Systematic review of variable and functional form selection in Covid-19 prognostic models

JSM 2025

6 August 2025

Michael Kammer, Gregor Buch, Marc Henrion and Georg Heinze on behalf of STRATOS TG2

STRATOS INITIATIVE

STRATOS
Topic Group 2

STRATOS (TG2)

STRengthening Analytical Thinking for Observational Studies

https://www.stratos-initiative.org/

Topic Group 2:

Selection of variables and functional forms in multivariable analysis

TG2 Aim: Derive guidance for variable and function selection in multivariable analysis.

Chairs: Georg Heinze, Aris Perperoglou, Willi Sauerbrei

Interrelated challenges (Harrell 2001, Sauerbrei et al. 2007)

- Selection of variables for inclusion in a multivariable model → identification of influential variables.
- Choice of the functional forms for continuous variables → insight into relationship with the outcome.

STRATOS (TG2)

First challenge:

Selection of variables for inclusion in a multivariable explanatory model.

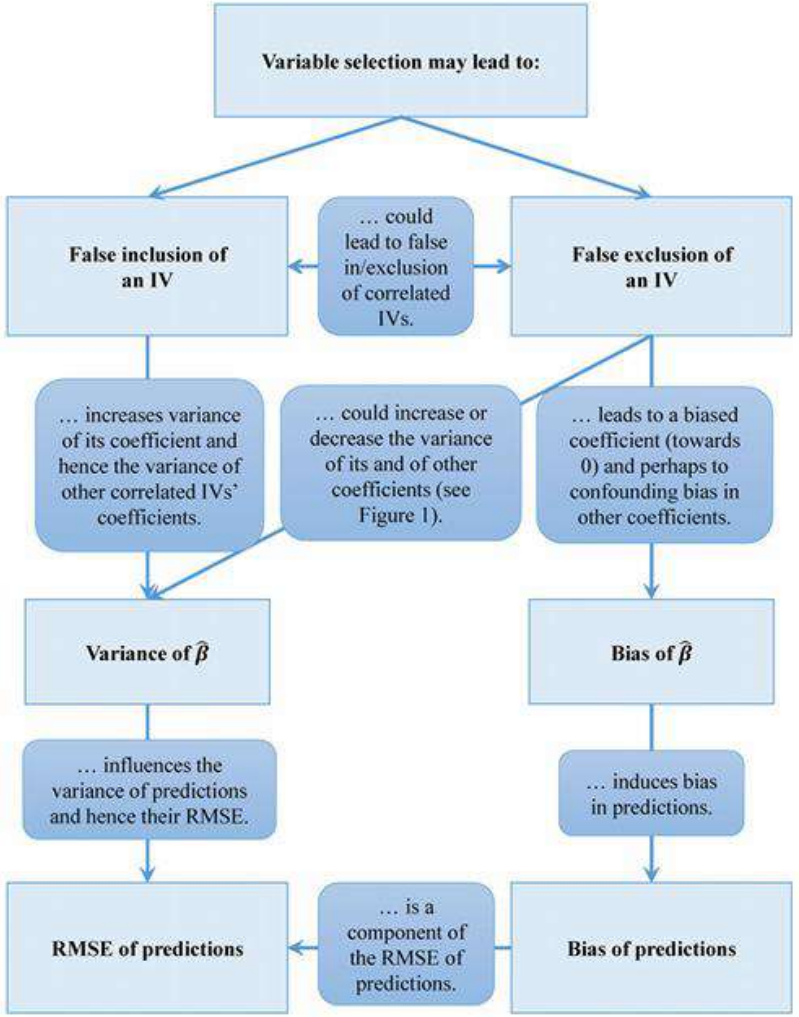
Multivariable models typically built through a combination of

- A priori inclusion of well established ‘predictors’.
- A posteriori data-driven selection of variables.

Consensus that **all model building strategies have weaknesses** (Miller 2002), but no consensus on the relative advantages and disadvantages of particular strategies.

Advanced methods (e.g. regularization techniques, resampling based methods, …) exist, but

- **No agreement, no state of the art.**
- **Need for clearer guidance and neutral, systematic comparisons.**



Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal*, *60*(3), 431-449.

STRATOS (TG2)

Second challenge:

Choice of the functional forms for continuous variables.

The effects of continuous predictors are typically modeled by

- Assuming linear relationships (possibly after simple transformations).
- Categorizing.

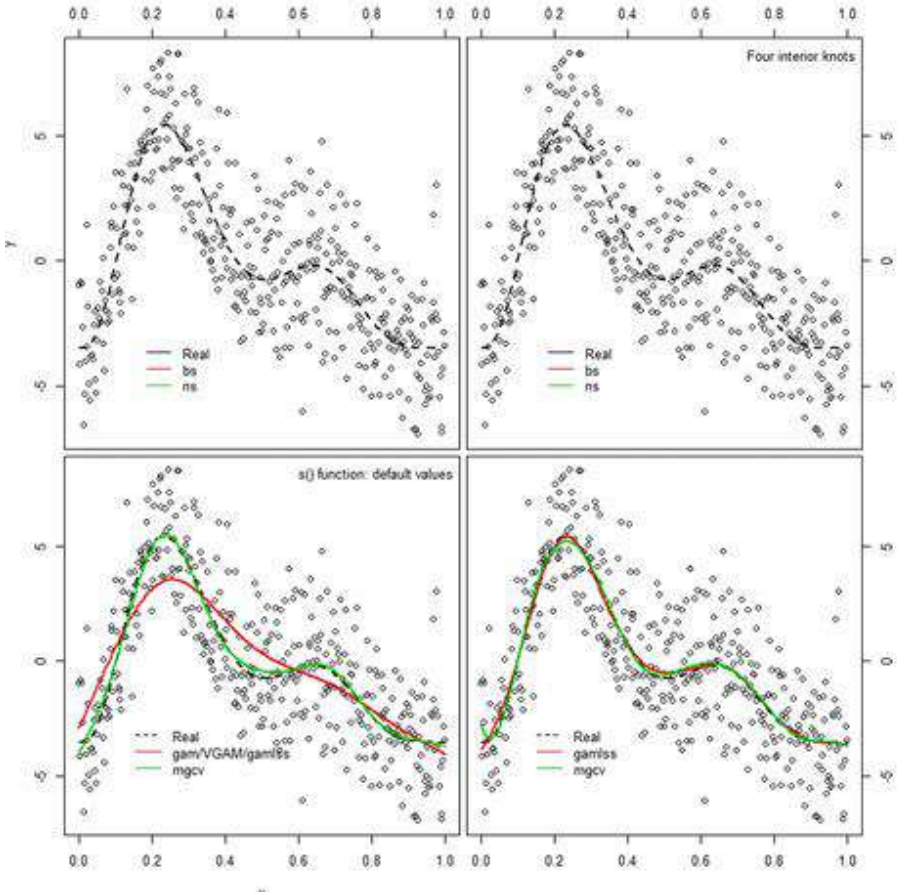
Problematic if reasons for and assumptions of such conventional approaches are not discussed and assessed.

