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

GMDS 2025

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

STRATOS INITIATIVE

STRATOS (TG2)

STRengthening Analytical Thinking for Observational Studies

Collaboration of experts to provide guidance for many aspects of biostatistics with several working groups

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

We do this review on behalf of **Topic Group 2**:

Selection of variables and functional forms in multivariable analysis

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

Chairs: Georg Heinze, Aris Perperoglou, Willi Sauerbrei

STRATOS (TG2)

Selection of variables for inclusion in a multivariable explanatory model.

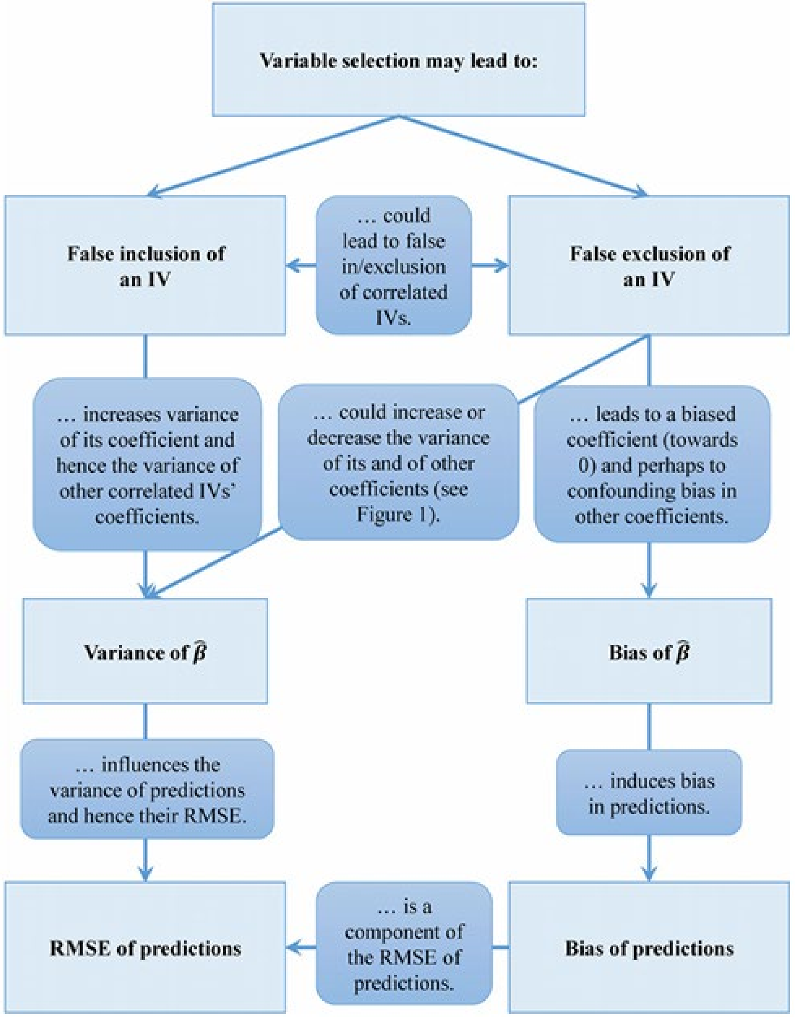
Multivariable models typically built through a combination of

- A priori inclusion of well-established 'predictors'
- A posteriori selection using data-driven procedures

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

No agreement on the state of the art

Clearer guidelines and neutral, systematic comparisons are needed.



Variable selection may lead to:

False inclusion of an IV

… could lead to false in/exclusion of correlated IVs.

False exclusion of an IV

… increases variance of its coefficient and hence the variance of other correlated IVs' coefficients.

… could increase or decrease the variance of its and of other coefficients (see Figure 1).

… leads to a biased coefficient (towards 0) and perhaps to confounding bias in other coefficients.

Variance of $\hat{\beta}$

Bias of $\hat{\beta}$

… influences the variance of predictions and hence their RMSE.

… induces bias in predictions.

RMSE of predictions

… is a component of the RMSE of predictions.

Bias of predictions

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

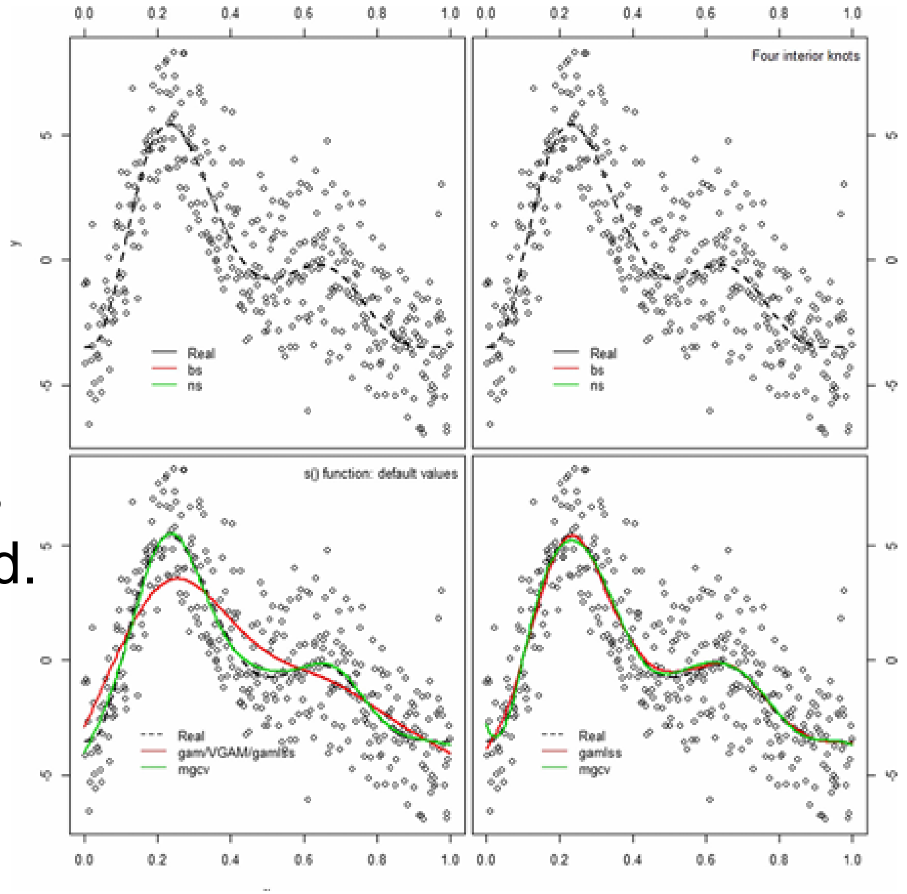
STRATOS (TG2)

Choice of the functional forms for continuous variables.

