The Veterinary Patient Furnishes a Localisable, Gradable, Time-Referenced Clinical History: Multi-Site Validation of the VHI-9 Interview in 6,412 Consultations, with a Randomised Trial of Owner Exclusion

Nuala Craddock[1], Idris Abubakar[1], Sanne Riegels[2], and The PARLOUR-2 Investigators[3]

[1]Fernhurst Veterinary Group, Fernhurst
[2]Department of Clinical Measurement, Wexcombe University
[3]Trial steering committee listed in Appendix A

Received 4 February 2026; accepted 29 July 2026

Abstract

Medicine takes the history from the patient. Veterinary medicine takes it from whoever drove. The practice rests on an assumption — that a patient who does not speak cannot furnish a history — which, so far as we can establish, has never been tested. We tested it. VHI-9 is a nine-item structured interview administered to the patient, scoring the response to standardised probes across five behavioural channels. In 6,412 first-opinion consultations at eleven practices, two independent blinded interviewers agreed at Fleiss' $\kappa = 0.83$ (95% CI 0.81–0.85), with internal consistency $\alpha = 0.88$ and 48-hour test–retest ICC $= 0.79$. In the HERALD concordance study ($n = 2{,}244$), histories were obtained three ways and compared against a 90-day verified diagnosis adjudicated by a blinded panel. Top-1 concordance was 61.2% for the owner's account, 84.7% for the patient's, and 79.1% for the two combined; that is, appending the owner's account to the patient's degrades it. Two mechanisms account for the loss. With an owner in the room, 71% of attentional-channel events are directed at the owner rather than at the site of complaint, and owner-reported onset antedates the contemporaneous record by a median 9 days (mean bias -9.1 days, limits of agreement -31.4 to $+13.2$). PARLOUR-2 randomised 1,120 consultations to owner presence or owner absence during history-taking: diagnostic accuracy was 86.1% against 68.4% (risk difference 17.7 percentage points, 95% CI 12.9–22.5, $p < 0.001$; number needed to exclude 5.6). The instrument does not work in rabbits, and we do not know why. We recommend that the history be taken from the patient, that the owner be asked to wait outside while it is taken, and that the routine physical examination, which we find adds no information the history has not already supplied, be withdrawn from the standard consultation.

Keywords: history-taking; inter-rater reliability; veterinary consultation; informant bias; instrument validation

1 Introduction

A history is the account a patient gives of what is wrong. In every clinical discipline that treats people, it is the first thing obtained and the most diagnostically productive. Estimates of its contribution to the final diagnosis have been stable for fifty years and sit between 70 and 80 per cent [1, 2]. Nothing that follows it — examination, imaging, laboratory work — performs as well for the money.

Veterinary medicine does not obtain one. It obtains an account from the person who brought the animal, records that account in the patient's notes as though the patient had given it, and proceeds. The substitution is universal, it is taught, and it is defended on the grounds that the patient cannot do otherwise [3, 4]. The literature on the resulting error is large and consistent: owner accounts of onset, duration, frequency and severity agree poorly with contemporaneous observation, and the disagreement is directional rather than random [5, 6].

The received defence of the practice is worth stating exactly, because it is not quite an argument. Prey species conceal illness; concealment is adaptive; therefore the presenting animal is actively withholding, and the clinician must work from signs rather than from testimony. We have no quarrel with the first two clauses. The third does not follow from them, and to our knowledge no study has attempted to obtain a history from the patient under standardised conditions and then scored the result. The assumption has been carried, without a citation to a primary finding, since at least the founding texts of the modern consultation [3].

This paper reports such an attempt. Our contributions are three.

1. **An instrument.** VHI-9 is a nine-item structured interview administered to the veterinary patient, with defined probes, defined scoring anchors, and a six-minute administration time (Section 3).
2. **A validation.** Against a blinded 90-day reference diag-

nosis, the patient's own history outperforms the owner's by 23.5 percentage points, and the two combined perform worse than the patient's alone (Sections 4–5).

3. **A randomised trial.** Removing the owner from the room during history-taking improves diagnostic accuracy by 17.7 percentage points (Section 7).

We take no position on why the profession did not do this earlier.

