The UK Biobank Study: Systematic Review of the Reporting Quality in UK Biobank Studies

Emma Anderson, George Davey Smith, Katie Drax, Mark Gibson, Marcus Munafò, Robert Thibault, Benjamin Woolf, and Rebecca Richmond

25 November 2021

# 1 Background

There is a general move towards making data freely accessible, or ‘open’, by academic and governmental institutions alike (Vasilevsky, Minnier, Haendel, & Champieux, 2017; World Wide Web Foundation, 2018). Completely open data allows any individual, regardless of their motivation, discipline or skillset, to access and analyse a dataset for free. This accessibility offers researchers new opportunities but there is also uncertainty about how the nature of open data may impact the reliability of research findings. One method researchers can use to investigate the reliability of results from open data is replication. In the case of secondary data analyses replications will involve new researchers retrieving the same dataset as the previous researchers and repeating their analyses. Researchers can make such direct replications of their work easier by publishing a detailed description of their statistical analysis. However, recent evidence suggests that it is unclear whether a detailed plan alone is sufficient to ensure direct replicability of findings (Hardwicke et al., 2018; Naudet et al., 2018).

To investigate the reliability of findings from open data, we will attempt to replicate all eligible research articles that use UK Biobank (UKB) data. The UKB is a prospective cohort study which collected a wide range of health data from 500,000 people aged 40-69 in the UK during 2006-2010 (UK Biobank Coordinating Centre, 2007). Importantly, UKB data is “controlled data”, meaning it is only available to bona fide researchers conducting research in the public interest on a cost-recovery basis. The UKB reserves the right to reject access applications for projects it deems inappropriate (UK Biobank Coordinating Centre, 2011). We plan to replicate UKB studies specifically because there is an abundance of published articles using the data and they are all contemporary, which increases the likelihood that authors will be contactable. Furthermore, they all use the same, professionally collected, maintained and diverse open resource which we will be able to access and use.

We first conducted a systematic review to (1) assess the completeness of UKB study reporting, (2) identify articles that are suitable for replication, and (3) inform the development of reporting guidelines for other articles using the UKB and similar resources. Many previous systematic reviews of reporting quality across a range of disciplines used guidelines to assess the completeness of reporting [e.g. Agha et al. (2014); Cook, Levinson, & Garside (2011); Plint et al. (2006)). We used the STROBE (STrengthening the Reporting of Observational Studies in Epidemiology) Statement, which provides reporting guidelines for observational studies, to evaluate articles arising from UKB and to make comparisons between the articles. The STROBE Statement has been endorsed by over 100 journals and, as no major changes have been made to the original version published in 2007 (Cevallos & Egger, 2014), these guidelines are relevant to all studies conducted using the UKB. Given that studies conducted using UKB may be cohort, cross-sectional or ‘nested’ case-control, the review used the combined STROBE Statement for cohort, cross-sectional and case-control designs, and its extensions, to assess the reporting quality in UKB articles.

We conducted a systematic review of articles reporting studies using data from the UKB with the aim of answering the following questions:

1. What is the quality of reporting of findings from the UKB (in terms of the detail, clarity and completeness with which authors report their design, analysis and results)?
2. Does reporting quality vary across the type of studies conducted?

# 2 Methods

## 2.1 Protocol registration

This study was preregistered before we began to screen the search results (Drax, Richmond, Woolf, Smith, & Munafo, 2019). For a description of the deviations from the registered protocol see the Supplementary Material.

## 2.2 Inclusion criteria

The inclusion criteria were:

* Publication date: After 30 March 2012. This is the date the UKB was launched and UKB data was first available for researchers to use.
* Publication type: Full research articles (including simulation and modelling articles).
* Language: English. Relevant studies written in another language may be relevant but, to ensure accurate and precise evaluation, the article would require a professional translation which is not feasible given the review’s resources.
* Data source: Article’s findings were produced from analysis(es) of data from the UKB. This includes studies which reported findings from other studies alongside findings from UKB data.
* Study type: Observational epidemiology studies
* Study design: Cross-sectional, cohort, or case-control study design.

The exclusion criteria were:

* Article findings were produced from analysis(es) of pooled data from multiple studies. These are studies which have used data from participants from the UKB and other datasets in the same analysis.
* Meta-analyses; narrative and systematic reviews; preprints; post-prints; replies; letters to the editor; corrections; and any other publication types that are not full research articles.
* Retracted articles
* Studies with inaccessible supplementary material. This was to ensure our estimates of reporting completeness included the entire version of record.

## 2.3 Information sources and search strategy

We searched four databases on 15/01/2019 for UKB articles: PubMed, EMBASE, Web of Science Core Collection (WoSCC) and PsycINFO. The search strategy was limited to articles published from 2012 onwards and there was no restriction on language or publication type.

The full search strategy for each database is included in Appendix A of the protocol (Drax et al., 2019). To summarise, the terms “UK Biobank”, “UKB”, “UKBiobank”, and “UKB Resource” were included in the “Title”, “Abstract”, “Keywords”, and “All Fields” fields or equivalent of each database.

## 2.4 Study selection

We conducted three stages of screening for eligibility. First, using Endnote, KD removed duplicate results, checked the results for completeness, updated the metadata if necessary, and retrieved the full-texts of all articles. KD updated the metadata by checking each article’s Endnote metadata contained the article’s DOI, and adding them if they were missing. KD checked completeness using a list of published full research articles that have used the UKB Resource, provided by the UKB on 27 November 2018. We conducted a systematic search despite having this list because we wished to capture all articles from 2018. Using the full-texts of the results, KD assessed if results were full research articles, and excluded reviews, corrections, conference abstracts, etc.