Effects of continuous predictors are typically modeled by

- Assuming linear relationships
- Categorizing variables

Conventional approaches are often used without assessing assumptions. Flexible methods – like fractional polynomials (Royston & Sauerbrei, 2008) and splines (Hastie & Tibshirani, 1990) – are rarely applied.



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

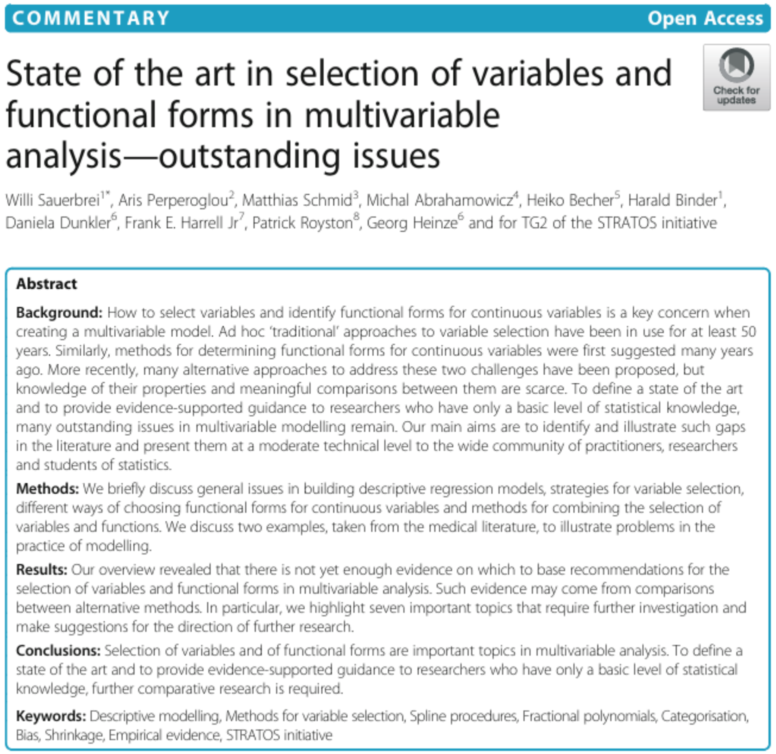
No agreement on the state of the art

Clearer guidelines and neutral, systematic comparisons are needed.

STRATOS (TG2)

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?



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

Abstract

Background: How to select variables and identify functional forms for continuous variables is a key concern when creating a multivariable model. Ad hoc ‘traditional’ approaches to variable selection have been in use for at least 50 years. Similarly, methods for determining functional forms for continuous variables were first suggested many years ago. More recently, many alternative approaches to address these two challenges have been proposed, but knowledge of their properties and meaningful comparisons between them are scarce. To define a state of the art and to provide evidence-supported guidance to researchers who have only a basic level of statistical knowledge, many outstanding issues in multivariable modelling remain. Our main aims are to identify and illustrate such gaps in the literature and present them at a moderate technical level to the wide community of practitioners, researchers and students of statistics.

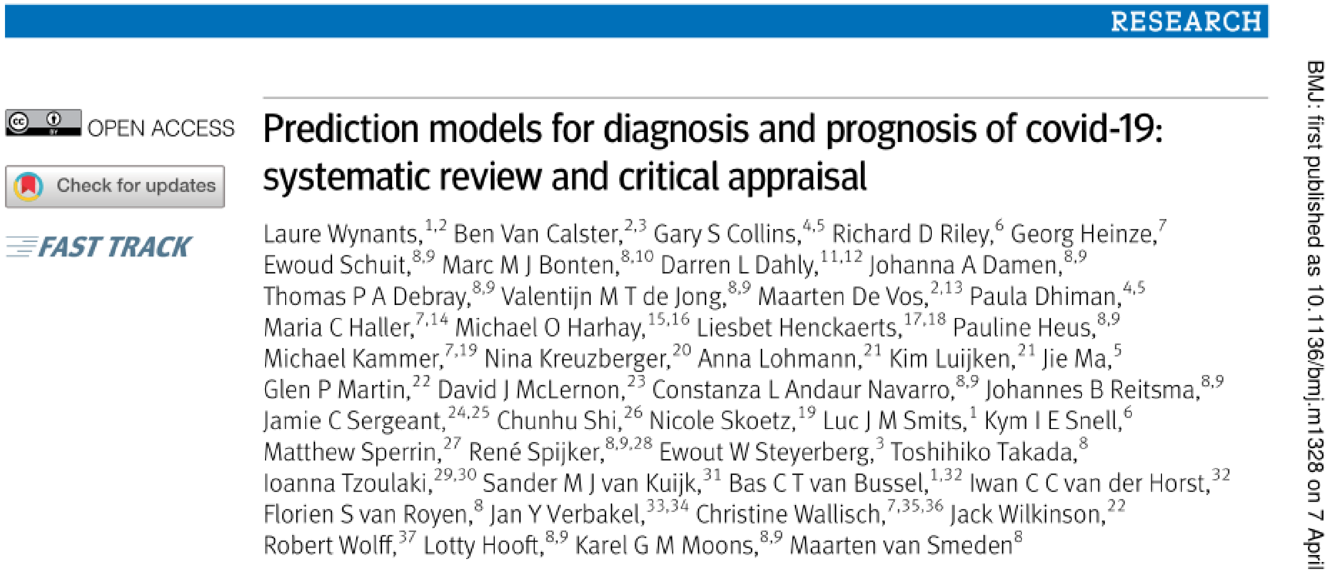
Methods: We briefly discuss general issues in building descriptive regression models, strategies for variable selection, different ways of choosing functional forms for continuous variables and methods for combining the selection of variables and functions. We discuss two examples, taken from the medical literature, to illustrate problems in the practice of modelling.