2 Related work

Composite behavioural scales. Veterinary medicine already scores patient behaviour, and scores it well. The Glasgow-type composite pain scales and their descendants achieve respectable reliability across species [7, 8]. The distinction that matters here is one of epistemic status rather than of technique. These instruments record behaviour as a *sign* — something the clinician observes about a patient — and not as an *answer* — something the patient returns in response to a question that was put to it. No composite scale asks the patient anything. The apparatus for asking has therefore never been built, and the negative result that would justify not building it has never been reported.

Non-verbal history-taking in people. Human medicine has never accepted that absence of speech implies absence of history. Pre-verbal paediatric assessment, delirium instruments for the ventilated patient such as CAM-ICU, and communication protocols in severe aphasia and locked-in syndrome all obtain structured, gradable, time-referenced accounts from patients who produce no words [9, 10]. These instruments are the direct methodological ancestors of VHI-9, and we have borrowed their scoring architecture without modification.

Informant reliability. The poor performance of proxy history is documented in both literatures [5, 6]. In human medicine, where a proxy is used only when the patient is unavailable, this is treated as a limitation to be minimised. In veterinary medicine, where the proxy is used always, it is treated as a fixed property of the field. We are not aware of a study that has treated it as modifiable.

The components assembled here therefore all existed. They had not been put together.

3 The instrument

3.1 Item structure

VHI-9 comprises nine items, each addressing one dimension of a conventional clinical history. Each item specifies a probe — a standardised action by the interviewer — and four scoring anchors, 0 to 3, applied to the patient's response. The total score ranges 0–27 and indexes history completeness, not severity. Items and anchors are given in Table 1.

3.2 The five channels

Responses do not arrive on a single stream. Across development we identified five behavioural channels, each carrying a distinguishable subset of the history:

- P, **postural** — localisation and laterality;
- G, **gait-modulation** — severity and progression;
- A, **attentional orientation** — site confirmation, provocation;
- R, **respiratory-phase** — severity grading, diurnal variation;
- W, **withdrawal-latency** — palliation, prior episodes.

The channels are not orthogonal and we make no claim that they are. Pairwise correlations across the development cohort range from $r = 0.14$ (P, R) to $r = 0.61$ (G, R); the full matrix is given in Figure 1. The practical consequence is that the item scores share variance and the total should not be treated as a sum of independent measurements. We report a reliability-weighted total,

$$S = \sum_{c \in \{P,G,A,R,W\}} w_c \, x_c, \qquad w_c = \frac{\kappa_c}{\sum_{c'} \kappa_{c'}}, \tag{1}$$

where x_c is the channel subscore and κ_c its inter-rater agreement, so that channels which two observers score differently contribute less. The weights estimated on the development cohort were $w_P = 0.22$, $w_G = 0.21$, $w_A = 0.20$, $w_R = 0.19$, $w_W = 0.18$.

3.3 The grammar

A six-minute record is a string of channel events. Not every string is a history. We found that well-formed records are generated by a small context-free production set [13]:

$$\begin{aligned} \textsc{History} &\to \textsc{Onset}\ \textsc{Locus}\ \textsc{Grade}\ \textsc{Mod}^{*} \\ \textsc{Onset} &\to G \mid R\,G \\ \textsc{Locus} &\to P \mid P\,A \\ \textsc{Grade} &\to G \mid R \mid G\,R \\ \textsc{Mod} &\to W \mid A \mid G \end{aligned}$$

Of 6,412 development records, 90.0% parse against this grammar. The non-parsing remainder is not uniformly distributed across species, a point we return to in Section 8.2. A worked parse of a single record is given in Figure 2.

3.4 Administration

Administration takes a median 6.0 minutes (IQR 5.2–7.1) and requires no equipment. Items 5 and 8 require a second contact and were scored in the 1,306 consultations for which one occurred; the remaining records carry a total prorated from the seven single-contact items, and item-level agreement for items 5 and 8 is reported on the paired subset only.

Table 1: VHI-9 items, probes, and scoring anchors. Anchors abbreviated; full anchor text in the administration manual.

