that may not be supportable. Instead of striving to bring out the worst in dogs in the stressful and temporary environment of a shelter, and devoting scarce resources to inherently flawed and unvalidated formal evaluations, how much more productive might it be to focus our energies on giving every dog the opportunity to be at his or her best? It may be far better for dogs, shelters, and communities if effort was spent regularly interacting with every shelter dog in normal and even enjoyable ways involving activities in which they will be expected to engage (e.g., walking, socializing with people, playgroups with other dogs, games, and training) to enrich their experience and minimize the adverse effect of being relinquished and confined to an unfamiliar environment, rather than investing additional resources in what is likely a losing proposition for all concerned. With proper training of staff about normal dog behavior, those activities are likelier to identify any additional dogs whose behavior may be of concern, to be more indicative of the typical personality and behavior of dogs, and help make dogs better candidates for adoption in the process.

Acknowledgments

The authors thank Elizabeth Arps of the National Canine Research Council for help with proofreading and formatting this article and Dr. Amy Marder for providing helpful comments.

There was no funding source beyond the relationships noted in the conflict of interest statement mentioned previously.

The idea for this article was conceived by both authors and both authors contributed to the writing and review, and approve this submission.

Ethical considerations

Not required.

Conflict of interest

Gary J. Patronek is a paid consultant to the National Canine Research Council, a subsidiary of Animal Farm Foundation. Janis Bradley is an employee of the National Canine Research Council.

References

Akobeng, A.K., 2007. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr. 96, 338–341.

Arealis, G., Galanopoulos, I., Nikolaou, V.S., Lacon, A., Ashwood, N., Kitsis, C., 2014. Does the CT improve inter- and intra-observer agreement for the AO, Fernandez and Universal classification systems for distal radius fractures? Injury 45, 1579–1584.

Banks, E., Reeves, G., Beral, V., Bull, D., Crossley, B., Simmonds, M., Hilton, E., Bailey, S., Barrett, N., Briers, P., English, R., Jackson, A., Kutt, E., Lavelle, J., Rockall, L., Wallis, M.G., Wilson, M., Patnick, J., 2004. Influence of personal characteristics of individual women on sensitivity and specificity of mammography in the Million Women Study: cohort study. BMJ 329, 477.

Bennett, C.M., Guo, M., Dharmage, S.C., 2007. HbA(1c) as a screening tool for detection of Type 2 diabetes: a systematic review. Diabet. Med. 24, 333–343.

Bobo, J.K., Lee, N.C., Thames, S.F., 2000. Findings from 752,081 clinical breast examinations reported to a national screening program from 1995 through 1998. J. Natl. Cancer Inst. 92, 971–976.

Bollen, K.S., Horowitz, J., 2008. Behavioral evaluation and demographic information in the assessment of aggressiveness in shelter dogs. Appl. Anim. Behav. Sci. 112, 120–135.

Buchsbaum, D.G., Buchanan, R.G., Centor, R.M., Schnoll, S.H., Lawton, M.J., 1991. Screening for alcohol abuse using CAGE scores and likelihood ratios. Ann. Intern. Med. 115, 774–777.

Casey, R.A., Loftus, B., Bolster, C., Richards, G.J., Blackwell, E.J., 2014. Human-directed aggression in domestic dogs (Canis familiaris): occurrence in different contexts and risk factors. Appl. Anim. Behav. Sci. 152, 52–63.

Christensen, E., Scarlett, J., Campagna, M., Houpt, K.A., 2007. Aggressive behavior in adopted dogs that passed a temperament test. Appl. Anim. Behav. Sci. 106, 85–95.

Coe, J.B., Young, I., Lambert, K., Dysart, L., Nogueira Borden, L., Rajić, A., 2014. A scoping review of published research on the relinquishment of companion animals. J. Appl. Anim. Welf. Sci. 17, 253–273.

Collins, J.F., Lieberman, D.A., Durbin, T.E., Weiss, D.G., Veterans Affairs Cooperative Study #380 Group, 2005. Accuracy of screening for fecal occult blood on a single stool sample obtained by digital rectal examination: a comparison with recommended sampling practice. Ann. Intern. Med. 142, 81–85.

D'Arpino, S., Dowling-Guyer, S., Shabelansky, A., Marder, A.R., Patronek, G.J., 2012. The use and perception of canine behavioral assessments in sheltering organizations. In: Proceedings of the American College of Veterinary Behaviorists/American Veterinary Society of Animal Behavior Veterinary Behavior Symposium, San Diego, CA, pp. 27–30.