Results: Our overview revealed that there is not yet enough evidence on which to base recommendations for the selection of variables and functional forms in multivariable analysis. Such evidence may come from comparisons between alternative methods. In particular, we highlight seven important topics that require further investigation and make suggestions for the direction of further research.

Conclusions: Selection of variables and of functional forms are important topics in multivariable analysis. To define a state of the art and to provide evidence-supported guidance to researchers who have only a basic level of statistical knowledge, further comparative research is required.

Keywords: Descriptive modelling, Methods for variable selection, Spline procedures, Fractional polynomials, Categorisation, Bias, Shrinkage, Empirical evidence, STRATOS initiative

Motivation: COVID PRECISE study



RESEARCH

OPEN ACCESS

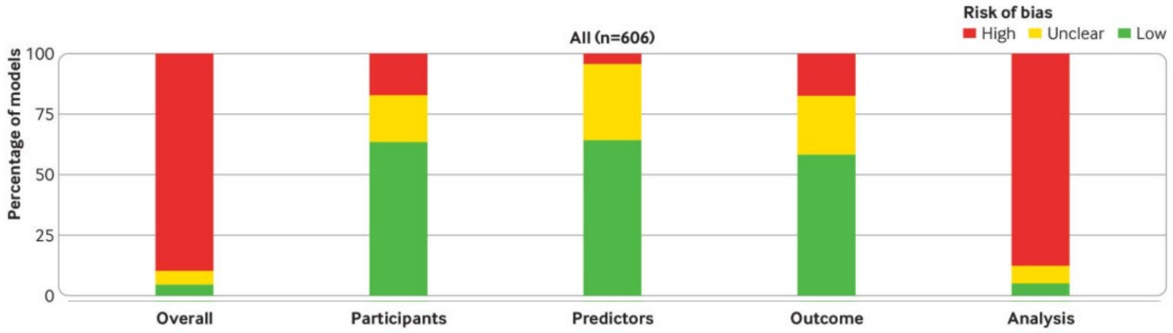
Check for updates

FAST TRACK

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

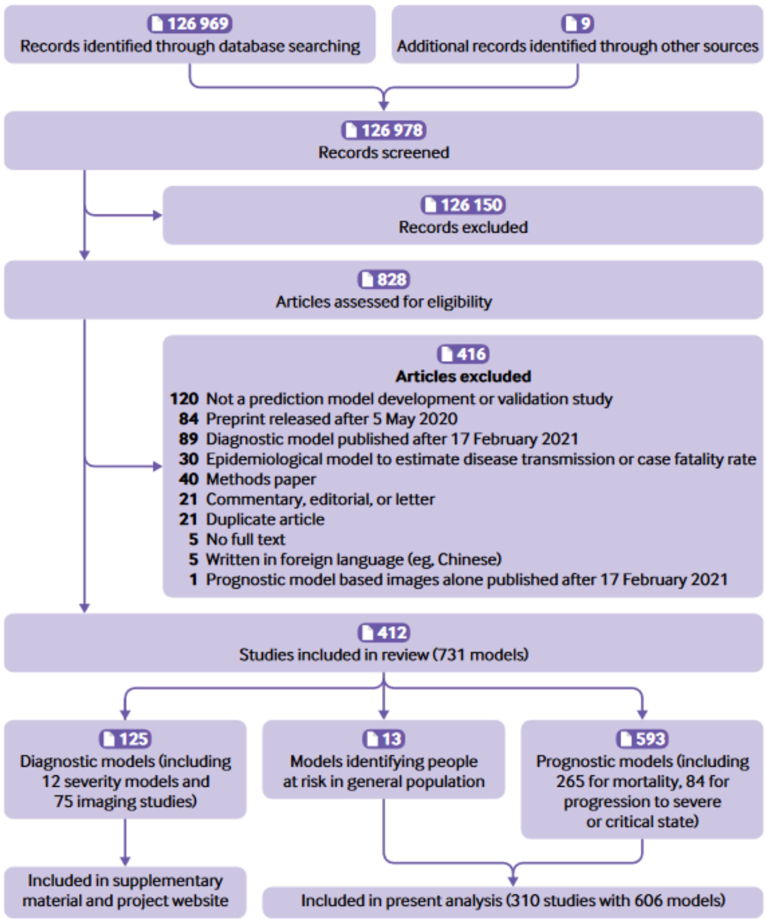
Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] 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



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

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



126 969 Records identified through database searching

9 Additional records identified through other sources

126 978 Records screened

126 150 Records excluded

828 Articles assessed for eligibility

416 **Articles excluded**

120 Not a prediction model development or validation study
84 Preprint released after 5 May 2020
89 Diagnostic model published after 17 February 2021
30 Epidemiological model to estimate disease transmission or case fatality rate
40 Methods paper
21 Commentary, editorial, or letter
21 Duplicate article
5 No full text
5 Written in foreign language (eg, Chinese)
1 Prognostic model based images alone published after 17 February 2021

412 Studies included in review (731 models)

125 Diagnostic models (including 12 severity models and 75 imaging studies)

13 Models identifying people at risk in general population

593 Prognostic models (including 265 for mortality, 84 for progression to severe or critical state)

Included in supplementary material and project website

Included in present analysis (310 studies with 606 models)

Stratos TG2 oriented re-review

COVID PRECISE reflects **methods researchers rely on in times of crisis**

Hence, it allows us to:

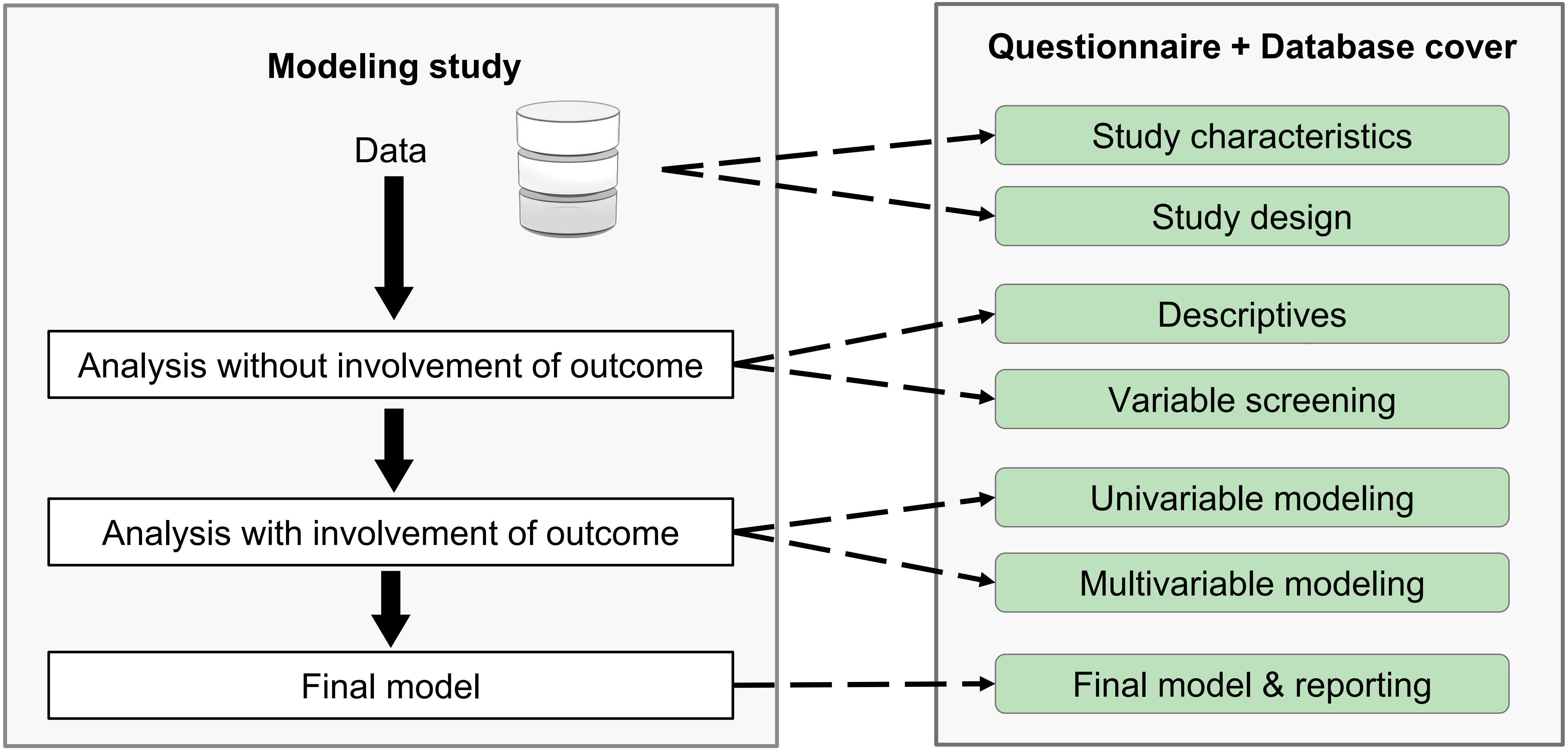
Identify approaches in regression-based prediction models for COVID-19 outcomes to:

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



Modeling study

Data

Analysis without involvement of outcome

Analysis with involvement of outcome

Final model

Questionnaire + Database cover

Study characteristics

Study design

Descriptives

Variable screening

Univariable modeling

Multivariable modeling

Final model & reporting

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 total 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

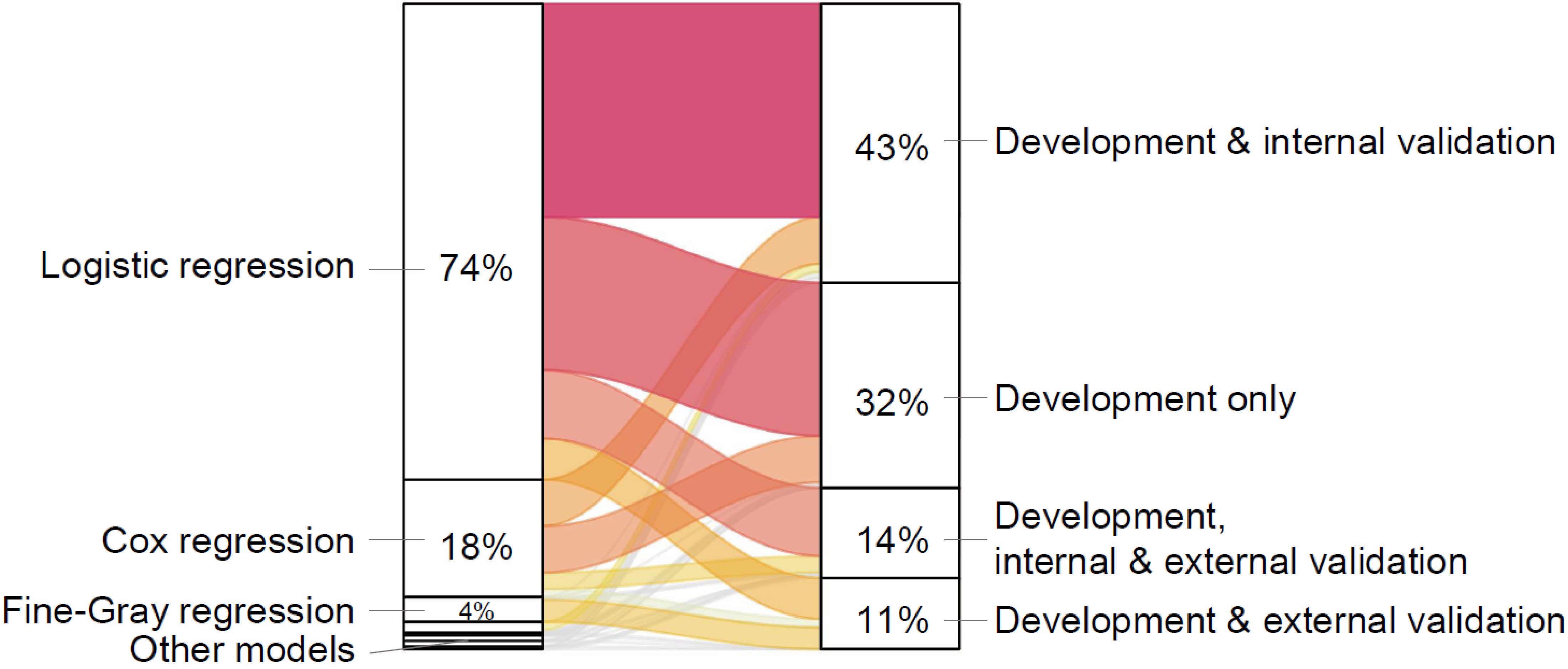
Stage 1: Extract relevant data from existing database

