# SeroTracker-ROB: reproducible decision rules for risk of bias assessment of seroprevalence studies

Niklas Bobrovitz*[1,2], Kim Noël*[3], Zihan Li[4], Christian Cao[5], Gabriel Deveaux[6], Anabel Selemon[5], Mercedes Yanes Lane[7], Tingting Yan[1], Rahul Arora[6,8]

*Contributed equally as co-first authors

**Affiliations:**

1. Temerty Faculty of Medicine, University of Toronto
2. Department of Critical Care Medicine, University of Calgary
3. Faculty of Medicine and Health Sciences, McGill University
4. Faculty of Engineering, University of Waterloo
5. Cumming School of Medicine, University of Calgary
6. Centre for Health Informatics, University of Calgary
7. COVID-19 Immunity Task Force, McGill University
8. Institute of Biomedical Engineering, University of Oxford

**Corresponding author:**
Niklas Bobrovitz
niklas.bobrovitz@mail.utoronto.ca
Medical Sciences Building, 1 King's College Cir
Toronto, ON M5S 1A8

## ABSTRACT

**Background:** Conducting risk of bias assessments for seroprevalence studies is a crucial component of infection surveillance but can be a time-consuming and subjective process. We aimed to develop and evaluate decision rules for transparent and reproducible risk of bias assessments of seroprevalence studies.

**Methods:** We developed the SeroTracker-ROB decision rules, which generate risk of bias assessments for seroprevalence studies from an adapted version of the Joanna Briggs Institute Critical Appraisal Checklist for Prevalence Studies. The decision rules were developed using published guidance on risk of bias assessment for prevalence studies, and the consensus opinions of researchers that have critically appraised thousands of prevalence studies. The decision rules were evaluated against SeroTracker's living systematic review database of SARS-CoV-2 seroprevalence studies. We determined decision rule coverage by calculating the proportion of database studies for which SeroTracker-ROB yielded a risk of bias assessment, and reliability by calculating intraclass correlations between SeroTracker-RoB assessments and the consensus manual judgements of two independent reviewers.

**Results:** The SeroTracker-ROB decision rules for risk of bias assessment classified 100% (n = 2,070) of prevalence studies in SeroTracker's database and showed good reliability compared to manual review (intraclass correlation 0.77, 95% CI 0.74 to 0.80). We developed a tool that implements these decision rules for use by other researchers.

**Conclusions:** The SeroTracker-ROB decision rules enabled rapid, transparent, and reproducible risk of bias assessment of seroprevalence studies, and may serve to support infection surveillance.

**Keywords:** decision rule; automation; prevalence; seroprevalence; risk of bias; critical appraisal

## INTRODUCTION

Seroprevalence studies (serosurveys) are important infection surveillance tools. They use antibody tests to identify previous infection or vaccination,[1] and are therefore crucial tools to measure the true extent of infection, quantify protection as a result of previous infection or vaccination, and inform decisions around public health measures.

Serosurveys have been used for decades, however, the COVID-19 pandemic resulted in an unprecedented increase in the utilization of serosurveys. While these studies have been crucial for the pandemic response, study methods have varied widely.[2] Accordingly, risk of bias (ROB) assessments are crucial to synthesizing and utilizing seroprevalence evidence to inform the scientific response and public health decision-making.

There are several tools for risk of bias assessment of prevalence studies yet there is no consensus on the definitive tool for use in evidence synthesis.[3] Some of the available tools have been validated however, utilizing them can be time-consuming and labor-intensive. Furthermore, even among validated tools there is inherent variation in assessments due to evaluator judgment.[4–6] These are barriers to rapid and reproducible synthesis of seroprevalence studies. Meta-epidemiological reviews of prevalence studies have shown that a unified approach to risk of bias assessment is needed, with emphasis on use of the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Prevalence studies, which has been validated and is the most commonly used critical appraisal tool for prevalence studies.[3]

Therefore, we developed decision rules based on the Joanna Briggs Institute (JBI) Critical Appraisal Checklist to generate transparent and reproducible risk of bias assessments for seroprevalence studies (the SeroTracker-ROB approach) and evaluated their coverage and reliability.

## METHODS

### The SeroTracker-ROB approach and tool

The SeroTracker-ROB approach for risk of bias assessment of seroprevalence involves: (1) a modified version of the standardized Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Prevalence Studies,[7] which can be completed based on extracted data for a given seroprevalence study; (2) decision rules that can be applied to the JBI checklist ratings to generate an overall risk of bias assessment.

