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 JR
  - 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

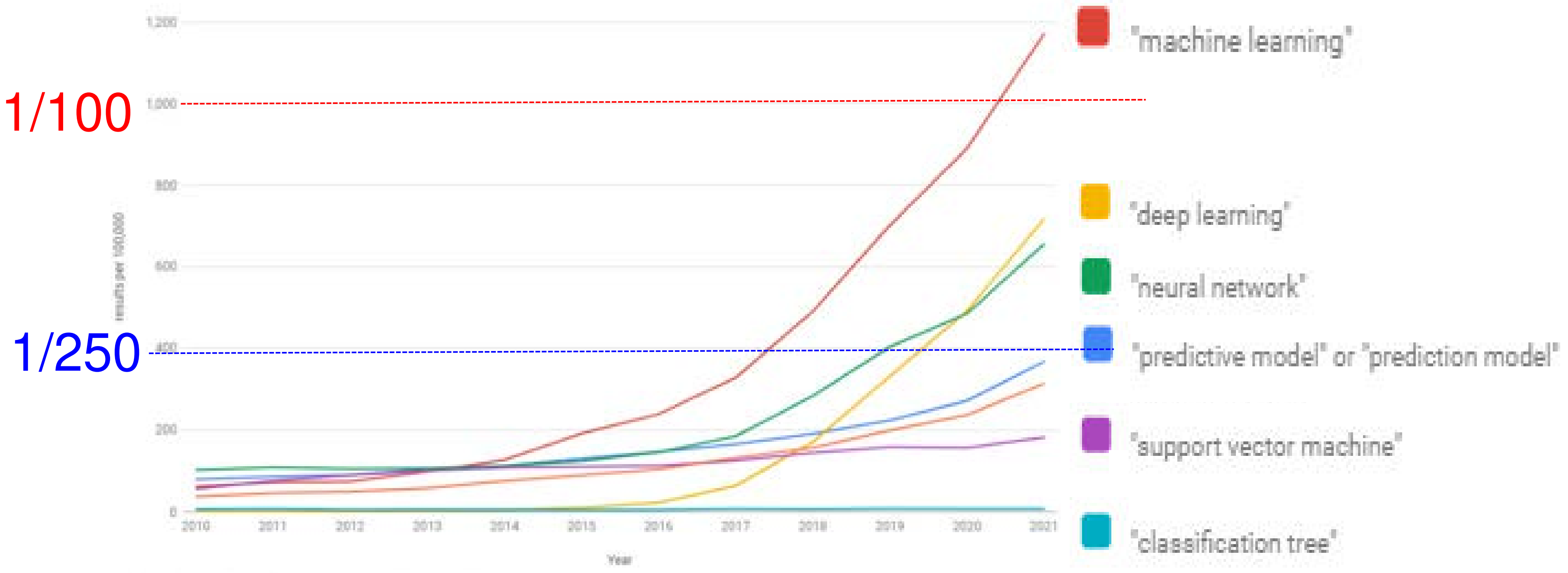
  “As clinical research catches up with other fields and finds itself immersed in the era of big data, the opportunity to apply more computational and data-driven techniques increases.” Goldstein et al., 2018, Health Informatics

- 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?

