Enoncé et corrigé du TD sur l'analyse critique de démarches scientifiques (séance n°7)

Objectifs d'apprentissage visés du TD n°7

(A) **Définir** l'inférence statistique ;

(A) **Fournir** la signification de la Standard Error de toute estimation, et son lien avec l'intervalle de confiance à 95% de l'estimation et avec la taille d'échantillon ;

(A) **Calculer** l'intervalle de confiance à 95% d'une estimation (moyenne, pourcentage) ;

(A) **Citez** dans une phrase les pourcentages qui vont être comparés puis testés, à partir d'effectifs fournis dans un tableau, dans l'objectif de tester l'association entre deux variables binaires ;

(A) **Interpréter** les informations contenues dans un tableau issu d'une étude clinique comprenant des pourcentages, moyennes, médianes, et degrés de signification ;

(A) **Communiquer** de façon pertinente à l'issue d'un résultat du test statistique non significatif ;

(A) **Citer** l'hypothèse nulle H0 d'un test statistique, qu'il soit ou non présenté dans un tableau issu d'une étude clinique ;

(A) **Identifier** la situation dans laquelle l'utilisation d'un test statistique (et ses conclusions qui vont avec) n'est pas pertinente ;

(B) **Appliquer** la définition du risque d'erreur de première espèce (α) d'un test statistique significatif présenté dans une étude clinique ;

(B) **Définir** le risque d'erreur de seconde espèce (β) d'un test statistique.

Descriptif général du TD

Comme les TD précédents, ce TD va se baser sur des tableaux d'articles parus dans des articles de recherche clinique vétérinaire. Ce TD a pour objectif de continuer à vous faire prendre conscience qu'il ne faut pas prendre pour « argent comptant » la communication scientifique délivrée par les chercheurs sous prétexte qu'elle est délivrée par des chercheurs mondialement connus ou parce que publiée dans des revues dites de « prestige ». Vous devez développer votre esprit critique, et pour le faire, vous devez entre autres avoir acquis les bases en biostatistique appliquée à la médecine (vétérinaire).

1er article : Nagata et al, J Vet Intern Med, 2017

L'objectif de l'étude était de montrer que, chez les chiens adultes souffrant d'un syndrome de Cushing[1], la mise sous traitement par du Trilostane au moment du diagnostic a un effet bénéfique sur le taux de survie des chiens dans les mois ou années qui suivent le diagnostic, comparé à une absence de traitement au moment du diagnostic.

Les auteurs ont utilisé des données collectées dans les dossiers médicaux dans l'hôpital vétérinaire Yuki, au Japon. Dans une telle étude qui vise à montrer que la mise sous traitement par du Trilostane (*versus* pas de traitement) est associée à la survenue d'un décès par la suite moins rapide que si l'on ne reçoit pas de traitement, dans la population des chiens souffrant d'un syndrome de Cushing, il est *au préalable* fondamental de vérifier que les deux groupes (« Trilostane » et « absence de traitement ») sont cliniquement comparables (dans le sens de « cliniquement similaires ») dans l'échantillon au moment du diagnostic sur des critères cliniques pertinents[2]. Pour cela, les auteurs ont dressé le tableau ci-dessous. « Baseline » veut dire de façon générale « à l'inclusion dans l'étude » ; ce terme signifie ici « au moment du diagnostic » (les chiens sont en effet inclus dans l'étude dès lors qu'ils sont diagnostiqués avec un syndrome de Cushing). Par conséquent, les valeurs des variables dans le tableau sont les valeurs recueillies au moment du diagnostic du syndrome de Cushing, donc juste avant que les 17 chiens ne soient mis sous Trilostane (les 26 autres, de toute façon, n'ont rien reçu comme traitement le jour du diagnostic).

Table 1. Baseline characteristics according to treatment group.