Flexible modeling techniques have been developed and, for multivariable analysis, incorporated in GAMs:

- Fractional polynomials (Royston and Altman 1994, Royston and Sauerbrei 2008).
- Splines (many 'flavours'; Boer 2001, Harrell 2001, Wood 20017, Hastie and Tibshirani 1990).

But:

- **No agreement, no state of the art.**
- **Need for clearer guidance and neutral, systematic comparisons.**



Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC medical research methodology*, *19*(1), 46.

STRATOS (TG2)

Selected outputs

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and Prognostic Research

COMMENTARY Open Access

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei[1*], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], Georg Heinze[6] and for TG2 of the STRATOS initiative

Heinze *et al.*
BMC Medical Research Methodology (2024) 24:178
https://doi.org/10.1186/s12874-024-02294-3

BMC Medical Research Methodology

RESEARCH Open Access

Regression without regrets –initial data analysis is a prerequisite for multivariable regression

Georg Heinze[1*], Mark Baillie[2], Lara Lusa[3,4], Willi Sauerbrei[5], Carsten Oliver Schmidt[6], Frank E. Harrell[7], Marianne Huebner[8] on behalf of TG2 and TG3 of the STRATOS initiative

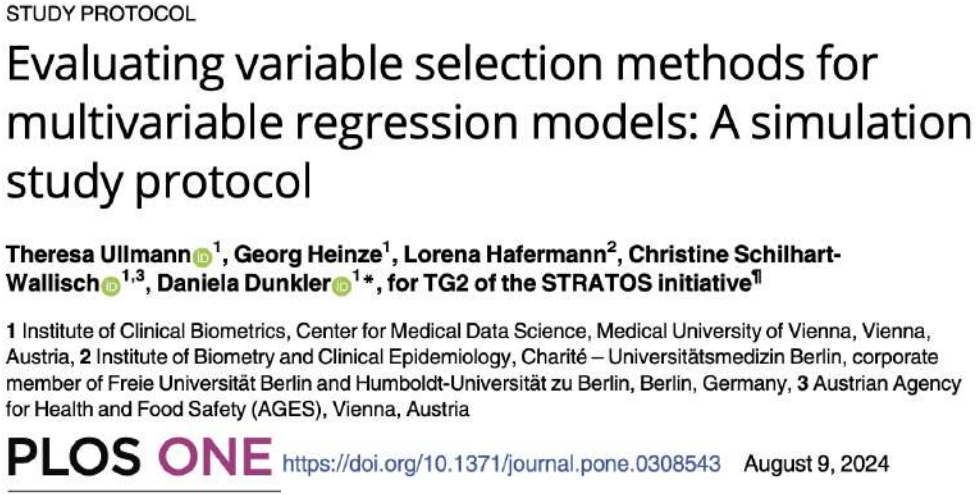
REVIEW Open Access

A review of spline function procedures in R

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

Perperoglou *et al. BMC Medical Research Methodology* (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

BMC Medical Research Methodology



STUDY PROTOCOL

Evaluating variable selection methods for multivariable regression models: A simulation study protocol

Theresa Ullmann[1], Georg Heinze[1], Lorena Hafermann[2], Christine Schilhart-Wallisch[1,3], Daniela Dunkler[1]*, for TG2 of the STRATOS initiative[¶]

1 Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria, **2** Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany, **3** Austrian Agency for Health and Food Safety (AGES), Vienna, Austria

PLOS ONE https://doi.org/10.1371/journal.pone.0308543 August 9, 2024

STRATOS (TG2)

Towards recommendations / guidelines:

Research needed!

1. Investigation and comparison of the **properties** of **variable selection strategies**
2. Comparison of **spline procedures** in **univariable and multivariable contexts**
3. How to model one or more variables with a **‘spike-at-zero**’?
4. Comparison of **multivariable procedures** for **model and function selection**
5. Role of **shrinkage to correct for bias** introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. **Adaptation** of procedures for **very large sample sizes** needed?

Covid-19 Prognostic Modelling Review

Motivation: COVID PRECISE study

RESEARCH

OPEN ACCESS

Check for updates

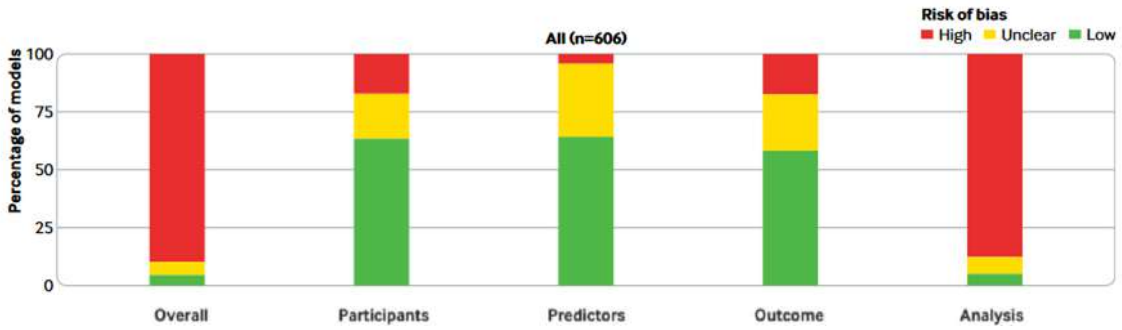
FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[21] Kim Luijken,[21] Jie Ma,[5] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32] Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch,[7,35,36] Jack Wilkinson,[22] Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