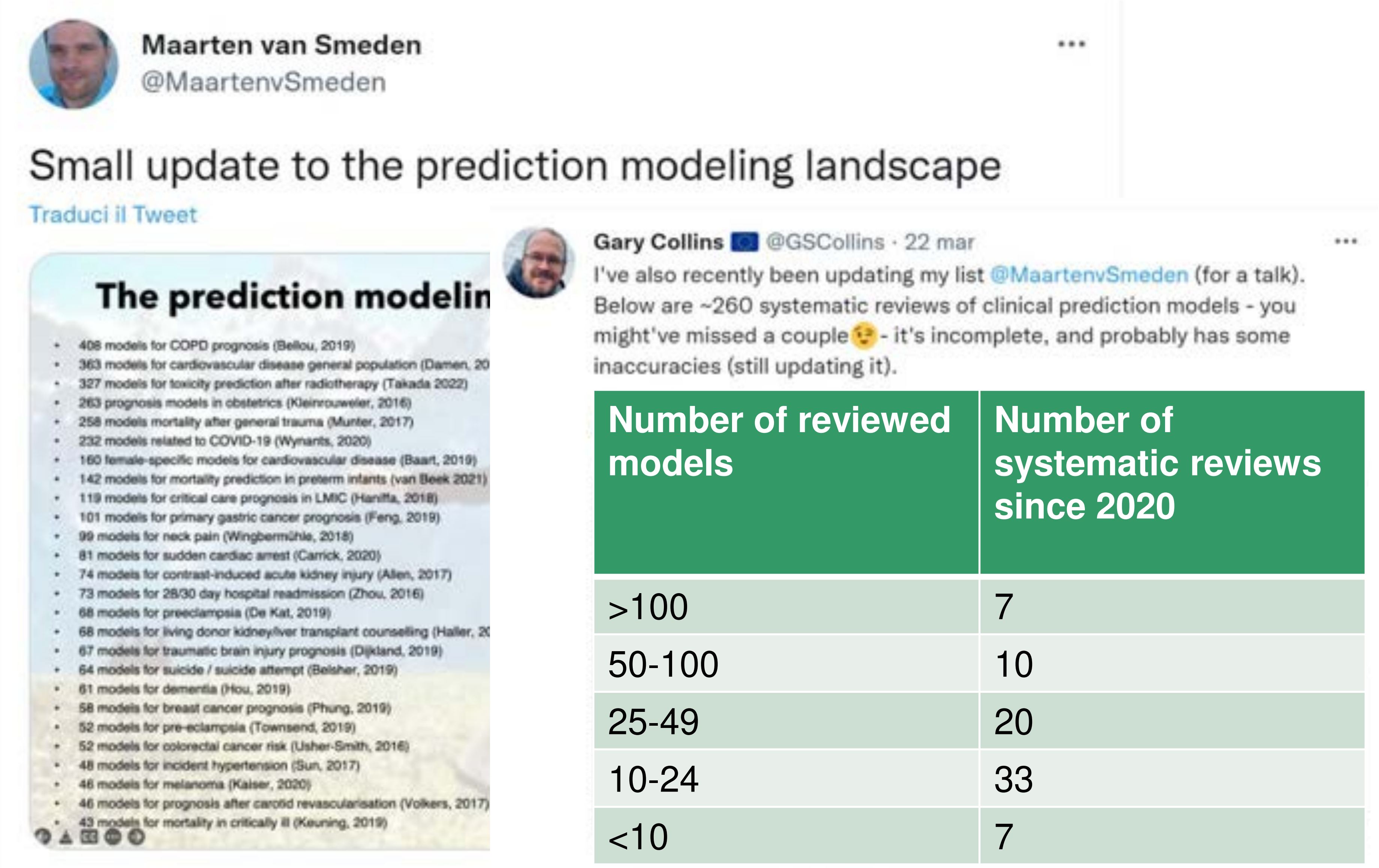
Results per 100,000 citations in Pubmed



Made with PubMed by Year: http://esperr.github.io/pubmed-by-year

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

The prediction modelin

- 408 models for COPD prognosis (Bellou, 2019)
- 363 models for cardiovascular disease general population (Damen, 20
- 327 models for toxicity prediction after radiotherapy (Takada 2022)
- 263 prognosis models in obstetrics (Kleinrouweler, 2016)
- 258 models mortality after general trauma (Munter, 2017)
- 232 models related to COVID-19 (Wynants, 2020)
- 160 female-specific models for cardiovascular disease (Baart, 2019)
- 142 models for mortality prediction in preterm infants (van Beek 2021)
- 119 models for critical care prognosis in LMIC (Haniffa, 2018)
- 101 models for primary gastric cancer prognosis (Feng, 2019)
- 99 models for neck pain (Wingbermühle, 2018)
- 81 models for sudden cardiac arrest (Carrick, 2020)
- 74 models for contrast-induced acute kidney injury (Allen, 2017)
- 73 models for 28/30 day hospital readmission (Zhou, 2016)
- 68 models for preeclampsia (De Kat, 2019)
- 68 models for living donor kidney/liver transplant counselling (Haller, 20
- 67 models for traumatic brain injury prognosis (Dijkland, 2019)
- 64 models for suicide / suicide attempt (Belsher, 2019)
- 61 models for dementia (Hou, 2019)
- 58 models for breast cancer prognosis (Phung, 2019)
- 52 models for pre-eclampsia (Townsend, 2019)
- 52 models for colorectal cancer risk (Usher-Smith, 2016)
- 48 models for incident hypertension (Sun, 2017)
- 46 models for melanoma (Kaiser, 2020)
- 46 models for prognosis after carotid revascularisation (Volkers, 2017)
- 43 models for mortality in critically ill (Keuning, 2019)

Gary Collins 🇪🇺 @GSCollins · 22 mar

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	10
25-49	20
10-24	33
<10	7

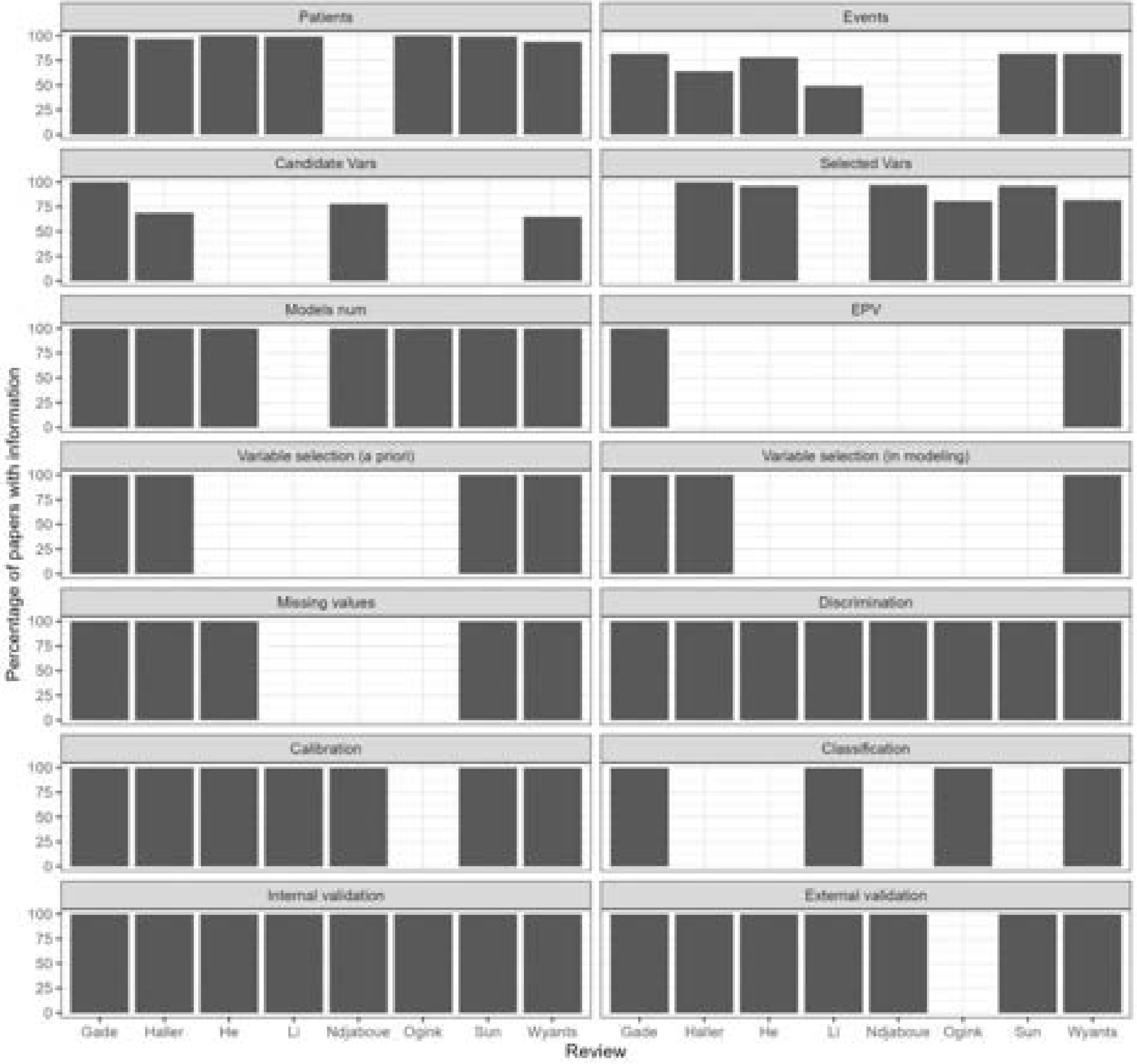
Selection of reviews

- 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

Selected reviews

Review	Population	Index model	Outcome	Year of publication of the papers included in the review	Models	Papers
Wyants 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	1991 to 2021	212	212
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	180	78
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
Ogink et al	Surgical orthopaedic population	Prognostic models from studies that included at least one ML-based prediction	Orthopaedic surgical outcomes	1996 to 2020	234	59
He et al.	Patients diagnosed with cervical cancer	All available models	Clinical outcome (recurrence, metastasis, death, etc.)	1987 to 2020	77	55
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	49	36
Gade et al.	Community-dwelling older adults (60+) of the general population	All available models	Falls	1994 to 2019	70	28

