Introduction to Bayesian analysis (applied to diagnostic test evaluation)

Julio Alvarez

VISAVET Health Surveillance Centre, Animal Health Department

Universidad Complutense

Quick outline

- The problem of evaluation of diagnostic tests: the gold standard dilemma
- Overview of Bayesian statistics
- Use of Bayesian statistics for evaluation of diagnostic tests

Diagnostic test evaluation

	Reference +	Reference -	**Total**
Test +			
Test -			
Total			**N**

Pos **TN**

Sensitivity is the ability of the test to correctly identify diseased individuals **FP**

Neg

Specificity is the ability of the test to correctly identify healthy individuals **TP** **FN** **Sick**

Healthy

$$Se = \frac{TP}{Sick} \qquad PPV = \frac{TP}{pos} \qquad AP = \frac{Pos}{N}$$

$$Sp = \frac{TN}{Healthy} \qquad NPV = \frac{TN}{Neg} \qquad TrP = \frac{Sick}{N}$$

The problem of diagnostic test evaluation

- Typically new/alternative tests are compared with well established/standardized tests ("gold standard"/ "reference test")
- This only informs of the performance of the new test **relative** to the old one!...

… unless we assume the gold standard is perfect (what we typically do)

Perfect test???

Perfect test????

Case example: mycobacterial diseases

- No perfect test (not even close…)
- Chronic slow diseases → delayed (non-protective) (humoral) immune response
- All known tests lack sensitivity and/or specificity
- Usual gold standard is culture… with a sensitivity <50% in multiple cases

Alternatives?

- Choose individuals of known status:
  - Negative: never saw the pathogen (never: different location/epidemiological settings)
  - Positive: experimental infections
- External validity???

Be careful when using certain terms!

Highly Accurate Antibody Assays for Early and Rapid Detection of Tuberculosis in African and Asian Elephants[▽]

Rena Greenwald,[1] Olena Lyashchenko,[1] Javan Esfandiari,[1] Michele Miller,[2] Susan Mikota,[3] John H. Olsen,[4] Ray Ball,[4] Genevieve Dumonceaux,[4] Dennis Schmitt,[5] Torsten Moller,[6] Janet B. Payeur,[7] Beth Harris,[7] Denise Sofranko,[8] W. Ray Waters,[9] and Konstantin P. Lyashchenko[1]*

recognized by elephant antibodies during disease. The serologic assays demonstrated 100% sensitivity and 95 to 100% specificity. Rapid and accurate antibody tests to identify infected elephants will likely allow

on disease status and history of exposure (Table 1). The TB-infected group included 26 animals from 17 herds with culture-confirmed TB due to *M. tuberculosis* (n = 25) or *M. bovis* (n = 1). Of the 26 elephants, 7 died and 11 were humanely euthanized. TB was not necessarily the cause of death or the reason for euthanasia. Disease was diagnosed antemortem by trunk wash culture (n = 15; 58%) or only postmortem by isolating *M. tuberculosis* or *M. bovis* from various tissues (n = 11; 42%). Ten elephants were treated with first-line anti-TB drugs

Development and Evaluation of an Enzyme-Linked Immunosorbent Assay for Use in the Detection of Bovine Tuberculosis in Cattle[▽]†

W. R. Waters,[1]* B. M. Buddle,[2] H. M. Vordermeier,[3] E. Gormley,[4] M. V. Palmer,[1] T. C. Thacker,[1] J. P. Bannantine,[1] J. R. Stabel,[1] R. Linscott,[5] E. Martel,[5] F. Milian,[6] W. Foshaug,[7] and J. C. Lawrence[5]

TABLE 1. Sensitivity of IDEXX *M. bovis* antibody ELISA with sera collected from naturally infected cattle

M. bovis-infected serum type		n[d]	No. of herds	Sensitivity (%)		
Source	ID or characterization[c]			Lot 1	Lot 2	Lot 3
Great Britain	AHVLA-2[a]	134	31	74.6	72.4	73.1
	AHVLA-1[b]	50	>5	86	88	86
Ireland	No visible lesions[b]	50	>5	48	44	46
	With visible lesions[b]	50	>5	72	70	70
	Skin test positive, Bovigam positive, with visible lesions[b]	30	22	96.7	86.7	90.0
New Zealand	AgResearch[b]	42	7	42.9	40.5	35.7
USA	Colorado[a]	81	1	44.4	45.7	49.4
	NVSL serum bank[a]	31	12	48.4	48.4	48.4
	Michigan[a]	10	1	30	30	30
Overall value		478	>89	63.6	61.9	62.6

[a] Infection status determined by histopathology with IS*6110* PCR and/or culture.
[b] Infection status determined by culture or presence of gross lesions and/or from a tuberculosis-affected herd.
[c] ID, identification. NVSL, National Veterinary Service Laboratory.
[d] n, number of animals.