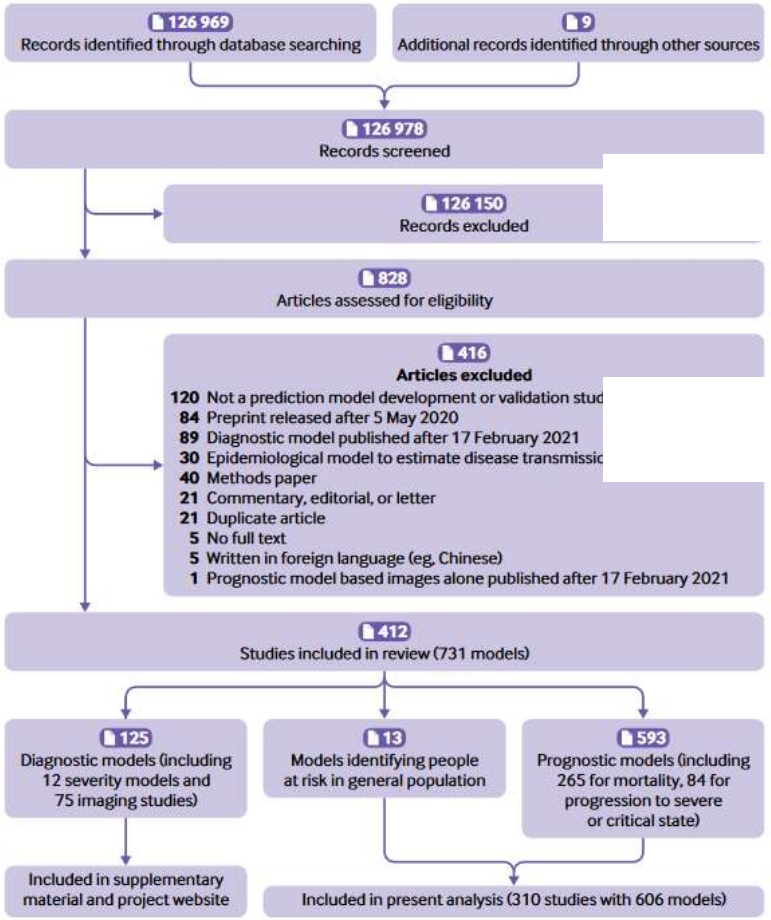
BMJ: first published as 10.1136/bmj.m1328 on 7 April

(Wynants et al 2020)



Full results database available
https://www.covprecise.org/

- 731 models from 412 studies
- Repeated updates during epidemic
- Risk of bias assessment (ROB)
- > 3000 citations



Stratos TG2 oriented re-review

COVID PRECISE reflects **methods researchers rely on in times of crisis,** when robust, reliable models are needed.

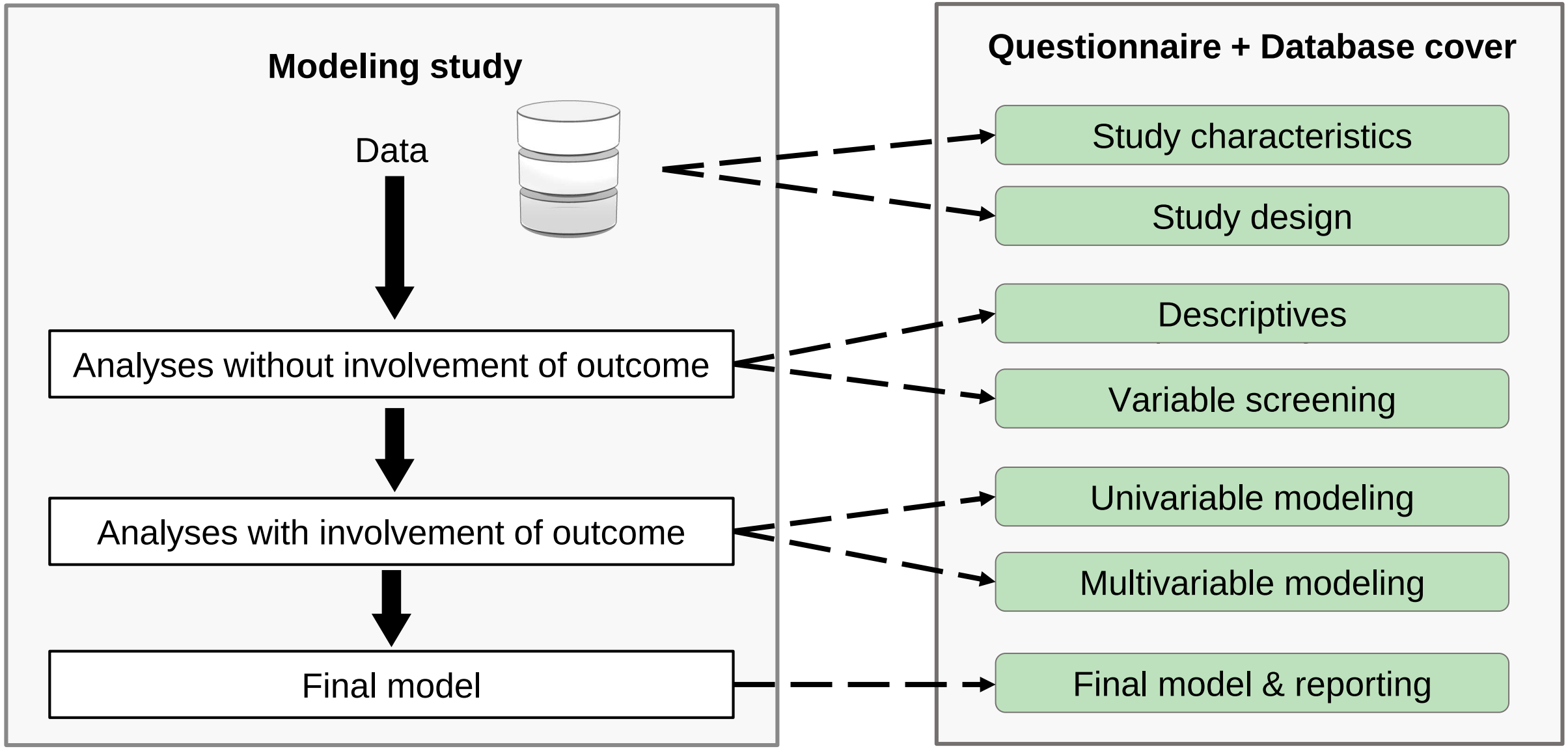
Hence, it allows us to:

Identify approaches in regression-based prediction models for COVID-19 outcomes to:

1) **select predictors** for regression models, and
2) **model the effects** of predictors, in particular the use of **non-linear functional forms** and the use of **interactions** between predictors.