Second, we imported the results into Rayyan (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016). BW and KD independently screened the results and indicated if they met the inclusion criteria. At this point we included all articles whether or not they had a cross-sectional, cohort, or case-control study design because the “Study design” inclusion critera was added later. They examined full texts if inclusion was unclear from the abstract alone. They discussed any conflicts and resolved them by mutual consent if possible, if not resolved after discussion then RR assigned the category. If insufficient information was available in the full text to determine a paper’s eligibility, we contacted the authors and requested the required information, if it was unobtainable we excluded the article. Time constraints meant we excluded XXX articles with insufficient information without contacting the authors.

## 2.5 Data collection process

### 2.5.1 Study Design Form

Different STROBE items are relevant for cohort, case-control and cross-sectional study designs. To ensure coders completed the same STROBE items for each article KD and MG independently identified the study designs used by each of the 178 “observational epidemiology” articles by completing the Study Design Form for each article. In the form, coders also indicated if all supplementary material can be accessed via University of Bristol subscriptions. Any conflicts were discussed by the two coders and resolved by mutual consent if possible. If not resolved after discussion, RR determined the study designs. The Study Design Form was hosted on Qualtrics but a PDF version can be found at <https://osf.io/jfk24/>. During this process we also identified studies we did consider to be observational epidemiological studies but could not consider them to have cohort, case-control or cross-sectional designs - for example those using prediction models. Any articles we could not classify as cohort, case-control, or cross-sectional were excluded at this point.

The articles were randomly assigned an ID number from 1-178. We selected the articles with an ID number between 1 and 80 for the first round of data extraction. We assigned KD as the first coder for all 80 articles and randomly assigned MG, RR, or BW as the second coder. We initially randomly selected 80 articles with the hope that we would randomly select more studies once data extraction for the first 80 were complete. Unfortunately we did not have time to assess more studies so we do not have data on the remaining 98 articles. Figure 1 shows the flow of records identified from our searches and reasons for exclusion.

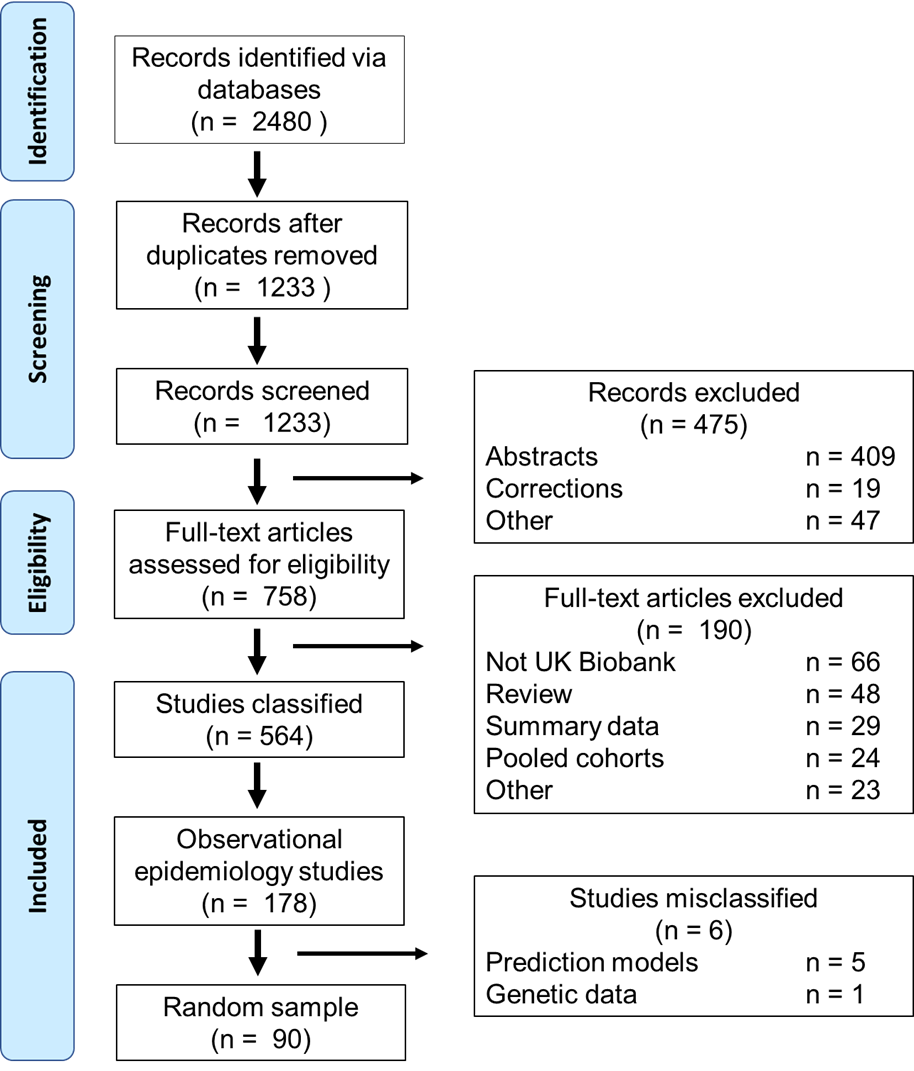


Figure 1. Flow diagram of included articles

### 2.5.2 Data Extraction Form

Two coders independently completed the data extraction form for each article assigned to them. Almost all variables in the form were categorical so coders assigned the most appropriate category for each variable. We conducted two pilots. In the first, MG, KD and/or RR piloted 5 observational epidemiology articles. We made major changes to the Data Extraction Form so we completed it again for the articles in the first pilot. In the second, KD piloted the changed form on 3 articles. We made no major changes so we kept the responses for articles in the second pilot. The PDF version of the data extraction form contains full information on our variables and how we measured them and is available at <https://osf.io/hngcj/>.