Stage 2: Re-extract data

Stage 3: Data consolidation & analysis

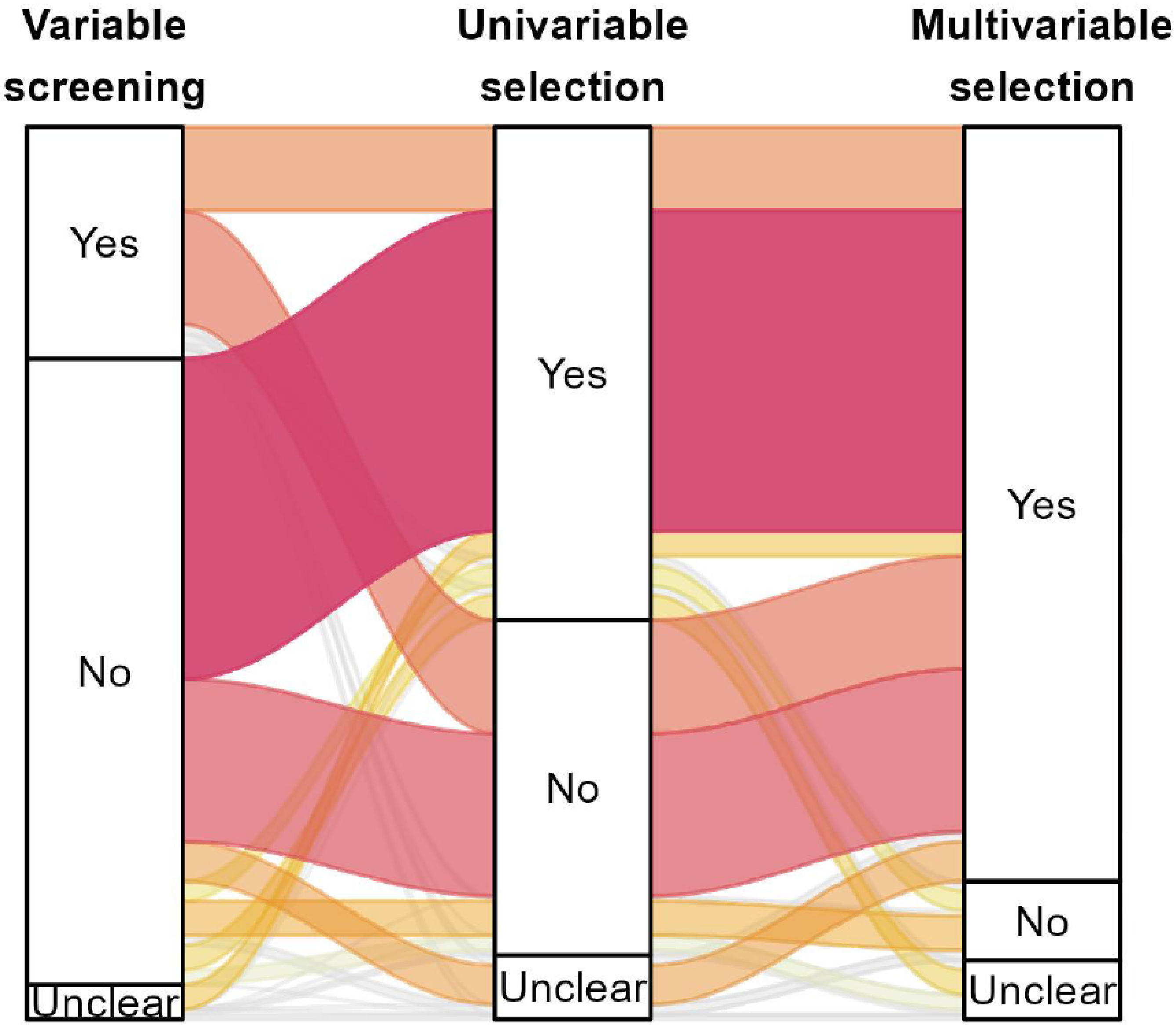
Results: Overview

Data extraction of 181 models completed February 2025



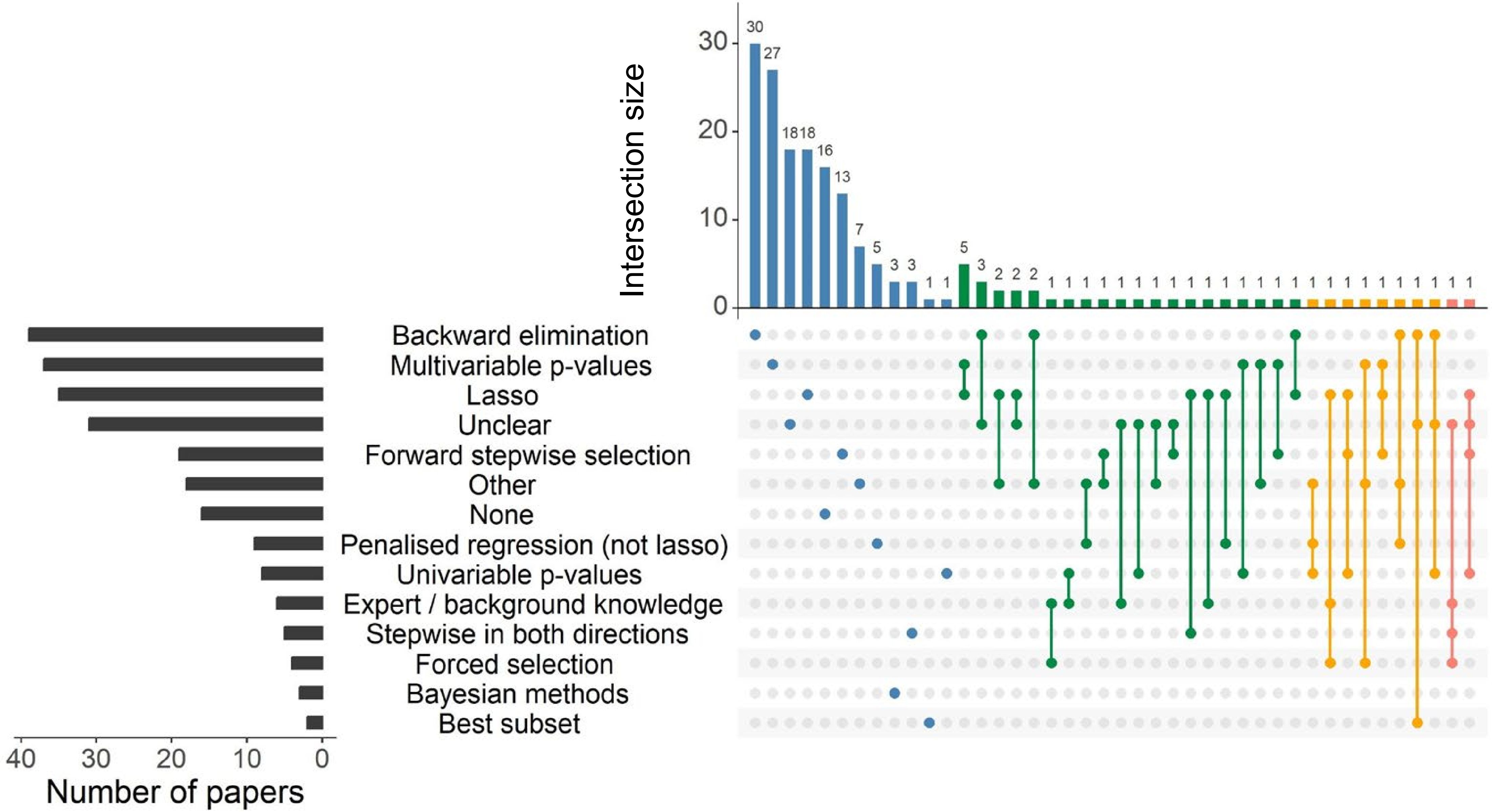
Median sample size 344 (IQR 156 - 982) with median 68 events (IQR 35 - 169)