This extends the data with details on the procedures which were not recorded for ROB.

Our model of a modelling workflow



Our re-review

Stage 0: Develop protocol and extraction sheet

- Input from original study authors and TG2 members
- Two pilot studies with 4 papers and several reviewers to test protocol
- Focus on regression based **prognostic** models. Excluded (from 731):
  - 124 diagnostic models,
  - 442 machine learning / non-parametric methods,
  - 232 external validations of existing models.

181 studies remain for re-review

For each a primary model was chosen by pre-defined criteria

Our re-review

Stage 0: Develop protocol and extraction sheet

Stage 1: Extract relevant data from existing database

- Study characteristics, Basic model characteristics, Reporting
- Provides background info for further extraction stages
- Done by core team

Our re-review

Stage 0: Develop protocol and extraction sheet

Stage 1: Extract relevant data from existing database

Stage 2: Re-extract data

- Invite reviewers for double review followed by consensus
- Extract details on variable selection & functional forms
- Done in pairs as double-review followed by consensus

Our re-review

Stage 0: Develop protocol and extraction sheet

Stage 1: Extract relevant data from existing database

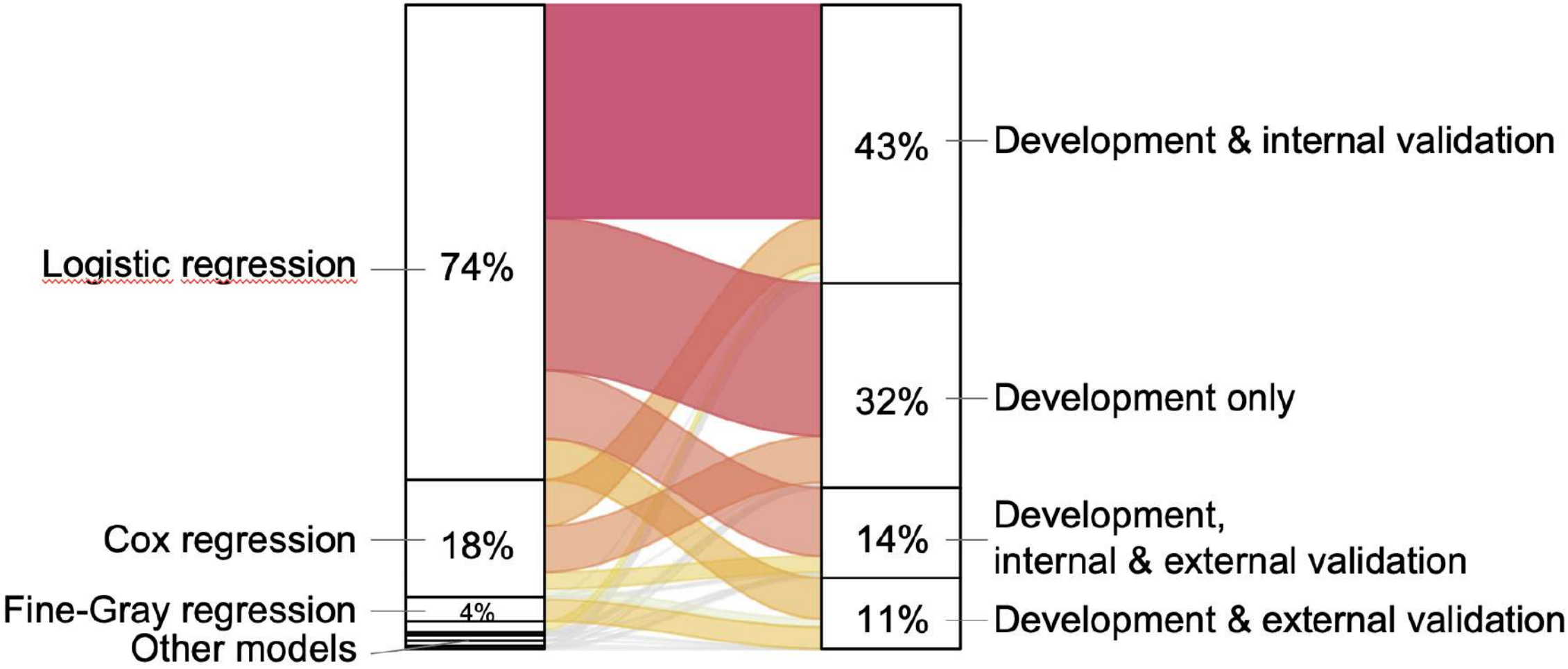
Stage 2: Re-extract data

Stage 3: Data consolidation & analysis

- Done by core team

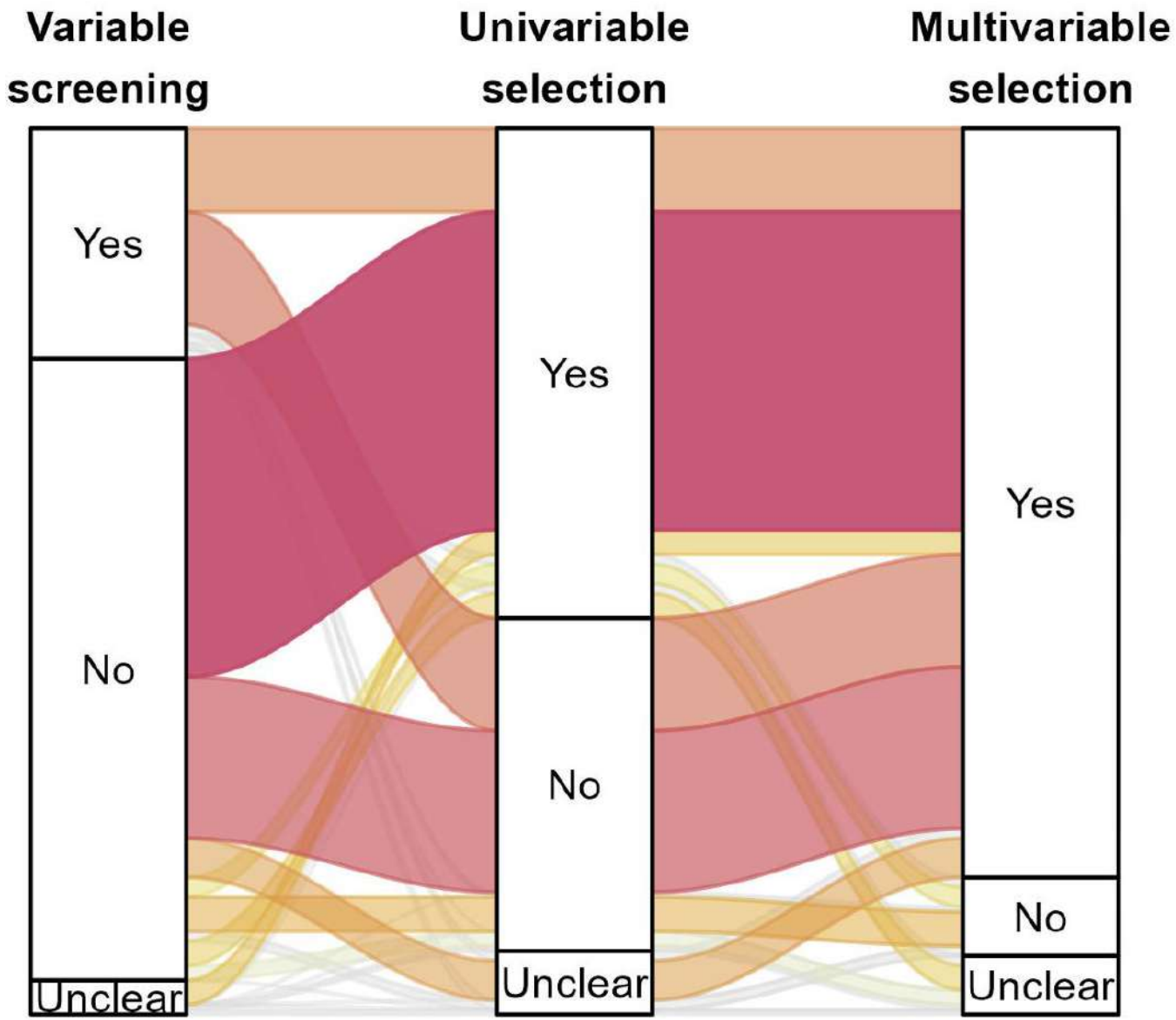
Results: Overview

Data extraction of 181 models completed February 2025



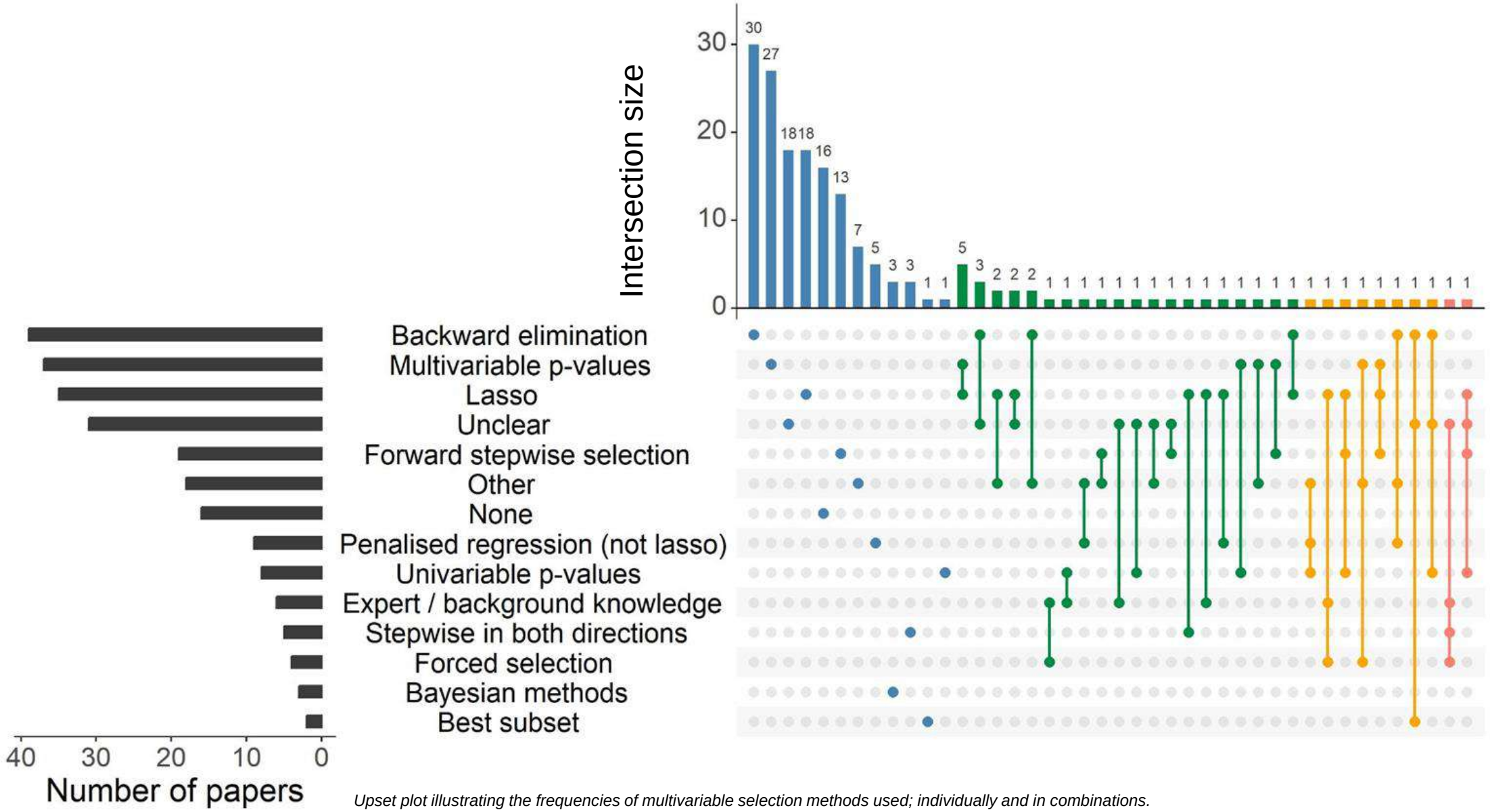
Median sample size 344 (IQR 156 - 982) with median 68 events (IQR 35 - 169)

