Data Extraction Manual

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Row 1 – Heading 1 | Row 2 – Heading 2 | Row 3 – Heading 3 | What to input into the cell | Example |
|  | No. |  | The study number you are working on | 1 |
|  |  | bioRxiv - if yes find the published paper and use that (put 'yes' in next column) if no published paper put 'no' in next column | Is the paper from bioRxiv?  If it is then put ‘yes’.  If it isn’t then put ‘no’ | Yes |
|  |  | Published | If the paper is from bioRxiv look for the published version. If you find the published version use that version and write ‘yes’.  If you cannot find the published version use the bioRxiv version and write ‘no’. | Yes |
| Article and author info | Title |  | What is the title of the study. | A comprehensive evaluation of the genetic architecture of sudden cardiac arrest |
| Article and author info | Publication year |  | What is the publication year of the study. | 2018 |
| Article and author info | Full Journal name |  | What is the full journal name; not the abbreviated version | European Heart Journal |
| Article and author info | Doi |  | What is the doi of the study | 10.1093/eurheartj/ehy474 |
| Article and author info | Corresponding author full name, institution, and country of institution (if different from first author) |  | What is the corresponding authors details? Include all affiliations for that author given in the study. | Dan E. Arking; arking@jhmi.edu; McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins, 733 N Broadway, Baltimore, MD 21205, USA |
| Introduction | Hypothesis |  | What is the stated hypothesis of the study? Copy and paste from the study if available. | We aim to identify potential loci associated with sudden cardiac arrest (SCA) and to identify risk factors causally associated with SCA |
| Introduction | Rationale |  | What is the rationale of the study? That is, why are they doing this study. | Sudden cardiac arrest (SCA) is a major cause of cardiac mortality, affecting over 300 000 people in the US every year. Family history of SCA is a strong risk factor for SCA in the general population, suggesting that genetic variation may influence SCA risk. |
| Methods | MR study design (One sample, Two sample etc.) |  | What is the MR study design of the study.  **Drop down list from data in ‘Auto\_fill’ sheet.** | Two-sample summary-level |
| Methods | Outcome data | Outcome | What is the outcome phenotype in the MR analysis. If they have multiple outcomes then each outcome must go on a separate row.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | Sudden cardiac arrest |
| Methods | Outcome data | Outcome definition | What is the definition of the outcome in the study you are reading? If they do not say then say they do not say. If they have taken the outcome data from an external GWAS use the definition from this. If they have meta-analysed many GWASs then the outcome will be defined in each GWAS, if they haven’t given a definition in the study you are reading then do not read all of the definitons of the individual GWASs that comprise the meta-analysis and instead put ‘Defined in individual GWASs’.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | Defined in individual GWASs |
| Methods | Outcome data | Study outcome from | Does the outcome data come from the study you are reading (if yes put ‘Current study’), if it doesn’t then you need to add the study they got the data from to the ‘Outcome\_GWASs’ sheet – this will require you going to the original GWAS paper and reading and extracting data from that.  **Drop down list from ‘Outcome\_GWASs’ which auto-fills a range of other cells based on the study selected.** | Current study |
| Methods | Outcome data | Age range | What is the age range of the people in the outcome data. If this is not reported in the study you are reading then you need to go to the original GWAS paper and extract the data. If, in the original GWAS paper it is many studies take the youngest and oldest ages and use those as your age-range. If it is not reported in the original GWAS paper what the age range is the state ‘Not reported’. If you are given a mean age then use this.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | 19-101 |
| Methods | Outcome data | Sex | What is the sex of the outcome data population? This will be autofiled if you have filled in the Ooutcome\_GWASs sheet with the data and selected the study outcome. If this is not the case then report the sex. If not reported in the study you are reading go to the original GWAS paper and report this.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | both |
| Methods | Outcome data | In your study: size/case | What is the number of people in total (or cases) in the outcome data that is reported in the study you are reading.  This is the number the study reports.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | 3939 |