Variable	Trilostane (n = 17)	Untreated (n = 26)	*P* value
Median age, years (IQR)	10 (9–13)	12 (10–12.8)	.50
Median weight, kg (IQR)	7.3 (5.4–14.5)	7.9 (5.3–10.9)	.56
Female, number (%)	10 (59)	18 (69)	.53
Median ALT, U/L (IQR)	98 (72–173)	91 (51–162)	.41
2 → Median ALP, U/L (IQR)	1651 (1008–2363)	1303 (824–2387)	.69 ← 2
Median pre-ACTH cortisol, μg/dL (IQR)	11.2 (7.7–17.3)	4.7 (3.7–7.7)	.006
Median post-ACTH cortisol, μg/dL (IQR)	36.4 (26.4–50)	31.4 (25.1–41.5)	.22
1 → Hepatomegaly, number (%)	8 (47)	17 (65)	.34 ← 1
Polyuria and polydipsia, number (%)	5 (29)	13 (50)	.22
Abdominal distension, number (%)	5 (29)	9 (35)	1
Alopecia, number (%)	6 (35)	6 (23)	.49
Panting, number (%)	2 (12)	4 (15)	1
Visit[a], number (IQR)	21 (11–36)	20 (8.8–27)	.35
Cost[b], US$ (IQR)	1554 (776–3502)	1262 (761–2110)	.51

IQR, interquartile range; ALT, alanine aminotransferase activity; ALP, alkaline phosphatase activity; pre-ACTH cortisol, cortisol concentrations before ACTH stimulation; post-ACTH cortisol, cortisol concentrations after ACTH stimulation.

[a]Apart from the monitoring visits.

[b]Apart from the cost of trilostane and its monitoring.

Dans le tableau ci-dessus, pour les variables binaires (par exemple, « Female »), « number » veut dire « nombre de chien » et « (%) » indique que les chiffres dans le tableau qui sont entre parenthèses sont des %. Par exemple, il y a 10 chiens femelles parmi les 17 chiens traités par du Trilostane, soit 59%.

1.1) Quels sont les deux indicateurs que teste le test statistique concernant le syndrome de polyurie-polydipsie (PU-PD) dont la valeur du degré de signification est de « 0,22 » (cf. flèches n°1) ? (Pour information, la PU-PD est définie par une diurèse supérieure à 50 ml/kg/j (polyurie) associée à une augmentation des apports liquidiens (polydipsie).)

Le test statistique teste l'association entre la présence d'une PU-PD au moment du diagnostic et le type de traitement reçu (Trilostane versus aucun traitement). Ces deux variables sont binaires, donc le test compare deux pourcentages. Parmi les 4 couples de % que l'on peut citer, je vais citer les deux que l'on

[1] Ensemble de symptômes résultant d'une imprégnation cortisolique chronique de l'organisme, ce qui sous-entend les origines hypophysaires, les origines surrénaliennes (« Cushing spontané »), et les Cushing iatrogéniques (Cushing provoqué par un acte médical).

[2] Nous reverrons en détails cette condition et les raisons de cette condition dans l'UC-0324 l'année prochaine.

retrouve dans le tableau : le pourcentage de chiens avec une PU-PD parmi les 17 chiens sous Trilostane (5/17=29%) et le pourcentage de chiens avec une PU-PD parmi les 26 chiens non traités (13/26=50%).

Le test statistique est soit le test du Chi-2, soit le test de Fisher, en fonction des effectifs attendus sous H_0. Il se trouve que les effectifs attendus sous H_0 sont tous > 5. Cela dit, le test du Chi-2 conduit à un degré de signification de 0,18 tandis que celui de Fisher conduit à un degré de signification de … 0,22, qui est la valeur de l'article… Bizarre…

1.2) Dans l'échantillon, le groupe des 17 chiens sous Trilostane était-il cliniquement comparable (ou « cliniquement similaire ») juste avant d'être mis sous Trilostane au groupe des 26 chiens non traités sur la présence d'une polyurie-polydipsie (PU-PD) au moment du diagnostic ? Quel est ou quelles sont les informations du tableau que vous allez utiliser pour répondre à cette question ?

La ligne du tableau correspondant à la PU-PD comprend cinq nombres : 5, 29, 13, 50, et 0,22. Parmi ces cinq nombres, quels sont ceux qui vont permettre de répondre à la question ?

