Understanding Society

Working Paper Series

2025 – 06

May 2025

Two new solutions to the zero weights problem

Jamie C. Moore* and Paul Clarke

Institute for Social and Economic Research, University of Essex, UK

Understanding Society

Economic and Social Research Council

Non-technical summary

To correct for non-response bias, household (HH) panel surveys release inverse propensity non-response (IP-NR) weights that adjust selection weights for non-response. However, often some respondents lack selection weights, so cannot be assigned IP-NR weights (the 'zero weights' problem). Previous solutions to this issue, which reduces weighted dataset quality, have limitations. Sharing with unweighted respondents the existing selection weights of HH members then estimating IP-NR weights requires an existing selection weight in the HH. Predicting selection probabilities for such respondents then estimating IP-NR weights is model based, uses only responding HH probabilities, and requires assumptions when adjusting for multiple HH selection paths. Hence, two new procedures are introduced. Both form clusters of unweighted and existing weight individuals with similar characteristics, and split the existing weights among cluster members. Split Selection (SS) weighting splits selection weights, then re-estimates IP-NR weights. Split IP-NR (SIP-NR) weighting splits estimated IP-NR weights. Procedure performance is then evaluated, using the UK Household Longitudinal Study COVID-19 Study datasets. Both performed well in increasing dataset size and subgroup analysis feasibility, and in reducing non-response biases and precision loss: indeed, precision loss was lower with SIP-NR weights than IP-NR weights. The use of these procedures is then discussed.

Two new solutions to the zero weights problem

Jamie C. Moore* & Paul S. Clarke

Institute for Social and Economic Research, University of Essex, UK

* Corresponding Author: Institute of Social and Economic Research, University of Essex, Essex, UK CO4 3SQ. Email: moorej@essex.ac.uk

Abstract:

Household panel surveys release inverse propensity non-response (IP-NR) weights that adjust selection weights for non-response. However, often some respondents lack selection weights, so cannot be assigned IP-NR weights. Previous solutions to this issue have limitations. We introduce two new procedures for weighting unweighted individuals. Both split the existing weights of individuals with unweighted individuals with similar characteristics. We evaluate procedure performance using data from the UKHLS COVID-19 Study, and find both perform well in increasing dataset sizes and reducing non-response bias and precision loss. We then discuss their broader use in survey design.

Keywords: Survey methodology, non-response, non-response weighting, dataset quality.

Acknowledgements:

The Understanding Society survey is funded by ESRC grants awarded to Nick Buck (RES-586-47-0002, ES/K005146/1) and Michaela Benzeval (ES/N00812X/1, ES/S007253/1, ES/T002611/1, ES/Y003071/1, ES/Y010469/1), with co-funding from various Government Departments. The data are collected by the National Centre for Social Research (NatCen) and Verian (formerly Kantar Public). The Understanding Society survey is funded by ESRC grants awarded to Nick Buck (RES-586-47-0002, ES/K005146/1) and Michaela Benzeval (ES/N00812X/1, ES/S007253/1, ES/T002611/1, ES/Y003071/1, ES/Y010469/1), with co-funding from various Government Departments. The data are collected by the National Centre for Social Research (NatCen) and Verian (formerly Kantar Public). The COVID-19 Study was funded by the Economic and Social Research Council (ES/K005146/1) and the Health Foundation (2076161). Fieldwork for the survey was carried out by Ipsos MORI and Kantar Public (now Verian). The data are distributed by the UK Data Service (main survey: 10.5255/UKDA-SN-6614-20; COVID-19 Study: 10.5255/UKDA-SN-8644-11).

1. Introduction

Reducing non-response bias is a task faced by all survey designers (Groves et al. 2001). Bias-reduction measures may be undertaken during data collection, such as ensuring the survey sample reflects the study population and obtaining responses from all sub-groups ('bias prevention measures': Groves & Heeringa 2006; Wagner 2008). They may also be undertaken post data collection ('bias-adjustment measures'), of which the most common is the supply of non-response weights that map respondents to study populations (Valliant & Dever 2013; see Little & Rubin 2014 for alternative methods).

If non-respondent information exists for probability surveys, Inverse Propensity Non-Response (IP-NR) weights are normally supplied. These adjust selection weights (the inverse of the selection probability) for the inverse of the probability of responding following selection, that is, IP-NR weight = selection weight × NR weight. They perform better in terms of bias reduction and precision loss (unequal weights over-inflate estimate variances: Little & Vartivarian 2005) than, for example, weights that calibrate respondents to population totals (e.g. Moore et al. 2024). The use of IP-NR weights, however, is not without its issues. One is that, in household (HH) panel surveys, it may not be possible to weight all respondents (Schonlau et al. 2013), which is termed in this paper the 'zero weights problem'. Data from these surveys are often re-purposed by supplying weights enabling cross-sectional analyses. However, sample members are selected within HHs, with selection probabilities derived from HH values, so such weights cannot be assigned to new HH entrants or when HHs split: the selection probabilities would need adjusting given

previous HH membership due to multiple selection paths. In addition, longitudinal adjustments for non-enumeration in items detailing HH characteristics may be made to individual selection weights (e.g. University of Essex & Institute for Social and Economic Research 2019), meaning that those not enumerated at all previous waves cannot be weighted (see also section 2.4).

Numbers of non-IP-NR weighted respondents can be substantial, for example, 28% in the 2009 wave of the Panel Study of Income Dynamics (Heeringa et al. 2011). This decreases weighted dataset quality as smaller datasets have larger sampling errors, make accurate sub-group analyses less feasible and, if they reflect study populations less well, can lead to increased non-response bias and precision loss after weighting (Schouten et al. 2016; Moore et al. 2024). Moreover, interviews avoidably excluded from datasets raise concerns given the costs for survey organizations and participants of collecting data.

Solutions to the zero-weight problem which permit survey estimation are used by established surveys (Schonlau et al. 2013). The first is HH weight sharing (Ernst 1989; Lavallée 1995; Heeringa et al 2011; Taylor et al. 2018; University of Essex & Institute for Social and Economic Research 2019). With this method, original sample members with (existing) selection weights share them with unweighted respondents in the same HH. Movers with existing weights retain them to share with others in their new HHs. Then, IP-NR weights are estimated for the larger dataset using the shared selection weights as inputs. The second is selection probability prediction (Haisken-Denew & Frick 2005; Watson 2012). With this method, wave 1 HH selection probabilities are predicted for unweighted respondents after modelling existing

values using survey information. These probabilities are then used to adjust current wave equivalents for multiple selection paths, individual probabilities/weights derived from them, and IP-NR weights are estimated for the larger dataset using the new selection weights as inputs. However, beyond requiring HH membership information to collected, both methods have limitations. HH weight sharing produces unbiased estimates (see also Lavallée 2007; Zhang 2021) but requires an existing weight in the HH to share, so respondents not in such HHs remain unweighted (for example, 16% in the 2008 wave of the British Household Panel Survey: University of Essex & Institute for Social and Economic Research 2023). Selection probability prediction weights all respondents, but is model based with predictions made using only responding HH probabilities, and adjustments for multiple selection paths involve assumptions.

These limitations are a major drawback for the survey we focus on in this paper, namely, the COVID-19 Study carried out as part of the UK Household Longitudinal Survey (UKHLS). The UKHLS is an annual, multi-mode, multi-domain HH panel survey of people living in the UK. It is based on probability sampling and users are supplied with IP-NR weights which can be used to make high-quality population inferences (Benzeval et al. 2020). Consequently, the UKHLS is widely used by decision-makers and researchers. During the COVID-19 pandemic, the team also fielded the UKHLS COVID-19 Study, a series of surveys that captured information from main survey participants more frequently. For full details, see section 2.2 and the Study User Guide (Institute for Social and Economic Research 2021) and website: https://www.understandingsociety.ac.uk/topic/covid-19.

The UKHLS team also supply IP-NR weights mapping COVID-19 Study respondent sample to the UK population (see section 2.3). The Study sample was those enumerated in HHs participating at either waves eight or nine at that time or in the time up to Study inception (i.e. at wave 10 / or in the early parts of wave 11 data collection) for whom email contact details existed). The IP-NR weights adjust the main survey non-response weights (in effect, the COVID-19 Study selection weights) for these individuals for (COVID-19) Study non-response (Moore et al. 2024). However, depending on the survey wave, between 15-25% of Study respondents are not IP-NR weighted because they lack main survey weight inputs (see sections 2.3 and 5.1). Both cross-sectional and longitudinal Study weights are affected, as with the latter only response to the Study was considered. Moreover, existing solutions to the zero-weight problem are not useful. Neither HH weight sharing nor selection probability prediction is possible because HH structure was not enumerated in the COVID-19 Study (it is not formally a HH panel survey like the main survey), and HH information from the main survey cannot be used instead because some was two or more years out of date when the pandemic started. It should also be noted that no alternative natural links exist between weighted and unweighted respondents, so other methods similar to HH weight sharing that utilise such links, collectively known as indirect sampling methods (see Lavallée 2007; Zhang 2021 for details), cannot be employed either.

In this paper, two new procedures that address the zero weights problem are outlined. Both are based on forming 'matching-clusters' of unweighted and existing weight individuals with similar characteristics, and splitting the existing weights

among cluster members. The first, Split Selection (SS) weighting, splits original sample member (existing) selection weights, then estimates the survey wave IP-NR weights. The second, Split IP-NR (SIP-NR) weighting, splits survey wave IP-NR weights estimated for those with existing selection weights. Then, procedure performance is evaluated, using the COVID-19 Study datasets. The sizes of datasets weighted by each procedure are reported, along with the characteristics of included respondents, to enable sub-group analysis feasibility to be studied. In addition, non-response bias reduction and precision loss due to weighting is quantified.

The paper proceeds as follows. In section 2, the two UKHLS surveys and existing weighting procedures are described. In section 3, the two new procedures are detailed. In section 4, methods used to evaluate procedure performance are outlined. In section 5, evaluation results are reported. In section 6 the implications of our research for the COVID-19 Study and survey design in general are discussed.

2. Considered surveys

2.1. *Understanding Society:* The UK Household Longitudinal Study (UKHLS)

The UKHLS main survey annually follows - up a sample of people living in the UK (Institute for Social and Economic Research 2022). Interviews are sought from all adults in participating HHs. The survey began in 2009, and includes respondents from the preceding British Household Panel Survey, which began in 1991. It has a sequential mixed-mode design: some sample members are allocated to web and others to face-to-face interview, with follow up in other modes. Its sample is constructed from

probability samples, with non-response carefully modelled (Lynn & Kaminska 2010; Lynn et al. 2012). Research shows that the survey continues to support valid population inference (Benzeval et al. 2020).

2.2. UKHLS COVID-19 Study

The UKHLS main survey is not set up to provide information at pace, so when the COVID-19 pandemic began a more frequent web survey was fielded to record how it and associated policy responses were affecting respondents. The COVID-19 Study sample was all adults (16+) in HHs responding at main survey Waves 8 or 9, who had not dropped out, died or emigrated as of April 2020. Wave 1 of the Study was fielded in April 2020, with eight further surveys undertaken (some non-respondents were also followed up by telephone, but the focus here is on web respondents: see University of Essex & Institute of Social and Economic Research 2021 for full details of the COVID-19 Study).

2.3. COVID-19 Study IP-NR weight construction

In the December 2021, IP-NR weights were released which map COVID-19 Study respondents to the UK population at the time of main survey wave 10 (2019-20), updated for mortality and emigration but not immigration. Cross-sectional forms are the product of respondents main survey wave 10 cross-sectional non-response weights and the inverses of their estimated probabilities of responding to the Study wave; the longitudinal forms are the product of a chain of weights derived given the probability of wave 1 response conditional on possessing the main survey weight, of

wave 2 response conditional on wave 1 response, and so on. The main survey wave 10 weight was used as the selection weight (before, the wave 9 weight was used) as data collection was almost complete in March 2020 and more of the Study sample possessed it. IP-NR weighting depends on possessing this weight, so its derivation (and who does not possess it) is detailed in section 2.4.

Regression was used to estimate response propensities, with main-survey wave 10 predictors for cross-sectional weights and first weights in the longitudinal chains. For later weights in longitudinal chains, some predictors were replaced with the same variables from the COVID-19 Study because only previous wave information was needed. Predictors included demographics, HH structure, economic, health and survey design variables. If many predictors exist, model overfitting can occur (Harrell 2001; Burnham & Anderson 2002), causing precision loss, so they were selected using logistic regression with a Least Absolute Shrinkage and Selection Operator (Lasso) procedure (in the Stata package 'lassopack': Ahrens et al. 2020), which excludes variables by shrinking unstable coefficient estimates towards zero without the need for statistical tests (Tibshirani 1996; Steyerberg et al. 2001). See the Appendix for more details of these procedures. Finally, weights were trimmed to restrict precision loss (see Valliant & Dever 2018): values more than 25 times the median were replaced with the threshold value, which limited precision loss to acceptable levels while still reducing biases (Moore et al. 2024).

2.4. UKHLS main survey cross-sectional non-response weight construction

The main survey weight acting as the selection weight in COVID-19 Study IP-NR weight construction is the wave 10 cross-sectional non-response weight. This weight is the wave 10 cross-sectional enumerated sample member weight, adjusted for non-response after modelling wave 10 response using HH questionnaire predictors. The wave 10 cross-sectional enumerated sample member weight is the wave 10 longitudinal enumerated sample member weight shared with (some) HH members (see next paragraph). The wave 10 longitudinal enumerated sample member weight its previous wave value, adjusted after modelling (non-) enumeration in the wave 10 HH questionnaire, using predictors from the similar wave 9 questionnaire. The waves 7-9 equivalents are computed similarly, as is the wave 6 weight, except the input is the wave 6 inclusion weight, which combines an inclusion weight for refreshment sample members entering the survey at the wave with the wave 6 longitudinal enumeration weight computed as outlined above for others enumerated at the wave, The waves 3-5 longitudinal enumeration weights are also computed as above, as is the wave 2 weight, except the input is the wave 2 inclusion weight, the inverse of sample member selection probability. For the derivation of these latter probabilities, see University of Essex & Institute for Social and Economic Research (2019).

HH weight shared cross-sectional enumeration weights are used as selection weights when estimating the cross-sectional non-response weights because weighted dataset sizes are larger than when their longitudinal counterparts are used. New entrants to original sample member (OSM) HHs (temporary sample members: TSMs) lack longitudinal enumeration weights, preventing non-response weight estimation. So do children born to OSM mothers or non-OSM fathers (permanent sample

members: PSMs),, OSMs who move from sample HHs, and OSMs in HHs not enumerated at one or more previous waves. HH weight sharing provides enumeration weights and enables non-response weight estimation for some of these individuals by, in HHs in which only some members have longitudinal weights, instead assigning all HH members as their weight an equal share of the existing longitudinal weight HH sum. In addition, newborns are assigned their mother's longitudinal weight, and, if they possess them, OSMs who move to new HHs take their longitudinal weight with them (to share with other HH members). It should be noted though, that these methods do not enable non-response weights to be estimated for all respondents. Those in HHs not enumerated at all waves remain without enumeration weights and cannot be weighted.