| Methods | Outcome data | In your study: control | What is the number of controls in the outcome data that is reported in the study you are reading.  This is the number the study reports.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | 25989 |
| Methods | Outcome data | In the consortium/OG GWAS: size/case | If the study has used data from a consortium what is the reported number of people in the consortium (or cases) – this should be filed in in addition to the previous population numbers.  This is the number reported in the original GWAS that your study got its data from if applicable, if not applicable NA.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | NA |
| Methods | Outcome data | In the consortium/OG GWAS: control | If the study has used data from a consortium what is the reported number of controls in the consortium – this should be filed in in addition to the previous population numbers.  This is the number reported in the original GWAS that your study got its data from if applicable, if not applicable NA.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | NA |
| Methods | Outcome data | Reason for different Ns in your study and the consortium/OG GWAS | If the study has used a consortium or other GWAS for their outcome data there may be a discrepancy in the number of people the study and the consortium/GWAS quote – does the study explain why there is a difference?  If no difference NA. | NA |
| Methods | Outcome data | health status | If the outcome is case/control state ‘case/control’ otherwise report whether the health status of the population is ‘healthy’ or whether they have a specific pre-selected condition – e.g. did they do an MR of x in people with prostate cancer of smoking, in this case health status of the outcome data would be ‘prostate cancer’ | Case/control |
| Methods | Outcome data | Population | What is the underlying population the outcome data comes from.  **Auto-filled based on ‘Study outcome from’ cell, otherwise filled in manually.** | European ancestry |
|  |  |  |  |  |
| Methods | Selection of genetic variants | Exposure | What is the exposure in your study? If there are multiple exposures only use those which are measures of adiposity. Each different exposure:outcome pair should go on a separate row.  **Auto-filled based on ‘Study exposure from’ cell, otherwise filled in manually.** | BMI |
| Methods | Selection of genetic variants | Study exposure from | Does the exposure come from the current study? If yes then select ‘current study’ – this will auto-fill a number of cells with ‘Manually fill’, you must fill these cells in yourself.  If the study comes from an external consortium/GWAS then select this in the dropdown list. If not in the drpdown list you will need to add the study to the ‘Exposure\_GWASs’ sheet and then select it.  **Drop down list from ‘Exposure\_GWASs’ which auto-fills a range of other cells based on the study selected.** | Locke 2015 BMI All ancestries |
| Methods | Selection of genetic variants | Age | What is the age of the population used to identify exposure instruments?  If you fill in ‘Study exposure from’ with current study you need to enter this manually. Otehrwise it will be auto filled by Matt at the end.  **Auto-filled based on ‘Study exposure from’ cell, otherwise filled in manually.** |  |
| Methods | Selection of genetic variants | Sex | What is the sex of the population used to identify exposure instruments?  If you fill in ‘Study exposure from’ with current study you need to enter this manually. Otehrwise it will be auto filled by Matt at the end.  **Auto-filled based on ‘Study exposure from’ cell, otherwise filled in manually.** |  |
| Methods | Selection of genetic variants | N reported in study | What is the N reported in the study you are reading? |  |
| Methods | Selection of genetic variants | N reported in original consortium/OG GWAS | What is the size of the consortium/GWAS that the exposure data was collected from?  **Auto-filled based on ‘Study exposure from’ cell, otherwise filled in manually.** |  |
| Methods | Selection of genetic variants | Reason for difference in N | If there is a difference between the Ns that the study reports and the original consortium/GWAS reports what is the reason for this? If they do not state a reason and you can not work it out (e.g. it is not obvious they have used the All ancestries GIANT data even though they say they use the European ancestry data) state ‘Not discussed’.  If no difference NA. | NA |
| Methods | Selection of genetic variants | Health status | As with the same column for ‘Outcome’.  If the exposure is case/control state ‘case/control’ otherwise report whether the health status of the population is ‘healthy’ or whether they have a specific pre-selected condition – e.g. did they do an MR of BMI measured in people with prostate cancer on smoking in people with prostate cancer, in this case health status of the exposure data would be ‘prostate cancer’ | Healthy |