#### *Part 1: Completion of the modified JBI checklist*

The JBI checklist was selected as the foundation for the approach as it is a validated and commonly used critical appraisal tool for prevalence studies.[3] In the original JBI checklist, independent reviewers evaluate nine categorical items, which involve judgment about: (1) sample frame appropriateness; (2) sampling method; (3) sample size/calculation; (4) reporting of age/sex; (5) representativeness of sample within analysis; (6) consistent test use; (7) test sensitivity and specificity; (8) appropriate statistical adjustment; and (9) response rate.

We adapted the checklist to seroprevalence studies and added reviewer guidance for each criterion to improve reproducibility, ease decision making burden, and improve specificity (Table 1, Supplementary file 1).

#### *Part 2: Decision rules for risk of bias assessment*

We developed decision rules that could be applied to the ratings from the nine-JBI items to generate an overall risk of bias assessment (low, moderate, or high) for each study (Figure 1). The decision rules were developed based on published guidance on estimating disease prevalence, reports on the evaluation of prevalence studies,[3,8] opinions of experts in evidence synthesis and infectious disease epidemiology, and the consensus of researchers at SeroTracker after evaluation of thousands of seroprevalence studies (Supplementary file 2).[2,9]

### Evaluating the SeroTracker-RoB approach

#### *Dataset*

The SeroTracker-ROB approach was evaluated using data from SeroTracker's living systematic review database of SARS-CoV-2 seroprevalence studies.[9] The protocol for the review was registered and published (PROSPERO: CRD42020183634, version July 21, 2021). Two independent reviewers screen title/abstracts and full texts for inclusion, as well as perform data extraction and manual risk of bias assessments. Disagreements are resolved by consensus or a senior-level arbitrator.

As of November 17, 2021, this database included 2,070 heterogenous studies that had risk of bias assessments from two reviewers, including published and unpublished peer-review literature, pre-prints, institutional reports, and media articles.

### *Statistical analysis*

The coverage of the SeroTracker-ROB decision rules was determined by calculating the proportion of studies for which the approach yielded a risk of bias assessment. Reliability of the ordinal risk of bias ratings was assessed using a two-way random-effects average-measures intraclass correlation measuring absolute agreement that compared the SeroTracker-ROB decision rules to the manually derived ratings agreed upon by two independent reviewers.[10–12] Reliability between the two independent reviewers was also calculated using a two-way random-effects single-measures intraclass correlation measuring absolute agreement. Analyses were conducted using STATA 14 (College Station, TX: StataCorp LP).

## RESULTS

### Modified Joanna Briggs Institute Critical Appraisal Checklist

Four of the nine JBI checklist items were modified to include conditions for binary judgement (i.e., item 4, item 6, item 8, item 9). Item 8 (appropriate statistical analysis) was split in two parts (8a, 8b) to differentiate studies that adjusted prevalence estimates for test characteristics (sensitivity/specificity) but not population demographics and vice versa. As a result, seven of the nine JBI items could be completed automatically from extracted data (Supplementary file 1).

Human reviewers were still required to judge two items: item 1 (whether the sample frame is representative of the target population) and item 5 (whether the characteristics of the sample are representative of the target population in both the main and sub-group analyses). Both these items required consideration of the characteristics of the sample and target populations within the contexts and settings where a given study was conducted. We were unable to automate completion of these items given the context-specific nature of these judgements and heterogeneity between studies leading to many unique circumstances requiring codification. Furthermore, in rare cases where a complex testing algorithm had no clear sensitivity or specificity, consideration of the validity of methods for identification of the condition (Item 6) was determined by manual review of the testing algorithm (Supplementary file 1).

**Decision Rules**

We developed a decision tree to classify overall study risk of bias (Figure 1). The decision tree considered different pathways based on the categorical ratings for each item of the JBI in a forward stepwise manner (i.e, item 1, item 2, etc.).

The tree included 61 decision rules, which could lead to low (n=6), moderate (n=29), or high (n=26) risk of bias. Permutations were grouped in pathways with upward and downward mobility of the overall risk of bias based on differences in critical JBI items along the paths (Figure 1). In general, risk of bias assessments from this tree emphasized whether the seroprevalence estimate for the sample was likely representative of the target population (Items 1, 2, 3, 5 and 8b), and whether the antibody test validly measured seropositivity (Items 6 and 8a). Item 4 was classified as a reporting item and judged to not impact the risk of bias. Some of the 61 decision rules include multiple possible responses to “non-critical items”, which were items that could be rated either as yes, no, or unclear without influencing the overall risk of bias.