L'erreur serait d'utiliser la valeur de degré de signification de « 0,22 », qui teste l'association entre la présence de PU-PD et l'appartenance au groupe de traitement (Trilostane versus sans traitement), pour répondre à la question.

Tout d'abord, un test statistique ne sert qu'à une seule chose : faire de l'inférence lorsque $p \leq 0,05$. C'est-à-dire, conclure au niveau de la population cible. Ici, les auteurs se moquent (et ils ont bien raison) de savoir si les deux groupes sont réellement (dans la population des millions de chiens souffrant d'un syndrome de Cushing) différents dans la population cible, ils veulent en effet savoir si dans l'échantillon, les deux groupes sont comparables. Et lorsque l'on ne souhaite pas faire d'inférence, on ne fait tout simplement pas de test statistique.

De plus, contrairement à ce que pensent de très nombreux chercheurs, ce n'est pas parce qu'une différence est non significative qu'elle correspond à une différence « faible » d'un point de vue clinique (dans les TD n°3 et n°4, le pourcentage de chiens décédés parmi les chiens non obèses à 25%) n'était pas significativement différent du % de chiens décédés parmi les chiens obèses (45%), et pourtant, la différence entre ces deux % estimés dans l'échantillon ne peut pas être considérée comme faible, cliniquement parlant).

Une fois que l'on a dit qu'il ne fallait pas utiliser le degré de signification pour répondre à la question, quels sont les nombres que nous devons utiliser pour y répondre ?

Les nombres « 5 » et « 13 » sont des effectifs, c'est-à-dire un nombre de chiens avec une PU-PD dans chacun des deux groupes. Mais pour savoir si, dans l'échantillon, les deux groupes sont cliniquement comparables (similaires) sur la présence de PU-PD, ce ne sont pas des nombres qui doivent être comparés, mais des pourcentages. En effet, ici, ce n'est pas le nombre de chiens avec une PU-PD (n=5) parmi les chiens traités sous Trilostane qui doit être comparé au nombre de chiens avec une PU-PD (n=13) parmi les chies non traités, mais le pourcentage de chiens avec une PU-PD parmi les chiens sous Trilostane (5/17=29%) à comparer au pourcentage de chiens avec une PU-PD parmi les chiens non traités (13/26=50%).

Ainsi, peut-on dire que 29% et 50% sont des pourcentages cliniquement comparables (similaires) dans l'échantillon ? Je pense que l'on peut difficilement le dire. Et pourtant, ces deux pourcentages ne sont pas significativement différents (p=0,22 ; en effet, le test statistique testait si ces deux pourcentages étaient significativement différents). Donc, écrire « les deux groupes de chiens (sous Trilostane et sans traitement) sont cliniquement comparables/similaires sur la présence de PU-PD parce que le pourcentage de PU-PD parmi les 17 chiens sous Trilostane (29%) n'était pas significativement différent

de celui parmi les 26 chiens non traités (50%) » est une grosse erreur de raisonnement que de nombreux chercheurs font malheureusement [3].

La première partie de cette question « Dans l'échantillon, le groupe des 17 chiens sous Trilostane était-il cliniquement comparable (ou « cliniquement similaire ») juste avant d'être mis sous Trilostane au groupe des 26 chiens non traités sur la présence d'une polyurie-polydipsie (PU-PD) au moment du diagnostic ? » ne vous sera jamais posée à l'examen telle quelle. En revanche, la 2ème partie de la question « Quel est ou quelles sont les informations du tableau que vous allez utiliser pour répondre à cette question ? » peut tout à fait tomber à l'examen.

1.3) Le groupe des 17 chiens sous Trilostane était-il cliniquement comparable (ou « cliniquement similaire ») juste avant d'être mis sous Trilostane au groupe des 26 chiens non traités sur la concentration en cortisol pré-stimulation ACTH (flèches n°2) ?