Devadas, A.A., Razulis, M.H., Shulman, K.S., 2013. Maryland Department of Legislative Services. Dog Bites in Maryland and Other States: Data, Insurance Coverage, and Liability. Available at: http://dlslibrary.state.md.us/publications/OPA/I/DogBite_2013.pdf. Accessed May 12, 2016.

Diederich, C., Giffroy, J.M., 2006. Behavioural testing in dogs: a review of methodology in search for standardization. Appl. Anim. Behav. Sci. 97, 51–72.

Diesel, G., Brodbelt, D., Pfeiffer, D.U., 2008. Reliability of assessment of dogs' behavioural responses by staff working at a welfare charity in the UK. Appl. Anim. Behav. Sci. 115, 171–181.

Fazel, S., Singh, J.P., Doll, H., Grann, M., 2012. Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. BMJ 345, e4692.

Frances, A., 2014. Saving Normal: An Insider's Revolt against Out-of-Control Psychiatric Diagnosis, DSM-5, Big Pharma, and the Medicalization of Ordinary Life. William Morrow Paperbacks, New York, NY.

Gilchrist, J., Sacks, J.J., White, D., Kresnow, M.J., 2008. Dog bites: still a problem? Inj. Prev. 14, 296–301.

Gurol, Y., Akan, H., Izbirak, G., Tekkanat, Z.T., Gunduz, T.S., Hayran, O., Yilmaz, G., 2010. The sensitivity and the specificity of rapid antigen test in streptococcal upper respiratory tract infections. Int. J. Pediatr. Otorhinolaryngol. 74, 591–593.

Guy, N.C., Luescher, U.A., Dohoo, S.E., Spangler, E., Miller, J.B., Dohoo, I.R., Bate, L.A., 2001. Demographic and aggressive characteristics of dogs in a general veterinary caseload. Appl. Anim. Behav. Sci. 74, 15–28.