More specifically, studies were considered low risk of bias if the sample was likely representative of the target population (i.e., appropriate sample frame, probability sampling, adequate sample size, and statistical adjustment for population characteristics) and the antibody test was valid (i.e., adequate sensitivity/specificity and/or adjustment for test characteristics). Studies were considered moderate risk of bias if the sample was somewhat representative of the target population (i.e., inappropriate sample frame but probability sampling, adequate sample size, and statistical adjustment for population characteristics) or the test was valid (i.e., adequate sensitivity/specificity and/or adjustment for test characteristics). Studies were considered high risk of bias if the sample was not likely representative of the target population (i.e., non-probability sampling and inadequate sample size) and the test used was invalid (i.e., poor sensitivity/specificity and lack of adjustment for test characteristic).

**Coverage**

The SeroTracker-ROB approach yielded a risk of bias assessment for 100% of the SARS-CoV-2 seroprevalence studies (n= 2,070) in the SeroTracker database.

**Reliability**

Intraclass correlation for the reliability of the risk of bias assessment was 0.77 (95% CI 0.74 to 0.80) between the SeroTracker-ROB approach compared to the manual reviewer rating, and 0.74 (95% CI 0.71 to 0.76) between the two manual independent reviewers’ ratings.

### SeroTracker-ROB Excel Tool

An Excel tool that implements the above-described algorithm was developed (Supplementary file 3). This tool takes user inputs for the sampling frame and each JBI item, and computes an automated overall risk of bias assessment. The coding was done using Visual Basic for Applications in Excel.

## DISCUSSION

The SeroTracker-ROB approach to risk of bias assessment of seroprevalence studies had excellent coverage (100%) and good reliability (ICC 0.77) compared to consensus assessments yielded by two independent reviewers.

With manual review, the inter-rater reliability for risk of bias for two independent reviewers was moderate (ICC 0.74), emphasizing the heterogeneity of risk of bias assessments.[4,5] As such, manual review cannot be considered a perfect standard, and the 0.77 intra-class correlation between the SeroTracker-RoB approach and manual review in part reflects the heterogeneity and inconsistency of manual assessment. This highlights the benefit of an automated approach that yields transparent and reproducible assessments.

There is no consensus on the most valid approach to automate risk of bias assessments.[3] Summary scores of critical appraisal checklists have been utilized to determine overall risk of bias ratings. However, this approach is limited because each item of a given critical appraisal tool does not equally impact on risk of bias.[13–15] For example marking six out of nine JBI items “yes” could be achieved by different combinations of item responses, each of which have different implications for risk of bias. Weighted averages overcome some limitations of simple summary scores, as they may better reflect the relevance of each item; however, items cannot always be considered independently, thereby introducing complexity in the derivation of a weighted score. Considering combinations of items may provide a more nuanced perspective on bias. For example, statistical adjustment for antibody test sensitivity and specificity may introduce less bias when sensitivity and specificity are very high, as opposed to low.

Some studies have trained multi-task machine learning algorithms to identify text relevant to risk of bias in randomized controlled trials and predict a risk of bias assessments using the text, with reasonable accuracy compared to human reviewers.[16,17] However, key limitations of the algorithms include training on small datasets, imperfect reliability with manual reviewers, inaccessible software to employ

these algorithms, and limited interpretability and transparency of the “black-box” decision making of the models.

A decision rule approach, on the other hand, provides a transparent and interpretable model for automation of study risk of bias decisions.[18] The tree structure is an ideal approach for capturing interactions between features in the data and clustering data points into distinct groups that are fairly easy to understand given that they can be visualized.

The decision rules developed as part of the SeroTracker-ROB approach may be of value to other investigators conducting risk of bias assessments for SARS-CoV-2 seroprevalence studies. This may be of importance given the endemic trajectory of SARS-CoV-2 and likely ongoing value of seroprevalence studies as infection surveillance tools.[19] The decision rules may also provide value for investigators conducting prevalence studies for other diseases. Therefore, we have embedded these decision rules into the SeroTracker-ROB Excel Tool (Supplementary file 3) which aims to bring value to researchers and public health stakeholders in terms of methodological robustness, research reproducibility, and efficiency. However, users should be aware that some of the JBI items in the tool may need to be adapted to suit user needs. For example, valid methods of detecting SARS-CoV-2 were judged according to assays meeting the 90% sensitivity and 97% specificity thresholds established by the World Health Organization (REF), but this criterion will vary for other conditions.