Completeness of reviews: Information available at paper level

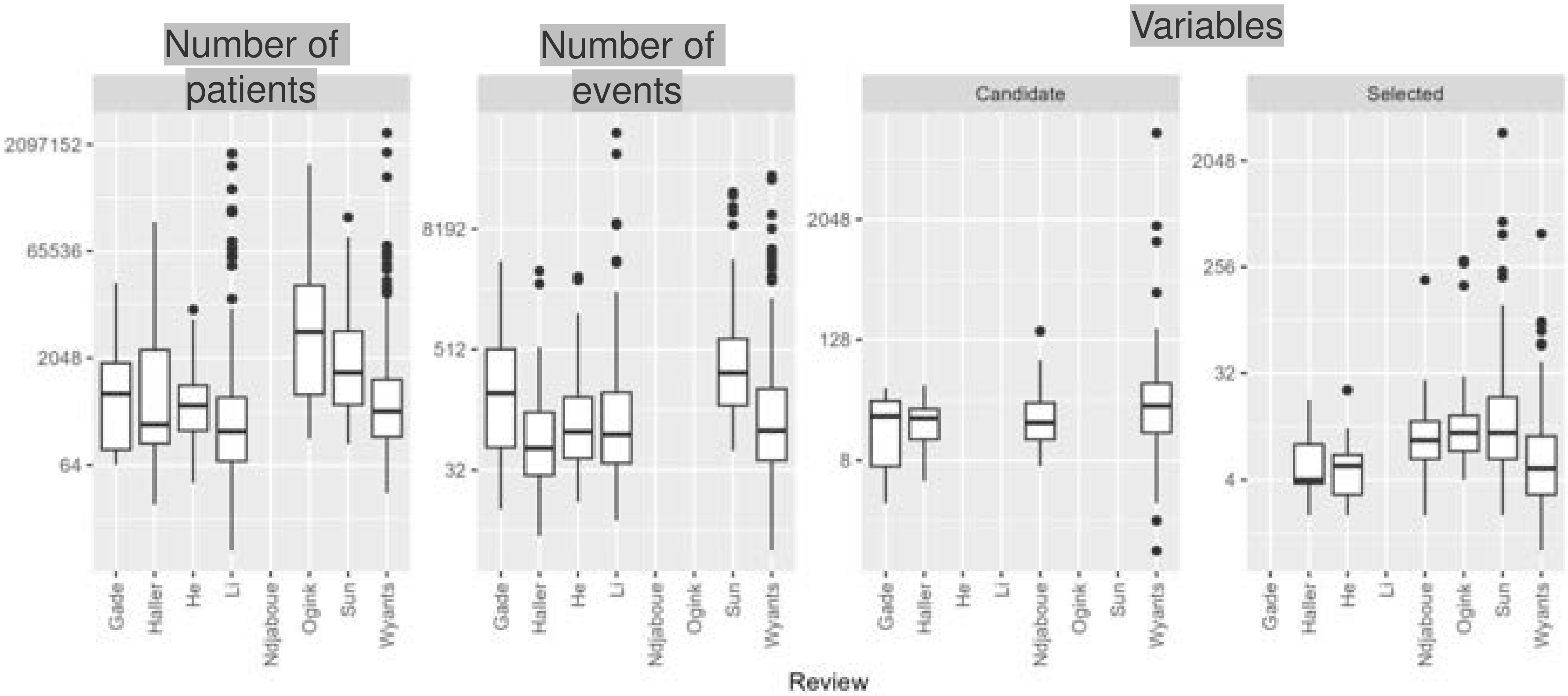


Some information is missing systematically in some reviews

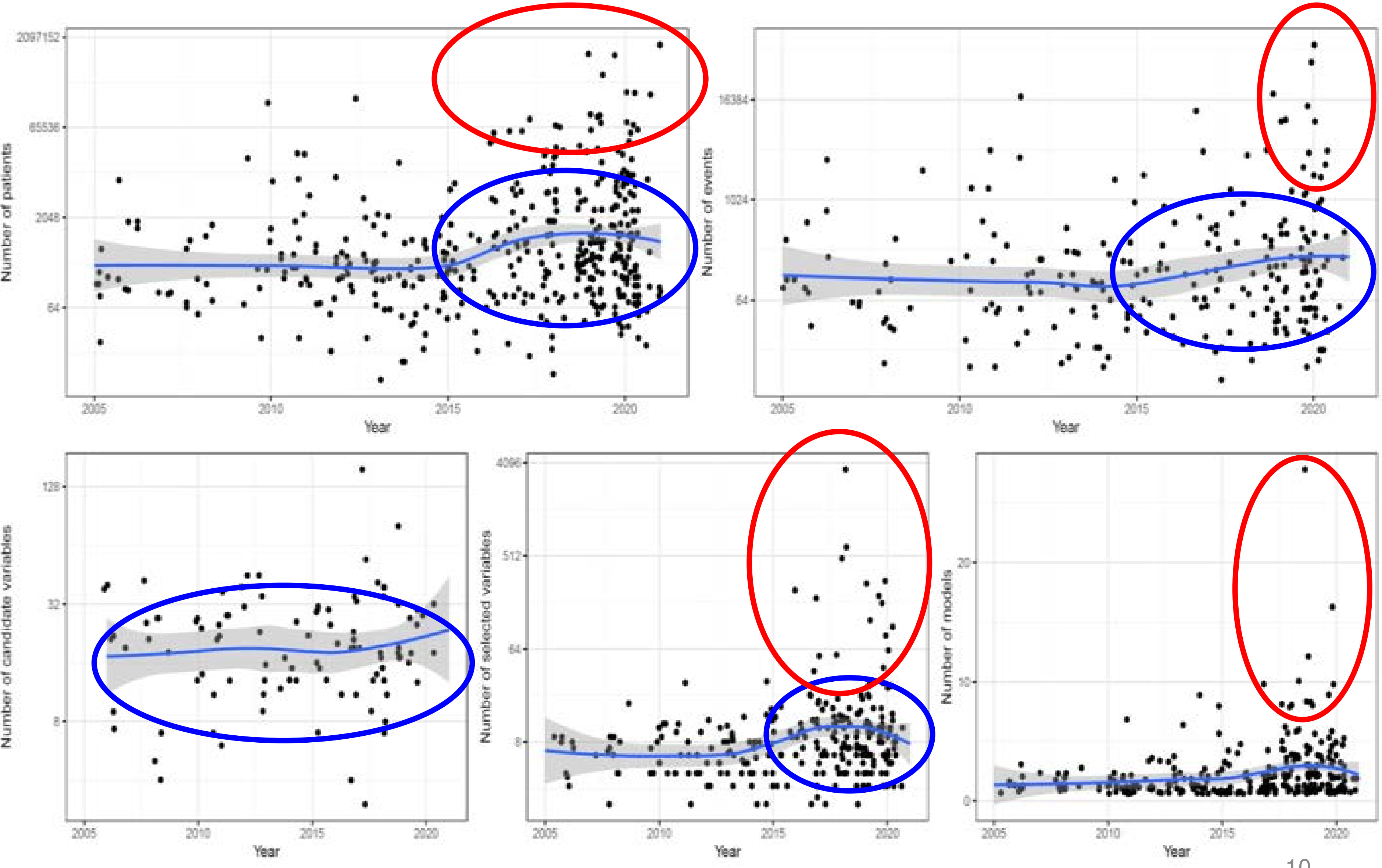
Number of events and of variables can be difficult to retrieve