3. Procedures to weight to non-IP-NR weighted COVID-19 Study respondents

Our two new procedures address the problem of weighting the zero-weight individuals mentioned at the end of the last section who remain without non-response weights after IPNR weighting and HH weight sharing has been undertaken. Without loss of generality, particular attention is paid to weighting these individuals in the UKHLS COVID-19 Study. In the Study (see also Introduction), they are joined in the sample by other similarly unweighted individuals such as those enumerated in main survey wave 8 / 9 HHs at wave 10 or during the early part of wave 11 data collection, and some TSMs and PSMs, who are not provided with main survey weights due to UKHLS sample considerations. Together, this group provides 15-25% of Study respondents. Neither (further) HH weight sharing nor selection probability prediction

can be used to provide weights for individuals in this group because HH membership was not enumerated in the Study and similar information from the main survey can be two or more years out of date. In addition, no alternative natural links exist between them and weighted individuals. so other indirect sampling methods cannot be used.

Both procedures involve forming matched-clusters, that is, clusters formed by matching weighted and unweighted respondents with similar characteristics, and then splitting the matched weights between the cluster members (the term 'splitting' is used to distinguish these procedures from weight sharing methods). As explained earlier, the 'selection' weights used in COVID-19 Study IP-NR weight estimation are the non-response weights for respondents to wave 10 of the main survey. The sample of individuals with these weights can be viewed as arising from an indirect sampling scheme (Lavallee 2007) or, more generally, a bipartite incidence graph sampling (BIGS) scheme (Zhang and Patone 2017; Zhang 2022). However, the final Study sample is not BIGS because it also includes individuals for whom information on links with existing HHs is unavailable (who hence would not be weightable).

Informally, let j index the sample individuals with selection weights, and l index hose left-out individuals without weights. The task our procedures perform is to calculate the COVID-19 Study weights (for a given Study wave taken to be wave 1 without loss of generality). These are w_j/ρ_j and w_l/ρ_l, where w_j is the available selection weight, w_l is an the unknown selection weight to estimated by splitting the w_j, and ρ_j and ρ_l are response propensities to be estimated using suitable regression

model common to both groups. The two procedures are alternative ways of estimating the Study non-response weights.

An informal framework within which we set out the assumptions under which our procedures lead to valid inference is set out in Appendix A.2. In summary, we show that our procedures can be used to obtain moderately conservative and design-consistent inference provided the analyst is able to choose a selection of variables Z where a) exchangeability holds: that is, for unit l with $Z_l = z$, we can find at least one j with $Z_j \approx z$ such that selection weight w_j can be split with l; b) response to each wave of the Study is missing at random given Z, that is, the response propensity ρ - the probability of responding to questionnaire - is independent of the survey variables; and c) the split weight is regular consistent estimator of the true unknown weight, The different sets of weights we refer to in the subsequent discussion are described in Table 1.

3.1. Split selection (SS) weighting

The first procedure is Split Selection (SS) weighting. Matched clusters are created using variables from main survey wave 10, the (existing) selection weights w_j are split among cluster members, and then (COVID-19) Study IP-NR weights $1/\rho_j$ and $1/\rho_l$ are estimated for the larger dataset. In step one, the clusters are defined as follows: synthetic selection weights are predicted for unweighted respondents based on the linear regression of the logs of respondent existing selection weights on the main survey wave 10 characteristics (weight logs are modelled to prevent negative values);

then, the synthetic weights of unweighted respondents are transformed back to the natural scale, and these weights used along with the existing weights to define clusters of respondents, with the weights effectively acting as proxies for their characteristics, in the following way. The existing selection weight w_j (or weights if there is more than one of the same value) closest to the value of the synthetic weight for l is identified, and pair (j, l) are taken to form a cluster. If a synthetic weight is equally close to two (or more) existing selection weights, one (or more) larger and one (or more) smaller, then the unweighted respondent is defined as forming a cluster with the respondent(s) with the larger existing weight(s) to prevent existing weight respondents from being included in more than one cluster. A graphical depiction of these different ways in which clusters may be defined is presented in Fig. 1.

In step two, the existing selection weights in each cluster can be split among all the cluster members as follows:

$$w_{split} = \sum_{j=1}^{n_c} w_j / n_c = w_j / n_c \quad (1)$$

where n_c is the number of units in matched-cluster c and is equal to 2 unless there are ties. Informally, both w_j and w_l in the cluster now receive w_{split}. This calculation is analogous to the weight splitting proposed by Lavallee (2007). A justification of it can also be based on the simultaneous sampling scheme proposed by Robbins et al. (2021, section 2.1.2) by assuming that the sample with selection weights and the sample without weights were obtained using different selection procedures, both of which are ignorable given the wave 10 variables, but drawn from the same population (a further working assumption is made that, given the wave 10 variables, the probability of unit having a selection weight is 0.5). An alternative scheme is 'weight

donation' in which j keeps its weight and w_j is donated to l to estimate unknown w_l. This leads to consistent inference if the selection probabilities for j and l are equal given the wave 10 variables. In practice, neither set of assumptions is likely to hold exactly so the pragmatic issue is whether the procedure, using either splitting or donation, preserves the estimates based on IP-NR weighted analyses while bringing the greatest improvement in precision.

In step 3, the same main survey wave 10 predictors used to create the synthetic selection weights are used to model the COVID—19 Study wave response propensities among those with the new (split-) selection weights estimated in step 1, where model selection is again undertaken using Lasso and post-selection prediction using OLS (see section 2.3 for details). The product of the new selection weight

and the inverse of individual estimated response propensities gives the COVID-19 Study SS weights.

3.2. Split IP-NR (SIP-NR) weighting

An issue the SS procedure in the COVID-19 Study is that it is unable to provide weights for the non-trivial number of Study respondents (see below) who did not respond to main survey wave 10 and so do not possess the information used to predict synthetic main survey (selection) weights and Study response propensities (see section 4.2 for numbers of such respondents). Hence, a second procedure, SIP-NR weighting, is also introduced that splits the IP-NR weights estimated for respondents to the Study wave

with existing selection weights with unweighted respondents using information present for everyone who responded in the COVID-19 Study, not just those with main survey information.

The SIP-NR procedure proceeds as follows: in the first step, the COVID-19 Study IP-NR weights are constructed as in section 2.3, by estimating the Study response propensities $\rho_j$by regressing the response to the Study wave on the wave 10 variables among those with the (original and not split) main survey wave 10 'selection' weights (using Lasso for model selection as before) then computing the IP-NR weights w_j/ρ_j; in the second step, the IP-NR weights from step one are split with Study respondents lacking such weights, using the Study variables from the relevant wave to match using the same matching procedure as set out in section 3.1, to give the Study wave SIPNR weight. .

In the application of SIP-NR weighting to the COVID-19 Study, predictors from the Study wave in question are used, with three exceptions. Sex and Age are from the main survey basic characteristics file, due to fewer missing values, and Education is not asked in the Study, so an analogue is derived from the main survey responses. This analogue is constructed as follows. Initially, a response is sought from 2020 calendar year dataset i.e. wave 10 year 2 and wave 11 year 1 data. If such a response does not exist, then one is sought from successively the wave 10 year 1 dataset, the wave 9 dataset, the wave 8 dataset, and the (part later collected) wave 11 year 2 dataset. If a response is still absent (~300 at wave 1), then one is imputed using existing response category probabilities. Following this, as with the Study IP-NR weights (see section 2.3), the split weights are scaled to have a mean of 1 and trimmed.

4. Evaluation methods

4.1. Dataset sizes and respondent sociodemographic characteristics

COVID-19 Study weighted dataset sizes given the IP-NR, SS and SIP-NR procedures are reported. In addition, the Study measured sociodemographic characteristics of respondents weighted by each procedure are presented, enabling evaluation of sub-group analysis feasibility. Main survey wave 10 information could not be used in the latter comparisons because some SIP-NR weighted respondents lacked it (see section 3.2). Similarly, since COVID-19 Study non-respondents lacked Study measured information, it was not possible to compare respondents to the Study sample.

4.2. Weight performance

4.2.1. Non-response bias reduction

The aim of weighting is to eliminate non-response bias, but unequal weights can lead to inefficient estimators, so bias reduction may come with considerable loss of precision (Little & Vartivarian 2005). Hence, both these elements must be considered in performance evaluations. Concerning bias reduction, COVID-19 Study IP-NR and SS weights are evaluated by quantifying the ability of weighted mean estimates of main survey wave 10 measured characteristics to recover similarly measured benchmarks computed for the Study sample using the main survey wave 10 weights. This approach avoids difficulties associated with obtaining external benchmarks (e.g. Hand 2018). The characteristics considered include both those in weight response propensity models and those not. However, it is not possible to evaluate SIP-NR weighted mean estimates on a similar basis because the main survey wave 10 non-respondents lack

relevant auxiliary information (see section 3.2). Hence, these weights are instead evaluated by quantifying their ability to recover benchmark Study IP-NR weighted mean estimates of Study wave measured characteristics: such information exists for all respondents. It should be noted though, that this means that it is not possible to investigate whether they reduced biases more than Study IP-NR weights (SS weights are also similarly evaluated for comparison).

Comparing benchmark and comparator weighted estimates is problematic in terms of statistical testing because the individuals in the former datasets are a subset of those in the later i.e. there are partial dependencies. Had dependencies been complete (i.e. datasets consisted of the same individuals), a suitable paired test could have been used. A test that does account for partial dependencies has recently been proposed (Crossley et al. 2021; Moore et al. 2024), but its derivation relies on the comparator dataset being a subset of the benchmark dataset rather than vice versa. Hence, while it can be used to compare IP-NR weighted estimates to main-survey weighted benchmarks (IP-NR weighted respondents are a subset of main survey weighted respondents), it is not strictly appropriate to use it to evaluate SS and SIP-NR weighted estimates because datasets include respondents not in the benchmark main survey or COVID-19 Study IP-NR weighted datasets. In this paper, unpaired T tests are used to compare benchmark and comparator estimates even though it means that their dependencies are not accounted for. Despite the use of such a test leading to an increased false-positive rate of false positives, we note that a) in the current context it is less of an issue than an increased rate of false negatives, and b) substantively similar results to those reported in section 5.3.1 are obtained using the

test mentioned previously (unpublished results). In addition, as overall performance measures, for each weight means across all studied characteristics of absolute estimate differences compared to benchmarks, standardized by benchmark estimate standard deviations, are reported.

4.2.2. Precision loss

To reduce precision loss, the last step in each of the weighting procedures is to replace weights more than 25 times the weight median with the threshold value (trimming: see section 2.3). A potential consequence of weight splitting is that if the extra weighted respondents have similar characteristics to those with weights above this threshold, the amount of trimming and therefore its likely impact on bias reduction may be decreased. Hence, to test this possibility, first numbers and (due to differing dataset sizes) proportions of weights trimmed given each weighting procedure are quantified.

Second, the DEFF (Kish 1965) is used to quantify precision loss due to the trimmed weights. This metric provides a conservative estimate (weighting variables and outcomes of interest are assumed to be uncorreleatd) of the extent to which survey sampling error is expected to depart from that under simple random sampling with a 100% response rate. It is calculated as follows:

$$DEFF = 1 + (SD(weights)/mean(weights))^2 \qquad (7)$$

where $SD(weights)$ is the weight standard deviation. A larger value implies greater precision loss. DEFFs are also transformed (= N / DEFF, where N is dataset size) to estimate effective dataset size.

5. Results

5.1. Dataset sizes

COVID-19 Study dataset sizes are reported in Table 2. Cross-sectional datasets include all weighted respondents to the wave. Longitudinal datasets include weighted respondents to the wave and all waves prior, so are smaller in size. Both decrease in size over waves due to attrition, except for the waves 8 and 9 cross-sectional datasets, which are larger than the wave 7 equivalent (incentives to complete the survey were offered at both waves). SS weighted datasets (see Table 1 for a summary of the considered weighting procedures and the respondents that they weight) are 7-15% larger than IP-NR weighted datasets. SIP-NR weighted datasets are 15-25% larger than IP-NR weighted datasets.

5.2. Respondent sociodemographic characteristics

With nine COVID-19 Study waves, there are too many weight type (cross-sectional or longitudinal) / wave combinations to report, so Table 3 focusses on the Study measured characteristics of the waves 2 cross-sectional and 8 longitudinal datasets. The former is reported in columns (i) to (iv)). IP-NR weighted respondent (the largest dataset element) characteristics (column (i)) reflect those of all respondents (column

(iv)), though younger Age category proportions are slightly lower, and older Age, 'Tenure: Owned' and 'Long term health condition: Yes' category proportions are slightly higher. SS but not IP-NR (column (ii)) and SIP-NR weighted only respondents (column (iii)) differ from IP-NR weighted respondents. Especially for SIP-NR only, compared to all respondents, Male, older Age, Degree, 'HH type: Single, no kid(s)', 'HH type: Couple, no kid(s)', 'Tenure: Owned' and 'Long term health condition: Yes' category proportions are lower, and younger age (sometimes much), A level, No Qualifications, 'Ethnic minority: Yes', 'Tenure: Mortgage', 'Tenure: Rented' and 'HH type: Couple, kid(s)' category proportions are higher. Though explicit comparisons were not possible (non-respondents lacked Study information), these differences should mean that the weight shared datasets better resemble the Study sample than IP-NR weighted datasets.

Similar occurs with the wave 8 longitudinal datasets (columns v) to viii)). Differences between the all respondent dataset (column viii)) and its wave 2 cross-sectional equivalent are due to non-random attrition. There is also clear evidence with these datasets that weight sharing makes subgroup analyses more feasible, with 44 IP-NR weighted respondents aged 16-19 (rounded, 0.006 * 7325: column v) & see Table 1 for dataset sizes), and 94 (0.011 * 8569) in the SIP-NR weighted (all respondent) dataset.

5.3. Weighted dataset performance

5.3.1. Non-response bias reduction

In this section, waves two cross-sectional and eight longitudinal weight performance is reported. Waves five and eight cross-sectional and two and five longitudinal weight performance is reported in the Appendix and mentioned below. Wave two cross-sectional weights perform well. IP-NR and SS weight performance in the statistical tests evaluating the recovery of benchmark main survey measured and weighted means is reported in columns i) to iii) in Table 4. Main survey estimates (for 10 characteristics in response propensity models, 5 not) are in column (i). IP-NR estimate differences are small (column (ii)), with three statistically significant (maximum (max) = 0.023). SS weights perform worse (column(iii)), with seven significant (max = 0.023). SS and SIP-NR weight performance in recovering benchmark Study measured IP-NR weighted means is reported in columns i) to iii) in Table 5. IP-NR estimates (for 10 characteristics in response probability models, five not) are in column (i). SS estimate differences are small (column (ii), with none significant (max = 0.009). With SIP-NR weights (which can only be evaluated this way: see section 4.2.1) (column (iii)), one difference is significant (max = 0.013).