TABLE 2. Specificity of IDEXX *M. bovis* antibody ELISA with sera collected from noninfected cattle from various geographic regions

Source of noninfected sera[a]	n[b]	No. of herds	Specificity (%)		
			Lot 1	Lot 2	Lot 3
Maine	126	2	99.2	98.4	99.2
Maine	126	2	99	98.4	99.2
Pennsylvania	79	1	88.6	92.4	98.7
Arkansas	39	1	100	100	97.4
New York	84	1	98.8	100	98.8
North Dakota	110	1	96.3	97.2	99.1
Washington	84	2	98.8	98.8	98.8
South Dakota	84	1	98.8	100	100
Missouri	92	>2	94.5	95.7	98.9
Texas	96	>2	93.7	93.8	95.8
Michigan	92	2	100	100	100
Iowa	8	1	100	100	100
Colorado	121	11	99.2	100	99.2
Great Britain (AHVLA)	50	>5	94	98	96
Ireland (UCD/DAFF)	92	16	100	100	95.7
Austria	316	>10	97.5	99.1	98.1
Overall value	1,473	>58	97.4	98.2	98.4

[a] Samples obtained from cattle from tuberculosis-free herds. AHVLA, Animal Health and Veterinary Laboratories Agency; UCD, University College Dublin; DAFF, Department of Agriculture, Fisheries and Food.
[b] n, number of animals.

So in reality…

	Test 2 +	Test 2 -	**Total**
Test 1 +	a	b	a+b
Test 1 -	c	d	c+d
Total	a+c	b+d	**N**

Sensitivity 1 and Sensitivity 2???
Specificity 1 and Specificity 2???

Alternatives to the “gold standard”-based approach

- “Hui and Walter model”(Hui and Walter, 1980)

  “If two tests are applied simultaneously to the same individuals from two populations with different disease prevalences, then assuming conditional independence of the errors of the two tests, the error rates of both tests and the true prevalences in both populations can be estimated by a maximum likelihood procedure”.

- Not originally Bayesian: can be solved using ML methods… but require 1) large sample size, 2) assume tests are independent and 3) their accuracy is assumed to be constant across populations

The Hui-Walter paradigm

- Hui-Walter model implementation to be further discussed in the next session

	Test 2 +	Test 2 -	**Total**
Test 1 +	a	b	a+b
Test 1 -	c	d	c+d
Total	a+c	b+d	**N**

Diseased population (P)

	Test 2 +	Test 2 -
Test 1 +	$p \times Se_1 \times Se_2$	$p \times Se_1 \times (1 - Se_2)$
Test 1 -	$p \times (1 - Se_1) \times Se_2$	$(1 - p) \times (1 - Se_1) \times (1 - Se_2)$

Non-diseased population (1-P)

	Test 2 +	Test 2 -
Test 1 +	$(1 - p) \times (1 - Sp_1) \times (1 - Sp_2)$	$(1 - p) \times (1 - Sp_1 \times Sp_2$
Test 1 -	$(1 - p) \times Sp_1 \times (1 - Sp_2)$	$(1 - p) \times Sp_1 \times Sp_2$

The Hui-Walter paradigm

- T1+T2+= P*Se1*Se2 + (1-P)*(1-Sp1)*(1-Sp2)
- T1+T2-= P*Se1*(1-Se2) + (1-P)*(1-Sp1)*Sp2
- T1-T2+= P*(1-Se1)*Se2 + (1-P)*Sp1*(1-Sp2)
- T1-T2-= P*(1-Se1)*(1-Se2) + (1-P)*Sp1*Sp2

5 parameters to estimate (P, Se1, Se2, Sp1, Sp2) vs. 3 degrees of freedom
Non-identifiable model ☹

The Hui-Walter paradigm

"If two tests are applied simultaneously to the same individuals from two populations with different disease prevalences, then assuming conditional independence of the errors of the two tests, the error rates of both tests and the true prevalences in both populations can be estimated by a maximum likelihood procedure".