EPV is problematic

The reviews are heterogenous



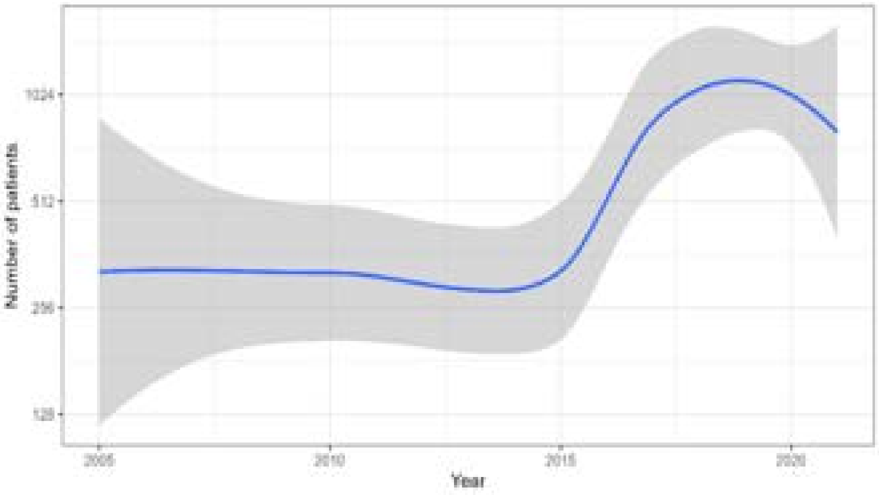
Main findings – time trends (without COVID-19)

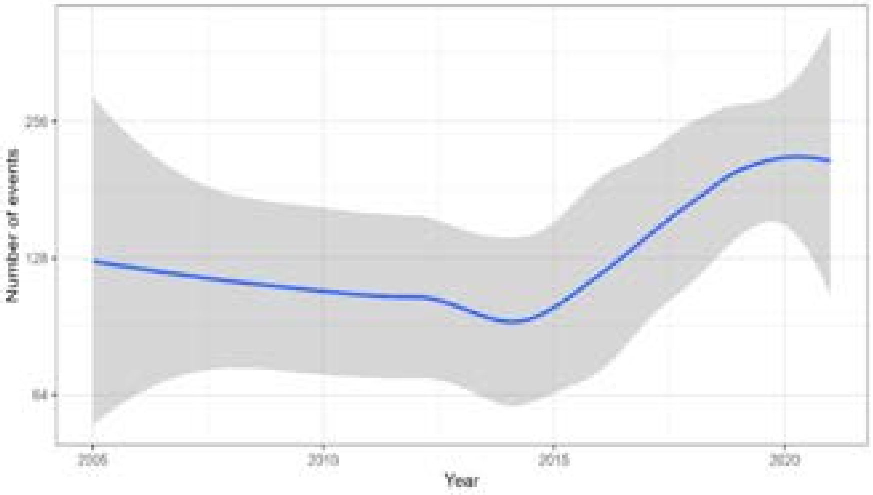


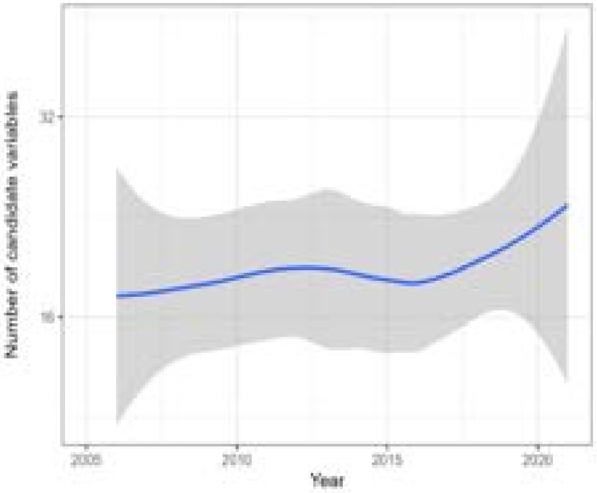
<u>From 2005</u>, graphs are mostly on log-2 scale on the y-axes (positively asymmetric distributions)

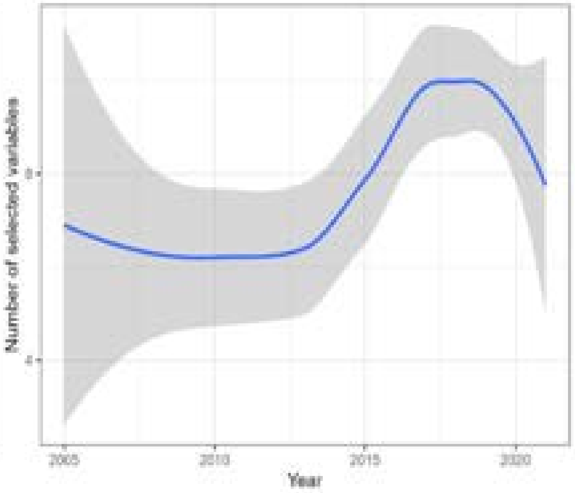
Zooming in – trends only

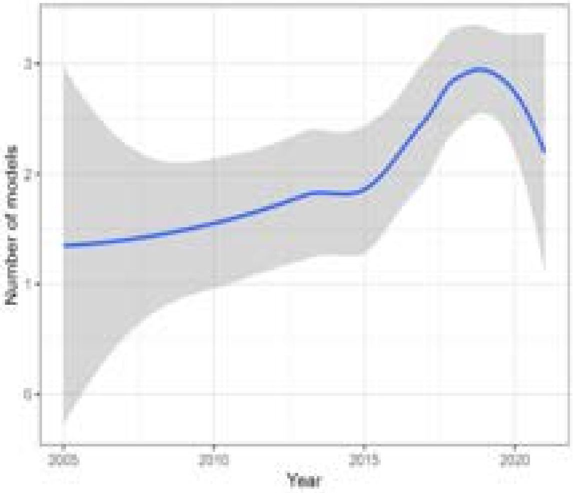
Data are bigger, more variables are used and more models are fitted
BUT the changes are smaller than might have been expected