Le raisonnement est bien entendu identique au précédent : il ne faut pas regarder le degré de signification du test statistique (ici, « 0,006 »), mais les valeurs des indicateurs comparés, ici les médianes de concentration en cortisol pré-stimulation ACTH de 11,2 µg/dL et 4,7 µg/dL respectivement parmi les 17 chiens sous Trilostane et parmi les 26 chiens non traités. Et là encore, ce n'est pas parce qu'une différence d'indicateurs est significative (ce qui est le cas ici : les deux médianes sont significativement différentes puisque p = 0,006) que les deux indicateurs comparés sont cliniquement très différents. La comparabilité clinique doit s'évaluer cliniquement et non pas statistiquement (ça semble pourtant évident, non ?! 😉). Le test statistique est au mieux inutile, et au pire générateur d'une erreur d'appréciation, dans le cas où l'on doit vérifier la comparabilité clinique entre deux (ou plusieurs groupes) sur des données d'inclusion (données de « baseline »).

La question est donc de savoir si, cliniquement parlant, 11,2 µg/dL et 4,7 µg/dL sont deux valeurs cliniquement différentes. Je ne sais pas répondre à la question, et je n'attends bien évidemment pas que vous sachiez y répondre. Ce que j'attends de vous, c'est de poser la démarche pour y répondre : utiliser les indicateurs (%, moyennes, médianes) et non pas le degré de signification pour vérifier que deux groupes d'animaux sont cliniquement comparables (similaires) sur des variables d'inclusion.

1.4) Mais alors, à quel moment dans l'étude de Nagata et coll. faut-il utiliser un degré de signification ?!

De façon générale, on réalise un test statistique lorsque l'on souhaite faire de l'inférence. Et en l'occurrence ici, les auteurs souhaitent montrer qu'il existe une différence <u>réelle</u> sur le taux de survie entre les millions de chiens sous Trilostane et les millions de chiens non traités, dans la population des millions de chiens souffrant d'un syndrome de Cushing. Et pour cela, dans l'échantillon, le taux de survie médian parmi les 17 chiens sous Trilostane devra être significativement différent (p < 0,05) du taux de survie médian parmi les 26 chiens non traités, ce qui est le cas dans cette étude (la courbe ci-dessous est une courbe dite de « Kaplan-Meier », que nous étudierons en détails l'année prochaine, dans l'UC-0313 😉) :

[3] D'où la phrase citée sur la page de couverture du polycopié de biostatistique en médecine vétérinaire 😉

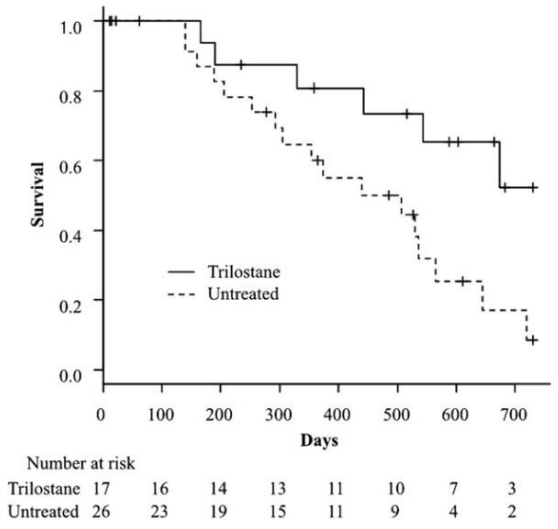

Fig 2. Kaplan–Meier survival curves for the trilostane group (solid line) and untreated group (dashed line). Median survival time for the trilostane group was not reached (95% confidence interval [CI], 443 days–not applicable), and was significantly longer than the 506 days (95% CI, 292–564 days; $P = .016$) for the untreated group. ←

2ème article : Borgarelli et al, J Vet Intern Med, 2015

L'objectif de l'étude était de montrer que, chez les chiens adultes souffrant d'une maladie valvulaire dégénérative mitrale (MVDM), la présence d'une hypertension pulmonaire (HP) augmente les risques de décès chez ces chiens. Les auteurs ont recruté 212 chiens au moment de la consultation au cours de laquelle le diagnostic de MVDM était posé. Parmi ces 212 chiens, 129 ne présentaient pas d'HP à la consultation de diagnostic et 83 en présentaient une. Le tableau ci-dessous permet d'étudier l'association entre la présence (*versus* absence) d'une HP et différentes caractéristiques des chiens.