The form contained questions to measure variables relating to the reporting quality of different information. We derived some variables from the STROBE guidelines. Some STROBE items were written as multiple-barrelled questions e.g " Report assessment methods for outcomes, exposures, predictors, confounder, and effect modifiers". We subdivided these questions into separate variables in the data extraction form. We could have assessed each item without dividing it up but collecting the reporting completeness at the subdivision level we have a more granular understanding of what information is or is not reported.

Variables derived from the STROBE items had the same six categories:

* “NA = this information was not relevant to any of the studies in the article”
* “Unsure = I am unsure what to respond”
* “No = this information was relevant but none of it was present”
* “Partially = some of the information was present”
* “Partially-External = some of the information was present and/or the authors report that additional information can be found in an external source (i.e. not in the article or supplementary material)”
* “Yes = all of the information was present”.

After two coders completed one data extraction form for each article, the two coders discussed any conflicts and resolved them by mutual consent if possible. If not resolved after discussion, RR resolved the conflict. However, given the number of conflicts, this practice could not be sustained. To reduce the time-burden on other authors KD and RR devised a series of “rules” and general guidelines KD could use to resolve conflicts by themself. This means KD resolved the majority of conflicts by these “Rules”. For example, for item 1a of the STROBE checklist we agreed that authors describing a study as a “longitudinal” is not sufficient according to the STROBE guidelines but it still gives some indication of the study design. Therefore, we agreed on the rule “Rule =”Partially" if “longitudinal” referred to" for item 1a. The rules and guidelines can be found at <https://osf.io/hngcj/>.

### 2.5.3 Other data collection

The items in Table 2 were extracted manually by KD or automatically by Endnote which contained the article’s metadata exported from the database the article was located in.

### 2.5.4 Unclear and Missing Information.

In our protocol we planned to locate and examine publications linked to the article, e.g. related full research articles, comments, corrections, etc, so that we could retrieve any missing items in the Data Extraction Form from them. If the missing information is not contained in a linked publication we planned to contact the article’s authors. However, we realised this would give an inaccurate assessment of the version of record. Also few articles had linked publications and there was a lot of missing or unclear information. Therefore, we did not attempt to retrieve any missing items in the Data Extraction Form, from either the authors or linked articles.

## 2.6 Data analysis

Using the data extraction form on Qualtrics, we collected data on the reporting quality on numerous variables, including those derived from the STROBE items. This mean our raw data estimated the reporting quality for all subdivisions of the STROBE items and for the “starred” items associated with any STROBE items that had an asterisk (\*) next to it. For the final analysis, we excluded all starred subdivisions because inter-rater reliability was so poor, this was primarily because raters could not agree when the items were or were not applicable. In the data extraction form raters could indicate items as “NA”, “Unsure”, “No”, “Partially”, “Partially-External” or “Yes”. After resolving conflicts, no “Unsure” ratings remained. We collapsed the “Partially-External” into the “Yes” catergory. This was because we agreed reporting there were very few “Partially-External” categories after resolving conflicts, to make the kappa statistics more interpretable, and we understood that reporting where the required information can be found could be considered complete reporting. We summed the subdivisions back together to calculate the reporting completeness of each STROBE item they corresponded to. This gave us an understanding of the reporting quality for the entire STROBE item and reduced the number of variables from 77 to 34.

We calculated Cohen’s kappa statistics to estimate the inter-rater reliability for each of the STROBE items. We treated any “Not Applicable” or “Unsure” values as missing data in the kappa calculations.  
For our subgroup analyses we repeated the analyses for the subdivisions. We conducted all our data cleaning and analysis in R.

# 3 Results

## 3.1 Study characteristics

Of the 70 included articles 4 contained case-control designs, 33 cohort, and 38 cross-sectional. These numbers add up to more than 70 because some articles contained cross-sectional and cohort studies. UK based researchers led the vast majority with 53 having authors affiliated with institutions in the UK. The mean number of authors was 7 (SD = 3). 3 studies were published in 2014, 6 in 2015, 10 in 2016, 17 in 2017, 30 in 2018, and 4 in 2019.

## 3.2 STROBE Reporting quality

Figure 1 presents the reporting completeness for each STROBE item, to ensure a comprehensive understanding of our results, Supplementary Table 1 provides the reporting completeness for each subdivision.