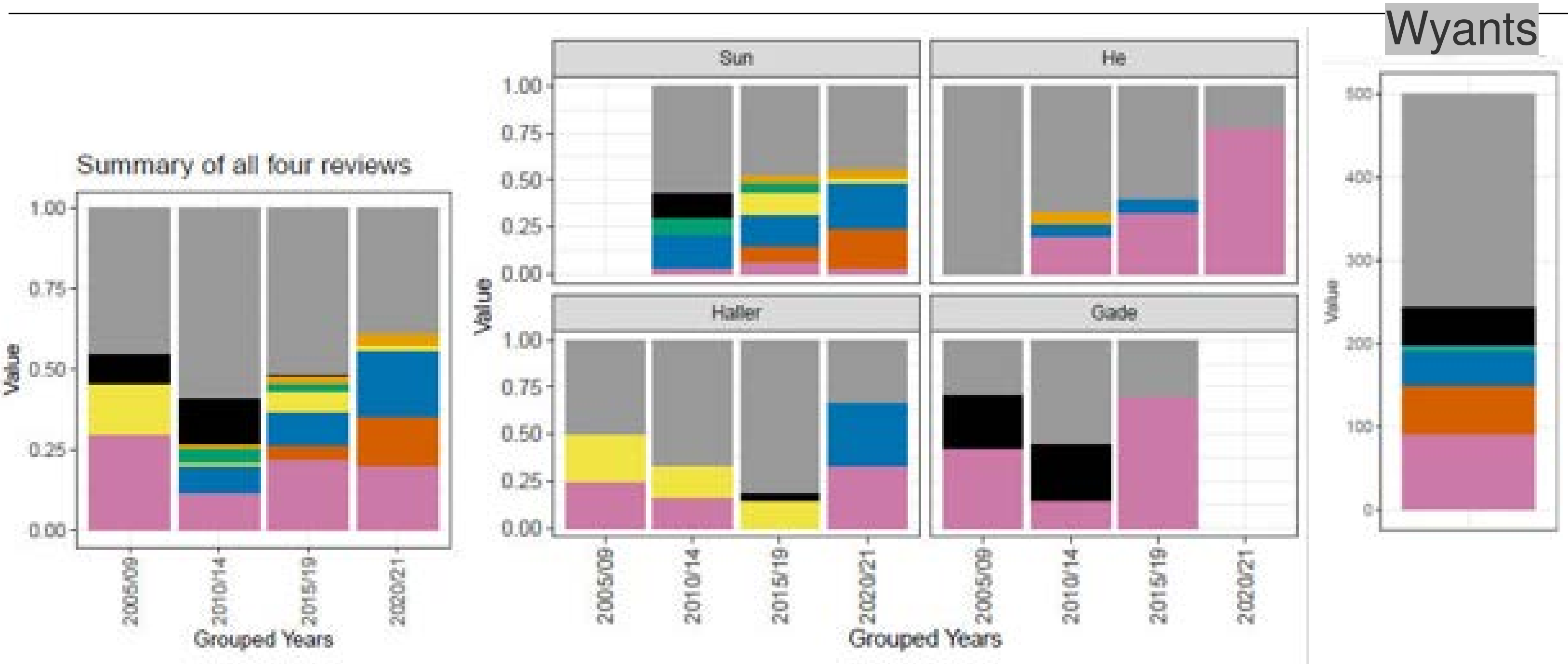






… and there is heterogeneity across reviews (not shown here)