Table 1. Descriptive statistics for all 212 dogs, dogs without pulmonary hypertension (PH), and dogs with PH

	All (n = 212)	No PH (129)	PH (n = 83)	*P*-Value
Age (years) (n = 212)	10.6 ± 2.6	10.6 ± 2.6	10.7 ± 2.7	.94
Sex (F/M) (n = 212)	91/121 (43%/57%)	55/74 (43%/57%)	36/57 (43%/57%)	.91
Weight (kg) (n = 212)	8.6 (1.2–80.7)	9.1 (1.2–80.7)	7.7 (1.6–67)	.19
ACVIM stage (n = 212)	100 B2 (47%)	76 B2 (59%)	24 B2 (29%)	.0022
	112 C (53%)	53 C (41%)	59 C (71%)	.0023
LA/Ao (n = 211)	2 (1.3–3.9)	1.9 (1.3–3.2)	2.3 (1.4–3.9)	<.0001
LVEDDn (n = 211)	2.02 (0.8–2.9)	1.9 (0.9–2.9)	2.1 (0.8–2.9)	.006
LVESDn (n = 211)	1.05 (0.2–1.9)	1.04 (0.4–1.9)	1.04 (0.2–1.8)	.877
E peak (m/s) (n = 117)	1.28 ± 0.40	1.3 ± 0.3	1.42 ± 0.37	.08
TRPG (mmHg) (n = 191)	33.2 (1.4–98.4)	24.1 (1.4–35.5)	46.2 (36.0–98.4)	<.0001
RV enlargement (yes/no) (n = 123) ↙	15 (12%)	7(47%)	8 (53%)	.067

P-value is referring to differences between dogs with and without PH. No PH, no pulmonary hypertension, ACVIM stage, class of heart failure according to ACVIM classification; LA/Ao, left-atrial to aortic root ratio; LVEDDn, normalized left-ventricle end-diastolic diameter indexed; LVESDn, normalized left-ventricle end-systolic diameter indexed; E peak, peak velocity of E wave of transmitral flow; TRPG, tricuspid regurgitation pressure gradient; RV, right ventricle.

2.1) La dernière ligne du tableau concerne la variable « RV enlargement » (augmentation de la taille du ventricule droit [VD], qui est une variable binaire : présence / absence d'une telle augmentation). On se rend compte en lisant cette ligne que l'on a l'information sur la taille du VD pour seulement 123 chiens parmi les 212 chiens de l'échantillon. C'est un peu embêtant, mais cela arrive quand il s'agit d'une variable qui n'est pas systématiquement recueillie. Que laissent spontanément penser la confrontation des valeurs de « 47% » et « 53% » sur cette ligne du tableau (ne vous posez pour l'instant pas la question de savoir comment ces valeurs ont été calculées) ?

Ces valeurs étant très proches l'une de l'autre, elles laissent spontanément penser que, dans l'échantillon, les deux groupes de chiens avec et sans HP semblent comparables (ou similaires) sur la

présence d'une augmentation de la taille du VD, ce qui signifie que dans l'échantillon, il y a une très faible association entre la présence d'une HP et la présence d'une augmentation du CD (« égalité des indicateurs » = « absence d'association » → « indicateurs très proches » = « très faible association »).

2.2) Dans l'échantillon des chiens dont l'information sur la taille du VD était connue, combien de chiens ont présenté une augmentation de la taille du VD ? Retrouvez le calcul qui a conduit à la valeur de « 12% » (colonne « All »).

Dans la colonne « All », on peut lire qu'il y a 15 chiens parmi les 123 qui ont présenté une augmentation de la taille du VD. Et 12% = 15/123.

2.3) Retrouvez les calculs qui ont conduit aux valeurs de « 47% » et « 53% », puis citez la phrase en français correspondant à chacun de ces deux calculs.

47% a été obtenu en faisant 7/15. La phrase en français est donc : le pourcentage de chiens ne présentant pas d'HP parmi les chiens de l'échantillon ayant présenté une augmentation du VD.

