The slowly changing landscape of predictive modeling in biomedicine

Lara Lusa

Franziska Kappenberg, Willi Sauerbrei, Matthias Schmid and Jörg Rahnenführer

STRATOS INITIATIVE

No conflict of interest

Project of the STRATOS initiative on prediction modeling

- Acknowledgements:
  - Project initiated by Willi Sauerbrei, Matthias Schmid and Jörg Rahnenführer
  - Comments from Federico Ambrogi, Riccardo de Bin, Anne-Laure Boulesteix, Ben van Calster, Mitch Gail, Frank Harrell, Marianne Huebner

- Which are the important steps for the development of useful prediction models?
  - Diagnosis, prognosis and treatment selection
- Which are the advantages/weaknesses of machine learning and statistical models?
- In the last few years:
  - considerable interest in similar topics emerged from several groups, due to important changes in the prediction modeling landscape

How is predictive modelling in medicine changing?

Suggestions from the editorials

- Data are more complex and more machine learning methods are used

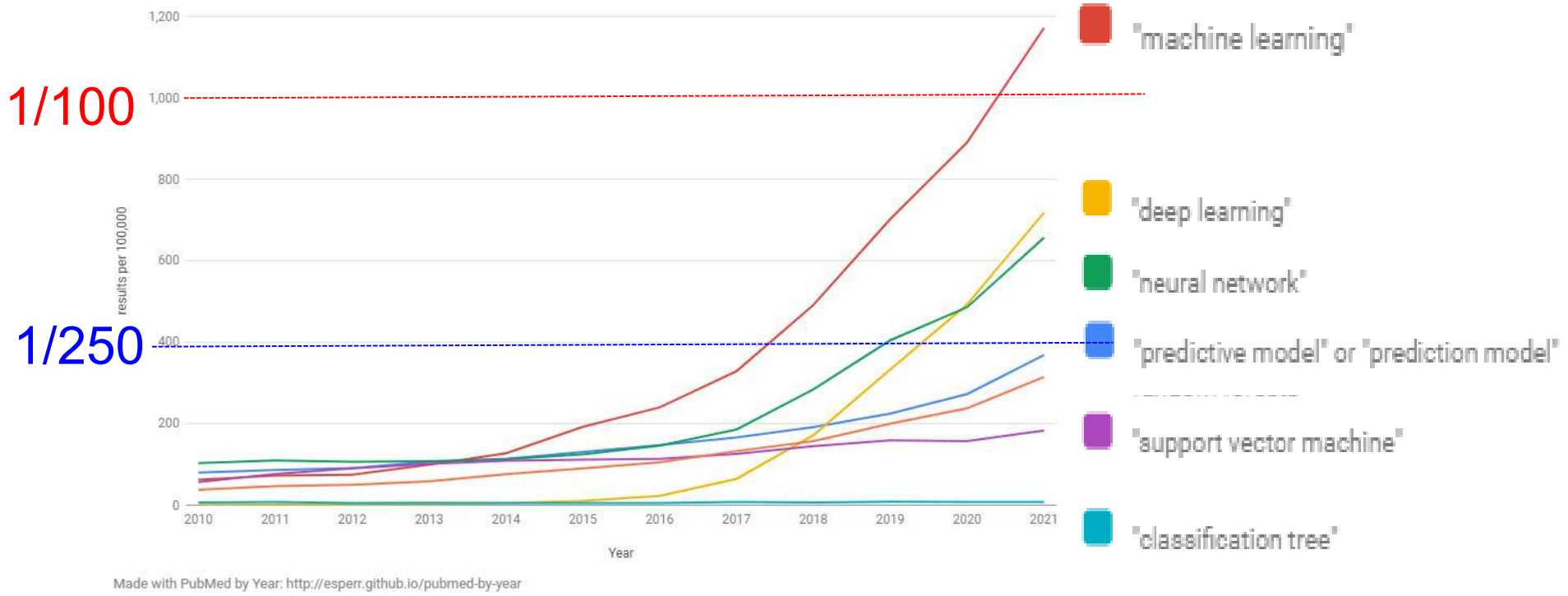
  "As clinical research catches up with other fields and finds itself immersed in the era of big data, the opportunity to apply more computational and data-driven techniques increases." Goldstein et al., 2018, Health Informatics

- The existing best practice recommendations from the traditional biostatistics and medical statistics literature are no longer sufficient to guide the use of predictive models.

  "while many best practice recommendations for design, conduct, analysis, reporting, impact assessment, and clinical implementation can be borrowed from the traditional biostatistics and medical statistics literature, they are not sufficient to guide the use of ML/AI in research." Vollmer et al., 2019, Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness

How is predictive modeling changing?

Results per 100,000 citations in Pubmed



The interest in predictive modeling in medicine is growing, and so is the use of machine learning methods

The “systematic review collectors”