Missing values

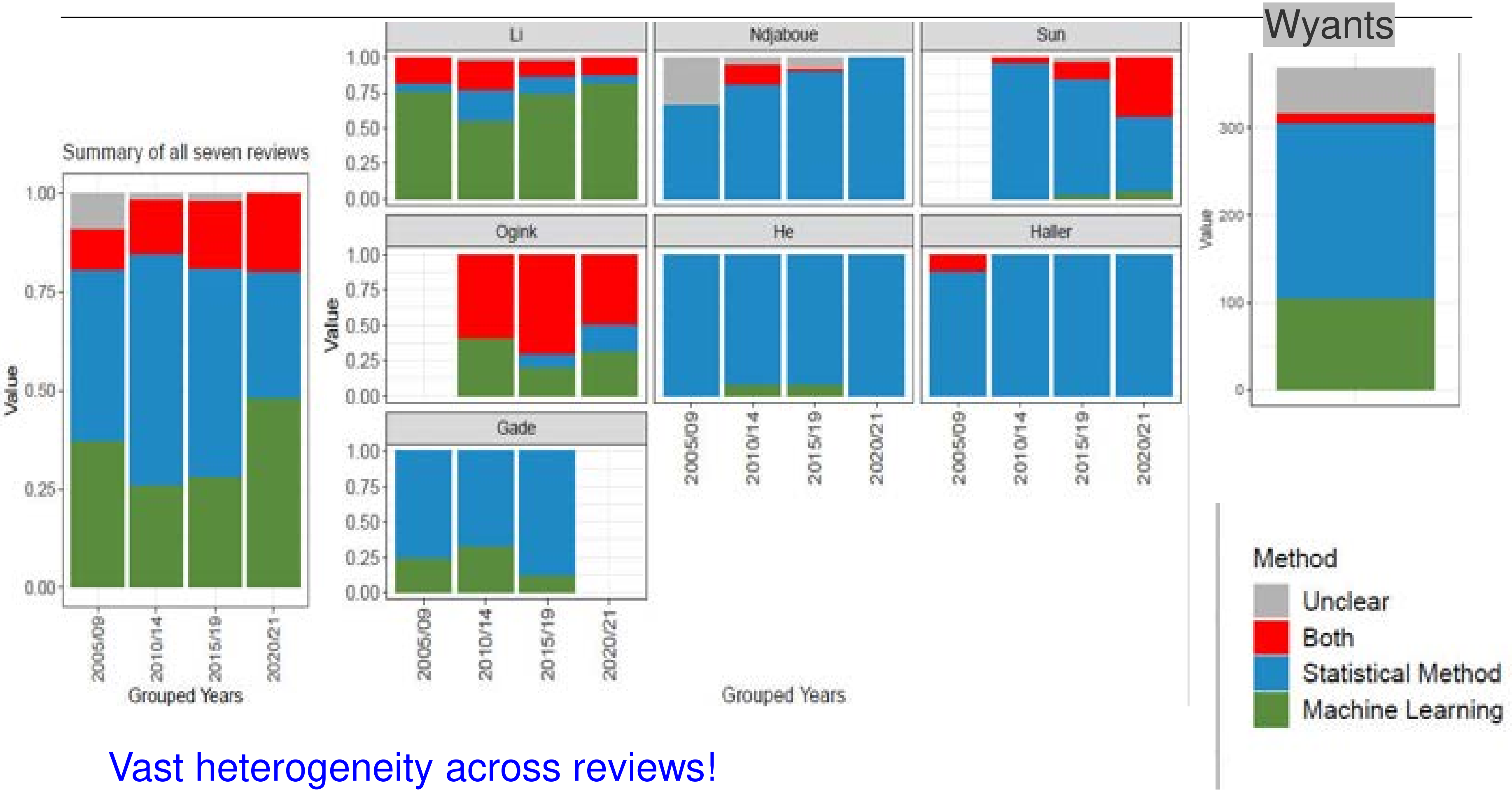


Method

- ⬛ Unclear / No information
- ⬛ Other
- 🟧 No Need To Report / None
- 🟦 Variable omission
- 🟩 Indicator methods / Dummy
- 🟨 Other imputation
- 🟦 Multiple imputation
- 🟧 Single imputation
- 🟪 Complete Case

3 reviews ignore the information
Information about missing data is still rarely reported in papers (gray)
Imputation methods (blue/red) are becoming more common
Complete case analysis is still the most common choice

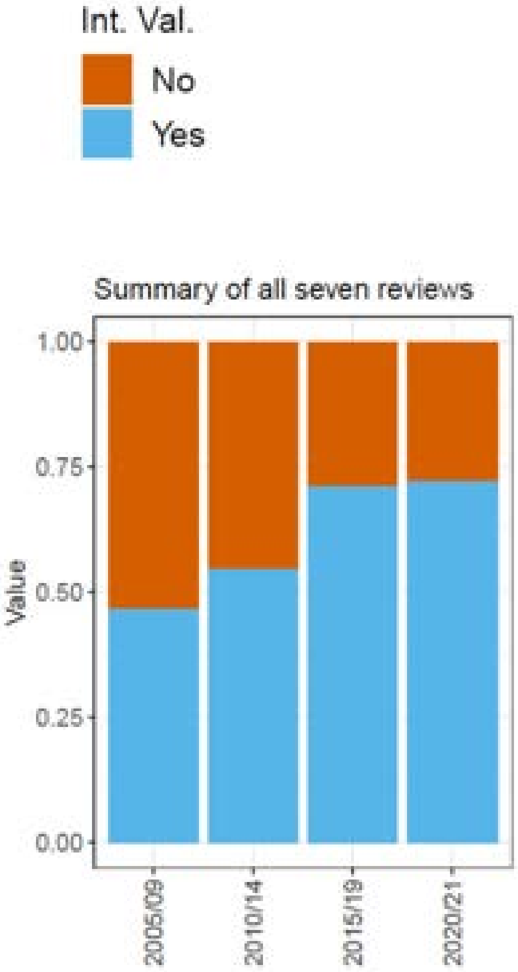
Type of model: ML vs statistical methods