53% a été obtenu en faisant 8/15. La phrase en français est donc : le pourcentage de chiens présentant une HP parmi les chiens de l'échantillon ayant présenté une augmentation du VD.

2.4) Commentez votre réponse à la question précédente, au regard de ce que vous avez écrit à la question 2.1.

*Vous vous rendez compte, grâce aux phrases en français, que ces deux pourcentages sont les mauvais pourcentages à citer pour savoir si deux groupes diffèrent ou non sur la présence d'un caractère binaire, ou autrement dit pour savoir s'il existe une association entre deux caractères. En effet, leur somme fait forcément 100%, puisque ces deux pourcentages ont été calculés parmi uniquement les 15 chiens qui ont présenté une augmentation de la taille du VD. Les auteurs n'ont pas suivi la règle du « le pourcentage de XX parmi les uns à comparer au pourcentage de XX parmi les autres », puisque les auteurs sont restés uniquement parmi les chiens de l'échantillon ayant présenté une augmentation du VD. La proximité des deux pourcentages « 47% » et « 53% » laissent donc, peut-être à tort, penser que dans l'échantillon, l'association entre la présence d'une HP et celle d'une augmentation du VD est faible. (Les pourcentages corrects qu'ils auraient dû présenter sont[4] : le pourcentage de chiens ayant présenté une augmentation du VD parmi les chiens **sans** HP et le pourcentage de chiens ayant présenté une augmentation du VD parmi les chiens **avec** HP. Malheureusement, les données ne nous permettent pas de calculer ces deux pourcentages : les dénominateurs ne sont pas 129 et 83, respectivement, puisqu'il y a de nombreuses données manquantes sur l'information de l'augmentation de la taille du VD.)*

*Par conséquent, la proximité de ces deux pourcentages (47% et 53%) ne veut **surtout pas** dire que, dans l'échantillon, les deux groupes de chiens avec et sans HP semblaient comparables ou similaires sur la présence d'une augmentation de la taille du VD, et donc que dans l'échantillon, l'association entre la présence d'une HP et celle d'une augmentation du VD était faible. (On remarquera au passage que les auteurs n'ont pas commis cette erreur pour le sexe des chiens…)*

Vous avez une belle preuve ici que si vous ne savez pas les règles de bases en biostat, vous pourriez très mal interpréter les résultats publiés dans des communications scientifiques ! Ou pire, croire ce que vous entendez dire par des personnes qui auront mal interprété les résultats d'une étude, sans vous, prendre la peine d'aller voir les choses par vous-même.

[4] Je vous présente ici un seul couple de pourcentages, mais vous savez (cf. TD n°2, 3, et 4) que trois autres couples de pourcentages auraient pu être cités.

3ème article : Arenas, J Vet Intern Med, 2014

3) Lisez le résumé de l'article ci-dessous, puis donnez la raison pour laquelle ce qui est écrit dans la conclusion n'est pas compatible avec les méthodes statistiques utilisées et leurs résultats ? (Ce qui ne veut pas dire que ce qui est écrit dans la conclusion est faux, mais cela signifie que l'étude de Arenas n'apporte aucun élément rigoureux pour laisser penser ce que les auteurs disent en conclusion.)

J Vet Intern Med 2014;28:473–480

Long-Term Survival of Dogs with Adrenal-Dependent Hyperadrenocorticism: A Comparison between Mitotane and Twice Daily Trilostane Treatment

C. Arenas, C. Melián, and M.D. Pérez-Alenza

Background: Treatment of adrenal-dependent hyperadrenocorticism (ADH) involves either surgical resection of the adrenal tumor or medical therapy. For many years, mitotane has been considered the medical treatment of choice for dogs with ADH.

Objectives: The aim of this study was to determine survival and prognostic factors for dogs with ADH treated with mitotane and trilostane.

Animals: Twenty-six dogs with ADH were included in the study.

Methods: Fourteen dogs were treated with mitotane and 12 dogs were treated with trilostane. Medical records were reviewed. Epidemiologic factors, signalment, clinicopathologic abnormalities, endocrine test results, and treatment protocols were evaluated to identify potential predictive factors of overall survival time.