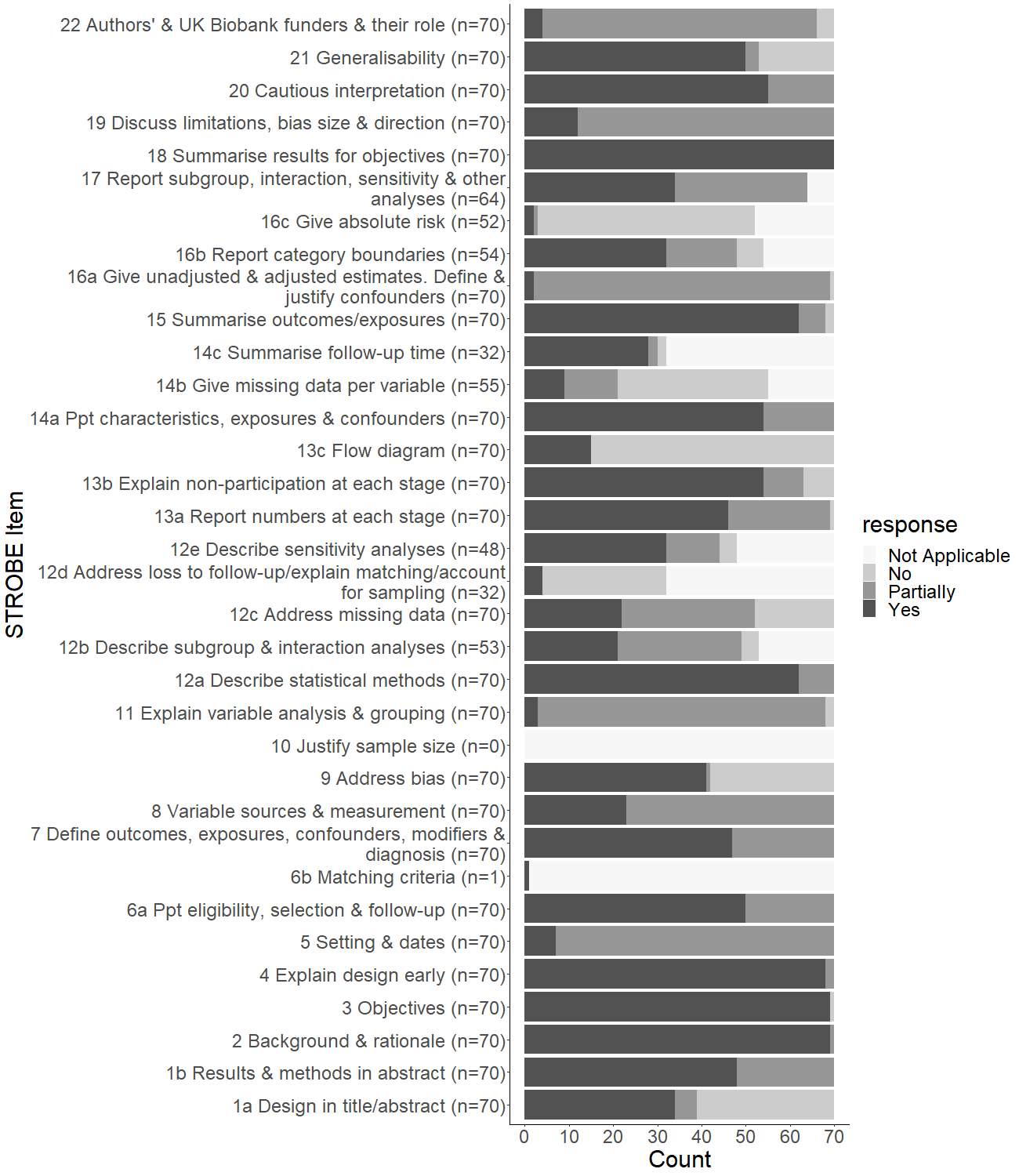


Figure 2. STROBE completion

Table 1 presents Cohen’s kappa statistics for each of the STROBE items. Many items had poor inter-rater reliability with 14 items having a kappa below 0.2. The main reason for this is that the data extraction form was subjective. Conflicts occured when one coder did not notice that the required information was reported and when coders interpretted the STROBE item differently. The list of “rules” we compiled give an indication for conflicts that occurred more than once, how we decided to resolve them and our justification for it.

Table 1. Inter-rater reliability for the STROBE items

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STROBEItem | Kappa | Zstatistic | Matches | Agreement | NotApplicableUnsure |
| 1a | 0.56 | 5.66 | 52 | 75 | 1 |
| 1b | 0.3 | 2.67 | 49 | 70 | 0 |
| 2 | -0.08 | -0.78 | 57 | 81 | 0 |
| 3 | 0.32 | 4.99 | 66 | 94 | 0 |
| 4 | -0.03 | -0.75 | 43 | 61 | 0 |
| 5 | 0.03 | 0.28 | 49 | 70 | 0 |
| 6a | 0.38 | 3.23 | 51 | 73 | 0 |
| 6b | 0 | NA | 0 | 0 | 69 |
| 7 | 0.14 | 1.19 | 41 | 59 | 0 |
| 8 | 0.07 | 0.9 | 30 | 43 | 0 |
| 9 | 0.24 | 2.27 | 41 | 59 | 1 |
| 10 | NA | NA | NA | NA | NA |
| 11 | 0.02 | 0.52 | 42 | 62 | 2 |
| 12a | 0.19 | 1.88 | 57 | 81 | 0 |
| 12b | 0.26 | 2.4 | 28 | 56 | 20 |
| 12c | 0.37 | 4.54 | 38 | 57 | 3 |
| 12d | 0.55 | 3.14 | 23 | 85 | 43 |
| 12e | 0.16 | 1.36 | 25 | 64 | 31 |
| 13a | 0.3 | 3.44 | 44 | 63 | 0 |
| 13b | 0.29 | 3.35 | 47 | 67 | 0 |
| 13c | 0.87 | 7.55 | 67 | 96 | 0 |
| 14a | 0.26 | 2.43 | 55 | 79 | 0 |
| 14b | 0.39 | 3.8 | 39 | 74 | 17 |
| 14c | 0.37 | 2.72 | 25 | 78 | 38 |
| 15 | 0.13 | 1.84 | 43 | 62 | 1 |
| 16a | 0.18 | 1.66 | 63 | 90 | 0 |
| 16b | 0.06 | 0.6 | 23 | 55 | 28 |
| 16c | 0.55 | 5.06 | 38 | 93 | 29 |
| 17 | 0.19 | 1.61 | 34 | 58 | 11 |
| 18 | 0 | 0 | 68 | 97 | 0 |
| 19 | 0.18 | 1.53 | 43 | 61 | 0 |
| 20 | 0.09 | 1.11 | 51 | 74 | 1 |
| 21 | 0.51 | 5.16 | 52 | 74 | 0 |
| 22 | 0.5 | 5.74 | 59 | 84 | 0 |