Maarten van Smeden
@MaartenvSmeden

Small update to the prediction modeling landscape

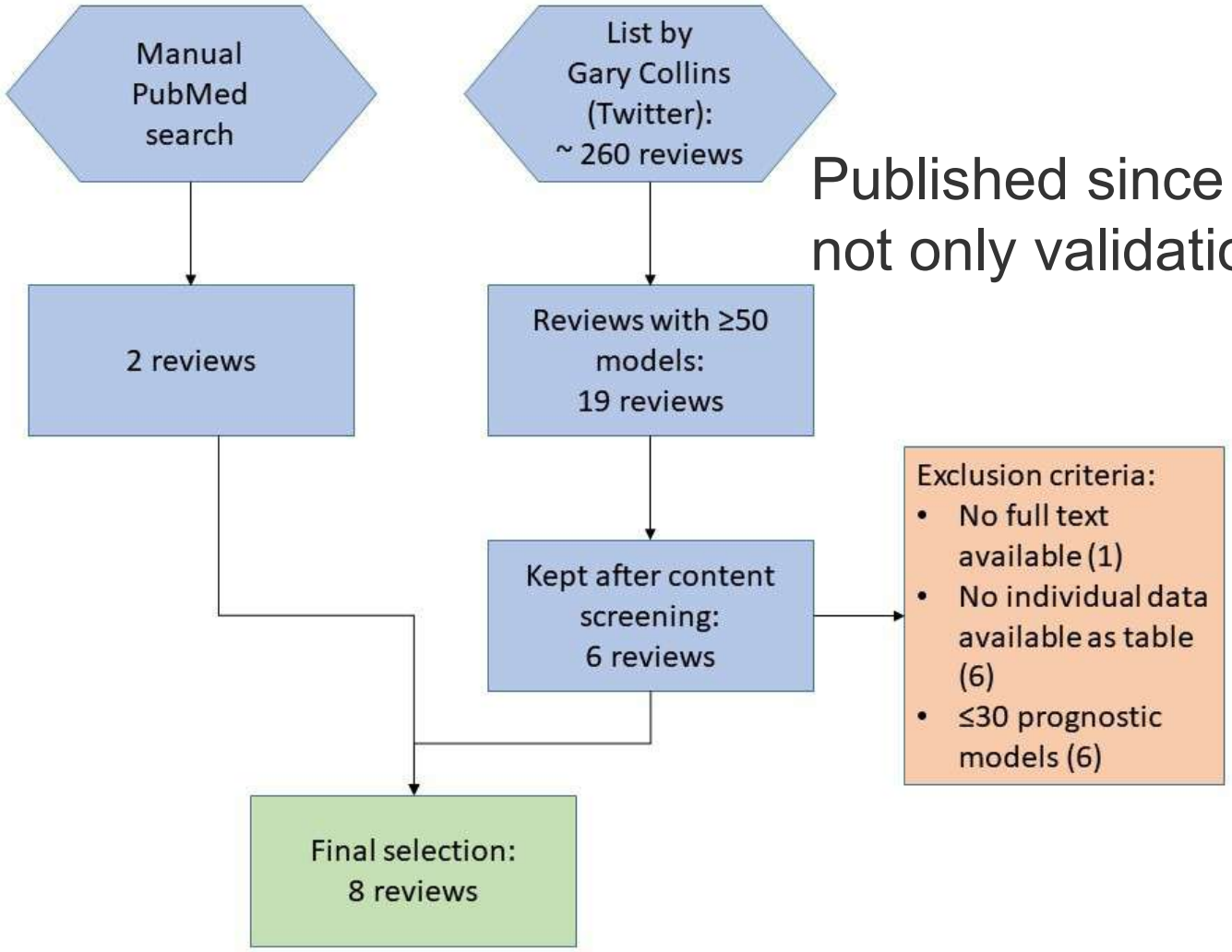
Traduci il Tweet

Gary Collins @GSCollins · 22 mar 2022

I've also recently been updating my list @MaartenvSmeden (for a talk). Below are ~260 systematic reviews of clinical prediction models - you might've missed a couple😉 - it's incomplete, and probably has some inaccuracies (still updating it).

Number of reviewed models	Number of systematic reviews since 2020
>100	7
50-100	13
25-49	22
10-24	33
<10	7

Selection of reviews



Published since 2020, not only validation (-1)

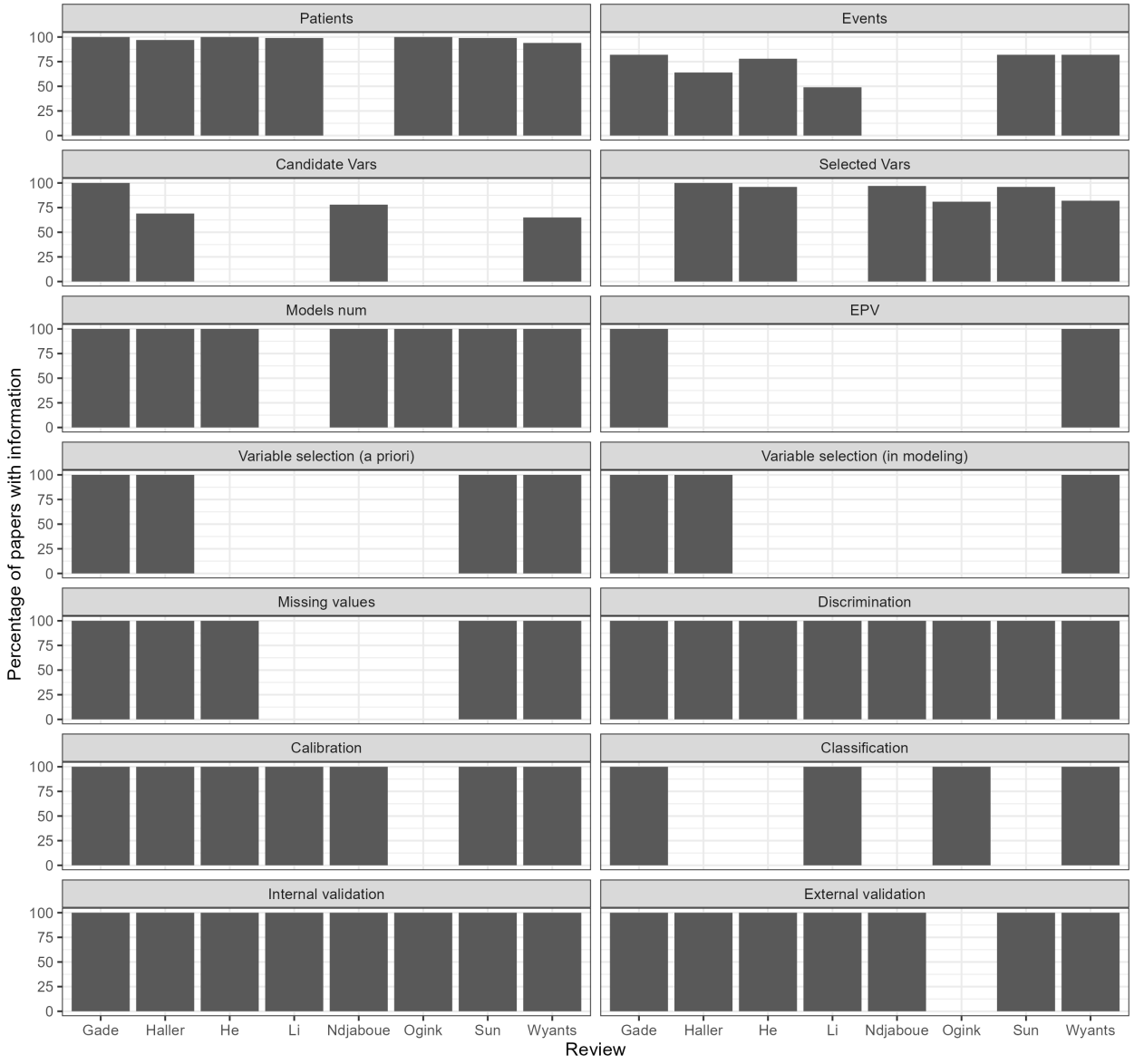
887 papers and 1448 models (after excluding pure validations and studies published before 2005)

Selected reviews

STRATOS INITIATIVE

Review	Population	Index model	Outcome	Year of publication of the papers included in the review	Models	Papers
Wynants et al.	Patient with confirmed COVID-19	All available prognostic models	All outcomes	2020 to 2022	501	368
Li et al.	Patients with vascular conditions	Predictive models that use ML methods	All outcomes (included segmentation)	1991 to 2021	202+	202
Sun et al.	Patient with heart failure	All available prognostic models from 2011	All-cause mortality or all-cause readmission of HF patients	2011 to 2021	176	78
Ndjaboue et al.	People with pre-diabetes and any type of diabetes, except gestational diabetes	All available models for which there was reported internal and/or external validation	Diabetes-related health conditions (complications)	2000 to 2020	175	75
Ogink et al	Surgical orthopaedic population	Prognostic models from studies that included at least one ML-based prediction	Orthopaedic surgical outcomes	1996 to 2020	218	56
He et al.	Patients diagnosed with cervical cancer	All available models	Clinical outcome (recurrence, metastasis, death, etc.)	1987 to 2020	74	52
Haller et al.	Recipients or donors in living organ transplantation	All available models	Any outcome occurring after transplantation donation in the recipient or donor	2004 to 2021	48	35
Gade et al.	Community-dwelling older adults (60+) of the general	All available models	Falls	1994 to 2019	54	21