This study had several strengths. Firstly, to our knowledge, this is the first validation of decision rules for risk of bias in seroprevalence studies. Secondly, the decision rules have a robust foundation in published guidance on estimating disease prevalence, opinions of methodological experts, and the experiences of researchers that have conducted thousands of risk of bias assessments for seroprevalence studies. Lastly, the approach was validated on a large sample of heterogeneous SARS-CoV-2 seroprevalence studies from SeroTracker’s living systematic review database, including peer-reviewed literature, pre-prints, institutional reports, and media articles.

This study had several limitations. First, the SeroTracker-ROB approach is not a fully objective process as it still requires reviewers to make subjective judgments on two JBI items, and, in rare circumstances, a third judgement for studies using complex antibody testing algorithms without a calculable sensitivity and specificity. However, the Joanna Briggs Institute provides guidance for making these assessments and researchers can comment on their decision making for these items to improve transparency. Second, the decision rule was derived using subjective judgement. Iteration of the decision rules may maximise their validity; however, the transparency of the process allows for

scientific debate and further refinement of the decision rules. Thirdly, the generalizability of the reliability estimates for the SeroTracker-ROB decision rules is unclear. Although the SeroTracker database includes a diverse range of studies, they all focus on SARS-CoV-2 seroprevalence. The SeroTracker-ROB approach may require validation using (sero)prevalence studies of other conditions.

## Conclusions

We developed and evaluated the SeroTracker-ROB decision rules to derive risk of bias assessments for seroprevalence studies. The decision rules had complete coverage and were reliable compared to two independent human reviewers. The SeroTracker-ROB decision rules can support transparent and reproducible evidence synthesis of infection prevalence studies, and may have particular value during outbreaks and pandemics.

## Declaration of interests

All authors report grants from the WHO COVID-19 Solidarity Response Fund (German Ministry of Health), grants from the Public Health Agency of Canada through Canada's COVID-19 Immunity Task Force, grants from the Robert Koch Institute, and grants from the Canadian Medical Association, during the conduct of the study.

## REFERENCES

1. Cheng MP, Yansouni CP, Basta NE, Desjardins M, Kanjilal S, Paquette K, et al. Serodiagnostics for Severe Acute Respiratory Syndrome-Related Coronavirus 2 : A Narrative Review. Ann Intern Med. 2020 Sep 15;173(6):450–60.
2. Bobrovitz N, Arora RK, Cao C, Boucher E, Liu M, Donnici C, et al. Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis. Khudyakov YE, editor. PLoS ONE. 2021 Jun 23;16(6):e0252617.
3. Borges Migliavaca C, Stein C, Colpani V, Barker TH, Munn Z, Falavigna M, et al. How are systematic reviews of prevalence conducted? A methodological study. BMC Medical Research Methodology. 2020 Apr 26;20(1):96.
4. Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. Journal of Clinical Epidemiology. 2013 Sep;66(9):973–81.
5. Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. Journal of Clinical Epidemiology. 2013 Sep;66(9):982–93.
6. Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C, et al. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. Journal of Clinical Epidemiology. 2012 Sep;65(9):934–9.
7. Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc. 2015 Sep;13(3):147–53.
8. Migliavaca CB, Stein C, Colpani V, Munn Z, Falavigna M. Quality assessment of