- Population 1

T1+T2+: P1*Se1*Se2+(1-P1)*(1-Sp1)*(1-Sp2)

T1+T2-: P1*Se1*(1-Se2)+(1-P1)*(1-Sp1)*Sp2

T1-T2+: P1*(1-Se1)*Se2+(1-P1)*Sp1*(1-Sp2)

T1-T2-: P1*(1-Se1)*(1-Se2)+(1-P1)*Sp1*Sp2

- Population 2

T1+T2+: P2*Se1*Se2+(1-P2)*(1-Sp1)*(1-Sp2)

T1+T2-: P2*Se1*(1-Se2)+(1-P2)*(1-Sp1)*Sp2

T1-T2+: P2*(1-Se1)*Se2+(1-P2)*Sp1*(1-Sp2)

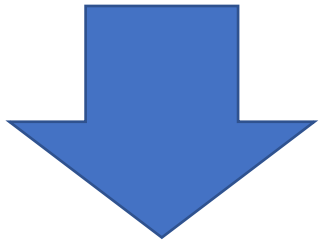
T1-T2-: P2*(1-Se1)*(1-Se2)+(1-P2)*Sp1*Sp2

6 parameters to estimate (P1, P2, Se1, Se2, Sp1, Sp2)

vs.

6 degrees of freedom

Identifiable model ☺



But has strong assumptions:

- 2 populations with different prevalences
- Se y Sp constant across populations
- Conditional Independence between tests

Conditional Independence?

- The result of one test is conditionally independent from the other (knowing if one diseased/non-diseased animal is positive/negative in one test gives no information on whether it will be positive/negative in the other (e.g., knowing that a card is a spades gives no information on whether it is a figure)
- Often not true!! (similar tests/tests based on similar principles
- Requires adding additional terms to equations (Vacek, 1985)

Considering conditional dependence

Number of parameters to estimate increases rapidly as the number of potentially dependent tests increases

TABLE 2. Maximum Number of Estimable Parameters and Number of Parameters to Be Estimated in the Absence of Conditional Independence and Under Conditional Independence as a Function of the Number of Tests per Subject

Number of Tests	Maximum Number of Estimable Parameters	Parameters to be Estimated Under Conditional Dependence	Parameters to Be Estimated Under Conditional Independence
1	1	3	3
2	3	7	5
3	7	15	7
4	15	31	9
5	31	63	11
h	$2^h - 1$	$2^{h+1} - 1$	$2h + 1$

Berkvens D et al. (2006) Estimating Disease Prevalence in a Bayesian Framework Using Probabilistic Constraints. doi: 10.1097/01.ede.0000198422.64801.8d

Alternatives to the “gold standard”-based approach

- If we could only incorporate some prior knowledge… we could get rid of large sample theory assumptions and “guide” our models
  - We typically have some information on diagnostic performance (i.e., test sensitivity is >20%)
  - We typically have some information on disease prevalence (i.e., proportion of infected is <50%)

Bayesians to the rescue!

- "How to revise our beliefs in the light of evidence" → probability of an event based on prior knowledge of conditions related to the event
- In its more general form:

$$Prior \times Likelihood = Posterior$$

- In its "primitive" definition: Bayes' rule

$$P(\theta|Y) = \frac{P(\theta) \times P(Y|\theta)}{P(Y)}$$

Bayesian applied to diagnostic test evaluation

- Joseph et al. (1995) **Bayesian** estimation of disease prevalence and diagnostic test evaluation in the absence of a gold standard

$$P(\theta|Y) = \frac{P(\theta) \times P(Y|\theta)}{P(Y)}$$

- Targets
  - Prevalence $\pi = P(D+)$
  - Sensitivity $Se_i = P(T_i+ \,|\, D+)$
  - Specificity $Sp_i = P(T_i- \,|\, D-)$
- Prior knowledge (beta distributions)
  - $\pi \sim Beta(a_\pi, b_\pi)$
  - $Se_i \sim Beta(a_{Sei}, b_{Sei})$
  - $Sp_i \sim Beta(a_{Spi}, b_{Spi})$

Application of Bayes' rule or Bayesian statistics?