Completeness of reviews: Information available at paper level

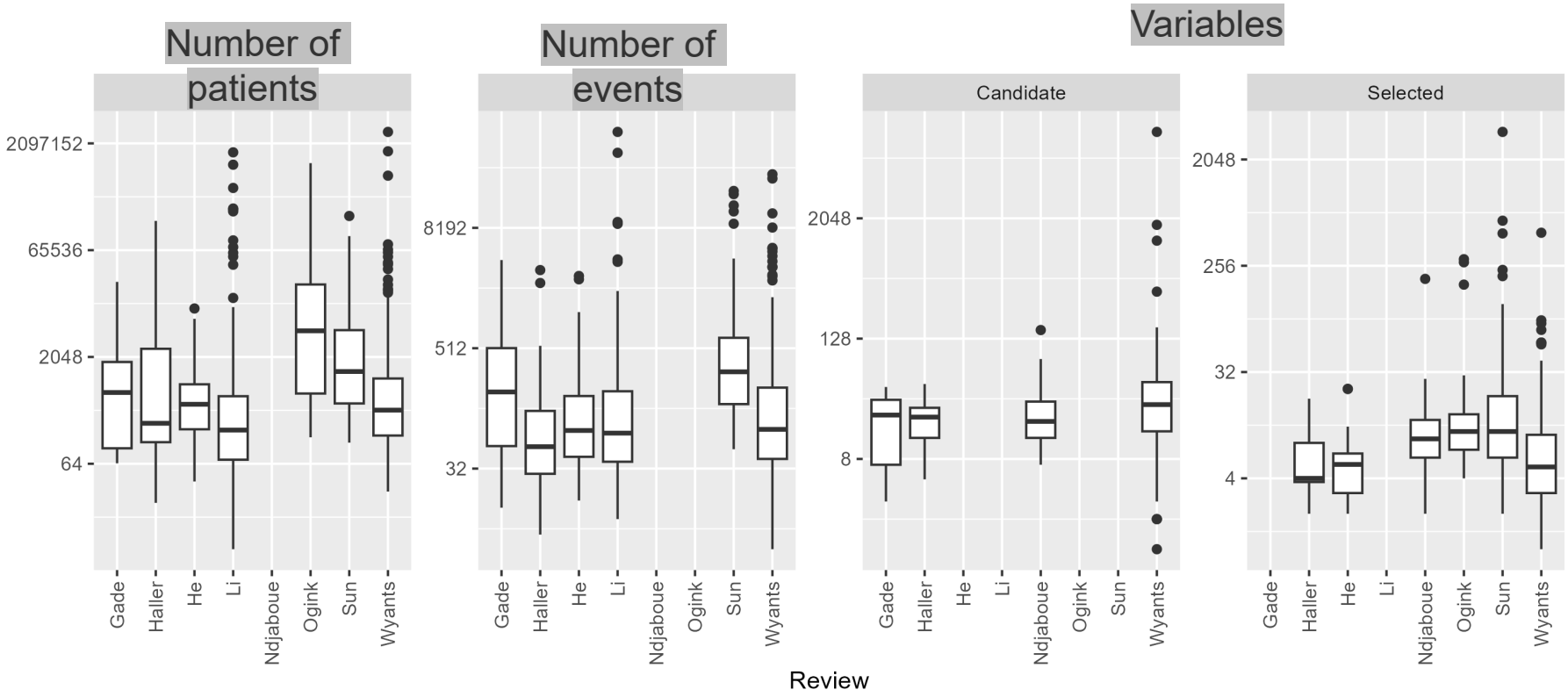


Some information is missing systematically in some reviews

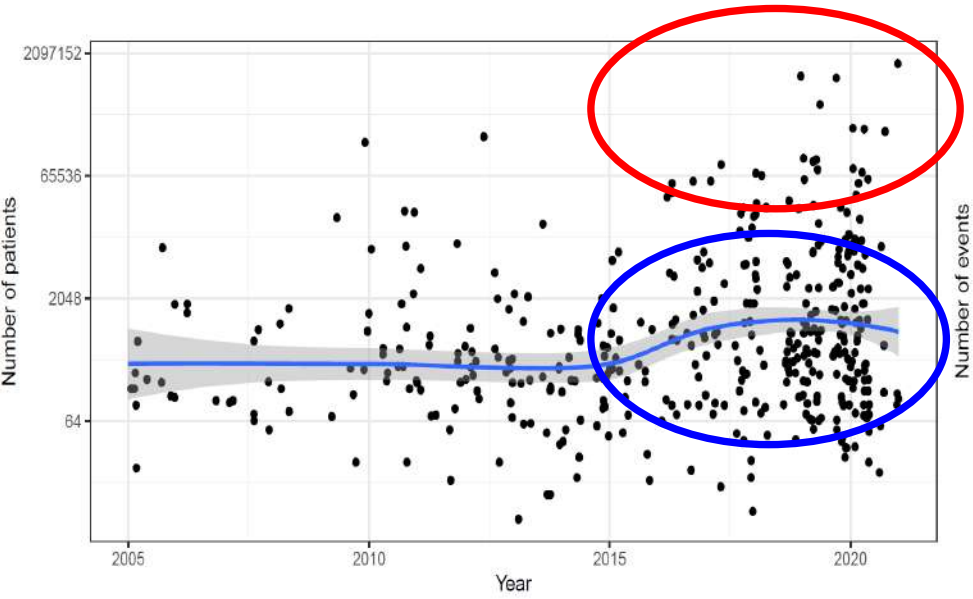
Number of events and of variables can be difficult to retrieve