Wave eight longitudinal weights also perform well. IP-NR and SS weight performance in recovering benchmark main survey measured and weighted means is reported in columns iv) to vi) in Table 4. Main survey estimates are in column (iv). IP-NR estimate differences are slightly larger than for the wave two cross-sectional weights (column (v)), with nine significant (max = 0.039). SS weights perform slightly better (column (vi)), with seven significant (max = 0.033). Wave eight longitudinal SS

and SIP-NR weight performance in recovering benchmark Study measured IP-NR weighted means is reported in columns iv) to vi) in Table 5. IP-NR estimates are in column (iv). SS estimate differences are small (column (v)), with none significant (max = 0.010). With SIP-NR weights, one difference is significant (max = 0.013). Results for the other weights studied are comparable (see Appendix, Tables 1 to 4). Cross-sectional SS weights perform slightly worse than IP-NR weights at recovering main survey measured and weighted benchmarks at waves five and eight. Longitudinal SS weights perform slightly worse than IP-NR weights at recovering such benchmarks at wave two, but the two sets of weights perform similarly at wave five. SIP-NR weights recover Study measured IP-NR weighted benchmarks slightly better than SS weights except in the wave two longitudinal dataset, where the opposite is found.

In addition, as overall performance measures, in Figs. 2 and 3 for each evaluated weight the means of absolute values of differences compared to benchmark weighted estimates of characteristics reported in Tables 3 and 4, standardized by benchmark estimate standard deviations, are presented. Results largely reaffirm those from the statistical tests. With wave two cross-sectional weights, differences compared to benchmark main survey measured and weighted means are reported in Fig 2a. The mean of differences given IP-NR weighted means is 0.018 (largest single value (LSV) = 0.066). That given SS weights is slightly larger (= 0.022), with the LSV slightly smaller (= 0.049). Differences compared to benchmark Study measured IP-NR weighted means are reported in Fig 3a. The mean of differences given SS weighted means is 0.011 (LSV = 0.023). That given SIP-NR weights is slightly smaller (= 0.007), with the LSV also slightly larger (= 0.026). With wave eight longitudinal weights,

differences compared to benchmark main survey measured and weighted means are reported in Fig 2b. The mean of differences given IP-NR weighted means is 0.043 (LSV = 0.111). That given SS weights is slightly smaller (= 0.034), with the LSV also slightly smaller (= 0.093). Differences compared to benchmark Study measured IP-NR weighted means are reported in Fig 3b. The mean of differences given SS weighted means is 0.012 (LSV = 0.027). That given SIP-NR weights is similar (= 0.012), with the LSV slightly larger (= 0.04).

Results for the other studied weights are comparable (see Appendix, Figs 1-4). With wave five cross-sectional weights, the mean difference between benchmark main survey measured and weighted means and IP-NR weighted means is 0.021 (LSV = 0.085). That given SS weighted means is slightly larger (= 0.024), with the LSV smaller (= 0.060). The mean difference between benchmark Study measured IP-NR weighted means and SS weighted means is 0.016 (LSV = 0.038). That given SIP-NR weights is slightly smaller (= 0.009), with the LSV also slightly smaller (= 0.027). For wave eight equivalents, the mean difference between benchmark main survey measured and weighted means and IP-NR weighted means is 0.018 (LSV = 0.075). That given SS weights is slightly larger (= 0.023), with the LSV smaller (= 0.058). The mean difference between benchmark Study measured IP-NR weighted means and SS weighted means is 0.013 (LSV = 0.029). That given SIP-NR weights is slightly smaller (= 0.0011), with the LSV also slightly smaller (= 0.025).

With wave two longitudinal weights, the mean difference between benchmark main survey measured and weighted means and IP-NR weighted means is 0.021 (LSV = 0.055). That given SS weights is slightly larger (= 0.026), with the LSV slightly larger

(= 0.053). The mean difference between benchmark Study measured IP-NR weighted means and SS weighted means is 0.010 (LSV = 0.02). That given SIP-NR weights is slightly smaller (= 0.009), with the LSV slightly larger (= 0.029). With wave five equivalents, the mean difference between benchmark main survey measured and weighted means and IP-NR weighted means is 0.027 (LSV = 0.086). That given SS weights is similar (= 0.027), with the LSV smaller (= 0.065). The mean difference between benchmark Study measured IP-NR weighted means and SS weighted means is 0.014 (LSV = 0.035). That given SIP-NR weightsis slightly smaller (= 0.007), with the LSV also slightly smaller (= 0.025).

It should be noted that comparable results in terms of bias reduction were obtained for the two procedures when 'weight-donation' (unweighted cluster members are given weighted cluster member undivided weights) instead of weight-splitting schemes were used in weight assignment (see also section ??). Procedure weighted estimate differences compared to benchmarks and mean standardized biases using weight donation were similar in size to when weight-splitting was used (unpubl. results).

5.3.2. Precision loss

In Table 6, the number and proportion of weights trimmed (replacing weights more than 25 times the weight median with the threshold value to reduce precision loss: see section 4.2.2) in the cross-sectional and longitudinal waves 2, 5 and 8 datasets are reported. Numbers are slightly greater for the SS than the IP-NR and SIP-NR datasets

(which are similar). However, proportions tend to be lowest in the SIP-NR datasets, implying that the extra respondents in these datasets are more similar to those with extreme weights, decreasing the amount of trimming and its impact on bias reduction.

DEFFs and effective dataset sizes given the same trimmed weights are also reported in Table 6. SS weight DEFFs are larger than IP-NR weight equivalents, but SIP-NR weight DEFFs are smaller than equivalents for both other weight types, implying that SIP-NR weights most reduced precision loss, followed by IP-NR then SS weights. Given also differences in real dataset sizes, this meant that effective dataset sizes were largest for the SIP-NR weighted datasets, then the SS weighted datasets, then the IP-NR weighted datasets.

It should be noted that findings differed when weight-donation instead of weight-splitting schemes were used in weight assignment. With weight donation, SIP-NR weight DEFFs were larger than IP-NR and SS weight DEFFs, with the consequence that SS weight effective dataset sizes were larger than SIPNR weight effective datasets for three of the six datasets considered (unpubl. Results). These differences occur because with weight splitting weighting otherwise unweighted individuals led to a reduction in extreme weight value sizes (due to being split with unweighted individuals with similar characteristics). With weight donation, the latter individuals are instead assigned the forementioned weight, inflating weight variability.

6. Discussion

We proposed two new procedures that assign weights mapping the survey respondents to the target population that include those who, lacking selection weights, cannot be weighted using Inverse Propensity Non-Response (IP-NR) methods (recalling that IP-NR weight = selection weight * non-response weight). Both form (matched) clusters of unweighted and existing weight individuals with similar characteristics, and assign weights to unweighted individuals split (sum of existing weights in cluster / total number of cluster members) or donate (assign as the estimated weight values) the existing weights of cluster members. The accuracy of inference for the different procedures and splitting/donation schemes depends on how well the underlying assumptions described in section 3.1 hold for UKHLS and the COVID-19 Study, which depends on the choice of variables used to match the weights and model the response propensities. The first procedure, SS weighting, splits / donates existing selection weights, then estimates IP-NR weights for the Study wave in question. The second, SIP-NR weighting, splits / donates IP-NR weights for the Study e estimated for those with existing selection weights (see Table 1 for a summary of weights estimated).

Procedure performance was evaluated using the UK Household Longitudinal Study (UKHLS) COVID-19 Study datasets. Evaluations considered weighted dataset sizes and respondent sociodemographic characteristics, also enabling subgroup analysis feasibility to be studied. In addition, non-response bias reduction and precision loss due to weight use (weights are inefficient: Little & Vartivarian 2005) was quantified. SS weighted datasets were 7-15% and SIP-NR weighted datasets 15-25%

larger than IP-NR weighted equivalents, reducing sampling error. SS and SIP-NR weighted dataset sizes differed due to some COVID-19 Study respondents not responding to main survey wave 10 and so lacking the information used by SS weighting to split selection weights (in contrast, SIP-NR weighting used information from the survey wave, which existed for all respondents, to split IP-NR weights, though SS weighting did benefit from it being possible to provide selection weights for Study non-respondents as well as respondents with main survey wave 10 information: see section 3). SS but not IP-NR and especially SIP-NR only weighted respondents differed (were younger, less educated, less likely to have children, more likely to be ethnic minorities) from IP-NR weighted respondents. Explicit comparisons were not possible because non-respondents lacked Study information, but these differences should mean that SS and SIP-NR weighted datasets better resemble the Study sample than IP-NR weighted datasets (see also following paragraph). Moreover, subgroup analyses will be more feasible: for example, there were more than twice as many respondents aged 16-19 in the wave 8 longitudinal SIP-NR weighted dataset than in the IP-NR weighted equivalent (as they only considered responses to the Study, longitudinal as well as cross-sectional IP-NR weights suffered from zero weights).

Non-response bias was evaluated by quantifying how well weighted estimates of respondent characteristics recovered benchmark estimates. Study IP-NR and SS weighted mean estimates of main survey wave 10 measured characteristics were statistically compared to main survey wave 10 weighted benchmarks. However, this was not possible for SIP-NR weights because, as noted previously, some respondents lacked main survey wave 10 information. Hence, though it meant that whether they

reduced biases more than Study IP-NR weights could not be studied (extra respondents may improve performance if datasets better resemble study populations: Schouten et al. 2016; Moore et al. 2024), they were instead evaluated by comparing mean estimates of Study measured characteristics to Study IP-NR weighted benchmarks (SS weights were also similarly evaluated). Focusing on 'split' weights, the tests showed that Study IP-NR weights recovered main survey weighted benchmarks slightly better than SS weights in four of the six wave / type (cross-sectional or longitudinal) combinations evaluated, in another the opposite was found, and the weights performed similarly in the final combination. SIP-NR weights recovered Study IP-NR weighted benchmarks slightly better than SS weights in three of the combinations, in another the opposite was found, and the weights performed similarly in the final two combinations. In addition, absolute differences were standardized by benchmark estimate standard deviations and means calculated to provide overall performance measures. Results were mostly similar to those from the statistical tests (IP-NR weights performed very slightly better than SS weights in the combination where the tests showed they performed equally well, and SIP-NR weights performed very slightly better than SS weights in the two combinations where the tests showed they performed equally well).

Precision loss was evaluated in several ways. First, weights were trimmed (values more than 25 times the weight median were replaced with threshold values) as a last step in weighting procedures to reduce precision loss. Trimming is likely to decrease bias reduction, but with weight splitting less may be needed if the extra respondents have the same characteristics as those with extreme weights. To study

this, numbers and (given differing dataset sizes) proportions of trimmed weights were quantified. Numbers trimmed increased with weight splitting, but proportions were lower in SIP-NR weighted datasets than IP-NR or SS weighted datasets. This implies that SIP-NR weighting enabled less trimming, decreasing impacts on bias reduction. Second, for trimmed weights, precision loss was quantified using DEFFs (Kish 1965), which estimate the extent to which sampling error departs from that given simple random sampling with 100% response. All SS weight DEFFs were larger than IP-NR equivalents, implying greater precision loss. However, SIP-NR weight DEFFs were always smaller than equivalents given both the IP-NR and SS weights, implying that SIP-NR weighting reduced precision loss compared to the other procedures. In addition, DEFFs were transformed to estimate effective weighted dataset sizes. Given also numbers of respondents in datasets, the SIP-NR procedure resulted in the largest effective dataset sizes, followed by the SS procedure, then the IP-NR procedure.

Findings concerning bias reduction when weight-donation rather bias spltting schemes were used to assign weights were similar to those outlined above. However, patterns in DEFFs and estimated effective dataset sizes differed, with SS weight datasets often being larger than SIPR weight datasets. These differences occur because with weight splitting weighting otherwise unweighted individuals led to a reduction in extreme weight value sizes (due to being split with unweighted individuals with similar characteristics). With weight donation, the latter individuals are instead assigned the forementioned weight, inflating weight variability.

This research has implications for both the UKHLS COVID-19 Study and survey design in general. For the COVID-19 Study, it shows that the new procedures weight

more respondents than IP-NR methods, reducing sampling error and increasing subgroup analysis feasibility, while maintaining and indeed improving weight performance in terms of reducing non-response bias and minimizing precision loss. Hence, as the procedure resulted in the most weighted respondents and the largest effective weighted dataset sizes, in the December 2021 dataset release SIP-NR weights were supplied (see Institute for Social and Economic Research 2021).

Concerning survey design more generally, the question arises as to whether the new procedures can be used in other surveys. Respondents without cross-sectional IP-NR weights occur in most HH panel surveys (Schonlau et al. 2013). Moreover, previous solutions to this issue have limitations. Sharing existing selection weights among HH members then estimating IP-NR weights provides unbiased estimates and is used in many surveys (Ernst 1989; Lavallée 1995; 2007; Heeringa et al 2011; Taylor et al. 2018; University of Essex & Institute for Social and Economic Research 2019; Zhang 2021). However, it cannot weight those in HHs without existing selection weighted members. Predicting wave 1 HH selection probabilities for unweighted respondents, using these to adjust current wave values for multiple selection paths, then estimating individual selection weights and IP-NR weights can weight all respondents and is used in several other surveys (Haisken-Denew & Frick 2005; Watson 2012). However, it is model based, predictions can only be made using responding HH selection probabilities, and adjustments for multiple selection paths require assumptions. By contrast, the new procedures can also (potentially: see below) weight all respondents and are more easily implemented than selection

probability prediction. In addition, in the COVID-19 Study they performed well at reducing non-response bias and precision loss due to weight use.

That said, it is advised that the new procedures only be used to weight all non IP-NR weighted survey respondents if there is no alternative i.e. if, as in the UKHLS COVID-19 Study, HH structure is not enumerated. As noted previously, HH weight sharing produces unbiased estimates, whereas weight splitting is justified on the basis of exchangeability of weighted and unweighted respondents with the same characteristics and so is model based. Moreover, SIP-NR weighting can only split the weights of current wave respondents, so when less than 100% of sample members respond, unweighted respondents may be assigned the (split) weights of those with less similar characteristics than if all had responded (see also section 3.2). To a lesser extent, this issue also occurs with SS weighting: only respondent selection weights can be split (as mentioned in the last paragraph, an analogous issue also arises with selection probability prediction). If the new procedures must be used, the evaluations reported here suggest that SIP-NR weighting should be utilised. However, if comparable auxiliary information exists for all respondents to the survey wave, for example because they all responded to a previous wave (one reason for supplying COVID-19 Study SIP-NR weights is that such information did not exist – see previously), SS weights should be estimated and performance compared. This performance comparison will also be of wider interest to survey designers: as noted previously, the fore-mentioned lack of comparable information on all Study respondents also affected the methods used to evaluate procedure performance in this paper.

In surveys in which HH structure is enumerated, a different role for the new procedures is envisaged: to weight those not weighted by HH weight sharing (see Schonlau et al. 2013 for the suggestion that a similar strategy involving selection probability prediction be used). If information on all HH members exists from HH or individual questionnaires, a variation of SS weighting should be evaluated. Selection weights or their analogues can be predicted for all enumerated sample members not weighted by HH weight sharing, then, using the same information, IP-NR weights can be computed after response propensity estimation for the larger sample (see section 2.4 for the use of the HH weight sharing element of this strategy in the UKHLS main survey). If this is not possible, SS weighting as utilized in this paper can instead be used in the outlined procedure (though new survey entrants will not be weighted due to lacking comparable auxiliary information), or, after IP-NR weight estimation using HH weight shared selection weights, SIP-NR weighting can be utilized. With all methods, weight performance should be evaluated, ideally compared to that of weights estimated by other methods, including selection probability prediction. However, now that multiple methods exist, it is also noted that given the benefits there is no reason for survey designers not to seek to ensure that all respondents are supplied with non-response weights.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020) lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, 20*:* 176-235.