Results: Survival times of dogs treated with mitotane (median, 15.6 months) or trilostane (median, 14.0 months) were not significantly different. Using univariate analysis, age and postadrenocorticotropic hormone cortisol concentrations were inversely correlated with survival time. The multivariate model also identified weakness at presentation as a negative prognostic indicator.

Conclusion and Clinical Importance: The type of medical treatment (mitotane versus trilostane) does not influence survival time in dogs with ADH; therefore, trilostane, a drug with less frequent and milder adverse effects, might be used as the primary medical treatment when adrenalectomy cannot be performed.

Key words: Cushing; Endocrine; Internal medicine.

Ce qui est écrit dans la conclusion est « The type of medical treatment (mitotane versus trilostane) does not influence survival time in dogs with ADH », à partir des résultats suivants présentés plus haut dans le résumé : « Survival times of dogs treated with mitotane (median, 15.6 months) or trilostane (median, 14.0 months) were not significantly different. »

La conclusion ci-dessus correspond à de l'inférence (présent utilisé, population cible citée (« in dogs with ADH »)), en parlant d'une égalité d'indicateurs (écrire que le type de traitement n'influence pas la survie est équivalent à écrire qu'il n'y a pas de différence de taux de survie entre les deux groupes). Statistiquement parlant, les auteurs sont donc en train d'écrire « H0 est vraie », à partir d'une différence non significative. Non seulement ils ne prennent pas de gants en inférant (souvenez-vous, quand on rejette H0, on prend des gants en disant « il y a des chances pour que H0 soit fausse »), mais aussi et surtout, il est interdit d'inférer une égalité d'indicateurs dans la population cible (ou bien dire que, dans la population cible, les indicateurs ont des chances d'être voisins) à partir d'une différence non significative de ces mêmes indicateurs dans l'échantillon.

4ème article : Watanabe, PLoS One, 2020

L'objectif de l'étude était de montrer une différence d'efficacité anesthésique entre deux protocoles de dosage utilisant deux concentrations différentes de buprénorphine chez des chats subissant des extractions dentaires (Simbadol versus Vetergesic). Pour cela, les auteurs ont en autres utilisé un score de douleur (le score CMPS-F).

4.1) Le tableau ci-dessous présente les moyennes des scores de douleurs dans chacun des deux groupes à différents temps. La colonne « p value between groups » provient d'un test statistique testant la différence de moyennes entre les deux groupes à chaque temps évalué (« baseline » = « 60 minutes avant la prémédication »).

Table 5. Pain scores using the Glasgow Composite Measure Pain Scale-Feline (CMPS-F) in cats undergoing dental extractions after the administration of Simbadol or Vetergesic. Values are expressed as mean (SEM).

Time points		Treatments	CMPS-F	*p* value between groups
Day 1	**Baseline**	Simbadol (n = 11)	0.7 (0.5)	0.858
		Vetergesic (n = 12)	0.8 (0.4)	
	Postoperative 0.5 h	Simbadol (n = 11)	0.9 (0.5)	0.558
		Vetergesic (n = 12)	0.5 (0.4)	
	Postoperative 1 h	Simbadol (n = 11)	1.5 (0.5)	0.148
		Vetergesic (n = 12)	0.6 (0.4)	
	Postoperative 2h	Simbadol (n = 11)	2.0 (0.5)	0.371
		Vetergesic (n = 12)	1.4 (0.4)	
	Postoperative 4 h	Simbadol (n = 11)	2.5 (0.5)	0.920
		Vetergesic (n = 11)	2.4 (0.4)	
	Postoperative 8 h	Simbadol (n = 9)	2.6 (0.5)	0.759
		Vetergesic (n = 10)	2.8 (0.5)	
Day 2	**8 am**	Simbadol (n = 9)	2.3 (0.5)	0.234
		Vetergesic (n = 9)	3.1 (0.5)	
	4 pm	Simbadol (n = 8)	1.7 (0.5)	0.775
		Vetergesic (n = 7)	1.9 (0.5)	
	Midnight	Simbadol (n = 8)	1.4 (0.5)	0.883
		Vetergesic (n = 7)	1.3 (0.5)	
Day 3	**8 am**	Simbadol (n = 8)	1.2 (0.5)	0.596
		Vetergesic (n = 7)	1.5 (0.5)	
	4 pm	Simbadol (n = 8)	1.6 (0.5)	0.297
		Vetergesic (n = 7)	0.9 (0.5)	
	Midnight	Simbadol (n = 7)	1.4 (0.5)	0.276
		Vetergesic (n = 7)	0.6 (0.5)	