| Methods | Selection of genetic variants | Population for exposure SNPs | What is the underlying population that the exposure SNPs come from?  If data is from the current study then use the population they quote. If the data is from an external GWAS/consortium state that population used for the GWAS. If the Ns reported in the study and the original consortium do not match this is likely because they have reported the wrong N – in this situation it is hard to tell what population they have used so state ‘Unclear – GIVE REASON FOR WHY UNCLEAR’ | Unclear – presumed All ancestries based on N reported in current study |
| Methods | Selection of genetic variants | Description of covariates included in model to adjust genetic variant-exposure estimates and standard errors | Have they adjusted their model?  If not discussed – ‘Not discussed’ |  |
| Methods | Selection of genetic variants | Data set used for genotyping (1K, HapMap, HRC, Bespoke, Sequence data, Unimputed array data, other) | What data set did the use for genotyping?  If not discussed – ‘Not discussed’ |  |
| Methods | Selection of genetic variants | Determination of independence - LD (Genetic variants with r² less than a specified threshold selected;  - Genetic variants with r² less than a specified threshold selected and LD matrix applied to remaining variants) | What was their LD threshold for determining independent SNPs?  If not discussed – ‘Not discussed’ | ≥ 0.01 |
| Methods | Selection of genetic variants | Evaluation of potential pleiotropy | What did they do to evaluate potential pleiotropy? State the method(s) used.  If not discussed – ‘Not discussed’ | HEIDI-outlier method |
| Methods | Selection of genetic variants | P-val threshold for selecting genetic variants for MR | What P value threshold have they used for selection of genetic variants? This should really be 5x10-8, however some studies may use additional SNPs in a separate analysis that do not reach this threshold. State the threshold | 5x10-8; additional SNPs below the GWS threshold and is based on the rationale that using a less stringent P-value cut-off will allow for incorporation of false-positives that will downward bias the GRSA effect estimates. However, we know that this approach reduces power, since power for Mendelian Randomization is a function of the variance explained of the SNPs in the GRSA, and for the complex traits in our study, the majority of the genetic signal is contained in SNPs that do not meet GWS. We recognize that while incorporating many additional true positives (increasing power), that many false-positives are included as well (decreasing the GRSA estimate). Because of this, we wanted to identify a P-value cutoff in which we limited the number of SNPs included, therefore limiting the number of false positive SNPs included, but which the set of SNPs closely reach maximal power. Therefore, for each trait, we plotted the variance of the trait explained vs. P-value threshold to determine the P-value cut-off for which the variance explained was maximized and included the fewest number of SNPs (Figure S4 and Table S9). Due to the downward bias of the GRSA estimate, we use this secondary analysis to only identify potential associations by evaluating the significance of the GRSA estimate at maximal power (Pmax), and do not draw any conclusions from the magnitude of the GRSA estimate (GRSAmax). |
| Methods | Selection of genetic variants | identification of palindromic SNPs | Did they identify palindromic SNPs? If yes, did they use palindromic SNPs and how did they use them? It is unlikely they will discuss this.  If not discussed – ‘Not discussed’ | Not discussed |
| Methods | Selection of genetic variants | Use of proxy SNPs | Did they use proxy SNPs?  If not discussed – ‘Not discussed’ | Not discussed |
| Methods | Selection of genetic variants | Testing for Hardy-Weinberg equilibrium to ascertain random assignment of genetic variants | Did they test for Hardy-Weinberg equilibrium?  If not discussed – ‘Not discussed’ | Not discussed |
|  | Instrumental variables - how many | Single | Did they use a single SNP as exposure?  If yes, state ‘1’  If no – ‘NA’ | NA |
|  | Instrumental variables - how many | Multiple | Did they use multiple SNPs as exposure?  If yes, state how many  If no – ‘NA’ | NA |
|  | Instrumental variables - how many | GRS | Did they use a genetic risk score composed of multiple SNPs as exposure?  If yes, state how many  If no – ‘NA’ | 72 |
|  | SNPs in GWAS |  | How many SNPs in the original GWAS/consortium reported as independent and reaching genome wide significance (5x10-8)  **Auto-filled based on ‘Study exposure from’ cell, otherwise filled in manually.** | 97 |