Benzeval, M., Bollinger, C. R., Burton, J., Crossley, T.F. & Lynn, P. (2020) *The representativeness of Understanding Society.* Understanding Society Working Paper Series 2020–08, Institute for Social and Economic Research.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Crossley, T. F., Fisher, P. & Low, H. (2021) The Heterogeneous and Regressive Consequences of COVID-19: Evidence from High Quality Panel Data. *Journal of Public Economics*, 193: 104344.

Dever, J. & Valliant, R. (2018) *Survey Weights: A Step-by-Step Guide to Calculation*. Stata Press.

Ernst, L.R. (1989) Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G. & Singh, M.P. (eds), p. 135–159. Wiley & Sons, New York,

Groves, R. M. & Heeringa, S. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs*. J. Roy. Stat. Soc. Ser. A.*, 169, 439-457.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. (eds.) (2001) *Survey Nonresponse*. Wiley Series in Survey Methodology.

Hand D.J. (2018) Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society, Series A*, 181: 555-605.

Haisken-DeNew, J.P. & Frick, J.R. (2005) *Desktop Companion to the German socio-economic panel study (SOEP).* Technical report, German Institute for Economic Research (DIW).

Harrell, F.E. Jr. (2001) *Regression modelling strategies: with applications to linear models, logistic regression and survival analysis.* New York: Springer.

Heeringa, S., Berglund, P., Khan, A., Lee, S., & Gouskova, E. (2011). *PSID Cross-sectional individual weights, 1997–2009.* Ann Arbor, MI: Institute for Social Research, University of Michigan.

Institute for Social and Economic Research (2021) Understanding Society COVID-19 User Guide. Version 10.0, October 2021. Colchester: University of Essex.

Institute for Social and Economic Research (2022) *Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009, User Guide,* Colchester: University of Essex.

Kalton, G. & Brick, J.M. (1995) Weighting schemes for household panel surveys. *Survey Methodology*, 21:33–44.

Kish, L. (1965) *Survey Sampling.* Wiley: New York.

Lavallee, P. (1995) Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21:25–32.

Lavallée, P. (2007) *Indirect Sampling,* New York: Springer

Little, R. J. A. & Vartivarian, S. (2005) Does weighting for nonresponse increase the variance of survey means? *Survey Methods.* 31: 161-168.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*, Wiley: New York.

Lynn, P. (2009) *Sample Design for Understanding Society.* Understanding Society Working Paper Series, 2009–01, Institute for Social and Economic Research.

Lynn, P. & Kaminska, O. (2010) *Weighting Strategy for Understanding Society.* Understanding Society Working Paper Series 2010–05, Institute for Social and Economic Research.

Lynn, P., Burton, J., Kaminska, O., Knies, G., & Nandi, A. (2012) *An initial look at non-response and attrition in Understanding Society.* ISER, University of Essex.

Moore, J.C., Burton, J., Crossley, T. F., Fisher, P., Gardiner, C., Jäckle, A., & Benzeval, M. (2024) *Assessing Bias Prevention and Bias Adjustment in a Sub-Annual Online Panel Survey*. Understanding Society Working Papers Series ??

Peytchev, A., Riley, S., Rosen, J., Murphy, J. & Lindblad, M. (2010) Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods,* 4: 21-29.

Robbins, M.W., Ghosh-Dastidar, B., & Ramchand, R. (2021) Blending probability and non-probability samples with applications to a survey of military caregivers. *J. Surv. Statist. Methodol.*, 9: 1114-1145.

Schonlau, M., Kroh, M., & Watson, N. (2013) The implementation of cross-sectional weights in household panel surveys. *Statistics Surveys*, 7: 37-57.

Schouten, B., Cobben, F., Lundquist, P. & Wagner, J. (2016) Does more balanced survey response imply less non-response bias? *J. Roy. Stat. Soc. Ser. A,* 179: 727-748. DOI: 10.1111/rssa.12152

Steyerberg, E.W., Eijkemans, M.J.C. & Habbema, J.D.F. (2001) Application of shrinkage techniques in Logistic regression analysis: a case study. *Statistica Neerlandica,* 55: 76–88. DOI: 10.1111/1467-9574.00157.

Taylor, M.F. (ed). with Brice, J., Buck, N. & Prentice-Lane, E. (2018) British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. Colchester: University of Essex

Tibshirani, R. (1996) Regression and shrinkage via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

University of Essex & Institute for Social and Economic Research (2019). *Understanding Society: The UK Household Longitudinal Study, Waves 1-9 User Guide.* Institute for Social and Economic Research.

University of Essex & Institute for Social and Economic Research. (2021). *Understanding Society: COVID-19 Study, 2020-2021*. [data collection]. *11th Edition.* UK Data Service. SN: 8644, DOI: 10.5255/UKDA-SN-8644-11

University of Essex & Institute for Social and Economic Research. (2023). *Understanding Society: Waves 1-12, 2009-2021 and Harmonised BHPS: Waves 1-18, 1991-2009*. [data collection]. *17th Edition.* UK Data Service. SN: 6614, DOI: http://doi.org/10.5255/UKDA-SN-6614-18

Wagner, J. R. (2008) *Adaptive Survey Design to Reduce Nonresponse Bias.* PhD diss., University of Michigan, Michigan.

Watson, N. (2012) *Longitudinal and cross-sectional weighting methodology for the HILDA survey.* Technical report, University of Melbourne.

Wooldridge, J. M. (2010) *Econometric analysis of cross section and panel data*. 2nd edition. MIT press.

Zhang, L.-C., & Patone, M. (2017) Graph sampling. *Metron*, 75: 277-299. DOI: 10.1007/s40300-017-0126-y

Table 1: The different forms of non-response weight estimated and evaluated in this paper.

Table 2: Cross-sectional and longitudinal COVID-19 Study weighted dataset sizes. Cross-sectional datasets contain respondents to the mentioned wave. Longitudinal datasets contain respondents to all waves up to and including the mentioned wave i.e. wave 4 includes respondents to all of waves 1, 2, 3 and 4. IP-NR datasets contain only respondents with the UKHLS main survey wave 10 weight required for IP-NR weight production. SS datasets contain the same respondents plus those with main survey information, which together can be weighted by the SS procedure. SIP-NR / N datasets contain all respondents, which together can only be weighted by the SIP-NR procedure.

Table 3: COVID-19 Study focal wave measured sociodemographic characteristics of the components of the wave 2 cross-sectional and wave 8 longitudinal datasets. We present the characteristics of IP-NR weighted respondents (respectively columns (i) & (v)), of respondents weighted by SS but not IP-NR methods (columns (ii) & (vi)), of respondents weighted only by SIP-NR methods (columns (iii) & (vi)); and of all respondents combined i.e. (i) + (ii) + (iii) and (v) + (vi) + (vii) (columns (iv) & (vii)).

Table 4: COVID-19 Study wave 2 cross-sectional and wave 8 longitudinal weight performance in recovering means of main survey measured and weighted characteristics. Main survey weighted mean estimates and (in brackets) their standard errors ('wt. est.'; columns (i) & (iv)); tests of differences between such estimates and estimates given COVID-19 Study IP-NR weights (columns (ii) & (v)); and tests of differences between such estimates and estimates given COVID-19 Study SS weights (columns (iii) & (vi)) are reported. * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$. Differences exist between the two sets of main survey estimates due to more sample member deaths by wave 8.

Table 5: COVID-19 Study wave 2 cross-sectional and wave 8 longitudinal weight performance in recovering means of COVID-19 Study measured IP-NR weighted characteristics. COVID-19 Study IP-NR weighted mean estimates and (in brackets) their standard errors ('wt. est.'; columns (i) & (iv)), tests of differences between such estimates and estimates given SS weights ('wt. diff.'; columns (ii) & (v)), and tests of differences between such estimates and estimates given SIP-NR weights (columns (iii) & (vii)) are reported. * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$.

Table 6: Numbers of trimmed weights and their proportions, and DEFFs and effective dataset sizes given the COVID-19 Study waves 2, 5 and 8 cross-sectional and longitudinal IP-NR, SS and SIP-NR weights.

Table 1:

Weighting strategy	Methods
1) Inverse propensity non-response (IP-NR) weight	a) For original survey sample members with (existing) selection weights, model and estimate response propensities for given survey wave. b) IP-NR weight = selection weight * (1 / response propensity).
2) Split selection (SS) weight	a) Split the existing selection weights of original sample members with unweighted survey sample members with similar characteristics. b) For this larger sample with the new weight, estimate response propensities and compute IP-NR weights for given survey wave as in 1).
3) Split IP-NR (SIP-NR) weight	a) For survey original sample members with existing selection weights, estimate response propensities and compute IP-NR weights for given survey wave as in 1). b) Split the IP-NR weights computed for original sample members in step a) with unweighted survey respondents with similar characteristics.

Table 2:

	COVID-19 Study wave								
	1	2	3	4	5	6	7	8	9
Cross-sectional:									
IP-NR	13994	12013	11515	11261	10616	9989	9915	10615	10489
SS	16604	14086	13453	13112	12332	11556	11471	12129	12250
SIP-NR / N	17761	14811	14123	13754	12876	12035	11968	12680	12818
Longitudinal:									
IP-NR		11220	10293	9957	8857	8102	7610	7325	6857
SS		13106	11968	11061	10202	9286	8697	8335	7801
SIP-NR / N		13698	12437	11458	10541	9574	8947	8569	8009

Table 3:

	Wave 2 cross-sectional				Wave 8 longitudinal			
	IP-NR	SS not IP-NR	SIP-NR only	All	IP-NR	SS not IP-NR	SIP-NR only	All
	(i)	(ii)	(iii)	(iv) = (i) + (ii) + (iii)	(v)	(vi)	(vii)	(viii) = (v) + (vi) + (vii)
Gender: Male	0.418	0.395	0.374	0.413	0.420	0.404	0.363	0.417
Age: 16-19	0.019	0.011	0.178	0.025	0.006	0.003	0.201	0.011
Age: 20-29	0.073	0.136	0.145	0.086	0.051	0.094	0.098	0.057
Age: 30-39	0.111	0.174	0.163	0.123	0.088	0.139	0.090	0.094
Age: 40-49	0.167	0.207	0.167	0.173	0.138	0.194	0.171	0.146
Age: 50-59	0.221	0.224	0.181	0.219	0.215	0.241	0.231	0.218
Age: 60-69	0.215	0.156	0.114	0.202	0.259	0.202	0.154	0.249
Age: 70-79	0.161	0.074	0.040	0.143	0.205	0.105	0.051	0.189
Age: 80-89	0.030	0.018	0.012	0.028	0.036	0.023	0.004	0.033
Age: 90+	0.002	0.000	0.000	0.002	0.002	0.000	0.000	0.002
Qualifications: Degree	0.506	0.518	0.440	0.505	0.514	0.530	0.461	0.515
Qualifications: A-level	0.197	0.219	0.235	0.202	0.193	0.227	0.199	0.197
Qualifications: GCSE or lower	0.290	0.256	0.311	0.286	0.288	0.240	0.330	0.283
Family type: Single, no kid(s)	0.109	0.120	0.092	0.109	0.110	0.130	0.141	0.113
Family type: Single, kid(s)	0.020	0.036	0.034	0.023	0.015	0.038	0.030	0.018
Family type: Couple, no kid(s)	0.302	0.302	0.200	0.297	0.349	0.362	0.239	0.347
Family type: Couple, kid(s)	0.202	0.245	0.308	0.213	0.162	0.207	0.295	0.171
Ethnic minority: Yes	0.103	0.124	0.171	0.109	0.076	0.082	0.126	0.078
Country: England	0.815	0.799	0.797	0.812	0.831	0.807	0.799	0.827
Country: Wales	0.060	0.057	0.057	0.060	0.054	0.045	0.060	0.053
Country: Scotland	0.084	0.100	0.099	0.087	0.078	0.105	0.103	0.082
Country: Northern Ireland	0.041	0.044	0.047	0.042	0.037	0.044	0.038	0.038
Tenure: Owned	0.473	0.330	0.241	0.442	0.547	0.409	0.302	0.524
Tenure: Mortgage	0.356	0.431	0.477	0.372	0.310	0.397	0.474	0.324
Tenure: Rented	0.023	0.038	0.059	0.027	0.018	0.034	0.026	0.020
Long-term illness: Yes	0.530	0.478	0.404	0.517	0.597	0.555	0.479	0.589

Table 4:

	Wave 2 cross-sectional			Wave 8 longitudinal		
	Main	Covid		Main	Covid	
		IP-NR	SS		IP-NR	SS
	wt est.	wt diff.	wt. diff	wt est.	wt diff.	wt. diff
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
In IPW model:						
Subjective financial situation (SFS): comfortable or OK	0.716 (0.003)	0.005	0.015**	0.716 (0.003)	-0.011	-0.006
SFS: just about getting by	0.203 (0.002)	-0.008	-0.013**	0.203 (0.002)	-0.003	-0.002
SFS: finding it quite / very difficult	0.081 (0.002)	0.003	-0.002	0.081 (0.002)	0.015***	0.008*
Tenure: Owned	0.343 (0.003)	0.009	0.023***	0.343 (0.003)	-0.020**	-0.005
Tenure: Mortgage	0.341 (0.003)	-0.021***	-0.015**	0.341 (0.003)	-0.021***	-0.019**
Tenure: Rented	0.119 (0.002)	0.004	-0.004	0.119 (0.002)	0.007	-0.003
Tenure: Social Housing	0.195 (0.002)	0.006	-0.004	0.195 (0.002)	0.033***	0.027***
Low skill occupation	0.362 (0.004)	-0.003	-0.005	0.362 (0.004)	0.011	0.014
Savings income?	0.372 (0.003)	-0.002	0.012*	0.372 (0.003)	-0.034***	-0.032***
Behind with some or all bills	0.059 (0.001)	-0.001	-0.008***	0.059 (0.001)	0.006*	0.003
Not in IPW model:						
Income poverty	0.155 (0.002)	0.010*	0.005	0.155 (0.002)	0.027***	0.025***
Receives core benefit	0.054 (0.001)	0.003	0.001	0.054 (0.001)	0.010***	0.008**
Visited GP	0.700 (0.003)	-0.001	-0.003	0.700 (0.003)	0.004	0.005
Smoker	0.144 (0.002)	0.023***	0.016***	0.144 (0.002)	0.039***	0.033***
Hospital outpatient	0.433 (0.003)	0.002	0.005	0.433 (0.003)	0.004	0.009

Table 5:

	Wave 2 cross-sectional			Wave 8 longitudinal		
	IP-NR	SS	SIP-NR	IP-NR	SS	SIP-NR
	wt est.	wt. diff	wt diff.	wt. est.	wt diff.	wt. diff
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
In IPW model:						
Subjective financial situation (SFS): comfortable or OK	0.756 (0.004)	0.009	-0.001	0.767 (0.005)	0.006	-0.001
SFS: just about getting by	0.182 (0.004)	-0.005	-0.002	0.189 (0.005)	-0.002	0.001
SFS: finding it quite/very difficult	0.062 (0.002)	-0.004	0.003	0.044 (0.002)	-0.004	-0.001
Employed?	0.607 (0.004)	-0.009	-0.001	0.601 (0.006)	-0.010	0.001
Behind with some or all bills	0.073 (0.002)	-0.006	-0.001	0.048 (0.003)	-0.006	-0.004
Behind with housing payments	0.065 (0.002)	-0.004	0.002	0.089 (0.005)	0.001	-0.001
HH type: Couple with children	0.203 (0.004)	-0.005	-0.007	0.179 (0.004)	-0.002	-0.009
HH type: Single, no children	0.121 (0.003)	-0.002	0.002	0.119 (0.004)	-0.006	0.001
Covid test?	0.040 (0.002)	0.000	-0.001	0.226 (0.005)	-0.001	-0.017*
Clinically vulnerable	0.389 (0.004)	0.007	0.013*	0.427 (0.006)	0.010	0.006
Not in IPW model:						
Advised to shield	0.068 (0.002)	0.002	0.003	0.069 (0.003)	-0.000	0.001
Gave or received money	0.151 (0.003)	0.002	0.001	0.116 (0.004)	-0.001	0.000
Less sleep than usual	0.218 (0.004)	-0.003	-0.000	0.181 (0.005)	-0.006	-0.009
More depressed than usual	0.286 (0.004)	-0.001	0.002	0.232 (0.005)	-0.005	-0.006
More lonely than usual	0.086 (0.003)	-0.000	-0.001	0.079 (0.003)	-0.003	-0.004

Table 6:

		IP-NR		SS		SIP-NR	
Type	Wave	Trimmed (Prop.)	DEFF (E.D. Size)	Trimmed (Prop.)	DEFF (E.D. Size)	Trimmed (Prop.)	DEFF (E.D. Size)
Cross-sectional	2	26 (0.0022)	2.779 (4333)	32 (0.0023)	3.047 (4623)	27 (0.018)	2.726 (5433)
	5	27 (0.0025)	2.932 (3621)	35 (0.0028)	3.209 (3843)	26 (0.0020)	2.802 (4595)
	8	18 (0.0017)	2.777 (3822)	25 (0.0021)	3.083 (3934)	26 (0.0021)	2.763 (4639)
Longitudinal	2	28 (0.0025)	2.968 (3780)	40 (0.0031)	3.225 (4064)	29 (0.0021)	2.902 (4720)
	5	23 (0.0026)	2.988 (2964)	33 (0.0032)	3.301 (3091)	19 (0.0018)	2.847 (3702)
	8	18 (0.0025)	2.932 (2498)	25 (0.0030)	3.189 (2614)	17 (0.0020)	2.873 (2983)

Figure 1: A graphical representation of cluster definition in the two new weight sharing procedures. $w_{(1)}, w_{(2)}, w_{(3)}, w_{(4)}$ and $w_{(5)}$ are existing (selection or IP-NR) weights ordered according to size (x-axis). $\widehat{w}_{(1)}, \widehat{w}_{(2)}, \widehat{w}_{(3)}$ and $\widehat{w}_{(4)}$ are similarly ordered synthetic weights. Three clusters of existing and synthetic weights are depicted that show the different ways in which clusters may be defined. Cluster 1 consists of the synthetic weight $\widehat{w}_{(1)}$ and two existing weights that are closest to it, $w_{(1)}$ and $w_{(2)}$, which are of the same size. If instead $w_{(1)} < w_{(2)}$, then this cluster would be formed of $\widehat{w}_{(1)}$ and $w_{(2)}$ only. Cluster 2 consists of the synthetic weight $\widehat{w}_{(2)}$ and the existing weight $w_{(4)}$, which is same distance away from $\widehat{w}_{(2)}$ as $w_{(3)}$ (i.e. $a = b$ where $a = \widehat{w}_{(2)} - w_{(3)}$ and $b = w_{(4)} - \widehat{w}_{(2)}$), but is the larger of the two mentioned existing weights. Cluster 3 consists of the synthetic weights $\widehat{w}_{(3)}$ and $\widehat{w}_{(4)}$, and the existing weight $w_{(5)}$, which is the closest existing weight to both the mentioned synthetic weights. See main text for further explanation.

Figure 2: Box plots of absolute values of the tests of COVID-19 Study weights reported in Table 4, standardised by benchmark estimate standard deviations. In a), tests compare wave 2 cross-sectional dataset IP-NR (white bars) and SS (light grey bars) weighted estimates of main survey measured characteristics to main survey weighted benchmarks. In b), tests compare wave 8 longitudinal dataset IP-NR (white bars) and SS (light grey bars) weighted estimates of main survey measured characteristics to main survey weighted benchmarks. In plots, bars indicate the inter-quartile range, the line within the median value, and the cross the mean value. Whiskers indicate minimum / maximum values, unless values exist that are smaller or larger than the

inter-quartile range, in which case they indicate the smallest / largest value within this range, and the outlying values are indicated by circles.

Figure 3: Box plots of absolute values of the tests of COVID-19 Study weights reported in Table 5, standardised by benchmark estimate standard deviations. In a), tests compare wave 2 cross-sectional dataset SS (light grey bars) and SIP-NR (dark grey bars) weighted estimates of COVID-19 Study measured characteristics to COVID-19 Study IP-NR weighted benchmarks. In b), tests compare wave 8 longitudinal dataset SS and SIP-NR weighted estimates of COVID-19 Study measured characteristics to COVID-19 Study IP-NR weighted benchmarks. In plots, bars indicate the inter-quartile range, the line within the median value, and the cross the mean value. Whiskers indicate minimum / maximum values, unless values exist that are smaller or larger than the inter-quartile range, in which case they indicate the smallest / largest value within this range, and the outlying values are indicated by circles.

Fig. 1

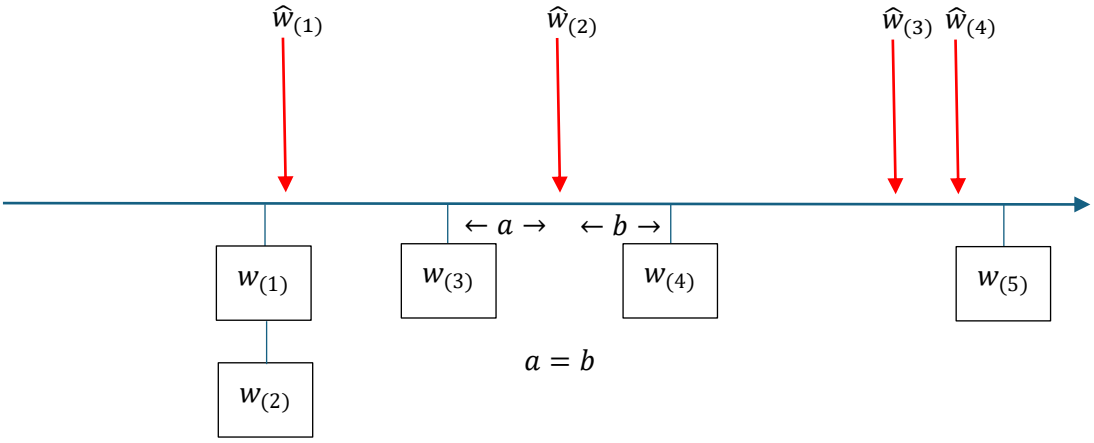

$\widehat{w}_{(1)}$
$\widehat{w}_{(2)}$
$\widehat{w}_{(3)}$
$\widehat{w}_{(4)}$
$\leftarrow a \rightarrow$
$\leftarrow b \rightarrow$
$w_{(1)}$
$w_{(3)}$
$w_{(4)}$
$w_{(5)}$
$w_{(2)}$
$a = b$

Fig. 2

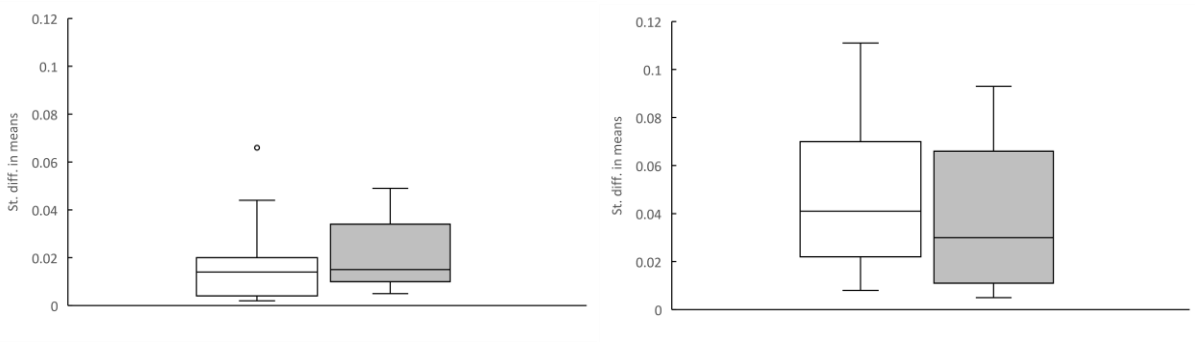
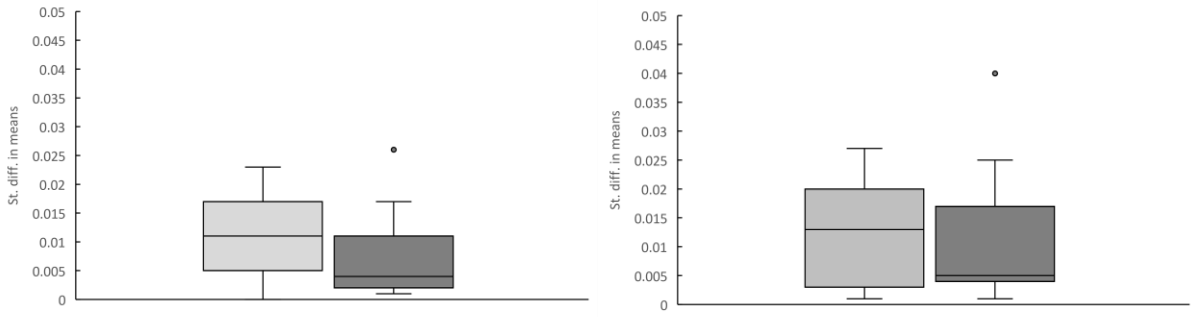

St. diff. in means
0.12
0.1
0.08
0.06
0.04
0.02
0
St. diff. in means
0.12
0.1
0.08
0.06
0.04
0.02
0

Fig. 3


St. diff. in means
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
St. diff. in means
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0

Supplementary Information for 'Two new solutions to the zero non-response weight problem' by J.C. Moore and P.S. Clarke

Appendix A

Lasso variable selection methods: details and use

Lasso procedures are regularised regression methods. As with other regularised regression methods, they minimise the sum of squared deviations between predicted and observed values similar to Ordinary Least Squares (OLS), but in addition impose a regularisation penalty on model complexity (Ahrens et al. 2020). Due to the imposition of this penalty, such methods tend to outperform OLS in terms of out of sample prediction, as reducing model complexity and inducing shrinkage bias decreases prediction error. In doing so, they also address the problem of model overfitting: high in-sample fit, but poor prediction performance on unseen data.

Regularised regression methods incorporate tuning parameters that determine the amount and form of regularisation penalty. With Lasso procedures (Tibshirani 1996; Steyerberg et al. 2001), the mean squared error is minimised subject to a penalty on the absolute size of coefficient estimates:

$$\hat{\beta}_{lasso}(\lambda) = \arg\min \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j\left|\beta_j\right|, \quad (1)$$

where $\hat{\beta}_{lasso}(\lambda)$ are the Lasso estimated coefficients for each predictor in the considered set *p* given the tuning parameter λ that determines the overall penalty level, *n* is sample size, y_i is the value of the response variable for subject *i* = 1,..*n*, x_i' are the values of the predictors for the same subjects, β are the OLS estimated

coefficients for the predictors, and ψ_j are (given λ) predictor-specific penalty loadings. A λ of zero results in the OLS model. Increasing λ ultimately results in an empty model, with all coefficients set to zero. It is this setting of some coefficients to zero and removal of predictors from models that enables Lasso to be used as a model selection technique. Note that in this paper we assume that predictors are uncorrelated and hence that Lasso-type penalisation is all that is necessary, enabling us (after standardising predictors so that they have equal variances) to set ψ_j all to unity: for methods suitable when this assumption is relaxed, see Zhou & Hastie (2005) & Ahrens et al. (2020).

Several techniques exist to choose the value of the tuning parameter λ. The first of these is cross-validation, which explicitly evaluates out of sample prediction performance. The data in question are split into training and validation datasets. The models for different values of λ are then estimated and variables selected using the training dataset. Next, they are fitted to the validation dataset, and mean squared prediction errors calculated to quantify performance (Ahrens et al. 2020). For example, with the commonly used *K*-fold cross-validation technique datasets are split into *K* groups of approximately equal size (Geisser 1975). One group is treated as the validation dataset, and the others combined as the training dataset. Then, for each value of λ, models are identified and their performance quantified multiple times in a process that involves each data point being used for validation once.

The second technique is the use of information criteria. Information criteria are closely related to regularised regression methods, being interpretable as likelihood methods that penalise the number of parameters in models. Again, models

for different λ are estimated and variables selected, then the best performing is chosen based on information criteria value. The Akaike Information Criterion (Akaike 1974) or the Bayesian Information Criterion (Schwarz 1978) may be used, along with their extensions (for small *n* / high *p* relative to *n*) the corrected AIC (AIC_c: Sugiura 1978) and the Extended BIC (EBIC: Chen & Chen 2008).

When producing the inverse propensity weights released with the UKHLS Covid-19 Study and in the main text of this paper, we use information criteria techniques to choose values of λ and identify models for estimating subject response probabilities. Specifically, we utilise the EBIC (in the Stata 16 package 'lassologit': see Ahrens et al. 2020), which is:

$$EBIC_{\xi}(\lambda) = n\,log\big(\hat{\sigma}^2(\lambda)\big) + df(\lambda)\,log(n) + 2\xi df(\lambda)\,log(p), \quad (2)$$

where $\hat{\sigma}^2(\lambda) = n - 1\ \sum_{i=1}^{n} \varepsilon_i^2$ and ε_i are the residuals. df is the effective degrees of freedom, the penalisation parameter common to all information criteria, and in this case is quantified as the number of coefficients estimated to be non-zero. $\xi[0,1]$ is a second penalisation parameter included in the EBIC to prevent over-selection of variables when *p* is relatively large, and is quantified as:

$$\xi = 1 - log(n)/\big(2\,log(p)\big) \quad (3)$$

We ultimately utilise EBIC techniques because in simulations Ahrens et al. (2020) show that in the majority of scenarios they perform best out of those mentioned earlier (all of which are available in the 'lassologit' package and its sister package 'lassopack': see Ahrens et al. 2020) in terms of model identification, that is, in terms of lowest rates of false positives (identifying predictors not correlated with the response variable) and

false negatives (not identifying actual correlates of the response variable). We note though, that some of the findings replicate earlier work: see, for example, Chen & Chen (2008) for simulations showing that the EBIC performs better than the BIC. Moreover, there are theoretical reasons why such findings might be expected. First, supporting their use for model identification, BIC techniques are the only ones of those tested that are model consistent, that is, will select the 'true' model (if in the potential set) with a probability nearing one as sample size tends to infinity (Yang 2005; Zhang et al. 2010). Second, when model identification is the goal, theory indicates that cross-validation training datasets should be small and validation datasets should be close to *n*, because more data are required to identify the correct model than to reduce bias and variance (Yang 2006). This does not occur with the *K*-fold cross-validation technique included in 'lassopack' (and in most other Lasso software packages: see, for example, StataCorp 2017), with which the training dataset is ~ n/K (see earlier). We note here that given this, intuitively at least the relatively small size of most survey datasets may preclude the use of more appropriate cross-validation techniques for identifying response probability models anyway: a sufficiently large evaluation dataset may lead to too small a training dataset for initial model selection to reliably take place.