- Probabilities referred to frequency of events
- In a more general framework: from "probability of the data given the model" (classic significance testing) to "probability of the model given the data" through an extension of Bayes' theorem.

$$P(\theta|D) = \frac{P(\theta|M) \times P(\theta)}{P(D)}$$

Bayes' theorem applied to models

$$P(\theta|D) = P(D|\theta) \times P(\theta) \Big/ P(D)$$

Posterior

Likelihood

Prior

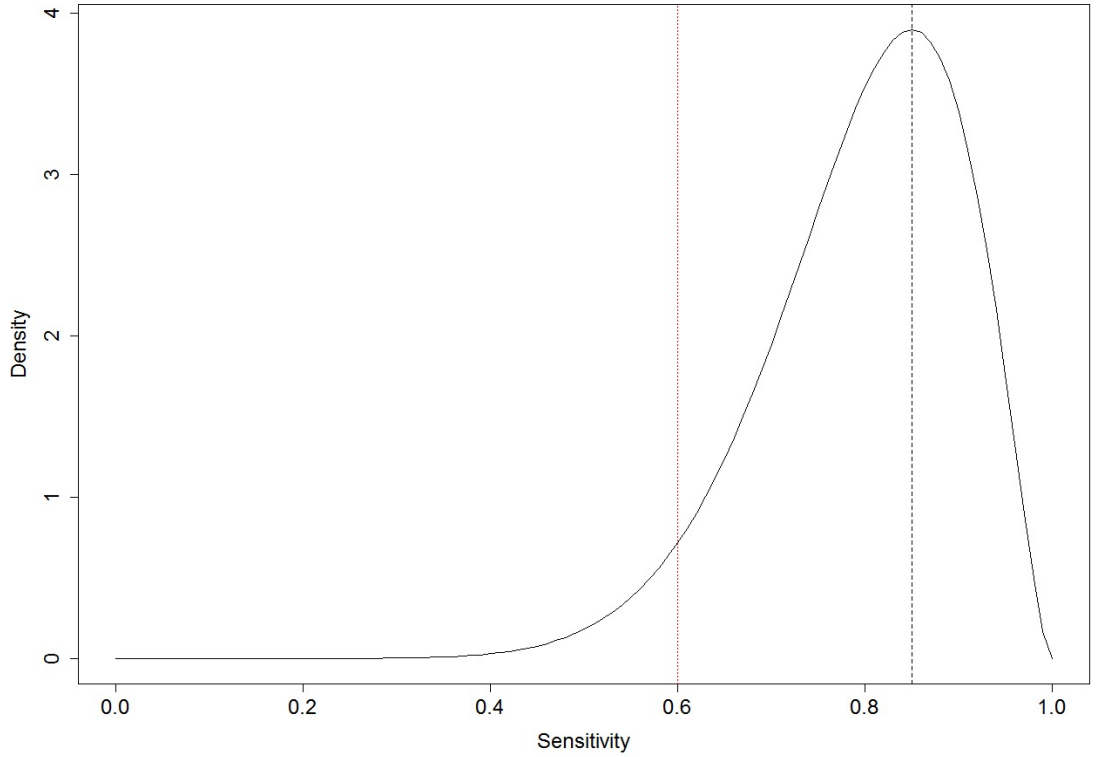
Evidence

Bayesian “crucial step”

- Frequentist statistics assumes something about θ (true, and with certain properties)
- Bayesian analysis considers θ as a random variable with a probability distribution $p(\theta)$ (prior) and $p(\theta|D)$ once that I have accounted for the data (conditional on D, posterior)

What is the sensitivity of test X?

- Option 1: 85%
- Option B:
  - Typically ~85%
  - Usually >60%



So it is a perfect match!!