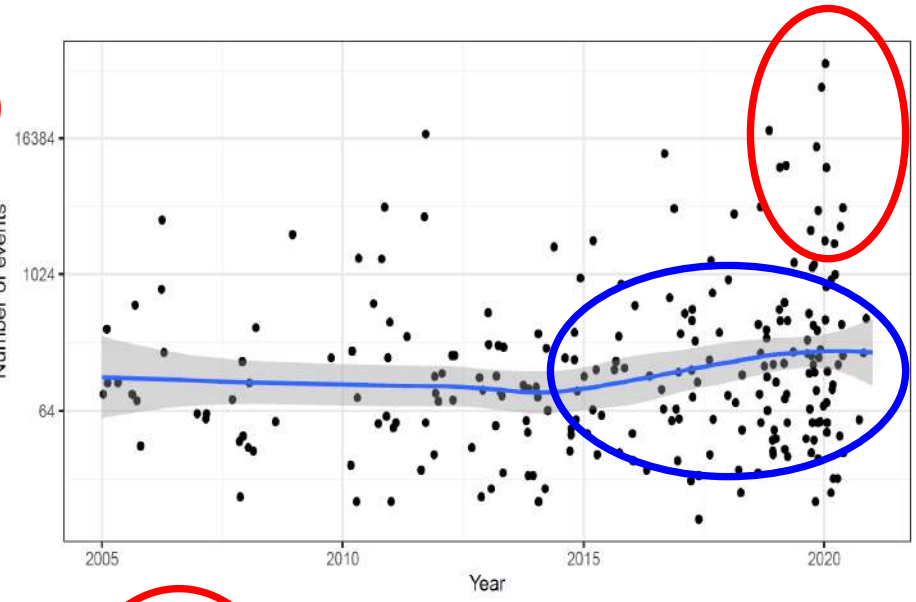
EPV is problematic

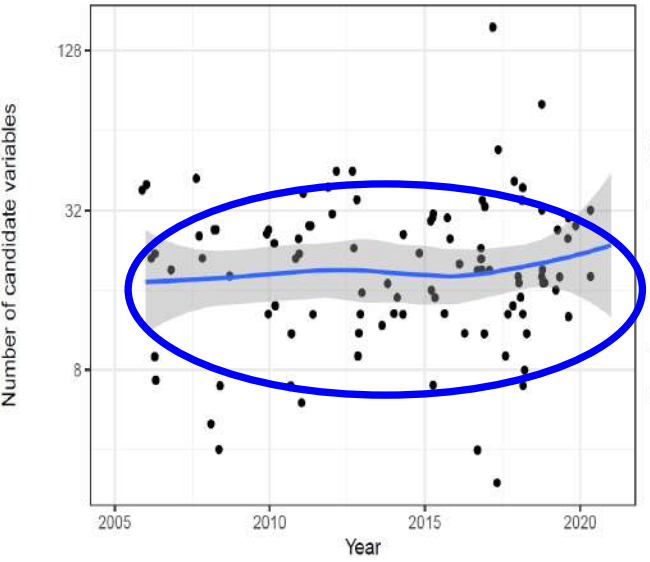
The reviews are heterogenous

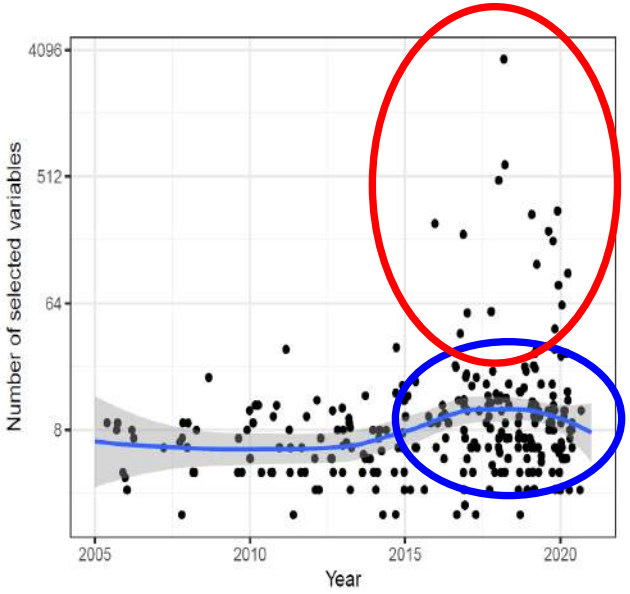


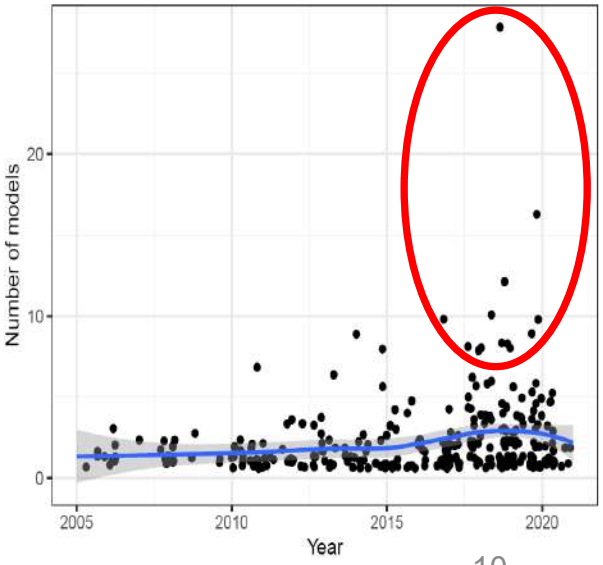
Main findings – time trends (without COVID-19)







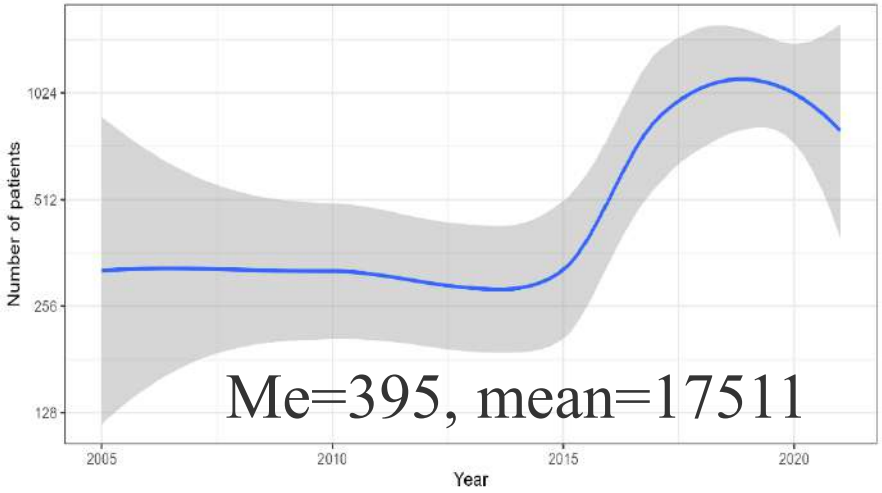


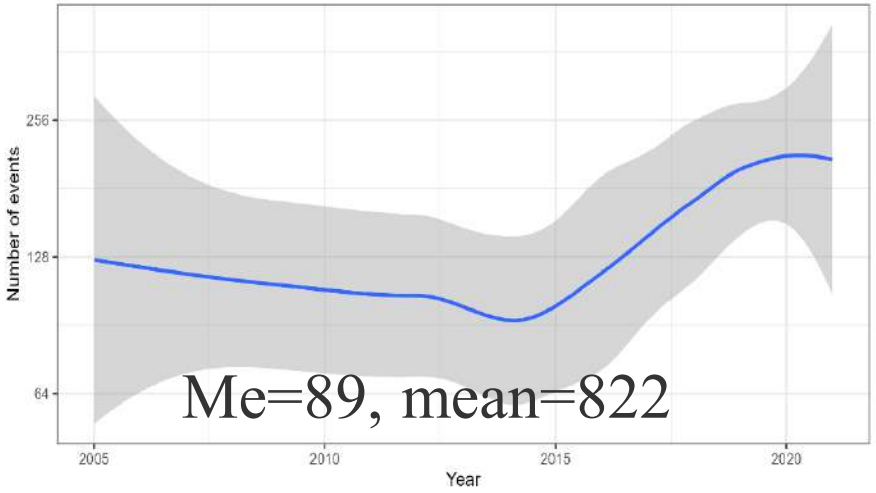


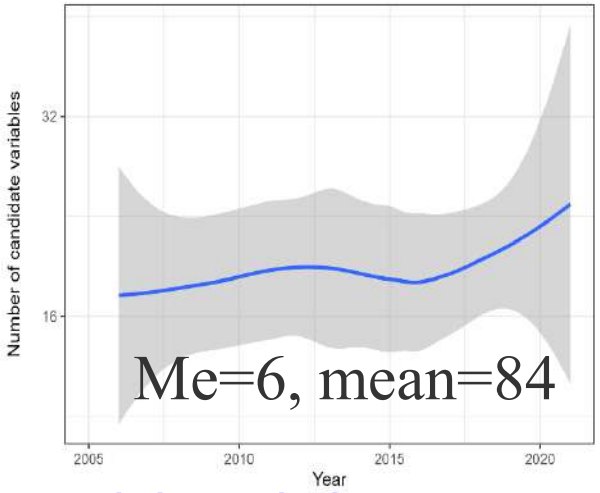
From 2005, graphs are mostly on log-2 scale on the y-axes (positively asymmetric distributions)