Results: Modelling patterns



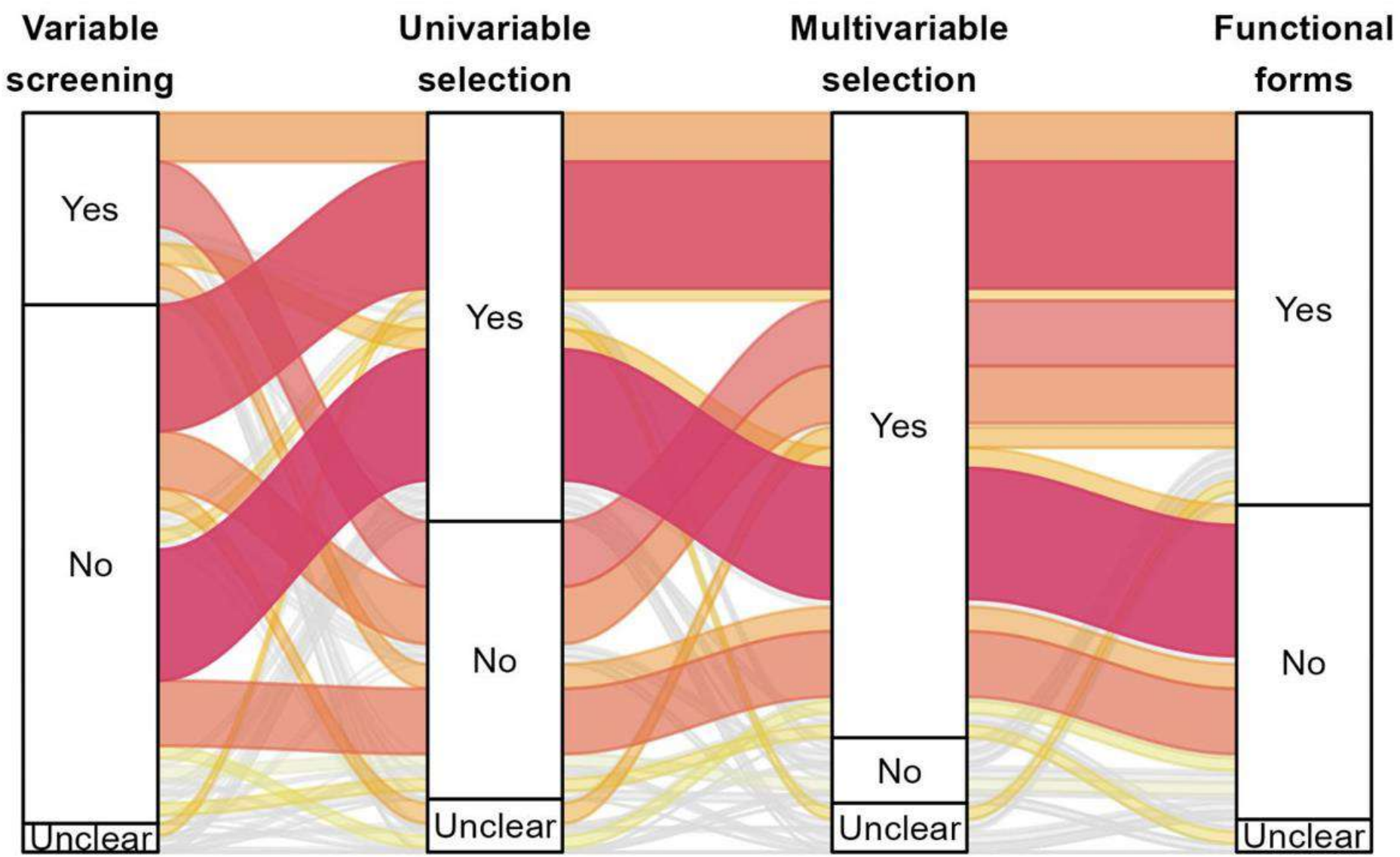
Alluvial plot illustrating the flow of modelling decisions. Flows are color-coded for distinct pathways.