## 3.3 Other reporting quality

We collected reporting quality data in addition to the STROBE items. We intended to conduct exploratory analyses using these variables to answer our second research question, for example comparing articles that reported using some form of reporting guidelines to those who did not. We did not conduct an exploratory analyses into the relationships between these variables because subgroup analyses for all of these variables would have very small subgroup samples. Instead we report the descriptive statistics for each variable in Table 2 below.

*Table 2. Descriptive statistics for additional reporting quality variables*

|  |  |
| --- | --- |
| **Variable** | **N (%)** |
| Report shared analysis code = Yes | 2 (2.9) |
| Exposure variables clearly labelled as exposures = Yes | 40 (57.1) |
| Report using reporting guidelines other than STROBE = Not Applicable | 70 (100.0) |
| Outcome variables clearly labeled as outcomes = Yes | 51 (72.9) |
| Report analyses registered = Yes time of registration unknown | 2 (2.9) |
| Report using STROBE guidelines = yes | 3 (4.3) |
| Cite STROBE guidelines if used | |
| No | 2 (2.9) |
| Not Applicable | 67 (95.7) |
| Yes | 1 (1.4) |
| Report UK Biobank credit statement correctly | |
| no | 11 (15.7) |
| Yes - almost exact wording | 22 (31.4) |
| Yes - exact wording | 37 (52.9) |
| Report UK Biobank variable IDs | |
| Always | 1 (1.4) |
| Most | 2 (2.9) |
| Never | 60 (85.7) |
| Sometimes | 7 (10.0) |
| Has supplementary material = Yes | 49 (70.0) |
| Published as open access | |
| No | 7 (10.0) |
| No but publicly available on another platform | 11 (15.7) |
| Yes | 52 (74.3) |
| Has a correction = Yes | 3 (4.3) |
| Report UK Biobank application number = Yes | 41 (58.6) |
| Has email address for corresponding author = Yes | 69 (98.6) |
| Has keywords = Yes | 40 (57.1) |
| Report conflict of interest statement = Yes | 62 (88.6) |

# 4 Discussion

## 4.1 Summary

Our results indicate that the STROBE items best reported by UK Biobank studies are those regarding backround (item 2), objectives (item 3), explaining design early (item 4), main statistical methods (item 12a), and summarising results (item 18) with nearly all articles reporting each item fully. The good reporting of the main statistical methods is encouraging for the potential reproducibility of the analysis methods. Most articles also fully reported information on generalisability (item 21).

Partial reporting dominated items regarding setting, recruitment, location, and dates (item 5); assessment methods (item 8); variable analysis and grouping (item 11); addressing missing data (item 12c); main results and definitions of confounders (item 16a); limitations and bias (item 19); and funder details (item 22). For some items partial reporting is not a major concern. Although it does not meet the STROBE guidelines, all information for item 5 and 8 is fully reported in the publicly available UK Biobank documentation. For others, partial reporting casts doubt on the ability of new researchers to know how to repeat the included studies. Authors rarely fully reported how they handled their variables (item 11) and missing data (item 12c). Another potential area of concern is the reporting of sensitivity, subgroup, interaction, and other analyses. Most reporting of the methods (items 12b and 12e) and results (item 17) from these analyses was incomplete. Both of these issues could limit how accurately the UK Biobank analysis methods can be repeated.

Other items that were poorly reported carry different concerns to those of reproducibility. Most authors partially reported item 16a. Supplementary Table 1 shows that while nearly all authors fully report their confounder-adjusted estimates most are failing report their unadjusted elements and to justify why they included the confounders they did.  
This information may be less vital to reproducing analyses than other items, but it does suggest a widespread norm of researchers including confounders without a stated justification. Similarly the significant lack of reporting for item 9, 19 and 22 suggests it is acceptable for researchers not to provide detailed analysis of potential bias in their studies or report the role of funders in their studies. Lastly, many authors failing to mention the design of their study in their title or abstract (item 1a) could create problems in retrieving studies during searches.

The worst reported items were 12d, 13c, 14b and 16c, with most authors reporting no information about them. Again the implications are of varying concern for different items. Figure 2 shows few authors reported addressing loss to follow up in item 12d. This could be because many studies used NHS data linked to the UK Biobank as follow up data so the studies are likely to have low loss to follow up. Although confirming this will require further investigation into the individual studies. The impact of rare reporting of flow diagrams (item 13c) is also tempered by the good results for 13a and 13b, as it suggests authors do report the information just not in a flow diagram. Authors are poor at reporting absolute risk estimates (16c) and missing data (14c).