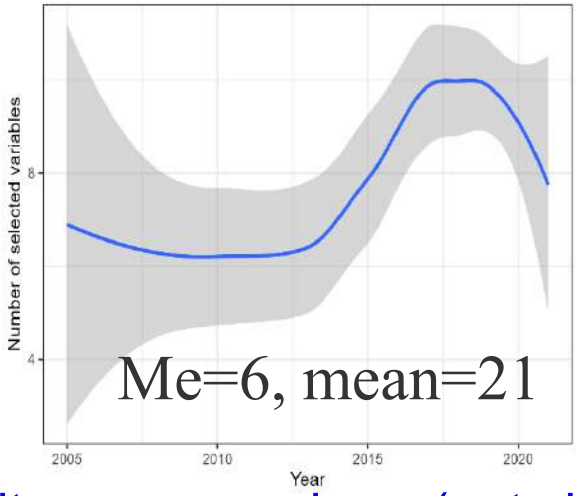
Zooming in – trends only

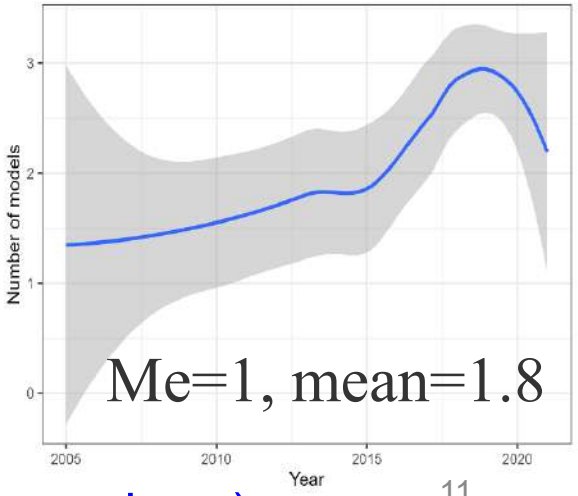
Data are bigger, more variables are used and more models are fitted
BUT the changes are smaller than might have been expected





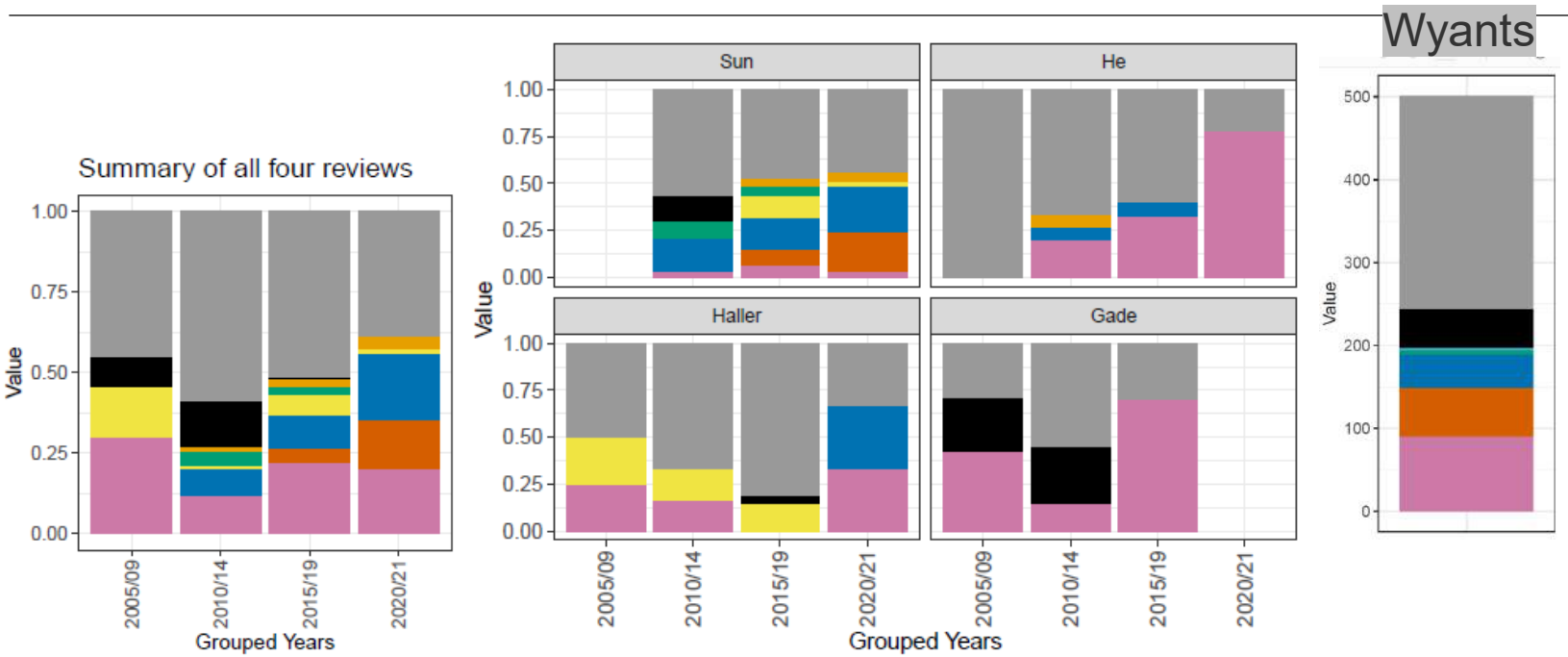


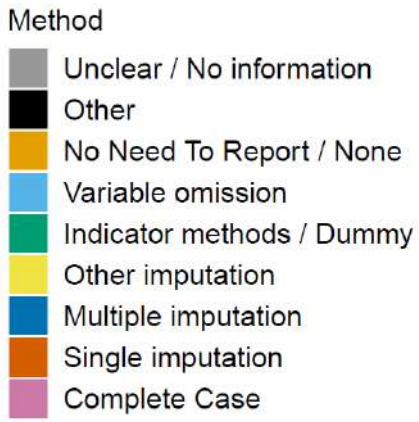




… and there is heterogeneity across reviews (not shown here)