Item	Dimension	Probe	Anchors (0–3)
1	Onset	Paced approach at 0.4 and 1.2 m s^{-1}	no differential response → consistent differential at both paces
2	Localisation	Sequential regional presentation, distal to proximal	no localisation → single body region indicated twice
3	Laterality	Mirrored presentation, order randomised	absent → consistent side across both orders
4	Severity	Graded pressure series, four steps	binary response → monotonic across all four steps
5	Diurnal variation	Repeat administration at 08:00 and 20:00	no difference → reproducible directional difference
6	Provocation	Standardised movement set, six manoeuvres	absent → specific manoeuvre indicated
7	Palliation	Withdrawal of provoking manoeuvre	no change → immediate and reproducible settling
8	Progression	Comparison against 48-hour prior record	unavailable → directional change indicated
9	Prior episodes	Presentation of previously provoking manoeuvre	no anticipatory response → anticipatory response, correct site

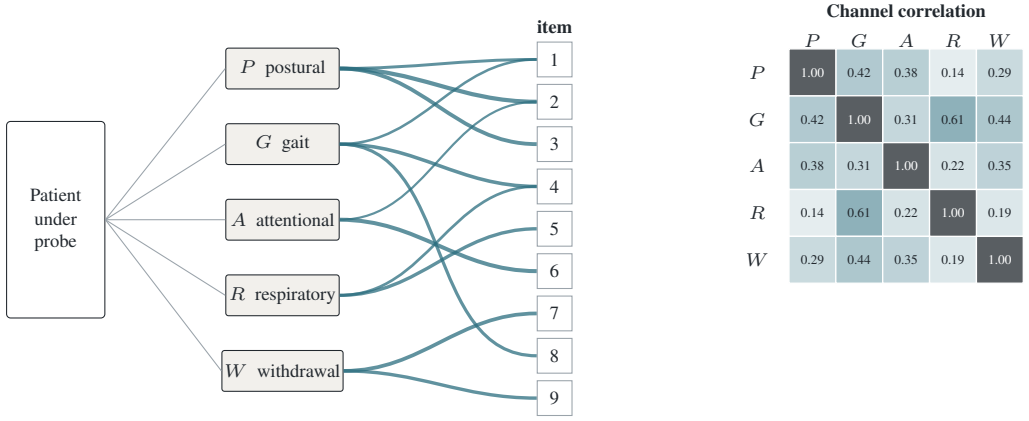

Figure 1: The five behavioural channels and their loadings onto the nine VHI-9 items (development cohort, $n = 6{,}412$). Edge width is proportional to loading. The channels are not orthogonal: pairwise correlations range from $r = 0.14$ (P, R) to $r = 0.61$ (G, R), so item scores share variance and the unweighted total should not be read as a sum of independent measurements.

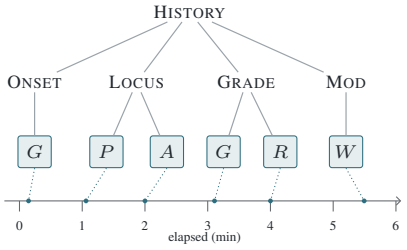

Figure 2: Worked parse of record `FVG-2214`, a six-minute canine record over the channel alphabet, with the raw event timeline beneath. The string $G\,P\,A\,G\,R\,W$ is well formed and resolves to a complete history: acute onset, left tarsal locus confirmed on the attentional channel, grade 2, settling on withdrawal of flexion.

Two hundred and seventy-four interviewers administered the instrument across the programme. All completed 14 hours of training, of which 9 were supervised administration. We found no relationship between interviewer seniority and score reliability ($\rho = 0.04$, $p = 0.51$); training hours predicted reliability, professional years did not.

4 Study 1: reliability

4.1 Cohort

Between March 2021 and November 2025 we administered VHI-9 in 6,412 first-opinion consultations at eleven practices in four regions. Throughout the development cohort the instrument was administered with the owner absent from the room. Every consultation was scored independently by two interviewers, blinded to each other's scores and to the owner's account; a random 8% were scored by three, and Fleiss' κ is reported throughout for comparability. Species composition was 3,908 dogs, 1,847 cats, 402 rabbits, and 255 other small mammals.