We coded additional variables to measure reporting quality of information not covered by the STROBE guidelines. 62 articles reported a conflict of interest statement, 41 provided their studies’ UK Biobank application number, and 59 reported the UK Biobank credit statement correctly. However, the reporting was poor for the variables we included to assess the reproducibility of the analysis methods. Only 2 papers shared any part of their analysis code, 2 reported having registered their analyses though whether these registrations were before or after seeing the data was unclear, and 10 reported at least one UK Biobank variable ID.

## 4.2 Limitations

Our study has several limitations. Firstly, many variables in the data extraction form we used to assess reporting quality had poor inter-rater reliability. The kappa statistics, use of “Rules” to solve conflicts, and exclusion of the starred subdivisions suggest as much. This questions the appropriateness of our data extraction form, and the STROBE guidelines it is a based on, as a measure of reporting quality since it does not seem to accurately estimate reporting quality across coders. This poor reliability may be because the STROBE guidelines were not designed as a measure of reporting quality. The STROBE guidelines are reporting guidelines designed for authors to use to check their manuscript contains all the relevant information, while they are writing up. The guidelines are not designed as an assessment of reporting by people other than the authors. Reverse engineering the guidelines for a purpose they were not designed for may be why the inter-rater reliability is poor. The appropriateness of the STROBE guidelines for secondary data analyses, such as those using the UK Biobank, is also questionable. We used the STROBE guidelines to create our data extraction form because it was the most appropriate reporting guidelines available. However, the STROBE guidelines did not necessarily capture the information we needed, leading to us excluding several STROBE items and including additional variables. For example, the guidelines were written in 2007, therefore there is no question on whether or how authors reported analysis code. We also excluded several items or subdivisions of items because they requested reporting of the same information or information not applicable to a seondary data analysis, specifically item 10, two subdivisions of item 6a, one subdivision of item 5, and one subdivision of item 8. See the rule guidelines for further information on these duplicate and inapplicable variables (<https://osf.io/hngcj/>).

## 4.3 Conclusions

Good reporting of statistical methods will aid replications of the main results. Comparing replication results to the original may be difficult given the poor reporting of unadjusted estimates. The lack of shared code will also make replications more difficult.

# 5 Funding Source

This review is being funded by the John Climax Benevolent Fund. The funder will support the conduct of the review by paying KD’s stipend. The funder had no input on any aspect of the project, such the protocol design, data collection, data analysis nor interpretation or publication of results.

# 6 Conflicts of Interest

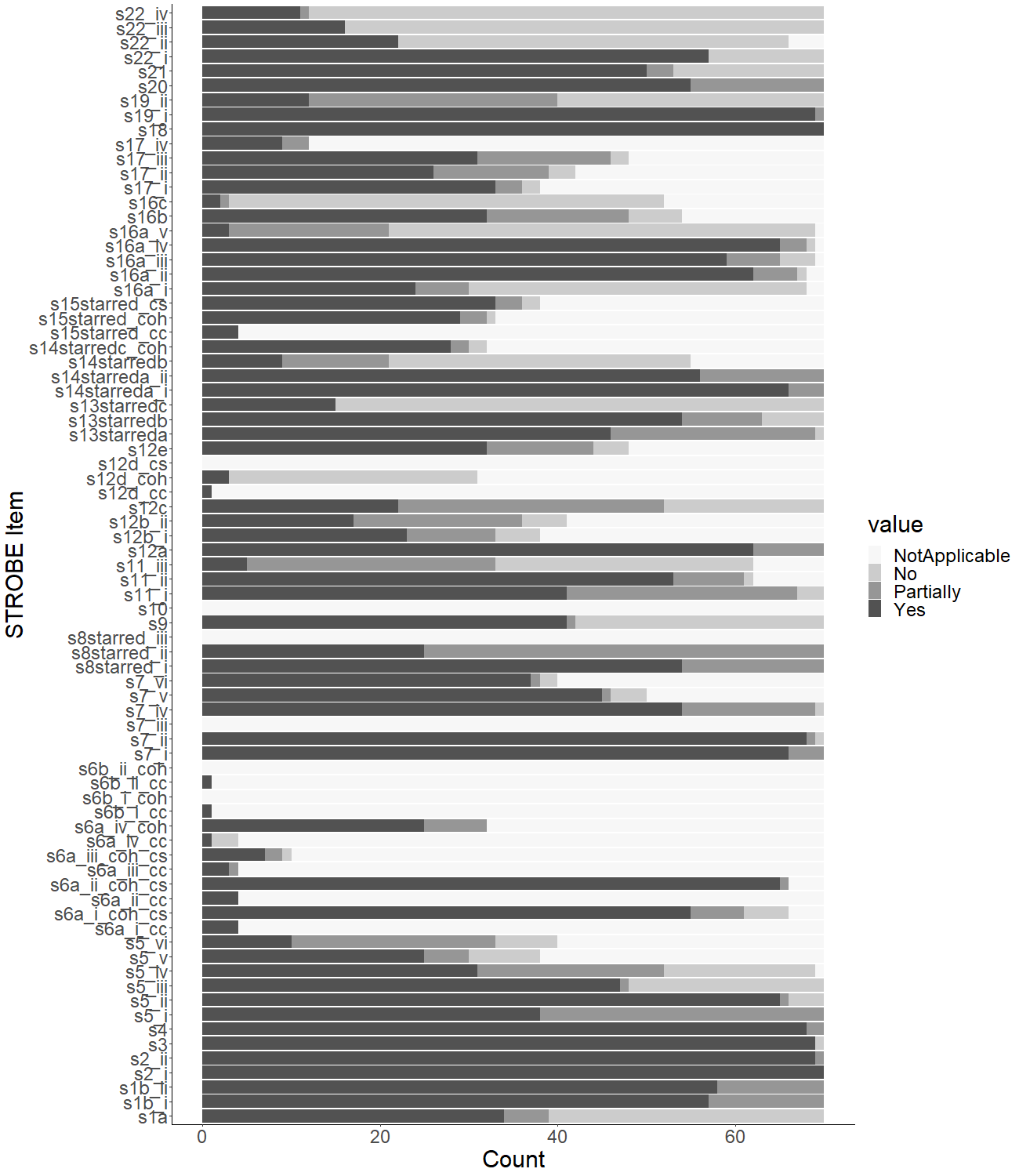
There are no conflicts of interest to report.