Missing values





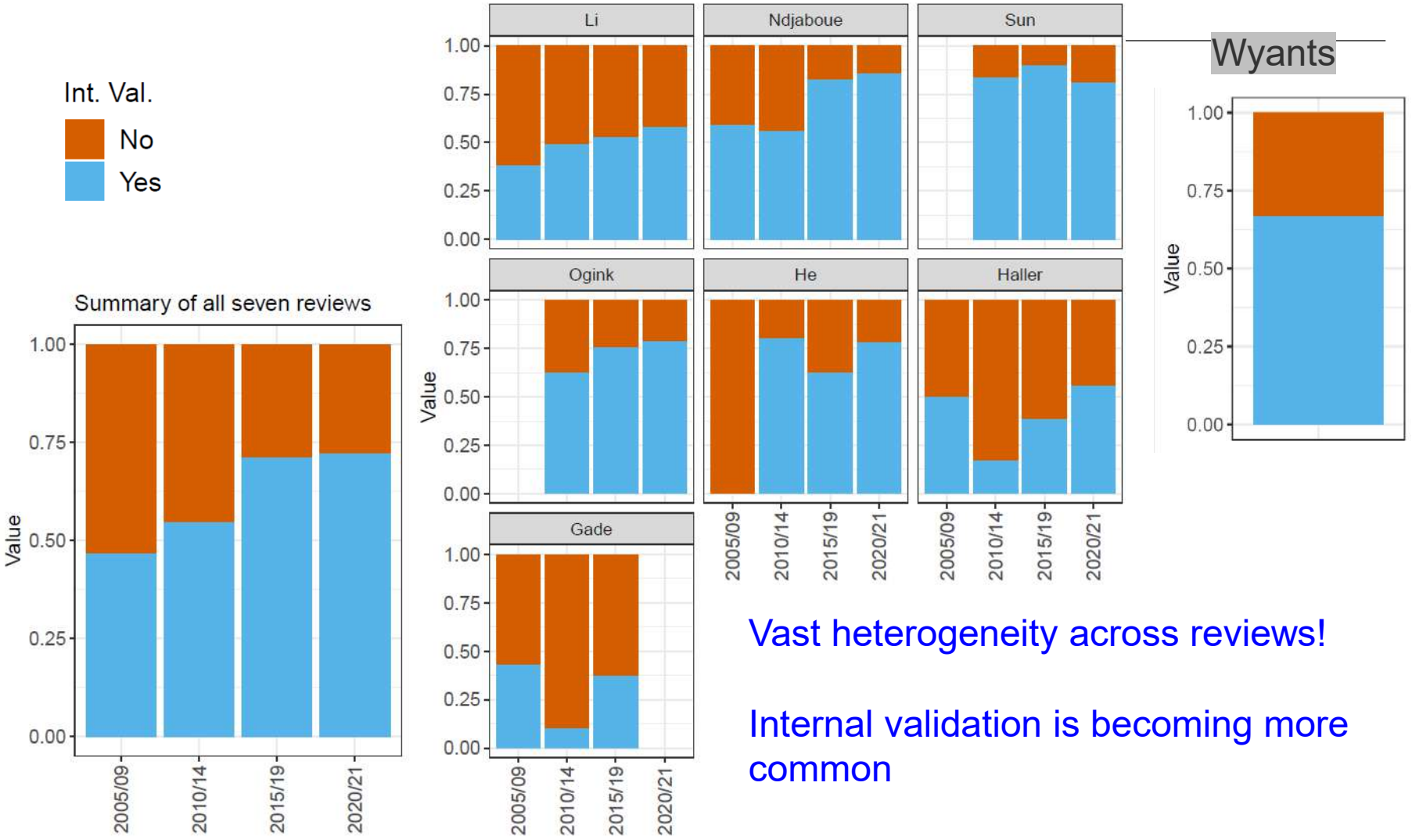
3 reviews ignore the information

Information about missing data is still rarely reported in papers
(gray, 54% of papers with no information)

Imputation methods (blue/red) are becoming more common
Complete case analysis is still the most common choice

The quality of reporting did not improve substantially in time

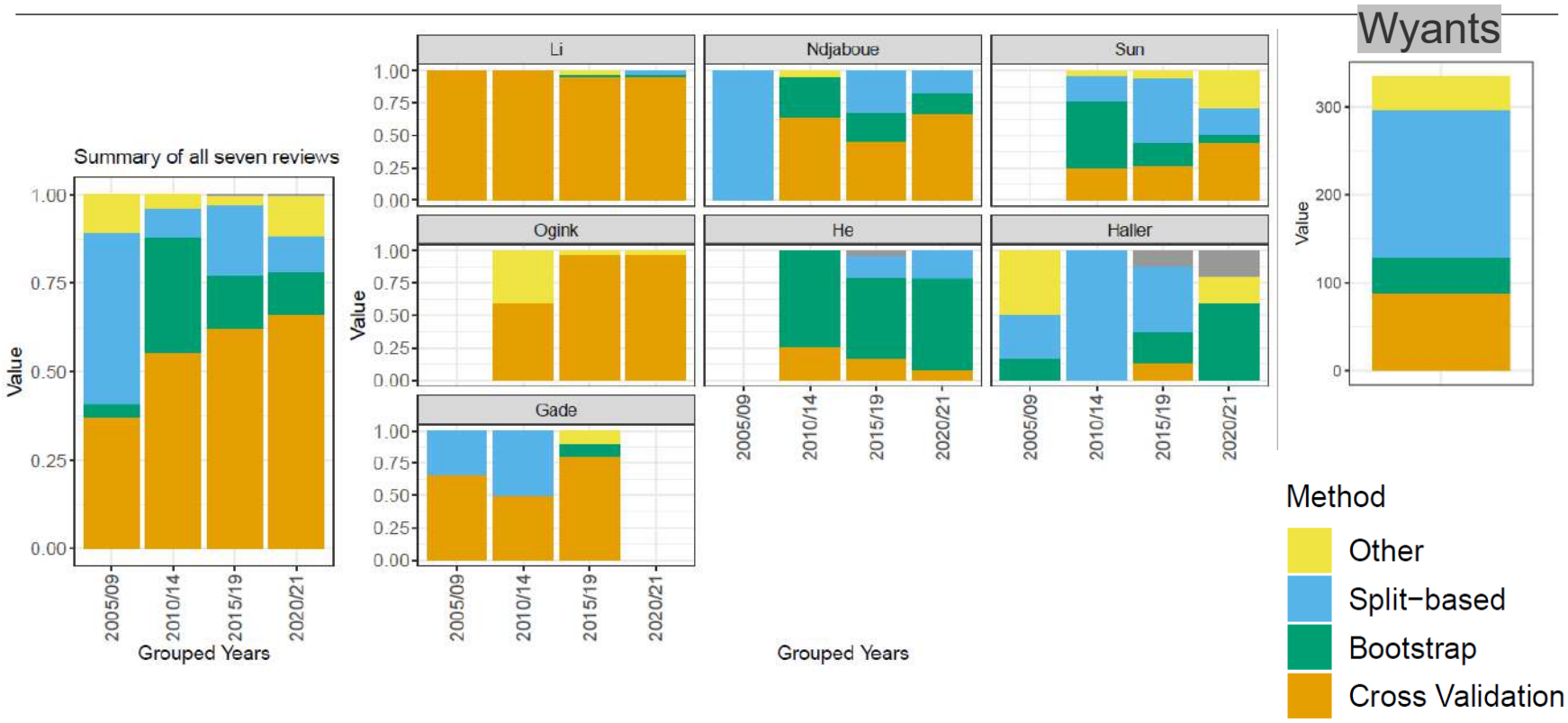
Internal validation



Vast heterogeneity across reviews!