As mentioned in the second paragraph of this section, for the above techniques to be used as described predictors must first be standardised so that they have unit variance. Hence, when modelling Covid-19 Study response probabilities we first converted all multi-category predictors and interactions into dummy variables. We also set up models so that the predictors 'Gender', 'Age' and 'Education' and their

interactions could not be removed during Lasso procedures, and included in the final selected predictor sets all dummy variables associated with Lasso-selected predictors: this approach reduced biases in weighted estimates compared to main survey wave 10 values (unpublished results). After model identification, we utilised post-Lasso OLS estimation to estimate subject response probabilities for weight calculation. This is because Lasso estimated coefficients are subject to attenuation bias (Ahrens et al. 2020). We fitted probit models including the Lasso-selected predictors, then computed estimated response probabilities using model coefficients and subject characteristics.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020) lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, 20*:* 176-235.

Akaike, H. (1974) A new look at the statistical model identification*. IEEE Transactions on Automatic Control* 19: 716–723.

Chen, J. & Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95: 759–771. DOI: 10.1093/biomet/asn034

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70: 320–328. DOI: 10.2307/2285815.

StataCorp (2017) Stata Lasso Reference Manual Release 17. StataCorp LLC, College Station, Texas.

Steyerberg, E.W., Eijkemans, M.J.C. & Habbema, J.D.F. (2001) Application of shrinkage techniques in Logistic regression analysis: a case study. *Statistica Neerlandica,* 55: 76–88. DOI: 10.1111/1467-9574.00157.

Sugiura, N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics—Theory and Methods* 7: 13–26. DOI: 10.1080/03610927808827599.

Tibshirani, R. (1996) Regression and shrinkage via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92: 937–950. DOI: 10.1093/biomet/ 92.4.937.

Yang, Y. (2006) Comparing learning methods for classification. *Statistica Sinica* 16: 635–657.

Zhang, Y., Li, Y. & Tsai, C-L. (2010) Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105: 312–323. DOI: 10.1198/jasa.2009.tm08013.

Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67: 301–320. DOI: 10. 1111/j.1467-9868.2005.00503.x.

Appendix B

We now sketch a framework in which the combined sampling design of UKHLS and its Covid-19 Study sits and use this to set out the conditions under which the SS and SIP-NR procedures we propose will give us valid first-order inference.

The structure is as follows:

- Appendix B.1 sets up notation and uses it to describe the changing population;
- Appendix B.2 describe the UKHLS main survey sample as a bipartite incidence graph sampling (BIGS) scheme.
- Appendix B.3 reviews the pragmatic assumptions made for UKHLS when making inference.
- Appendix B.4 finally comes to the Covid-19 Study and sets out the conditions under which the SS and SIP-NR procedures described by Moore and Clarke (2024, sec. 3) can be used to make valid inferences about the population at the time of the Covid-19 Study.

It should be noted that the development represents a simplification of the UKHLS and UKLS Covid-19 Study designs but these simplifications do not undermine the results.

B.1 BIGS Notation and Change in UKHLS Target Population

We take the UKHLS main survey at Waves 9 to be obtained by direct and indirect sampling (Lavallee 2007). This means it can also be characterized as a special case of Bipartite Incidence Graph Sampling (BIGS) scheme (e.g. Zhang 2022; Zhang and Patone

2017). Within the BIGS framework, the target population is characterised by $\mathcal{B} = \{F, \Omega, \mathcal{H}\}$ where, in our case, F is the primary (directly sampled) population, Ω is the secondary (indirectly sampled) population, and $(i\kappa) \in \mathcal{H}$ is the set of structural links in the population between pair $i \in \mathcal{F}$ and $\kappa \in \Omega$ such that the direct selection of i leads to forward/indirect selection of κ if (and only if) $(i\kappa) \in \mathcal{H}$, else $(i\kappa) \notin \mathcal{H}$.

To characterise the clusters formed by the links in $\mathcal{H}$, define also the set $\alpha(i) = \{\kappa: (i\kappa) \in \mathcal{H}\}$ of individuals who would be indirectly selected were $i \in \mathcal{F}$ selected into the sample, and the set $\beta(\kappa) = \{i: (i\kappa) \in \mathcal{H}\}$ of those in $\mathcal{F}$ who, if directly selected, induce the inclusion of $\kappa \in \Omega$ in the sample. The number of units in these sets is denoted by $|\alpha(i)|$ and $|\beta(\kappa)|$, respectively.

But before framing our sampling design as a BIGS scheme, we describe the evolution of the population between UKHLS Wave 1 (incorporating the Ethnic Minority Boost Sample) and the time of the Covid-19 Study

Let $\mathcal{P}^0$ be the UK population at UKHLS Wave 1 or baseline $t = 0$. The available Postcode Address File (PAF) determines the partition $\mathcal{P}^0 = \mathcal{T}^0 \cup \bar{\mathcal{T}}^0$, where $\mathcal{T}^0$ is the subpopulation with non-zero selection probabilities and $\bar{\mathcal{T}}^0$ the left-out individuals whose selection probabilities are zero.

Now let $\mathcal{P}^1$ be the population at Wave 6 the time of the Immigrant and Ethnic Minority Boost (IEMB) sample. The available PAF at Wave 6 and focus on areas of high minority ethnic densities allows us to modify the partition to be $\mathcal{P}^1 = \mathcal{T}^1 \cup \bar{\mathcal{T}}^1$ where (ignoring survival for now) $\mathcal{T}^1$ includes $\mathcal{T}^0$ and all those whose selection probabilities are now

non-zero as a result of the IEMB design; likewise, $\bar{\mathcal{T}}^1$ includes those people in $\bar{\mathcal{T}}^0$ and those population newcomers $\mathcal{N}^1$ whose selection probabilities remain non-zero.

Note that all population sets are taken to include both eligible and ineligible individuals: it is assumed that the focus on eligible individuals will be done by the analyst through sample exclusions or zero weights. The weight construction Moore and Clarke (2024), for example, includes only eligible adults.

Now denote the time of Wave 9 as $t = 2$ (to simplify without loss of generality, we ignore that Wave 8 respondents who did not respond at Wave 9 were also included). The waves are combined without affecting the subsequent arguments. The cross-sectional population can be written

$$\mathcal{P}^2 = \mathcal{T}^1(2) \cup \bar{\mathcal{T}}^1(2) \cup \mathcal{N}^2, \quad \text{(A. 1)}$$

where $\mathcal{T}^1(2)$ and $\bar{\mathcal{T}}^1(2)$ are the 'survivors' from $\mathcal{T}^1$ and $\bar{\mathcal{T}}^1$ present at $t = 2$, respectively, and $\mathcal{N}^2$ is the population of newcomers present at $t = 2$ who were not present at $t = 1$. A survivor is taken to be someone who does not die or does not leave the UK.

Finally, denote the time of the Covid-19 Study by $t = 3$. Dropping the superscripts and subscripts for quantities related to $t = 3$, the cross-sectional population is

$$\mathcal{P} = \mathcal{T}^1(3) \cup \bar{\mathcal{T}}^1(3) \cup \mathcal{N}^2(3) \cup \mathcal{N}, \quad \text{(A. 2)}$$

where $\mathcal{T}^1(3)$, $\bar{\mathcal{T}}^1(3)$ and $\mathcal{N}^2(3)$ are survivors present at $t = 3$, and $\mathcal{N}$ is the population of newcomers present at $t = 3$ not present at $t = 2$.

B.2 UKHS Wave 9 as a BIGS Sample

We begin by formulating UKHLS Wave 9 ($t = 2$) as BIGS design $\mathcal{B} = \{F, \Omega, \mathcal{H}\}$ and relate its components to the target population.

A. F comprises everyone in $\mathcal{T}^1$ who survived and subsequently complied with UKHLS such that their survey weights are available at $t = 2$ or would have counterfactually complied had they been selected instead.

B. Ω comprises everyone in $\mathcal{T}^1$ who survived but have no survey weight available at $t = 2$ because they did not comply with UKHLS or would have counterfactually non-complied had they been selected instead. Furthermore, it includes those individuals in $\bar{\mathcal{T}}^1(2) \cup \mathcal{N}^2$ who are now part of households containing individuals in $\mathcal{T}^1$ according to the PAF.

C. $\mathcal{H}$ contains the links induced by the household structure at $t = 2$ between $i \in F$ and $\kappa \in \Omega$.

The target population is straightforwardly $\mathcal{F} \cup \Omega \subseteq \mathcal{P}$. This excludes those individuals in $\bar{\mathcal{T}}^1(2) \cup \mathcal{N}^2$ not in households containing individuals in $\mathcal{T}^1$.

Now the sample can be defined as follows:

D. $s \subset \mathcal{F}$ contains the (directly sampled) individuals with survey weights available for analysis.

E. $\Omega_s \subset \Omega$ contains is the corresponding sample of (indirectly sampled) individuals without survey weights available.

F. $\mathcal{H}_s \subset \mathcal{H}$ is the set of links between sample members in s and those in Ω_s.

Finally, note from section 2.4 that $s \cup \Omega_s$ for UKHLS also excludes (i) OSMs with incomplete wave response, (ii) PSMs, and (iii) non-(minority) ethnic TSMs from the IEMB sample if they are not resident in HHs containing a weight to be shared, despite being in $\mathcal{H}_s \subset \mathcal{H}$. These individuals will figure in appendix A.4 if they survive.

B.3 Shared Weight Estimation

Before moving on to set up the SS and SIP-NR procedures introduced by Moore and Clarke (2024), we review the implicit assumptions behind the use of the main survey weights for analysing survey variable(s) Y at $t = 2$.

Letting random variable S_i indicate whether unit $i \in \mathcal{F}$ appears in s, design-based/finite-population inference about the population total $T_2 = \sum_{i\in\mathcal{F}} y_i + \sum_{\kappa\in\Omega} y_\kappa$ for any survey variable Y can be based on

$$t_2 = \sum_{i\in\mathcal{F}} w_i y_i S_i + \sum_{\kappa\in\Omega} w_\kappa\, y_\kappa S_\kappa, \quad (\mathrm{A}.3)$$

where w_i is the survey weight for unit i and, for unit κ, indirect selection indicator $S_\kappa = \sum_{i\in\beta(\kappa)} S_i/|\beta(\kappa)|$ and shared weight $w_\kappa = \sum_{i\in\beta(\kappa)} w_i/|\beta(\kappa)|$.

Similarly, for model-based super-population inference, let $\psi_j(Y_j;\theta)$ be the score function based on the statistical model chosen by the analyst. The score is taken to

satisfy $E\{\psi_j(Y_j;\theta)\} = 0$ where θ is the true parameter value and expectation is with respect to the true infinite-population model. Without loss of generality, we focus on the target population's mean $\theta = \mu_2 = E(Y_j)$ based on the simplest possible score $\psi_j(Y_j;\mu_2) = Y_j - \mu_2$. The pseudolikelihood estimator $\hat{\mu}$ for μ_2 is then chosen such that

$$\sum_{i\in s} w_i(Y_j - \hat{\mu}) + \sum_{\kappa\in\Omega_s} w_\kappa\,(Y_\kappa - \hat{\mu}) = 0. \quad \text{(A.4)}$$

Both (A.3) and (A.4) will be unbiased for respective parameters T_2 and μ_2 if the true survey weights are available for analysis (generally, only consistency holds but is unbiased for scores linear in parameters). The pragmatic approach to variance estimation for (A.3) and (A.4) is based on the following assumption:

> Assumption A.1: Survey weights w_i treated as known rather than estimated quantities and the shared weights w_κ can be treated as survey weights.

The general form of variance estimator for (A.3) is complex and beyond the scope of this paper even under Assumption A.1. However, some modification to the form of (A.3) and (A.4) in terms of clusters can be used to simplify calculations as follows:

Clusters partition the population $\mathcal{F} \cup \Omega$ indexed by $c \in \{1, \dots, C\}$ defined as follows:

> Definition A.1: The target population can be partitioned into C clusters $\mathcal{F} \cup \Omega = \cup_{c=1}^{C} \gamma_c$ such that cluster c contains a) all i satisfying $\alpha(i) = \alpha_c$, and b) all κ satisfying $\beta(\kappa) = \beta_c$, and $\gamma_c = \alpha_c \cup \beta_c = \{i \in \mathcal{F}: \alpha(i) = \alpha_c\} \cup \{\kappa \in \Omega: \beta(\kappa) = \beta_c\}$.

Now, for example, (A.3) can be rewritten in terms of clusters as

$$t_2^{clus} = \sum_{c=1}^{C} S_c w_c \left\{ \sum_{i \in \beta_c} y_i + \sum_{\kappa \in \alpha_c} y_\kappa \right\} = \sum_{c=1}^{C} S_c \sum_{j \in \gamma_c} y_j w_j^*, \quad \text{(A. 5)}$$

where random variable S_c indicates whether cluster c is selected, $w_j^* = w_c$ if $j \in \gamma_c$ and $w_c = 1/\pi_c$ is the weight based on the selection probability $\pi_c = \Pr(S_c = 1)$ for direct selection (and response) of the units in β_c (a short discussion of how it is calculated is given below).

The variance formula can then be based on the design $\mathcal{D}$ induced by the UKHLS sampling design (incorporating the main and boost surveys) with the subsequent response process treated as additional stages of selection with known selection probabilities. Model-based inference can be based on the linearized estimator.

However, the challenge for analysts is that, strictly speaking, they need to derive and calculate π_c. The simplest approach to this is to assume that households remain intact and linked to the PAF addresses in which case π_c can be based on the original household selection and household non-response probabilities.

B.4 Covid-19 Study Inference

The population $\mathcal{P}$ at $t = 3$ (the time of the Covid-19 Study) is defined in (A. 2). Sample selection involves inviting each $i \in s$ and $\kappa \in \Omega_s$ to participate online at each wave. The additional complication comes from the appearance of unweighted 'left-out' individuals $l \notin s \cup \Omega_s$ in the Covid-19 sample. Despite knowing that these new individuals were at some point members of HHs containing UKHLS members, no information on HH membership is available for them because no attempt was made

to create $\mathcal{H}_s$ at $t = 3$. Hence, in contrast to appendix A.2, inference cannot be based on the theoretical results of indirect sampling or BIGS sampling.