Haverbeke, A., Pluijmakers, J., Diederich, C., 2015. Behavioral evaluations of shelter dogs: literature review, perspectives, and follow-up within the European member states's legislation with emphasis on the Belgian situation. J. Vet. Behav.: Clin. Appl. Res. 10, 5–11.
Jackson, J.L., O'Malley, P.G., Kroenke, K., 2003. Evaluation of acute knee pain in primary care. Ann. Intern. Med. 139, 575–588.
Khan, L., Mitera, G., Probyn, L., Ford, M., Christakis, M., Finkelstein, J., Donovan, A., Zhang, L., Zeng, L., Rubenstein, J., Yee, A., Holden, L., Chow, E., 2011. Inter-rater reliability between musculoskeletal radiologists and orthopedic surgeons on computed tomography imaging features of spinal metastases. Curr. Oncol. 18, e282–e287.
Kikuchi, M., Hogue, T., Mills, D.S., 2014. Definition and management of human directed aggressive behavior of dogs in the UK and Japan. J. Vet. Behav. Clin.: Appl. Res. 9 (6), e9.
King, T., Marston, L.C., Bennett, P.C., 2012. Breeding dogs for beauty and behaviour: why scientists need to do more to develop valid and reliable behaviour assessments for dogs kept as companions. Appl. Anim. Behav. Sci. 137, 1–12.
Lutz, B.L., 2009. Liability 'hysteria.' Don't Let Liability 'hysteria' Keep You From Sending Good Dogs Home. Available at: http://animalfarmfoundation.org/files/Liability_Hysteria_Viewpoint_7.pdf. Accessed May 12, 2016.
Marder, A.R., Shabelansky, A., Patronek, G.J., Dowling-Guyer, S., Segurson D'Arpino, S., 2013. Food-related aggression in shelter dogs: a comparison of behavior identified by a behavior evaluation in the shelter and owner reports after adoption. Appl. Anim. Behav. Sci. 148, 150–156.
Matsunaga, F.T., Tamaoki, M.J., Cordeiro, E.F., Uehara, A., Ikawa, M.H., Matsumoto, M.H., dos Santos, J.B., Belloti, J.C., 2009. Are classifications of proximal radius fractures reproducible? BMC Musculoskelet. Disord. 10, 120.
Medeiros, L.R., Duarte, C.S., Rosa, D.D., Edelweiss, M.I., Edelweiss, M., Silva, F.R., Winnnikow, E.P., Simões Pires, P.D., Rosa, M.I., 2011. Accuracy of magnetic resonance in suspicious breast lesions: a systematic quantitative review and meta-analysis. Breast Cancer Res. Treat 26, 273–285.
Messing, J.T., Campbell, J., Sullivan Wilson, J., Brown, S., Patchell, B., 2015. The lethality screen: the predictive validity of an intimate partner violence risk assessment for use by first responders. J. Interpers. Violence. [Epub ahead of print].
Mohan-Gibbons, H., Weiss, E., Slater, M., 2012. Preliminary investigation of food guarding behavior in shelter dogs in the United States. Animals (Basel) 2, 331–333.
Mornement, K.M., Coleman, G.J., Toukhsati, S., Bennett, P.C., 2010. A review of behavioral assessment protocols used by Australian animal shelters to determine the adoption suitability of dogs. J. Appl. Anim. Welf. Sci. 13, 314–329.
Overall, K.L., 2015. The mismeasure of behavior: identifying tests meaningful to the species studied. J. Vet. Behav.: Clin. Appl. Res. 10, 1–4.
Patronek, G.J., Slavinski, S.A., 2009. Animal bites. J. Am. Vet. Med. Assoc. 234, 336–345.
Planta, J.U.D., De Meester, R.H.W.M., 2007. Validity of the Socially Acceptable Behavior (SAB) test as a measure of aggression in dogs towards non-familiar humans. Vlaams Diergen. Tijds. 76, 359–368.
Pocklington, C., Gilbody, S., Manea, L., McMillan, D., 2016. The diagnostic accuracy of brief versions of the Geriatric Depression Scale: a systematic review and meta-analysis. Int. J. Geriatr. Psychiatry 31, 837–857.
Rayment, D.J., De Groef, B., Peters, R.A., Marston, L.C., 2015. Applied personality assessment in domestic dogs: limitations and caveats. Appl. Anim. Behav. Sci. 163, 1–18.
Sacks, J.J., Kresnow, M., Houston, B., 1996. Dog bites: how big a problem? Inj. Prev. 2, 52–54.
Salman, M.D., Hutchison, H., Ruch-Gallie, R., Kogan, L., New, J.C., Kass, P.H., Scarlett, J.M., 2000. Behavioral reasons for relinquishment of dogs and cats to 12 shelters. J. Appl. Anim. Welf. Sci. 3, 93–106.
Schuetz, G.M., Zacharopoulou, N.M., Schlattmann, P., Dewey, M., 2010. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. Ann. Intern. Med. 152, 167–177.
Sheppard, G., Mills, D.S., 2003. Construct models in veterinary behavioural medicine: lessons from the human experience. Vet. Res. Commun. 27, 175–191.
Stamm, W.E., Counts, G.W., Running, K.R., Fihn, S., Turck, M., Holmes, K.K., 1982. Diagnosis of coliform infection in acutely dysuric women. N. Engl. J. Med. 307, 463–468.
Taylor, K.D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. J. Vet. Behav.: Clin. Appl. Res. 1, 94–108.
Terasawa, T., Blackmore, C.C., Bent, S., Kohlwes, R.J., 2004. Systematic review: computed tomography and ultrasonography to detect acute appendicitis in adults and adolescents. Ann. Intern. Med. 141, 537–546.
Valsecchi, P., Barnard, S., Stefanini, C., Normando, S., 2011. Temperament test for re-homed dogs validated through direct behavioral observation in shelter and home environment. J. Vet. Behav.: Clin. Appl. Res. 6, 161–177.
van der Borg, J.A.M., Beerda, B., Ooms, M., Silveira de Souza, A., van Hagen, M., Kemp, B., 2010. Evaluation of behaviour testing for human directed aggression in dogs. Appl. Anim. Behav. Sci. 128, 78–90.
van der Borg, J.A.M., Netto, W.J., Planta, D.J.U., 1991. Behavioural testing of dogs in animal shelters to predict problem behaviour. Appl. Anim. Behav. Sci. 32, 237–251.
Weiss, E., 2016. Breaking Up Is Hard to Do. ASPCA PRO, New York, New York. Available at: http://www.aspcapro.org/blog/2016/02/10/breaking-hard-do. Accessed August 29, 2016.
Weiss, E., Gramann, S., Spain, C.V., Slater, M., 2015. Goodbye to a good friend: an exploration of the re-homing of cats and dogs in the US. Open J. Anim. Sci. 5, 435–456. Available at: http://file.scirp.org/pdf/OJAS_2015100914300959.pdf. Accessed August 29, 2016.
Wise, J.K., Heathcott, B.L., Gonzalez, M.L., 2002. Results of the AVMA survey on companion animal ownership in US pet-owning households. American Veterinary Medical Association. J. Am. Vet. Med. Assoc. 221, 1572–1573.