Internal validation is becoming more common

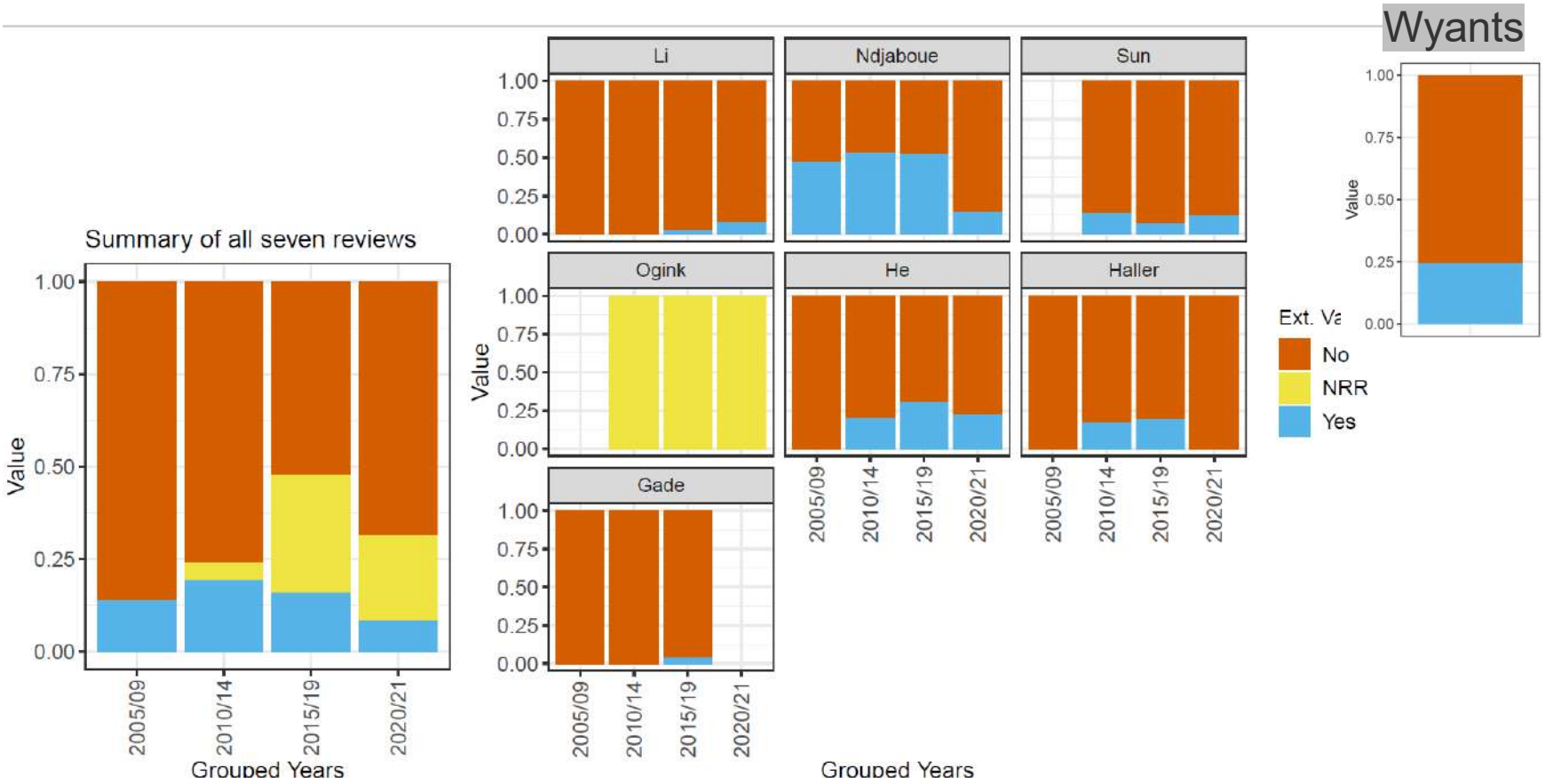
Internal validation – if performed, which?



Vast heterogeneity across reviews!

Cross-validation is gaining popularity, split-based methods are common only in some reviews

External validation



External validation is still uncommon (might be performed in subsequent papers)
More common in the review from Njdaboue, due to inclusion criteria

Time trends in reporting

Year	Internal val.	External val.	Discrimination	Classification	Calibration
2005/09	21/46 (46%)	8/46 (17%)	20/46 (43%)	20/25 (80%)	12/46 (26%)
2010/14	82/148 (55%)	27/141 (19%)	98/148 (66%)	47/72 (65%)	47/141 (33%)
2015/19	142/200 (71%)	26/167 (16%)	149/200 (74%)	62/100 (62%)	50/167 (30%)
2020/21	82/125 (66%)	14/109 (13%)	87/125 (70%)	42/82 (51%)	30/109 (28%)
COVID-19	248/368 (67%)	95/368 (26%)	209/368 (57%)	114/368 (31%)	82/368 (22%)

Increase of reporting of internal validation and discrimination measures, poor reporting of calibration

Type of model: ML vs statistical methods

Vast heterogeneity across reviews!

Li (using ML models) and Ogrink (at least one ML model) selected models based on the use of ML – but many of their models were classified (by us) as statistical

Time changes are not visible within reviews (the overall summary are influenced by weight of each review)

Comparison of statistical and ML models

	n	Me	Mean	Range	IQR
Number of patients					
Statistical	381	421	11,891	4 to 1,621,149	160 to 1475
ML	260	347	19,753	8 to 1,567,636	130 to 1071
Number of events					
Statistical	292	84	591	7 to 28,140	41 to 288
ML	166	95	689	10 to 46,163	48 to 214
Number of candidate variables					
Statistical	225	23	33	1 to 1224	14 to 37
ML	68	33	289	7 to 15,000	24 to 49
Number of selected variables					
Statistical	370	6	12	1 to 488	4 to 10
ML	113	6	19	2 to 618	3 to 13
Number of models					
Statistical	400	1	2	1 to 10	1 to 2
ML	157	1	1	1 to 8	1 to 1

The size of the datasets was similar (more extremes for ML methods)

Many more candidate variables for ML, but similar number of selected

Summary statistics of predictive performance measures by type of model

	Discrimination	Calibration	Classification
Statistical	334/449 (74%)	171/441(39%)	96/255(38%)
ML	124/277 (45%)	24/261(9%)	129/262(49%)
Both	80/99 (81%)	13/69(19%)	41/78(53%)
Unclear	25/62 (40%)	13/60(22%)	19/52(37%)

Very poor reporting of calibration of ML models, poor for discrimination.

Conclusions

- Quantitative assessment of changes is important, but it is not straightforward
- Reviews
  - Not many that include many prediction models and have (complete relevant) publicly available data
  - wide heterogeneity across reviews in almost all the aspects
    - Example: COVID/Wynants vs the other 2020/21 papers
    - Truly reflecting differences in the fields or somehow related to the review process?
- The changes in predictive modeling are not as substantial as it might have been anticipated
  - Growth in the 2015-19 period in number of patients and variables, followed by a stabilization, the centers of the distributions are stable, extreme values are more common