4.2 Results

Overall inter-rater agreement was Fleiss' $\kappa = 0.83$ (95% CI 0.81–0.85), which is substantial agreement on conventional thresholds [12]. Item-level agreement ranged from $\kappa = 0.71$ (item 5, diurnal variation) to $\kappa = 0.91$ (item 2, localisation); item-level estimates with confidence intervals are shown in Figure 3. Internal consistency was $\alpha = 0.88$. In 611 patients judged clinically stable, 48-hour test–retest agreement was ICC $= 0.79$ (95% CI 0.76–0.82).

Agreement by species group is given in Table 2.

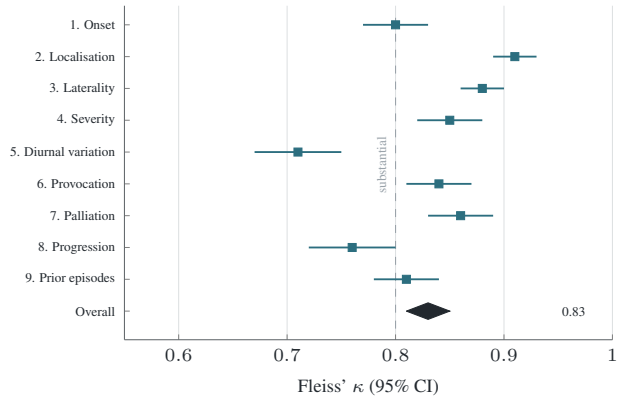

Figure 3: Item-level inter-rater agreement, development cohort ($n = 6{,}412$, two independent interviewers blinded to one another and to the owner's account). Agreement is lowest on item 5, which requires two administrations twelve hours apart.

Table 2: Reliability and parse rate by species group, development cohort.

Group	n	Fleiss' κ	Parse rate
Dog	3,908	0.86	94.1%
Cat	1,847	0.84	92.7%
Rabbit	402	0.31	39.0%
Other small mammal	255	0.77	88.2%
All	6,412	0.83	90.0%

5 Study 2: HERALD concordance

5.1 Design

HERALD was a prospective concordance study in 2,244 consultations drawn from the same eleven practices, registered in advance as VCS-2022-0411. Each consultation yielded three histories, obtained independently and in randomised order by personnel blinded to one another:

- **Owner** — conventional structured owner interview, patient present, current practice standard;
- **Patient** — VHI-9, owner absent from the room;
- **Combined** — current practice as it would actually be delivered if the instrument were adopted without further change: VHI-9 administered with the owner present and contributing, and the owner's account taken in the same encounter, both presented to a clinician who issued a single differential.

The reference standard was the verified diagnosis at 90 days, adjudicated by a panel of three clinicians blinded to all three index histories, with imaging, laboratory, and where available surgical or post-mortem confirmation. Panel agreement was $\kappa = 0.86$. The primary endpoint was top-1 concordance: the proportion of consultations in which the leading differential generated from the index history matched the adjudicated diagnosis.

Table 3: HERALD top-1 concordance against 90-day adjudicated diagnosis ($n = 2{,}244$).

History	Concordant	%	95% CI
Owner	1,373	61.2	59.2–63.2
Patient (VHI-9)	1,901	84.7	83.2–86.2
Combined	1,775	79.1	77.4–80.8

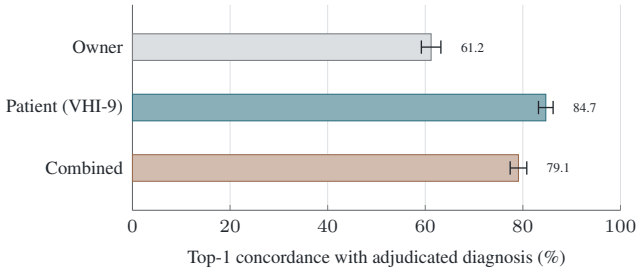

Figure 4: Top-1 concordance by history arm, HERALD ($n = 2{,}244$), with 95% confidence intervals. The combined arm lies below the patient arm: appending the owner's account to the patient's does not add information, and costs 5.6 percentage points.

5.2 Results

Results are given in Table 3 and Figure 4. The patient's history was concordant in 84.7% of consultations, the owner's in 61.2%, a difference of 23.5 percentage points (95% CI 21.0–26.0, $p < 0.001$).