- Bayesian statistics allows us to obtain a posterior distribution that can incorporate some prior belief (how sensitive do I think the test is) and some data (what result did I have in population X?)
- Problem: we need to compute the posterior distribution

$$P(\theta|D) = {P(D|\theta)\times P(\theta)}\Big/{P(D)}$$

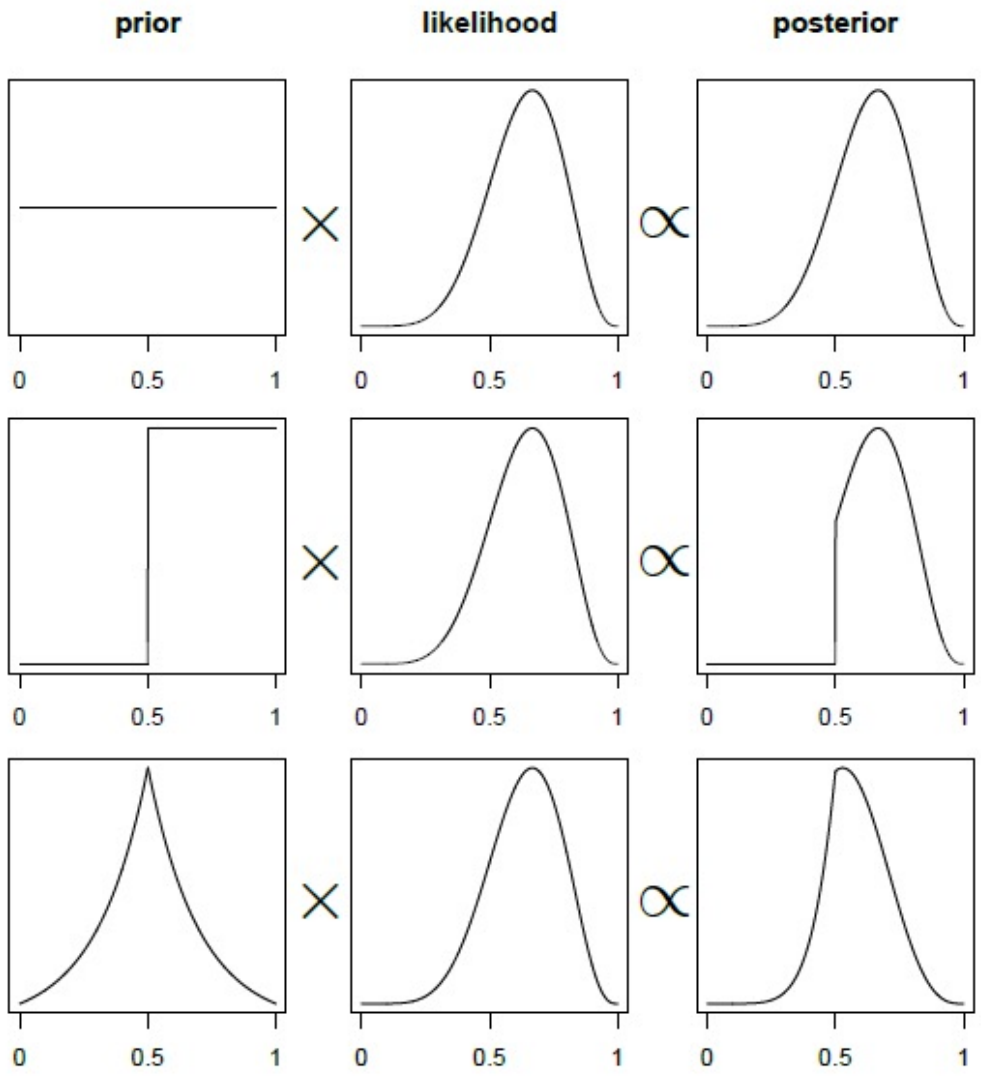
… and that is not necessarily easy!

A word about the evidence

- Also called “average likelihood”, “marginal likelihood”, “prior predictive” or “probability of the data”
- Expresses the probability of the data across all parameter values weighted by the prior (“average likelihood of the data averaged over the prior”)
- Standardizes the posterior so that it adds (integrates) up to one
- Typically, we don’t need it to estimate the posterior and we write just

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Posterior is a function of the prior and the likelihood



So how do we estimate the posterior?