Results: Multivariable selection methods



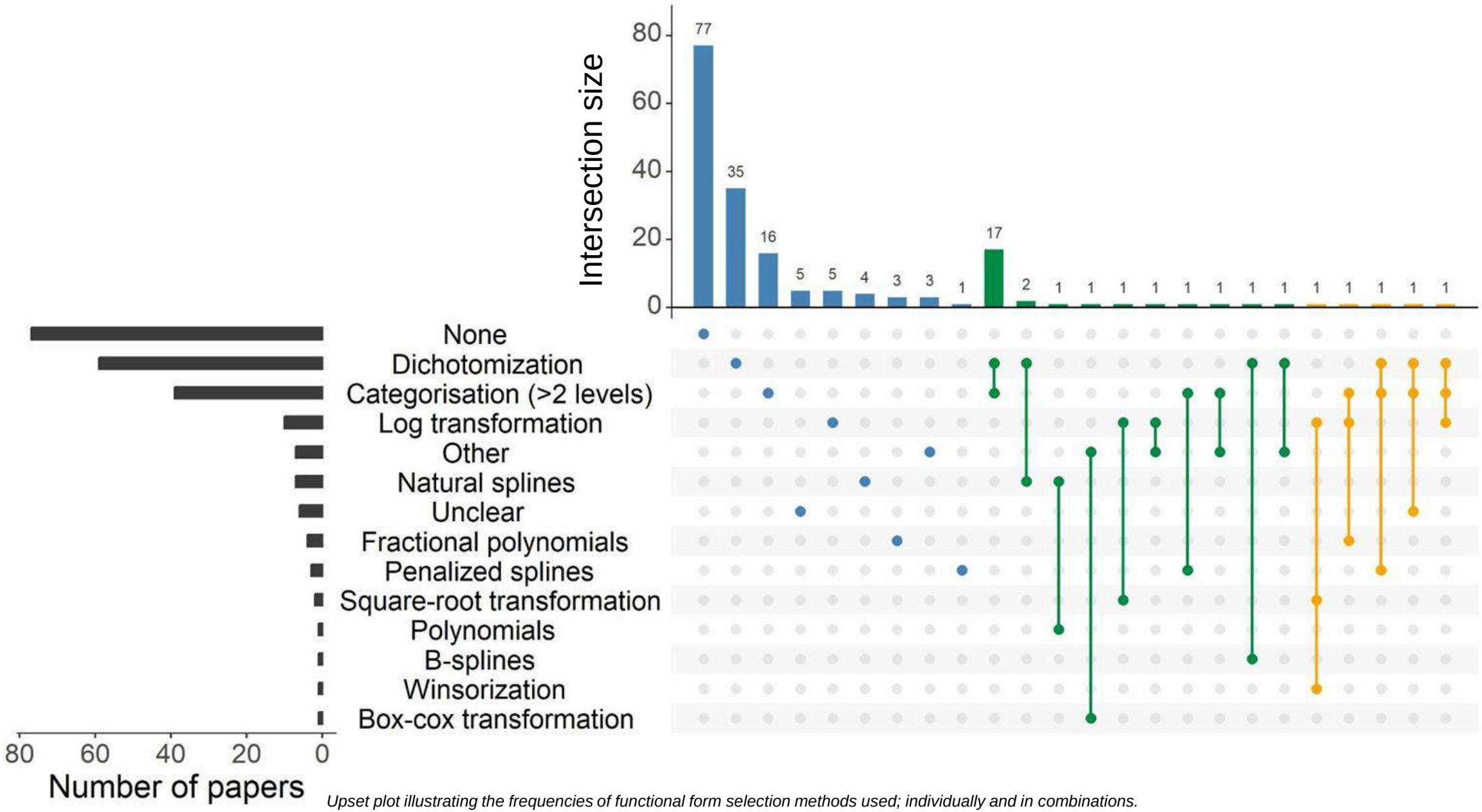
Upset plot illustrating the frequencies of multivariable selection methods used; individually and in combinations.

Results: Modelling patterns



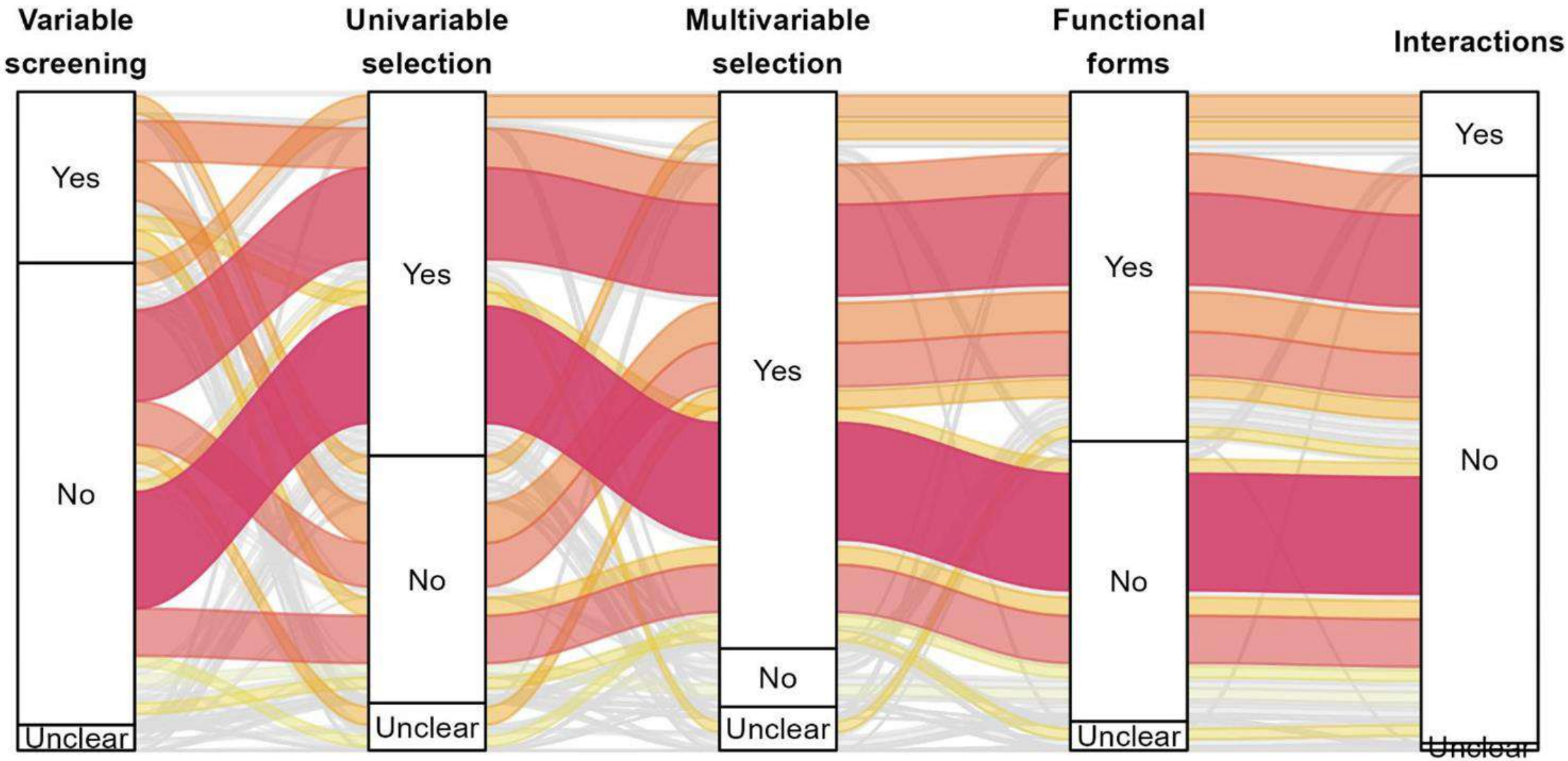
Alluvial plot illustrating the flow of modelling decisions. Flows are color-coded for distinct pathways.

Results: Functional form selection



Upset plot illustrating the frequencies of functional form selection methods used; individually and in combinations.

Results: Modelling patterns



Alluvial plot illustrating the flow of modelling decisions. Only combinations occurring more than once are visualized. Flows are color-coded for distinct pathways.