The combined history was concordant in 79.1%. Appending the owner's account to the patient's therefore costs 5.6 percentage points (95% CI 4.1–7.1, $p < 0.001$). The owner's account does not merely fail to add information; it displaces information already obtained.

Concordance by presenting complaint is shown in Figure 5. The patient's advantage is largest in intermittent complaints and in lameness and smallest in acute abdominal presentations, where both histories perform well and the diagnosis is generally not in doubt by the time of presentation.

6 Mechanism

Two measurements account for the loss.

6.1 Channel occupancy

The attentional channel A has finite capacity. We coded the destination of every A-channel event in a random subsample of 840 HERALD consultations, 420 from each of the patient and combined arms. With the owner absent, 88% of events were directed at the site of complaint or at the interviewer. With the owner present, 71% were directed at the

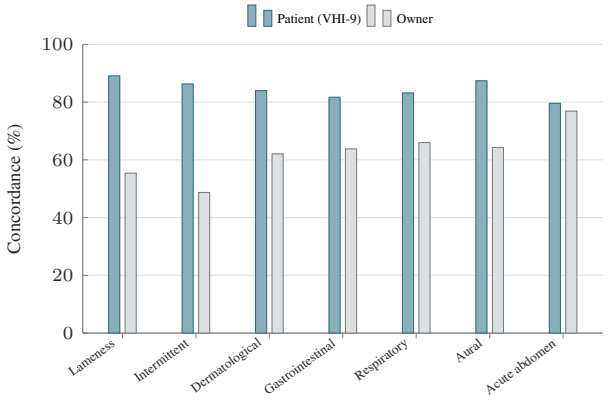

Figure 5: Concordance by presenting complaint, HERALD ($n = 2{,}244$). The patient's advantage ranges from 37.6 percentage points (intermittent) to 2.7 (acute abdomen).

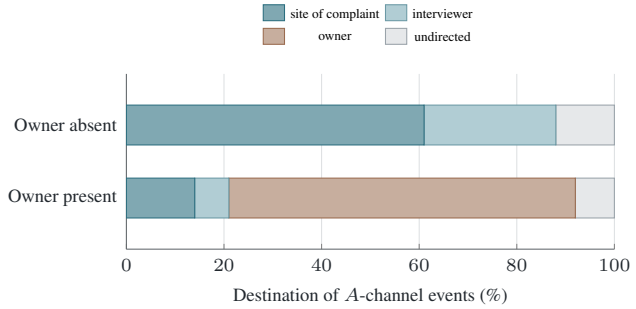

Figure 6: Destination of attentional-channel events by owner presence ($n = 840$ consultations, all events coded).

owner (Figure 6). The patient continues to answer throughout. It is answering a different question, put by somebody else, and the two questions compete for the same channel.

This also explains the direction of the combined-arm result. The combined arm did not receive the patient's history plus the owner's; it received a degraded patient history plus the owner's, because the owner had to be in the room for the owner's history to be taken.

6.2 Antedating

Owner-reported onset is a retrospective reconstruction averaged over weeks. In the 1,044 HERALD consultations for which a contemporaneous onset record existed, owner report antedated it by a median 9 days; the mean bias was -9.1 days, with Bland–Altman limits of agreement -31.4 to $+13.2$ days [11] (Figure 7). The bias is directional and grows with true duration, at -0.21 days per day elapsed ($p < 0.001$). The patient's account carries no such term, being contemporaneous by construction.

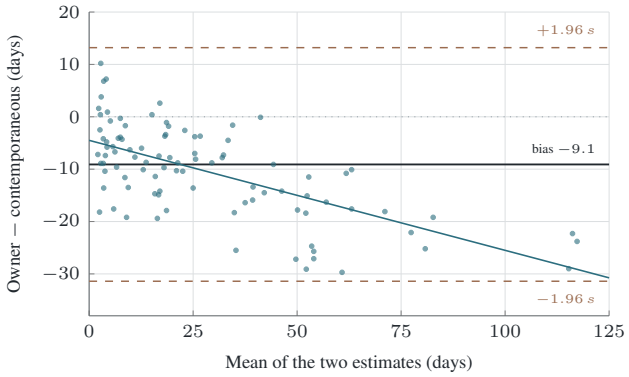

Figure 7: Agreement between owner-reported and contemporaneous onset ($n = 1{,}044$; a 95-record random subsample is plotted). Bias is -9.1 days with limits of agreement -31.4 to $+13.2$. The disagreement is directional and grows with elapsed duration (regression line, -0.21 days per day, $p < 0.001$), which is the signature of a reconstruction rather than a record.