# 7 Supplementary Material

Only including observational epidemiological studies is a major departure from the protocol in which we stated that studies would be “classified according to their data type” during data extraction but that all classifications would be included. This was so we could analyse articles according to their appropriate STROBE extension, such as the strengthening the reporting of genetic association studies (STREGA) statement (Little et al., 2009). Given that it would be too time-consuming to extract data from all studies in all included articles and to design multiple data extraction forms, we decided to extract data and assess the reporting quality of studies that could be assessed using the original STROBE guidelines only, i.e. traditional epidemiology studies.

The final data extraction form combined the data items described as part of the “Data Extraction Form” and “Reporting Quality Assessment” in the protocol, but not all were kept.

## 7.1 Supplementary Table 1

*Supplementary Table 1. Descriptive statistics for reporting quality of subdivisions of all STROBE items* 

## 7.2 Supplementary Table 2

Supplementary Table 2. Kappa statistics for reporting quality of subdivisions of all STROBE items

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STROBEItem | Kappa | Zstatistic | Matches | Agreement | NotApplicableUnsure |
| s1a | 0.56 | 5.66 | 52 | 75 | 1 |
| s1b\_i | 0.26 | 2.56 | 57 | 81 | 0 |
| s1b\_ii | 0.24 | 2.24 | 53 | 76 | 0 |
| s2\_i | 0 | NA | 62 | 89 | 0 |
| s2\_ii | -0.04 | -0.54 | 61 | 87 | 0 |
| s3 | 0.32 | 4.99 | 66 | 94 | 0 |
| s4 | -0.03 | -0.75 | 43 | 61 | 0 |
| s5\_i | 0 | 0 | 36 | 51 | 0 |
| s5\_ii | 0.27 | 2.85 | 60 | 86 | 0 |
| s5\_iii | 0.4 | 4.1 | 51 | 73 | 0 |
| s5\_iv | -0.05 | -0.57 | 21 | 30 | 0 |
| s5\_v | 0.52 | 4.34 | 27 | 75 | 34 |
| s5\_vi | NA | NA | NA | NA | NA |
| s6a\_i\_cc | NA | NA | 4 | 100 | 66 |
| s6a\_i\_coh\_cs | 0.45 | 4.83 | 53 | 80 | 4 |
| s6a\_ii\_cc | 0 | NA | 3 | 75 | 66 |
| s6a\_ii\_coh\_cs | -0.02 | -0.41 | 58 | 89 | 5 |
| s6a\_iii\_cc | NA | NA | NA | NA | NA |
| s6a\_iii\_coh\_cs | NA | NA | NA | NA | NA |
| s6a\_iv\_cc | 0 | NA | 3 | 75 | 66 |
| s6a\_iv\_coh | -0.09 | -0.65 | 20 | 62 | 38 |
| s6b\_i\_cc | 0 | NA | 0 | 0 | 69 |
| s6b\_i\_coh | NA | NA | NA | NA | NA |
| s6b\_ii\_cc | 0 | NA | 0 | 0 | 69 |
| s6b\_ii\_coh | NA | NA | NA | NA | NA |
| s7\_i | 0.22 | 2.18 | 62 | 89 | 0 |
| s7\_ii | 0.14 | 1.62 | 61 | 87 | 0 |
| s7\_iii | NA | NA | NA | NA | NA |
| s7\_iv | 0.06 | 0.53 | 40 | 61 | 4 |
| s7\_v | 0.12 | 0.8 | 28 | 80 | 35 |
| s7\_vi | 0.23 | 1.87 | 22 | 81 | 43 |
| s8starred\_i | -0.05 | -0.41 | 52 | 74 | 0 |
| s8starred\_ii | 0 | 0.02 | 26 | 37 | 0 |
| s8starred\_iii | NA | NA | NA | NA | NA |
| s9 | 0.24 | 2.27 | 41 | 59 | 1 |
| s10 | NA | NA | NA | NA | NA |
| s11\_i | 0.16 | 2.2 | 40 | 59 | 2 |
| s11\_ii | -0.07 | -1 | 26 | 50 | 18 |
| s11\_iii | 0.24 | 3.05 | 28 | 53 | 17 |
| s12a | 0.19 | 1.88 | 57 | 81 | 0 |
| s12b\_i | 0.31 | 2.77 | 19 | 66 | 41 |
| s12b\_ii | 0.26 | 2.59 | 21 | 52 | 30 |
| s12c | 0.37 | 4.54 | 38 | 57 | 3 |
| s12d\_cc | NA | NA | 1 | 100 | 69 |
| s12d\_coh | 0.47 | 2.69 | 22 | 85 | 44 |
| s12d\_cs | NA | NA | NA | NA | NA |
| s12e | 0.16 | 1.36 | 25 | 64 | 31 |
| s13starreda | 0.3 | 3.44 | 44 | 63 | 0 |
| s13starredb | 0.29 | 3.35 | 47 | 67 | 0 |
| s13starredc | 0.87 | 7.55 | 67 | 96 | 0 |
| s14starreda\_i | 0.12 | 1.38 | 64 | 91 | 0 |
| s14starreda\_ii | 0.25 | 2.6 | 56 | 80 | 0 |
| s14starredb | 0.39 | 3.8 | 39 | 74 | 17 |
| s14starredc\_coh | 0.37 | 2.72 | 25 | 78 | 38 |
| s15starred\_cc | 0 | NA | 3 | 75 | 66 |
| s15starred\_coh | 0.09 | 1.68 | 13 | 41 | 38 |
| s15starred\_cs | 0.21 | 1.88 | 28 | 76 | 33 |
| s16a\_i | 0.46 | 4.64 | 46 | 70 | 4 |
| s16a\_ii | 0.13 | 1.13 | 56 | 84 | 3 |
| s16a\_iii | NA | NA | NA | NA | NA |
| s16a\_iv | -0.05 | -0.43 | 60 | 90 | 3 |
| s16a\_v | 0.28 | 3.33 | 47 | 69 | 2 |
| s16b | 0.06 | 0.6 | 23 | 55 | 28 |
| s16c | 0.55 | 5.06 | 38 | 93 | 29 |
| s17\_i | 0.02 | 0.18 | 18 | 60 | 40 |
| s17\_ii | 0.03 | 0.36 | 15 | 38 | 31 |
| s17\_iii | 0.48 | 4.17 | 30 | 75 | 30 |
| s17\_iv | 0 | NA | 0 | 0 | 68 |
| s18 | 0 | 0 | 68 | 97 | 0 |
| s19\_i | 0.25 | 2.11 | 65 | 93 | 0 |
| s19\_ii | 0.12 | 1.56 | 28 | 40 | 0 |
| s20 | 0.09 | 1.11 | 51 | 74 | 1 |
| s21 | 0.51 | 5.16 | 52 | 74 | 0 |
| s22\_i | 0.66 | 5.87 | 61 | 87 | 0 |
| s22\_ii | 0.83 | 6.76 | 60 | 92 | 5 |
| s22\_iii | 0.92 | 7.51 | 65 | 97 | 3 |
| s22\_iv | 0.78 | 6.4 | 63 | 94 | 3 |