Results: Modelling patterns

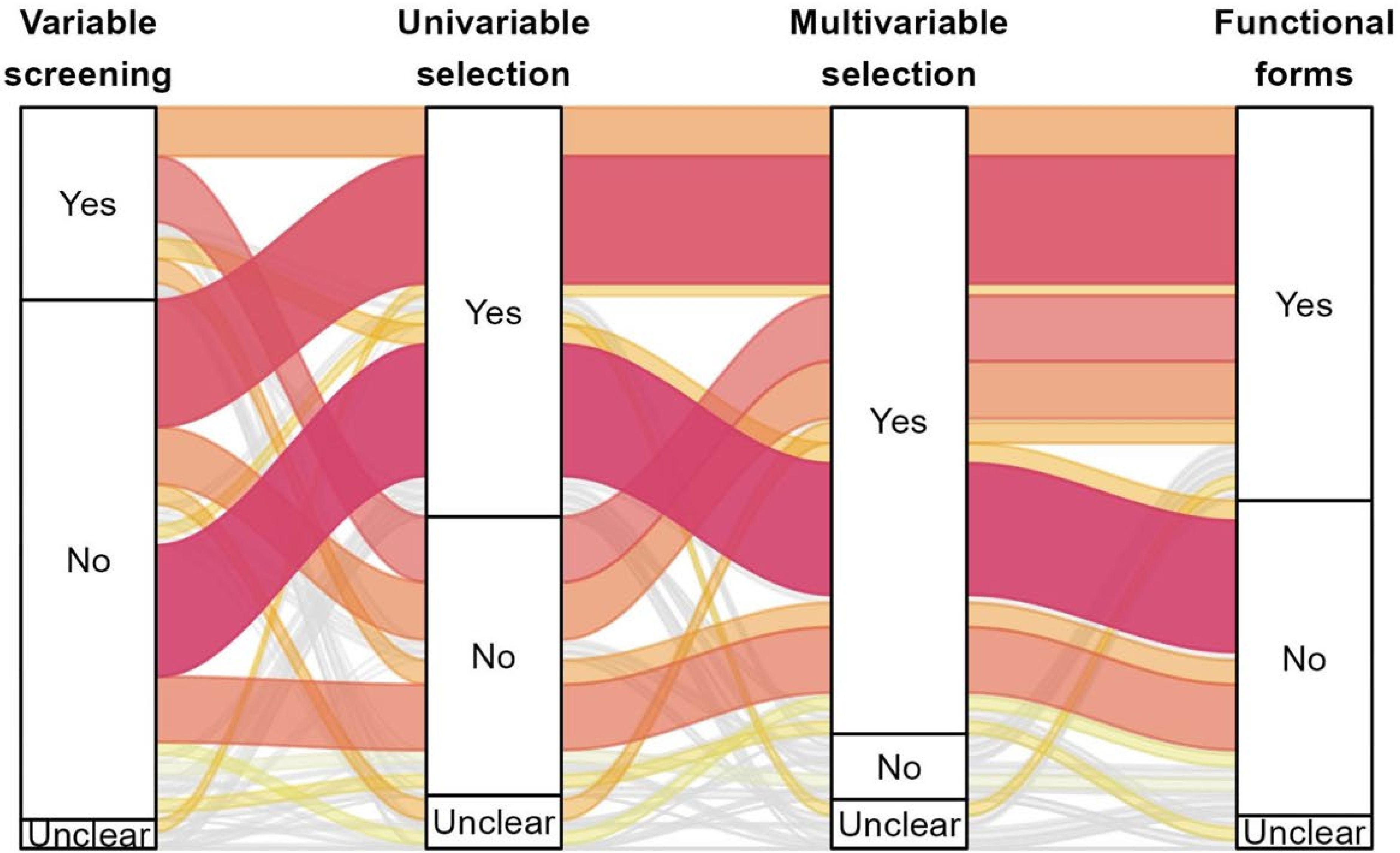


Alluvial plot illustrating the flow of modelling decisions. Flows are color-coded for distinct pathways.