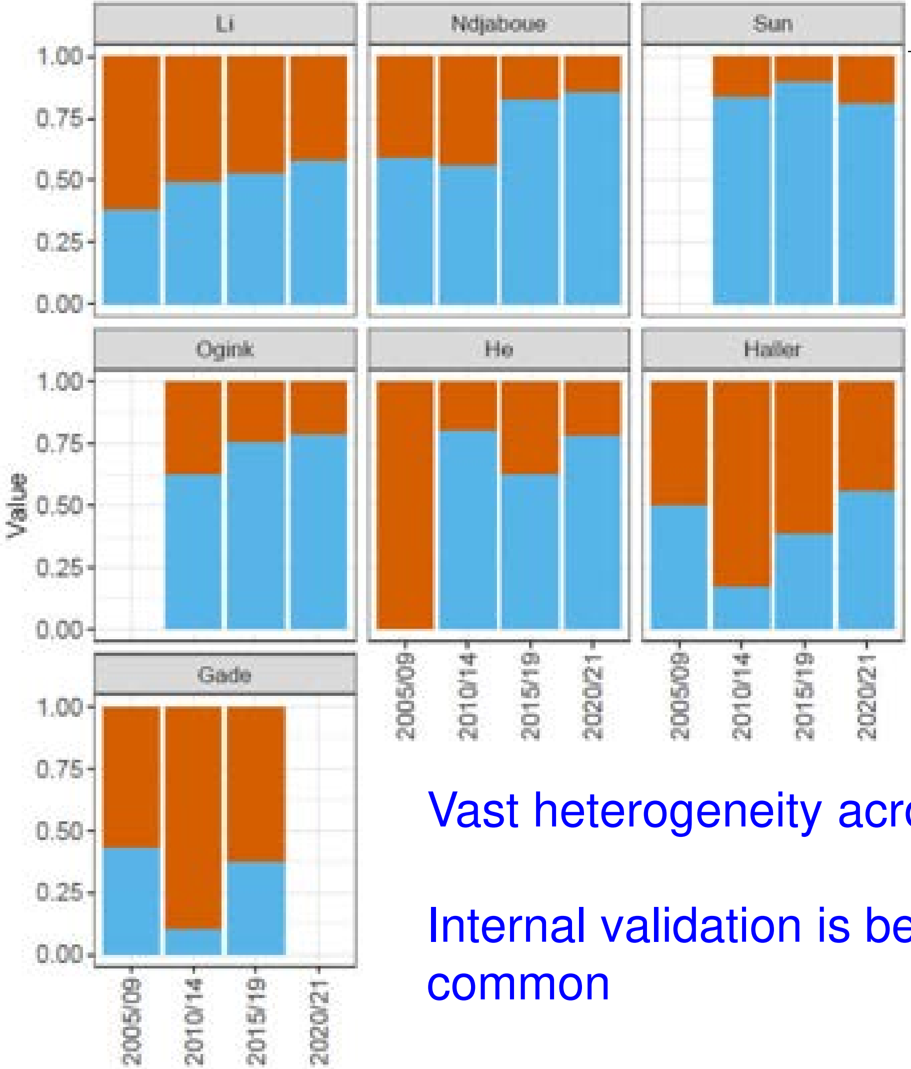


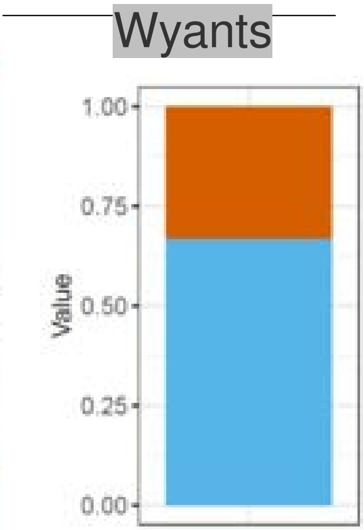
Vast heterogeneity across reviews!

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

Internal validation



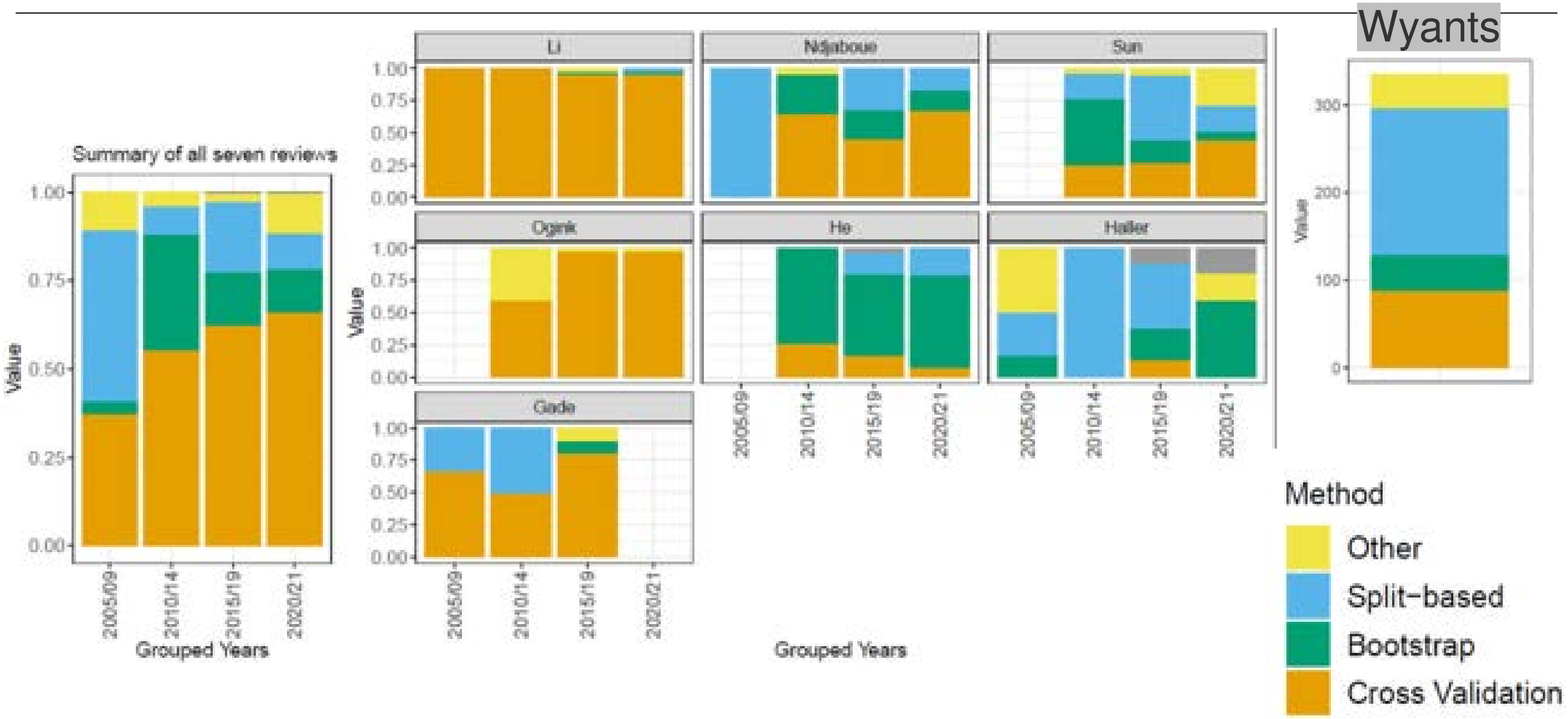




Vast heterogeneity across reviews!