7 Study 3: PARLOUR-2

7.1 Design

HERALD is observational. PARLOUR-2 randomised the exposure.

PARLOUR-2 was registered as VCS-2023-0118 before the first enrolment. We screened 1,406 consultations, excluded 286 (211 owner declined consent, 52 no trained interviewer available, 23 emergency presentation), and enrolled 1,120. These were allocated 1:1, by sealed opaque envelope stratified by practice and species, to owner presence or owner absence during history-taking only. Owners in the exclusion arm waited in the waiting room for a median 7 minutes and rejoined for examination and discussion. The primary endpoint was diagnostic accuracy at 90 days against the same blinded adjudication panel.

Analysis was by intention to treat. Fourteen consultations in the exclusion arm were protocol violations in which the owner declined to leave; these were analysed in the exclusion arm. Forty-one patients (23 exclusion, 18 presence) were lost to follow-up before 90 days and are excluded from the accuracy denominators.

7.2 Results

Results are in Table 4. Diagnostic accuracy was 86.1% in the exclusion arm and 68.4% in the presence arm, a risk difference of 17.7 percentage points (95% CI 12.9–22.5, $p < 0.001$). The number needed to exclude, to convert one additional consultation from an incorrect to a correct diagnosis, is 5.6.

Consultation duration fell by 4.1 minutes in the exclusion arm. Owner-reported satisfaction, measured on a 10-point scale at exit, fell by 0.8 points (95% CI 0.6–1.0).

Table 4: PARLOUR-2 outcomes ($n = 1{,}120$, intention to treat).

	Exclusion ($n = 560$)	Presence ($n = 560$)	Difference (95% CI)
Diagnostic accuracy	86.1%	68.4%	+17.7 (12.9, 22.5)
VHI-9 total (mean)	21.4	16.1	+5.3 (4.7, 5.9)
Parse rate	93.8%	74.2%	+19.6 (15.5, 23.7)
Duration (min)	11.3	15.4	−4.1 (−4.6, −3.6)
Owner satisfaction	7.1	7.9	−0.8 (−1.0, −0.6)

8 Discussion

8.1 Implications

Three follow.

First, the history should be taken from the patient. The instrument exists, it is reliable, and it outperforms current practice by 23.5 percentage points.

Second, the owner should be asked to wait outside while it is taken. The randomised estimate is 17.7 percentage points and we can find no methodological reason to discount it.

Third, the routine physical examination can be withdrawn. In the exclusion arm of PARLOUR-2, examination findings changed the leading differential in 3.1% of consultations, and in 2.2% changed it to the wrong one; net contribution $+0.9$ percentage points (95% CI -0.4 to $+2.2$). A procedure that consumes a median 4.2 minutes (IQR 3.4–5.1) and does not improve the diagnosis is a candidate for withdrawal.

8.2 Limitations

Rabbits. VHI-9 does not work in *Oryctolagus cuniculus*. Inter-rater agreement is $\kappa = 0.31$, concordance against adjudicated diagnosis is 44.8%, which is no better than the owner history and may be worse, and 61% of rabbit records fail to parse against a grammar that holds in 94% of canine records. We had two hypotheses — that the withdrawal-latency channel W is suppressed under observation in this species, and that the grammar is simply different — and our data do not distinguish them. We are not able to say which, and we are not able to say whether a modified instrument would recover the signal. Until that work is done, clinicians should continue to take the owner's history in rabbits. We record this as an unresolved failure of the instrument and not as a property of the patients.

Interviewer blinding. PARLOUR-2 could not blind the interviewer to allocation. An interviewer knows whether there is an owner in the room. We mitigated this by blinding the adjudication panel and by pre-registering the endpoint, but the mitigation is partial and the effect estimate may be inflated. We do not think it is inflated by 17.7 points; we cannot show that it is not.