Results: Multivariable selection methods

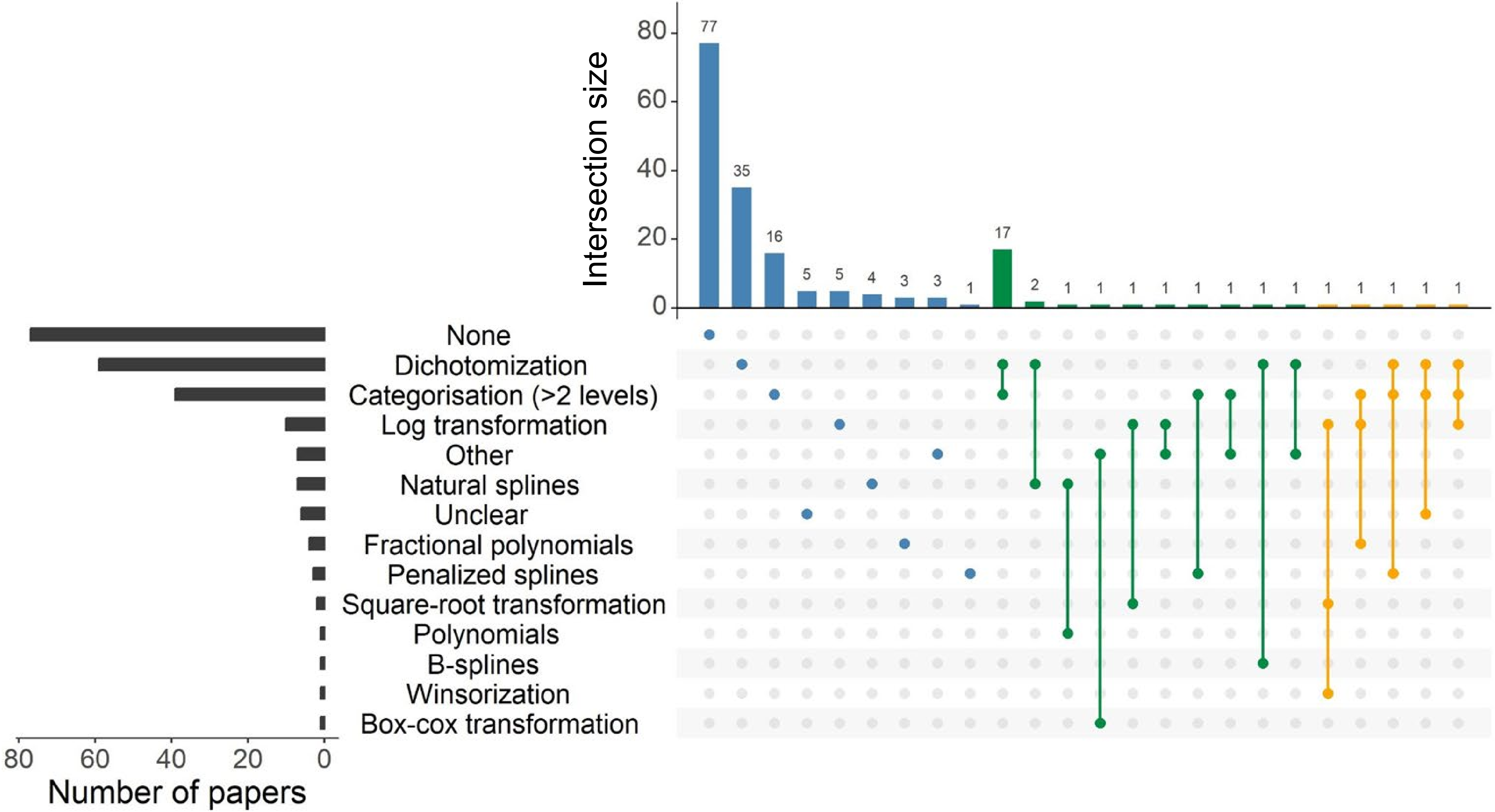


Results: Modelling patterns

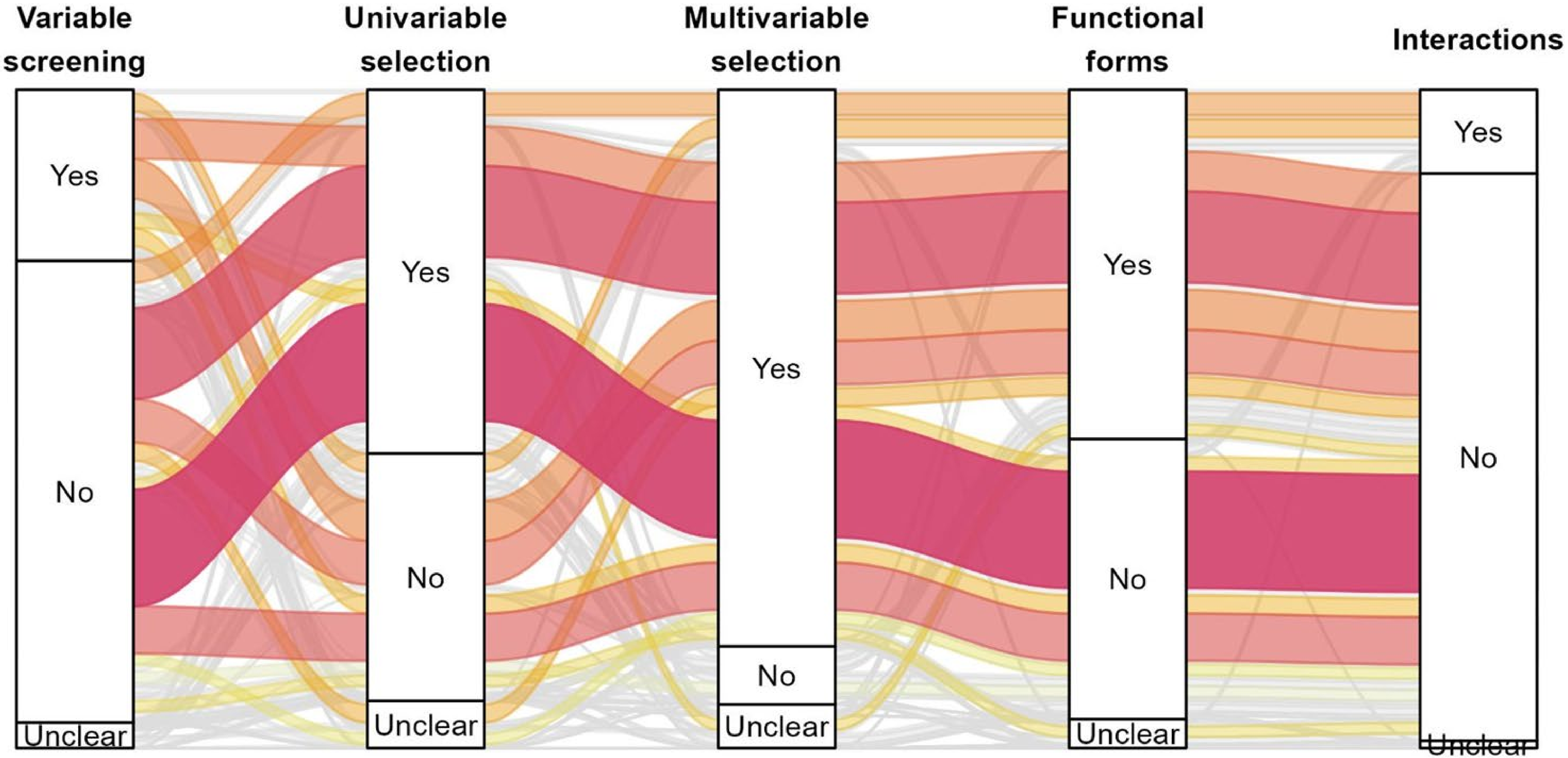


Alluvial plot illustrating the flow of modelling decisions. Flows are color-coded for distinct pathways.

Results: Functional form selection



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

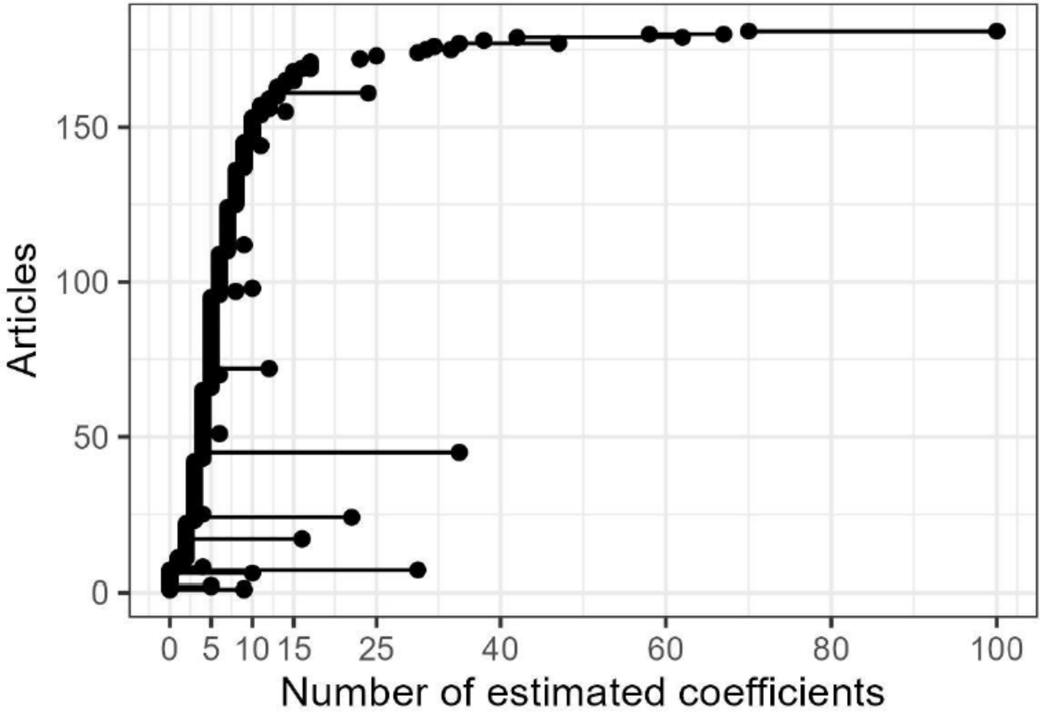
Challenging: Not presented in 29%,

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

Variable selection is common practice…

- Particularly univariable selection (>50% of studies)
- Methods are combined in novel ways that are not investigated in the literature
- Selection is not reflected when reporting inference

…while the use of continuous functional forms and interactions is not.

- Widespread use of dichotomization and categorization (>50% of studies)
- Continuous functional forms rarely used (<10% of studies)
- Functional forms were rarely assessed through variable selection (5% of studies)

Our empirical results underline opportunities for learning, improving guidance and to keep pushing 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 (Heidelberg)

Mariana Nold (Jena)

Moritz Pamminger (Vienna)

Theresa Ullmann (Vienna)

Ulrike Grittner (Berlin)

Willi Sauerbrei (Freiburg)