Internal validation is becoming more common

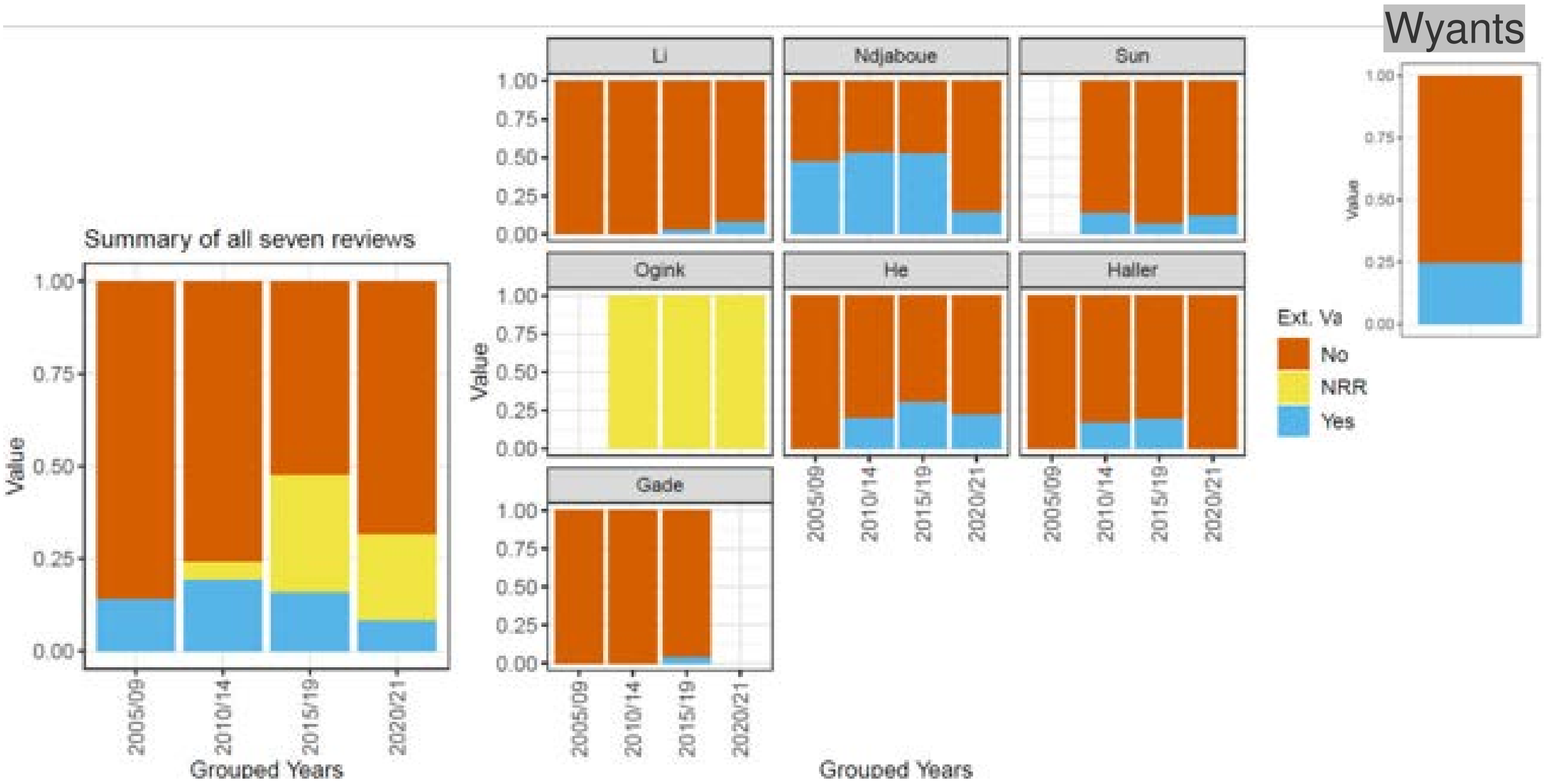
Internal validation – if performed, which?



Vast heterogeneity across reviews!

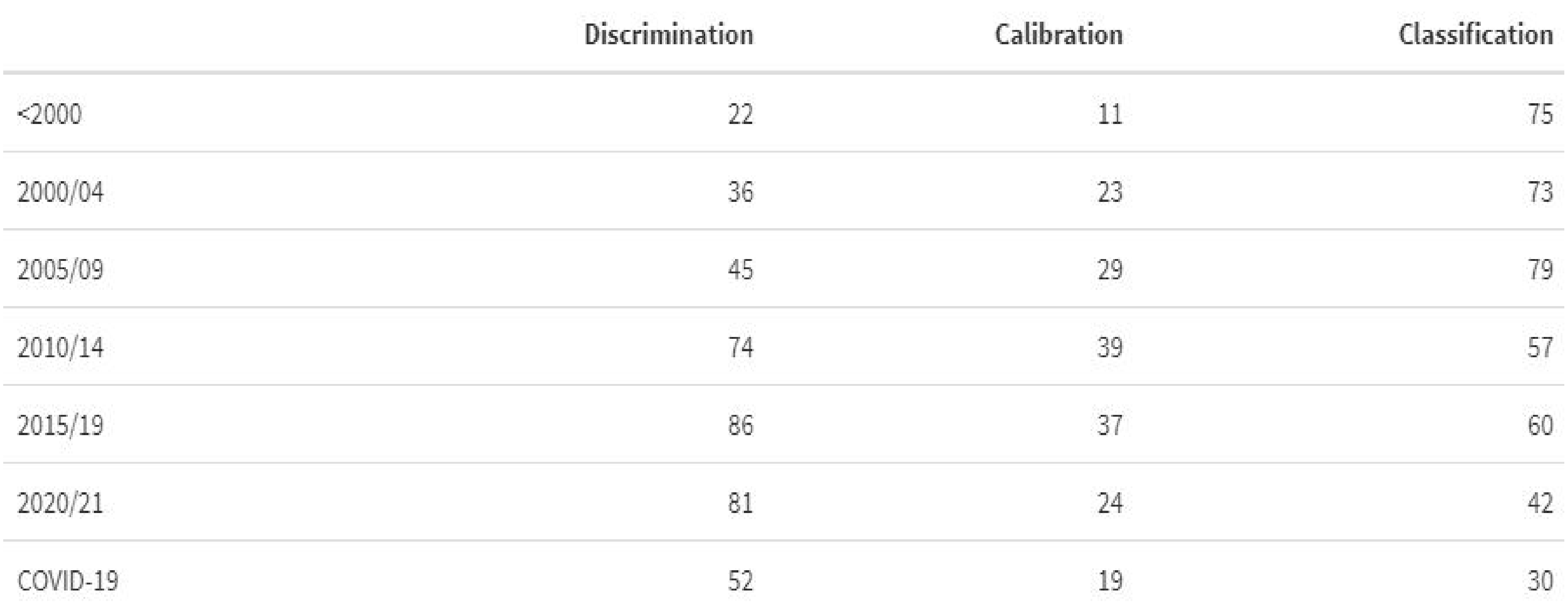
Cross-validation is gaining popularity, split-based methods are common only in some reviews

External validation



External validation is still uncommon
More common in the review from Njaboue, due to inclusion criteria

Time trends for the measures



	Discrimination	Calibration	Classification
<2000	22	11	75
2000/04	36	23	73
2005/09	45	29	79
2010/14	74	39	57
2015/19	86	37	60
2020/21	81	24	42
COVID-19	52	19	30

Steady increase of reporting of discrimination measures, less

Conclusions

- Quantitative assessment of changes is important, but it is not straightforward
- Reviews
  - Not many that include many predictors and have (complete relevant) publicly available data
  - There is a lot of heterogeneity across reviews in almost all the aspects
    - Truly reflecting differences in the fields or somehow related to the review process?

Changes in predictive modeling

- Larger sample sizes (and number of events)
- Larger mean number of selected variables
    - but similar median number of variables (candidate and selected)
- More imputation methods for missing values
    - but still poor reporting
- More models per paper
- Not a very clear increase of the use of ML methods in all reviews
    - Increase of the use of both methods
- More use of internal validation, but sill limited use of external validation
- Discrimination is better reported than calibration
    - and its reporting improved with time, more than for calibration

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”