prevalence studies: a systematic review. Journal of Clinical Epidemiology. 2020 Nov;127:59–68.
9. Arora RK, Joseph A, Wyk JV, Rocco S, Atmaja A, May E, et al. SeroTracker: a global SARS-CoV-2 seroprevalence dashboard. The Lancet Infectious Diseases [Internet]. 2021 Apr 1 [cited 2021 Mar 14];21(4). Available from: https://doi.org/10.1016/S1473-3099(20)30631-9
10. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016 Jun;15(2):155–63.
11. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):30–46.
12. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979;86(2):420–8.
13. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. International Journal of Epidemiology. 2007 Jun 1;36(3):666–76.
14. Jüni P. The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. JAMA. 1999 Sep 15;282(11):1054.
15. Greenland S. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001 Dec 1;2(4):463–71.
16. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association. 2016 Jan 1;23(1):193–201.
17. Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. Int J Epidemiol. 2016 Feb;45(1):266–77.
18. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. [Internet]. 2019. Available from: https://christophm.github.io/interpretable-ml-book/
19. European Centre for Disease Prevention and Control. COVID-19 surveillance guidance: Transition from COVID-19 emergency surveillance to routine surveillance of respiratory pathogens [Internet]. Stockholm; 2021 Oct. Available from: https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-surveillance-guidance.pdf
20. Bewley KR, Coombes NS, Gagnon L, McInroy L, Baker N, Shaik I, et al. Quantification of SARS-CoV-2 neutralizing antibody by wild-type plaque reduction neutralization, microneutralization and pseudotyped virus neutralization assays. Nat Protoc. 2021 Jun;16(6):3114–40.

## Table 1. Modified Joanna Briggs Institute Checklist for Prevalence Studies

| **Joanna Briggs Institute checklist items[a]** | **SeroTracker-ROB standardized guidance for checklist items [response options]** | **Notes** |
|---|---|---|
| Item 1: Was the sample frame appropriate to address the target population? | Sample frame described and it approximated the target population [yes / no] | Requires subjective evaluation. See Supplementary file 1 of JBI Guidance for full details |
| Item 2: Were study participants recruited in an appropriate way? | Probability sampling (simple or stratified random) OR entire sample (e.g., an entire town) [yes / no / unclear] | |
| Item 3: Was the sample size adequate? | ≥599 OR sample size calculation was provided AND the required sample for 80% power was below the threshold (n<599) OR precision was 1.25% [yes / no / unclear] | See Supplementary file 1 for sample size calculation. |
| Item 4: Were the study subjects and setting described in detail? | Average age AND proportion of gender/sex provided [yes / no] | |
| Item 5: Was data analysis conducted with sufficient coverage of the identified sample? | The demographic characteristics (gender/sex, age, and ethnicity) of the sample are at least somewhat representative of the population in both the main and sub-group analyses [yes / no / unclear] | This item requires subjective review of demographic breakdown in the study, coverage within subgroup estimates, and review of author comments on representativeness of the sample |
| Item 6: Were valid methods used for the identification of the condition? | The serological test used did meet the WHO Unity Study Criteria for serological tests: sensitivity minimum 90% AND specificity minimum 97%. The combined sensitivity of multiple testing algorithms must also meet this threshold, or the testing algorithm must be logical and robust for algorithms without a calculable combined sensitivity and specificity [yes / no] | Exceptions are made for algorithms using a commercial or in-house binding assay with confirmatory testing using virus neutralization assay, as they constitute the gold standard in serological evaluation.[20] |
| Item 7: Was the condition measured in a standard, reliable way for all participants? | The same serology test was used for all participants [yes / no / unclear] | |
| Item 8A: Provides statistical adjustment for test characteristics and provides the information necessary to determine the numerator, denominator, prevalence estimate, and confidence interval | Conducted test adjustment AND reported relevant information OR Item 6 was “Yes” [yes / no / unclear] | |

| | | |
|---|---|---|
| Item 8B Provides statistical adjustment for population characteristics or the sample is somewhat representative of the population [probability sampling] and provides the information necessary to determine the numerator, denominator, prevalence estimate, and confidence interval | Conducted population adjustment AND reported relevant information OR Item 1, 2, and 3 were all "Yes" [yes / no / unclear] | |
| Item 9: Was the response rate adequate, and if not, was the low response rate managed appropriately? | Response rate > 60% or the demographics of the sample were a reasonable match to those of the target population [yes / no / unclear] | |
| **Overall risk of bias** | **Low:** The estimates are very likely correct for the target population. To obtain a low risk of bias classification, all criteria must be met or departures from the criteria must be minimal and unlikely to impact on the validity and reliability of the prevalence estimate. These include sample sizes that are just below the threshold when all other criteria are met, reporting only some of characteristics of the sample, test characteristics below the threshold but corrections for the test performance, and response rates that are just below the threshold in the context of probability based sampling of an appropriate sampling frame with population weighted seroprevalence estimates.<br>**Moderate:** The estimates are likely correct for the target population. To obtain a moderate risk of bias classification, most criteria must be met and departures from the criteria are likely to have only a small impact on the validity and reliability of the prevalence estimates.<br>**High:** The estimates are not likely correct for the target population. To obtain a high risk of bias, many criteria must not be met or departures from criteria are likely to have a major impact on the validity and reliability of the prevalence estimates.<br>**Missing:** There was insufficient information to assess the risk of bias. | |
| **[a]Item 1-9 are a modified version of the Joanna Briggs Institute Critical Appraisal Checklist for prevalence studies** | | |