# 8 References

Agha, R. A., Camm, C. F., Doganay, E., Edison, E., Siddiqui, M. R. S., & Orgill, D. P. (2014). Randomised controlled trials in plastic surgery: A systematic review of reporting quality. *European Journal of Plastic Surgery*, *37*(2), 55–62. <https://doi.org/10.1007/s00238-013-0893-5>

Cevallos, M., & Egger, M. (2014). STROBE (STrengthening the Reporting of Observational studies in Epidemiology). In D. Moher, D. G. Altman, K. F. Schulz, I. Simera, & E. Wager (Eds.), *Guidelines for Reporting Health Research: A User’s Manual* (pp. 169–179). <https://doi.org/10.1002/9781118715598.ch17>

Cook, D. A., Levinson, A. J., & Garside, S. (2011). Method and reporting quality in health professions education research: A systematic review. *Medical Education*, *45*(3), 227–238. <https://doi.org/10.1111/j.1365-2923.2010.03890.x>

Drax, K., Richmond, R., Woolf, B., Smith, G., & Munafo, M. (2019). *UK Biobank Study - Systematic Review*. <https://doi.org/10.17605/OSF.IO/SF5VJ>

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, *5*(8), 180448. <https://doi.org/10.1098/rsos.180448>

Little, J., Higgins, J. P. T., Ioannidis, J. P. A., Moher, D., Gagnon, F., Von Elm, E., … Birkett, N. (2009). STrengthening the REporting of genetic association studies (STREGA)- An extension of the STROBE statement. *Genetic Epidemiology*, *33*(7), 581–598. <https://doi.org/10.1002/gepi.20410>

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in *The BMJ* and *PLOS Medicine*. *BMJ*, k400. <https://doi.org/10.1136/bmj.k400>

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. <https://doi.org/10.1186/s13643-016-0384-4>

Plint, A., Moher, D., Morrison, A., Schulz, K., Altman, D., Hill, C., & Gaboury, I. (2006). Does the CONSORT checklist improve the quality of reports of ... *Medical Journal of Australia*, *185*(5), 263. <https://doi.org/10.5694/J.1326-5377.2006.TB00557.X>

UK Biobank Coordinating Centre. (2007). *UK Biobank: Protocol for a large-scale prospective epidemiological resource UK Biobank Coordinating Centre Stockport* (No. March; pp. 1–112).

UK Biobank Coordinating Centre. (2011). *Access Procedures: Application and review procedures for access to the UK Biobank Resource* (pp. 1–36).

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ*, *5*, e3208. <https://doi.org/10.7717/peerj.3208>

World Wide Web Foundation. (2018). *Open Data Barometer - Leaders Edition*. Washington DC: World Wide Web Foundation.