- Other than analytical mathematics (always unattractive for vets, and sometimes non-computable depending on the model) there are several numerical techniques, such as:
  - Grid approximation
  - Quadratic approximation
  - **Markov Chain Monte Carlo (beloved MCMC) → several tools developed in the last decades that allow its easy implementation, e.g.**
    - **Bayesian inference Using Gibbs Sampling: BUGS (1997) → WinBUGS, OpenBUGS**
    - **Just Another Gibbs Sampler: JAGS (2007)**
    - **Stan (based on Hamiltonian/hybrid Monte Carlo)**

Nowadays

- BLCMs widely applied for the evaluation of (veterinary) diagnostic tests: endorsed by the WOAH

2.2.2. Samples from animals of unknown status

When the so-called reference standard is imperfect, which is the rule with any diagnostic tests, estimates of DSe and DSp for the candidate assay based on this standard will be flawed. A way to overcome this problem is to perform a latent class analysis of the joint results of the two tests assuming neither test is perfect.

Latent-class models do not rely on the assumption of a perfect reference test but rather estimate the accuracy of the candidate test and the reference standard with the joint test results (Branscum et al., 2005; Enøe et al., 2000; Georgiadis et al., 2003; Hui & Walter, 1980). If a Bayesian latent class analysis is used, prior knowledge about the performance of the reference test and the candidate test can be incorporated into the analysis.

Because these statistical models are complex and require critical assumptions, statistical assistance should be sought to help guide the analysis and describe the sampling from the target population(s), the characteristics of other tests included in the analysis, the appropriate choice of model and the estimation methods based on peer-reviewed literature (see *Terrestrial Manual* Chapter 3.6.5 [footnote [14]] for details).

Nowadays

- Guidelines available for adequate reporting of the use of BLCMs in the context of diagnostic test evaluation

Preventive Veterinary Medicine 138 (2017) 37–47

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models

Polychronis Kostoulas[a,*], Søren S. Nielsen[b], Adam J. Branscum[c], Wesley O. Johnson[d], Nandini Dendukuri[e], Navneet K. Dhand[f], Nils Toft[g], Ian A. Gardner[h]

[a] Laboratory of Epidemiology, Biostatistics and Animal Health Economics, Faculty of Veterinary Medicine, University of Thessaly, Karditsa GR43100, Greece
[b] Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark
[c] Biostatistics Program, Oregon State University, Corvallis, OR, 97331, USA
[d] Department of Statistics, University of California, Irvine, CA, 92697, USA
[e] McGill University Health Centre, McGill University, Montréal, QC, Canada
[f] Faculty of Veterinary Science, The University of Sydney, 425 Werombi Road, Camden, 2570 NSW, Australia
[g] Technical University of Denmark, National Veterinary Institute, Bülowsvej 27, DK-1870 Frederiksberg C, Denmark
[h] Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, Prince Edward Island C1A4P3, Canada