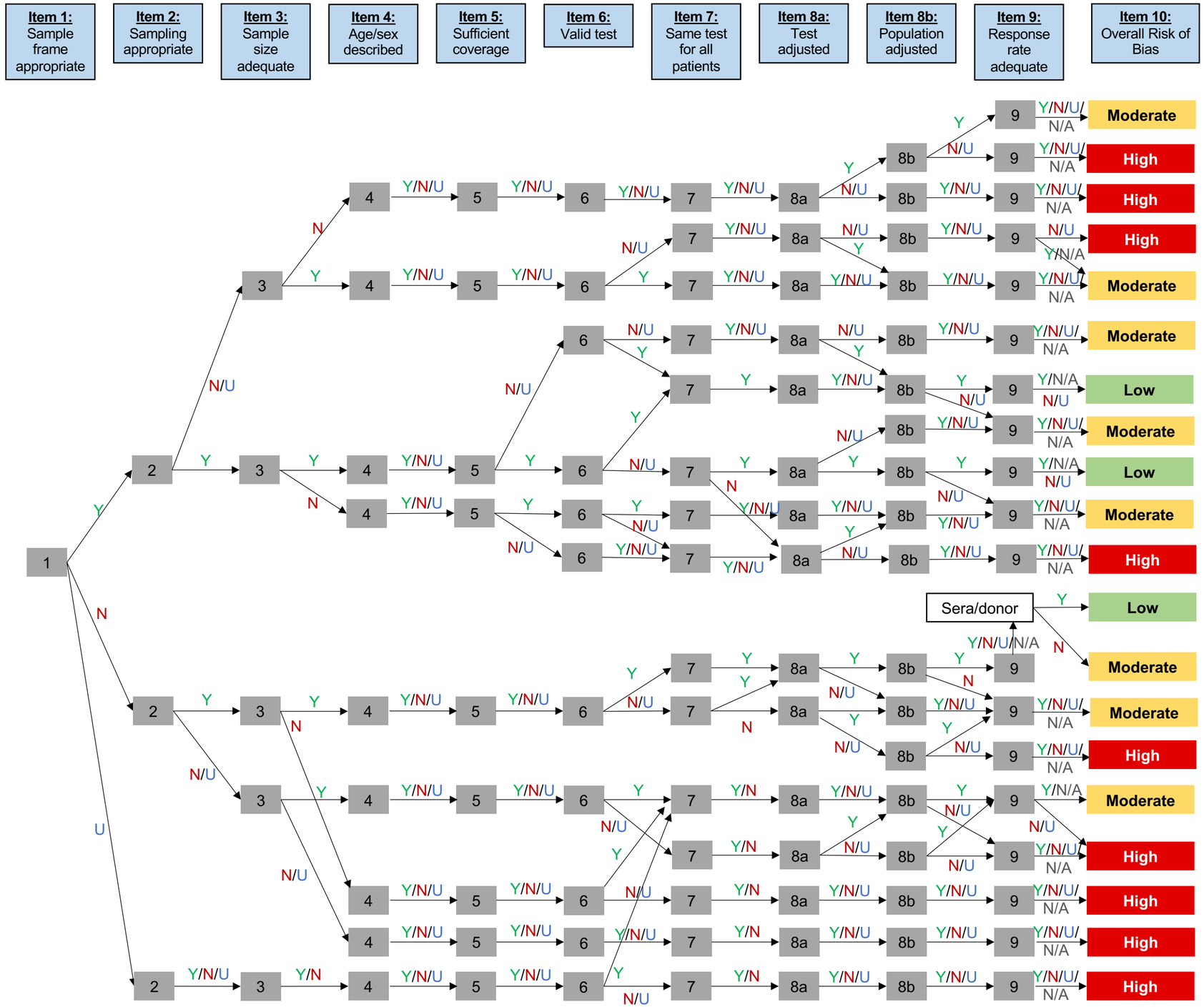


Figure 1. Decision rules to determine risk of bias for seroprevalence studies