Changes in predictive modeling

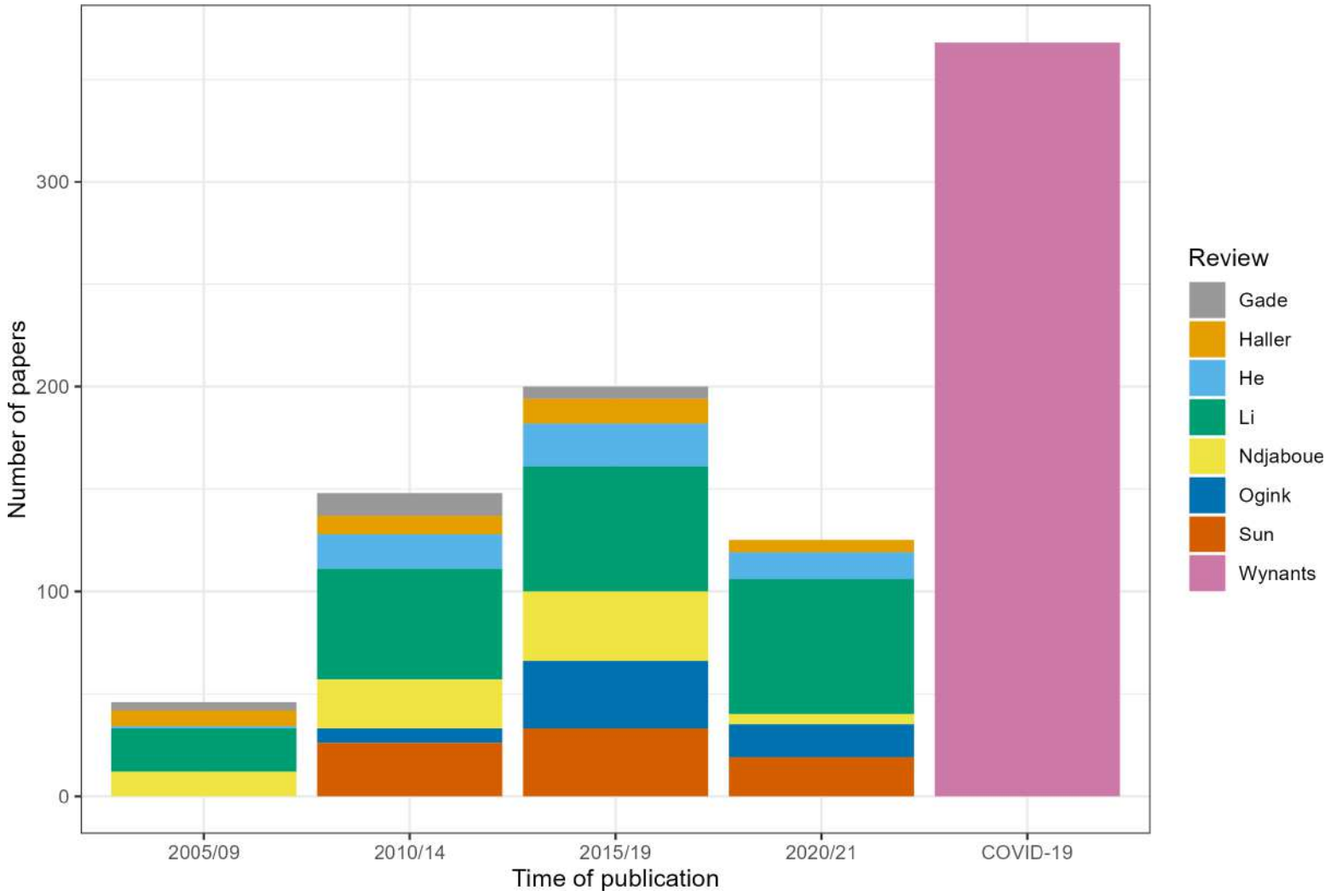
- Larger sample sizes (and number of events) and larger mean number of selected variables, more models per paper, but
  - the central part of the distributions are mostly stable in the 2020/21 period
  - similar median number of variables (candidate and selected)
    - Underreported pre-selection?
    - Under-representation of imaging models in our data?
    - Time lag makes the growth not yet observable?
    - Simpler models are still preferred (even if more candidates are available)?
- There might be a trend towards increasingly following guidelines
  - performing/reporting internal validation,
  - using resampling methods instead of sample splitting
  - reporting discrimination
  - more imputation methods for missing values (but still poor reporting for missing data)
- Not a very clear increase of the use of ML methods (with)in reviews and not striking differences in the characteristics of data being used

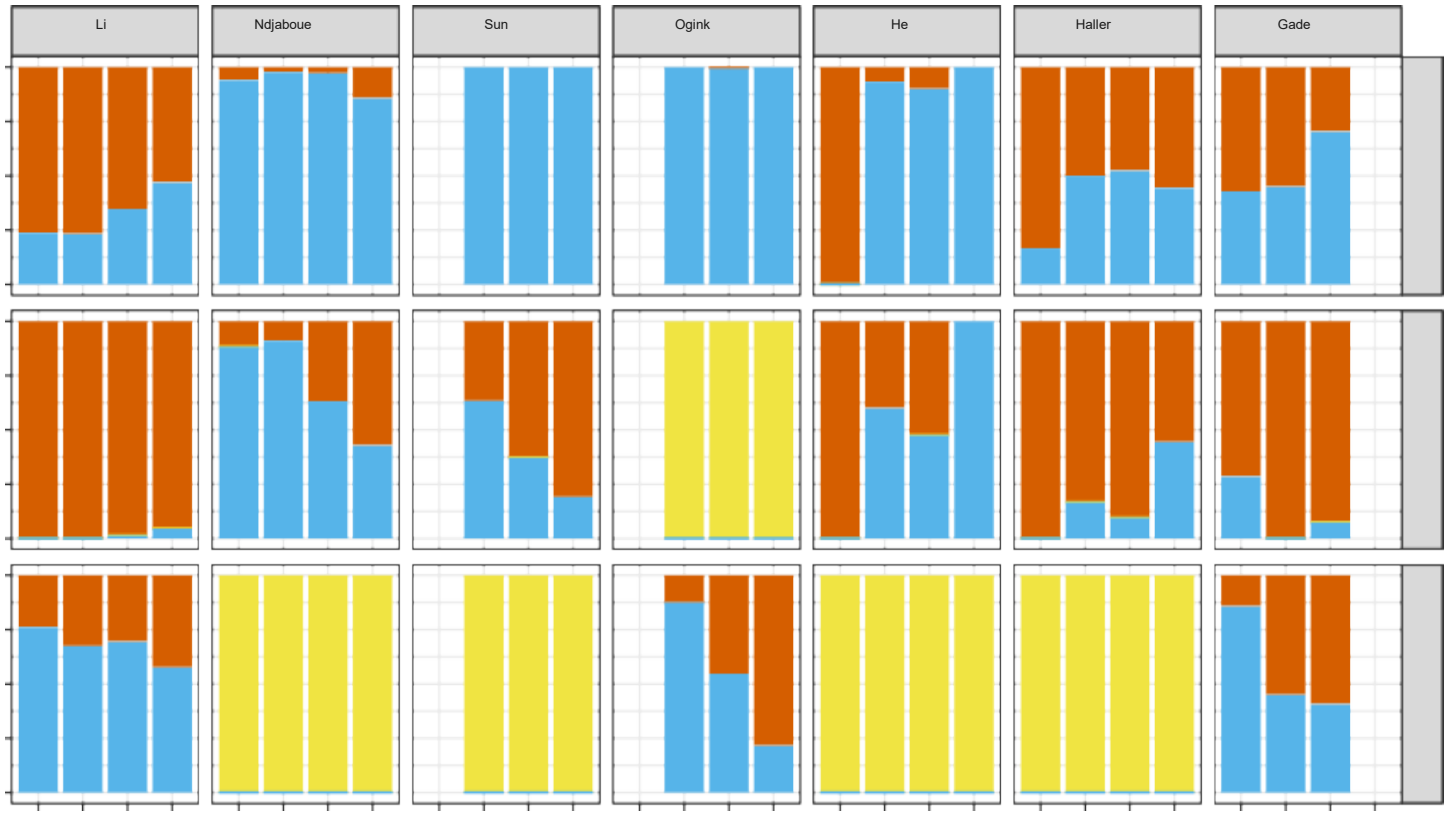
Beyond our draft…

- … and back to the original idea of the project
- ML vs statistical models
  - The distinction is difficult. Should we focus on model complexity?
  - Comparison of the performance -> reliable?
- Assessment of bias
  - Larger for ML/complex models
  - how reliable is the assessment?
- Need for new guidelines?
  - All the basic principles apply and are still not always used
  - Comprehensible understanding of methods needed to identify specificities related to “new models”

Selection of reviews

- Type of prognostic models
  - Multivariate prognostic models (more than 2 candidate predictors)
  - Developmental studies
- Type of review
  - Published in 2020 or after
  - Includes development models (not only validation)
  - Per paper/per model data are available (table format) for most of the information suggested in the CHARMS checklist
  - Includes at least 30 models/papers
  - Source: the lists of the “systematic review collectors” and additional PubMed search
- 8 selected reviews, including 841 papers and 1499 models


Number of papers
300
200
100
0
2005/09
2010/14
2015/19
2020/21
COVID-19
Time of publication
Review
Gade
Haller
He
Li
Ndjaboue
Ogink
Sun
Wynants


Li
Ndjaboue
Sun
Ogink
He
Haller
Gade