Nowadays

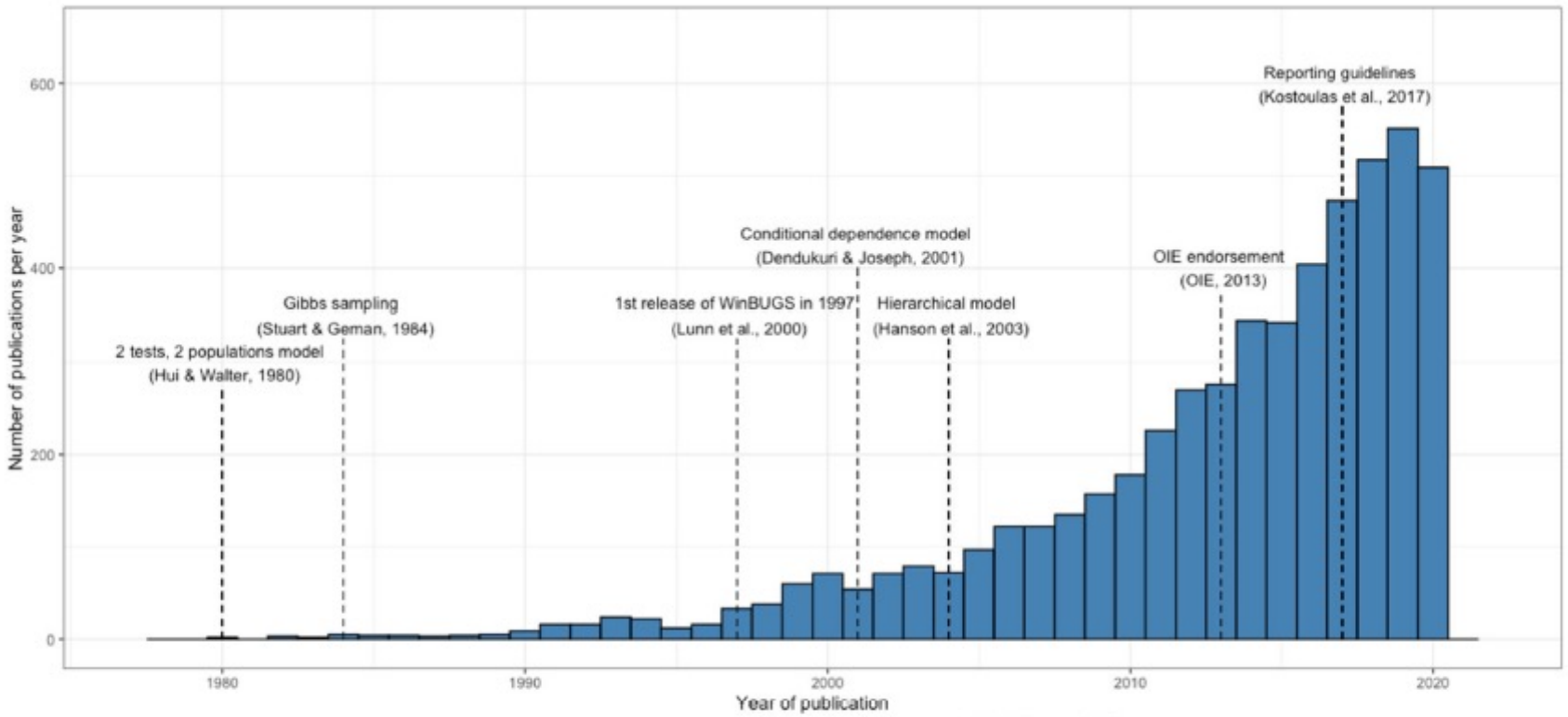
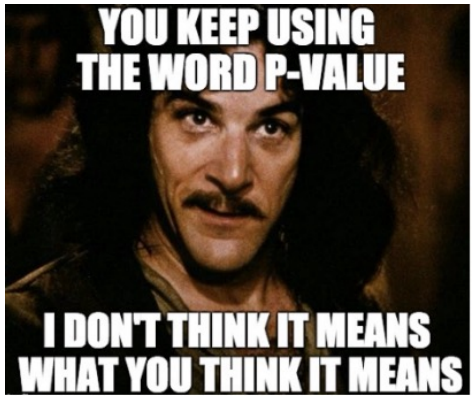


Fig. 2

Frequency histogram of the number of peer-reviewed articles published on latent class analysis when there is an imperfect reference test

Frequentist or Bayesian?

- Criticism to Bayesian approaches
  - Overreliance on priors → cheating!
  - Lack of robustness when data and prior conflict
- Criticism to frequentist approaches
  - Inefficiency/inflexibility/lack of realism → multiple dependences in spite of not using priors!
  - 99.9% there is prior information available… why not using it?
  - Less intuitive: “Bayesian interpretation”



References

- Hui, S. L., & Walter, S. D. (1980). Estimating the error rates of diagnostic tests. Biometrics, 167-171.
- Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 959-968.The effect of conditional dependence on the evaluation of diagnostic tests
- Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American journal of epidemiology, 141(3), 263-272.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. J. (2002). WinBUGS: Bayesian Inference Using Gibbs Sampling Manual, Version 1.4. London: Imperial College; Cambridge, UK: MRC Biostatistics Unit.
- Enøe, C., Georgiadis, M. P., & Johnson, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Preventive veterinary medicine, 45(1-2), 61-81.
- Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, No. 125.10, pp. 1-10).
- Toft, N., Innocent, G. T., Gettinby, G., & Reid, S. W. (2007). Assessing the convergence of Markov Chain Monte Carlo methods: an example from evaluation of diagnostic tests in absence of a gold standard. Preventive veterinary medicine, 79(2-4), 244-256.
- Kostoulas, P., Nielsen, S. S., Branscum, A. J., Johnson, W. O., Dendukuri, N., Dhand, N. K., . . . & Gardner, I. A. (2017). STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models. Preventive veterinary medicine, 138, 37-47.