Instead, we propose a matching estimators based on an exchangeability assumption to be set out below. Let s^* denote the set of extra-sample linked individuals. Each $l \in s^* \subset \mathcal{L}$ falls into one of the following three categories:

1. Individuals in Ω but not Ω_s who are now included by virtue of being members of households containing at least one Covid-19 respondent from $s \cup \Omega_s$.
2. Individuals in $\mathcal{F}$ but not s who are now included by virtue of being members of households containing at least one Covid-19 respondent from $s \cup \Omega_s$.
3. Surviving left-out individuals and newcomers $\mathcal{L} \subset \bar{\mathcal{T}}^1(3) \cup \mathcal{N}^2(3) \cup \mathcal{N}$ (from (A.2)) who are included by dint of being members of households containing at least one Covid-19 respondent in $s \cup \Omega_s$. As described at the end of appendix A.2 for UKHLS, the set of unweighted individuals also includes those present prior to wave 9 for whom it is not possible to share a weight using conventional sharing procedures (or for whom it would have counterfactually not been possible to share a weight).

Note that 1-3 imply that the target population excludes left-out individuals and newcomers not in $\mathcal{L}$, that is, left-out and newcomer individuals who are not in households containing at least one Covid-19 respondent from $\mathcal{F} \cup \Omega$: the probability of being included in the Covid-19 Study is zero for these people.

Non-response to the Covid-19 Study

All those for whom the contact details were available to UKHLS (even those without weights and even if details about HH links were not available about them) at the time of the Covid-19 Study were invited to participate in the Covid-19 Study. The impact of refusals and no-replies is ignored in the following development.

Subsequently, at the start of each wave of the Covid-19 Study, those who agreed to participate were sent a request and a link to the online questionnaire. In the following development, we take the wave in question to be wave 1 without loss of generality.

The response is assumed to satisfy the following assumption:

> Assumption A.2 (Missing at Random Questionnaire Non-response): For any survey variable Y, the probability that individual $i \in \mathcal{F} \cup \Omega \cup \mathcal{L}$ fills in the online questionnaire depends on the survey variables such that $\Pr(R_i = 1|Y_i = y_i) > 0$ depends non-trivially on y_i. However, variables Z can be found satisfying $\Pr\left(R_i = 1\middle|Y_j = y_i, Z_i = z_i\right) = \Pr\left(R_i = 1\middle|Z_i = z_j\right) = \rho_i$ exist and are known by and available to the analyst.

Assumption A.2 allows us to estimate the response propensities provided suitable variables are available. It also takes the response process to be the same across the three subpopulations $\mathcal{F}$, Ω and $\mathcal{L}$.

The problem

Incorporating $l \in s^*$ into the estimator requires further assumptions. Consider the following biased estimator of the population total $T = \sum_{i\in\mathcal{F}} y_i + \sum_{\kappa\in\Omega} y_\kappa + \sum_{l\in\mathcal{L}} y_l$ of Covid-19 Study variable Y (note this is different to the Y in appendix A.3 which was from the main survey):

$$t^{biased} = \sum_{i\in\mathcal{F}} w_i y_i S_i R_i + \sum_{\kappa\in\Omega} w_\kappa y_\kappa S_\kappa R_\kappa + \sum_{l\in\mathcal{L}} y_l L_l R_l,$$

where S_i, S_κ, w_i and w_κ are defined as in (A.3), R_i, R_κ and R_l are response indicators for whether individuals i, κ and l, respectively, fill in the questionnaire, and $L_l = I(l \in s^*)$ indicates whether $l \in \mathcal{L}$ is selected into the unweighted left-out sample s^* (indicator function $I(E) = 1$ if event E is true or zero otherwise).

Estimator t^{biased} is appropriately named because the weights included in it do not adjust for non-random response to the invitation, and the unweighted sample individuals require a weight because they were not selected using simple random sampling.

This motivates the 'unbiased' estimators of the form

$$t = \sum_{i\in\mathcal{F}} y_i w_i^* S_i R_i + \sum_{\kappa\in\Omega} y_\kappa w_\kappa^* S_\kappa R_\kappa + \sum_{l\in\mathcal{L}} y_l \widehat{w}_l^* L_l R_l, \quad \text{(A. 6)}$$

for the population total, and estimator $\hat{\mu}$ of $\mu = E(Y)$ the solution to

$$\sum_{i\in s} w_i^* (Y_j - \hat{\mu}) + \sum_{\kappa\in\Omega_s} w_\kappa^* (Y_\kappa - \hat{\mu}) + \sum_{l\in s^*} \widehat{w}_l^* (Y_l - \hat{\mu}) = 0, \quad \text{(A. 7)}$$

where, in section3, w_j is referred to as a selection weight, $w_j^* = w_j/\rho_j$ as an IP-NR weight, $\widehat{w}_l^* = w_l/\rho_l$ and $\widehat{w}_l$ is an estimator of the unknown true survey weight $\widetilde{w}_l$.

An alternative approach can be based on the simultaneous weighting approach proposed by Robbins et al. (2021, sec. 2.1.2). They suppose sample $s \cup \Omega_s = S_1$ is drawn from the target population with known probabilities, and $s^* = S_2$ is drawn from the same population with unknown probabilities. Their simultaneous propensity score weight for $i \in S_1 \cup S_2$ is

$$w_i^* = (1 - \gamma_i)\, w_i/\rho_i\,, \quad \text{(A.8)}$$

where we set $\gamma_i = \Pr(i \in S_2 | i \in S_1 \cup S_2) = 0.5$. For $l \in S_2 = s^*$, the selection weight w_l must also be 'estimated' as in (A.7).

The SS and SIP-NR procedures are alternative ways of estimating $\widetilde{w}_l$.

Matching

Both SS and SIP-NR procedures described in section 3 involve splitting the selection or IP-NR weight for $j \in s \cup \Omega_s$ with $l \in s^*$. Splitting the response propensities is straightforwardly allowed under assumption A.2. However, splitting the selection weights requires a further assumption.

> Assumption A.3 (Exchangeable ignorable selection): The analyst chooses variables Z such that, for any pair $l \in \mathcal{L}$ and $j \in \mathcal{F} \cup \Omega$, if $Z_l = Z_i = z$ then

$$\Pr(L_l = 1 | Y_l, Z_l = z) = \Pr(L_l = 1 | Z_l = z) = \Pr\big(S_j = 1 \big| Z_j = z\big)$$
$$= \Pr\big(S_j = 1 \big| Y_j, Z_j = z\big)$$

That is, had $l \in \mathcal{L}$ counterfactually been included in the PAF at Waves 1 or 6, it would have the same UKHLS compliance behaviour as those actually included with the same auxiliary variable characterisation.

For example, (A.7) with known weights implies $E(\hat{\mu}) = \mu$ because $E\{w_i^*(Y_j - \mu)S_iR_i\} = E\{w_\kappa^*(Y_\kappa - \mu)S_\kappa R_\kappa\} = 0$, straightforwardly, and

$$E\{w_l^*(Y_l - \mu)S_lR_l\} = E_Z\left\{\frac{\Pr(L_l = 1|Z)}{\Pr(S_l = 1|Z)}E(Y - \mu|L = 1, Z)\right\} = 0,$$

where the final equality holds under assumptions A.1, A.2 and A.3 (assumption A.2 implies that the response process is the same for all three subpopulations and assumption A.3 that $E(Y_l - \mu|L_l = 1, Z) = E(Y_l - \mu|Z)$).

Similar arguments also hold for (A.7) in combination with (A.8) based on the simultaneous propensity score weights.

Inference

Restricting the discussion to model-based inference based on (A.7) and its generalisation, arguments given elsewhere (e.g. Chernozhukov et al. 2018, C3-C5) can be used demonstrate that the use of $\widehat{w}_l$ would lead to a small over-estimation of the standard error of $\hat{\mu}$.

To argue this, it is first necessary to make the further assumption that the matching estimator is a good predictor in large sample sizes.

Under assumption A.2, the response propensities ρ can be estimated from the available data on $j \in s \cup \Omega_s$ or on $j \in s \cup \Omega_s \cup s^*$. Treating $\hat{\rho}_j$ as the true response probability is well known to *over*-estimate the standard errors (e.g. discussion of the 'IPWCC' estimator by Tsiatis (2006, p.206)) and so is used in practice: conservative inference is worth the price of simplicity.

Moreover, in terms of the splitting, we require the following assumption to hold:

> Assumption 4 (Good estimator): Matching leads to a regular estimator satisfying
>
> $$\widehat{w}_l = \Pr(S_l = 1 | Z_l = z) + o_p\left(n^{-1/2}\right), \quad \text{(A. 9)}$$
>
> where $n = \sum_{j \in s \cup \Omega_s} R_j$ and $o_p\left(n^{-1/2}\right)$ represents omitted variables converging to zero at rate $\sqrt{n}$.

Setting $\widehat{w}_l = w_l(Z; \hat{\eta})$, a Taylor series expansion of $\psi_l(Y_l;\, \hat{\mu}, \hat{\eta}) = (Y_l - \hat{\mu}) w_l(Z; \hat{\eta})$ around μ and $w_l(Z_l; \eta) = \Pr(S_l = 1|Z)$ depends on the additional term $n^{-1} \sum \partial w_l(Z; \eta)/\partial \eta \, (\hat{\eta} - \eta)$ if $\widehat{w}_l$ is treated as an estimated parameter. If $\hat{\eta}$ was estimated on a completely independent data sample, this additional term would be $o_p\left(n^{-1/2}\right)$ too because of (A.9) and so the asymptotic distribution of $\hat{\mu}$ would be unaffected. However, we know that $\hat{\eta}$ is, in effect, estimated from data on donor individuals known to be clustered with those being donated to. This situation is less acute than that for $\hat{\rho}_j$, which is estimated from exactly the same sample individuals as those it is used for and, as already discussed, leads to an (in practice acceptable) over-estimation of the standard errors, but we can expect not accounting for matching-estimator imprecision to contribute less to standard-error over-estimation.

Note that the above argument is based on treating of $\psi(Y;\hat{\mu},\hat{\eta})$ as independent and identically distributed random variables, but the same conclusion follows for inference based on the design-robust linearized variance estimator.

References

Chernozhukov, V., Chetzverikov, D., Demirer M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21, C1-C68.

Moore, J.C. and Clarke, P.S. (2024) Two new solutions to the zero non-response weight problem. Submitted manuscript.

Robbins, M.W., Ghosh-Dastidar, B. and Ramchand, R. (2021) Blending probability and non-probability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology* 9: 1114-1145.

Tsiatis, A.A. (2006) *Semiparametric Theory and Missing Data*. Springer: London.

Zhang, L.-C. (2022) *Graph Sampling*. CRC Press: Abingdon.

Zhang, L.-C. and Patone, M. (2017) Graph sampling. *Metron*, 75: 277-299.

Appendix C: Evaluation results.

Table 1: COVID-19 Study wave 5 and wave 8 cross-sectional weight performance in recovering estimated means of main survey measured characteristics. We present main survey weighted mean estimates ('wt. est.'; respectively columns (i) & (iv)); tests of differences between such estimates and estimates given COVID-19 Study IP-NR weights (columns (ii) & (v)); and tests of differences between such estimates and estimates given COVID-19 Study SS weights (columns (iii) & (vi)). * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$.

Table 2: COVID-19 Study wave 5 and wave 8 cross-sectional weight performance in recovering estimated means of COVID-19 Study wave measured characteristics. We present COVID-19 Study IP-NR weighted estimated means ('wt. est.'; respectively columns (i) & (iv)), tests of differences between such estimates and estimates given Mod1 weights ('wt. diff.'; columns (ii) & (v)), and tests of differences between such estimates and estimates given SIP-NR weights (columns (iii) & (vii)). * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$. Note that differences exist between the two sets of main survey weighted estimates are due to a greater number of subject deaths by wave 8.

Table 3: COVID-19 Study wave 2 and wave 5 longitudinal weight performance in recovering estimated means of main survey measured characteristics. We present main survey weighted mean estimates ('wt. est.'; respectively columns (i) & (iv)); tests

of differences between such estimates and estimates given COVID-19 Study IP-NR weights (columns (ii) & (v)); and tests of differences between such estimates and estimates given COVID-19 Study SS weights (columns (iii) & (vi)). * equals *P*<0.05, ** equals *P*<0.01, *** equals *P*<0.001.

Table 4: COVID-19 Study wave 2 and wave 5 longitudinal weight performance in recovering estimated means of COVID-19 Study wave measured characteristics. We present COVID-19 Study IP-NR weighted estimated means ('wt. est.'; respectively columns (i) & (iv)), tests of differences between such estimates and estimates given SS weights ('wt. diff.'; columns (ii) & (v)), and tests of differences between such estimates and estimates given SIP-NR weights (columns (iii) & (vii)). * equals *P*<0.05, ** equals *P*<0.01, *** equals *P*<0.001. Note that differences exist between the two sets of main survey weighted estimates are due to a greater number of subject deaths by wave 5.