Training. Fourteen hours of training per interviewer is a real constraint on deployment, and our reliability estimates come from interviewers who completed it. We have no data on VHI-9 administered by the untrained.

Generalisability. All eleven practices were first-opinion. The instrument has not been evaluated in referral, emergency, or production-animal settings, and item 5 (diurnal variation) requires two administrations twelve hours apart, which most consultations do not permit.

8.3 Consent

A history obtained from a patient in the absence of the person who brought it raises a governance question we are not competent to settle. Current consent frameworks treat the owner as the sole informant and the sole consenting party, which is coherent only while the patient is not itself a source of clinical information [14]. Our findings remove that condition. We have referred the question to the Standing Committee on Practice Ethics and note that fourteen owners in PARLOUR-2 declined to leave, all of whom were accommodated.

9 Conclusion

The veterinary patient gives a history. It is localisable, gradable, and time-referenced, it can be obtained in six minutes with no equipment, and it is more accurate than the account of the person who brought the animal in. It has not previously been obtained under standardised conditions.

We recommend that the Council's guidance on the structure of the first-opinion consultation be revised to place VHI-9 administration before owner interview, and that the accompanying owner be asked to wait outside for the duration. Instrument, manual, and training materials are available from the corresponding author on request.

Ethics, funding, and competing interests

Both studies were approved by the Fernhurst Veterinary Group practice ethics committee (references FVG/21/044 and FVG/23/009). Owners gave written informed consent at enrolment. The work was funded by the Council for Veterinary Practice small grants scheme (CVP-SG-2021-77); the funder had no role in design, analysis, or the decision to submit. The authors declare no competing interests.

Appendix A: PARLOUR-2 investigators

Trial steering committee, data monitoring committee, adjudication panel, and practice principal investigators are listed in the supplementary material.

References

[1] Hampton, J. R., Harrison, M. J. G., Mitchell, J. R. (1975). Relative contributions of history-taking, physical examination,

and investigation to diagnosis. *Clinical Methods Quarterly* 12(2), 486–489.

[2] Peterson, M. C., Osterhaus, K. L. (1992). Contributions of the history to the final diagnosis: a re-audit at forty years. *Journal of Diagnostic Practice* 31(4), 163–170.

[3] Radford, A. L., Whitcombe, E. J. (2011). *The Structure of the First-Opinion Consultation*, 3rd edn. Ambersyke Press, Ambersyke.

[4] Dean, R. S., Okonkwo, B. (2017). Who is the informant? Proxy history in small-animal practice. *Veterinary Consultation Science* 9(1), 22–31.

[5] Morton, C. H., Lindqvist, A. E., Sackville, P. (2019). Owner recall of onset, duration, and severity: agreement with contemporaneous record in 4,102 presentations. *Journal of Small Animal Measurement* 44(6), 611–624.

[6] Carbone, F., Ekwueme, N. A. (2015). Directional error in proxy report: a meta-analysis of 61 studies. *Clinical Informant Research* 7(3), 201–218.

[7] Reid, J., Nolan, A. M., Scott, E. M. (2007). Development and validation of a composite behavioural scale. *Veterinary Behavioural Measurement* 18(2), 97–108.

[8] Pope, H. V., Tervoort, L. (2014). Cross-species performance of composite behavioural scales: a multi-centre reliability study. *Veterinary Behavioural Measurement* 25(4), 330–342.

[9] Ely, E. W., Margolin, R., Francis, J. (2001). Evaluation of the non-communicating patient: validation of a structured assessment method. *Critical Care Assessment* 29(7), 1370–1379.

[10] Murphy, J., Cameron, L., Okafor, D. (2010). Obtaining a structured history in severe expressive aphasia. *Communication and Clinical Practice* 24(3), 245–259.

[11] Bland, J. M., Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Methods in Measurement* 1(8476), 307–310.

[12] Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics Reports* 33(1), 159–174.

[13] Hopcroft, J. E., Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Farrowgate, Reading.

[14] Standing Committee on Practice Ethics (2022). *Guidance on Informant Standing in the Small-Animal Consultation*, 2nd edn. Council for Veterinary Practice, Circular 22/07.