Results: Model reporting is challenging

Guidance documents rarely cited

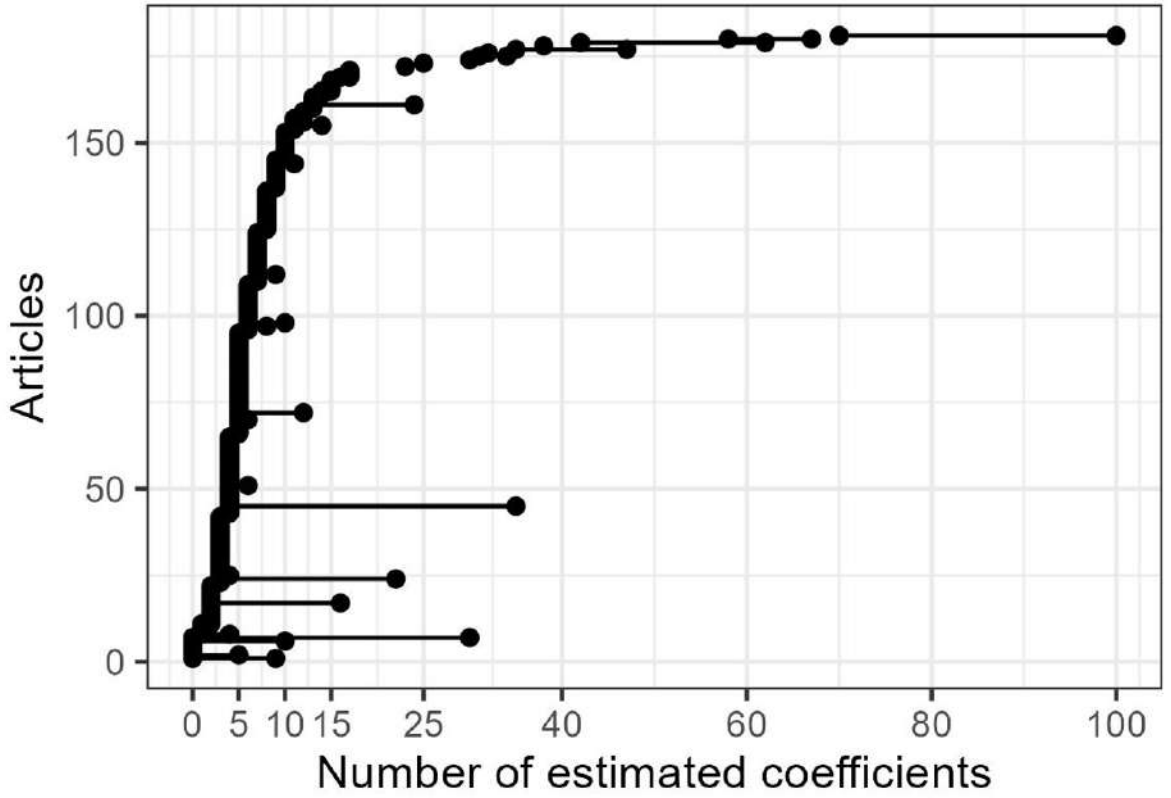
COVID PRECISE review cited in 23%, TRIPOD in 15%, others ≤ 3 times

Full, final models often not reported

Challenging: Not presented in 29%,
as sum score 11%, as online tool 7%

Easier: Nomogram 25%,
(partial) regression formula 17%

Considerable uncertainty even about e.g. number of coefficients



Results: Unusual approaches

There were quite a few unusual approaches for variable and functional form selection that reviewers struggled with during extraction.

- Unclear reporting.
- ‘Expected’ unusual choices [e.g. interesting p-value cut-offs, unorthodox stepwise selections, creative categorisation cut-offs].
- Fairly complex procedures [often unclear rationale, often badly reported].
- Genuinely creative applications [e.g. lasso as part of a stepwise elimination strategy].

→ **A need for more comprehensive / authoritative guidance?**

→ **An opportunity to learn?**

Conclusions: Modeling workflows are diverse

- **No standard modelling workflow.**
- **Variable selection is common practice.**
  - Particularly multivariable selection (>80% of models) but also univariable (>50%).
  - Methods are combined in novel ways that are not investigated in the literature.
  - Selection is not reflected when reporting inference.
- **The use of continuous functional forms and interactions is not.**
  - Widespread use of dichotomization and categorization (>50% of models).
  - Continuous functional forms rarely used (<10% of models).
  - Functional forms were rarely assessed through variable selection (5% of models).

Our empirical results underline opportunities for learning, improving guidance and to keep pushing for better reporting

Find the protocol at https://osf.io/2afuz/

A big thank you to all our reviewers and supporters

Alexander Gieswinkel (Mainz)

Alice Schneider (Berlin)

Andreas Klinger (Vienna)

Daniel Schulze (Berlin)

Daniela Dunkler (Vienna)

David McLernon (Aberdeen)

James Chirombo (Blantyre)

Johannes Vey (Heidelberg)

Laure Wynants (Maastricht)

Linard Hoessly (Basel)

Lorena Hafermann (Berlin)

Manuel Feißt (Berlin)

Mariana Nold (Jena)

Moritz Pamminger (Vienna)

Theresa Ullmann (Vienna)

Ulrike Grittner (Berlin)

Willi Sauerbrei (Freiburg)

References

- De Boor, C., 2001. A practical guide to splines: with 32 figures, Rev. ed. ed, Applied mathematical sciences. Springer, New York.
- Harrell, F.E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3462-1
- Hastie, T., Tibshirani, R., 1999. Generalized additive models. Chapman & Hall/CRC, Boca Raton, Fla.
- Miller, A., 2002. Subset Selection in Regression, Chapman and Hall/CRC. https://doi.org/10.1201/9781420035933
- Royston, P., Altman, D.G., 1994. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Applied Statistics 43, 429. https://doi.org/10.2307/2986270
- Royston, P., Sauerbrei, W., 2008. Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for continuous variables. Wiley, Chichester.
- Sauerbrei, W., Royston, P., Binder, H., 2007. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statist. Med. 26, 5512–5528. https://doi.org/10.1002/sim.3148
- Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279
- Wynants, L., Van Calster, B., Collins, G.S., et al. 2020, Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ, 7:369:m1328. https://doi.org/10.1136/bmj.m1328