|  | Reason for difference in SNPs |  | If there is a difference between the number of SNPs used and the number reported in the original GWAS that the study says they use the instruments from state the reason gives for this discrepancy? E.g. did they exclude x number of SNPs for not being in Hardy Weinberg equilibrium, or where x SNPs not present (and proxys not used) in the outcome.  If no difference - NA  If not discussed – ‘Not discussed’ | Not discussed |
|  | A priori power calculation (including how calculated, how r² determined, etc.) | Yes/No | Did they perform an a priori power calculation?  Yes or no – only state ‘no’ if they explicitly state they did not do one.  If not discussed – ‘Not discussed’. | Not discussed |
|  | A priori power calculation (including how calculated, how r² determined, etc.) | How | How did they perform their a priori power calculation? Include all information here.  If no a priori power calculation performed – ‘NA’ | NA |
|  | Methods used for estimating causal effect |  | What method did they use to estimate the causal effect?  **Drop down list from ‘Autofill\_data’. If the method used is not in the drop down list add it to the ‘Autofill\_data’ sheet.** | Inverse variance weighted |
|  | Methods used for testing for pleiotropy (MR-Egger, Cochrans Q, Ruckers Q etc) |  | What method did they use to test for pleiotropy?  **Drop down list from ‘Autofill\_data’. If the method used is not in the drop down list add it to the ‘Autofill\_data’ sheet.** |  |
|  | Methods used for testing for pleiotropy (MR-Egger, Cochrans Q, Ruckers Q etc) | Notes | If they do not report the results state ‘Results not reported’.  If they report results from the pleiotropy tests report the summary of these results (i.e. what did they conclude from these tests) | they did not report these |
|  | Assessment of instrument strength (mean F statistic, I2 etc) |  | Did they report the instrument strength for the method used to determine causal effect?  State the measure and result. Otherwise, if not discussed – ‘Not discussed’ |  |
|  | Software used to perform MR (R, Stata etc.) |  | What statistical software did they use to perform the MR analysis (NOT THE DATA MANIPULATION)? | R |
|  | Program used to perform MR (MR-Base etc) |  | What programme/package did they use to perform the MR analysis? | The associations of these SNPs with the risk factors and the SCA outcome are used to calculate an inverse-variance weighted multi-SNP GRSA as implemented in the R-package ‘gtx’. |
|  | Quality control | Replication by independent researcher | Did they have an independent researcher replicate the study – yes/Not discussed  Only state no if explicitly stated they did not do this. | Not discussed |
|  | Quality control | Was the analysis replicated in the same paper e.g. did they subset the data for discovery and replication analysis | Did they perform discovery/replication within their data set? Explain what they did in terms of replicating their results.  If not discussed – ‘Not discussed’ | No - they did an overall MR and then subsetted men and women and repeated |
|  | Multiple testing |  | What did they do for multiple testing? If they did not correct for multiple testing and state a significance of 0.05 then state ‘No’  If they performed multiple testing state exactly what they did. | false discovery rate (FDR) cut-off of FDR < 0.05. |
|  | Exposure units (SD, mmol etc.) |  | What units is the exposure GWAS reported in |  |
|  | Outcome units (SD, mmol etc.) |  | What units is the outcome GWAS reported in |  |
|  |  |  |  |  |
| Results | Instrumental variable analysis | Expressed as: OR/beta/etc - GET THE BETA IF POSSIBLE | What are the units you are reporting expressed as? |  |
| Results | Instrumental variable analysis | Estimate | What is the effect estimate? |  |
| Results | Instrumental variable analysis | SE | What is the standard error of the effect estimate |  |
| Results | Instrumental variable analysis | CI % | What is the confidence interval % |  |
| Results | Instrumental variable analysis | CI lower | What is the lower bound of the confidence interval |  |
| Results | Instrumental variable analysis | CI upper | What is the upper bound of the confidence interval |  |
| Results | Instrumental variable analysis | P | What is the raw p value of the causal effect |  |
| Results | Instrumental variable analysis | Adjustment method | What is the multiple testing/p-value adjustment method |  |
| Results | Instrumental variable analysis | P-adjusted | What is the adjusted p-value |  |
|  |  |  |  |  |
| Discussion | Limitations - do the authors discuss limitations | Limitations overall |  |  |
| Discussion | Limitations - do the authors discuss limitations | Horizontal pleiotropy |  |  |
| Discussion | Limitations - do the authors discuss limitations | Statistical power (post hoc power calc?, selection bias, etc.) |  |  |
|  |  |  |  |  |
| Other | Notes |  | Any notes or extra information you think pertinent to the interpretation or data extraction of the study  Leave blank if no |  |
|  |  |  |  |  |
|  |  |  |  |  |