Supposons qu'en vrai, les deux régimes de traitement (Simbadol et Vetergesic) ont des effets identiques sur la réduction de la douleur. Quelle est, sous cette hypothèse, la probabilité d'observer une différence significative de moyennes de score de douleurs 30 minutes après l'opération ? En déduire la probabilité de ne pas observer de différence significative, sous cette même hypothèse.

Lorsque H_0 est vraie (ce qui est le cas sous l'hypothèse énoncée dans la question), il existe 5% de risque de rejeter (à tort donc) H_0 : c'est la définition du risque d'erreur de 1ère espèce α. Par conséquent, la probabilité d'observer une différence significative entre les moyennes du score de douleur 30 minutes après l'opération, sous l'hypothèse que les deux régimes de traitement (Simbadol et Vetergesic) ont des effets identiques sur la réduction de la douleur, est de 5%.

Evidemment, la probabilité de ne pas observer de différence significative sous cette même hypothèse est 100% – 5% = 95%.

4.2) Onze tests statistiques testant la différence de moyennes de score de douleur ont été réalisés : 30 minutes, 1h, 2h, 4h, 8h, au jour 2 (à trois moments), et au jour 3 (à trois moments). Faisons la même hypothèse que ci-dessus (les deux régimes de traitement (Simbadol et Vetergesic) ont des effets identiques sur la réduction de la douleur). Quelle est, sous cette hypothèse, la probabilité d'observer au moins une différence significative parmi les onze testées ? Aidez-vous de la réponse à la question précédente, et commentez ce résultat.

Pr(au moins une différence significative) = 1 – Pr(aucune différence significative parmi les 11) = 1 – Pr(pas de différence signif. à 30 minutes ET à 1h ET à 4h, et …, et à minuit du jour 3) = 1 – 0,95x0,95x…x0,95x0,95 = 1 – $0,95^{11}$ = 0,43 = 43%

Ainsi, si les deux régimes de traitement (Simbadol et Vetergesic) ont des effets identiques sur la réduction de la douleur, il y avait 43% de chances (« de risques », en fait) d'observer malgré tout une différence significative à au moins l'un des moments étudiés. C'est énorme, et c'est beaucoup plus que le seuil de 5% attendu : normalement, si en vrai H_0 est vraie, on ne doit rejeter H_0 que dans 5% des cas. Dans le cas de l'article, si les deux régimes de traitement (Simbadol et Vetergesic) ont des effets identiques sur la réduction de la douleur, il n'y avait non pas 5% mais 43% de risques de rejeter H_0 au moins une fois : à 30 minutes, 1h, …, ou minuit du $3^{ème}$ jour.

Cette situation s'appelle la situation de « multiple testing » (tests statistiques multiples) et doit être repérée dans un article. Il est interdit de multiplier les tests statistiques pour maximiser ses chances d'obtenir une différence significative (ce serait trop simple), sauf si l'on prend en compte la situation de tests statistiques multiples.

Ainsi, deux solutions pour éviter cette situation de tests statistiques multiples : soit on ne part pas à la pêche au degré de signification < 0,05 (autrement dit, on ne teste qu'une seule différence, à un seul moment), soit on utilise des tests statistiques qui prennent en compte la situation de tests statistiques multiples. Et pour information, les auteurs ont utilisé cette seconde solution 😉 (donc les valeurs des degrés de signification de la table ont été modifiées pour qu'elles puissent être correctement comparées à 5%).

Si vous pensez vous trouver dans cette situation lors de votre thèse (je sais, c'est dans quelques années !), j'ai rédigé un petit guide pour mieux comprendre les choses, ici (« Les tests statistiques multiples (doc HAL) »).