Table 1

	Wave 5 cross-sectional			Wave 8 cross-sectional		
	Main	Covid		Main	Covid	
		IP-NR	SS		IP-NR	SS
	wt est.	wt diff.	wt. diff	wt est.	wt diff.	wt. diff
	(i)	(iii)	(iii)	(iv)	(v)	(vi)
In IPW model:						
Subjective financial situation (SFS): comfortable or OK	0.717 (0.003)	-0.002	0.012*	0.716 (0.003)	0.005	0.014**
SFS: just about getting by	0.203 (0.002)	-0.005	-0.013**	0.203 (0.002)	-0.008	-0.012**
SFS: finding it quite / very difficult	0.081 (0.002)	0.007*	0.001	0.081 (0.002)	0.003	-0.002
Tenure: Owned	0.343 (0.003)	0.011*	0.029***	0.343 (0.003)	0.010	0.028***
Tenure: Mortgage	0.341 (0.003)	-0.024***	-0.021***	0.341 (0.003)	-0.023***	-0.020***
Tenure: Rented	0.119 (0.002)	0.000	-0.008*	0.119 (0.002)	0.003	-0.007
Tenure: Social Housing	0.195 (0.002)	0.012*	0.000	0.195 (0.002)	0.009*	-0.001
Low skill occupation	0.362 (0.004)	0.003	0.001	0.362 (0.004)	0.005	0.002
Savings income?	0.372 (0.003)	-0.002	0.012*	0.372 (0.003)	-0.000	0.013*
Behind with some or all bills	0.059 (0.001)	0.002	-0.009**	0.059 (0.001)	0.000	-0.008**
Not in IPW model:						
Income poverty	0.155 (0.002)	0.015***	0.013**	0.155 (0.002)	0.006	0.004
Receives core benefit	0.054 (0.001)	0.001	-0.002	0.054 (0.001)	-0.001	-0.004
Visited GP	0.700 (0.003)	-0.005	-0.003	0.700 (0.003)	-0.004	-0.002
Smoker	0.144 (0.002)	0.030***	0.020***	0.144 (0.002)	0.026***	0.019***
Hospital outpatient	0.433 (0.003)	-0.005	0.001	0.433 (0.003)	-0.005	-0.002

Table 2

	Wave 5 cross-sectional			Wave 8 cross-sectional		
	IP-NR	SS	SIP-NR	IP-NR	SS	SIP-NR
	wt est.	wt. diff	wt diff.	wt. est.	wt diff.	wt. diff
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
In IPW model:						
Subjective financial situation (SFS): comfortable or OK	0.752 (0.004)	0.012*	-0.001	0.751 (0.004)	0.009	-0.005
SFS: just about getting by	0.190 (0.004)	-0.008	-0.002	0.194 (0.004)	-0.003	0.003
SFS: finding it quite/very difficult	0.058 (0.002)	-0.004	0.003	0.055 (0.002)	-0.005	0.002
Employed?	0.594 (0.005)	-0.011	-0.001	0.600 (0.005)	-0.014*	-0.000
Behind with some or all bills	0.067 (0.002)	-0.009**	0.003	0.060 (0.002)	-0.006	0.003
Behind with housing payments	0.065 (0.002)	-0.006	0.001	0.076 (0.004)	-0.005	0.002
HH type: Couple with children	0.209 (0.004)	-0.004	-0.011*	0.197 (0.004)	-0.006	-0.009
HH type: Single, no children	0.113 (0.003)	-0.003	0.003	0.111 (0.003)	-0.005	0.002
Covid test?	0.124 (0.003)	-0.001	0.001	0.278 (0.004)	-0.002	-0.011
Clinically vulnerable	0.415 (0.005)	0.009	0.008	0.436 (0.005)	0.006	0.012
Not in IPW model:						
Advised to shield	0.072 (0.003)	0.003	0.004	0.088 (0.003)	-0.001	0.003
Gave or received money	0.119 (0.003)	-0.001	0.001	0.122 (0.003)	0.001	0.004
Less sleep than usual	0.185 (0.004)	-0.001	-0.001	0.200 (0.004)	-0.002	0.002
More depressed than usual	0.221 (0.004)	-0.002	-0.002	0.250 (0.004)	-0.003	-0.000
More lonely than usual	0.062 (0.002)	-0.005	-0.003	0.082 (0.003)	-0.003	-0.002

Table 3

	Wave 2 longitudinal			Wave 5 longitudinal		
	Main	Covid		Main	Covid	
		IP-NR	SS		IP-NR	SS
	wt est.	wt diff.	wt. diff	wt est.	wt diff.	wt. diff
	(i)	(iii)	(iii)	(iv)	(v)	(vi)
In IPW model:						
Subjective financial situation (SFS): comfortable or OK	0.716 (0.003)	0.013*	0.023***	0.717 (0.003)	0.000	0.011*
SFS: just about getting by	0.203 (0.002)	-0.014**	-0.020***	0.203 (0.002)	-0.011*	-0.017***
SFS: finding it quite / very difficult	0.080 (0.002)	0.001	-0.003	0.080 (0.002)	0.010**	0.005
Tenure: Owned	0.344 (0.003)	0.008	0.023***	0.344 (0.003)	-0.002	0.016**
Tenure: Mortgage	0.340 (0.003)	-0.018***	-0.016**	0.340 (0.003)	-0.019**	-0.019***
Tenure: Rented	0.119 (0.002)	0.007	-0.002	0.119 (0.002)	0.003	-0.005
Tenure: Social Housing	0.195 (0.002)	0.003	-0.005	0.195 (0.002)	0.016***	0.007
Low skill occupation	0.362 (0.004)	-0.002	0.001	0.362 (0.004)	0.013	0.010
Savings income?	0.372 (0.003)	-0.001	0.010	0.372 (0.003)	-0.021***	-0.010
Behind with some or all bills	0.059 (0.001)	-0.006*	-0.012***	0.059 (0.001)	0.002	-0.007**
Not in IPW model:						
Income poverty	0.155 (0.002)	0.011**	0.009*	0.155 (0.002)	0.018***	0.018***
Receives core benefit	0.054 (0.001)	0.006*	0.005	0.054 (0.001)	0.009**	0.006*
Visited GP	0.700 (0.003)	0.003	0.001	0.700 (0.003)	0.002	0.002
Smoker	0.144 (0.002)	0.019***	0.015***	0.144 (0.002)	0.030***	0.023***
Hospital outpatient	0.433 (0.003)	0.003	0.005	0.433 (0.003)	-0.000	0.002

Table 4:

	Wave 2 longitudinal			Wave 5 longitudinal		
	IP-NR	SS	SIP-NR	IP-NR	SS	SIP-NR
	wt est.	wt. diff	wt diff.	wt. est.	wt diff.	wt. diff
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
In IPW model:						
Subjective financial situation (SFS): comfortable or OK	0.744 (0.004)	0.007	-0.003	0.738 (0.005)	0.011	-0.003
SFS: just about getting by	0.194 (0.004)	-0.005	0.000	0.206 (0.004)	-0.006	0.002
SFS: finding it quite/very difficult	0.062 (0.002)	-0.002	0.002	0.056 (0.002)	-0.005	0.001
Employed?	0.608 (0.005)	-0.006	0.005	0.596 (0.005)	-0.008	0.000
Behind with some or all bills	0.088 (0.003)	-0.006	0.003	0.075 (0.003)	-0.009*	0.003
Behind with housing payments	0.075 (0.003)	-0.001	0.002	0.068 (0.003)	-0.004	0.001
HH type: Couple with children	0.202 (0.004)	-0.003	-0.012*	0.201 (0.004)	0.001	-0.010
HH type: Single, no children	0.117 (0.003)	-0.004	0.003	0.114 (0.003)	-0.005	-0.000
Covid test?	0.041 (0.002)	0.001	-0.000	0.116 (0.003)	0.000	0.000
Clinically vulnerable	0.385 (0.005)	0.005	0.006	0.407 (0.005)	0.004	0.004
Not in IPW model:						
Advised to shield	0.066 (0.002)	0.001	0.002	0.065 (0.003)	0.002	0.003
Gave or received money	0.153 (0.003)	-0.002	0.003	0.118 (0.003)	-0.004	-0.000
Less sleep than usual	0.216 (0.004)	-0.003	-0.002	0.183 (0.004)	0.000	-0.001
More depressed than usual	0.284 (0.004)	-0.002	0.003	0.220 (0.004)	-0.005	-0.000
More lonely than usual	0.078 (0.003)	-0.003	-0.002	0.062 (0.003)	-0.005	-0.006

Figure 1: Box plots of absolute values of the tests of COVID-19 Study weights reported in Appendix Table 1, standardised by benchmark estimate standard deviations. In a), tests compare wave 5 cross-sectional dataset IP-NR (white bars) and SS (light grey bars) weighted estimates of main survey measured characteristics to main survey weighted benchmarks. In b), tests compare wave 8 cross-sectional dataset IP-NR and SS weighted estimates of main survey measured characteristics to main survey weighted benchmarks. In plots, bars indicate the inter-quartile range, the line within the median value, and the cross the mean value. Whiskers indicate minimum / maximum values, unless values exist that are smaller or larger than the inter-quartile range, in which case they indicate the smallest / largest value within this range, and the outlying values are indicated by circles.

Figure 2: Box plots of absolute values of the tests of COVID-19 Study weights reported in Appendix Table 2, standardised by benchmark estimate standard deviations. In a), tests compare wave 5 cross-sectional dataset SS (light grey bars) and SIP-NR (dark grey bars) weighted estimates of COVID-19 Study measured characteristics to COVID-19 Study IP-NR weighted benchmarks. In b), tests compare wave 8 cross-sectional dataset SS and SIP-NR weighted estimates of COVID-19 Study measured characteristics to COVID-19 Study IP-NR weighted benchmarks. In plots, bars indicate the inter-quartile range, the line within the median value, and the cross the mean value. Whiskers indicate minimum / maximum values, unless values exist that are smaller or larger than the inter-quartile range, in which case they indicate the smallest / largest value within this range, and the outlying values are indicated by circles.

Figure 3: Box plots of absolute values of the tests of COVID-19 Study weights reported in Appendix Table 3, standardised by benchmark estimate standard deviations. In a), tests compare wave 2 longitudinal dataset IP-NR (white bars) and SS (light grey bars) weighted estimates of main survey measured characteristics to main survey weighted benchmarks. In b), tests compare wave 5 longitudinal dataset IP-NR and SS weighted estimates of main survey measured characteristics to main survey weighted benchmarks. In plots, bars indicate the inter-quartile range, the line within the median value, and the cross the mean value. Whiskers indicate minimum / maximum values, unless values exist that are smaller or larger than the inter-quartile range, in which case they indicate the smallest / largest value within this range, and the outlying values are indicated by circles.

Figure 4: Box plots of absolute values of the tests of COVID-19 Study weights reported in Appendix Table 4, standardised by benchmark estimate standard deviations. In a), tests compare wave 2 longitudinal dataset SS (light grey bars) and SIP-NR (dark grey bars) weighted estimates of COVID-19 Study measured characteristics to COVID-19 Study IP-NR weighted benchmarks. In b), tests compare wave 5 longitudinal dataset SS and SIP-NR weighted estimates of COVID-19 Study measured characteristics to COVID-19 Study IP-NR weighted benchmarks. In plots, bars indicate the inter-quartile range, the line within the median value, and the cross the mean value. Whiskers indicate minimum / maximum values, unless values exist that are smaller or larger than the inter-quartile range, in which case they indicate the smallest / largest value within this range, and the outlying values are indicated by circles.

Fig 1

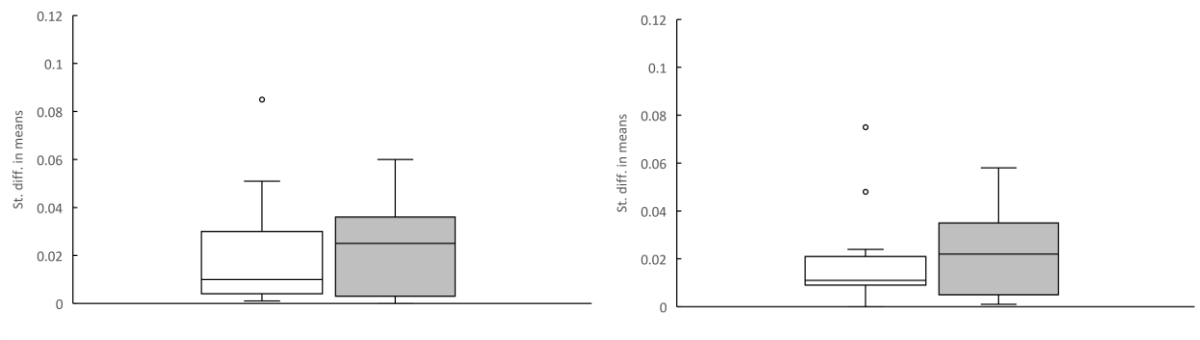

St. diff. in means
0.12
0.1
0.08
0.06
0.04
0.02
0
St. diff. in means
0.12
0.1
0.08
0.06
0.04
0.02
0

Fig 2

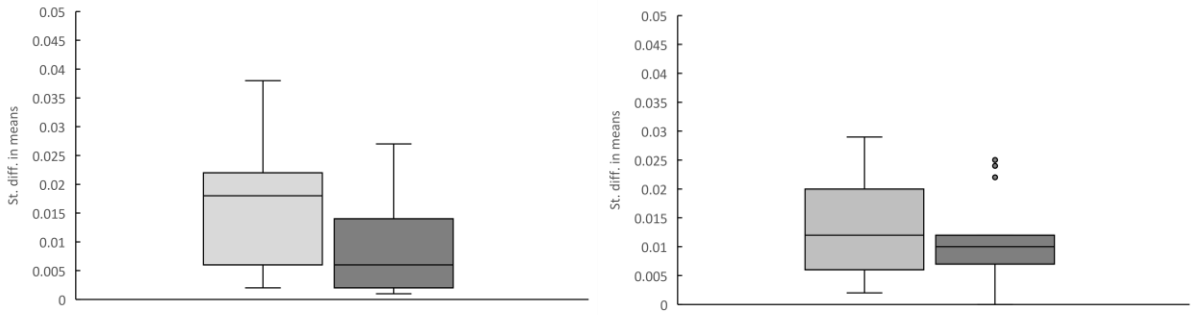

St. diff. in means
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
St. diff. in means
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0

Fig. 3

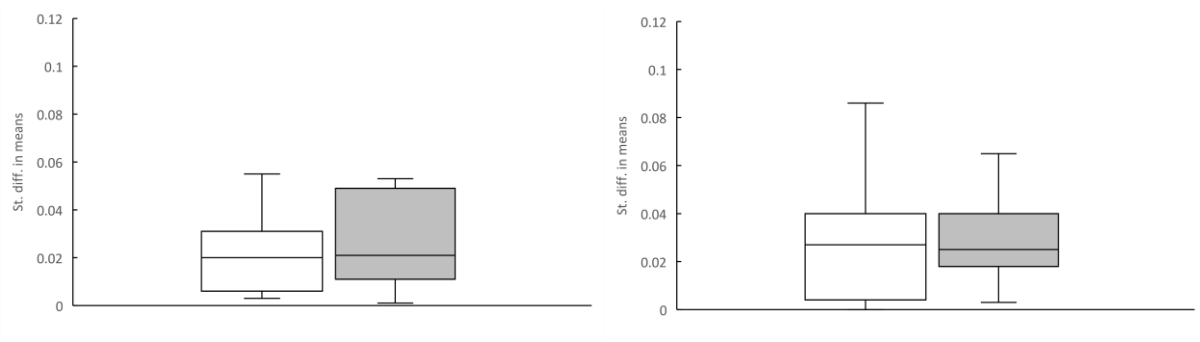
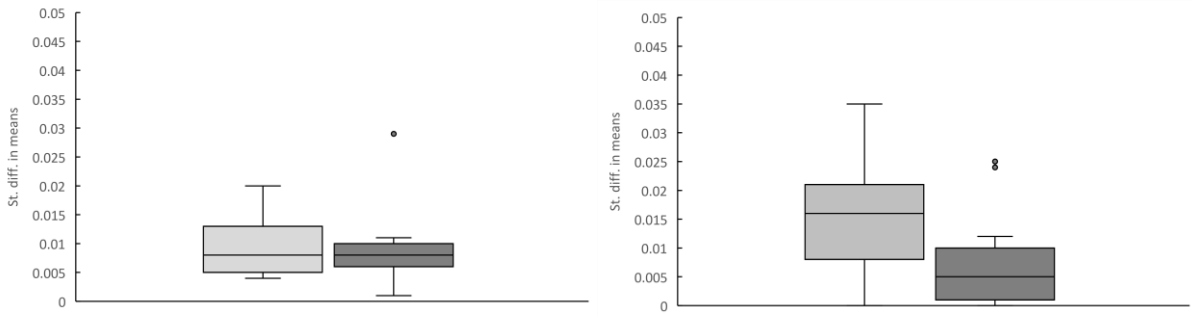

St. diff. in means
0.12
0.1
0.08
0.06
0.04
0.02
0
St. diff. in means
0.12
0.1
0.08
0.06
0.04
0.02
0

Fig. 4


St. diff. in means
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
St. diff. in means
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0